**Supplementary Material**

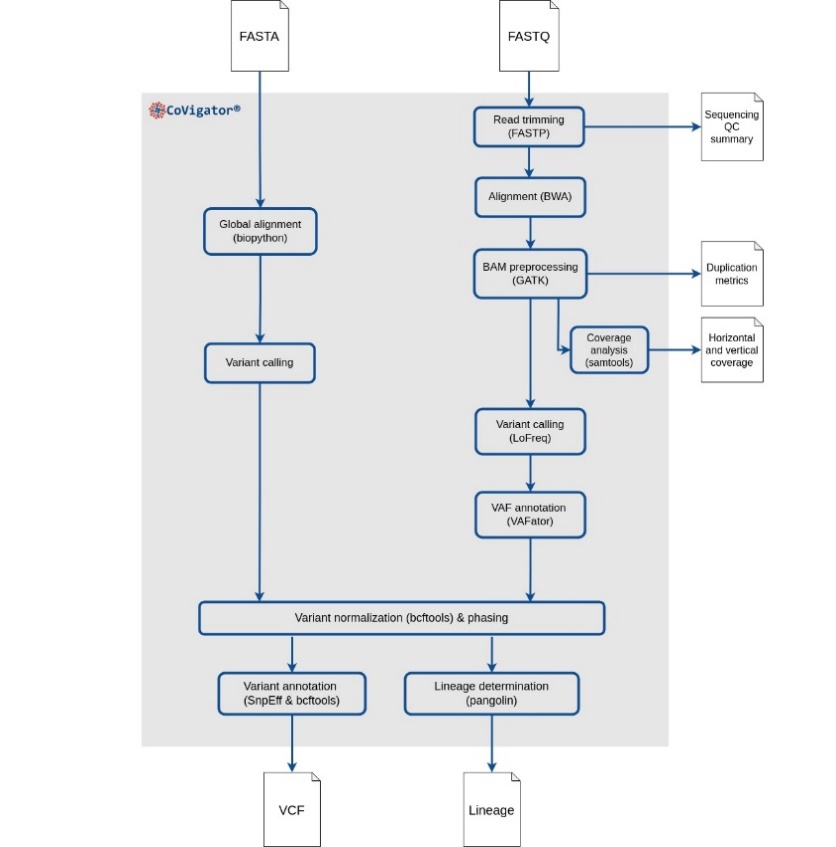
CoVigator - a knowledge base for navigating SARS-CoV-2 genomic variants

**Thomas Bukur 1&, Pablo Riesgo-Ferreiro1&, Patrick Sorn1, Ranganath Gudimella1, Thomas Rösler1, Martin Löwer1\*, Barbara Schrörs1 & Ugur Sahin1,2**

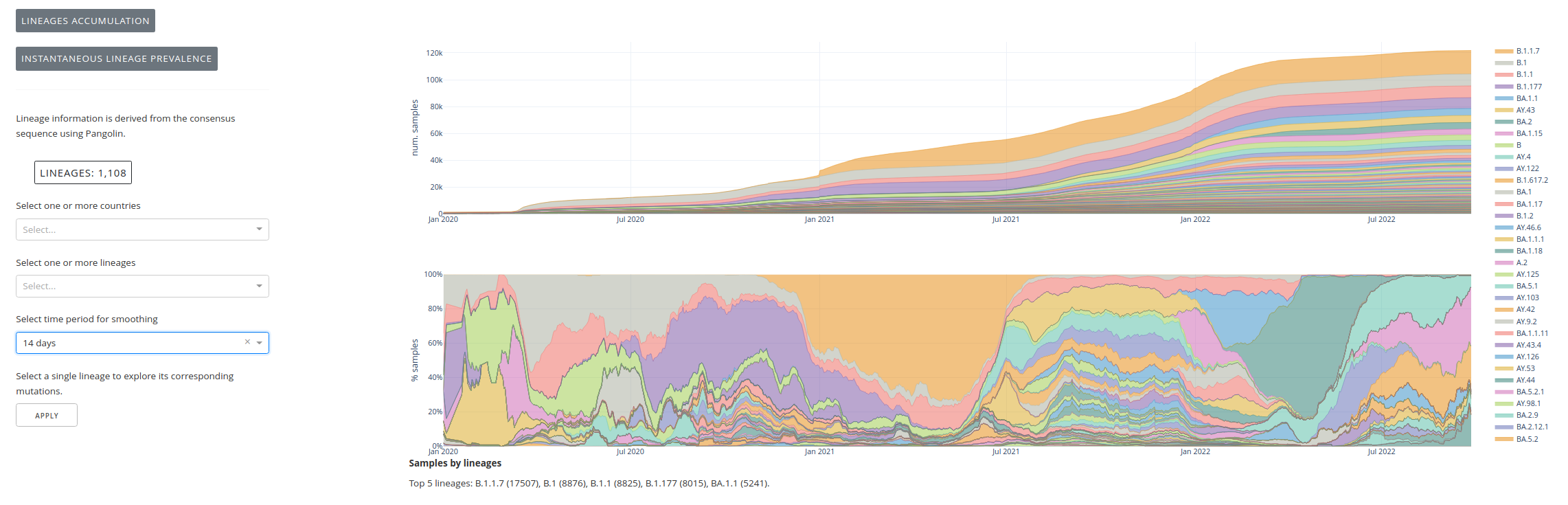
1. Translational Oncology at the Medical Center of the Johannes Gutenberg-University Mainz gemeinnützige GmbH, Mainz, Rhineland-Palatinate, Germany
2. BioNTech SE, Mainz, Rhineland-Palatinate, Germany

