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

Report No. 4 – 27 May 2021

Summary:

- Update on mobilisation of raw reads, now totaling sequencing data sets from 552,185 viral raw read sets from 64 countries, a 5% increase since the previous report, as of 26 April 2021 (data freeze 19 April).
- Information on the B.1.617.2 variant and the B.1.621 variant was added.
- A decision was made on how to perform variant calling of the Oxford Nanopore data, the current workflow will be adjusted soon after which the backlog of Oxford Nanopore data can be processed. We will continue to further benchmark Oxford Nanopore variant calling.

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.

In this report, the status of the development of the Oxford Nanopore workflow is described, and progress on the analysis of the backlog of raw reads is given.

Section I: Data mobilisation

The number of datasets released into the COVID-19 Data Portal since the previous data freeze (19 Apr. 2021) up to the current data freeze (4 May 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
Raw data sets	Total	301,378	354,106	438,112	525,348	552,185
	Illumina	255,431	302,409	367,462	446,375	469,142
	Oxford Nanopore	45,222	50,972	69,921	77,913	81,466
	Other	725	725	729	1,060	1,577
Source countries for raw data		54	54	58	61	64

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

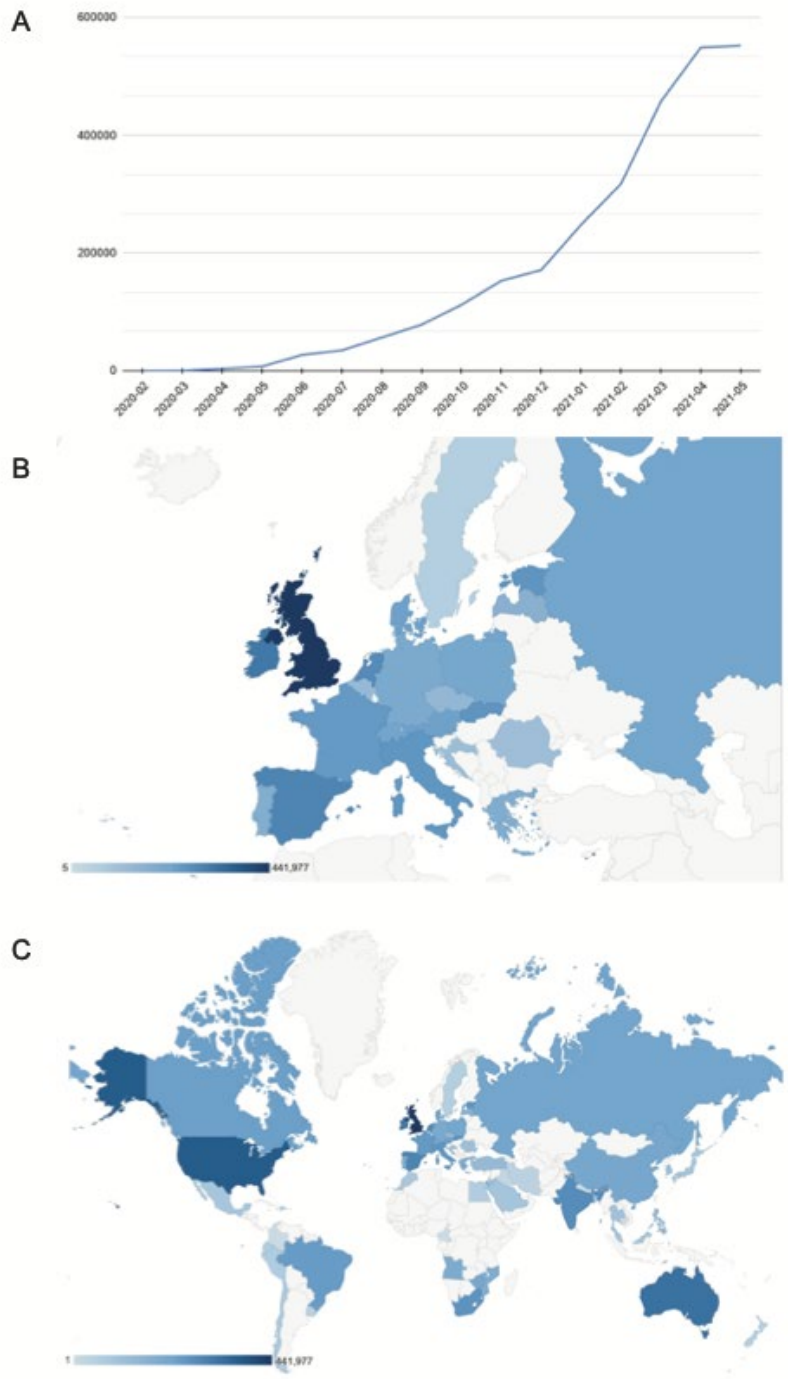


Figure I: Growth of raw SARS-CoV-2 data and distribution of sources, showing (A) sustained growth in raw data since launch of the mobilisation campaign by cumulative number of data sets, (B) and (C) geographical sources of European and global raw data, respectively, for which 85% of global data have been routed through the SARS-CoV-2 Data Hubs, with the remaining 15% 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).

