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

Report No. 2 – 11 Mar. 2021

Summary:

- Update on mobilisation of raw reads, now totaling sequencing data sets from 354,106 viral isolates from 60 countries, a 17% increase since the previous report (16 Feb. 2021).
- The majority of data submitted is from a limited number of countries (UK and USA).
- With the exception of the UK, there is a considerable delay in uploading data to the ENA. The majority of uploads deposited from 4 Jan. 2021 to 4 Mar. 2021 was from samples throughout 2020.
- The analysis of recently submitted data with recent sampling dates shows the ability to detect variants of concern in the raw sequence data.
- To make optimal use of the EU open data effort, member states need to be encouraged to start sharing raw sequence data as timely as possible.

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. An automated workflow for calling of mutations in raw Illumina read sets shared through the EU COVID-19 open data initiative has been finalised and is now processing the backlog of data. This is expected to be finalised by April after which reports for the full dataset will be available *(Note that this is conditional to intensity of new uploads. In case the number of new releases increases, those will be prioritized over re-analysis of historical data)*. In report No. 1, a first overview of some minor variant selection options was provided. In this report, focus is on a new addition to the analysis, focusing on combinations of mutations defining variants of concern, as specified by the WHO Evolution working group.

Section I: Data mobilisation

The number of datasets released into the data portal since the previous report (16 Feb. 2021) is shown in Table I, and summarized for raw reads in Figures I and II and for assembled sequences in Figure III (https://www.covid19dataportal.org). The plots

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

show the data submitted since the last report. The current VCF workflow is based on analysis of Illumina sequence read sets. A workflow for data generated using the second commonly used platform, nanopore, is under development. A short update on the progress will be given in the next report.

Table I: Previous number, number new submissions.

Date		**16 Feb. 2021**	**4 Mar. 2021**
Raw data sets	Total	301,378	354,106
	Illumina	255,431	302,409
	Oxford Nanopore	45,222	50,972
	Other	725	725
Source countries for raw data		59	60

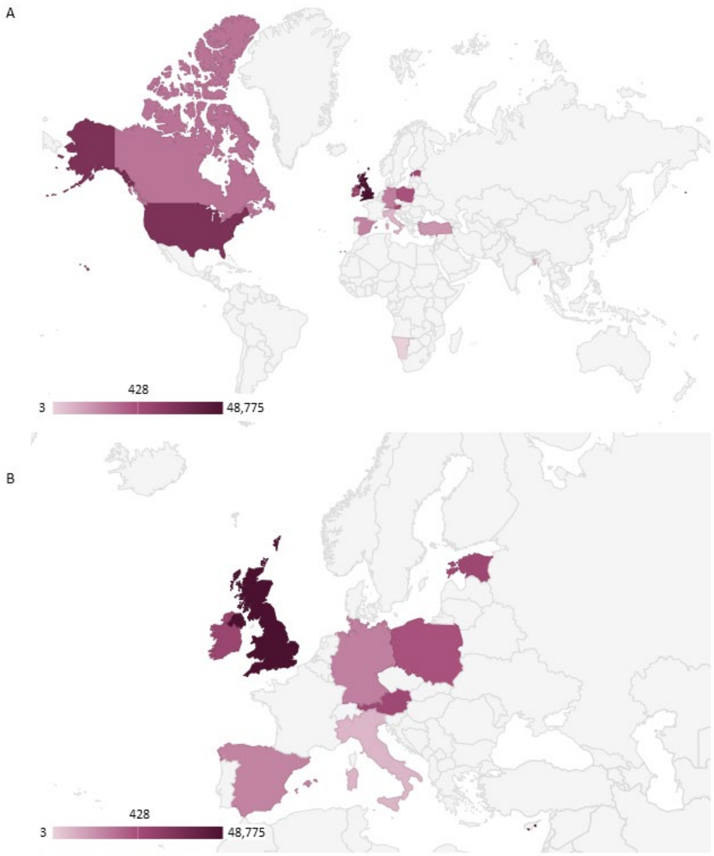


Figure I: Geographical sources of mobilised ***raw data since report No. 1****, spanning the period from 16 Feb. 2021 to 4 Mar. 2021 globally (A) and within Europe (B).*

