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

[vandermaaten08a.pdf](https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf?fbcl)

Has very clean looking figures and good overall layout

[Principal component analysis - Analytical Methods (RSC Publishing) DOI:10.1039/C3AY41907J](https://pubs.rsc.org/en/content/articlehtml/2014/ay/c3ay41907j)

Has very clean looking figures

Make figures big

Abstract

One line introduction:

Building on UMAP paper, how to tackle large datasets

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 ongoing 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.

What work has already been done?

**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. There is also insufficient information publicly available about how to analysis datasets of this to extract meaningful features. 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. Dimensionality reduction techniques will need to be utilised due to the size of the dataset, 3 methods were selected for this. 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 defining cluster 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: principal 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

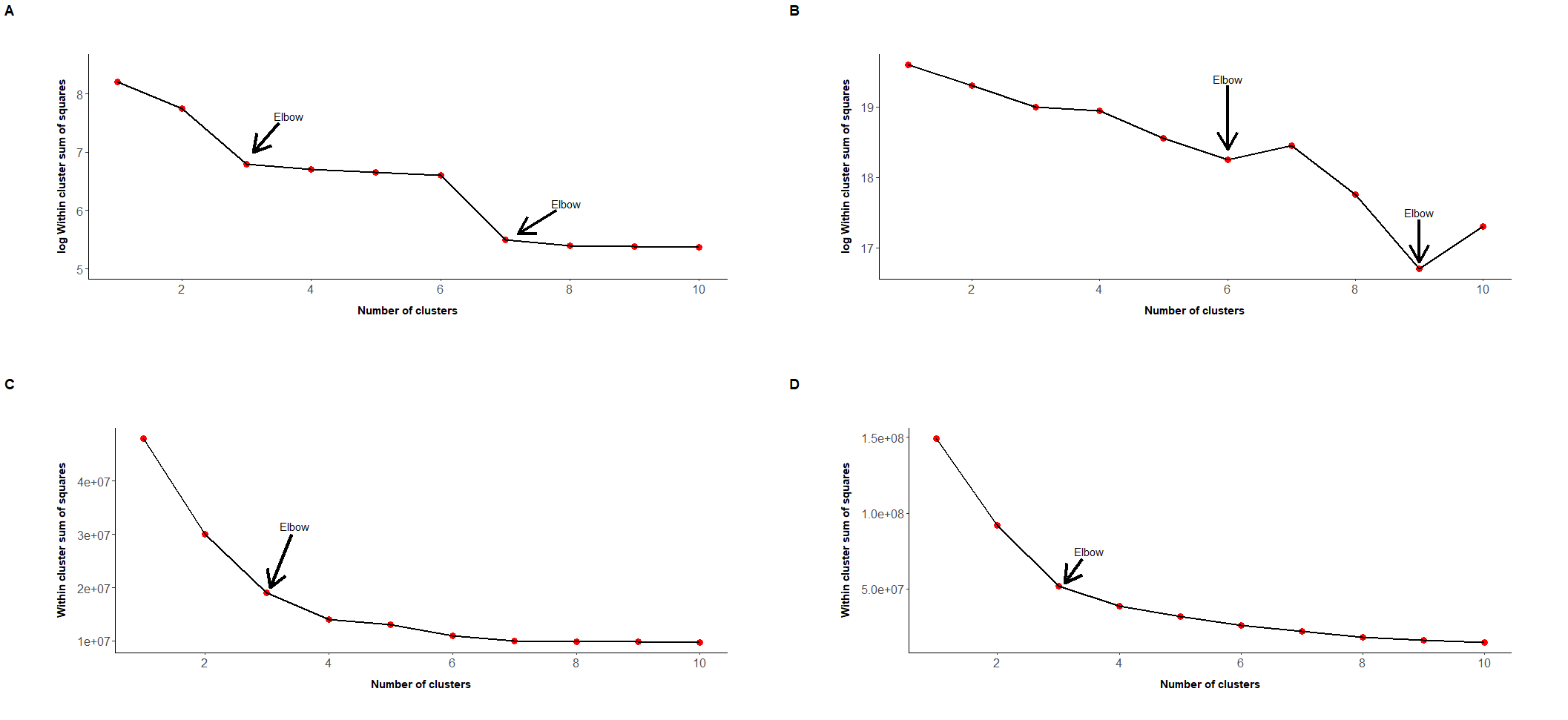
AI-generated content may be incorrect.