**\***Corresponding Author: martin.loewer@tron-mainz.de

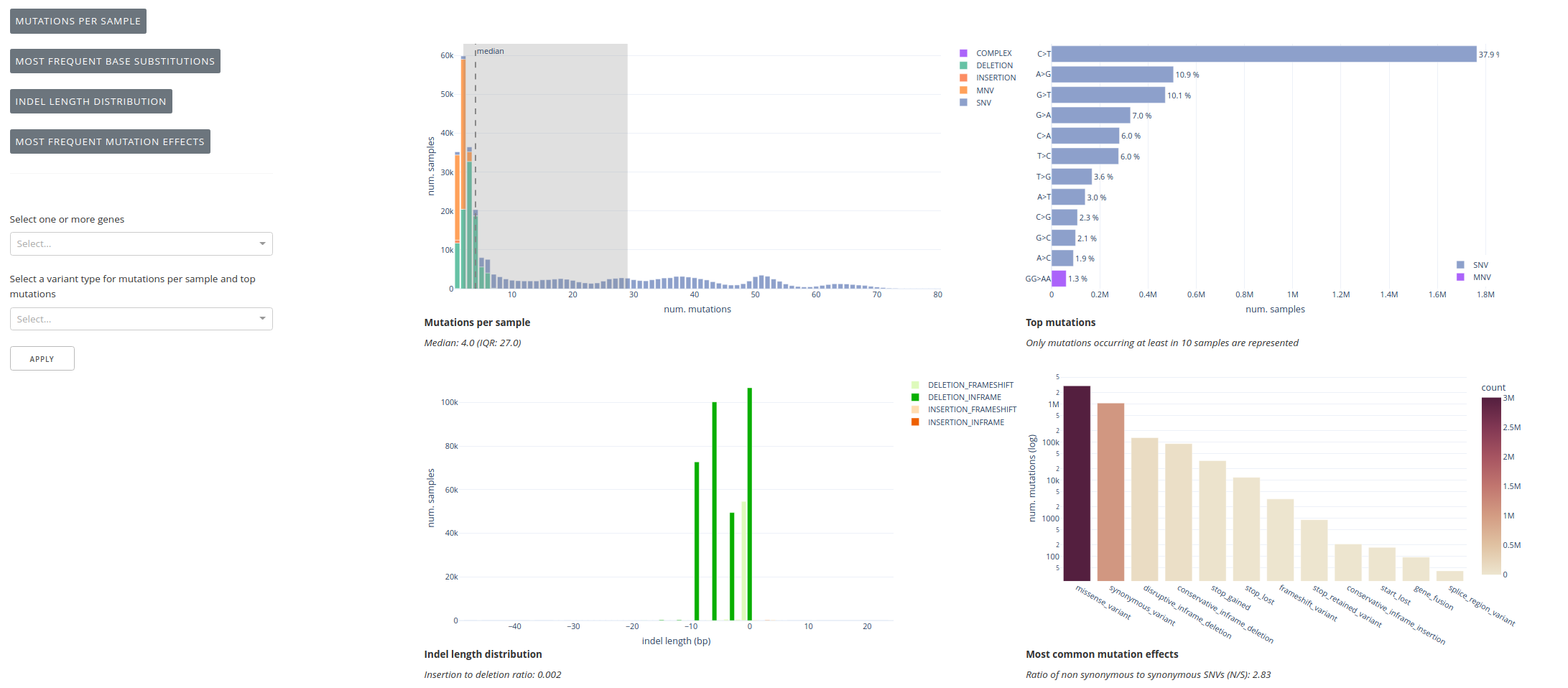
& Contributed equally to this work



**Figure S1**. Workflow of CoVigator pipeline. The input can be either FASTQ or FASTA files. The output are multiple VCF files (one per variant caller), a Pangolin lineage file, horizontal and vertical coverage results and the FASTP sequencing summary. Optionally, the BAM file and other intermediate files may be kept in the output.



