Analysis of SARS-CoV-2 spike protein mutations in the UK

Abstract

Introduction

**Background: COVID-19 pandemic**

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the strain of coronavirus responsible for the ongoing COVID-19 pandemic (Hu et al., 2021). SARS-CoV-2 was initially detected in Wuhan, China, and has now spread to all corners of globe (Huang et al., 2020). Approximately 770,000,000 people have contracted the virus and over 7,000,000 people have died due to complications caused by the virus (Anon). This figure is most likely even higher as a result of governments’ inability to report cases, as well as, corruption with countries attempting to hide the true values. An estimated value of the number of deaths sits at around 18,000,000 – 32,000,000 (Anon). The economic burden has been far-reaching, estimated losses are up to $16 trillion worldwide and unemployment levels have soared (Cutler and Summers, 2020). Nationwide lockdowns, that existed at the height of the pandemic, have not only had an economic effect, but, have also impacted education and literacy rates. It has been predicted that by 2030 less than 40% of pupils in the UK will achieve a pass in GCSE English and Mathematics (Major et al.). SARS-CoV-2’s effects will continue to resonate, even as the virus becomes less prevalent.

Coronaviruses are of the order Nidovirales, which is comprised of several families of related viruses (Fehr and Perlman, 2015). SARS-CoV-2 is a more recent virus that has impacted humanity, human coronaviruses (CoVs), HCoV-229E and HCoV-OC43, have co-existed for centuries (Pyrc et al., 2006). These viruses result in far more mild symptoms similar to that of the common cold. This is in stark comparison with severe acute respiratory syndrome coronavirus 1 (SARS-CoV-1), Middle East respiratory syndrome – related coronavirus (MERS-CoV) and severe acute respiratory coronavirus 2 (SARS-CoV-2), which are all highly pathogenic. This high pathogenicity exists due to the viruses targeting important cellular tissues, bronchial epithelial cells and tissue in the upper respiratory tract **(source)**. SARS-CoV-1, MERS-CoV and SARS-CoV-2 have emerged over the past 25 years, with several outbreaks occurring. Most notably the SARS-CoV-1 2002-2004 outbreak, 2015 MERS outbreak in South Korea and the COVID-19 pandemic as a result SARS-CoV-2 **(source)**. There appears to be a growing trend in the in the prevalence of coronaviruses outbreaks, even though the rates of SARS-CoV-2 have greatly diminished. This stresses the importance of studying SARS-CoV-2 and other related to viruses, so that globally we are better prepared to combat the next epidemic or pandemic.

**Background: SARS-CoV-2 structural and molecular information**

**Background: Spike glycoprotein**

**Aims of the investigation:**

Analysis was completed on x UK SARS-CoV-2 spike protein mutations that had been deposited in y.

1. **Investigate the clustering and distribution of SARS-CoV-2 spike protein mutations in the UK**
2. **Identify and characterise common spike mutations + non vs syn mutations**
3. **Analyse their potential effects on viral infectivity**, link to receptor binding domain/receptor binding motif

**Methods overview**

1. **Summarise the dataset used**
2. **Mention of the clustering techniques used, R-based analysis, various packages used and tools used for visualisation**

**Summary of the main findings**

Results

**Introduction to the dataset**

Figure 1, Number of sequences per day

Figure 2, Number of mutations over time

Use of linear regression model used, what does show?

Histogram of the most common number of mutations per sequence

Figure 3, Number of unique mutations over time

A graph of different colored lines

AI-generated content may be incorrect.

**Figure 1**. The number of sequences per day from 2020 – 2024. **A** The overall the number of sequences per day from years 2020 – 2024. The coloured zones represent the emergence of a new strain of SARS-CoV-2. Yellow – beta, red – alpha, purple – delta, pink – gamma, green - omicron. Also on the plot are specific dates showing the first use of that vaccine. A generalised linear model has been used. **B** The number of sequences per day for 2020. **C** The number of sequences per day for 2021. **D** The number of sequences per day for 2022. **E** The number of sequences per day for 2023. **E** The number of sequences per day for 2023. **F** The number of sequences per day for 2024.

Want to change this plot, to have curved lines showing the emergence of new strains

A colorful chart of a graph

AI-generated content may be incorrect.

Figure 1:

What is interesting about the figure?

