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

Samuel Aldous

Main body: 5,976

Abstract: 239

Abstract

We will build on the work by Hozumi et al., where they developed a method using dimensionality reduction and K-means clustering to analyse large-scale SARS-CoV-2 mutation datasets. This study focuses on synthesising a technique to tackle even larger datasets and extract meaningful information on the mutations and patterns that exist. Specifically, we were interested in which dimensionality reduction technique is best paired with K-means clustering regarding efficiency and visualisation. We explored the clusters divulged, their mutations and how infectivity is altered. 1,984,861 SARS-CoV-2 spike glycoprotein sequences were analysed using R. The temporal and general trends of the dataset were investigated. A binary dataset containing the 1,000 most common mutations was generated and dimensionality reduction, using 3 methods, paired with K-means was performed. The techniques for dimensionality reduction were Principal Component Analysis (PCA), t-distributed stochastic neighbour embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP). The defining mutations of each cluster were investigated, particularly those within the receptor binding domain and receptor binding motif regions, in terms of infectivity. t-SNE paired with K-means provided the clearest, most distinct clusters. Using the elbow method, a cluster number of 3 was selected. Clusters I and II had several of the same defining mutations and likely shared the same root. Cluster III shares similarities with the Omicron variant and was predicted to be the most infectious cluster. Overall, this provides an approach to dealing with large-scale datasets, to characterise clusters and their associated attributes.

1. Introduction

1.1. COVID-19 pandemic

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the strain of coronavirus (CoVs) responsible for the ongoing COVID-19 pandemic (Hu et al., 2021). SARS-CoV-2, initially detected in Wuhan, China, has spread worldwide (Huang et al., 2020). Approximately 770,000,000 people contracted the virus and over 7,000,000 people have died from complications of the virus (WHO, 2024). An estimated value of the number of deaths sits at around 18,000,000 – 32,000,000 (WHO, 2024). The economic burden has been far-reaching, with estimated losses are up to $16 trillion worldwide (Cutler and Summers, 2020). The global presence of SARS-CoV-2 has decreased, driven by vaccines and herd immunity; however, new emerging variants pose a threat.

Human CoVs, e.g. HCoV-229E, have co-existed alongside humans for centuries (Pyrc et al., 2006). These viruses result in milder symptoms similar to 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. 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 caused by SARS-CoV-2 (Arora et al., 2020).

There appears to be a growing trend in the prevalence of CoV outbreaks, even though the rates of SARS-CoV-2 have greatly diminished (WHO, 2024). This stresses the ongoing importance of studying SARS-CoV-2 and other related viruses so that globally we are better prepared to combat the next pandemic.

1.2. SARS-CoV-2 structural and molecular information

CoVs are of the order Nidovirales, comprising several families of related viruses (Fehr and Perlman, 2015). CoVs are enveloped positive-sense single-stranded RNA viruses (Yang and Rao, 2021). CoVs, unlike most RNA viruses, have an exonuclease genetic proofreading mechanism (Cui, Li and Shi, 2019). This genetic proofreading mechanism usually leads 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 cellular entry receptors through its spike protein (Hoffmann et al., 2020).

1.3. Spike glycoprotein

The spike glycoprotein, a homotrimer with 1273 residues, is one of the main structural components of SARS-CoV-2 (Zhang et al., 2021). The receptor binding domain (RBD), composed of S1 and S2 subunits, is vital for binding and cellular fusion with ACE2 (McCallum et al., 2020). This interaction plays a vital role in the infectivity of SARS-CoV-2. The interaction is key 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 RBD of the S1 subunit catalyses the attachment to the ACE2. Specifically, residues of the receptor binding motif (RBM) are involved directly in the binding (Chen et al., 2020). Many spike protein mutations that increase viral infectivity have been identified. N501Y, a known spike protein mutation, 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, enhancing 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 SARS-CoV-2 strains.

1.4. Aims

Previous studies highlighted individual spike protein mutations instead of looking at the overall mutational trends. Gaps remain in understanding the general distribution of mutations and the clustering patterns in the UK. There is insufficient information publicly available about how to analyse large datasets to extract meaningful information. Analysis was completed on 1,984,861 individual UK SARS-CoV-2 spike protein sequences stored in GISAID. This study aims to assess the temporal and general dataset trends to provide information on the sequencing activity, number of mutations and the emergence of novel mutations.

Due to the size of the dataset, dimensionality reduction techniques were utilised. 3 methods were selected for this: Principal Component Analysis (PCA), t-distributed stochastic neighbour embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP). The suitability of each dimensionality reduction method paired with K-means was assessed and the number of clusters that best represent the data was selected. Clustering was then performed to group sequences with similar SARS-CoV-2 spike protein mutations together. Further analysis characterised spike protein mutations, specifically within the RBD and RBM; to explore the impact on protein functionality and virus infectivity 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, how to extract meaningful information from large datasets and to elucidate possible changes in viral infectivity.

1.5. Main findings

Sequencing efforts increased with each major variant wave and as time progressed, sequences accumulated more mutations from 2020-2024. Assessing the three dimensionality techniques, t-SNE paired with K-means yielded the best representation of the clustering in the dataset. The elbow method identified three as the optimal cluster number. Two clusters shared similar sets of mutations, whilst the third cluster had a more distinct mutation set and more mutations residing on the RBM.

2. Methods and Materials

2.1. Data

The original dataset, provided by Dr Richard Bingham from the University of York, had been deposited in GISAID. The UK_seqs_msa_0522_spike_mutations dataset provided the basis for the majority of the analyses. This began as a text file containing the sequence identifiers and the specific mutations associated with each sequence.

UK_seqs_msa_0522_spike_mutations dataset was converted into a table format, consisting of two columns: the sequence information and all the mutations for that sequence in the second column. The dimensions of this dataset were 1984861 rows and 2 columns. This dataset, with 1,984,861 individual sequences and mutations, was then manipulated using R to achieve the analysis required.

For clustering, a second dataset was created, containing the 1000 most common mutations; each mutation was a column and the rows were the individual sequences. The values were either a 1 or 0, corresponding to the presence or absence of a given mutation in that sequence. The dimensions of this dataset were 1,984,612 x 1,001, one column for the sequence identifier and 1,000 mutation columns. The sequence identifier column had to be dropped as it is not a numeric feature. Clustering using all the mutations was too computationally time-consuming and I had limited access to the computers that were the best at this. Restricting the number of mutations made clustering more manageable.

2.2. R packages

A full list of the packages used can be found in the supplementary material, I will outline the main ones used here. Data manipulation and visualisation were performed using the core tidyverse packages along with ggplot2. Dimensionality reduction was performed via the packages: factoextra, Rtsne, umap and uwot. K-means clustering was performed using the base kmeans() function. Sources for the packages used and information on the scripts can be found in the supplementary material.

https://github.com/sha524/Spike_protein

2.3. Nucleotide change to amino acid change

The mutations of the centroids were in a nucleotide form, non_C21618G~T-R, this had to be converted to amino acid notation. The spike glycoprotein gene begins at nucleotide 21563 (Wu et al., 2020). To find the codon position, we calculate the position within the spike glycoprotein gene:

$$nucleotide\ position = \ nucleotide\ of\ interest - \ start\ nucleotide$$

The amino acid position was calculated using:

$$Amino\ acid\ position = \frac{nucleotide\ positon}{3} + 1$$

2.4. Lineage assignment

Potential lineage assignment was performed using a variety of web-based applications for exploring which spike protein mutations aligned with known strains of SARS-CoV-2. The applications used were covSpectrum, coVariants and outbreak.info (Chen et al., 2022; Elbe and Buckland-Merrett, 2017). Pangolin nomenclature was used to designate the names of variants (Rambaut et al., 2020).

2.5. Statistical analysis

The Kruskal-Wallis test assessed, whether the sequence date had a significant effect on the number of mutations per sequence. To further emphasise the relationship, Spearman's rank correlation was used to test for correlation between when a sample was taken and the number of mutations for that sequence.

GAM was fitted using the mgcv package and allowing us to depict the non-linear dynamic relationship of the number sequences per day across 2020 – 2024, Figure 1.

Linear regression model was used to assess the relationship between the median number of mutations per sequence for each year over time. The model chosen as the data appears linear, following an increase in the number of mutations per sequence over time. Then used Kruskal-Wallis test for statistical significance.

2.6. K-means clustering

For information on the R script used, see the research compendium. K-means clustering is an unsupervised learning technique that finds subgroups within a larger dataset (Bradley and Fayyad, 1998). K-means clustering breaks observations into a pre-defined number of clusters, calculated using the elbow method. First each point is randomly assigned to a predefined cluster. The centres of each of these subgroups is then calculated. Each point is then assigned to the nearest newly calculated cluster. Several iterations occur, until each point is assigned to the nearest cluster (Na, Xumin and Yong, 2010). 3 clusters were selected.

The elbow method for cluster determination involved calculating the within-cluster sum of squares (WCSS) for cluster numbers 1-10. nstart parameter, the number of iterations was 10. The WCSS was

then plotted against the number of clusters. The elbow point was determined subjectively as the value where the WCSS stops decreasing significantly. A high WCSS means that the clusters are more spread out and for a low WCSS the clusters are more compact (Cui, 2020). Once the number of clusters had been selected, K-means was performed with number of clusters.

2.7. Principal component analysis (PCA)

PCA is a linear dimensionality reduction method that finds a lower dimensionality representation of features, while maintaining as much variance in the data as possible (Jolliffe and Cadima, 2016). A covariance matrix is used to calculate where the data varies the most (Richardson). The eigenvectors and their eigenvalues are extracted from this matrix. The eigenvectors are the directions of the principal components and the eigenvalues explain the variance of each new principal component (Frost, 2022). The eigenvalues correspond to the importance of that principal component in terms of the original data (Sadrjavadi et al., 2015).