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

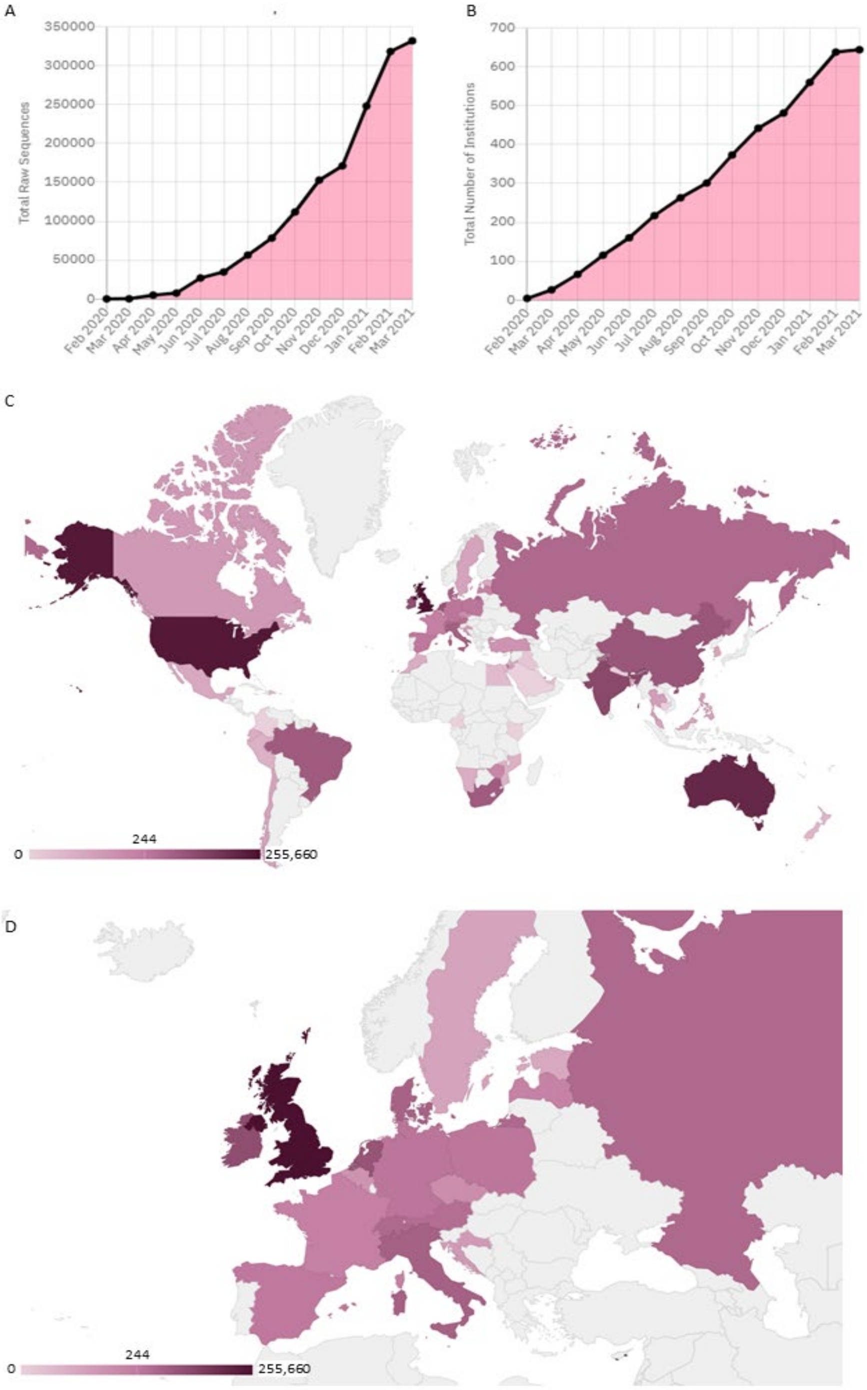


Figure II: Growth of raw SARS-CoV-2 data and distribution of sources, showing (A) sustained growth in raw data since launch of the mobilisation campaign, (B) a growing number of institutions providing data, (C) and (D) geographical sources of global and European raw data, respectively, for which 75% 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).