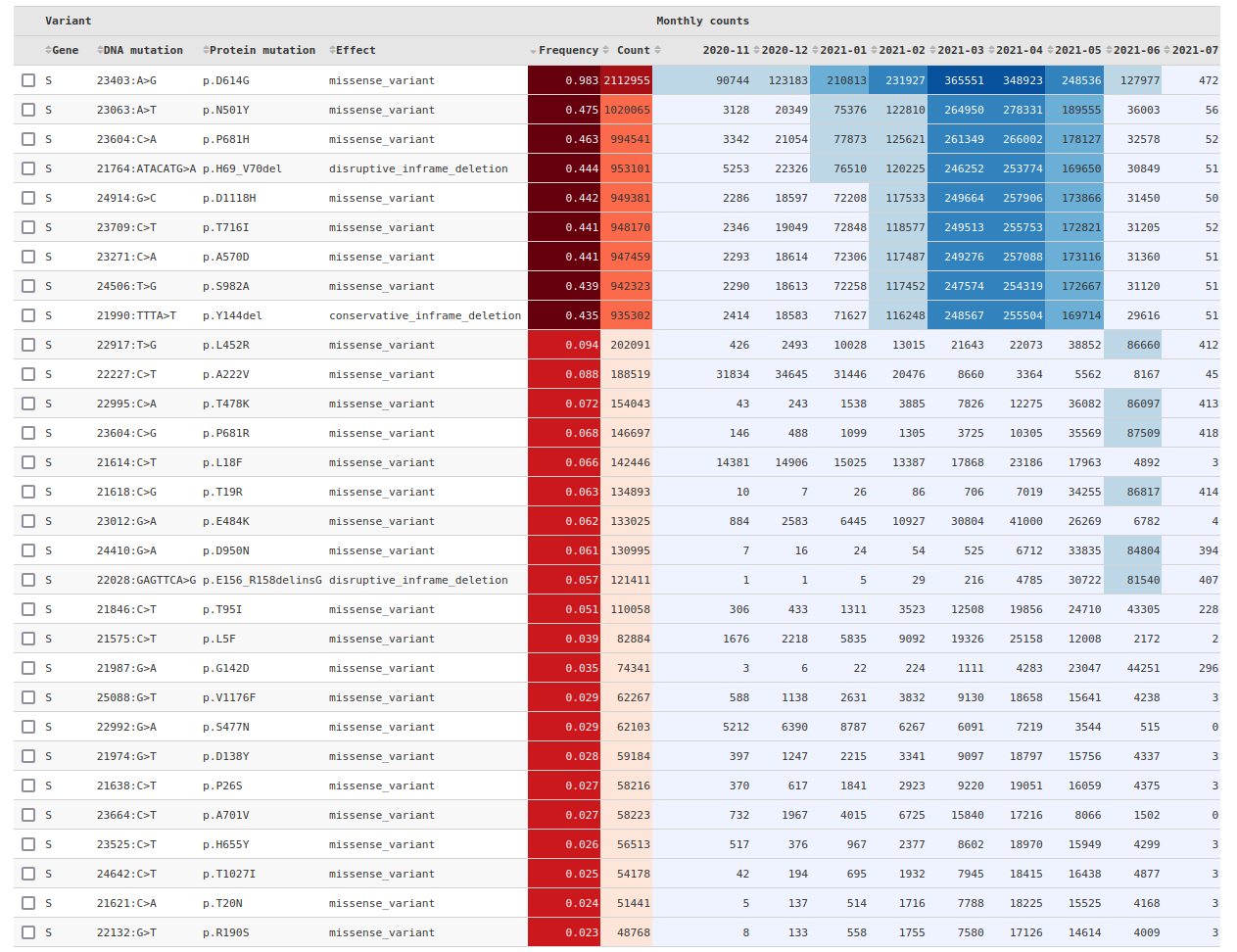
**Figure S2**. Screenshot of the lineages tab in the ENA dataset. On the left the filters for countries and lineages. On the top right the accumulation through time of samples by lineage. On the bottom right the percentage of samples by lineage.