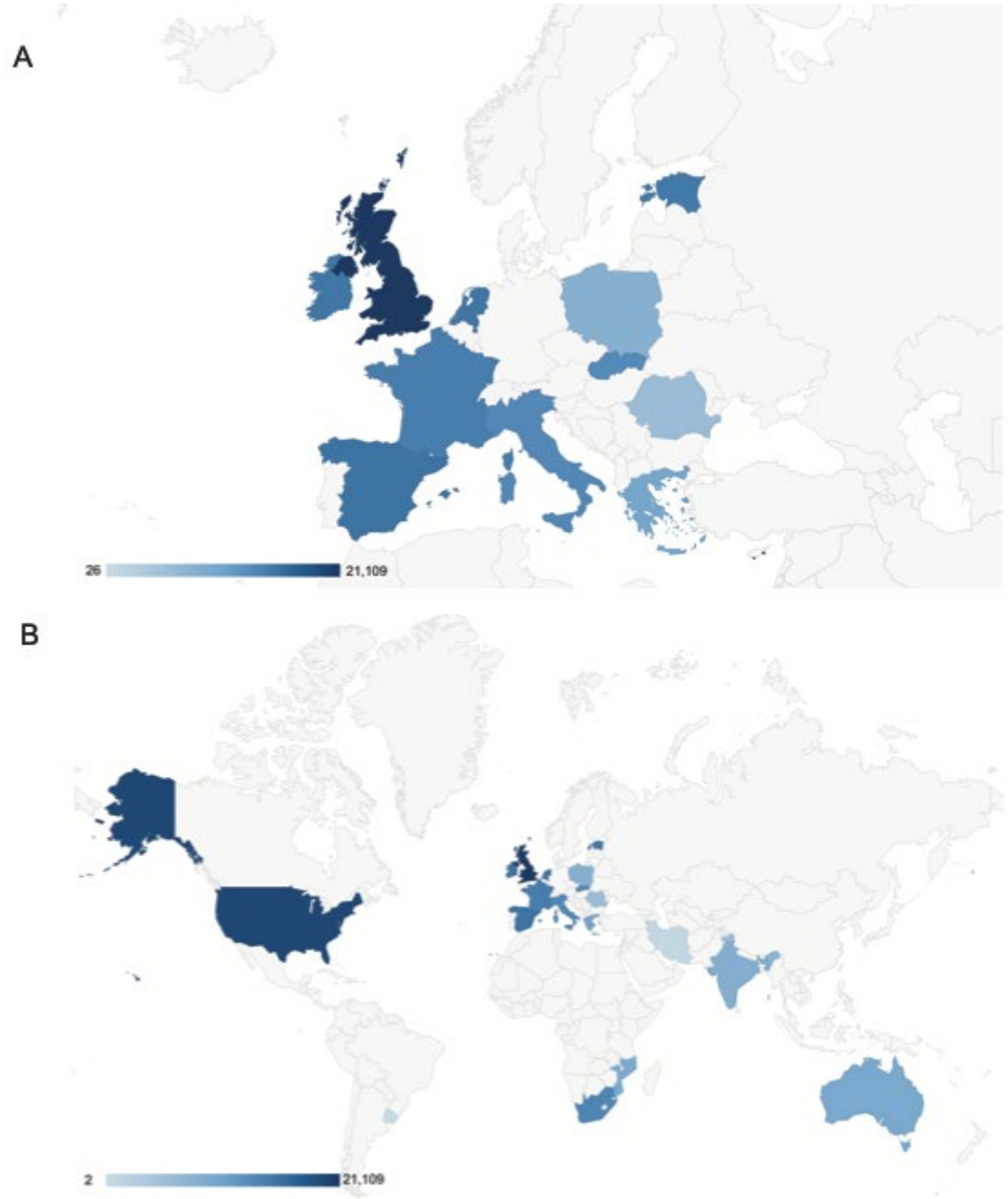


Figure II: New raw SARS-CoV-2 data and distribution of sources at European (A) and global (B) levels mobilised since 19 Apr. 2021. Note that the colour scales are logarithmic best to show the broad range across countries.