Section II: Analysis

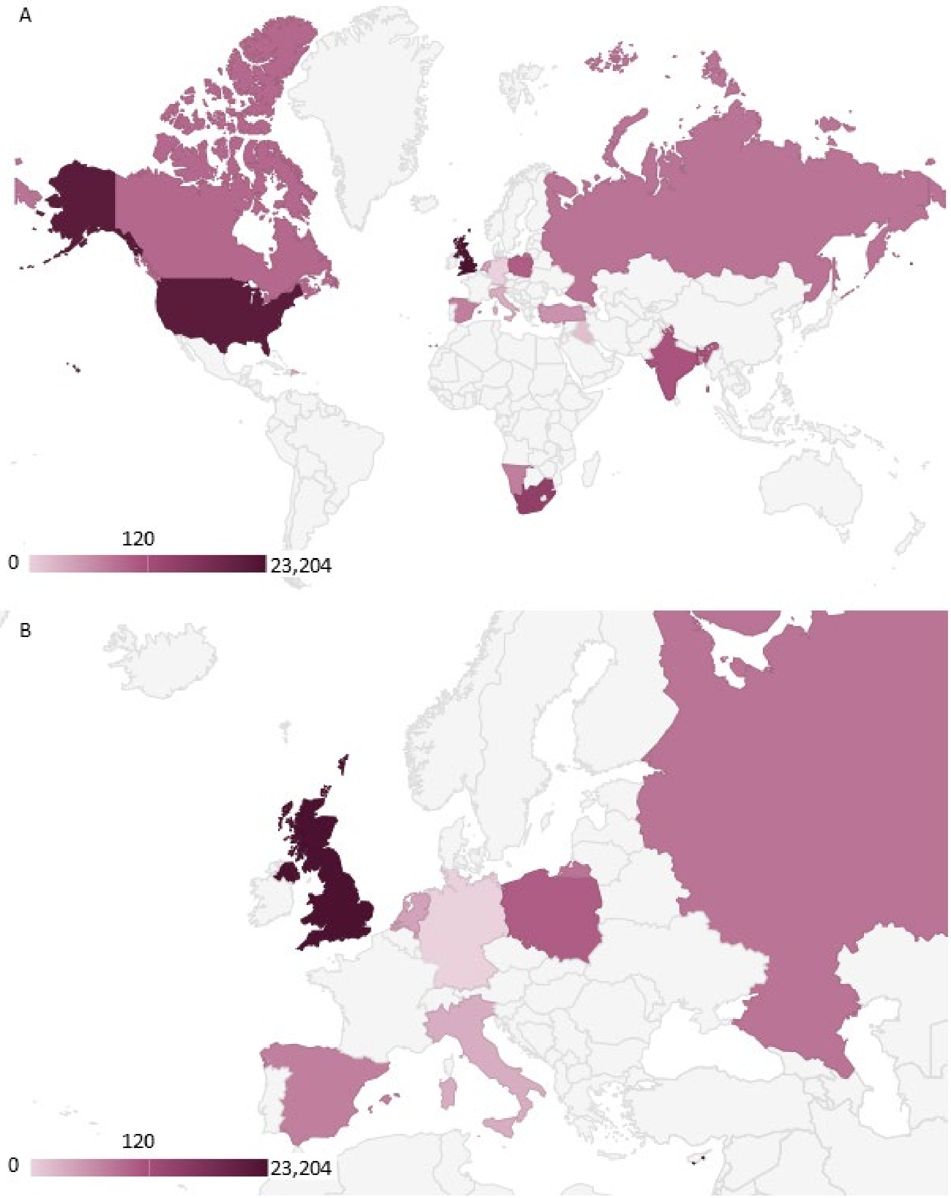


Figure III: Geographical sources of analysed raw data comprising 34,081 submissions spanning the period from 4 Jan. 2021 to 4 Mar. 2021 and covering Illumina data only, showing global (A) and European (B) views.

Results of variant calling, first version visualisation tool

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 from 4 Jan. 2021 to 4 Mar. 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.