**Figure S3**. Screenshot of mutation statistics tab in the ENA dataset. On the left filters for genes and variant types. On the top left the distribution of the number of mutations per sample. On the top right the most frequent substitutions. On the bottom left the indels length distribution. On the bottom right the distribution of the different types of mutations.



**Figure S4**. The recurrent mutations tab for the spike protein in the ENA dataset. On the left the filters gene, protein domain, number of top occurring mutations, abundance of mutation metric, date range, bin size for the genome view, pairwise co-occurrence metric, minimum pairwise co-occurrences and the minimum number of samples in the neighborhood to consider a mutation a core point during clustering. On the top, right the top ten occurring mutations in the spike protein with the number of observations per month in the defined date range. On the bottom left a view of the spike protein including mutations by their frequency in the ENA dataset, conservation metrics and protein domains.



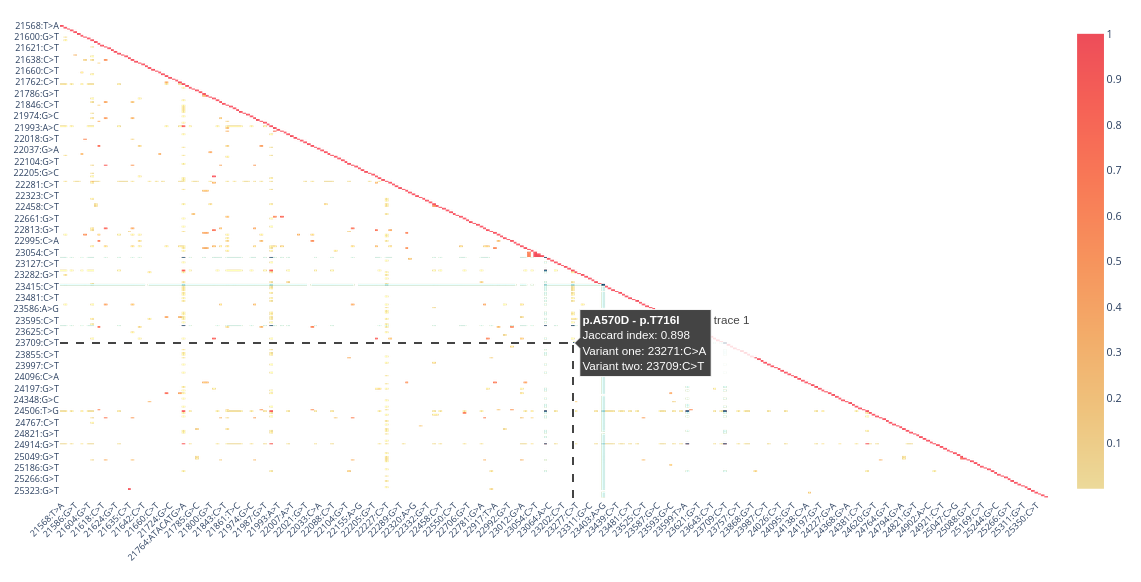
**Figure S5**. Top 30 mutations in the spike protein from COVID-19 Data Portal. We can observe the decline and raise of the alpha and delta lineages mutations since May 2021. The month ranges are configurable from the dashboard and the raw data in CSV format is downloadable.