For information on the R script used, see the research compendium. Because the most common 1000 mutation dataset being in a binary format, presence or absence of a mutation, there was no need for scaling or centring the data. PCA uses principal components, these are the new variables that are created by transforming the original data into a new coordinate system. Before K-means clustering was applied, principal components 1 and 2 were retained, PC1 and PC2. A scree plot used to select the number of principal components to retain Figure S1. Very little variance was explained by PC1 and PC2, 0.07, so PC1 to PC9 were selected as more variance was explained, 0.17.

2D visualisation of the dimensionality reduction was performed using ggplot2, Figure 5A. The cluster id's generated by K-means were assigned to each value's principal component. These values are plotted in this new PCA coordinate space, using PC1 and PC2 as the axes. Each point was coloured according to the assigned cluster.

2.8. t-distributed stochastic neighbour embedding (t-SNE)

t-SNE is a non-linear dimensionality reduction technique that reduces the amount of information in the data. The local and global structures are retained by keeping the neighbouring points close together (van der Maaten, 2008).

For each point in the high-dimensional space, the probability of a specific point neighbouring all the other points is calculated using the Euclidean distance metric (Arnoldi, 1951). t-SNE uses a normal distribution in the high-dimensional space (Hinton and Roweis). Points that are closer together have a higher probability than points further away. Every point is treated as the point of interest, the final result is a balancing of all these relationships. The algorithm then randomly initialises the points in a low-dimensional space (Chourasia, Ali and Patterson, 2022). Using a t-distribution, how similar the points are to each other is then calculated (Song et al., 2019). The points are iteratively adjusted until the low-dimensional data resembles the high-dimensional data (Cai and Ma, 2021).

t-SNE was first employed on the 1,000 most common mutation dataset. This failed due to duplicates, which disrupt the pairwise distance metric. The duplicates were removed and the algorithm was successful. The number of datapoints retained after the duplicates were removed was 54,657.

2D visualisation of the dimensionality reduction was performed using ggplot2, Figure 5B. The cluster id's generated by K-means were assigned to each value's new coordinates. The points were plotted according to their position in the new two dimensions. Each point was coloured according to the assigned cluster.

2.9. Uniform Manifold Approximation and Projection (UMAP)

UMAP is a non-linear graph-based dimensionality reduction technique designed to preserve the local and global structure of the data (McInnes, Healy and Melville, 2018). First, a graph in high-dimensional space is made and the k-nearest neighbours method is used to work out the probability of points neighbouring each other. A graph in low-dimensional space is constructed, where the points are randomly assigned positions. The positions of the points in the low-dimensional graph are changed to decrease the differences between the high-dimensional graph and the low-dimensional graph (McInnes, Healy and Melville, 2018).

UMAP was performed on the 1,000 most common mutation dataset. A sparse matrix, from the Matrix package, was employed as UMAP could not process the whole dataset. This sparse matrix was then converted back to a tibble, as UMAP cannot handle sparse data matrices. The sparse matrix method did not work; therefore, a random 500,000 sample was taken, this appeared to be computational limit.

Visualisation of the dimensionality reduction was performed using ggplot2, Figure 5C. The cluster id's generated by K-means were assigned to each values' new low-dimensional coordinates. The points were then plotted according to their position in the new two dimensions. Each point was coloured according to the assigned cluster.

3. Results

3.1. Introduction

The analysis is based on the complete SARS-CoV-2 genome sequences deposited in GISAID. The dataset includes sequence information such as a unique identifier, sample date, country of origin and specific mutations for each sequence. The dataset was manipulated to provide an overview and identify general trends in the data. 15213 unique single mutations in the spike protein were identified.

3.2. Temporal and general trends

3.2.1. Sequencing efforts

Sequencing activity varies across the UK from 2020 to 2024, Figure 1. 78% of sequences originated from England, 14% from Scotland and 8% from Wales. An overview of the sequencing activity, Figure 1A, highlights 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, sequencing efforts reduced after mid-2022 to 2024, as SARS-CoV-2 becomes less prevalent. This reduction may reflect reduced testing or changes in public policies. A baseline of testing continues across the UK. A gradual increase in sequencing activity persists in 2020, Figure 1B. A rapid increase in the number of sequences per day, Figure 1C, coincides with the emergence of new variants, such as Omicron. A sharp increase in sequencing activity, peaking at the start of the 2022, coincides with the emergence of the Omicron variant, followed by a rapid decline in the number of sequences per day, Figure 1D. Overall sequencing activity becomes very low, Figure 1E-F, 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.

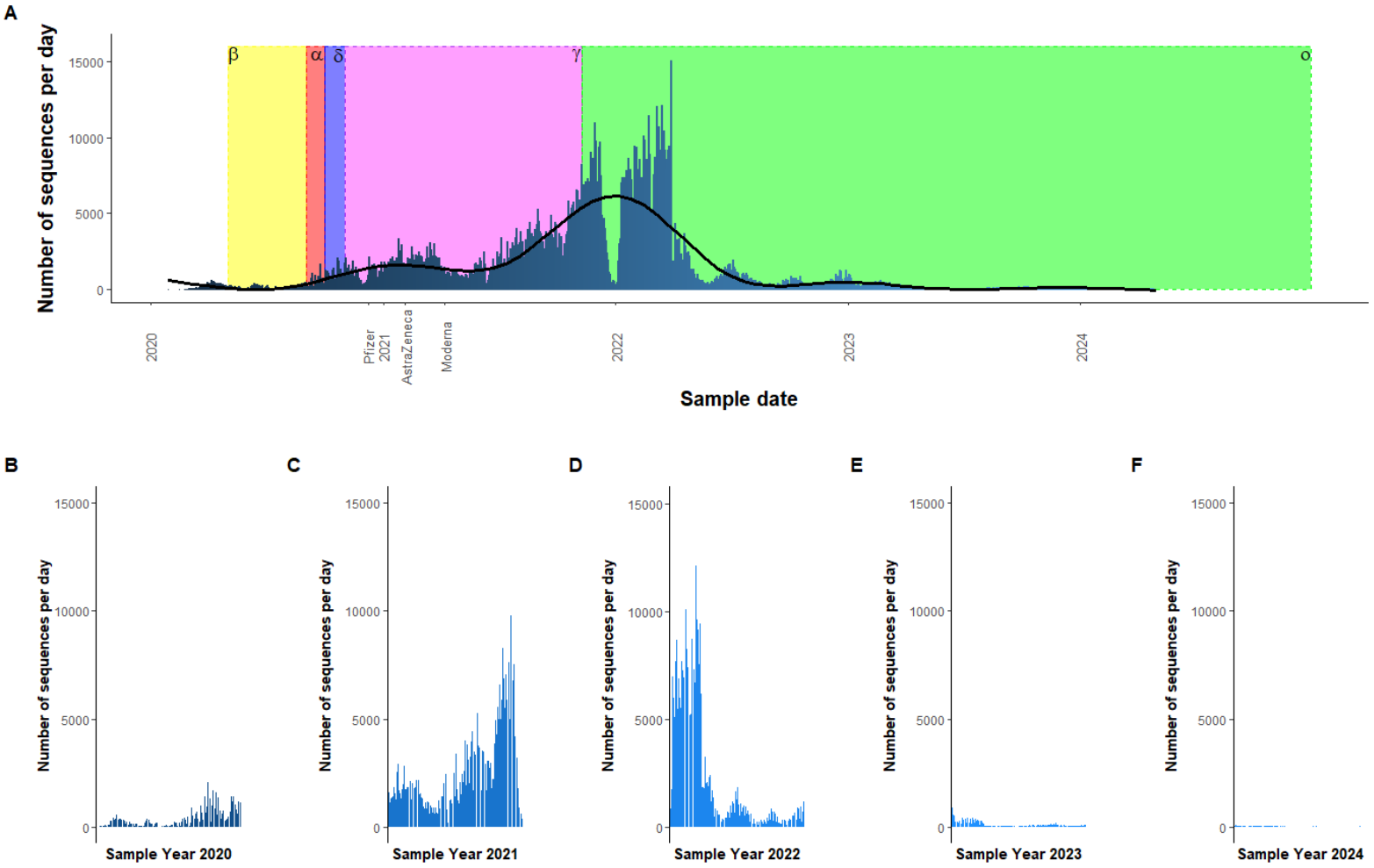


Figure 1. Sequencing activity for SARS-CoV-2 from 2020 – 2024. **A** The overall the number of sequences per day from years 2020 – 2024. The coloured zones represent the emergence of a new strain of SARS-CoV-2. Yellow – Beta (WHO, 2021), red – Alpha (Rambaut et al., 2020), purple – Delta (WHO, 2021), pink – Gamma (WHO, 2021), green – Omicron (Callaway, 2021). Also, on the plot are specific dates showing the first use of that vaccine. General additive model has been used to show the relationship between sample date and the number of sequences per day. **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.

3.2.2. Overview of the mutation

The general trends around the mutations were also investigated. The distribution of mutations per sequence over time, 2020 – 2024, was shown using a violin plot, Figure 2. The median number of mutations per sequence increases from 2020 – 2024, implies that as time progresses the sequences accumulate mutations. The maximum number of mutations for a sequence in 2021 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. Increased sequencing activity, Figures 1C-D, leads to a higher probability of obtaining extreme values, as more samples are taken. A Kruskal-Wallis test, $p > 2.2e\text{-}16$, followed by post-hoc analysis showed that there were significant differences between the years in terms of number of mutations per sequence; implying 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. This pattern correlates with Muller's ratchet, that if there is no recombination occurring, deleterious mutations will accumulate (Metzger and Eule, 2013).