Variants of concern

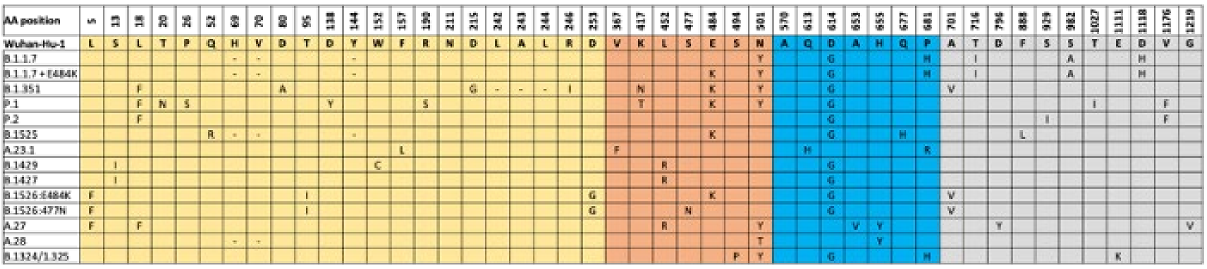
Below, the first summary is given of analysis of raw read datasets for presence of the combination of mutations that define the different variants of concern.

At the moment three Variants of Concern (VOC) have been described: the B.1.1.7, the B.1.351 and the B.1.1.28.1 (P1) variants. In addition, some B.1.1.7 sequences have been detected that contain the mutation E484K as well. All of these VOCs are defined by a remarkable number of mutations along the genome and in the spike protein. Some of the common mutations are the N501Y and the E484K. Several papers concerning the transmissibility, the possible effect of individual and combined mutations on antigenicity in relation to vaccination, and potential effect on disease severity, and impact on diagnostics have been described. In this report, data are presented for the analysis of presence of variants B.1.1.7, B.1.1.7 + mutation E484K, B.1.35.1, and B.1.1.28.1.

Variants of interest

In addition to the VOC there have been several reports of Variants of Interest (VOI) 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: the E484K mutation has been associated with reduced neutralization, the V367F mutation with increased expression, and the L452R mutation with increased infectivity and reduced neutralization. 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. Below is a table summarising all variants currently under consideration in the WHO virus evolution group. This list is not static, but WHO is in the process to formalise the process of assignment of a label (as VOC or VOI). Once that is available, the WHO designated variants can be included in the future versions of this report.

Table II. Overview of mutations in the spike region of the genome of several VOC/VOIs. Colors represent region within spike genome: green = signal peptide (SP), yellow = N-terminal domain (NTD), orange = receptor binding domain (RBD), blue = close to furin cleavage site, grey = other regions in S2 subunit



AA position	5	13	18	20	26	52	69	70	80	95	138	144	152	157	190	211	215	242	243	244	246	253	367	417	452	477	484	494	501	570	613	614	653	655	677	681	701	716	796	888	929	982	1027	1111	1118	1176	1219
Wuhan-Hu-1	L	S	L	T	P	Q	H	V	D	T	D	Y	W	F	R	N	D	L	A	L	R	D	V	K	L	S	E	S	N	A	Q	D	A	H	Q	P	A	T	D	F	S	S	T	E	D	V	G
B.1.1.7							-	-				-																	Y			G				H		I				A			H		
B.1.1.7 + E484K							-	-				-															K		Y			G				H		I				A			H		
B.1.351			F						A								G	-	-	-	I			N			K		Y			G					V										
P.1			F	N	S						Y				S									T			K		Y			G											I			F	
P.2			F																													G									I					F	
B.1525						R	-	-				-															K					G			H					L							
A.23.1														L									F								H					R											
B.1429		I											C												R							G															
B.1427		I																							R							G															
B.1526:E484K	F									I												G					K					G					V										
B.1526:477N	F									I												G				N						G					V										
A.27	F		F																						R				Y				V	Y					Y								V
A.28							-	-																					T					Y													
B.1324/1.325																												P	Y			G				H								K			

B.1.1.7 non-E484K variant (UK variant)

