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

Report No. 6 – 28 July 2021

Summary:

- Update on mobilisation of raw reads, now totaling sequencing data sets from 872,011 viral raw read sets from 69 countries, a 28% 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 66,993 samples of the total 106,732 have been processed so far.

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 (10 Jul 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	10 July 2021
Raw data sets	Total	301,378	354,106	438,112	525,348	552,185	679,693	872,011
	Illumina	255,431	302,409	367,462	446,375	469,142	575,481	703,104
	Oxford Nanopore	45,222	50,972	69,921	77,913	81,466	93,581	106,732
	Other	725	725	729	1,060	1,577	7,134	62,175
Source countries for raw data		54	54	58	61	64	66	69

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

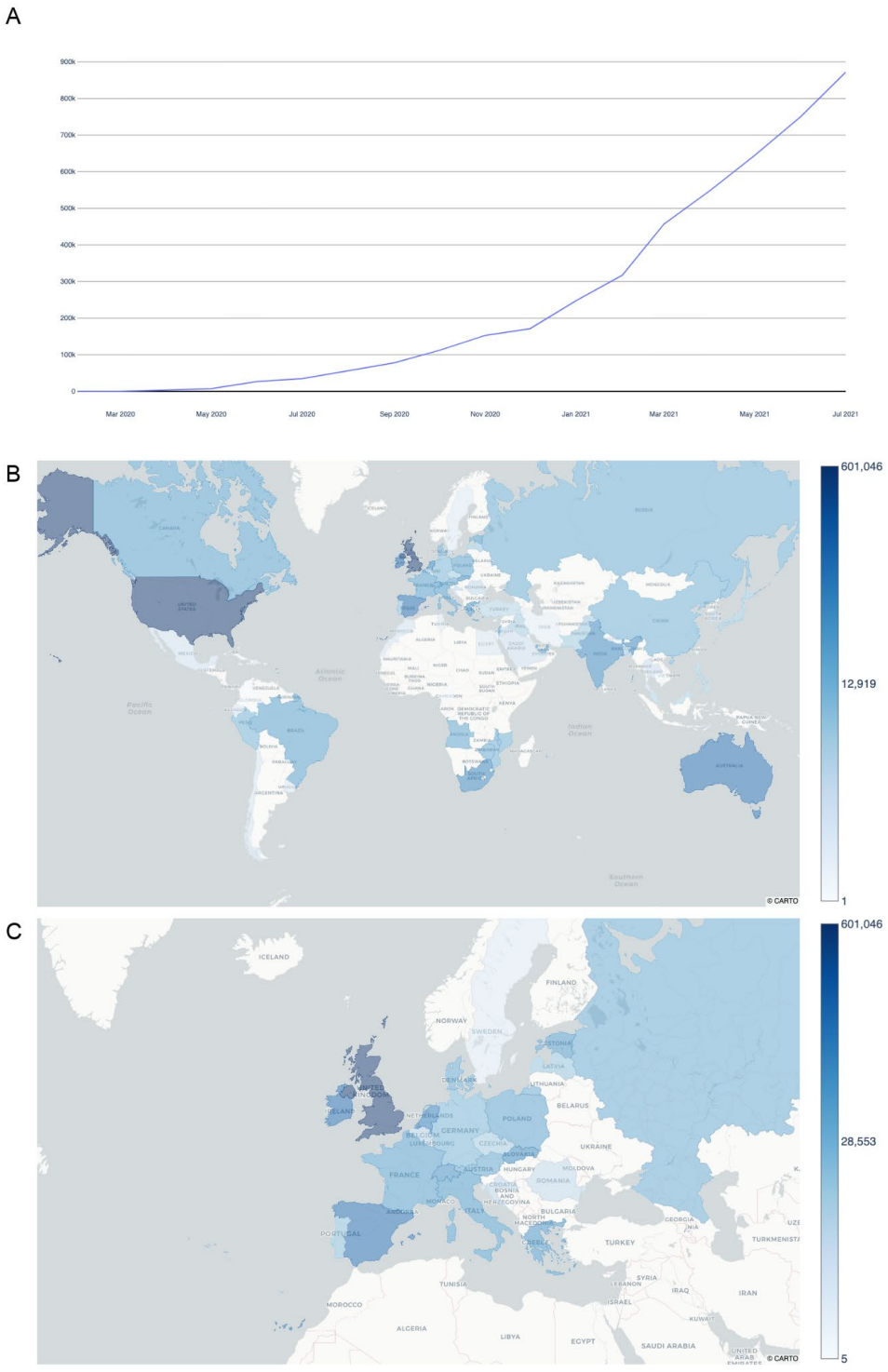


Figure I: Globally available raw SARS-CoV-2 data and distribution of sources, showing (A) sustained growth in raw data since launch of the mobilisation campaign by cumulative number of data sets, (B) and (C) geographical sources of global and European raw data, respectively, for which 75% of global data have been routed through the SARS-CoV-2 Data Hubs, with the remaining 25% 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).

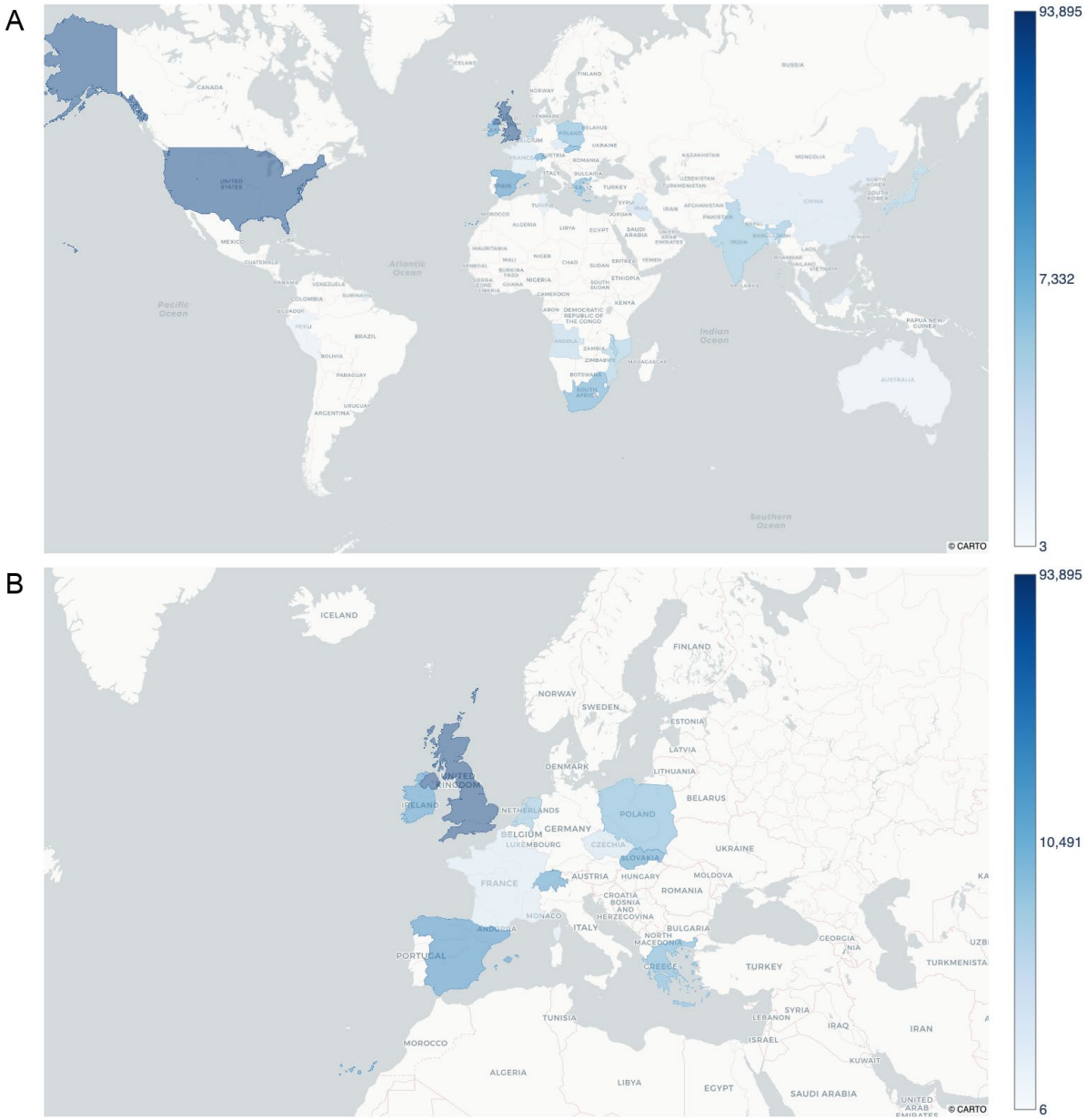


Figure II: New raw SARS-CoV-2 data and distribution of sources at global (A) and European (B) levels mobilised since 14 June 2021. Note that the colour scales are logarithmic best to show the broad range across countries.

