**Tracking SARS-CoV-2 mutations & variants through the COG-UK-Mutation Explorer**

Derek W Wright1, William T Harvey1, Joseph Hughes1, MacGregor Cox2, Thomas P Peacock3, Rachel Colquhoun4, Ben Jackson4, Richard Orton1, Morten Nielsen6, Nienyun Sharon Hsu9, The COVID-19 Genomics UK (COG-UK) consortium5, Ewan M Harrison2,7,8, Thushan I de Silva9, Andrew Rambaut4, Sharon J Peacock2, David L Robertson1#, Alessandro M Carabelli2#

1MRC-University of Glasgow Centre for Virus Research, University of Glasgow, Glasgow, UK

2Department of Medicine, University of Cambridge, Cambridge, UK

3Department of Infectious Disease, St Mary’s Medical School, Imperial College London, UK

4Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK

5https://www.cogconsortium.uk. Full list of consortium names and affiliations are in Appendix 1

6Universidad Nacional de San Martin, Argentina

7Wellcome Sanger Institute, Hinxton, UK

8Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

9The Florey Institute for Host-Pathogen Interactions and Department of Infection, Immunity and Cardiovascular Disease, Medical School, University of Sheffield, Sheffield, UK

#Contact: david.l.robertson@glasgow.ac.uk and amc257@medsch.cam.ac.uk

**Key words:** SARS-CoV-2; COVID-19; virus; spike; protein structure; antibody escape; antigenic variation; mutation; amino acid replacements; variants of concern; evasion; resistance; fitness; evolution

**Abstract**

COG-UK Mutation Explorer (COG-UK-ME, <https://sars2.cvr.gla.ac.uk/cog-uk/> - Last Accessed date 16/03/22) is a web resource that displays knowledge and analyses on SARS-CoV-2 virus genome mutations and variants circulating in the UK, with a focus on the observed amino acid replacements that have an antigenic role in the context of the human humoral and cellular immune response. This analysis is based on more than 2 million genome sequences (as of March 2022) for UK SARS-CoV-2 data held in the CLIMB-COVID centralised data environment. COG-UK-ME curates these data and displays analyses that are cross-referenced to experimental data collated from the primary literature. The aim is to track mutations of immunological importance that are accumulating in current variants of concern and variants of interest that could alter the neutralising activity of monoclonal antibodies (mAbs), convalescent sera and vaccines. Changes in epitopes recognised by T cells, including those where reduced T cell binding has been demonstrated, are reported. Mutations that have been shown to confer SARS-CoV-2 resistance to antiviral drugs are also included. Using visualisation tools, COG-UK-ME also allows users to identify the emergence of variants carrying mutations that could decrease the neutralising activity of both mAbs present in therapeutic cocktails, e.g., Ronapreve. COG-UK-ME tracks changes in frequency of combination of mutations and brings together curated literature on the impact of those mutations on various functional aspects of the virus and therapeutics. Given SARS-CoV-2 unpredictable nature as exemplified by yet another variant of concern, Omicron, continued surveillance of SARS-CoV-2 remains imperative to monitor virus evolution linked to the efficacy of therapeutics.

1. **Introduction**

As of March 2022, SARS-CoV-2, the causative agent of COVID-19, has accounted for over 450 million infections and 6 million deaths worldwide (<https://covid19.who.int/>). SARS-CoV-2 was first identified at the end of 2019 in the city of Wuhan, China and has since spread with unprecedented efficiency among humans (Hu et al. 2021). In contrast to other RNA viruses, the *Coronaviridae* family is characterised by relatively high replication fidelity due to the proofreading activity of their polymerases (Robson et al. 2020). Early analyses of SARS-CoV-2 genomes estimated an evolutionary rate of around 0.001 subs/site/year (2-3 mutations per month) (Duchene et al. 2020), however there is much deviation from this rate across the phylogeny with several outlier lineages, including variants of concern (VOCs), that have rapidly acquired several mutations at a much higher rate than this. The analysis of mutations from virus genome data is important for basic virology (Houldcroft et al. 2017), to identify evolutionary signals associated with mutations prior to experimental and real-world data on clinical outcomes or vaccine effectiveness, and to document and track changes that could alter the effectiveness of therapeutics. At present, almost 9 million genome sequences are now available via the GISAID Initiative, permitting near real-time surveillance of the unfolding pandemic (Shu and McCauley 2017, Meredith et al. 2020).