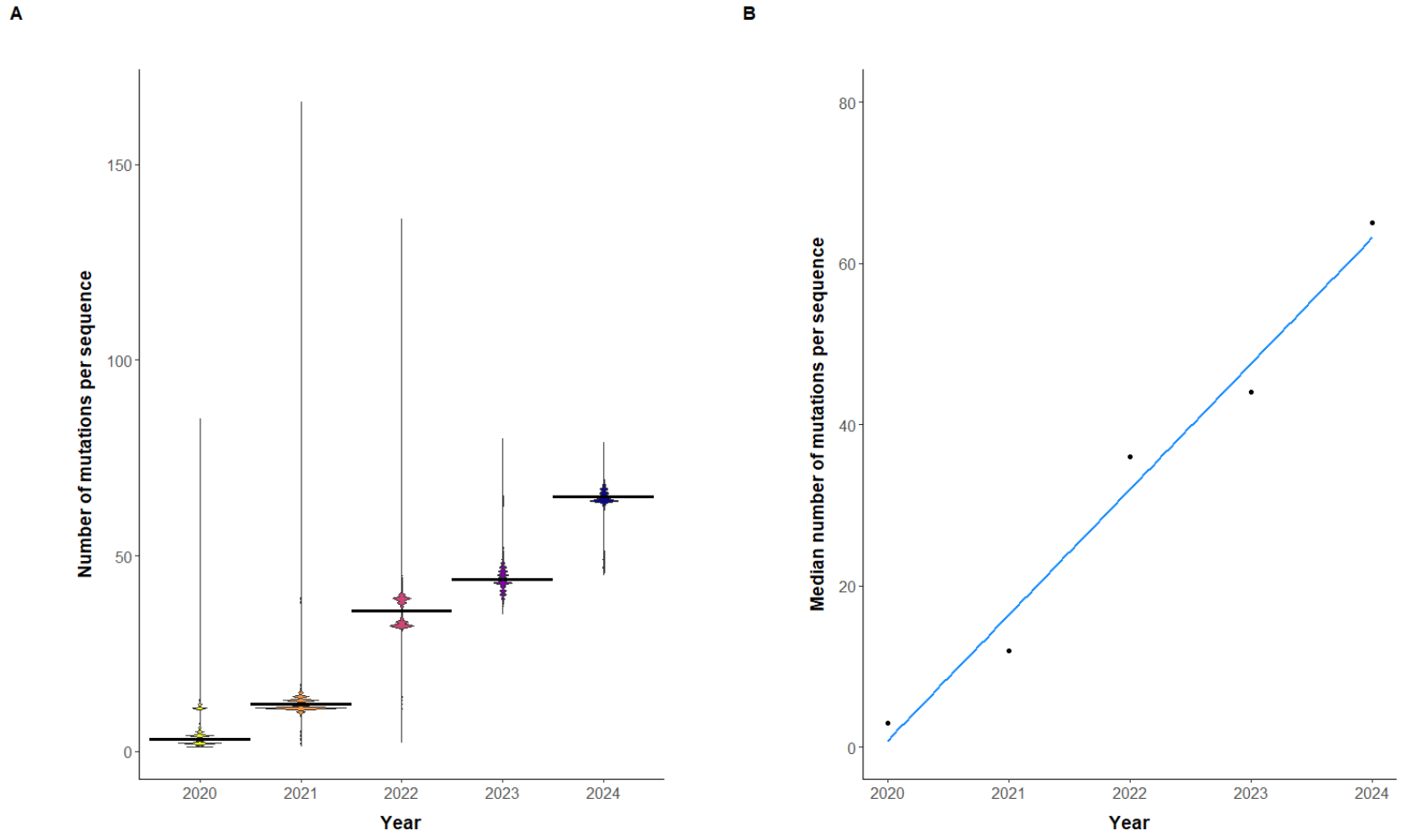


Figure 2. A Violin plot. Sequences accumulate mutations as time progresses. The sequences have been grouped by year. The thicker black line across each distribution represents the median number of mutations per sequence for that year. Median values are as follows: 2020 - 3, 2021 - 12, 2022 - 36, 2023 - 44, 2024 - 65. Variance levels: 2020 - 10.7, 2021 - 4.61, 2022 - 12.0, 2023 -16.7, 2024 - 25.4. There was a significant effect of the year the sequence was taken and the number of mutations per sequence (Kruskal-Wallis: $\chi 2 = 1558575$, d.f. = 4, $p > 2.2e\text{-}16$). Post-hoc comparison showed that there were significant differences between all the years. A significance level of 0.05 was used**. B** Scatter plot showing the median number of mutations per sequence over time (2020–2024). A strong positive correlation was observed between year and the median number of mutations per sequence (Spearman's $\rho = 0.895$). The linear regression line suggests a consistent increase in median mutations over time.

15,213 unique mutations were detected. The distribution of these newly discovered mutations provides insight into the emergence of new viral strains and the stability of the viral genome. The emergence of novel SARS-CoV-2 spike protein mutations from 2020 – 2024, illustrates the ongoing evolution of SARS-CoV-2, Figure 3. During 2021 – 2022, Figure 3A, 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 reduction in SARS-CoV-2 genome sequencing. The median number of novel mutations per year peaks in 2021 at 14 mutations, before decreasing, Figure 3B. The overall distribution of the data closely resembles Figure 1. When a new variant is detected, sequencing activity increases, causing the number of novel mutations detected to also 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.

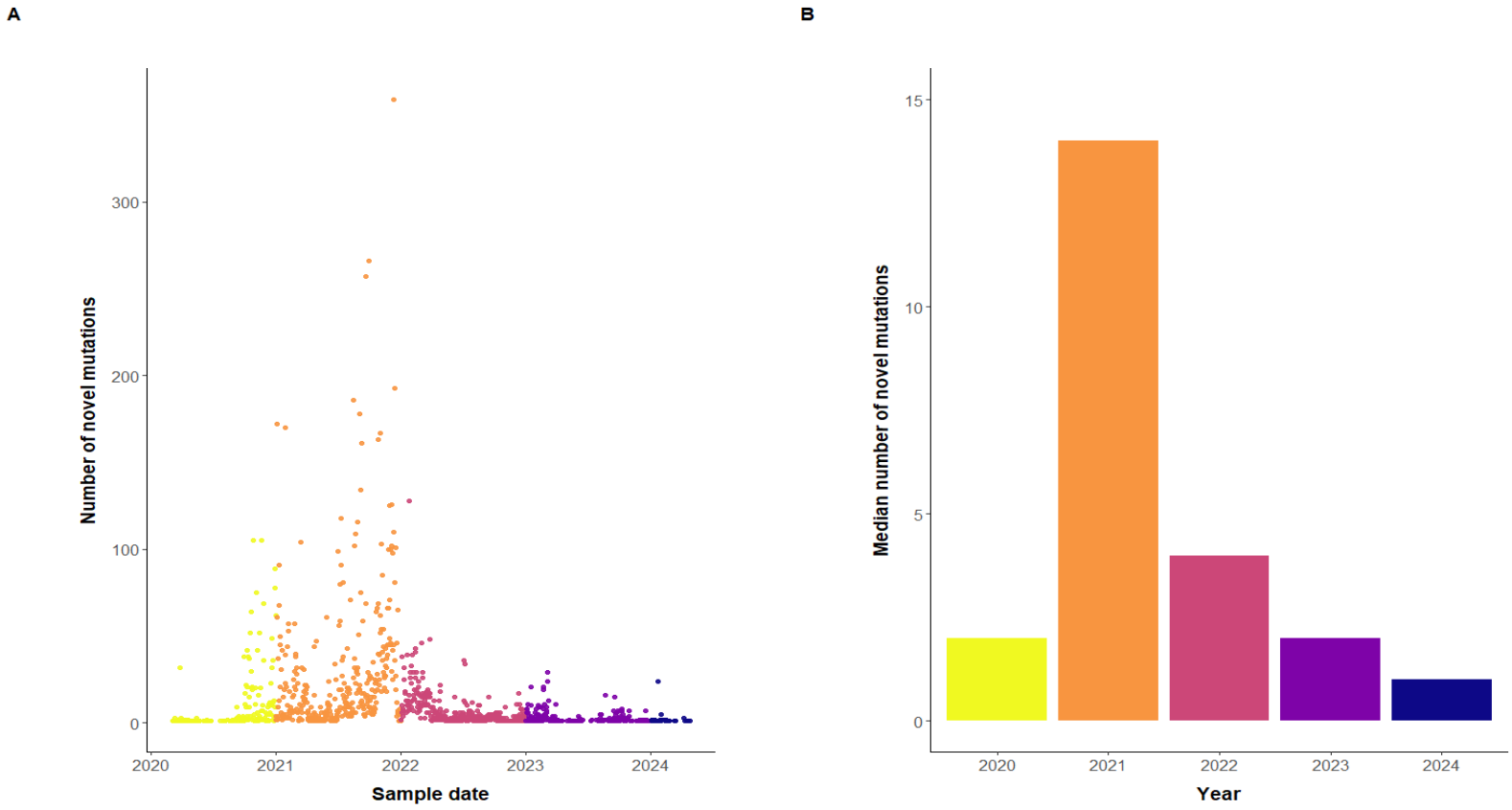


Figure 3. **A** The emergence of novel SARS-CoV-2 spike protein mutations from 2020 - 2024. Figure depicts the number of novel mutations for any given sample date. Mean values for each year: 2020 - 10.8, 2021 - 29.0, 2022 - 7.95, 2023 - 3.47, 2024 - 2.25. **B** The median number of novel mutations from 2020 - 2024. Median values for each year: 2020 - 2, 2021 - 14, 2022 - 4, 2023 - 2, 2024 - 1.