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

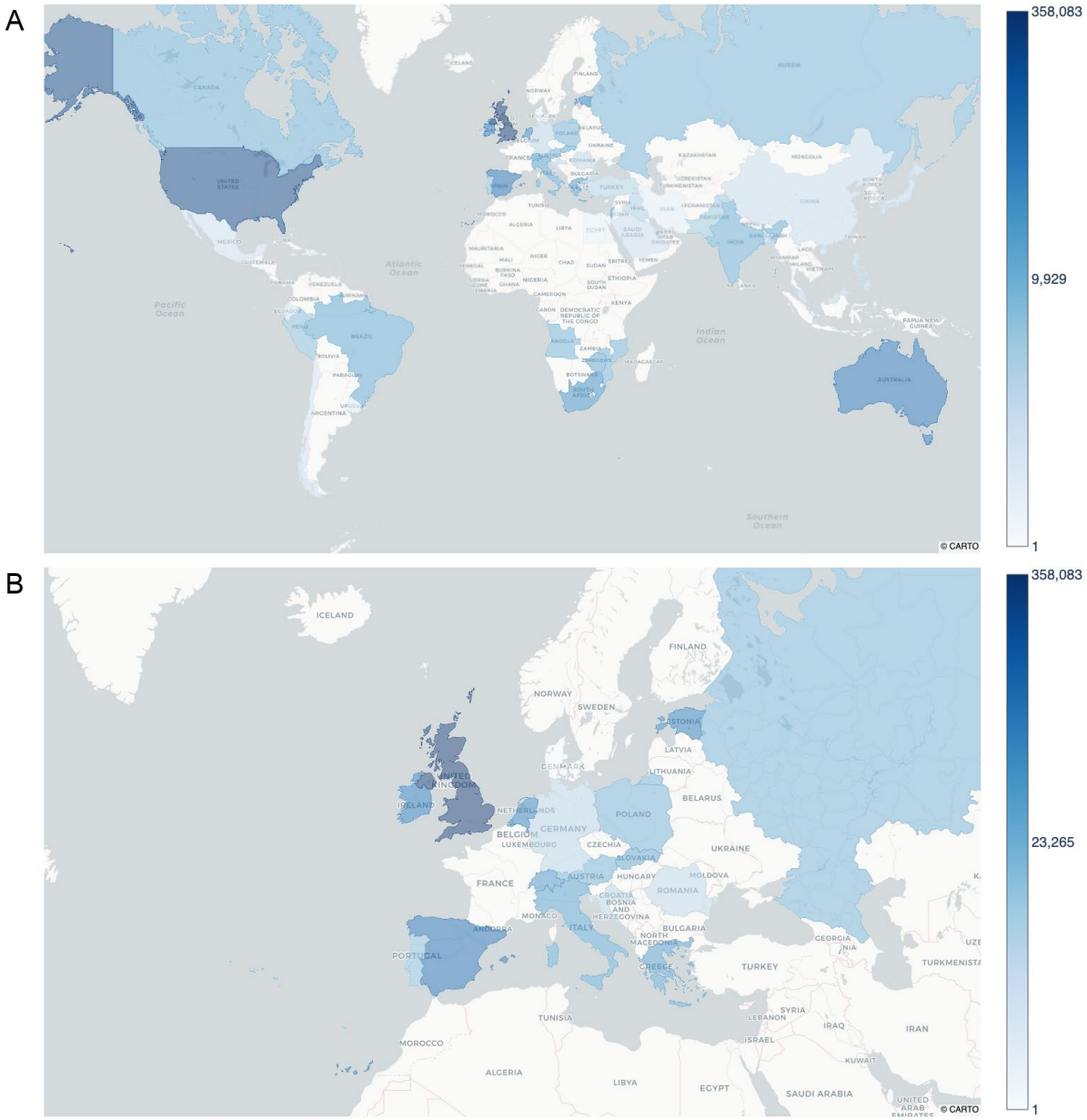


Figure III: Geographical sources of analysed raw data comprising 496,822 data sets spanning the period of data first published from 31 Jul. 2020 to 10 Jul. 2021 globally (A) and within Europe (B). Note that the colour scales are logarithmic best to show the broad range across countries.