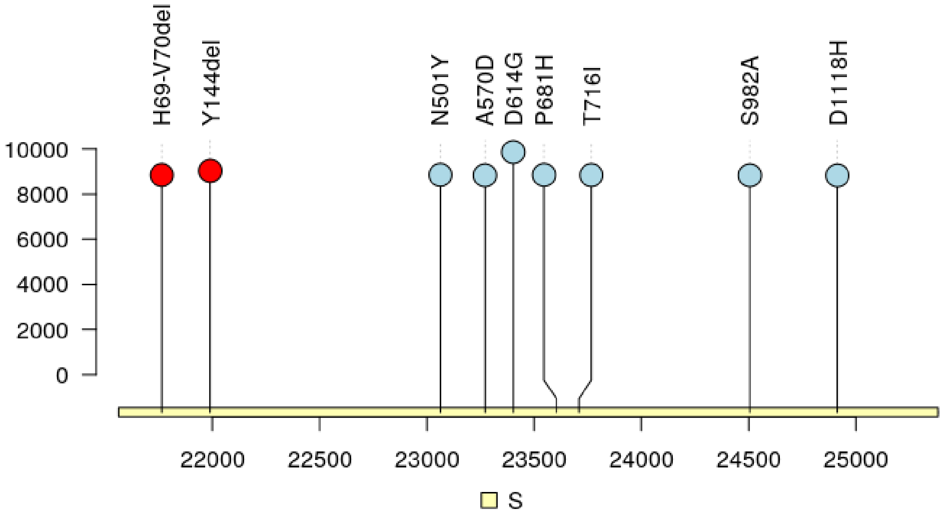


Figure IV: Variant of concern B.1.1.7 as defined by the mutations in the spike protein.

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 January 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 January, 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) or not at all (B.1.351), reflecting the delayed submissions. Therefore, the country plots are only shown for variant B. 1.1.7, as an example of visualisations available once data become more complete.