SARS-CoV-2 showed relatively inconsequential genetic change until late 2020 (MacLean et al. 2021). Subsequently, later months of 2020 were characterised by the emergence, across the globe, of VOCs possessing mutations that altered virus phenotype in terms of transmissibility and antigenicity (Harvey et al. 2021). Concurrently, shifts in the immune profile of the human population likely represented a change in selective environment evidenced by an increase in dN/dS ratios indicative at positive selection at codons across the genome and notable levels of convergence across the global phylogeny (Martin et al. 2021). The continuing emergence of SARS-CoV-2 variants exhibiting heightened transmissibility or antigenic novelty necessitates tools to detect, describe and track those antigenic changes and make this information accessible to researchers, public health agencies and drug and vaccine developers so that the information becomes actionable.

Since the beginning of the pandemic, several bioinformatics tools have been developed to analyse and generate outputs that support actionable information (e.g., Pangolin lineages <https://cov-lineages.org/index.html>; <https://filogeneti.ca/covizu/>; <https://outbreak.info>; COVID-19 CG [https://covidcg.org](https://covidcg.org/); <https://coval.ccpem.ac.uk/>; CoV-GLUE [http://cov-glue.cvr.gla.ac.uk](http://cov-glue.cvr.gla.ac.uk/), [https://nextstrain.org](https://nextstrain.org/) and [https://covariants.org](https://covariants.org/) - Last Accessed date 16/03/22). Although these tools have been essential for data curation, analysis research and public health impact (Hufsky et al. 2021), they have been mainly focusing on epidemiological aspects of the pandemic lacking the relevant information from the literature on the immunological effect of mutations.

This scientific need led us to create the COG-UK-Mutation Explorer (COG-UK-ME), a web resource that provides tracking of non-synonymous mutations in SARS-CoV-2 genome. COG-UK-ME is based on UK data and it has been developed by the COVID-19 Genomics UK (COG-UK) consortium - created to deliver large-scale and rapid whole-genome virus sequencing to local NHS centres and the UK government. COG-UK-ME relies on CLIMB-COVID, a data-centric bioinformatics environment for centralising UK SARS-CoV-2 sequences (Nicholls et al. 2021). Here, we describe COG-UK-ME and its main functionality. COG-UK-ME currently has around 5,000 users per month, with approximately 30% from the UK, 20% from the US and the remainder from other international locations.

COG-UK-ME has three aims: firstly, to make available amino acid mutations in a user-friendly way enabling data transparency; secondly, to report on amino acid variation present in SARS-CoV-2 sequences that have been shown to confer resistance against antibodies or disrupt T cell epitope binding. The third is to report on the emergence of new mutations that have the potential to reduce the effectiveness of some therapeutics that have been granted approval for use. Data accumulating over a time course can be analysed, so that trends can be detected and tracked.

1. **Data Analysis**

COG-UK-ME is a publicly accessible web resource that displays in-depth information and analyses of SARS-COV-2 virus genome mutations and variants. Sequence information is deposited daily on the MRC CLIMB-COVID platform (Nicholls et al. 2021), that has been generated by the COVID-19 Genomics UK (COG-UK) Consortium, Wellcome Sanger Institute, public health agencies and other approved providers. Virus lineages are assigned by using a phylogenetic framework to identify those lineages that contribute most to active spread (Rambaut et al. 2020, O’Toole et al. 2021). Mutations for UK sequences are then analysed on the CLIMB platform and linked with curated data on antigenicity, therapeutics and drug resistance. The prepared data files are then transferred from CLIMB to a web server and visualised.