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 16 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 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 22 of July 2021

No new VOCs and VOIs have been detected since the last update. The latest VOC Delta (B.1.617.2) is increasing in prevalence in multiple countries and in many countries becomes dominant, replacing earlier circulating lineages. As of 22 July, Delta (B.1.617.2) is found in 107 countries globally. The VOI Lambda (C.37) that mainly circulated in the South American countries (Argentina, Chile and Peru) has become the dominant variant in Peru and is found in 29 different countries globally. Some sub-lineages of the VOCs have been identified; these sub-lineages contain additional mutations that might be of biological importance. The sub-lineages of some of the VOCs are further explained in the section below. Also, the formed VOIs Epsilon (B.1.427/B.1.429), Zeta (P.2) and Theta (P.3) have been downscaled to Alerts for Further Monitoring and have been removed from the overview.

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, B.1.351.2 and B.1.351.3), Gamma (B.1.1.28.1 or P.1) and Delta (B.1.617.2, AY.1 and AY.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; e.g., AY.1 and AY.2 contain the additional mutation K417N when compared with its parent lineage. According to the WHO nomenclature, these sub-lineages AY.1 and AY.2 together with B.1.617.2 are still referred to as Delta. For the Beta variant, sub-lineage B.1.351.2 contains the additional L18F mutation, which is associated with escape from multiple N-terminal domain (NTD) binding monoclonal antibodies. The other Beta sub-lineage B.1.351.3 contains the additional A67V mutation in the NTD, which is also found in the VOI Eta (B.1.525).

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 the broad antibody response that is induced by vaccination. In this report, data are

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

presented for the analysis of the presence of variants Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1) and Delta (B.1.617.2). 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

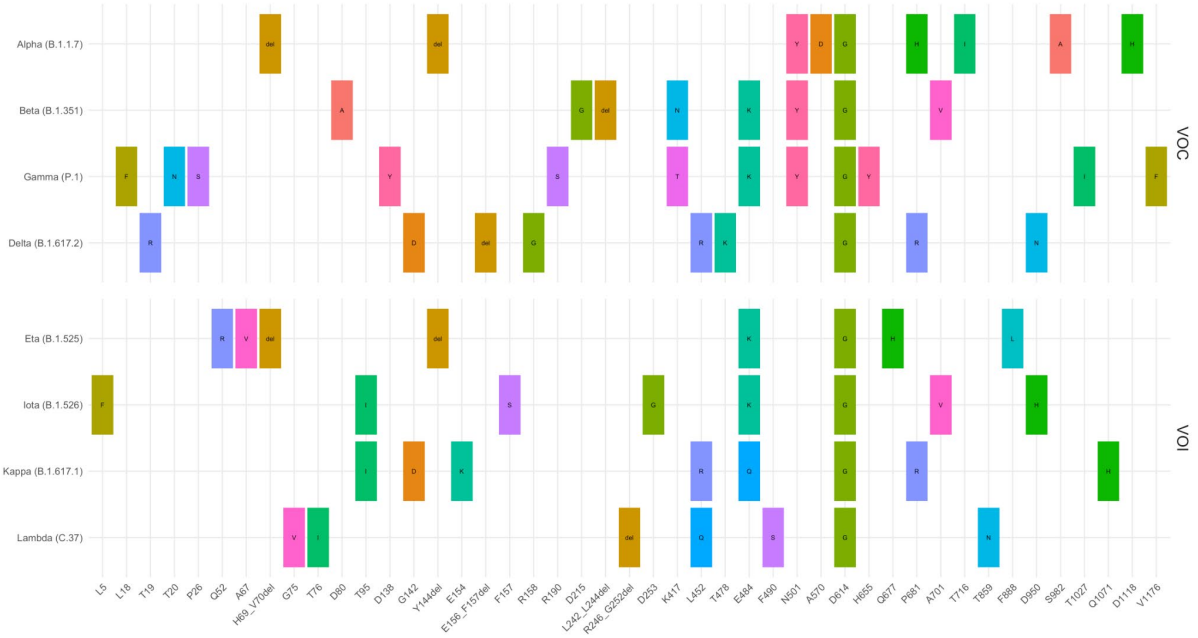
This work is supported by funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 874735 (VEO).

