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

Abstract

One line introduction:

Methods used:

General trends:

Dimensionality reduction and cluster selection:

Clustering and geographic distribution:

Mutational analysis:

Conclusion:

1. Introduction

**1.1. 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 the 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 due to governments’ inability to report cases and 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). The global presence of SARS-CoV-2 has decreased, as a result of the rollout of vaccines and the build-up of immunity, however, new emerging variants still pose a threat.

Human CoVs, HCoV-229E and HCoV-OC43, have co-existed alongside humans 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 to 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 is where the angiotensin-converting enzyme 2 (ACE2) 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, the 2015 MERS outbreak in South Korea and the COVID-19 pandemic as a result of SARS-CoV-2 (Arora et al., 2020). There appears to be a growing trend in the prevalence of coronavirus 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 viruses so that globally we are better prepared to combat the next epidemic or pandemic.

**1.2. 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 (Amicone et al., 2022). SARS-CoV-2 enters the human cell by binding to several different cellular entry receptors, such as ACE2, through its spike protein (Hoffmann et al., 2020).

**1.3. Spike glycoprotein**

The spike glycoprotein is one of the main structural components of SARS-CoV-2. The homotrimer is composed of two regions, the S1 and S2 regions, that are 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 to the ACE2. More specifically, residues of the receptor binding motif are involved directly in the binding (Chen et al., 2020). Many spike protein mutations that increase viral infectivity have been identified. N501Y is a known spike protein mutation that increases viral infectivity, by enhancing the affinity of the spike protein with host cellular receptors (Liu et al., 2022). D614G increases spike protein flexibility and stability, allowing for enhanced accessibility to the ACE2 cellular receptor (Korber 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.

**1.4. Aims of the investigation**

Previous studies have highlighted individual spike protein mutations and rationalised their effects, instead of looking at the overall mutational trends. Gaps remain in understanding the general distribution of mutations and the clustering patterns in the UK. Analysis was completed on 1984861 individual UK SARS-CoV-2 spike protein sequences stored in GISAID. This study aims to assess the general dataset trends to provide information on the sequencing activity, number of mutations and the emergence of novel mutations. The large dataset required dimensional reduction, 3 methods were used. The suitability of each dimensionality reduction method paired with K-means was assessed and the number of clusters that best represents the data was selected. Clustering and distribution analysis were performed to classify SARS-CoV-2 spike protein mutations based on sequence and track the geographical distribution of the virus strains. Further analysis characterised spike protein mutations, distinguishing between synonymous and non-synonymous mutations, to evaluate their impacts on protein functionality. The potential effects of the most common mutations on viral infectivity have also been considered. Outlining their location within the receptor binding domain (RBD) and receptor binding motif (RBM), to understand the impact on protein functionality following mutations at these vital residues. The study looks to provide insight into the trends of SARS-CoV-2 spike glycoprotein mutations in the UK and to elucidate possible changes in viral infectivity.

**1.** **Provide a complete overview of the general dataset trends**

**2.** **Dimensionality reduction and cluster selection**

**3.** **Investigate the clustering and distribution of SARS-CoV-2 spike protein mutations in the UK**

**4.** **Analyse their potential effects on viral infectivity, link to receptor binding domain/receptor binding motif**

**1.5. Methods overview**

Dataset used was provided by Dr Richard J Bingham, at the University of York, which was obtained from GISAID. Dimensionality reduction methods used: principle component analysis (PCA), t-distributed stochastic embedding, uniform manifold and projection (UMAP). K-means clustering was used, 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. Git, GitHub and Visual Studio Code were used alongside each to create the research compendium. A link to a GitHub repository containing the research compendium can be found in the methods section. It is recommended that you use the compendium for insight into the scripts and analysis used.

**1.6. Summary of the main findings**

**The number of mutations per sequence increases over time**

2. Results

**2.1. 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 1984861 genome sequences from across the UK, provided 15213 unique single mutations.

**2.2. General trends**

**2.2.1. 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. Notably, 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 as omicron. 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.

**2.2.2. Overview of the mutations**

Following this overview of the sequencing activity, the general trends linked around the mutations were also investigated. Figure 2 shows the distribution of mutations per sequence over time (2020 – 2024) using a violin plot. 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. In 2021, the maximum number of mutations for a sequence was 166. In 2022, a sequence had 136 mutations, yet, the maximum values for the other years was close to half of these values. This trend is possibly linked to the sequencing efforts during those years. Fig 1C and Fig 1D depict increased sequencing activity; this increased activity will lead to a higher probability of obtaining extreme values, due to the larger number of samples being taken. A Kruskal-Wallis test, p > 2.2e-16, followed by post-hoc analysis showed that there were significant differences between the years in terms of number of mutations per sequence; this implies that sequences are progressively accumulating mutations over time. A Spearman’s rank correlation value of 0.895 further supports the notion that as time progresses mutations per sequence increase.