**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. 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 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 binary dataset containing the 1000 most common mutations and their associated sequence identifier. 1’s in the dataset corresponded to the presence of a mutation and 0’s to the absence of that mutation. 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.

**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.

I employed a 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 principal 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. 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, components after PC9 add very little variance. 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 4A-B.

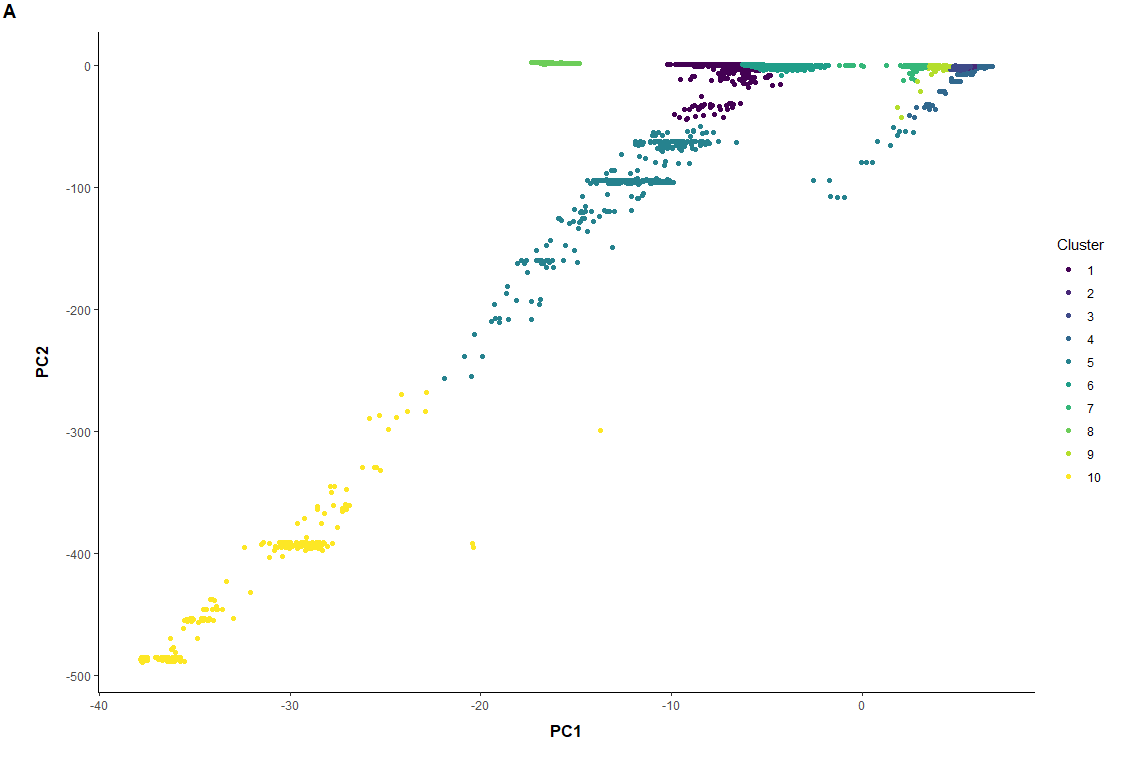
Utilising t-SNE to determine if the number of clusters identified by PCA is the same. t-SNE is a non-linear random dimensionality reduction algorithm that retains the local structure in high dimensional data, while also reflecting the global structure (van der Maaten, 2008). Duplicate values had to be removed from the dataset, as the t-SNE failed first time. 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 paired with PCA. Comparison of wcss across various cluster numbers using the elbow method discovered 3 clusters, Figure 4C.

