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.

SARS-CoV-2 is the most recent coronavirus (CoVs) 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 (Marik et al., 2021). This where the angiotensin-converting enzyme 2 is expressed (Li et al., 2020). 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 (Arora et al., 2020). 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 (Anon). 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**

Coronaviruses are of the order Nidovirales, which is comprised of several families of related viruses (Fehr and Perlman, 2015). CoVs are enveloped positive-sense single-stranded RNA viruses (Yang and Rao, 2021). Known for having a high mutation and recombination rate, CoVs unlike most RNA viruses have a genetic exonuclease proofreading mechanism (Cui, Li and Shi, 2019). This genetic proofreading mechanism would usually lead to a high fidelity rate and potentially lower mutation rate, however, a high mutation rate still exists. SARS-CoV-2 enters the human cell by binding to several different cellular entry receptors, such as, angiotensin-converting enzyme 2, through its spike protein (Hoffmann et al., 2020).

**Background: Spike glycoprotein**

The spike glycoprotein is one the main structural components SARS-CoV-2. A homotrimer composed of two regions, the S1 and S2 regions, vital for binding and cellular fusion (McCallum et al., 2020). The ACE2 – spike protein interaction plays a vital role in the infectivity of SARS-CoV-2. The interaction is vital as it has been shown that the binding free energy change between the host ACE2 and the spike protein is proportional to the infectivity of SARS-CoV-2 (Wang et al., 2021). The spike protein receptor binding domain (RBD) of the S1 subunit catalyses the attachment directly to the ACE2. More specifically, residues of the receptor binding motif are involved directly in the binding (Chen et al., 2020). This emphasises the biological importance of the spike protein as mutations to residues on the protein can potentially increase or decrease the infectivity of novel strains of SARS-CoV-2.

**Aims of the investigation:**

Analysis was completed on 1984861 individual UK SARS-CoV-2 spike protein sequences stored in GISAID. A conclusive overview of the dataset and the general trends was performed to provide information on the sequencing activity, number of mutations and the emergence of novel mutations. Clustering was performed to classify strains based on sequence similarity, to track the geographical distribution of the different strains and analyse the different viral variants. Further analysis using R was used to identify and characterise the spike protein mutations. The mutations potential effects on viral infectivity and their functionality have also been considered, through elucidating their location in regard to the receptor binding domain.

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

**Methods overview**

Clustering techniques used were k-means and t-distributed stochastic neighbour embedding (t-SNE), to assess the distribution of viral variants. Many R packages were used, including the tidyverse and ggplot, which provided the basis for the majority of the R analysis. A link to a GitHub repository containing the research compendium can be found in the methods section.

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**

The analysis is based on the complete SARS-CoV-2 genome sequences deposited in GISAID, as previously mentioned. The dataset includes sequence information such as a unique identifier, sample date, and country of origin, along with specific mutations for each sequence. The dataset was manipulated to provide an overview and identify general trends in the data. The resulting 1984861 genome sequences from across the UK, provided 15213 unique single mutations.

**Sequencing efforts**

Figure 1A-F illustrates the number of sequences per day from 2020 – 2024. 78% of sequences originated from England, 14% from Scotland and 8% from Wales. Figure 1A provides an overview of the sequencing activity, highlighting peaks associated with the emergence of new SARS-CoV-2 variants and a decline in sequencing as COVID-19 cases decreased. Sequencing activity peaked at the end of 2021, Fig 1C, and the beginning of 2022, Fig 1D, reflecting the rapid spread of new variants and the need to characterise them. It is interesting to note the reduced sequencing efforts after mid-2022 to 2024 as SARS-CoV-2 becomes less prevalent. This reduction may be in relation to reduced testing, herd immunity or changes in public policies. A baseline of testing continues across the UK. Fig 1B, gradual increase in sequencing activity. Fig 1C: rapid increase in the number of sequences per day, coincides with the emergence of new variants, **such x and y**. Fig 1D sharp increase in sequencing activity, peaking at the start of the year, this peak is likely due to the emergence of the omicron variant, followed by a sharp decline in the number of sequences per day. Fig 1E-F, overall sequencing activity is very low, indicating reduced viral spread. By highlighting the sequencing trends, this figure provides a timeline of the spike protein mutations and insights into the emergence of novel SARS-CoV-2 variants in the UK.

A screenshot of a graph

AI-generated content may be incorrect.

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

Use a model to show to the relationship between the number of sequences per day and the sample date

A colorful chart of a graph

AI-generated content may be incorrect.

**Overview of the mutations**

Following this overview of the sequencing activity, the general trends linked around the mutations were also investigated. The violin plot in Figure 2 shows the distribution of the number of mutations per sequence over time (grouped by year, 2020 – 2024). A violin plot helps to illustrate the overall shape of the data and identify the variation that exists within this dataset. The shape of the violin can be interpreted in that the wider sections indicate more observations and the narrow sections represent fewer observations. The median number of mutations per sequence increases from 2020 – 2024, implies that as time progresses the sequences accumulate more mutations. Variance also increases across the years, with 2024 having the highest variance, 25.4. In 2020 and 2021, a number of sequences had high mutation counts, however, by 2024, the distribution had stabilised and the data is less dispersed

The number of mutations per sequence increases over time

Use of linear regression model used, what does show?

Is there a link between the number of mutations per sequence and the sample date?

Use correlation

Variation

Does a later sample date result in a sequence with more mutations?

Want to use statistics, compare theA graph of different colored squares

AI-generated content may be incorrect. values between each year, statistically significant

Figure 3, Number of unique mutations over time

Number of novel mutations per day

15213 unique individual mutations

**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

How many of these mutations are on the receptor binding domain?

How many of these mutations are on the receptor binding motif?

Why is there a syn mutation?

What is a synonymous mutation and what is nonsynonymous mutation?

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:** [**sha524/Spike\_protein**](https://github.com/sha524/Spike_protein)

Arora, P. et al. (2020). Learning from history: Coronavirus outbreaks in the past. *Dermatologic therapy*, 33 (4), p.e13343.

Chen, J. et al. (2020). Mutations strengthened SARS-CoV-2 infectivity. *Journal of molecular biology*, 432 (19), pp.5212–5226.

Cui, J., Li, F. and Shi, Z.-L. (2019). Origin and evolution of pathogenic coronaviruses. *Nature reviews. Microbiology*, 17 (3), pp.181–192. [Accessed 21 February 2025].

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.

Hoffmann, M. et al. (2020). SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell*, 181 (2), pp.271-280.e8.

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.

Li, M.-Y. et al. (2020). Expression of the SARS-CoV-2 cell receptor gene ACE2 in a wide variety of human tissues. *Infectious diseases of poverty*, 9 (1), p.45.

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].

Marik, P. E. et al. (2021). A scoping review of the pathophysiology of COVID-19. *International journal of immunopathology and pharmacology*, 35, p.20587384211048024.

McCallum, M. et al. (2020). Structure-guided covalent stabilization of coronavirus spike glycoprotein trimers in the closed conformation. *Nature structural & molecular biology*, 27 (10), pp.942–949. [Accessed 23 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.

Wang, R. et al. (2021). Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants. *Communications biology*, 4 (1), p.228. [Accessed 27 October 2024].

Yang, H. and Rao, Z. (2021). Structural biology of SARS-CoV-2 and implications for therapeutic development. *Nature reviews. Microbiology*, 19 (11), pp.685–700. [Accessed 21 February 2025].

*COVID-19 cases*. [Online]. datadot. Available at: https://data.who.int/dashboards/covid19/cases?n=c [Accessed 21 February 2025a].

*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 2025b].