15213 unique individual mutations were detected in the dataset, illustrating the ongoing evolution of SARS-CoV-2. The distribution of these novel mutations provides insight into the emergence of new viral strains and the stability of the viral genome. Figure 3, depicts the emergence of novel SARS-CoV-2 spike protein mutations from 2020 – 2024. Panel A of Figure 3, shows during 2021 – 2022, there was an increase in the number of novel mutations, potentially driven by the virus obtaining advantageous mutations. 2022- 2024 shows a gradual decline and stabilisation in the number of novel mutations detected. This decrease could be attributed to vaccines and a general reduction in SARS-CoV-2 genome sequencing. Panel B reinforces this trend, showing the median number of novel mutations per year, peaking in 2021 at 14 mutations per sample date, before decreasing. The overall distribution of the data points closely resembles Figure 1. This similarity is a result of variant emergence and sequencing activity. When a new variant is detected, sequencing activity increases, causing the number of novel mutations detected also to increase. New variants with all these novel advantageous mutations are outcompeting the older variants. Increased genomic sequencing efforts detect these novel mutations, to characterise the newly adapted viral variants.

A screenshot of a graph

AI-generated content may be incorrect.

**2.3. UK SARS-CoV-2 mutation clustering**

**2.3.1. Dimensionality reduction and cluster selection**

A large feature space can lead to time consuming computations, high memory usage and poor clustering performance (Hozumi et al., 2021). This becomes particular problematic given the large dataset containing over a 1 million SARS-CoV-2 spike protein sequences being used.

Before clustering techniques such as K-means or hierarchal clustering can be performed, there is a need for dimensionality reduction to make the data more manageable. Initial K-means clustering was performed on the dataset, see below for information on this dataset, before dimensionality reduction. This was reaching computational times of over 6 hours and due to time constraints with access to equipment, dimensionality reduction was essential. Dimensionality reduction techniques involve converting high multi-dimensional data to a lower dimensional space, while retaining important features of the data (Alkhayrat, Aljnidi and Aljoumaa, 2020). Hozumi et al reported that dimension-reduced k-means clustering methods outperformed the original K-means clustering (Hozumi et al., 2021). This supports the notion that dimensionality reduction on this dataset will improve the accuracy of cluster selection.

To improve the efficiency, I generated a dataset containing the 1000 most common mutations and their associated sequence identifier. I created this dataset to divulge the most accurate number of clusters for the original data. The 1000 most common mutation dataset was still extremely large and required further processing. The objective was to assess which dimensionality reduction method conserves the most important features of the data and which method is better paired with K-means to select the number of clusters. The efficiency of dimensionality reduction method has also been considered. K-means was used as the clustering method to assess if this approach using dimensionality reduction provided a more accurate representation for the number of clusters than without.

I employed a principle component analysis (PCA), a linear dimensional reduction method on the 1000 most common mutations dataset. The main role of PCA is to find a lower dimensional representation of the features, by transforming the original data into a new coordinate system using the principle components (Jolliffe and Cadima, 2016). The elapsed time for the dimensionality reduction was 1 hour 28 mins. Subsequent K-means clustering on the PCA data took an elapsed time of 69.98 seconds; in total 1 hour 29 minutes. K-means paired with PCA demonstrated a fourfold increase in efficiency than just K-means alone. Comparison of within cluster sum of squares (wcss) across different clusters using the elbow method revealed 3, 6, 7 or up to 9 clusters, Figure **B**.

Utilising t-SNE to determine if the number of clusters identified by PCA is the same. t-SNE is a non-linear random dimensionality reduction algorithm that retains the local structure in high dimensional data, while also reflecting the global structure (van der Maaten, 2008). Duplicate values had be removed from the dataset, as the t-SNE failed the first time. This is because t-SNE calculates pairwise distance, if duplicates exist then the pairwise value would be zero, which disrupts the algorithm. This reduced the dataset to 54657elements. The elapsed time for the dimensionality reduction was 4 minutes. Following this, K-means clustering on the t-SNE data took an elapsed time of 3.6 seconds. K-means paired with t-SNE showed over a thousand-fold increase in efficiency than K-means alone. Comparison of wcss across various cluster numbers using the elbow method discovered 3 clusters, Figure **B**.