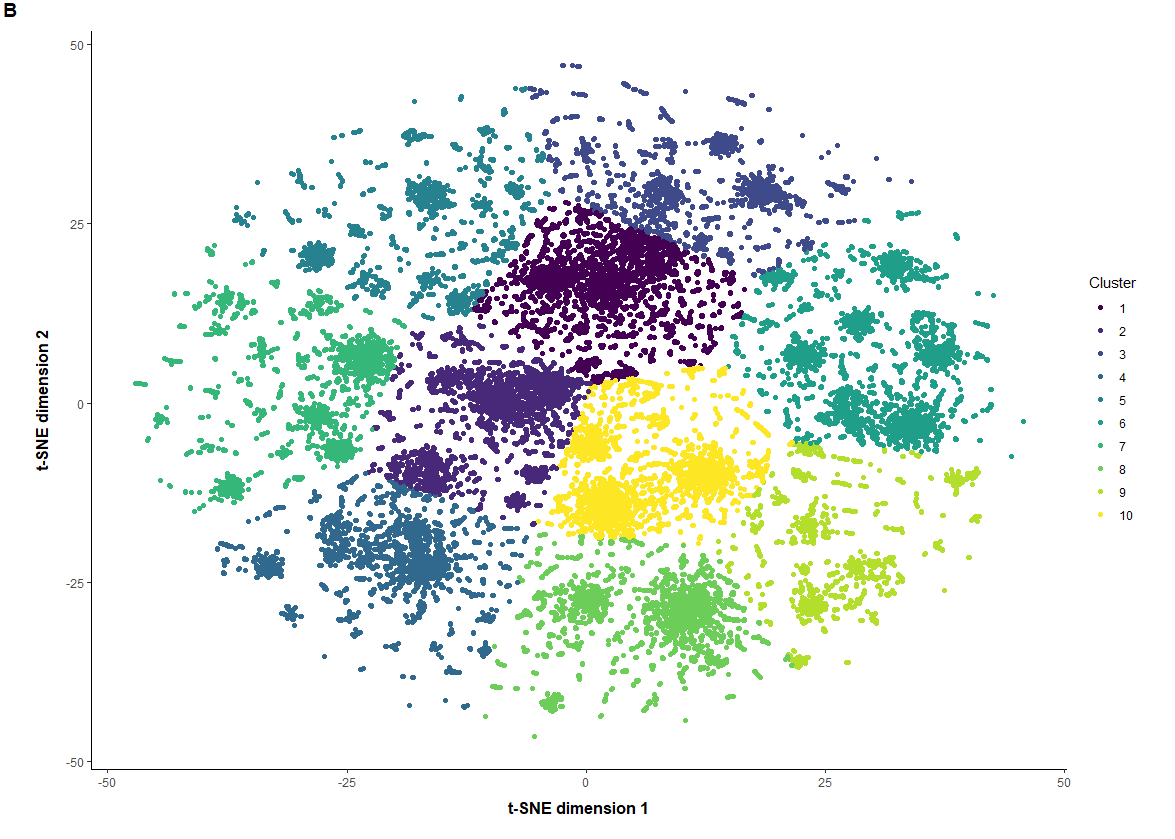
The final dimensionality reduction technique used was UMAP. Another non-linear reduction method, UMAP has been reported to be more efficient than t-SNE and is potentially better at keeping more of the global structure than t-SNE (McInnes, Healy and Melville, 2018). The UMAP algorithm was struggling with the size of the dataset. The dataset was converted to a sparse matrix then back into tibble format. A sparse matrix 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 random 500000 sample was taken and UMAP was performed on this. This was less than ideal as important features and relationships may be lost by taking a sample. 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 twofold increase in efficiency than K-means paired with PCA. Comparison of wcss across various cluster numbers using the elbow method discovered 3 clusters, Figure 4D.