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

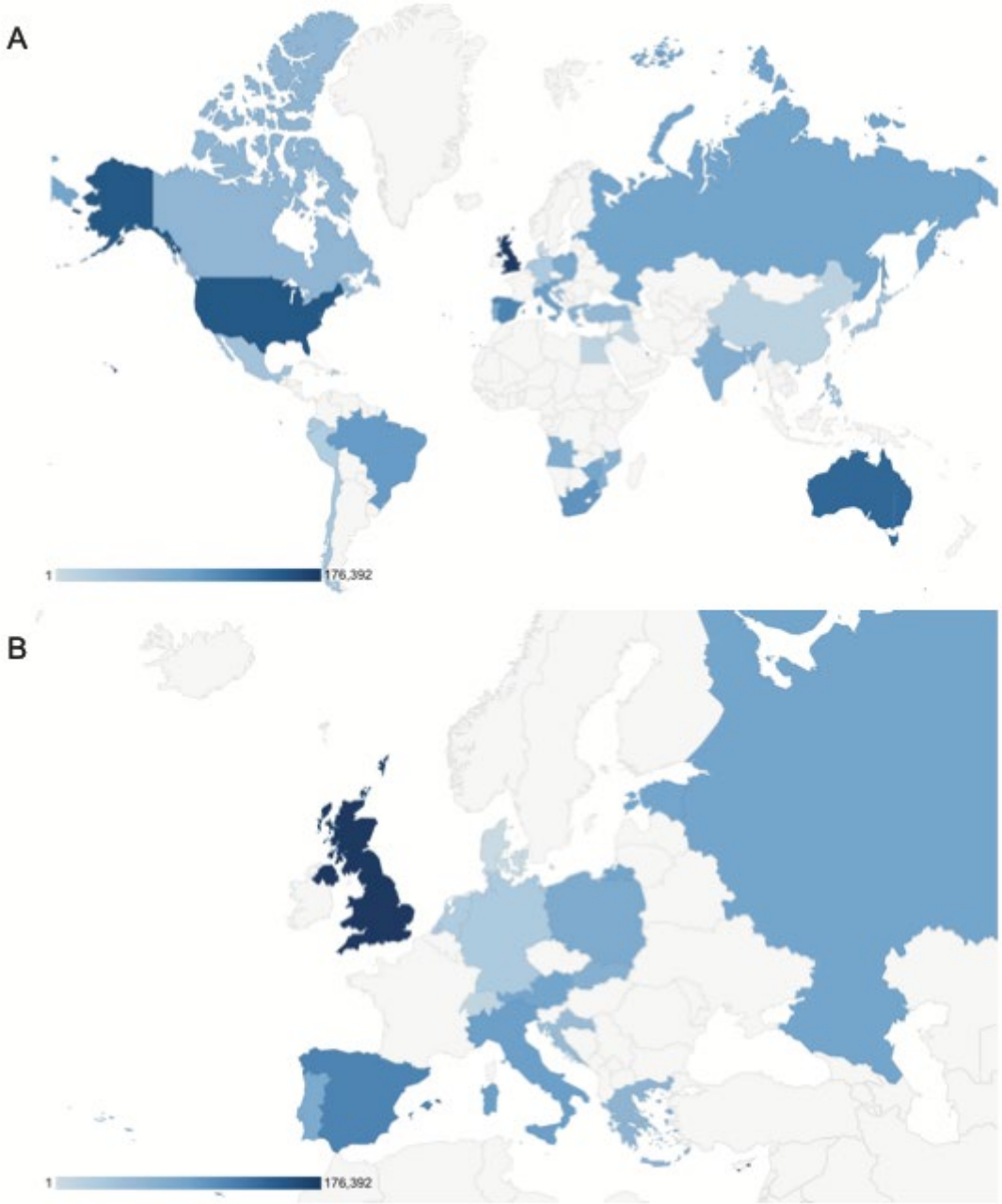


Figure III: Geographical sources of analysed raw data comprising 180,982 data sets spanning the period of data first published from 31 Jul. 2020 to 18 Apr. 2021 globally (A) and within Europe (B). Note that the colour scales are logarithmic best to show the broad range across countries.

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, 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 and/or made public from 31 Jul. 2020 to 30 Apr. 2021

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 26 Apr. 2021

The B.1.617 lineage has now been classified as VOC given the rapid spread of the novel variant globally. All three sublineages of B.1.617 -- B.1.617.1, B.1.617.2 and B.1.617.3 -- are classified as VOC by the WHO. All three sublineages have been increasingly detected in several countries and B.1.617.2 has shown to have increased transmissibility and some impact on vaccine effectiveness.

In addition, the B.1.621 is added as VOI. There is not much known about this novel VOI but this variant was originally found in Columbia and has since rapidly spread in a relatively short time across the globe. Among other countries, this virus has been detected in Germany, the Netherlands, Spain, Switzerland, and the USA.

