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

Report No. 1 – February 16th 2021

Summary:

- Mobilisation of raw SARS-CoV-2 sequencing data sets from almost 300,000 viral isolates from 59 countries.
- 60,000 assembled sequences from 78 countries.
- First version visualisation of mutations in 31,474 sequences
- Several variants including mutations in sites of immunological relevance were only observed as minority variants and would be missed when only looking at assemblies

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 GISAID data.

Data mobilisation

We have mobilised substantial data via the SARS-CoV-2 Data Hubs and made these available from the COVID-19 Data Portal https://www.covid19dataportal.org. (See Figure I). In addition to raw data, we have also mobilised some 60,000 assembled sequences from 78 countries.

Submitted raw reads are processed through a number of high-throughput computational workflows based on expert analytical processing developed by VEO partners. A first workflow allows assessment of quality metrics. QC-ed datasets are further processed for variant calling. Below are first results of a workflow that allows systematic variation calling from Illumina amplicon data and assembly.

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

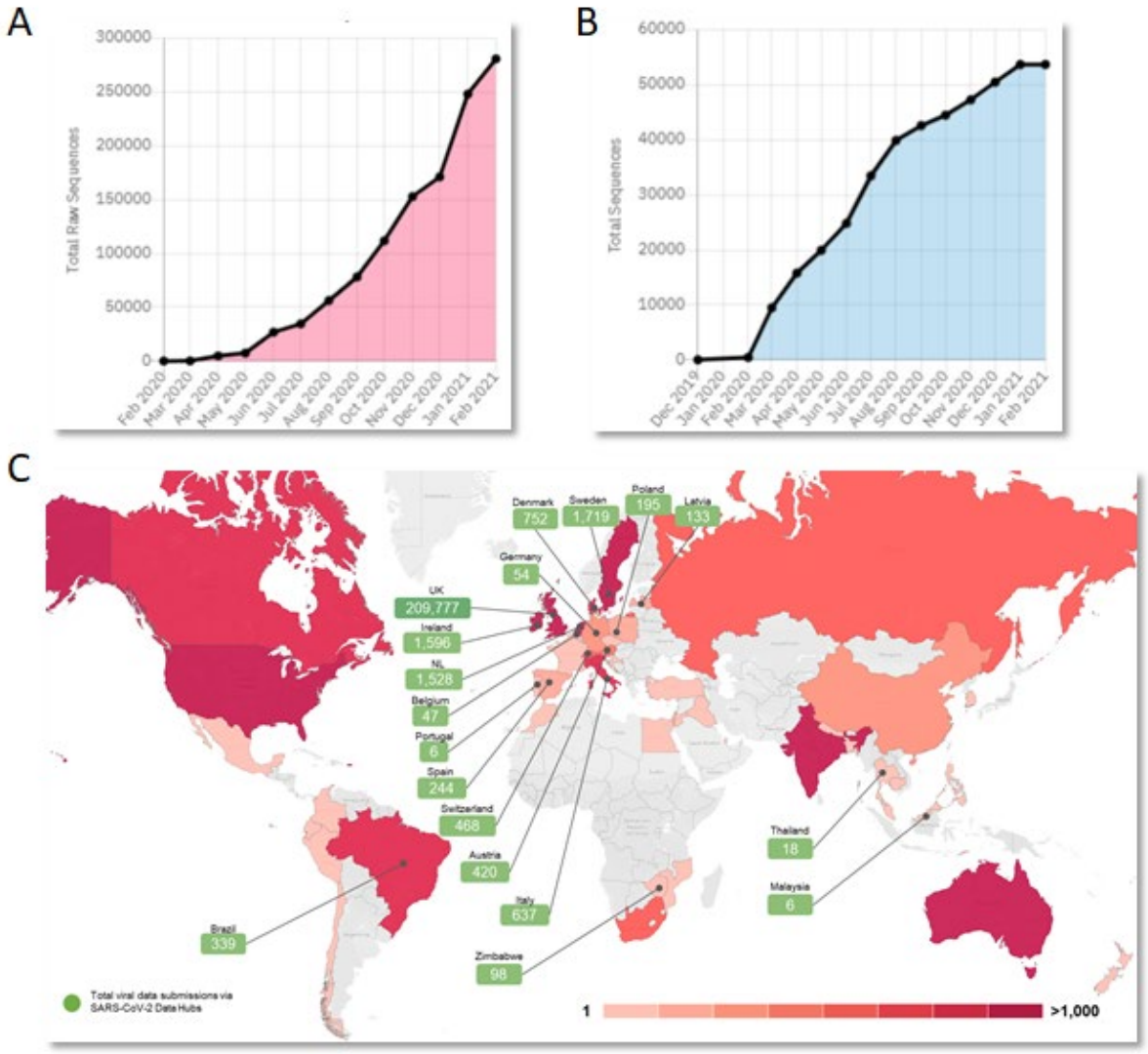


Figure I. Growth of raw SARS-CoV-2 data and distribution of sources, showing (A) sustained growth in raw data since launch of the mobilisation campaign, (B) growth in assembled sequence data and (C) geographical sources of raw data for which 75% of global data have been routed through the SARS-CoV-2 Data Hubs (see green boxes), with the remaining 25% arriving into the Platform from collaborators in the US and Asia.

Results of variant calling, first version visualisation tool

A workflow to analyse the submitted data has been established, and at this stage, 31,474 variant sets have been analysed. Now that the workflow has been established, the data will be more frequently updated in the future.

Examples of lineage defining mutations

With the evolving pandemic, the genomic diversity of SARS-CoV-2 has increased. Systems for standardized nomenclature have tracked the number of evolving lineages and identified lineage defining mutations. An example of that is given below. Lineage defining mutations are as defined by Nextstrain (https://nextstrain.org/).

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

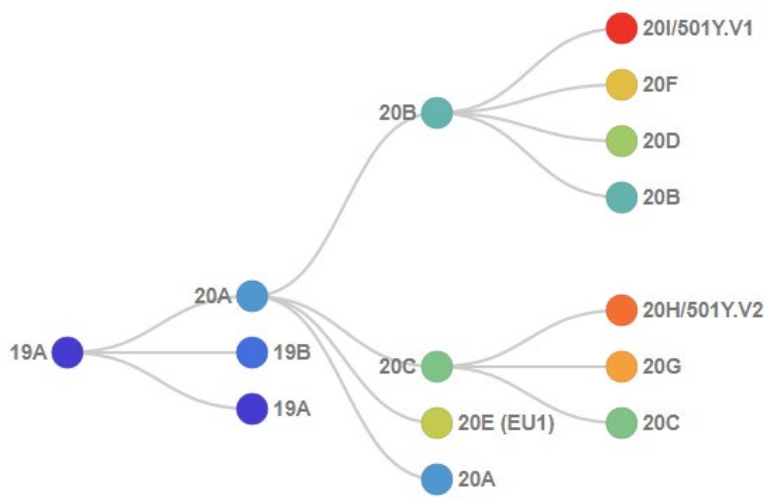
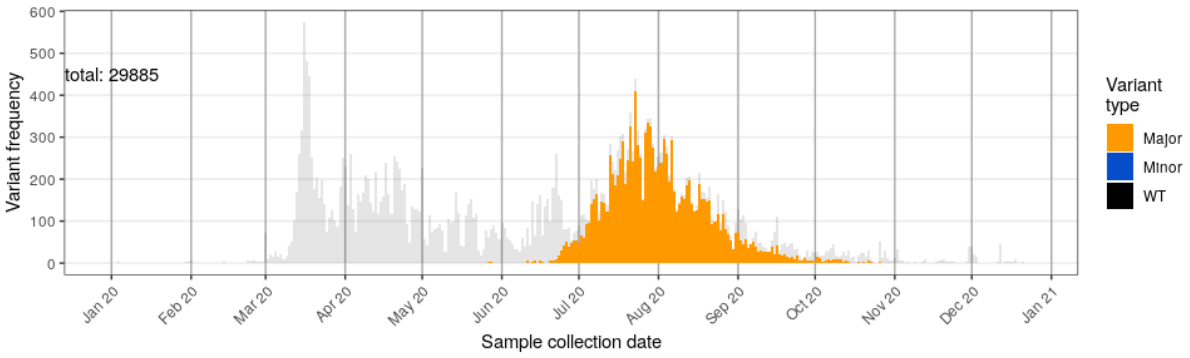


Figure II: Lineage defining mutations as determined by Nextstrain.