3.3. UK SARS-CoV-2 mutation clustering

3.3.1. Dimensionality reduction and cluster selection

A large feature space leads to time-consuming computations, high memory usage and poor clustering performance (Hozumi et al., 2021). This becomes particular problematic using a dataset containing over 1,000,000 SARS-CoV-2 spike protein sequences. Building on Hozumi et al work, I implemented the same 3 dimensionality reduction algorithms, PCA, t-SNE and UMAP on the UK SARS-CoV-2 spike protein data.

Before K-means can be performed, dimensionality reduction needs to make the data more manageable. Initial K-means clustering was performed on the dataset, before dimensionality reduction. This was reaching computational times of over 6 hours and given time constraints with access to equipment, dimensionality reduction was essential. Dimensionality reduction techniques involve converting high-dimensional data to a lower-dimensional space, while retaining important data features (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 suggests that dimensionality reduction on this dataset will improve the accuracy of cluster selection and visualisation.

I generated a binary dataset containing the 1000 most common mutations and their associated sequence identifier. 1s corresponded to the presence of a mutation and 0s to the absence of that mutation. I created this dataset to divulge the most accurate number of clusters for the original data. The 1,000 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 data features for visualisation and is better paired with K-means to select the number of clusters. The efficiency of the dimensionality reduction method has also been considered.

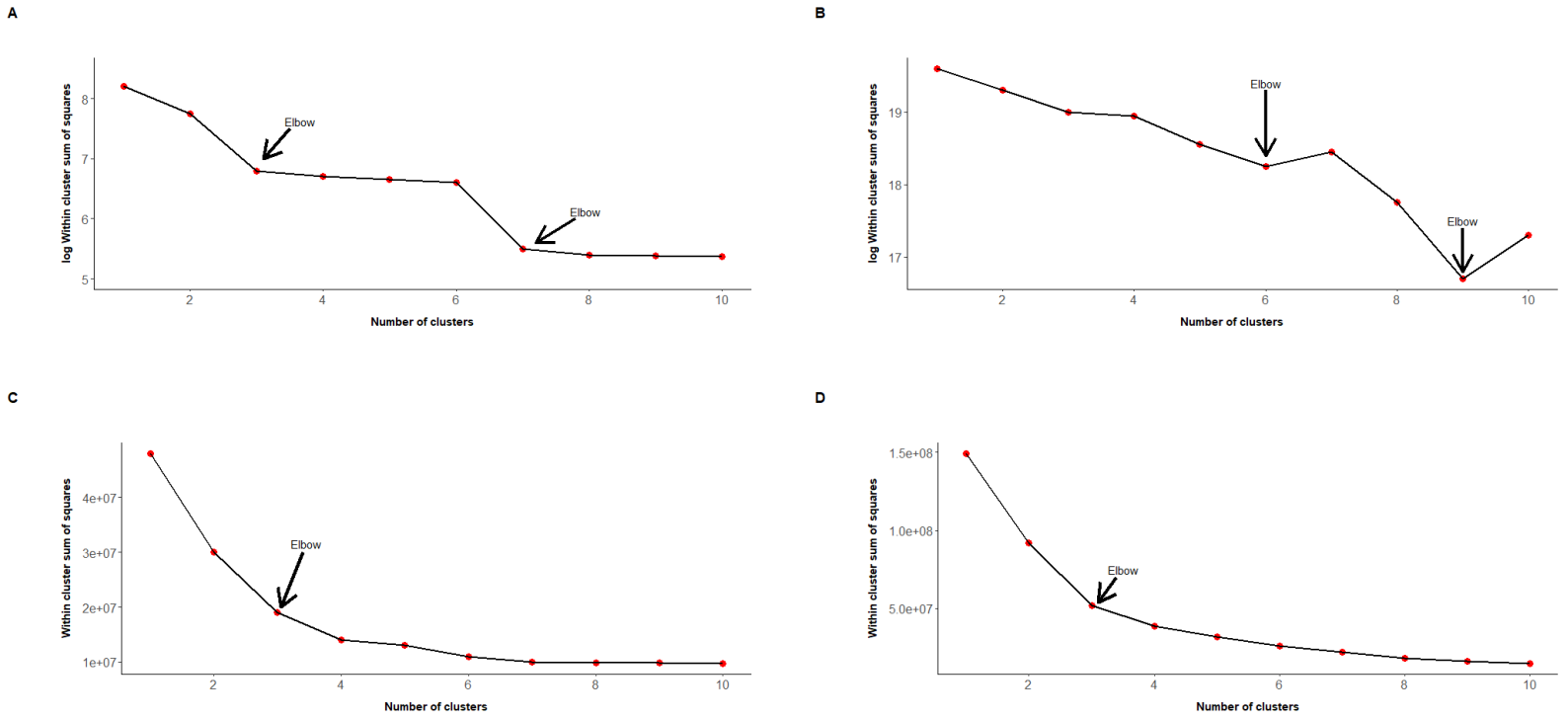


Figure 4. Elbow plots. The optimal number of clusters is shown as the turning point in the in the elbow plots. Within cluster sum of squares of the PCA assisted K-means clustering. **A** PC1 and PC2, then K-means for 1000 most common mutations on a log10 scale. **B** PC1:PC9, followed by K-means for the 1000 most common mutations on a log10 scale. **C** t-SNE elbow plot. **D** UMAP elbow plot. From the elbow method, the optimal number of clusters ranges from 3 up to 9.

PCA was applied to the 1,000 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 principal components (Jolliffe and Cadima, 2016). The elapsed time for the dimensionality reduction was 1 hour 28 minutes. 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. A scree plot was used to select the number of principal components to retain, Figure S1. Principal components 1 to 9 were selected for the elbow plot. The cumulative proportion of variance explained by PC1 to PC9 was 0.17. Principal components 1 to 9 were selected because, as subsequent components add very little variance. Comparison of WCSS across different clusters using the elbow method revealed 3, 6, 7 or up to 9 clusters, Figure 4A-B.

The second dimensionality reduction algorithm employed was t-SNE. t-SNE retains the local structure in high-dimensional data, while also reflecting the global structure; particularly useful for complex biological data (van der Maaten, 2008). t-SNE failed the first time, so duplicate values had to be removed from the dataset. t-SNE calculates pairwise distance, if duplicates exist then the pairwise value would be zero, disrupting the algorithm. This reduced the dataset to 54,657 elements. The elapsed time for the dimensionality reduction was 4 minutes. 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 4C.

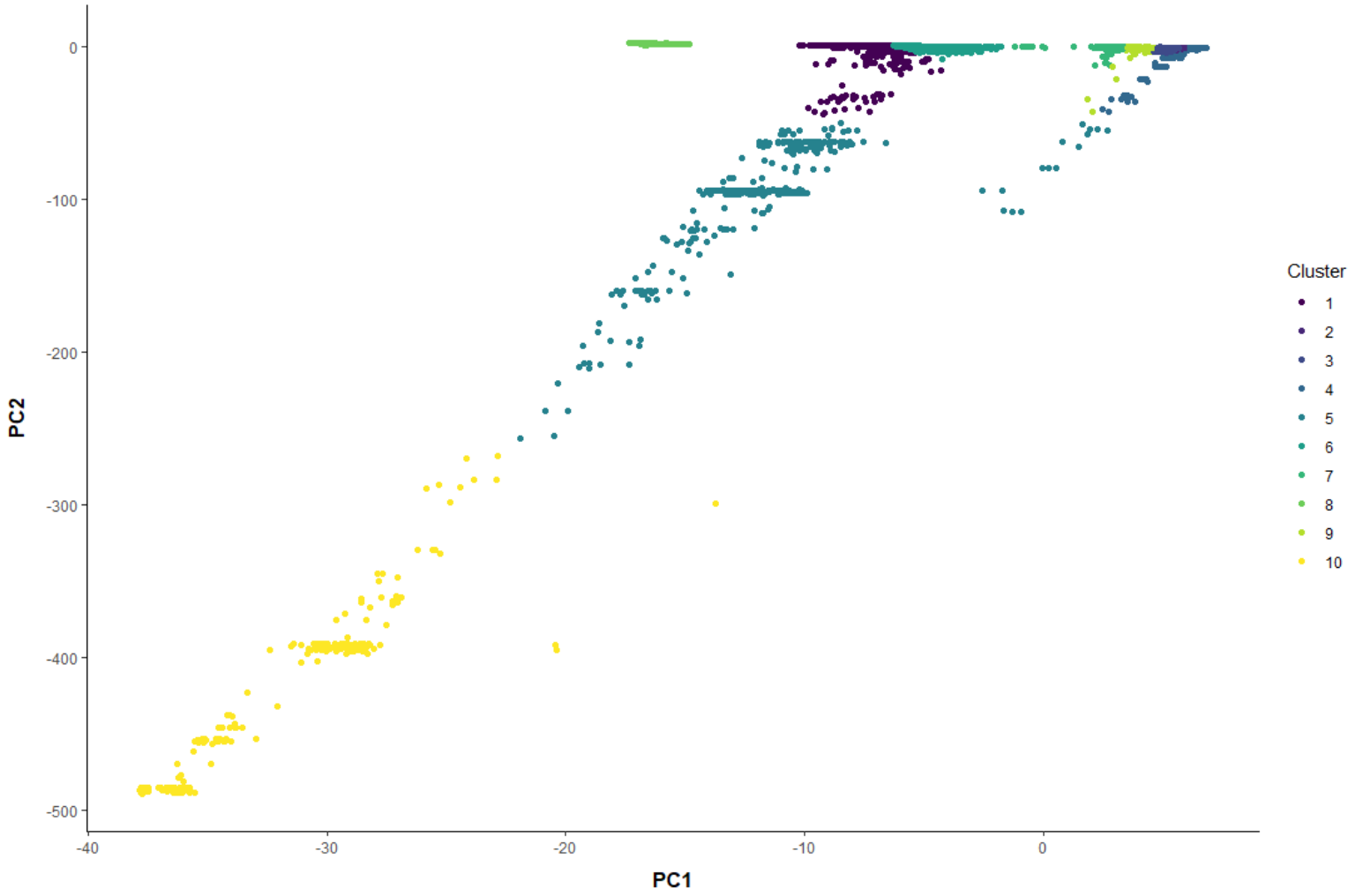
The final dimensionality reduction technique used was UMAP, a non-linear reduction method. UMAP has been reported to be more efficient than t-SNE and is potentially better at keeping more of the global structure than t-SNE (McInnes, Healy and Melville, 2018). UMAP was struggling with the size of the dataset. The dataset was converted to a sparse matrix, as it is more memory efficient for storage as only where the mutation appears is retained (Bates, Maechler and Jagan, 2000). This was also unsuccessful. The final option employed was to take a sample from the dataset. A 500,000 random sample was taken and UMAP was performed on this. By taking a sample, important features and relationships may be lost. The elapsed time for the dimensionality reduction was 37 minutes. K-means clustering on the UMAP dimension reduced data took an elapsed time of 36 seconds; in total 37 minutes and 37 seconds. K-means combined with UMAP showed around a ninefold increase in efficiency than K-means alone. Comparison of WCSS across various cluster numbers using the elbow method discovered 3 clusters, Figure 4D.