Variants of concern

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

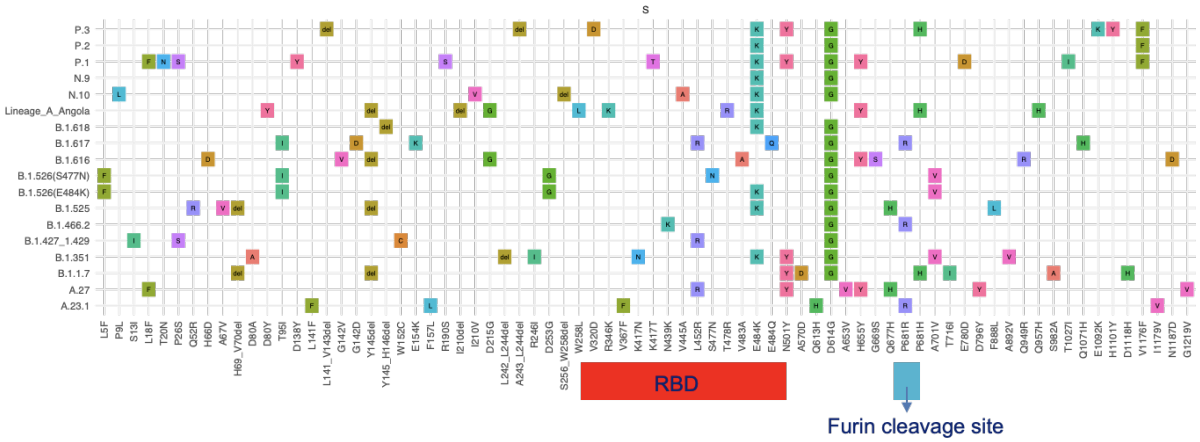
At the moment, four VOCs have been described: the B.1.1.7, the B.1.351,the B.1.1.28.1 (P1) and the B.1.617 variants. In addition, some B.1.1.7 sequences have been detected that contain the mutation E484K as well. All of these VOCs are defined by a remarkable number of mutations along the genome and in 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.

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, 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 of formalising 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 the different mutations of several VOC/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 vblu 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).

B.1.1.7 non-E484K variant (UK variant)

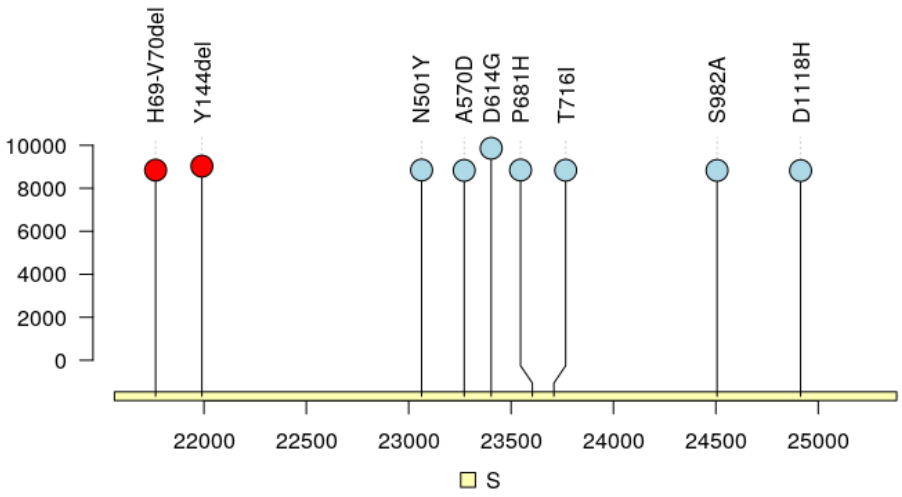


Figure IV: Variant of concern B.1.1.7 as defined by the mutations in the spike protein.

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 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).