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.

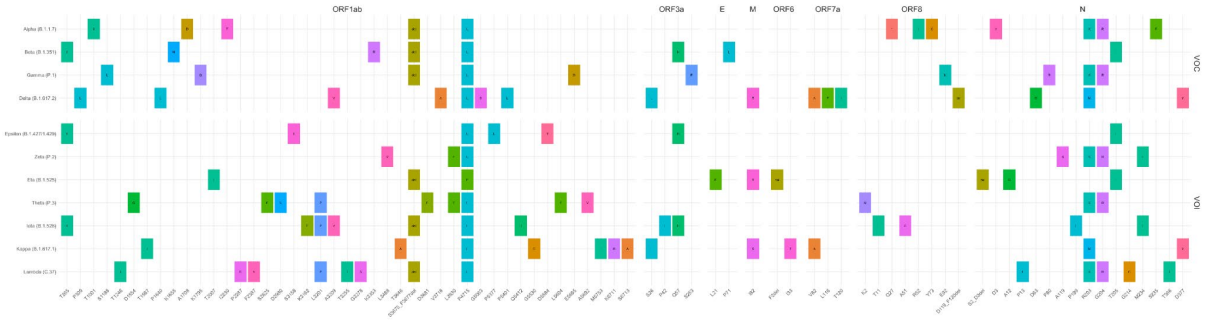
There is evidence that individual mutations may have some effect: for instance, the E484K mutation has been associated with reduced neutralization by convalescent and post-vaccine sera, the N501Y mutation with increased binding affinity to the hACE2 receptor, and the L452R mutation with increased infectivity 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.

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)

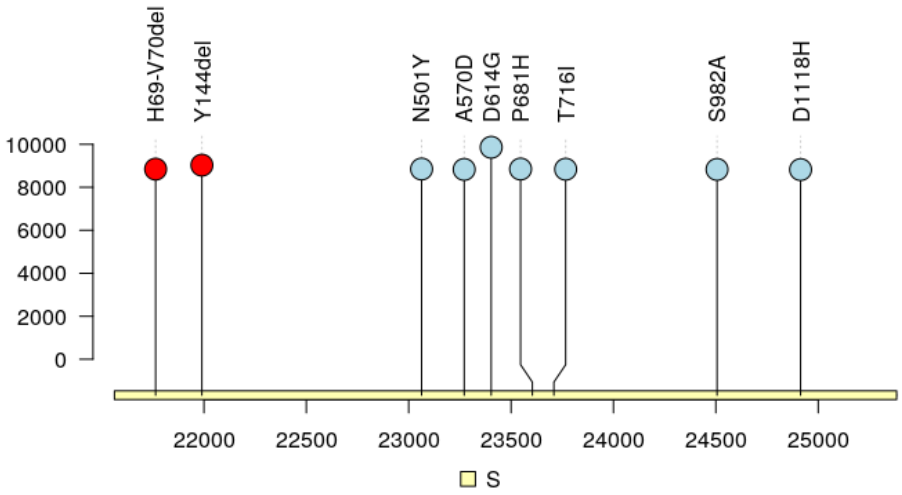


Figure IV: Variant of concern Alpha as defined by the mutations in the spike protein.

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).