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

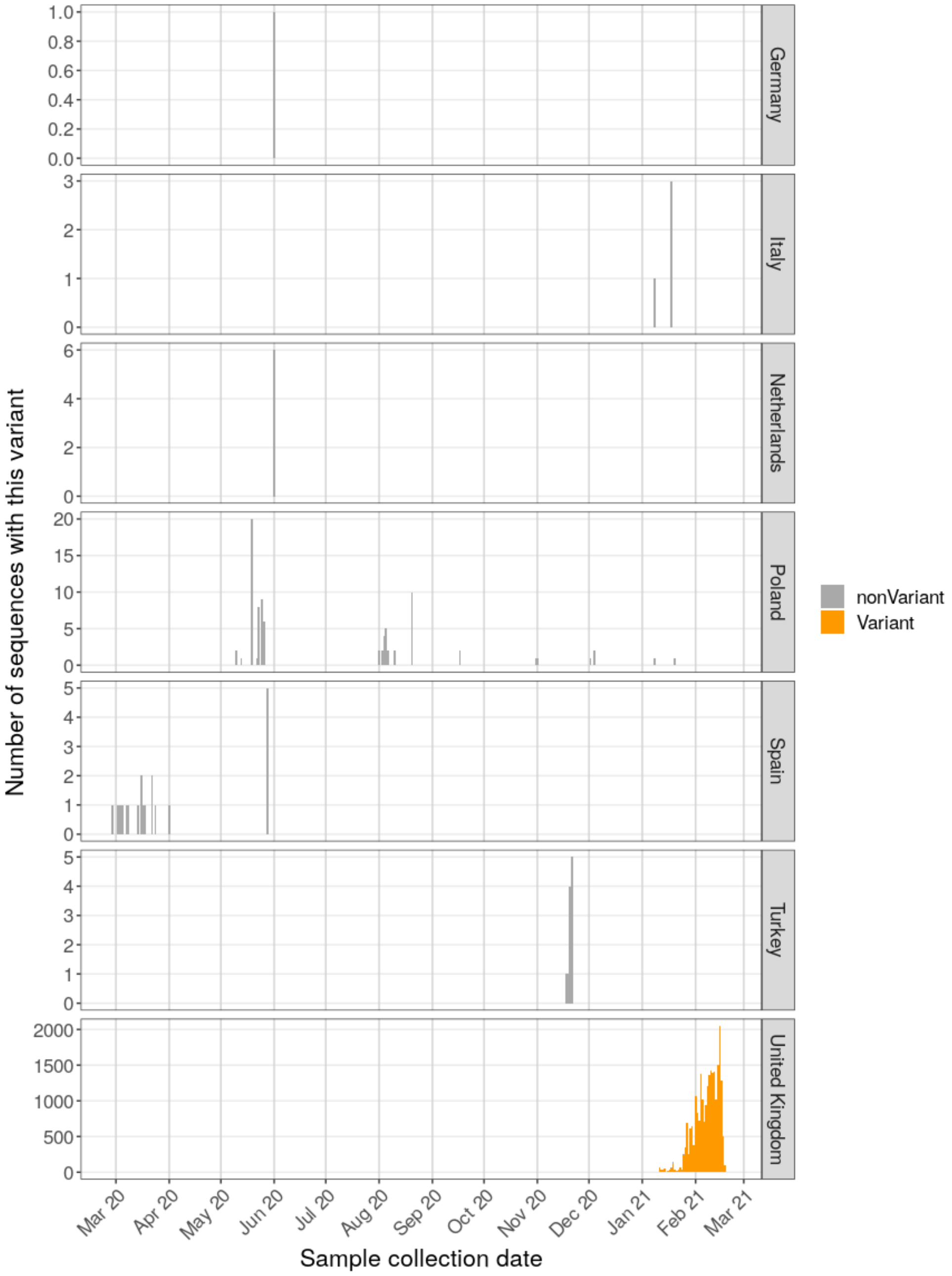


Figure V: Number of sequences by date of sampling for variant B.1.1.7 (orange) and non B.1.1.7 for countries in Europe and Turkey.