**2.1 Tracking Changes in the Mutation Count**

COG-UK-ME shows a browsable dataset of all the amino acid sequence variation in SARS-CoV-2 protein sequences. These are shown for all data, and in the recent past – over the last 28 days – in the United Kingdom and in the four UK nations (England, Scotland, Wales and Northern Ireland) (“Mutation Counts” and “Mutations by week” tabs). The “VOCs and VUIs in the UK” tab shows through tables and visualisations the number of sequences of variants under investigation (VUI) and VOCs as designated by the UK Health Security Agency (formerly Public Health England) (<https://www.gov.uk/government/publications/covid-19-variants-genomically-confirmed-case-numbers/variants-distribution-of-cases-data> - Last Accessed date 16/03/22) (Figure 1a). COG-UK-ME also provides visualisations of the spike protein structure showing the position of the VOC defining mutations (Figure 1b). Data are also placed in their geographical context by showing the number of sequences and percentage of variants per region (Nomenclature of Territorial Units for Statistics – NUTS1) (“Geographical distribution” tab).

**2.2 Spike Profile Tracking**

In addition to tracking the frequency of individual substitutions across the genome and of lineages identified as VOCs or VUIs, changes in the frequency of combinations of spike amino acid substitutions are tracked. Each spike profile is defined as the combination of substitutions compared with the original genotype (Wuhan-Hu-1). Profiles may represent monophyletic lineages or they may have arisen convergently across the phylogeny. Changes in profile frequency over the latest 56-day period are considered. For currently circulating profiles (those sampled within the latest seven days), a sortable and searchable table includes information on the pango lineage(s) for which the profile has been associated, the number of substitutions comprising the profile, and the count of sequences across the latest 56-day and 28-day periods. The average growth rate (plotted on the y-axis in Figure 2A) is calculated as the mean percentage change in frequency between each 2-week period within the 56-day period. As growth rates are sensitive to potentially stochastic changes at very low frequencies, we also calculate a statistic that estimates recent expansion or contraction of each profile, calculated over the 56-day period (plotted on the y-axis in Figure 2B). For each profile, $i$, the absolute value for this statistic, $X\_{i}$, is calculated using the observed frequency, $O\_{i,j}$, of each profile, $i$, in each of the most recent two-week periods, $j$, according to

$$X\_{i}=\sum\_{}^{}\frac{(O\_{i,j}-E\_{i})^{2}}{E\_{i}}$$

where $E\_{i}$ is the frequency of profile $i$ over the full eight-week period under consideration. Thus, the value calculated is influenced by both the rate of change in profile frequency and the overall frequencies of a given profile and is more robust to stochastic differences in profile frequency that tend to occur at low frequencies.

This monitoring of spike profiles allows the detection of emerging, potentially advantageous, spikes that might not be detected by surveillance methods conditioned on mutations previously determined to be noteworthy through experimentation or other means. This simple approach is complementary to more sophisticated phylogenetic approaches for the estimation of lineage-specific growth rates. One advantage of this simple non-phylogenetic approach is that the convergent accumulation of a substitution or combination of substitutions on a particular background is identified. Such a scenario could arise when there is strong selective pressure on a genotype (e.g. the introduction of a therapeutic). For example, this approach would quickly alert to the growth of a profile such as Delta + E484K emerging convergently across the Delta phylogeny in response to within-host, immune-mediated selection, even if the instances of this profile are interspersed across the phylogeny.

**2.3 Antigenic changes**

The “Antigenic changes” tab shows a table listing all mutations in the spike protein present in the UK sequence dataset that have individually been associated with some significant degree of weaker virus neutralisation by convalescent plasma, post-vaccination sera, or SARS-CoV-2 spike-specific mAbs (referred to as "Escape mutations" in Figure 3). Alongside links to the associated literature for each substitution, a confidence score representing the weight of evidence associated with each substitution is shown: “high”, whenever the antigenic role of mutation is supported by multiple studies including at least one that reports an effect observed with (post-infection serum) convalescent plasma; “medium”, if the antigenic role of the mutation is supported by multiple studies; and “low”, when the mutation is supported by a single study (Figure 3). In the “VOCs + Antigenicity” tab, COG-UK-ME reports the occurrence of additional amino acid substitutions or deletions linked to antigenic change within each VOC (Figure 4). Relative proportions (expressed as percentages) of sequences carrying specific mutations can give information about the antigenic diversity within a VOC lineage.