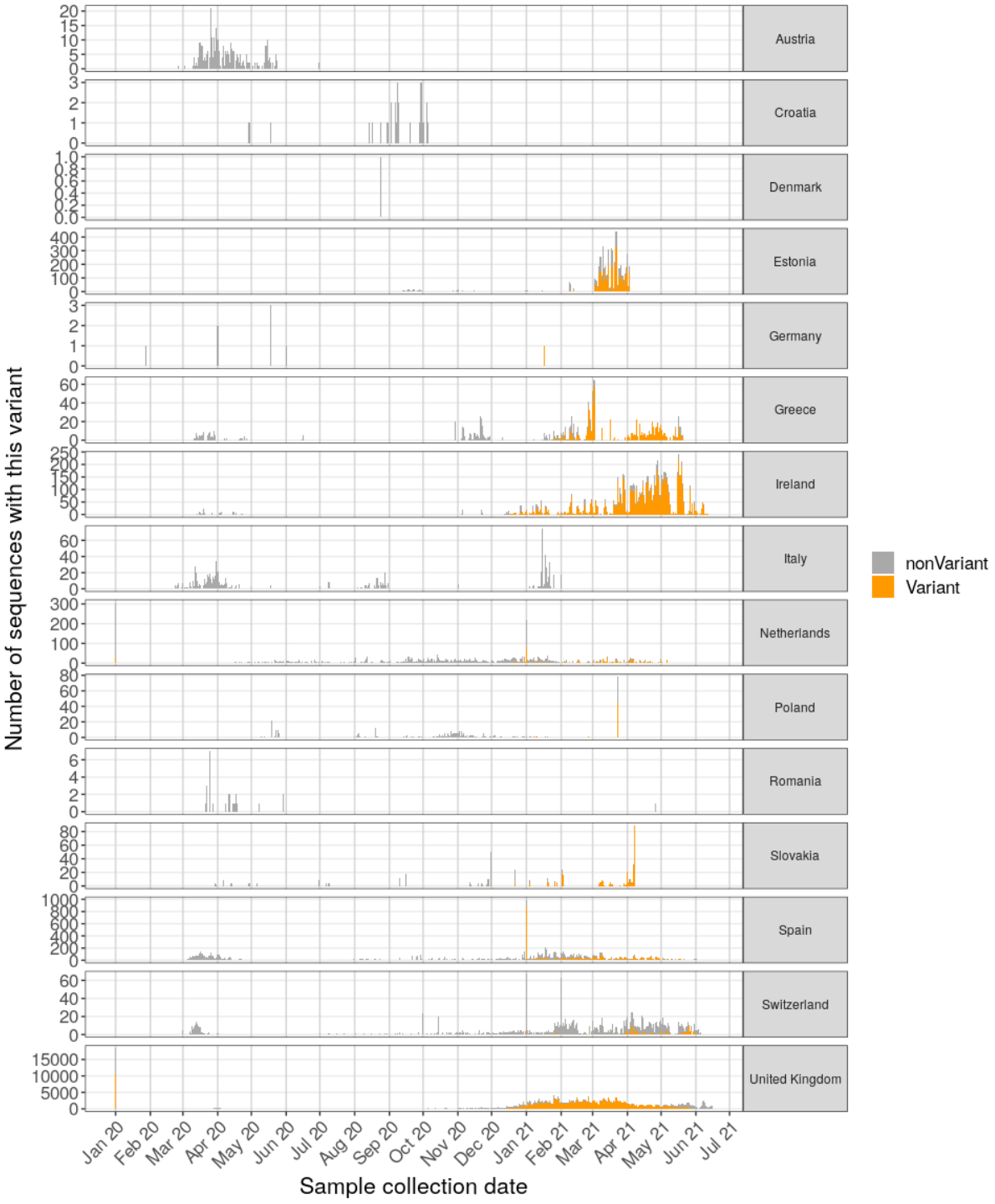


Figure V: Number of sequences by date of sampling for Alpha variant (orange) and non Alpha for countries in Europe and Turkey.

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)