The visualisations illustrate the dimensionality reduction techniques on the 1000 mutation SARS-CoV-2 data, Figure 5. All datasets were reduced to 2 dimensions to allow for better visualisation. PCA performs poorly, Figure 5A, with poor separation and erratic clustering. t-SNE provides much clearer and defined clustering than PCA, Figure 5B. UMAP gives more distinct clustering compared to PCA, however, there is a large amount of overlap between the clusters, Figure 5C. This makes it hard to distinguish the individual cluster groups.

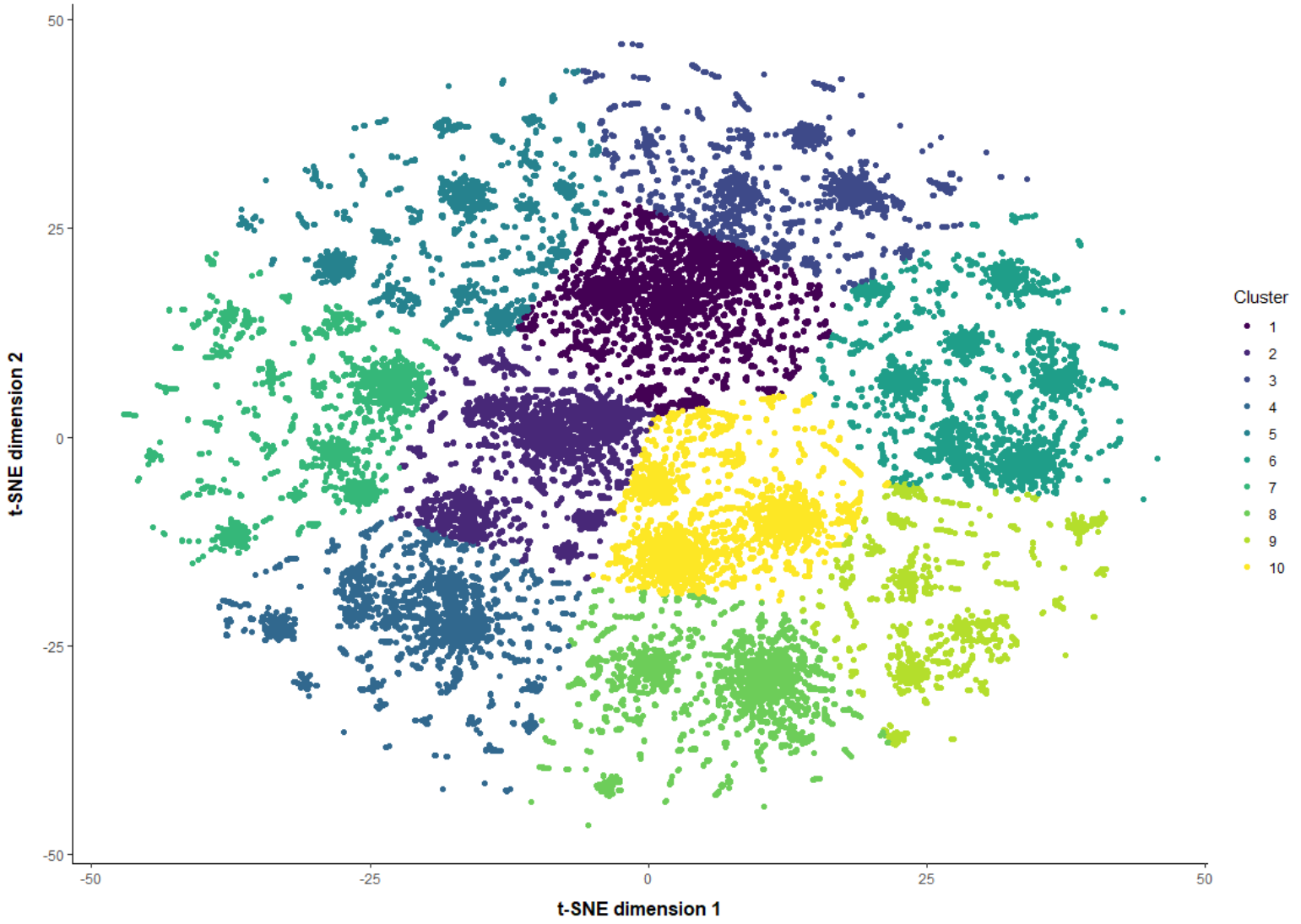
Large data sets are more computationally intensive and time-consuming; therefore, the efficiency of the dimensionality reduction method should be considered when selecting a method. t-SNE was up to 22 times faster than PCA and 9 times faster than UMAP. t-SNE demonstrated the most efficient overall performance, with UMAP having the second fastest and PCA performing the worst in terms of computational speed.

With this considered, t-SNE paired with K-means was selected. t-SNE dimensionality technique provides the clearest clusters and was the most efficient, however, t-SNE also retain the fewest original features. Time constraints and access to equipment were not too much of an issue, therefore, the best cluster representation was more important than the efficiency of the dimensionality reduction technique. 3 clusters were selected for further analysis.

A



B


50
25
0
-25
-50
t-SNE dimension 2
-50
-25
0
25
50
t-SNE dimension 1
Cluster
1
2
3
4
5
6
7
8
9
10

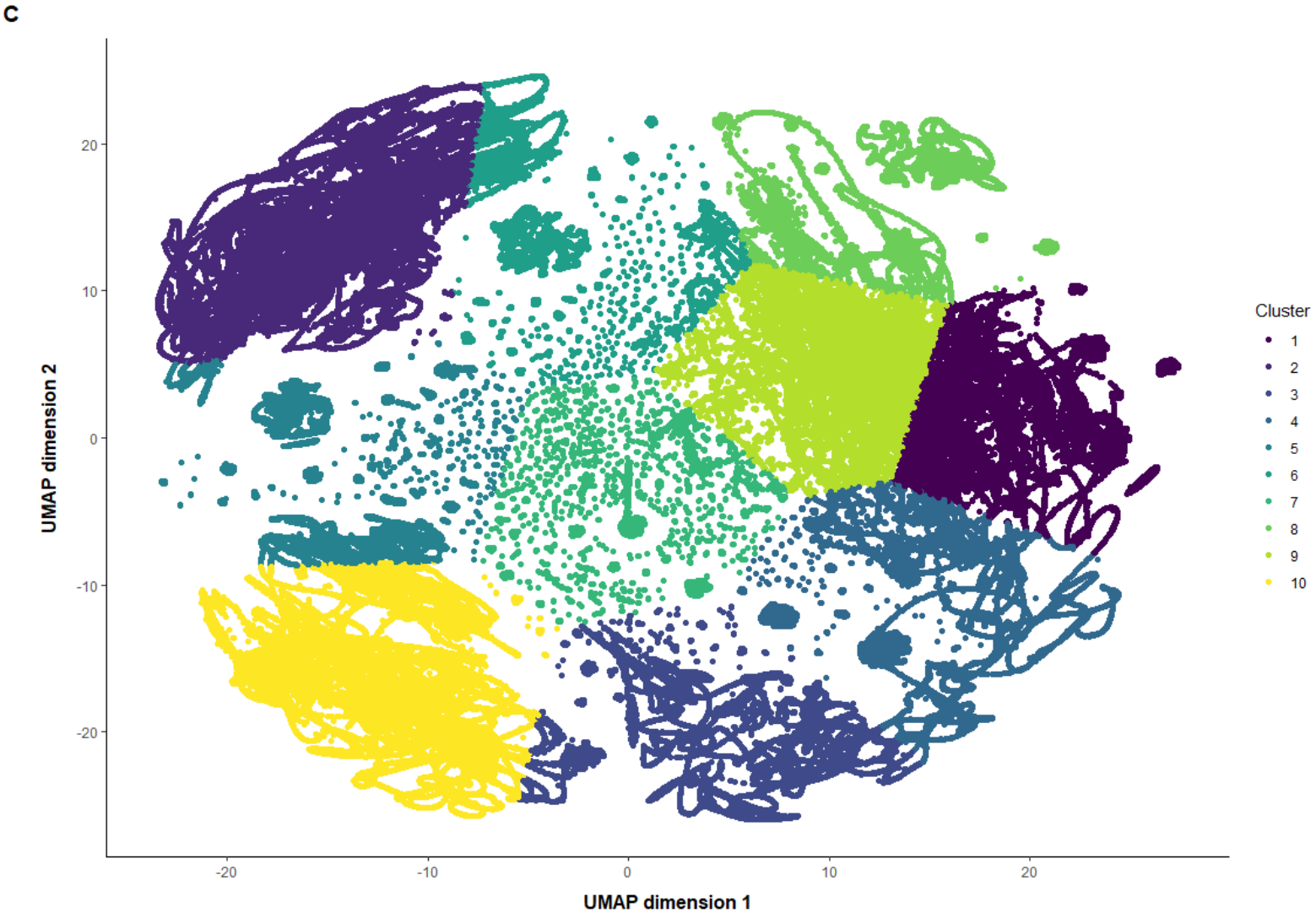


Figure 5. 2D visualisations illustrating the differences of the dimensionality reduction techniques used for the UK SARS-CoV-2 spike glycoprotein dataset, showing 1-10 distinct clusters. The 1-10 clusters were used to work out the actual number of clusters later on. **A** PCA visualisation. **B** t-SNE visualisation. **C** UMAP visualisation.