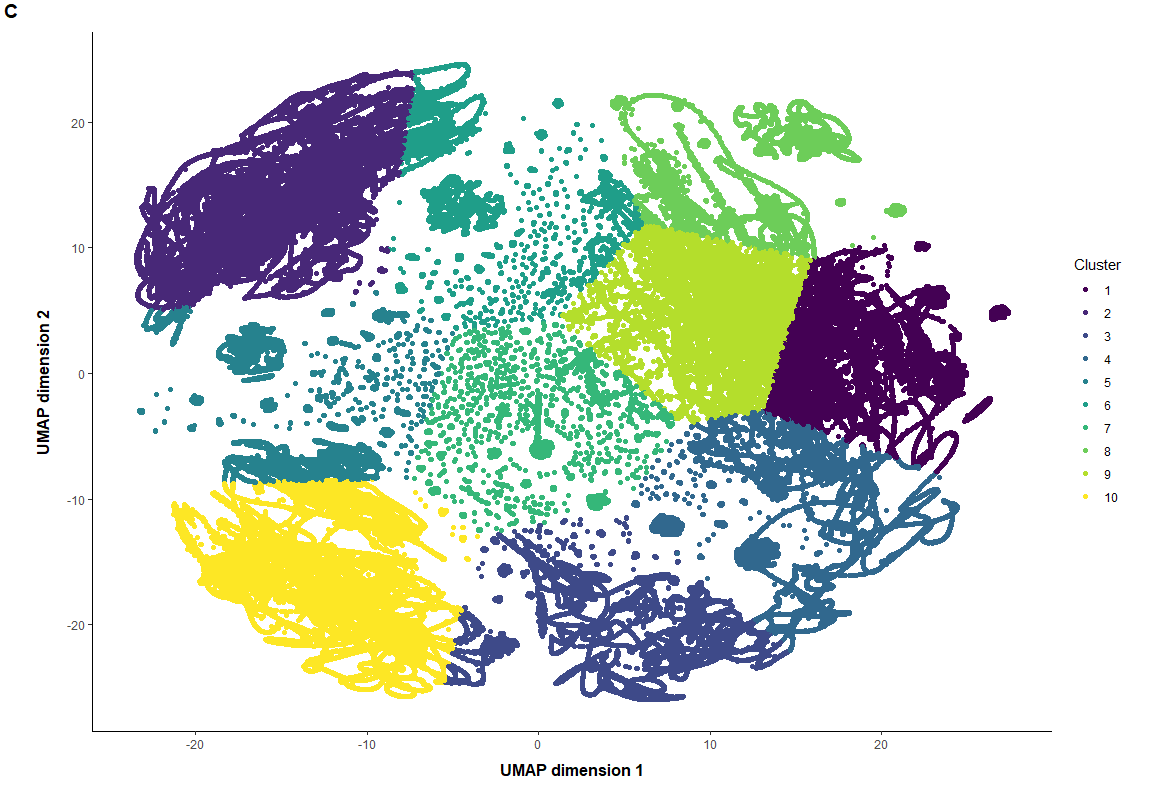
Figure 5, shows the visualisation of dimensionality reduction techniques on the 1000 mutation SARS-CoV-2 data. All the datasets were reduced to 2 dimensions to allow for better visualisation. PCA performs poorly, this can be seen in Figure 5A with very bad and erratic clustering. t-SNE provides much clearer and defined clustering than PCA, Figure 5B. Figure 5C, shows that UMAP gives the clear clustering compared to PCA, however, there is a large amount of overlap between the clusters. This makes it hard to distinguish the individual cluster groups.

Large data sets can be 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.