* The figure clearly shows the various SARS-CoV-2 waves over time
* Increase in sequencing activity around the emergence of a new SARS-CoV-2 variants and decreased as COVID-19 cases declined
* Reflects the rapid spread of these new variants and the need to investigate these new variants
* Reduced sequencing efforts after mid-2022 to 2024
* SARS-CoV-2 becomes less prevalent
* People stop testing regularly
* More people are vaccinated and immune to the virus
* Still a baseline of testing continues
* Omicron takes over as the most prevalent variant of SARS-CoV-2
* Lockdowns in the UK
* 26th March 2020 – June 2020
* 5th November 2020 – 2nd December 2020
* 6th January 2021 – 8th March schools open 2021
* 2020 (B) gradual increase in sequencing activity
* 2021 (C) rapid increase in the number of sequences per day, coincides with the emergence of new variants, such gamma and omicron
* 2022 (D) sharp increase in sequencing activity, peaking at the start of the year, this peak is likely due to the emergence of omicron variant, followed by a sharp decline in the number of sequences per day
* 2023 (E) overall sequencing activity is very low
* 2024 (F) overall sequencing activity is also very low, indication of reduced viral spread

**Clustering and the distribution of mutations**

1. **Geographic distribution of spike protein mutation**

Figure, map of the distributions

1. **Map visualising the distribution in the UK**

**Mutation analysis**

What I did?

1. **Summary of the most common mutations**

**Figure x**. The top 20 most common spike protein mutations. **A** Plot of the most common spike protein mutations. Blue colour indicates nonsynonymous and red colour indicates synonymous mutations. **B** The 3D structure of the SARS-CoV-2 spike protein with the top 20 most common mutations marked on their respective residues.

Top 20 most common mutations:

non\_A23403G~D-G, non\_C22995A~T-K, non\_G21987A~G-D, non\_A23063T~N-Y, non\_C23604A~P-H, non\_C21846T~T-I, non\_T22917G~L-R, non\_G22992A~S-N, non\_C23525T~H-Y, syn\_C25000T, non\_G23948T~D-Y, non\_T22679C~S-P, non\_T24469A~N-K, non\_T23599G~N-K, non\_A24424T~Q-H, non\_C23854A~N-K, non\_C22686T~S-F, non\_A23055G~Q-R, non\_T23075C~Y-H, non\_T22882G~N-K

1. **Non vs syn mutations**
2. **Frequency of mutations of most common mutations**

**Evolutionary Insights**

1. **Phylogenetic tree highlighting relationships among clusters**

**Impact of mutations**

1. **Link mutations to known functional effects**

Discussion

**Interpret the results**

1. **Correlation between geographical distribution and specific mutations**
2. **Mutation hotspots? Natural selection + selective pressure**

**Link back to previous studies**

**Implications for public health**

1. **Vaccine design**
2. **Impact of mutations on diagnostics and therapeutic intervention**

**Limitations and future directions**

1. **Limitations of the dataset**
2. **Limitations of techniques used**
3. **Future research**

Conclusion

**Summary of the main findings**

**Importance of studying SARS-CoV-2**

Methods and Materials

**Data sources**

**R packages used**

**Statistical analysis**

**Any additional stuff used**

**Link to GitHub**

Cutler, D. M. and Summers, L. H. (2020). The COVID-19 pandemic and the $16 trillion virus. *JAMA: the journal of the American Medical Association*, 324 (15), pp.1495–1496.

Fehr, A. R. and Perlman, S. (2015). Coronaviruses: an overview of their replication and pathogenesis. *Methods in molecular biology (Clifton, N.J.)*, 1282, pp.1–23.

Hu, B. et al. (2021). Characteristics of SARS-CoV-2 and COVID-19. *Nature reviews. Microbiology*, 19 (3), pp.141–154.

Huang, C. et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*, 395 (10223), pp.497–506.

Major, L. E. et al. *A generation at risk Rebalancing education in the post-pandemic era*. [Online]. Available at: https://www.nuffieldfoundation.org/wp-content/uploads/2022/02/A-generation-at-risk-rebalancing-education-in-the-post-pandemic-era-1.pdf [Accessed 9 February 2025].

Pyrc, K. et al. (2006). Mosaic structure of human coronavirus NL63, one thousand years of evolution. *Journal of molecular biology*, 364 (5), pp.964–973.

*Estimated cumulative excess deaths during COVID-19*. [Online]. Our World in Data. Available at: https://ourworldindata.org/grapher/excess-deaths-cumulative-economist-single-entity?focus=~Confirmed+deaths [Accessed 4 February 2025].