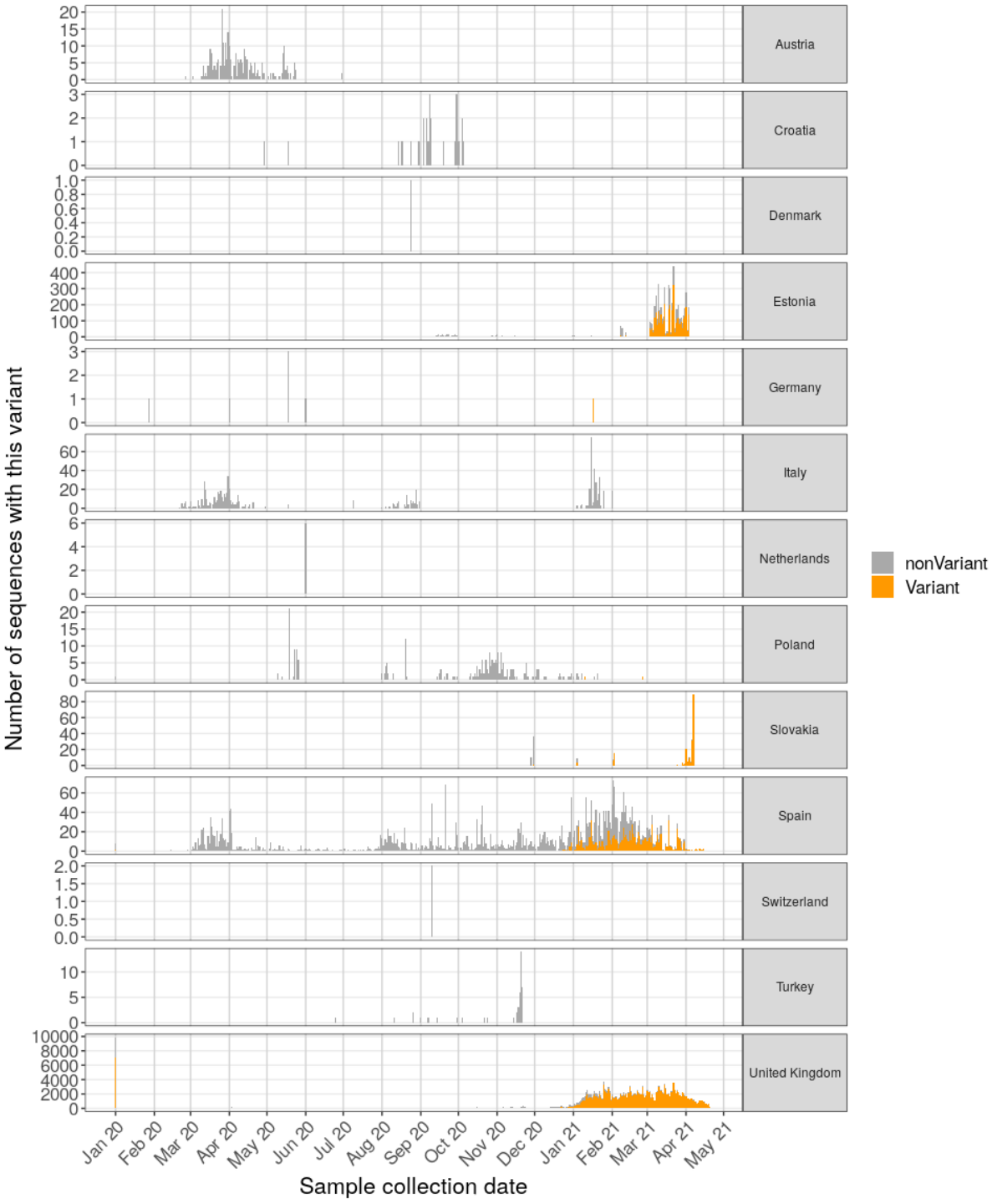


Figure V: Number of sequences by date of sampling for variant B.1.1.7 (orange) and non B.1.1.7 for countries in Europe and Turkey.