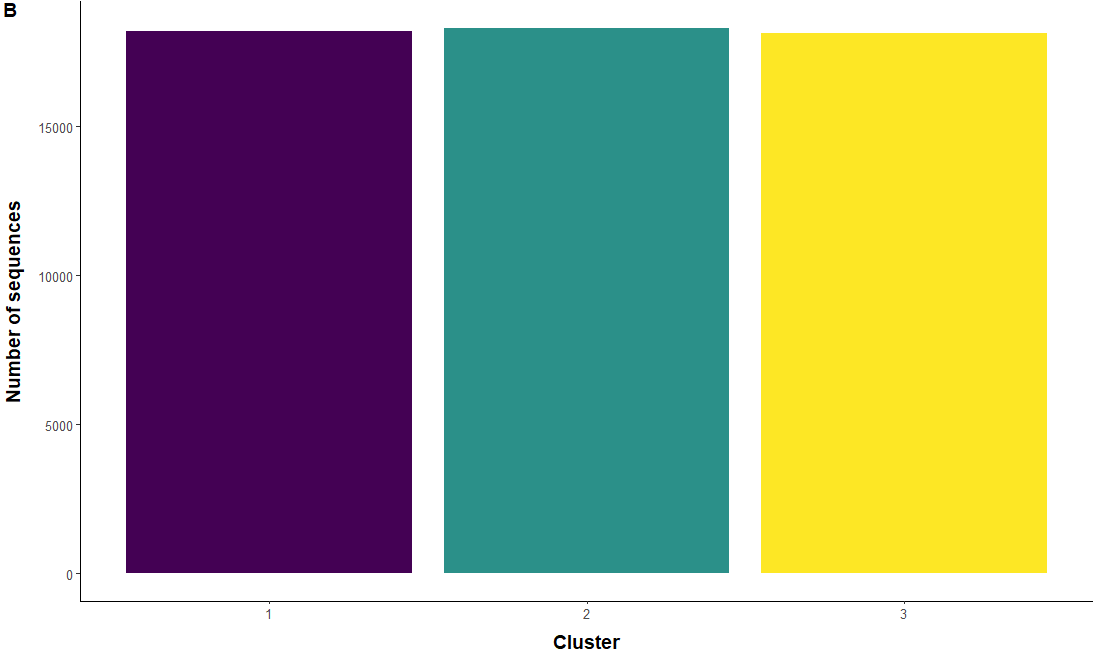
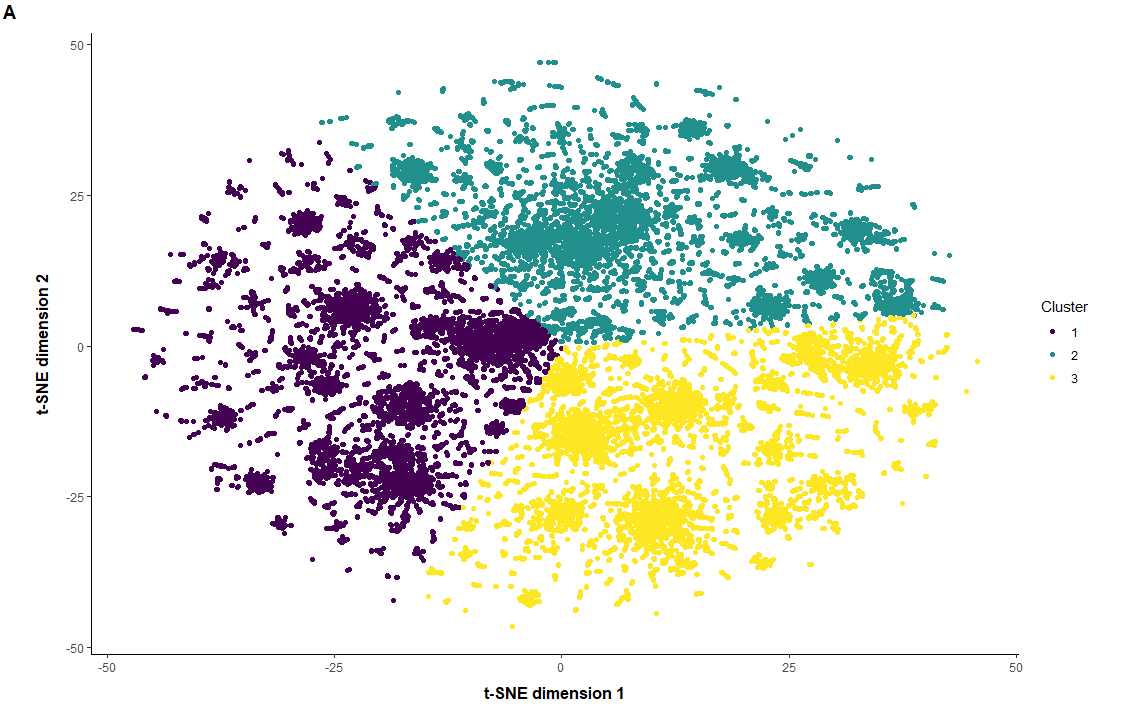
With this considered, t-SNE dimensionality reduction method paired with K-means was selected. t-SNE dimensionality technique provides the clearest clusters and was the most efficient, however, t-SNE does also retain the fewest original features. Time constraints and access to equipment were not too much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







**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.

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

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

Using t-SNE dimensionality reduction technique paired with K-means we 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. Figure 6A depicts distinct groups, with little to no overlap between the clusters.