3.3.2. Clustering of spike protein mutations

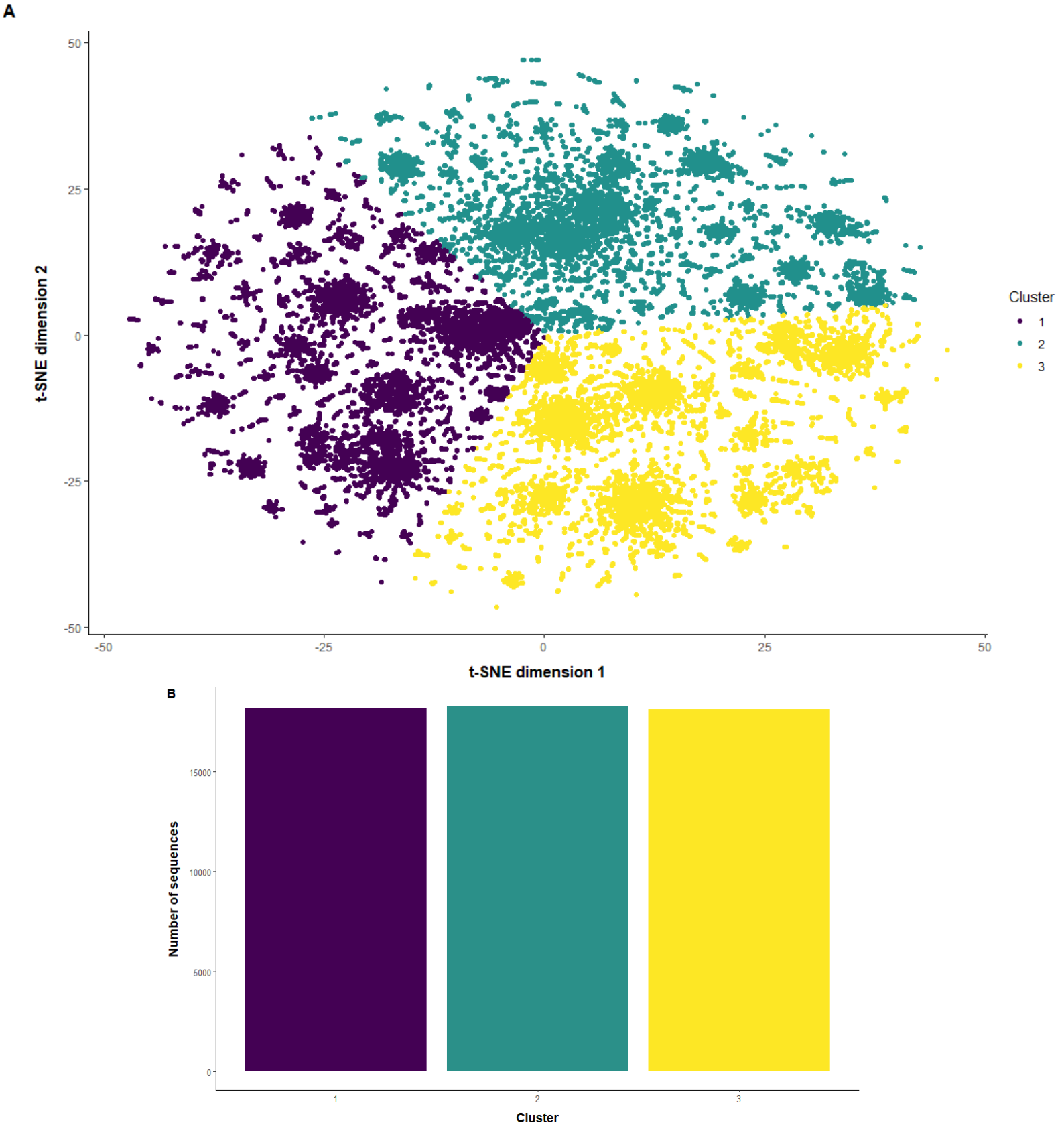


Figure 6. **A** Figure 2D visualisation of the SARS-CoV-2 spike protein mutations in the UK with 3 clusters using t-SNE dimensionality reduction method paired with K-means. **B** Box plot comparing the number sequences per centroid

t-SNE paired with K-means divulged 3 clusters. This was performed to investigate the clustering patterns of SARS-CoV-2 spike protein mutations in the UK. K-means clustering groups sequences with similar mutational patterns together, allowing us to visualise the distribution and diversity of mutations, Figure 6A. All the sequences in a cluster share a similar mutational pattern, with the centroid of each cluster representing the defining mutations for that specific cluster. There are distinct groups, with little to no overlap between the clusters, Figure 6A.

3 clusters that were identified using the elbow method and through assessing which dimensionality method provided the clearest clustering visualisation, Figure 4. This cluster number shows 3 distinct and well-separated groups. The clusters are similar in size, ranging from 18300 sequences in Cluster I to 18144 sequences in Cluster III, Figure 6B. The box plot of the number of sequences per cluster, illustrates the distribution of sequences across the clusters, Figure 6B. Cluster I had the most sequences, 18300, Cluster III had the fewest sequences, 18144. Cluster II had 18300 sequences. This likely suggests that Cluster I, Cluster II and Cluster III are potentially the dominant variants of SARS-CoV-2 in the UK.

The clusters are each defined by a specific set of mutations. The centroid of each cluster was used to explore these defining mutations. The mutations were initially in nucleotide notation, for example, non_C21618G~T-R, this was converted to amino acid form to provide a greater insight into the amino changes occurring. Possible co-mutations occurring each of the clusters were also identified. Co-mutations are mutations that occur simultaneously or near each other, they can provide an enhanced effect on protein structure and function (Phillips, 2008).

Cluster I has the fewest number of mutations, followed by Cluster II, with Cluster III having the most mutations associated with its centroid. Cluster I was defined by 10 mutations, 6 single nucleotide mutations and 3 co-mutations, Table 1. Cluster II is determined by 15 mutations, 12 single nucleotide mutations and 3 co-mutations, Table 1. Clusters I and II share a co-mutation involving E156G, F157EN and R158E. E156G is caused by a double-nucleotide substitution, whilst F157N and R158E arise from triple-nucleotide substitutions. Cluster III is primarily distinguished by 33 mutations, with a co-mutation at P25H and P26H. P25H is caused by a double-nucleotide substitution, whereas P26H results from a single-nucleotide substitution, Table 1. Cluster III has another potential co-mutation occurring at the residues: S371F, S373P, S375F and T376A. Cluster III is the only cluster that accumulated a synonymous mutation at residue 1146, where a cytosine is substituted for a thymine.

Two mutations were shared across all the clusters, G142D and D614G. Clusters I and II share several mutations: T19R, L452R, P681R, E156G, F157E and R158E. Cluster II and Cluster III share the mutation T478K. The sharing of mutations provides evidence that clusters may have a shared root. The presence of shared roots, conforms to the idea of divergent evolution of SARS-CoV-2 variants (Futuyma, 2013).

Cluster	Single nucleotide substitution	Double nucleotide substitution	Triple nucleotide substitution
Cluster I	T19R, G142D, **L452R**, T465K, **D614G**, P681R	E156G	F157E, R158E
Cluster II	T19R, T95I, **G142D**, L452R, T478K, **D614G**, P681R, D950N, Y145H, A222V, V1264L, S12T	E156G	F157E, R158E
Cluster III	**G142D**, T478K, **D614G**, N440K, N501Y, P680H, T19I, L24S, V213G, G339F, S371F, S373P, S375F, T376A, D405N, R408S, K417N, S477N, E484A, Q493R, Q498R, Y505H, H655Y, N679K, N764K, D796Y, Q954H, N969K, P26H, P1263	P25H	A27H

Table 1. The mutations associated with each cluster, categorised by their respective nucleotide substitution. Mutations shared across all the clusters are in bold. Coloured green are the mutations shared by Cluster I and Cluster II. In red are the mutations that overlap between Cluster II and Cluster III.

The mutations that are the drivers of separation between clusters were also identified. These mutations were unique to certain clusters and differentiated the clusters from one another. Cluster I is defined by the T465K mutation. Cluster II is distinguished by 6 unique mutations T95I, D950N, T145N, A222V, V1264L and S12T. Cluster III is characterised by 29 unique mutations, such as N501Y, Y505H and S477N. Identifying these mutations provides insight into specific characteristics that define a cluster and how each cluster could link to a known lineage of SARS-CoV-2.

4. Mutation analysis

For each cluster, I determined how many of its mutations are located within the RBD and RBM regions. The RBD and RBM are vital for spike protein binding to the host ACE2 receptor. Examining mutations in these regions provides insight into how mutations increase or decrease viral infectivity. The RBD is located between residues 319-541, with the RBM positioned at 437-508 within the RBD, Figure 7 (D'Ippolito et al., 2021). In Clusters I and II 2/10 centroid mutations, L452R and T478K, were located within the RBM of the RBD. Cluster III exhibited a larger proportion of mutations within the RBD, 16/33, with approximately half positioned within the RBM region.