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

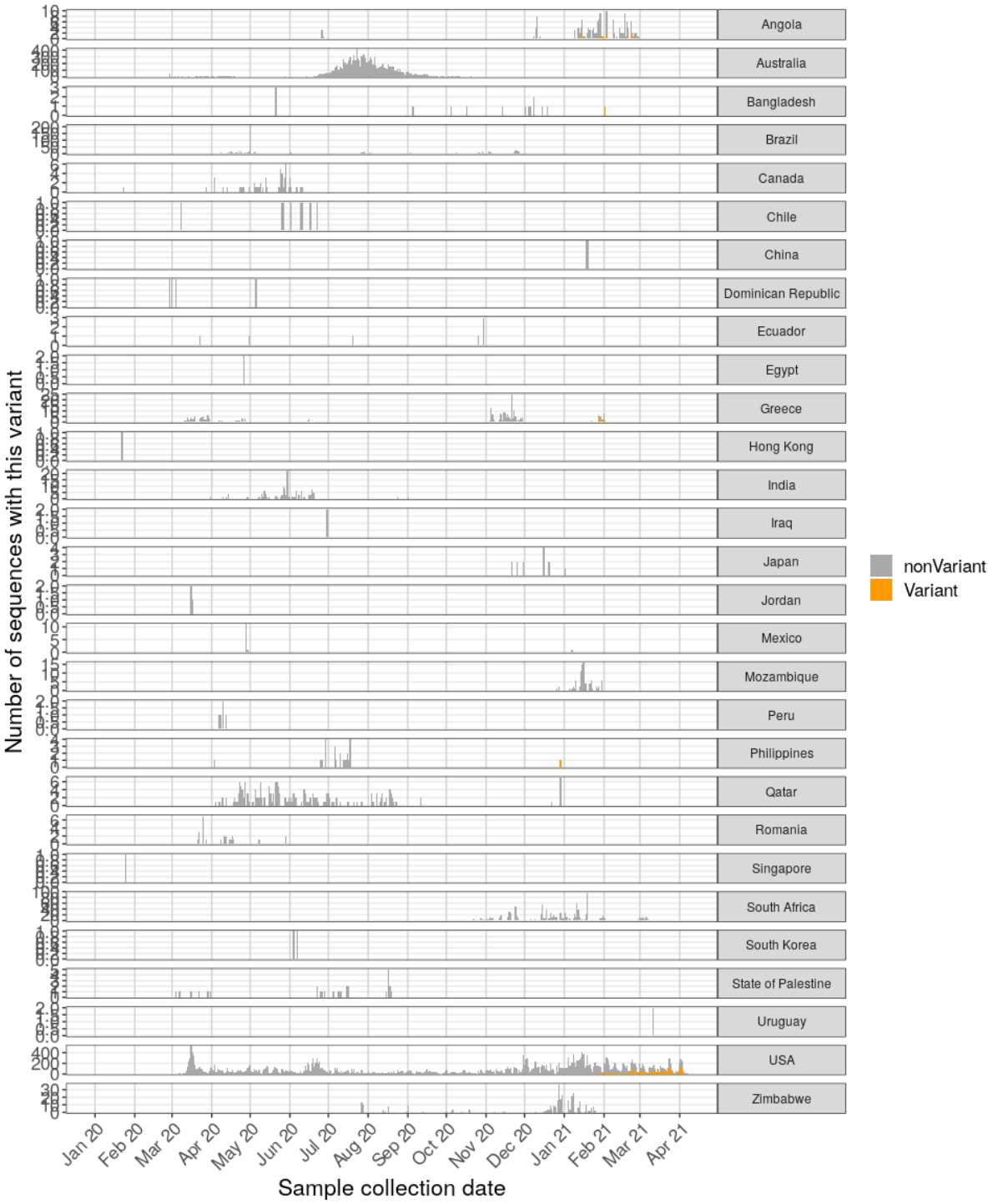


Figure VI: Number of sequences by date of sampling for variant B.1.1.7 (orange) for non-European countries.

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

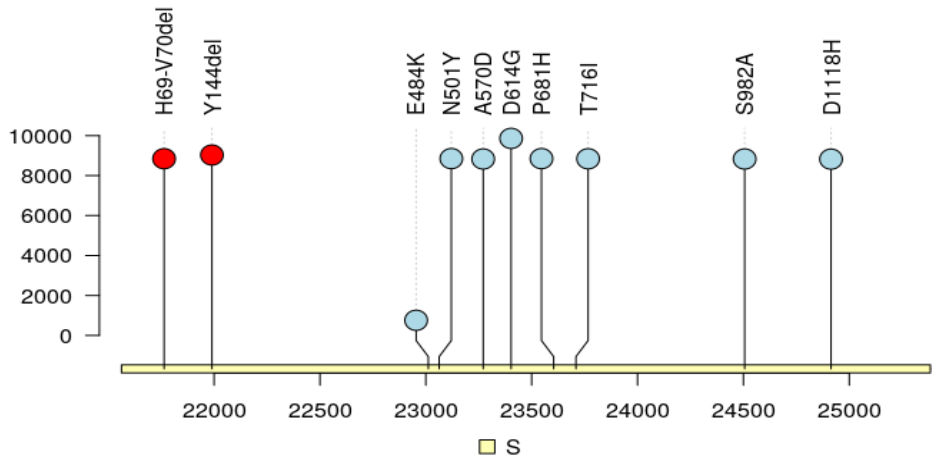


Figure VII: Variant of concern B.1.1.7 + E484K.

The only B.1.1.7 + E484K lineage samples are 121 from the UK and 3 in the US, but these are hardly visible against the large number of background sequences.