 **2.4 T Cell Epitope Mutations**

Similar to the “Antigenic changes” tab, the “T cell epitope mutations” tab shows amino acid replacements in experimentally proven T cell epitopes both in spike and in other proteins which have been described in the literature. Data are further filtered based on experimental studies just defining T cell epitopes (“Epitope studies”) or those reporting on the impact of specific mutations on T cell recognition (“Reduced T cell recognition”). Also shown are predicted antigen presentation likelihood percentile rank values to the experimentally proposed HLA restriction element based on the NetMHCpan (CD8) and NetMHCIIpan (CD4) 4.1 algorithms (<https://services.healthtech.dtu.dk/service.php?NetMHCpan-4.1> and [https://services.healthtech.dtu.dk/service.php?NetMHCIIpan-4.0](https://services.healthtech.dtu.dk/service.php?NetMHCpan-4.1) - Last Accessed date 16/03/22)(Reynisson et al. 2020) for both the wildtype and mutant peptide variants. Here, peptides with predicted percentile rank scores of less than 2.0 for CD8 and less then 5.0 for CD4 are likely HLA binders. Amino acid replacements in any epitope are visualised through logo plots, in which each letter represents an amino acid replacement present in a specific epitope and its height represents residue frequency. The number below the sequence logo shows the position relative to the start position of the epitope.

 **2.5 Drug Resistance**

The “Drug resistance” and “Ronapreve” tabs show tables and visualisations for those mutations associated with resistance of SARS-CoV-2 to antiviral treatments (e.g. Remdesivir) and therapeutic mAb cocktails that are currently used in clinical settings (e.g. Ronapreve, cocktail of *casirivimab* and *imdevimab*) (Beigel et al. 2020, Sidebottom and Gill 2021). The UpSet plot in the Ronapreve tab allows users to track amino acid substitutions known to affect either *casirivimab* or *imdevimab* mAbs and in combination (Figure 5). Other therapeutics will be added in the future.

 **3. Concluding Remarks**

Bioinformatics resources such as COG-UK-ME play an important role by providing clear and accessible information to those who are tackling the pandemic, including through public health actions and the development of vaccines and therapeutics. COG-UK-ME is unique in presenting data from a densely sequenced population with an emphasis on publicly available data (bioproject accession PRJEB37886 and public alignments <https://www.cogconsortium.uk/tools-analysis/public-data-analysis-2/> - Last Accessed date 16/03/22). COG-UK-ME also brings together curated literature on the impact of mutations on various functional aspects of the virus. The COG-UK-ME interface allows users to track mutations that are a potential threat based on a phenotypic impact on virus biology or by conferring resistance to the human immune response, including that boosted by vaccines or antiviral drugs. Rapid analyses of VOCs, for example the accumulation of any mutation, can also be obtained from the COG-UK-ME interface. Of particular interest to researchers and for therapeutics, are mutations that either have an antigenic role or that affect T cell binding. These mutations are intensely monitored by researchers and Public Health Agencies to identify any new variant that could escape immunity generated by vaccines. Timely identification of VOC/VUI samples can facilitate access to clinical specimens to isolate live virus and serum for further immunological evaluation.

Although amino acid sequence analyses are not sufficient to determine the functional effect of a single mutation on SARS-CoV-2 fitness when taken in isolation, COG-UK-ME strives to collate all the available literature on SARS-CoV-2 mutations and provides data to support experiments that investigate the change in phenotype that these mutations might confer on variants.

 **4. Methods**

Throughout COG-UK-ME, Wuhan-Hu-1 (NCBI RefSeq NC\_045512) is used as the reference sequence for nucleotide coordinates, codon numbering within viral proteins, and wild-type amino acid assignments. Sequences are regularly uploaded onto the MRC-CLIMB platform. Sequences with quality issues are excluded. Amino acid replacements and in-frame indels in each sequence are identified (Nicholls et al. 2021).

Source code is available at <https://github.com/wrightdw/COG-UK-ME> (Last Accessed date 16/03/22).

