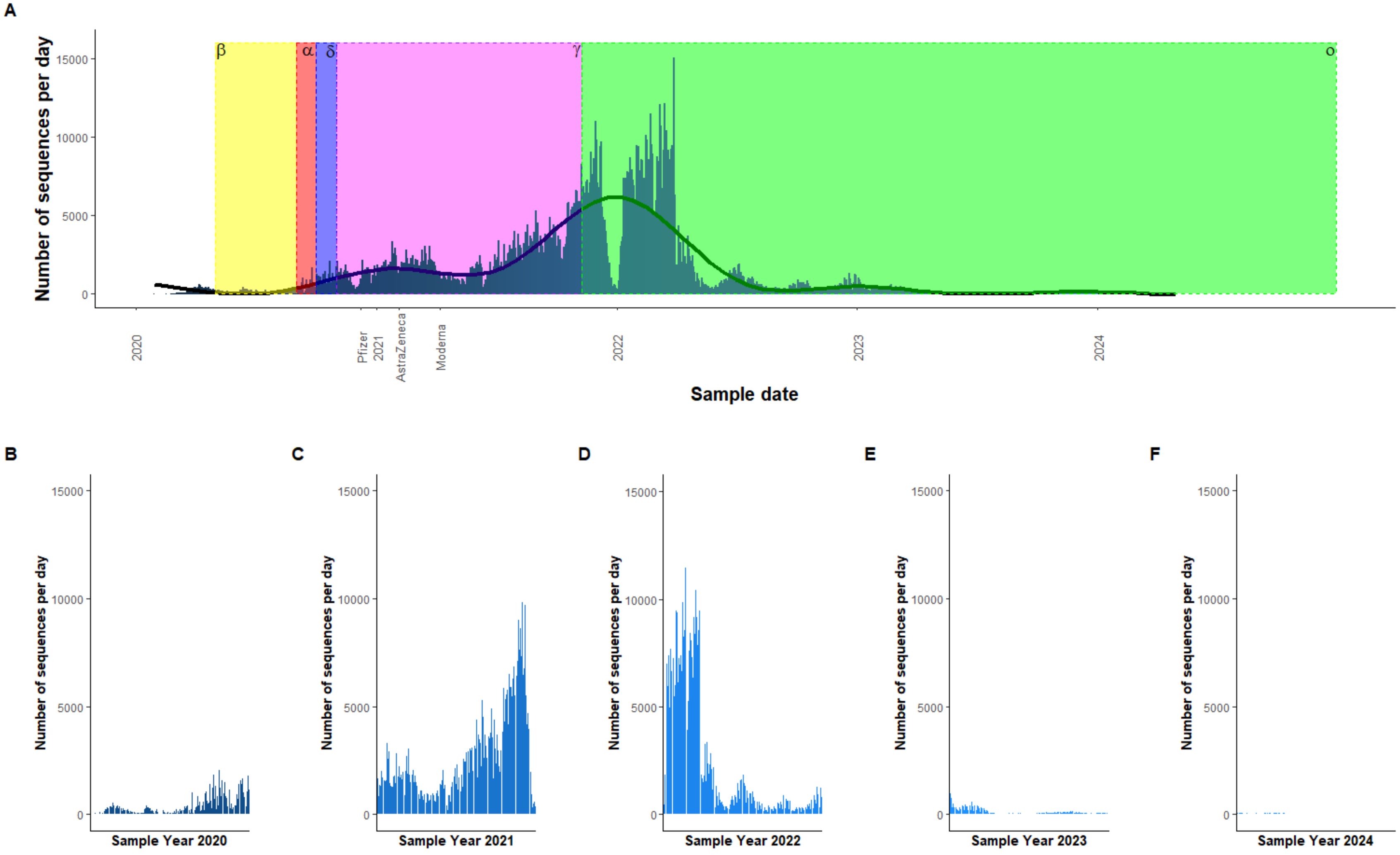

Figure 1. The number of sequences per day from 2020 – 2024. **A** The overall the number of sequences per day from years 2020 – 2024. The coloured zones represent the emergence of a new strain of SARS-CoV-2. Yellow – beta, red – alpha, purple – delta, pink – gamma, green - omicron. Also on the plot are specific dates showing the first use of that vaccine. A generalised 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.