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

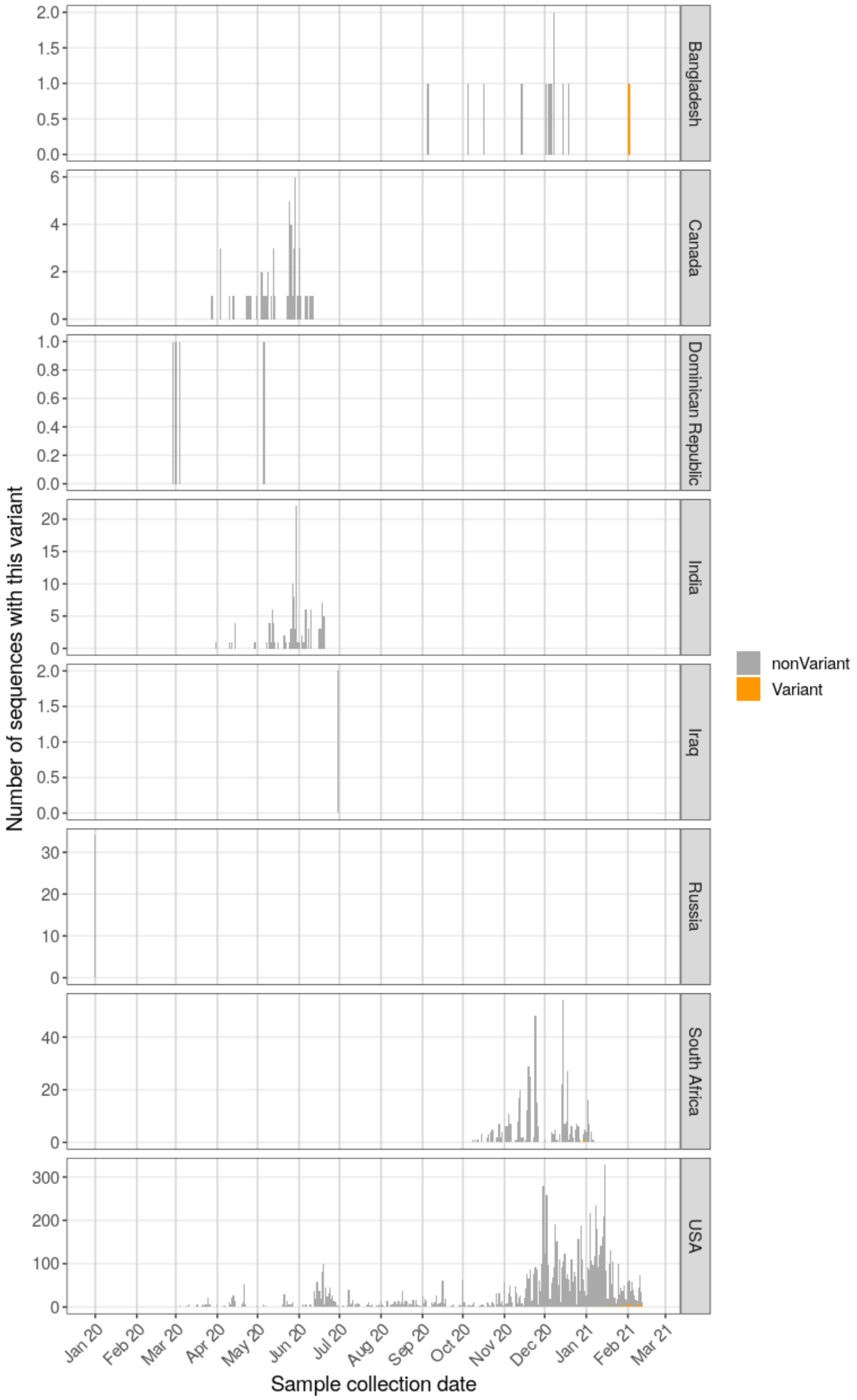


Figure VI: Number of sequences by date of sampling for variant B.1.1.7 (orange) for non-European countries.

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

B.1.1.7 + E484K variant

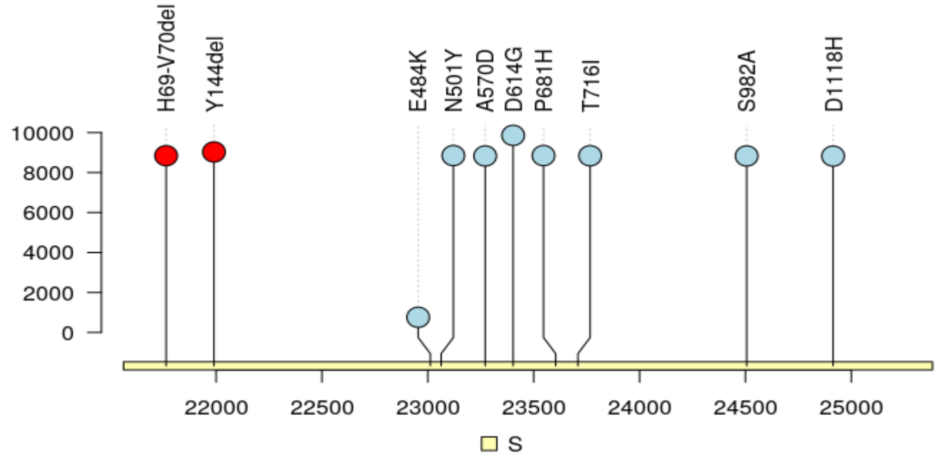


Figure VII: Variant of concern B.1.1.7 + mutation E484K.

B.1.351 variant (South Africa)