**Supplementary Method for co-occurance analysis**

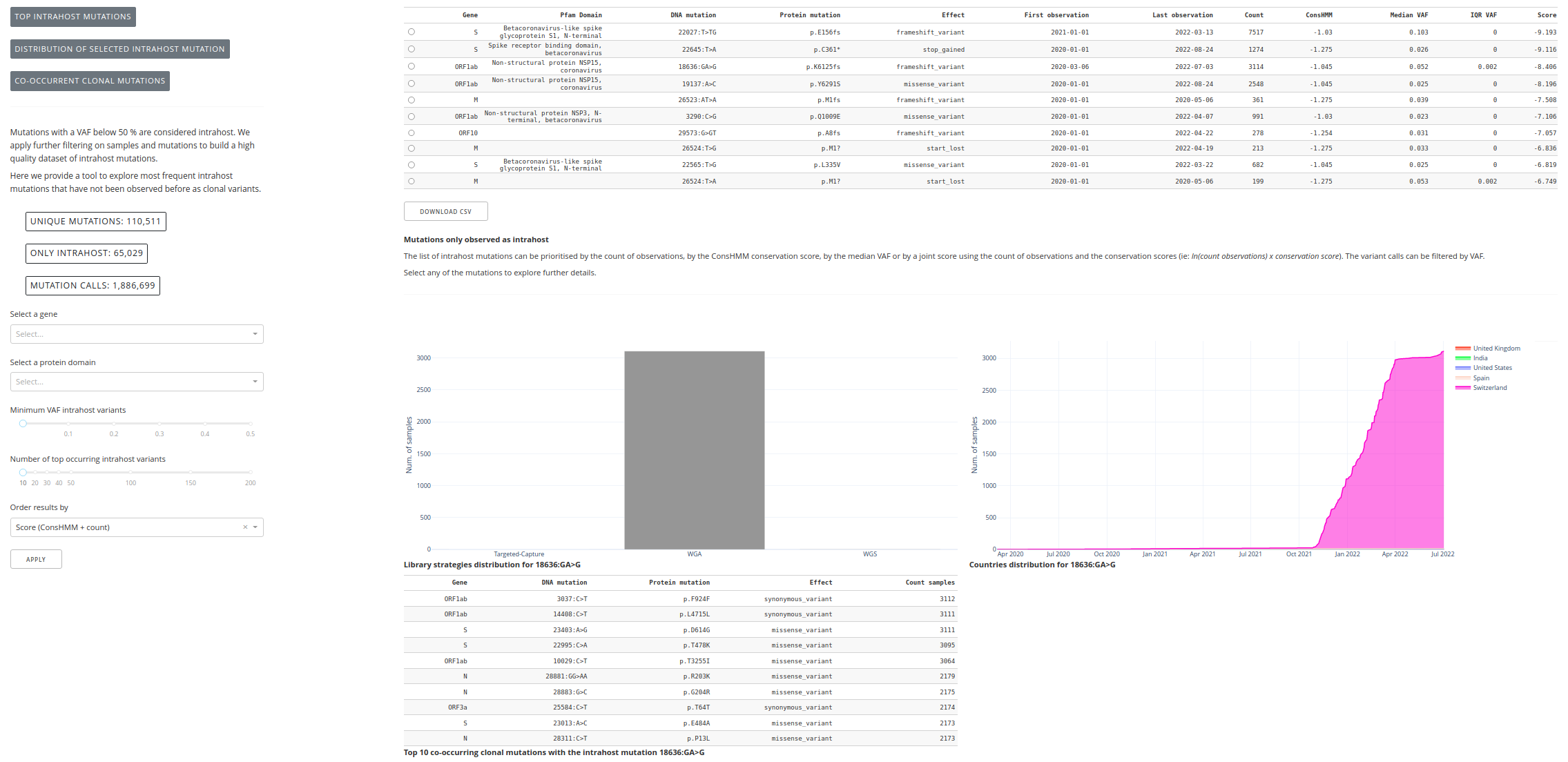
The co-occurrence analysis employs the similarity matrix with Cohen’s kappa coefficient between every pair of variants.

Where a and b are the count of observations of two variants. The intersection of a and b are the co-occurring observations of both variants and the union of a and b are the count of samples on where any of the two variants were observed without repetitions.

Clustering is performed using the OPTICS algorithm [1] as implemented in scikit-learn library version 0.24.2 [2] with max\_eps=1.4 and min\_samples=5, although the latter can be parametrized in the dashboard. Other parameters are set to default. Even though in the dashboard the heatmap can show different similarity metrics the clustering is always performed with Cohen’s kappa coefficient.



**Figure S6**. Jaccard index co-occurrence matrix on the spike protein

**

**Figure S7**. The intrahost mutations tab helps to prioritize recurrent intrahost mutations that undetected before in clonal state. When selecting a particular intrahost mutation it shows the distribution of this mutation over time, geography and different library strategies. Also, identifies the clonal mutations with which it co-occur

**Table S1**. Co-occurrence clusters on the spike protein with its matching lineage information (SARS-CoV-2 lineages fetched from <https://github.com/cov-lineages/constellations>). Clustering is performed considering only synonymous mutations with at least 10 pairwise co-occurrences. The minimum number of neighbours to call a cluster is five. Pairwise co-occurrence is measured with the Jaccard index corrected with the Cohen’s kappa coefficient. These parameters are configurable in the dashboard and the clustering can be applied over other genes or domains. The column “Cluster description” contains a manual annotation of the cluster based on resemblance to a particular given lineage. Only cluster 8 containing six mutations is not annotated with any lineage. The clustering results can be downloaded in a raw CSV from the dashboard.





**References**

1. Ankerst, M.; Breunig, M.M.; Kriegel, H.-P.; Sander, J. OPTICS. SIGMOD Rec. 1999, 28, 49–60, doi:10.1145/304181.304187.

2. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011, 12, 2825–2830.