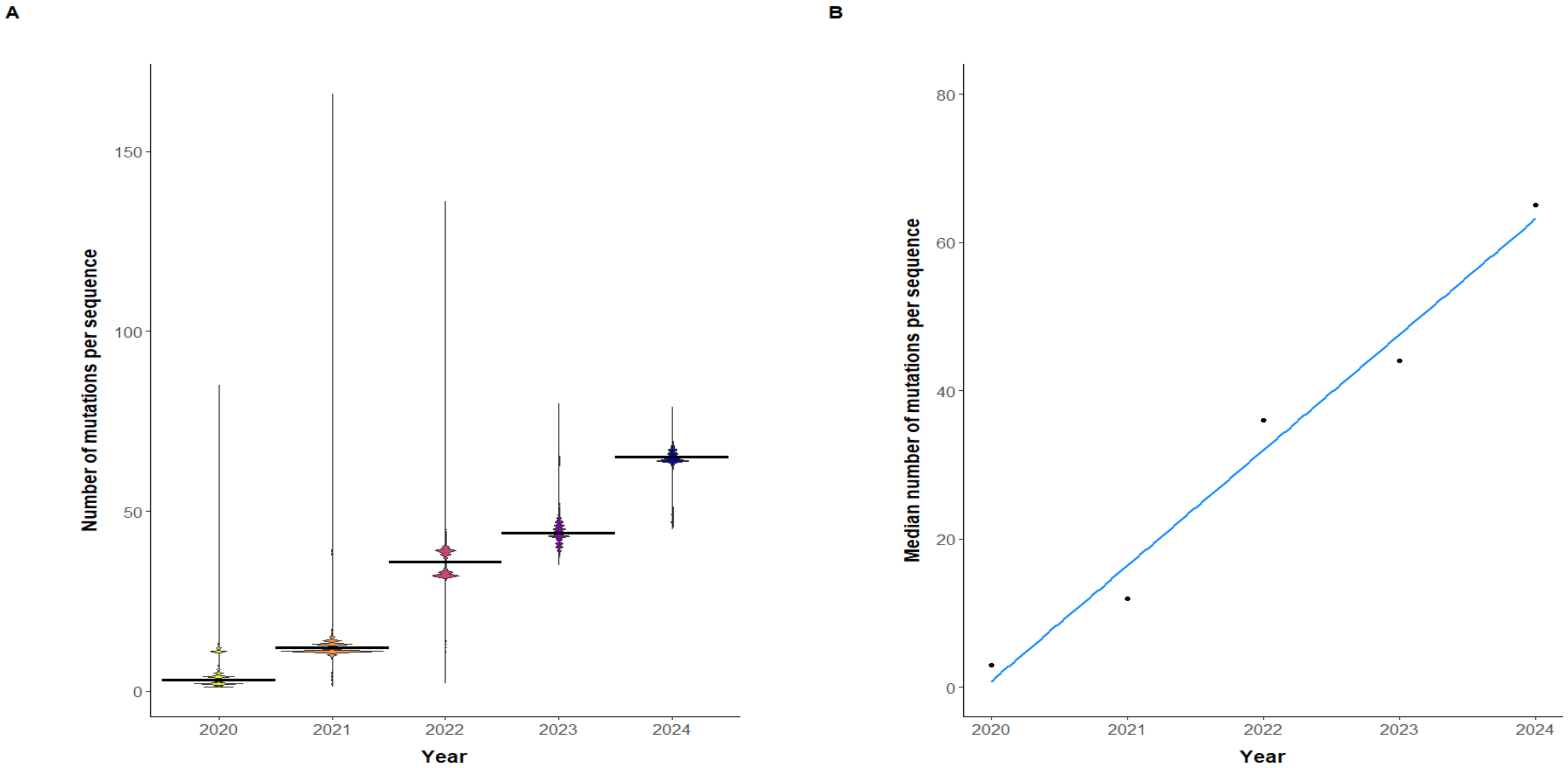

Figure 2. A Violin plot, showing the number of mutations per sequence over time. The sequences have been grouped by year. The thicker black line across each distribution represents the median number of mutations per sequence for that year. Median values are as follows: 2020 = 3, 2021 = 12, 2022 = 36, 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: χ2 = 1558575, d.f. = 4, p > 2.2e-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 ρ = 0.895). The linear regression line suggests a consistent increase in median mutations over time.