Figure 6A, depicts 3 clusters that were identified. This was determined using the elbow method, Figure 4, and through assessing which dimensionality method provided the clearest clustering visualisation. This cluster number shows 3 clearly 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. Figure 6B, shows a box plot of the number of sequences per cluster, illustrating the distribution of sequences across the clusters. Cluster I had the most sequences, 18300, Cluster III had the fewest sequences, 18144. Cluster II had 18300 sequences. A Kruskal-Wallis test was used, since the original data is non-parametric, to reveal that the cluster sizes were not significantly different from each other, p-value = 0.368. 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 as mentioned previously 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 had to be 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-mutation are mutations that are occur in close proximity to each other or simultaneously, they can provide an enhanced effect on protein structure and function **(source)**.

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. Cluster I and Cluster 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 also the only cluster that had 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. Cluster I and Cluster 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 **(source)**.

|  |  |  |  |
| --- | --- | --- | --- |
| 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

The mutations that are the drivers of separation between clusters were also identified. These were mutations that were unique to certain clusters and differentiate 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 defining 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.

Possible reasons for multiple clusters?

What is interesting?

Once the centroid of each cluster had been identified, Phylogenetic Assignment of Named Global Outbreak Lineages, pangolin, could be used (O’Toole et al., 2021). This involves assigning each cluster to a Pango lineage of SARS-CoV-2. Pango lineage assignment provides an insight in the characteristics and specific traits of the clusters. Cluster I matched with **X**. Cluster II aligned with **Y**. Cluster III showed similarity with **Z**.

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

What clusters share the same root?

Do certain clusters relate to known variants of SARS-CoV-2?

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

Are any of the clusters not assigned to a lineage?

**Figure 8.** Cluster evolution over time, scatter plot clusters over time, will show the evolution of the clusters, what clusters share the same root

Which clusters share the same root?

How do clusters evolve over time?

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

**Figure 9.** Geographical distribution was assessed by mapping clusters across the UK. There was no available data for Northern Ireland.

Proportion of clusters for each region?

What is the dominate cluster in each region?

Are any clusters found in just one region of the UK?

**3. Mutation analysis**

**Figure 9. 3D visualisation of spike glycoprotein binding to human ACE2 receptor. RBD residues have been coloured X and RBM residues have been coloured Y.**

For each cluster I determined how many of its mutations are located within the RBD region and RBM residue region. The RBD and RBM are vital for spike protein binding to the host ACE2 receptor, examining mutations in these regions can provide insight how mutations increase or decrease viral infectivity. The RBD is located between residues 319-541, with the RBM residues positioned at 437-508 within the RBD, Figure X **(source)**. In Cluster I and Cluster 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 of positioned within the RBM region.

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

UMAP has been reported to be the best?

Link to previous studies on dimensionality reduction methods

Which dimensionality technique is the best for data of this size

Why a Jaccard distance metric should have been used over Euclidean distance metric

Need to use that on all of them

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.

**How consistent are the clusters across different dimensionality reduction methods?**

Lack off

**4.4. Clustering**

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

What are the problems with unsupervised learning techniques?

Shared roots evidence of divergent evolution

Divergent and convergent evolution

Why is this important?

Common ancestor

Try different clustering values and what will they show

How the number of clusters was determined is very subjective

**4.5. Mutations**

Do any of the mutations identified have anything in common with these mutations?

What is a synonymous mutation and what is nonsynonymous mutation?

Presence of synonymous mutation in Cluster III

Which cluster could potentially be the most infectious?

What evidence is there for this?

What structural changes could the mutations make?

Amino acid changes

Which cluster could potentially be the least infectious?

What evidence is there for this?

What structural changes could the mutations make?

Amino acid changes

Make sure to link structure to function

Residue Role

Y449 ACE2 binding

L455 ACE2 interaction

F456 Structural core

F486 Strong ACE2 binder

N487 Polar interaction

Y489 ACE2 contact

Q493 Major ACE2 interaction

Q498 Hydrogen bonding

N501 Key mutation site (e.g., N501Y)

Y505 Important for receptor binding

**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**

The original dataset was provided by Dr Richard Bingham from the University of York, which had been deposited in GISAID. This was then imported into RStudio for analysis. 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 1984861 individual sequences and mutations, was then manipulated using R in various ways to achieve the analysis required.