**4.1 Data Preparation**

Sequence metadata files are processed on the CLIMB-COVID platform (Nicholls et al. 2021) using the R statistical programming language (Team 2021) and the Tidyverse collection of R packages (Wickham et al. 2019). Non-UK sequences are filtered out. Amino acid replacements and reference amino acids are counted for all time and for a 28-day period up to and including the latest sequence date, for the UK and the four UK nations. Counts are linked with data on antigenic changes, data on therapeutics, epitope data and predicted epitope binding percentile rank values. Counts of all amino acids across all positions in the spike protein are prepared for the visualisation of sequence logos.

PANGO lineage name aliases are resolved to the full lineage names, using the current designations at <https://github.com/cov-lineages/pango-designation> (Last Accessed date 16/03/22). VOC and VUI lineages are counted by day and by week, for the UK and the four UK nations, counting AY.x sub-lineages within the Delta VOC hierarchically. VOC lineages are also counted by week and by geographic region, according to the 12 Nomenclature of Territorial Units for Statistics first level regions of the UK (NUTS1). Antigenic amino acid replacements and deletions in the spike protein, are counted for VOC lineages, excluding lineage defining replacements, as defined by the UK Health Security Agency at <https://github.com/phe-genomics/variant_definitions> (Last Accessed date 16/03/22). Following data preparation, the resultant data files are transferred from CLIMB to a web server for visualisation.

**4.2 Literature search**

We searched PubMed, LitCovid, BioRxiv and MedRxiv using the search terms “SARS-CoV-2”, combined with “mAbs”, “monoclonal”, “convalescent”, “neutralisation/neutralization”, “epitope”, “antibody”, for studies published from January 2020 to July 2021, and manually searched the references of select articles for additional relevant articles (Figure S2). We also searched BioRxiv, and MedRxiv using combinations of the search terms; “COVID19”, “COVID-19”, “SARS-CoV-2”, “remdesivir”, “favipiravir”, “molnupiravir”, “nirmatrelvir”, “ritonavir”, “paxlovid”, “antiviral”, “binding”, “efficacy”, “effective”, “resistance”, “resistant”, “sensitivity”, “inhibit”, “evasion”, “mutation”, “variant”. Results reporting on SARS-CoV-2 mutations that cause resistance to antiviral drugs were recorded and published to the dashboard. This included many different types of assays and studies; neutralisation assays, receptor binding assays, clinical efficacy studies, transcriptional inhibition assays, and in silico indications of resistance. Antiviral drugs are included in the review if they are clinically approved somewhere in the world, or are in stage 3 clinical trials. This search is repeated each week, allowing the timely updating of the dashboard when new research arises.

**4.3 Data Visualisation**

The Shiny framework is used to create the COG-UK-ME web application, hosted in the Shiny Server environment (Chang et al. 2021). In order to maximise performance across multiple concurrent users, most values are pre-computed in the data preparation process on CLIMB, with the web application focussing on data visualisation.

The bar charts for VOC lineages and mutations, the geographical maps of VOC lineages and the scatter plot of spike profiles are generated using *ggplot2*, with interactive features added using *Plotly* (2015). The heatmap of antigenic changes in the spike protein is generated using the *ComplexHeatmap* package (Gu et al. 2016), using antigenic replacements and structural domain classifications. Amino acid replacements in epitopes are visualised as sequence logos using the *ggseqlogo* package (Wagih 2017). UpSet plots for mutations affecting Ronapreve are generated using the *UpsetR* package (Lex et al. 2014, Conway et al. 2017). The web application user interface is created using the *shinydashboard* (Chang and Ribeiro 2021), *shinydashboardPlus* (Granjon 2021), *shinyWidgets* (Perrier et al. 2021) and *shinyjs* packages (Attali 2020).

For the visualisations of the VOC spike mutations on the structure, The file 6vsb\_1\_1\_1.pdb containing a complete model of the full-length glycosylated spike homotrimer in open conformation with one monomer having the receptor-binding domain in the ‘up’ position was obtained from the CHARMM-GUI Archive (Woo et al. 2020, 2021). This model is itself generated based upon a partial spike cryo-EM structure (PDB ID: 6VSB (Wrapp et al. 2020)) For visualisation, the model was trimmed to the ectodomain (residues 14-1164) and the signal peptide (residues 1-13) and glycans were removed. Figures were prepared using PyMol (Schrödinger-LLC 2010).