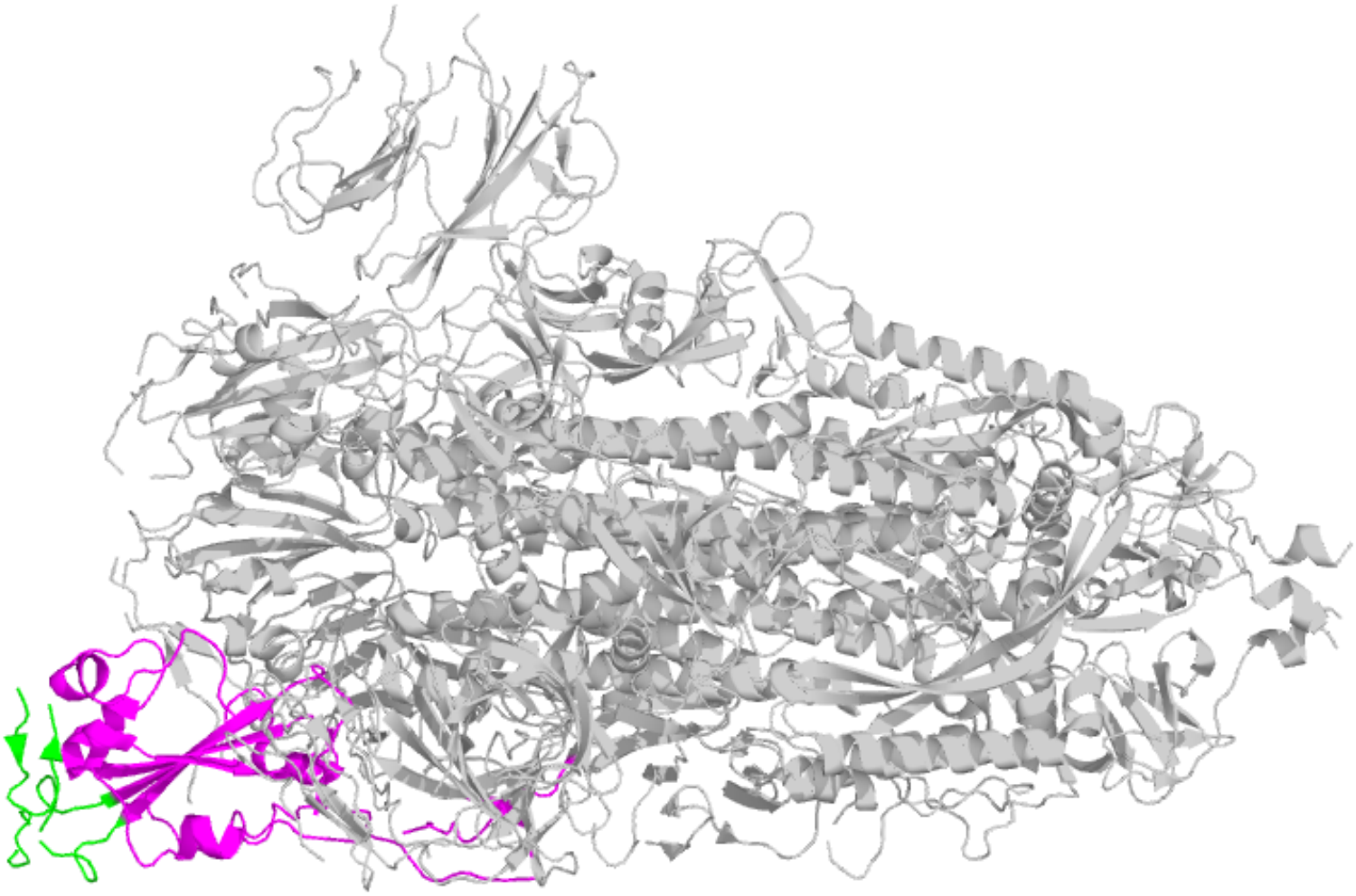


Figure 7. 3D visualisation of spike glycoprotein using PyMol. RBD residues have been coloured purple, RBM residues have highlighted in green and the spike protein core has been marked grey (Schrödinger, LLC, 2015).

5. Discussion

5.1. Key findings

Temporal and general trends:

- Increased sequencing activity aligns with the emergence of new variants
- Sequences accumulated mutations as time progressed from 2020-2024

Dimensionality reduction:

- Out of the three dimensionality reduction techniques assessed, t-SNE paired with K-means provided the clearest representation of the clusters
- The elbow method determined the most accurate number of clusters as 3

Cluster relationships:

- Clusters I and II share similar mutations, most likely share the same root
- Cluster III is the most genetically distinct cluster, with a higher proportion of mutations on the RBM

5.2. Temporal and general trends

Daily sequences peaked during the Gamma and Omicron variant waves, with a gradual increase for the other variant waves; driven by the need for identification of variants and their characteristics. After 2022, sequencing decreased dramatically, possibly reflecting increased vaccine coverage, as by July 2021 most individuals in the UK had received their 2nd (Mathieu et al., 2020). A decrease in viral decline and a lack of interest in sampling also play a role.

Sequences accumulated more mutations as time progressed, with the median rising from 3 in 2020 to 65 in 2024. This is supported by the estimation that SARS-CoV-2 has a mutation rate of 1x10-3 substitutions per base, reflecting a time-dependent increase in mutations per sequence (Abbasian et al., 2023). Selective pressure is a key factor. Mutations that enhance transmissibility and provide a survival advantage are more likely to be fixed in a population (Harvey et al., 2021).

A similar overall trend can be seen in Figure 1 and Figure 2, where increased sequencing during variant waves lead to the detection of more novel mutations. This introduces sampling bias; these mutations may have already been established and are only detected when sampling intensified. New mutations are still occurring after 2022, but are no longer being detected. Sampling bias may also exist in the uneven proportion of samples from the main regions of the UK. This could disrupt regional trends in terms of cluster presence in those areas.

5.3. Dimensionality reduction

PCA assumes the data is linear, which is untrue for this data; important non-linear structures may be lost as a result (Jolliffe and Cadima, 2016). PCA fails to retain the local structure of the data, losing relationships between points. Many principal components are generated, which makes interpretation hard and visualisation poor. These findings are consistent with Hozumi et al., who also reported PCA performing poorly on large datasets (Hozumi et al., 2021).

t-SNE and UMAP are non-linear methods that retain both the local and global structures. These techniques are considered more appropriate for large datasets as the data is non-linear (Linderman et al., 2019; McInnes, Healy and Melville, 2018). PCA has been reported to be computationally faster, however, the results here suggest otherwise: PCA fourfold increase, t-SNE thousandfold increase and UMAP ninefold increase (Hozumi et al., 2021). This may result from t-SNE reducing the feature size more and UMAP having to operate on a sample due to memory limitations.

UMAP has been reported to be the best method for reducing large datasets and for visualisation of the clustering (Hozumi et al., 2021). The results here do not support this, with less distinct clusters for UMAP, Figure 5C, whilst t-SNE shows well-separated clusters, Figure 5B. UMAP required a random sample, which introduces bias, as the sample may not reflect the overall structure of the data. t-SNE paired with K-means provides a faster, more dynamic approach, with improved cluster clarity.

The use of a Euclidean distance metric was a major limitation, as it values the presence and absence of a mutation equally. The presence of a mutation is more important, as shared absences are common (Chung et al., 2019). A Jaccard distance metric solves this problem as the algorithm is only interested in where mutations are present. To improve clustering visualisation, all three methods should be rerun using this metric.

5.4. Clustering

Multiple elbow points suggest the presence of subgroups within the clusters, Figure 4. To explore this, vary the cluster number to see what patterns are revealed. Once the clusters have been divulged you could also perform clustering on the clusters, to investigate the subgroups, which may reveal new strains. The elbow method is subjective; to determine a more accurate cluster number, Peter J. Rousseeuw suggests using a silhouette score as well (Rousseeuw, 1987).

Biological data is not spherical as K-means assumes (Jain, 2008). K-means performance was limited by the parameters used. The nstart value for centroid initialisation was low and the cluster number, k, was capped. This occurred as the dataset was large; values above the basic parameters were not computationally achievable with the equipment and limited access to it.

Hierarchical clustering has previously been performed successfully on biological datasets, such as for gene expression or protein clustering (Boratyn, Datta and Datta, 2006). There is no requirement to predetermine the number of clusters, allowing you to see how the data groups naturally (Bouguettaya et al., 2015). Hierarchical clustering may provide a better alternative when working with binary datasets.

5.5. Mutations

Likely Pango lineage assignment was performed using web-based applications, covSpectrum, coVariants and outbreak.info (Chen et al., 2022; Anon; Gangavarapu et al., 2023). Once the centroid of each cluster had been identified, defining mutations were input into covSpectrum. covSpectrum looks for similarities between the inputted mutations and known Pango lineages of SARS-CoV-2. Although this approach does not use a lineage assignment method such as Pangolin, it provides insights into the cluster identity and characteristics. Clusters I and II aligned with AY.4, a Delta strain, the presence of Delta mutations L452R and P681R supports this.

Cluster III showed similarity with the Omicron strain, which also has a large number of defining mutations (Papanikolaou et al., 2022). This supports the finding that later occurring variants, such as Omicron, accumulate more mutations, due to having more time to mutate and evolve.

Clusters I and II likely diverged from a common ancestor, as evidenced by the shared alignment of the Delta strain and the numerous overlapping defining mutations. Cluster III potentially displays convergent evolution, obtaining advantageous mutations separately from the other clusters while under the same selective pressures. Similar to Omicron, which was reported to have evolved independently from other variants (Ito et al., 2023).