B.1.351 variant (South Africa)

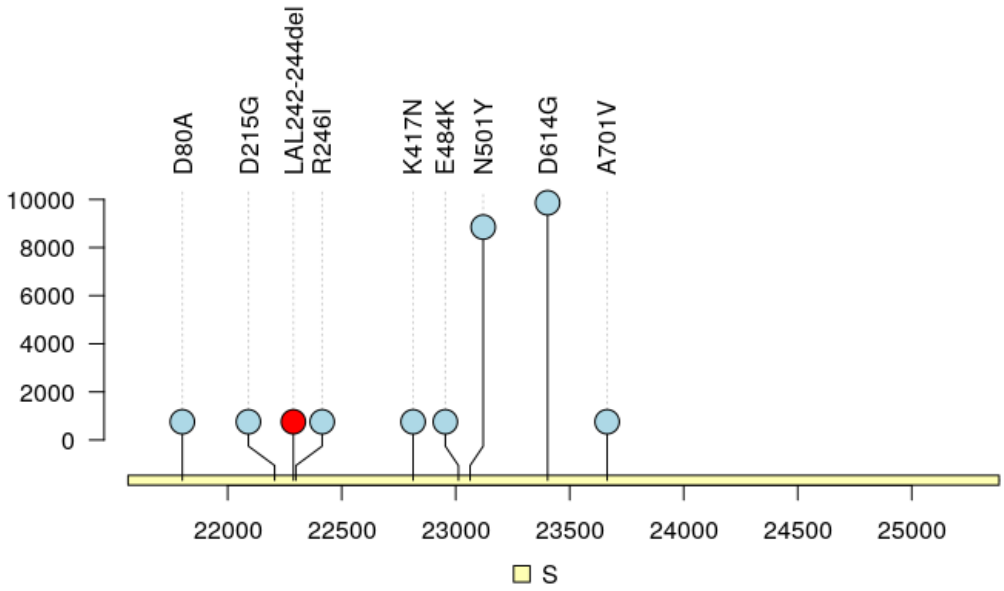


Figure VIII: Mutations in the spike protein defining variant B.135.1. This variant was not detected in the data uploaded since January.

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

The only B.135.1 lineage samples from Europe are 552 from the UK, 1 from Greece, 57 from Estonia and 3 from Spain, but these are hardly visible against the large number of background sequences.