**Acknowledgements**

COG-UK-ME is enabled by data from COG-UK and MRC-CLIMB. All UK data is available in GISAID and we thank all sequence generators for rapid data-sharing. We also acknowledge Omar Wagih for his support with his ggseqlogo sequence logo package used for visualising mutations in epitopes.

**Funding**

COG-UK is supported by funding from the Medical Research Council (MRC) part of UK Research & Innovation (UKRI), the National Institute of Health Research (NIHR) (MC\_PC\_19027)*,* and Genome Research Limited, operating as the Wellcome Sanger Institute. WTH and TPP are supported by the G2P-UK National Virology Consortium (MR/W005611/1) and WTH by the MRC (MR/R024758/1). DLR, DW and JH are funded by the MRC (MC\_UU\_12014/12) and DLR by the Wellcome Trust (220977/Z/20/Z).

**References**

(2015). "Plotly Technologies Inc. Collaborative data science." from <https://plot.ly>. Last Accessed date 16/03/22

(2021). "CHARMM-GUI Archive – COVID-19 proteins library." from <https://charmm-gui.org/?doc=archive&lib=covid19%202021>. Last Accessed date 16/03/22

Andreano, E., et al. (2021). "SARS-CoV-2 escape from a highly neutralizing COVID-19 convalescent plasma." Proceedings of the National Academy of Sciences **118**(36): e2103154118.

Attali, D. (2020). "shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds. ." from [https://CRAN.R-project.org/package=shinyjs](https://CRAN.R-project.org/package%3Dshinyjs).

Barnes, C. O., et al. (2020). "SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies." Nature **588**(7839): 682-687.

Baum, A., et al. (2020). "Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies." Science **369**(6506): 1014-1018.

Beigel, J. H., et al. (2020). "Remdesivir for the Treatment of Covid-19 — Final Report." New England Journal of Medicine **383**(19): 1813-1826.

Chang, W., et al. (2021). "Shiny: Web Application Framework for R." from [https://CRAN.R-project.org/package=shiny](https://CRAN.R-project.org/package%3Dshiny).

Chang, W. and B. B. Ribeiro (2021). "shinydashboard: Create Dashboards with 'Shiny'." from [https://CRAN.R-project.org/package=shinydashboard](https://CRAN.R-project.org/package%3Dshinydashboard).

Chi, X., et al. (2020). "A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2." Science (New York, N.Y.) **369**(6504): 650-655.

Conway, J. R., et al. (2017). "UpSetR: an R package for the visualization of intersecting sets and their properties." Bioinformatics **33**(18): 2938-2940.

Duchene, S., et al. (2020). "Temporal signal and the phylodynamic threshold of SARS-CoV-2." Virus Evolution **6**(2): veaa061.

Granjon, D. (2021). "shinydashboardPlus: Add More 'AdminLTE2' Components to 'shinydashboard'."

Greaney, A. J., et al. (2021). "Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies." Cell host & microbe **29**(3): 463-476.e466.

Gu, Z., et al. (2016). "Complex heatmaps reveal patterns and correlations in multidimensional genomic data." Bioinformatics **32**(18): 2847-2849.

Harvey, W. T., et al. (2021). "SARS-CoV-2 variants, spike mutations and immune escape." Nature Reviews Microbiology **19**(7): 409-424.

Houldcroft, C. J., et al. (2017). "Clinical and biological insights from viral genome sequencing." Nature Reviews Microbiology **15**(3): 183-192.

Hu, B., et al. (2021). "Characteristics of SARS-CoV-2 and COVID-19." Nat Rev Microbiol **19**(3): 141-154.

Hufsky, F., et al. (2021). "Computational strategies to combat COVID-19: useful tools to accelerate SARS-CoV-2 and coronavirus research." Briefings in Bioinformatics **22**(2): 642-663.