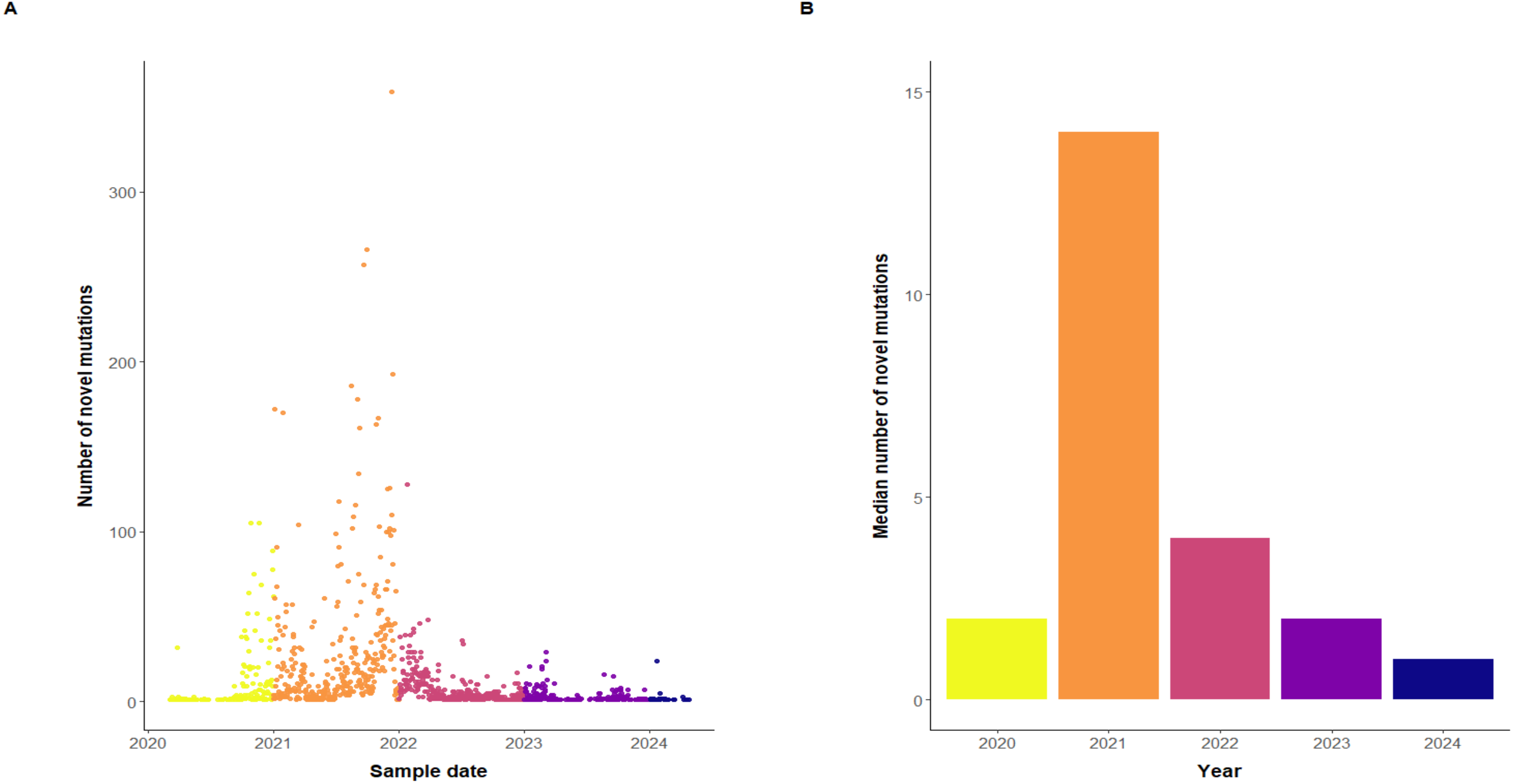

Figure 3. **A** The emergence of novel SARS-CoV-2 spike protein mutations from 2020 - 2024. Figure depicts the number of novel mutations for any given sample date. The colours represent the year the sample was taken: 2020 – yellow, 2021 - orange, 2022 - pink, 2023 - purple and 2024 – dark blue. Mean 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.