Nextclade 20F 1163T mutation



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

Overview of the top 25 minor variant mutations

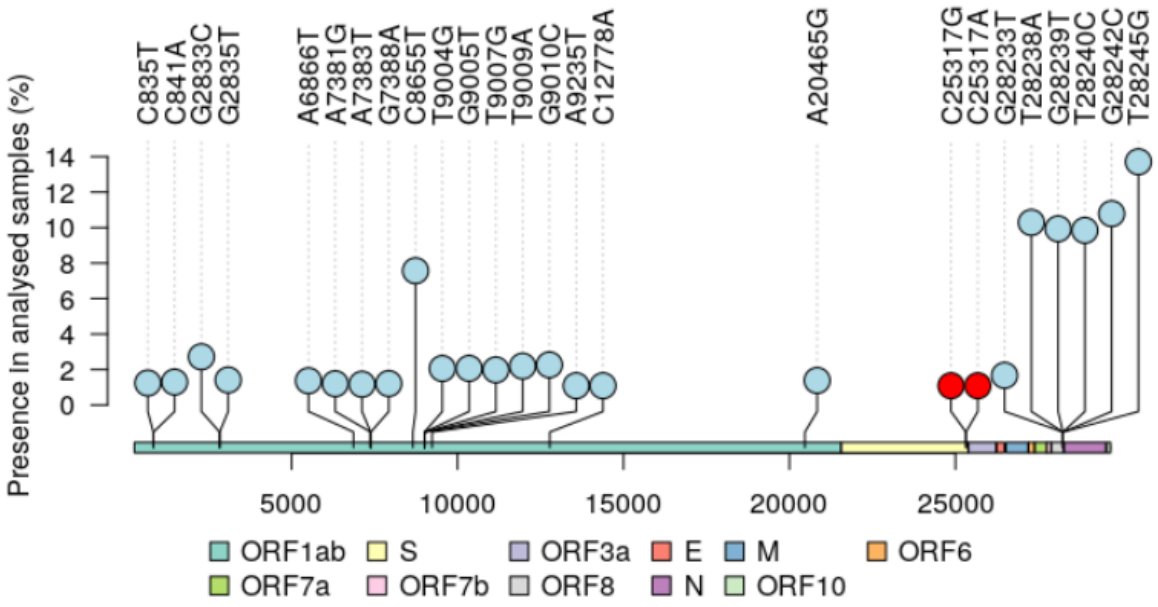


Figure III. Plot of the 25 most abundant (undetected) minor variants in the SARS-CoV-2 genome. Mutation in the spike region is highlighted by red.

Minor variant mutations in the spike protein

In several locations in the spike protein minor variant mutations have had a significant role in the total population of sequence data. Below are two examples, one example of a "silent" mutation and one example of a mixed minor/major variant mutation.

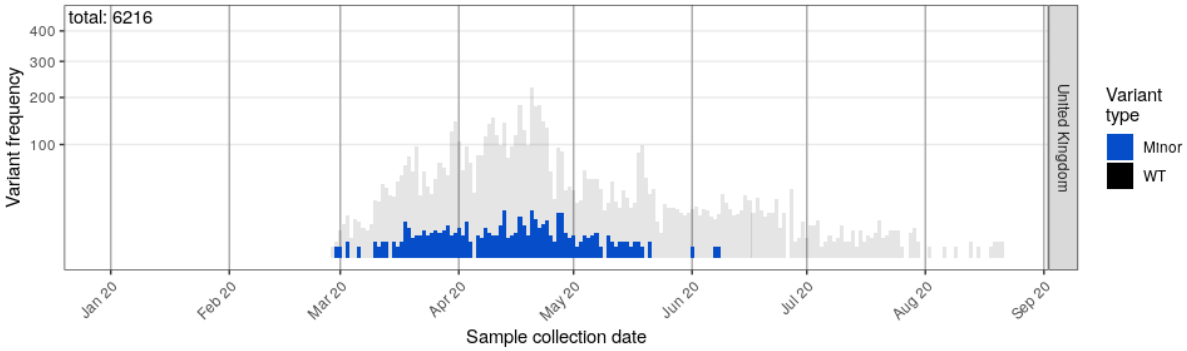


Figure IV. The spike mutation at position 25317 C>G (S 1252 Ser>Cys) in the first outbreak wave in the UK has been present as a minor variant only and is not detected as a major variant.