Lex, A., et al. (2014). "UpSet: Visualization of Intersecting Sets." IEEE transactions on visualization and computer graphics **20**(12): 1983-1992.

Li, Q., et al. (2020). "The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity." Cell **182**(5): 1284-1294.e1289.

Liu, Z., et al. (2021). "Identification of SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization." Cell host & microbe **29**(3): 477-488.e474.

MacLean, O. A., et al. (2021). "Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen." PLOS Biology **19**(3): e3001115.

Martin, D. P., et al. (2021). "The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages." Cell **184**(20): 5189-5200.e5187.

Meredith, L. W., et al. (2020). "Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study." Lancet Infect Dis **20**(11): 1263-1272.

Nicholls, S. M., et al. (2021). "CLIMB-COVID: continuous integration supporting decentralised sequencing for SARS-CoV-2 genomic surveillance." Genome biology **22**(1): 196-196.

Nicholls, S. M., et al. (2021). "CLIMB-COVID: continuous integration supporting decentralised sequencing for SARS-CoV-2 genomic surveillance." Genome biology **22**(1): 196.

O’Toole, Á., et al. (2021). "Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool." Virus Evolution **7**(2): veab064.

Perrier, V., et al. (2021). "shinyWidgets: Custom Inputs Widgets for Shiny."

Rambaut, A., et al. (2020). "A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology." Nature Microbiology **5**(11): 1403-1407.

Reynisson, B., et al. (2020). "NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data." Nucleic Acids Research **48**(W1): W449-W454.

Robson, F., et al. (2020). "Coronavirus RNA Proofreading: Molecular Basis and Therapeutic Targeting." Molecular Cell **79**(5): 710-727.

Schrödinger-LLC (2010). "The PyMOL Molecular Graphics System." from <https://www.pymol.org/>. Last Accessed date 16/03/22

Shu, Y. and J. McCauley (2017). "GISAID: Global initiative on sharing all influenza data - from vision to reality." Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin **22**(13): 30494.

Sidebottom, D. B. and D. Gill (2021). "Ronapreve for prophylaxis and treatment of covid-19." BMJ **374**: n2136.

Team, R. C. (2021). "R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.". from <https://www.R-project.org/>. Last Accessed date 16/03/22

Wagih, O. (2017). "ggseqlogo: a versatile R package for drawing sequence logos." Bioinformatics **33**(22): 3645-3647.

Wang, Z., et al. (2021). "mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants." Nature **592**(7855): 616-622.

Weisblum, Y., et al. (2020). "Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants." eLife **9**: e61312.

Wickham, H., et al. (2019). "Welcome to the {tidyverse}." Journal of Open Source Software **4**(43): 1686.

Woo, H., et al. (2020). "Developing a Fully Glycosylated Full-Length SARS-CoV-2 Spike Protein Model in a Viral Membrane." The Journal of Physical Chemistry B **124**(33): 7128-7137.

Wrapp, D., et al. (2020). "Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation." Science **367**(6483): 1260-1263.

**Figure Legend**

**Figure 1. (A)** Frequency plot showing the number of SARS-CoV-2 sequences per week for VOCs Alpha, Delta, Delta-AY4.2, Omicron BA.1 and BA.2, and “other” pre-VOC variants (see key) in the UK. The light grey box covering the two most recent weeks indicates a period in which sequence counts are low due to a lag (Figure S1); **(B)** Spike protein structure showing locations of Delta and Omicron specific spike mutations. Ectodomain of the spike homotrimer in open conformation with individual spike protein chains shown in different colours. On each monomer, highlighted spheres show the locations of amino acid substitutions, deletions (Δ), or insertions (ins) that distinguish the Omicron (BA.1) variant, relative to the original genotype (Wuhan-Hu-1). These are annotated on the monomer with an ‘up’ receptor-binding domain where they are highlighted in red on teal. The substitution D614G which is shared by common descent by all lineage B.1 descendants is italicised. The visualisation is made using a complete spike model (Woo et al. 2020) which is in turn based upon a partial cryo-EM structure (RCSB Protein Data Bank (PDB) ID: 6VSB (Wrapp et al. 2020)).