Cluster	Key mutations	Potential advantages
Cluster I	L452R, P681R, D614G, T478K, G142D	Increased receptor binding, enhanced spike protein stability and immune escape
Cluster II	L452R, P681R, D614G, T478K, G142D	Increased receptor binding, enhanced spike protein stability and immune escape
Cluster III	N501Y, Q498R, E484A, K417N, N679K, Y505H, D796Y, D614G, G142D, T478K	Enhanced transmissibility, immune escape, impaired antibody binding and increased receptor binding

Table 2. Key spike protein mutations provide potential advantageous functions for each cluster. L452R (Motozono et al., 2021), P681R (Saito et al., 2022), D614G (Korber et al., 2020), T478K (Di Giacomo et al., 2021), G142D (McCallum et al., 2021), N501Y (Liu et al., 2022), Q498R (Zahradník et al., 2021), E484A (Jangra et al., 2021), K417N (Cao et al., 2022, N679K (Vu et al., 2023), Y505H (Wang et al., 2023), D796Y (Elko et al., 2024).

We considered how many key mutations are present in each cluster, how many of these mutations occur within the RBD and RBM and their potential impact on viral infectivity. Cluster III may be the most infectious, as it contains the highest number of mutations within the RBD and RBM regions. These include many key mutations which will increase SARS-CoV-2 infectivity, Table 2. A key mutation, N501Y, is an asparagine to tyrosine substitution that introduces hydrophobic and π-π stacking interactions with ACE2, increasing the stability of the RBD-ACE2 complex and enhancing transmissibility (Liu et al., 2022). Clusters I and II may be less infectious than Cluster III, owing to the absence of key mutations present of the RBD and RBM, that are vital for ACE2 binding and cell entry. Cluster III contains a synonymous mutation at residue 1146. It has been reported that synonymous mutations can alter the structure of RNA, resulting in advantages for Cluster III, without changing the amino acid code (Boon, Sia and Ng, 2021; Goymer, 2007).

I propose validating these claims using the method described by Wang et al. Mutations are induced for each cluster on the wild type variant, the binding free energy change between RBD and ACE2 is then measured, as this is proportional to SARS-CoV-2 infectivity (Wang et al., 2021).

6. Conclusion

t-SNE-assisted K-means clustering revealed 3 clusters, each of which contains several key mutations that have the potential to increase the infectivity of the cluster variant. Cluster III most likely represents the most infectious strain. Spike protein mutations dictate how vaccines function and their effectiveness (Liu et al., 2021). Through clustering analysis and profiling, we provide an effective solution to deal with large datasets and reveal key mutations that can guide future vaccine development. Studying SARS-CoV-2 is now vital for the UK to manage the next pandemic more effectively.

Abbasian, M. H. et al. (2023). Global landscape of SARS-CoV-2 mutations and conserved regions. *Journal of translational medicine*, 21 (1), p.152.
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.
Arnoldi, W. E. (1951). The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quarterly of applied mathematics*, 9 (1), pp.17–29.
Arora, P. et al. (2020). Learning from history: Coronavirus outbreaks in the past. *Dermatologic therapy*, 33 (4), p.e13343.

Bates, D., Maechler, M. and Jagan, M. (2000). Matrix: Sparse and dense matrix classes and methods. *CRAN: Contributed Packages, The R Foundation*. [Online]. Available at: doi:10.32614/cran.package.matrix.
Boon, W. X., Sia, B. Z. and Ng, C. H. (2021). Prediction of the effects of the top 10 synonymous mutations from 26645 SARS-CoV-2 genomes of early pandemic phase. *F1000Research*, 10, p.1053.
Boratyn, G. M., Datta, S. and Datta, S. (2006). Biologically supervised hierarchical clustering algorithms for gene expression data. In: *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*. August 2006. IEEE. [Online]. Available at: doi:10.1109/iembs.2006.4398704.
Bouguettaya, A. et al. (2015). Efficient agglomerative hierarchical clustering. *Expert systems with applications*, 42 (5), pp.2785–2797.
Bradley, P. and Fayyad, U. (1998). Refining initial points for K-means clustering. *International Conference on Machine Learning*, pp.91–99. [Accessed 5 April 2025].

Cai, T. T. and Ma, R. (2021). Theoretical foundations of t-SNE for visualizing high-dimensional clustered data. *arXiv [stat.ML]*. arXiv [Online]. Available at: http://arxiv.org/abs/2105.07536.
Chen, C. et al. (2022). CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. *Bioinformatics (Oxford, England)*, 38 (6), pp.1735–1737.
Chen, J. et al. (2020). Mutations strengthened SARS-CoV-2 infectivity. *Journal of molecular biology*, 432 (19), pp.5212–5226.
Chourasia, P., Ali, S. and Patterson, M. (2022). Informative initialization and kernel selection improves t-SNE for biological sequences. In: *2022 IEEE International Conference on Big Data (Big Data)*. 17 December 2022. IEEE. pp.101–106.
Chung, N. C. et al. (2019). Jaccard/Tanimoto similarity test and estimation methods. *arXiv [stat.ME]*. arXiv [Online]. Available at: http://arxiv.org/abs/1903.11372.
COVID-19 cases. [Online]. *datadot*. Available at: https://data.who.int/dashboards/covid19/cases?n=c [Accessed 21 February 2025a].
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].
Cui, M. (2020). Introduction to the K-Means Clustering Algorithm Based on the Elbow Method. *Clausius Scientific Press, Canada*.
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.

D'Ippolito, R. A. et al. (2021). Refining the N-termini of the SARS-CoV-2 spike protein and its discrete receptor-binding domain. *Journal of proteome research*, 20 (9), pp.4427–4434.

Elbe, S. and Buckland-Merrett, G. (2017). Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global challenges (Hoboken, NJ)*, 1 (1), pp.33–46.
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].

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.
Frost, H. R. (2022). Eigenvectors from Eigenvalues Sparse Principal Component Analysis (EESPCA). *Journal of computational and graphical statistics*, 31 (2), pp.486–501.
Futuyma, D. (2013). *Evolution*. Sinauer.

Gangavarapu, K. et al. (2023). Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. *Nature methods*, 20 (4), pp.512–522. [Accessed 3 May 2025].
Goymer, P. (2007). Synonymous mutations break their silence. *Nature reviews. Genetics*, 8 (2), pp.92–92. [Accessed 6 May 2025].

Harvey, W. T. et al. (2021). SARS-CoV-2 variants, spike mutations and immune escape. *Nature reviews. Microbiology*, 19 (7), pp.409–424. [Accessed 30 April 2025].
Hinton, G. and Roweis, S. Stochastic neighbor embedding. [Online]. Available at: https://proceedings.neurips.cc/paper_files/paper/2002/file/6150ccc6069bea6b5716254057a194ef-Paper.pdf [Accessed 14 April 2025].
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.

Jain, A. K. (2008). Data clustering: 50 years beyond K-means. In: *Machine Learning and Knowledge Discovery in Databases*. Lecture notes in computer science. Berlin, Heidelberg: Springer Berlin Heidelberg. pp.3–4.
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.
Linderman, G. C. et al. (2019). Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nature methods*, 16 (3), pp.243–245. [Accessed 5 May 2025].
Liu, Y. et al. (2022). The N501Y spike substitution enhances SARS-CoV-2 infection and transmission. *Nature*, 602 (7896), pp.294–299.

Liu, Z. et al. (2021). Identification of SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization. *Cell host & microbe*, 29 (3), pp.477-488.e4.

Marik, P. E. et al. (2021). A scoping review of the pathophysiology of COVID-19. *International journal of immunopathology and pharmacology*, 35, p.20587384211048024.
Mathieu, E. et al. (2020). *COVID-19 Pandemic*. *Our World in Data*.
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].
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.
Metzger, J. J. and Eule, S. (2013). Distribution of the fittest individuals and the rate of Muller's ratchet in a model with overlapping generations. *PLoS computational biology*, 9 (11), p.e1003303.

Na, S., Xumin, L. and Yong, G. (2010). Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm. In: *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*. April 2010. IEEE. pp.63–67.

Papanikolaou, V. et al. (2022). From delta to Omicron: S1-RBD/S2 mutation/deletion equilibrium in SARS-CoV-2 defined variants. *Gene*, 814 (146134), p.146134.
Phillips, P. C. (2008). Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nature reviews. Genetics*, 9 (11), pp.855–867. [Accessed 2 May 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.

Rambaut, A. et al. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature microbiology*, 5 (11), pp.1403–1407. [Accessed 16 April 2025].
Richardson, M. Principal Component Analysis. [Online]. Available at: http://www.sdss.jhu.edu/~szalay/class/2024/etc/SignalProcPCA.pdf [Accessed 13 April 2025].
Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20, pp.53–65.

Sadrjavadi, K. et al. (2015). Comparison of correlation ranking and eigenvalue ranking unfolded principal component regression for direct determination of naproxen in human serum using excitation–emission matrix fluorescence spectroscopy. *Journal of the Iranian Chemical Society*, 12 (6), pp.967–977.
Song, W. et al. (2019). Improved t-SNE based manifold dimensional reduction for remote sensing data processing. *Multimedia tools and applications*, 78 (4), pp.4311–4326.

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].

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].
Wu, F. et al. (2020). A new coronavirus associated with human respiratory disease in China. *Nature*, 579 (7798), pp.265–269.

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].

Zhang, J. et al. (2021). Structure of SARS-CoV-2 spike protein. *Current opinion in virology*, 50, pp.173–182.

Online Resource
CoVariants. [Online]. Available at: https://covariants.org/ [Accessed 5 May 2025].