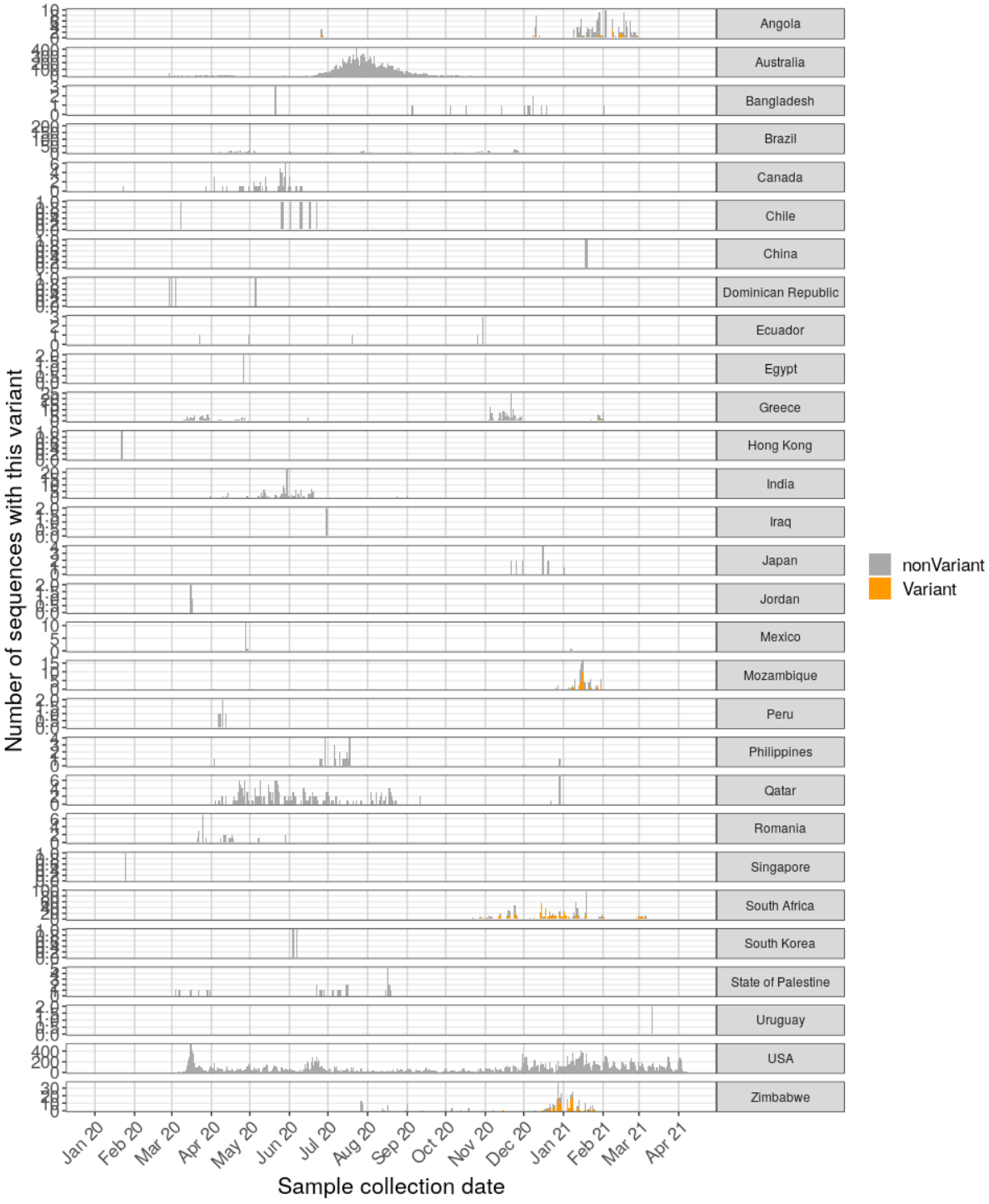


Figure IX: Number of sequences by date of sampling for variant B.135.1(orange) for non-European countries.

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

B.1.1.28.1 (P1) variant (Brazil)

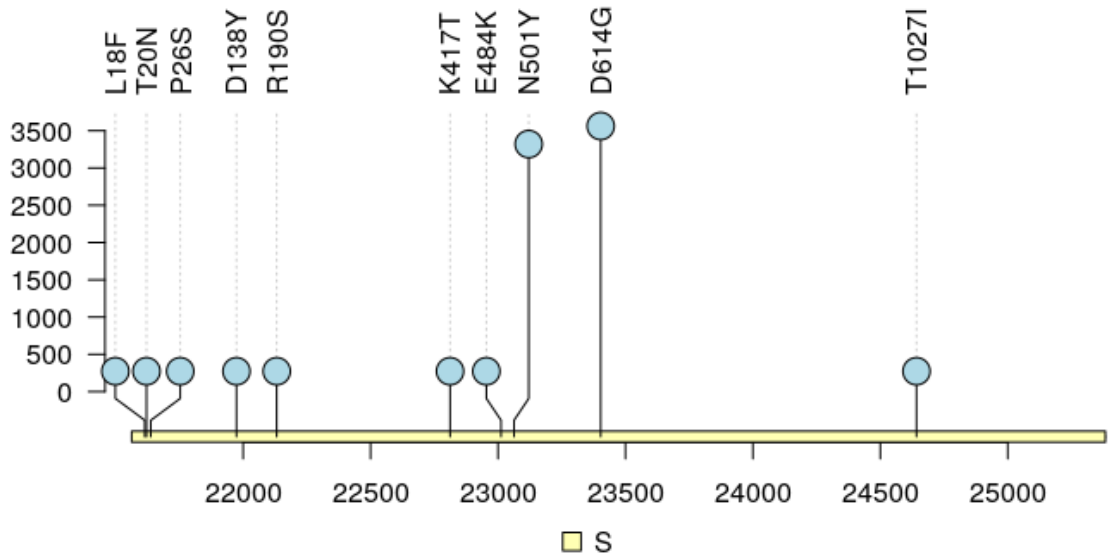


Figure X: Mutations in the spike protein defining variant P1.

B.1.1.28.1 was found in 58 UK samples, 3 Italian samples and 2 Spanish samples, which does not show up clearly in the bar charts. Outside of Europe, there is one sample in Japan and 81 in the USA, and one in Uruguay.

B.1.617

Samples containing all B.1.617 lineage defining spike protein mutations (G142D, E154K, L452R, E484Q, P681R, Q1071H) have not been found. However, 210 samples from the UK and 2 samples from the USA contained both the L452R and the E484Q mutation commonly known as the “double mutation”, which has been associated with immune escape and increased infectivity of the B.1.617 variant.

Oxford Nanopore sequencing data

While the majority of raw sequencing data are based on Illumina technology, an increasing amount of data are based on Oxford Nanopore sequencing (Table I). Automatically identifying minority variants in these data is more complicated than from Illumina data due to the higher error rate. While we are still evaluating and improving the bioinformatic workflows, we have implemented a first workflow for variant analysis to use and will start soon with its implementation and processing of the Nanopore read data.

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 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.

VEO will continue to analyse all publicly shared Illumina data for presence of variants. In addition, an Oxford Nanopore VCF calling workflow will be implemented soon, after which the backlog of Nanopore data can be processed, and variants derived from Nanopore data can be added to the variant database. We will continue to evaluate and further improve this in the coming periods. 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.

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).