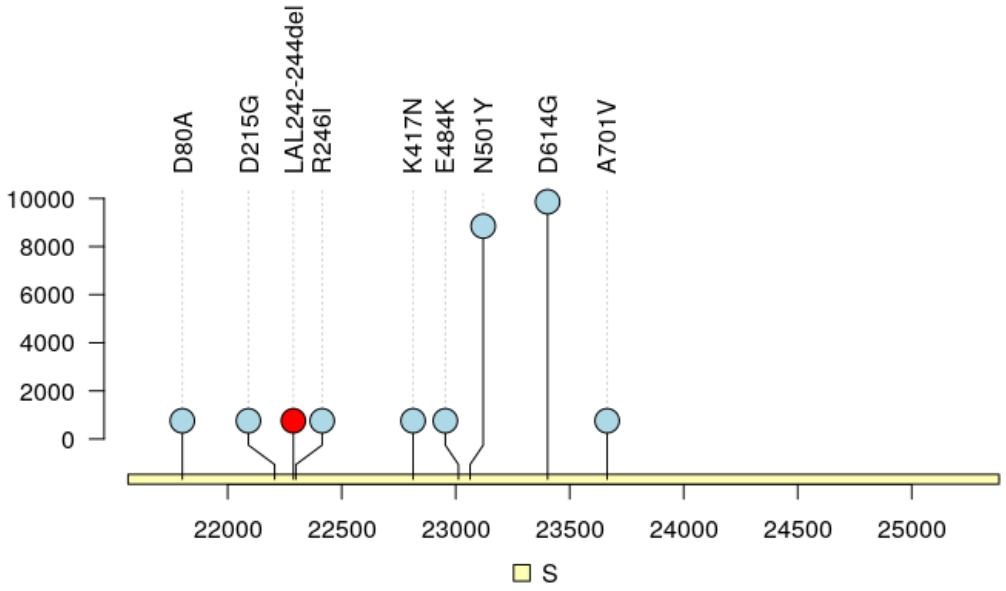


Figure VII: Mutations in the spike protein defining variant Beta. This variant was not detected in the data uploaded since January.

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	1052
Netherlands	92	697
Spain	19	628
Greece	15	46
Ireland	34	74
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)

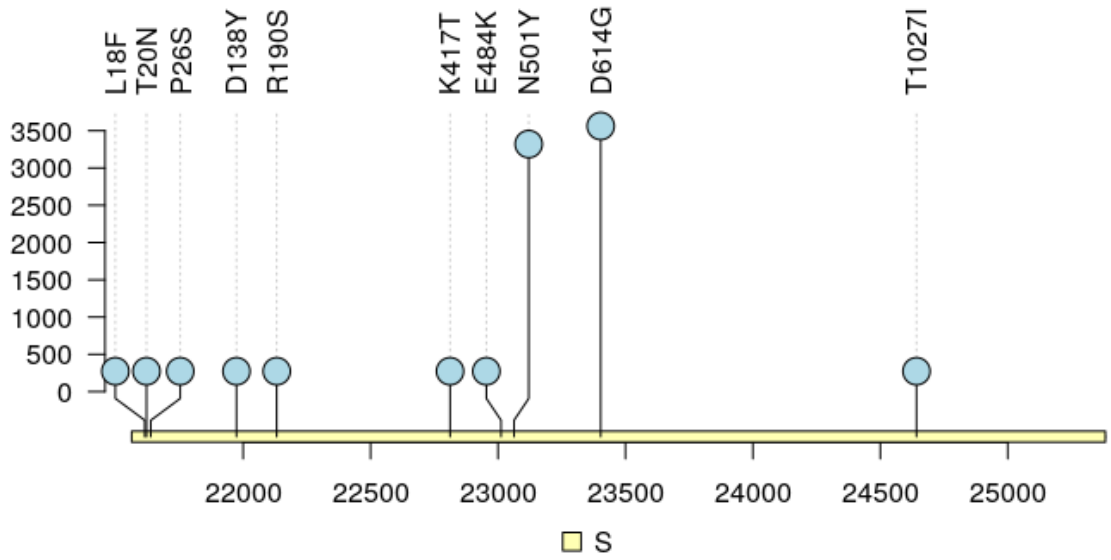


Figure IX: Mutations in the spike protein defining variant Gamma.

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	221
Spain	99	954
Japan	1	111
USA	1304	22143
Italy	3	2218
Netherlands	15	560
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	8	33
Bangladesh	1	1

Delta variant (B.1.617.2)

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 European countries in the bar charts.