**Figure 2. Spike profiles sampled within 7 days of the latest UK sequence are summarised.** Each spike profile is a set of amino acid substitutions listed relative to the original genotype (Wuhan-Hu-1). Figure prepared with data compiled on 27th February 2022 with a most recent sequence date of 24th February 2022. **A**) Points represent spike profiles positioned by the number of sequences in the latest 28-day period and the average growth rate calculated over the latest 56-day period. Points are coloured by an expansion/contraction statistic that takes both the rate of change in frequency and the overall frequencies of a profile into account. The cursor is hovering to show information associated with BA.2 + P1162L. **B**) Points represent spike profiles positioned by the number of sequences in the latest 28-day period and the expansion/contraction statistic used to colour points in **A**. Here points are coloured to show profiles associated with Delta, Omicron (BA.1/BA.1.x) and Omicron (BA.2) variants. The cursor is again positioned to highlight the position of BA.2 + P1162L. The plot has been zoomed to focus on profiles with 28-day counts between 100 and 1,000 **C**) Searchable table sorted to show the four profiles in the UK with highest values in the expansion/contraction column. Further columns show profile numbers in the latest 28- and 56-day periods, and average growth rate.

**Figure 3.** Amino acid substitutions in the spike protein identified in the UK dataset (referred to as "Escape mutations") that have been associated with weaker neutralisation of the virus by convalescent or post-vaccination plasma/serum or spike-specific monoclonal antibodies (mAbs), or that have been observed to emerge upon exposure to either mAbs or plasma in laboratory experiments.

\*High confidence (red) refers to an antigenic role supported by multiple studies including at least one that reports an effect observed with (post-infection serum) convalescent plasma; “medium” (orange), if the antigenic role is supported by multiple studies; and “low” (yellow), when the mutation is supported by a single study. The boxes above the table enable filtering by multiple criteria.

**Figure 4.** Heatmap showing frequency of spike amino acid substitutions and a deletion with a potential or confirmed antigenic role on top of BA.2 through time. The labelled structural domains are indicated in the left side: SP signal peptide, NTD N-terminal domain, RBD receptor-binding domain, RBM receptor-binding motif, S2 subunit, FP fusion peptide. Residues are also coloured according to the class of antibody that binds to an epitope. RBD antibody classes 1-4 (Barnes et al. 2020) are depicted by colours: green (class 1: ACE2 blocking, bind open RBD only), yellow (class 2: ACE2 blocking, bind open and closed RBD), blue (class 3: non-ACE2 blocking, bind open and closed RBD) or yellow (class 4: non-ACE2 blocking, bind open RBD only). Residues described in an NTD epitope (Chi et al. 2020) are coloured in magenta (class 5). Each residue is also classified as having evidence for mutations either affecting neutralisation by mAbs(Baum et al. 2020, Li et al. 2020, Weisblum et al. 2020, Liu et al. 2021) or serum from previously infected individuals (convalescent plasma) (Li et al. 2020, Weisblum et al. 2020, Andreano et al. 2021, Greaney et al. 2021, Liu et al. 2021) or vaccinated individuals (Wang et al. 2021) and emerging upon exposure to mAbs (Baum et al. 2020, Weisblum et al. 2020, Liu et al. 2021) or plasma (Weisblum et al. 2020, Andreano et al. 2021) in laboratory experiments.

**Figure 5.** UpSet plot showing the counts of mutations affecting Ronapreve constituent mAbs that have occurred individually and in combinations (Lex et al. 2014). Occurrence is shown in the full UK SARS-CoV-2 genome sequence dataset (**A**) and in a dataset compiled of sequences in the latest 28-day period (**B**). Spike amino acid substitutions known to affect either casirivimab or imdevimab mAbs were considered. The upper histogram shows the number of sequences per mutation (dots) or combination of mutations (lines) and the bottom left histogram presents the number of sequences with each specific substitution. Rows are coloured according to the mAb to which the greatest fold-decrease in binding was recorded (blue = casirivimab, orange = imdevimab), with a lighter shade indicating a fold-decrease of less than 100 and darker shade indicating 100 or greater.