The final dimensionality reduction technique used was uniform manifold approximation and projection (UMAP). Another non-linear reduction method, UMAP has been reported to be more efficient than t-SNE and is potentially better at keeping more of the global structure than t-SNE (McInnes, Healy and Melville, 2018). The elapsed time for the dimensionality reduction was **X**. K-means clustering on the UMAP dimension reduced data took an elapsed time of **Y**. K-means combined with t-SNE showed a **Z** increase in efficiency than K-means alone. Comparison of wcss across various cluster numbers using the elbow method discovered **A** clusters, Figure **B**.

UMAP has been reported to be the best?

Figure of elbow plots for PCA and t-SNE and UMAP

Figure of different dimensional reduction algorithms

PCA vs T-SNE vs UMAP?, which is better?

Which dimensionality technique provides the clearest clusters?

t-SNE is non-linear method, unlike PCA which is linear method, therefore, t-SNE is able to capture more complex relationships compared to PCA.

**2.3.2. Cluster selection**

3 clusters:

7 clusters:

9 clusters:

Figure 2D visualisation of the SARS-CoV-2 spike protein mutations in the UK with 3, 7 and 9 clusters. **A** 3 clusters. **B** 7 clusters. **C** 9 clusters.

What is interesting?

How many sequences per clusters?

Which cluster is the best?

Possible reasons for multiple clusters?

**2.3.3. Clustering and geographic distribution of spike protein mutations**

t-SNE for visualisation

bar plot for clustering visualisation

What mutations define each cluster?, centroid, which(), how do the clusters relate to known variants, pango

Put each centroid for each cluster into pango, get assignment, can see which cluster relates to which SARS-CoV-2 variant

scatter plot, date vs cluster points

Figure, map of the distributions, Map visualising the distribution in the UK

Use UMAP paper as the outline

Scatter plot clusters over time, will show the evolution of the clusters, what clusters share the same root

**3. 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**

**Impact of mutations**

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

4. Discussion

**4.1. Key findings**

**4.2. General trends**

Link back to previous studies

**4.3. Dimensionality reduction**

Limitations of the techniques used

Which was the best, UMAP has reported to be the best

Link to previous studies on dimensionality reduction methods

**4.4. Clustering**

1. **Correlation between geographical distribution and specific mutations**

**4.5. Mutations**

**4.6. Implications for public health**

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

**4.7. Further limitations and future directions**

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

5. Conclusion

**Summary of the main findings**

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

6. Methods and Materials

**6.1. Data sources**

**Link to GitHub:** [**sha524/Spike\_protein**](about:blank)

**Pango for SARS-CoV-2 lineages**

**6.2. R packages used**

Graphics packages used: ggplot2, hrbrthemes, cowplot, viridis, maps, mapproj, sf, ggthemes, rnaturalearth, rnaturalearthdata, rnaturalearthhires, ggfortify. Data manipulation and wrangling was mainly done using the core tidyverse packages. Statistical packages used included: mgcv for fitting generalised additive models (GAMs) as seen in figure 1, Rmisc and pgrimess for summarising data and descriptive statistics. The devtools and usethis packages were needed to install any packages from GitHub repositories. Dimensionality reduction was done using the packages: factoextra, Rtsne and UMAP. K-means clustering was performed using the base R function.

Need to add sources for each package used

**6.3. Statistical analysis**

**6.4. K-means clustering**

**6.5. Principle component analysis (PCA)**

**6.6. t-distributed stochastic neighbour embedding (t-SNE)**

**6.7. Uniform manifold and projection (UMAP)**

7. Acknowlegements

Creative Lab University of York, datacamp, riffomonas project and other online resources used for helping with R

Alkhayrat, M., Aljnidi, M. and Aljoumaa, K. (2020). A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. *Journal of big data*, 7 (1). [Online]. Available at: doi:10.1186/s40537-020-0286-0.

Amicone, M. et al. (2022). Mutation rate of SARS-CoV-2 and emergence of mutators during experimental evolution. *Evolution, medicine, and public health*, 10 (1), pp.142–155.

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.

Hozumi, Y. et al. (2021). UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets. *Computers in biology and medicine*, 131 (104264), p.104264.

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.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 374 (2065), p.20150202.

Korber, B. et al. (2020). Tracking changes in SARS-CoV-2 Spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell*, 182 (4), pp.812-827.e19.

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.

Liu, Y. et al. (2022). The N501Y spike substitution enhances SARS-CoV-2 infection and transmission. *Nature*, 602 (7896), pp.294–299.

van der Maaten, L. (2008). *Visualizing Data using t-SNE*. [Online]. Available at: https://lvdmaaten.github.io/publications/papers/JMLR\_2008.pdf [Accessed 28 March 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].

McInnes, L., Healy, J. and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]*. *arXiv* [Online]. Available at: http://arxiv.org/abs/1802.03426.