For the clustering a second dataset was created. As mentioned previously this dataset contained the 1000 most common mutations, each mutation was a column and the rows were the individual sequences. Each value was either a 1 or 0, corresponding to the presence or absence of a given mutation in that specific sequence. The dimensions of this dataset were 1984612x 1001, one column for the sequence identifier and 1000 mutation columns. The sequence identifier column had to be dropped as it is not a numeric feature. Clustering using all the mutations was too computational 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.

**6.2. R packages and R scripts 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, umap and uwot. K-means clustering was performed using the base R function. Matrix package

Sources for the packages used be found in the research compendium. A seed was set in RStudio for reproducibility, using set.seed(123). Information on the scripts used for the analyses see the provided research compendium in the supplementary material.

<https://github.com/sha524/Spike_protein>

**6.3. Nucleotide change to amino acid change**

Once the centroids had been identified the mutations associated with them were divulged. Initially the mutations were in a nucleotide form, non\_C21618G~T-R, the amino acid position had to be determined. The spike glycoprotein gene begins at nucleotide 21563 (Wu et al., 2020). To find the codon position, we first calculate the position within the spike glycoprotein gene:

The amino acid position was then calculated using the formula:

**6.4. Sequence analysis**

After clustering, Pangolin was used for sequence analysis (O’Toole et al., 2021). Pangolin’s main role here was assigning SARS-CoV-2 variants to specific clusters. Pango’s nomenclature for lineage assignment can be found at the following paper (Rambaut et al., 2020). The centroid of each cluster was used for reference, as this contains the mutations that define that cluster. Pangolin’s web interface was used to run the sequence analysis.

**6.5. Statistical analysis**

The Kruskal-Wallis test was used to assess statistical significance. In particular, does the sequence date have a significant effect on the number of mutations per sequence. Spearman’s rank correlation was used to test for correlation between different variables. Tested for correlation between when a sample was taken and the number of mutations for that sequence, to further emphasise the relationship. A Kruskal-Wallis test was also conducted to assess the statistical significance between the different cluster sizes, identified using **X** and K-means.

A generalised additive model (GAM) was used to show the relationship between non-linear variables, such as sample date and the number of sequences per day, Figure 1. GAM was fitted using the mgcv package and allows us to depict the dynamics of the number sequences per day across 2020 - 2024.

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.

**6.6. K-means clustering**

For information on the R script used, see the research compendium on the GitHub link. K-means clustering is an unsupervised learning technique, that involves finding subgroups within a larger dataset (Bradley and Fayyad, 1998). K-means clustering breaks observations into a pre-defined number of clusters, which was calculated using the elbow method. The algorithm works by first randomly assigning each point to one the pre-defined clusters. The centres of each of these subgroups is then calculated. Each point is then assigned to the nearest newly calculated cluster. This is then repeated for several iterations, until each point is assigned to the nearest cluster (Na, Xumin and Yong, 2010). The total number of clusters selected was 3**.**

The elbow method for cluster determination, involved calculating the within cluster sum of squares (wcss) for each cluster numbers 1-10. nstart parameter, the number of iterations used for each cluster 1-10 was 10. The within cluster sum of squares was then plotted against the number of clusters. The elbow point was then determined subjectively as the value where the wcss stops decreasing significantly. A high wcss means that the clusters are more spread out and a low wcss means that the clusters are more compact (Cui).

Once the number of clusters had been selected, K-means was performed with number of clusters. nstart, the number of iterations used was 10.

**6.7. Principal component analysis (PCA)**

PCA is a linear dimensionality reduction method; its main role in finding 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 on the GitHub link. Due to the most common 1000 mutation dataset being in a binary format, presence or absence of a mutations, there was no need for scaling or centring the data around zero. 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 I selected principal components 1 and 2, PC1 and PC2, to begin with. Figure S1 depicts the scree plot used to select the number of principal components to retain. As seen in 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 values’ principal component. These values are then plotted in this new PCA coordinate space, using PC1 and PC2/PC9 as the axes. Each point was coloured according to the assigned cluster.