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

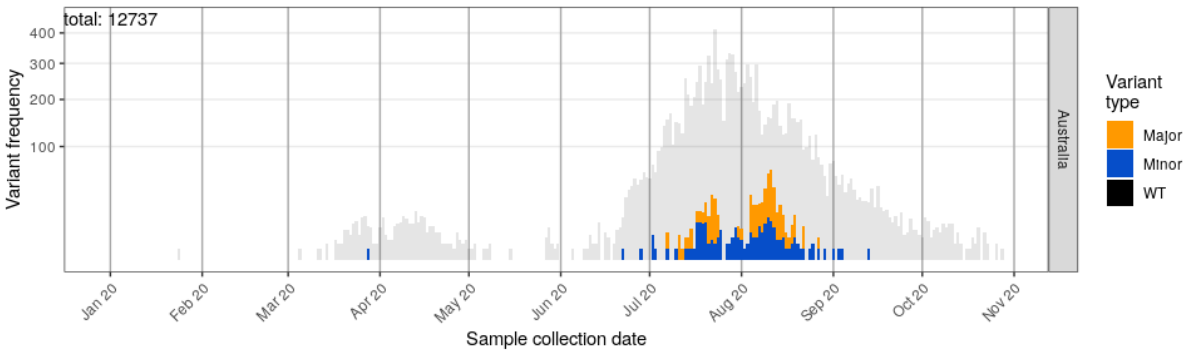


Figure V. The spike mutation at position 24933 G>T (S 1124 Gly>Val) existed as a mixture of major and minor variants in the Australian outbreak from June to November.

Abundant minor variants outside of the spike region

Apart from the spike region of the SARS-CoV-2 genome several mutations have been present as minor variants in the rest of the genome. Depicted in the figure below is a mutation at position 28245 T>G (ORF8 118 Leu>Val) which hardly shows up as a major variant, but is present in most sequences originating from the first outbreak wave in the UK and the US.

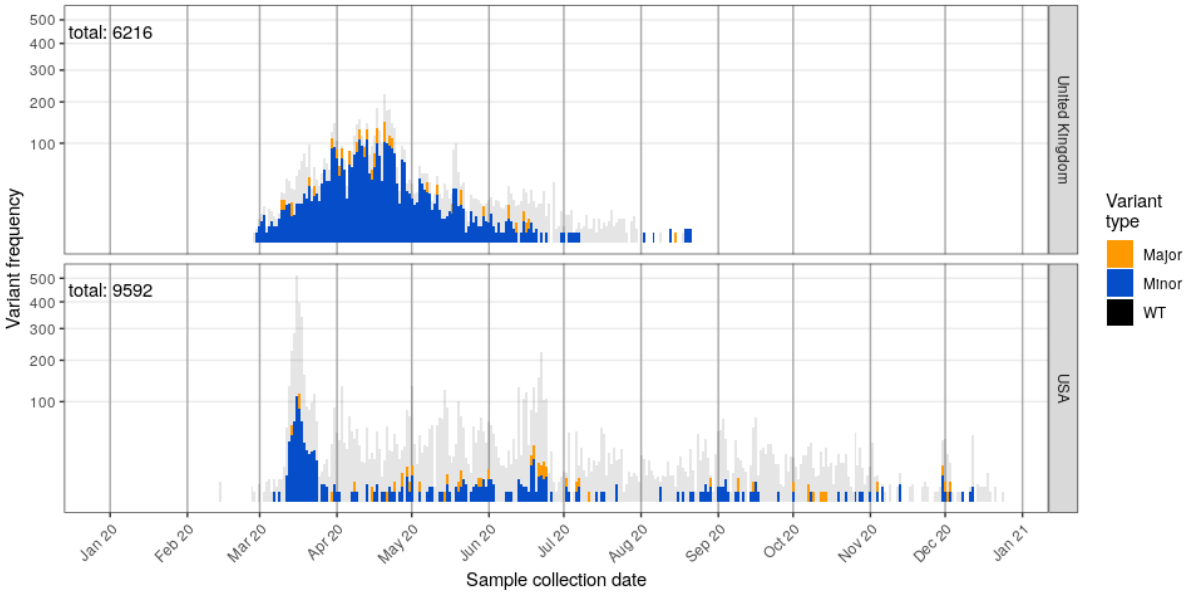


Figure VII. The most abundant minor variant mutation until now is located at position 28245 T>G (ORF8 118 Leu>Val) and was present in most of the SARS-CoV-2 strains in the first UK and US outbreaks, but existed predominantly as a minor variant.

Recommendations and next steps:

VEO will in the coming weeks process the remaining raw data and provide more in depth analyses of mutations across all data. VEO is also working on providing visualization of all data for easy overview of the data. In addition, a combined analytic workflow providing combined analyses for data from different sequencing platforms will be created.

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

The initial analyses already show that information is missed when only analysing assembled data.

The value of the analyses provided greatly depends on the amount of data that are shared as raw data. Thus, we strongly urge EU member states and all scientists and health authorities around the world to rapidly and openly share their raw sequencing data.

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).