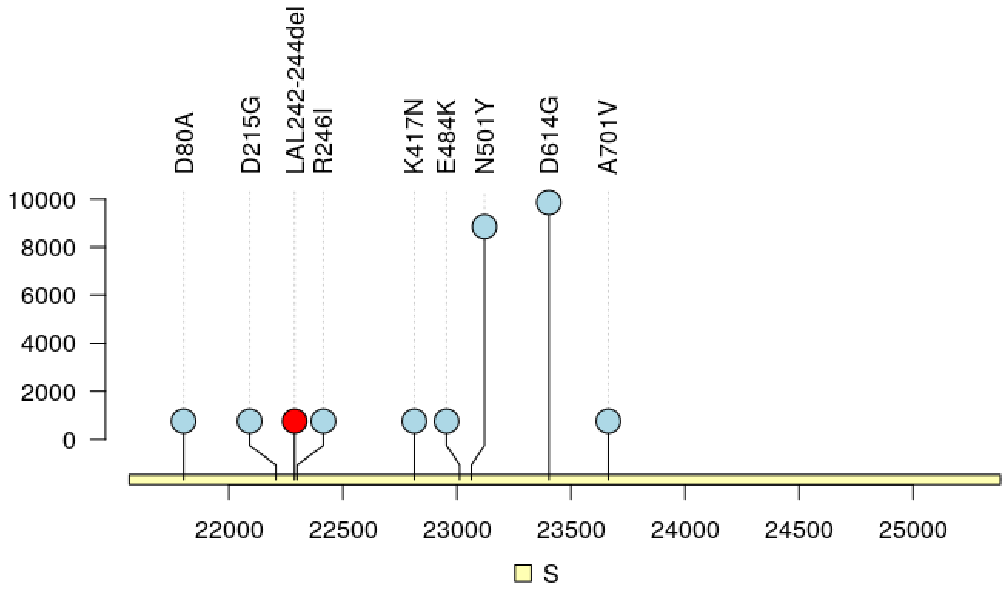


Figure VIII : Mutations in the spike protein defining variant B.135.1 . This variant was not detected in the data uploaded since January.

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

B.1.1.28.1 (P1) variant (Brazil)

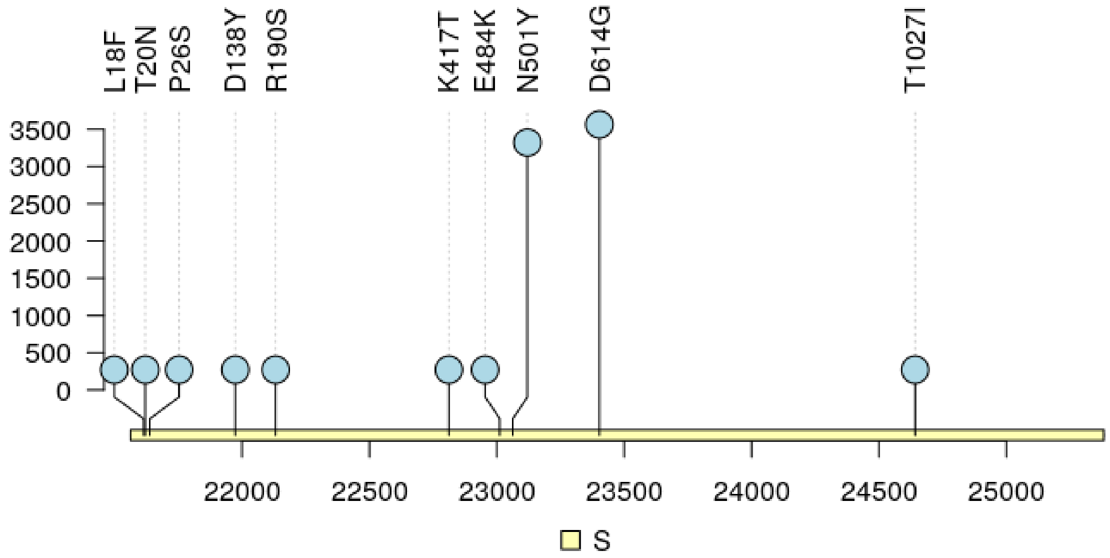


Figure IX : Mutations in the spike protein defining variant B.135.1

Recommendations and next steps:

The above report shows the first results of the automated mutation analysis on raw read datasets submitted to ENA, as well as visualisations of the data. This shows that a substantial number of raw reads has been publicly released but that the geographical distribution is highly skewed to a few countries, reflecting large-scale sequencing efforts. However, this does provide proof of concept of the function of the system. The number of raw sequencing data that are generated and shared from the EU member states are very limited and more data are needed to provide a timely overview of circulating variants. We currently are working 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.

In the following weeks, VEO expects to analyse all publicly shared Illumina data for presence of variants and to have a developed workflow in place for integrating nanopore data as well. 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.

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).