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

t-SNE is a non-linear dimensionality reduction technique, that maintains the local structure by keeping neighbouring points close together during the process. The algorithm reduces the amount of information in the data, while attempting to keep the local and global structures (van der Maaten, 2008). For each point in the high-dimensional space, t-SNE calculates the probability of specific point neighbouring all the other points 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 high probability, than points that are 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 then iteratively adjusted until the low-dimensional data resembles the high-dimensional data (Cai and Ma, 2021).

t-SNE was first employed on the original 1000 mutations dataset. This failed due to duplicates, which disrupt the pairwise distance metric. The duplicates were removed and the algorithm was successfully ran again. The number of datapoints retained after the duplicates were removed was **X**. The algorithm was performed using a perplexity of 30, 500 iterations and the data was reduced to two dimensions to aid visualisation.

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

**6.9. Uniform Manifold Approximation and Projection (UMAP)**

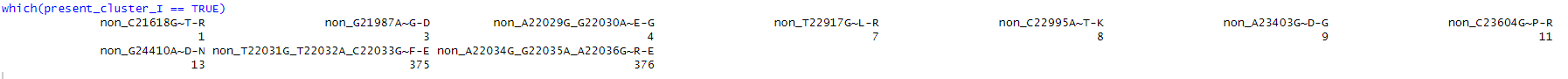
UMAP is a relatively new non-linear graph based dimensionality reduction technique, designed to preserve the local and global structure of the data (McInnes, Healy and Melville, 2018). First a graph in high-dimensional space is made and k-nearest neighbours method is used to work out the probability of points neighbouring each other. Then 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-dimension graph and the low-dimensional graph (McInnes, Healy and Melville, 2018).

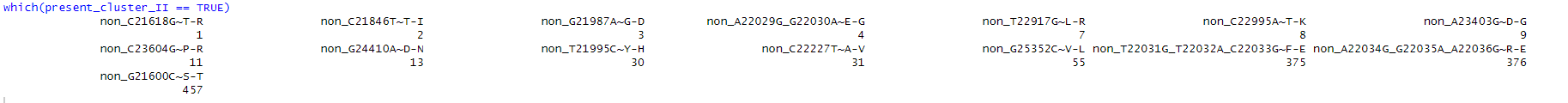
UMAP was performed on the 1000 most common mutation dataset. A sparse matrix, from the Matrix package, was employed as UMAP could not process the whole dataset. This sparse matrix was then converted back to a tibble, as UMAP cannot handle sparse data matrices. Sparse matrix method still do not work, therefore, a random 500000 sample was taken. 500000 was selected as this appeared to be the upper limit, without causing RStudio to crash. The following default parameters were used: n\_neighbors set to 15, min\_dist of 0.1 and the n\_components equal to 2.

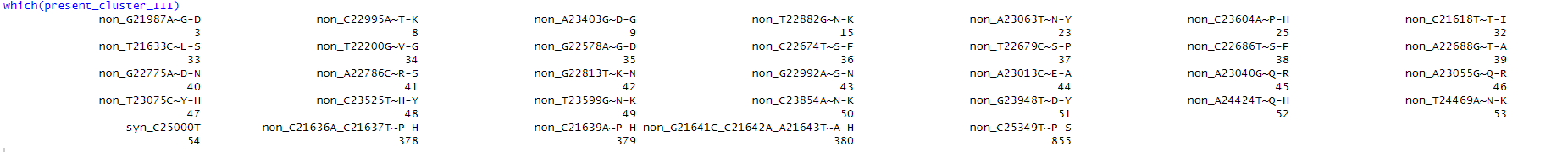
To visualise the results of the dimensionality reduction ggplot2 was again used, Figure 5C. The cluster id’s generated by K-means were assigned to each values’ new low-dimensional coordinates. The points were then plotted according to their position in the new two dimensions. Each point was coloured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

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].

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.

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.

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.

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.

Cui, M. *On the elbow method*. [Online]. Available at: doi:10.23977/accaf.2020.010102.

Frost, H. R. (2022). Eigenvectors from Eigenvalues Sparse Principal Component Analysis (EESPCA). *Journal of computational and graphical statistics: a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, 31 (2), pp.486–501.

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].

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.

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].

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.

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.

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

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].

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.

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