ENA		GISAID
United Kingdom	5544	156339
Netherlands	5	1373
USA	47	18771
Ireland	42	2184
Switzerland	4	841
Spain	25	2999

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

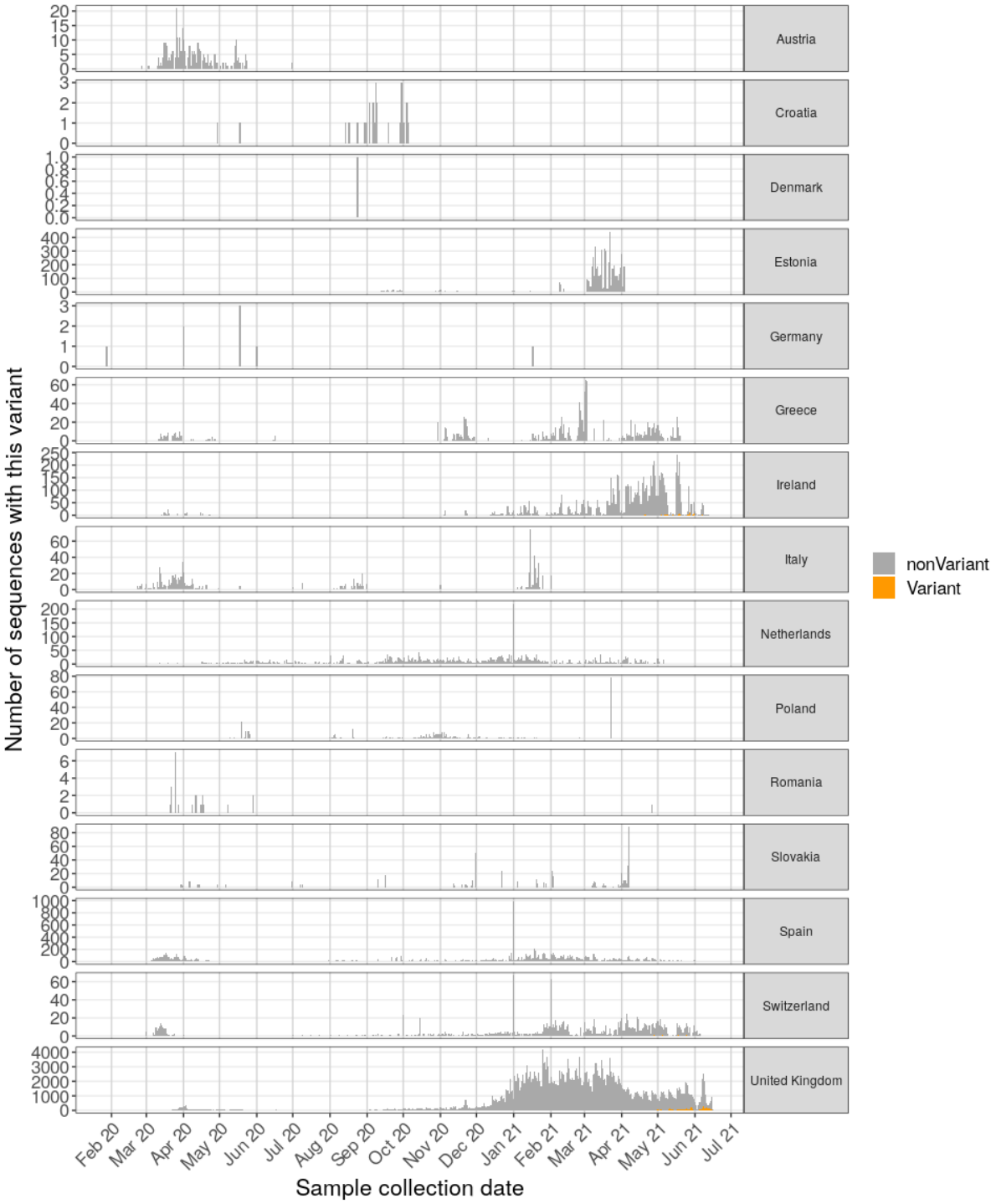


Figure X: Number of sequences by date of sampling for variant B.1.617.2 *(orange) for European countries.*

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

A minor issue with the Nanopore VCF analysis script called for an update of the script and a rerun of the Nanopore data, therefore the current snapshot contains 2,693 Nanopore data based VCF files. The total 109,616 Nanopore read datasets in the ENA public database will be processed in the coming weeks.

Phylogenetic visualization tool

In Report #5 (23 June 2021), the phylogenetic visualization tool (https://www.covid19dataportal.org/phylogeny-tree) was introduced. The SARS-CoV-2 phylogeny presented in this tool is constructed with reference-based mapping and distance methods, adapted from Szarvas et al. (2020).

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).