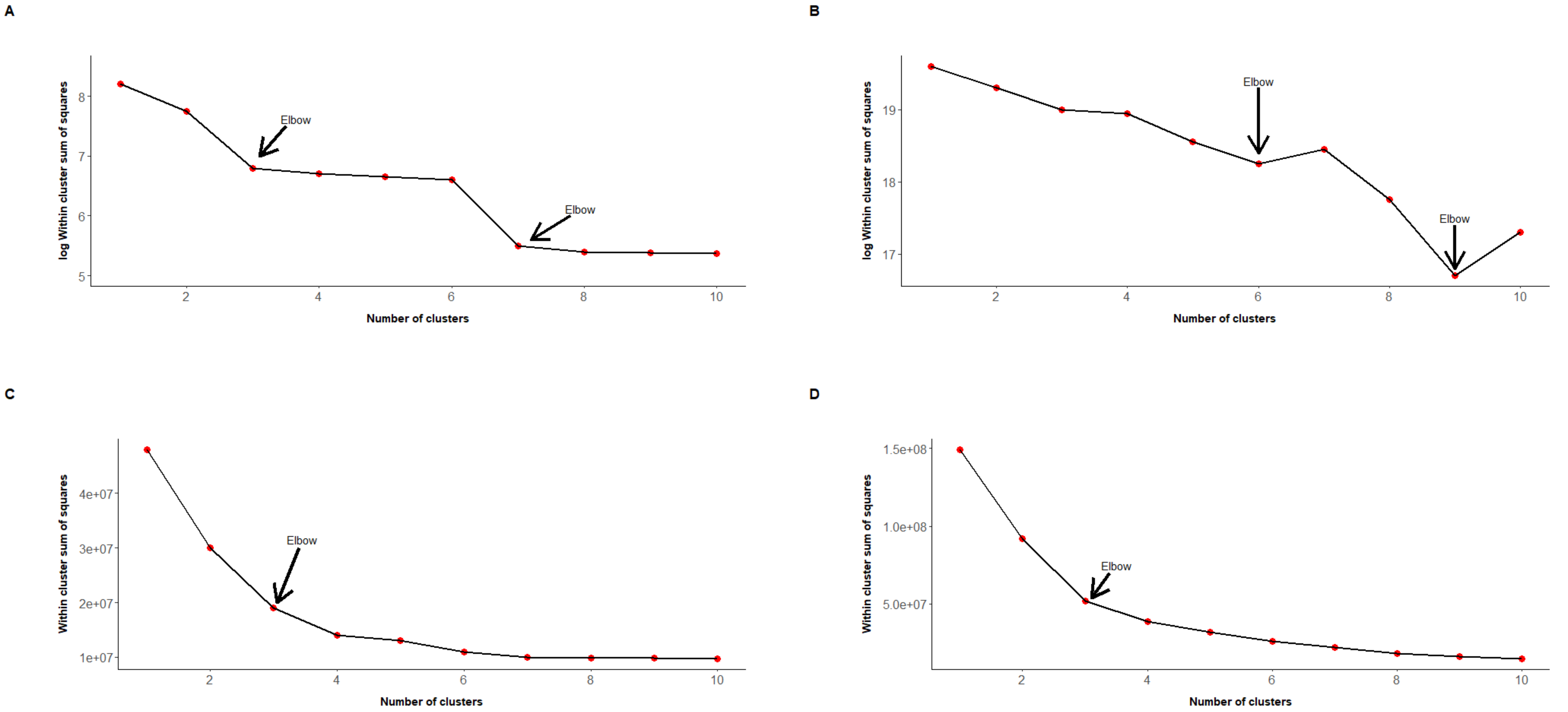

Figure 4. Elbow plots. The optimal number of clusters is shown as the turning point in the in the elbow plots. Within cluster sum of squares of the PCA assisted K-means clustering. **A** PC1 and PC2, then K-means for 1000 most common mutations on a log10 scale. **B** PC1:PC9, followed by K-means for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

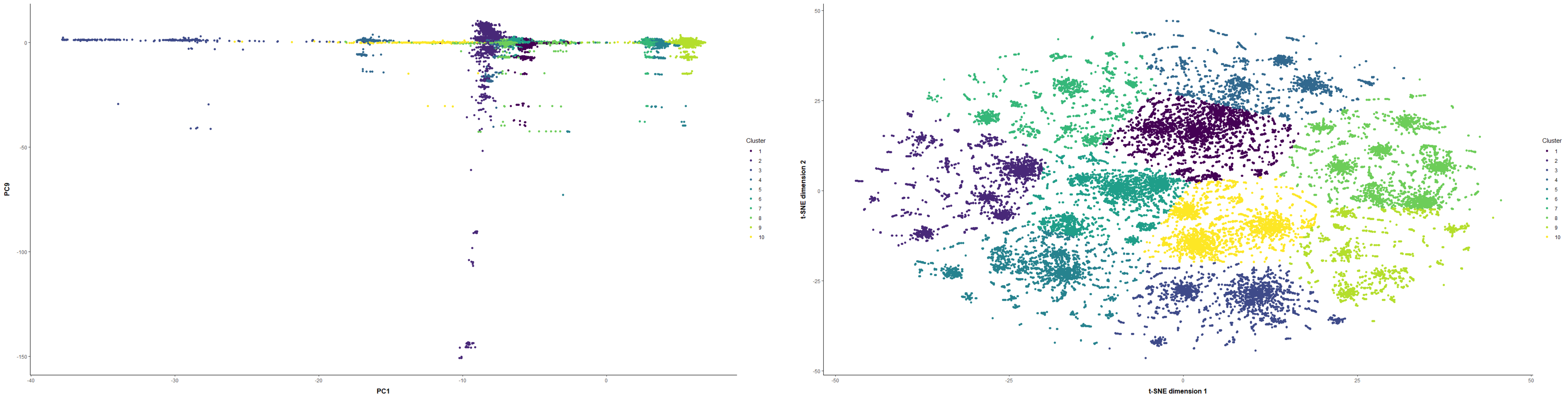

Figure 5. 2D visualisations illustrating the differences of the dimensionality reduction techniques used for the UK SARS-CoV-2 spike glycoprotein dataset, showing 1-10 distinct clusters. The 1-10 clusters were used to work out the actual number of clusters later on. **A** PCA visualisation. **B** t-SNE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

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

Figure 7. Cluster evolution over time. A scatter plot clusters over time, will show the evolution of the clusters, what clusters share the same root

Figure 8. Map of clusters across the regions of UK. Map was generated using R. Each country is coloured according to the dominant cluster. No data was available for Northern Ireland, which has been coloured white.

RCSB PDB - 6M0J: Crystal structure of SARS-CoV-2 spike receptor-binding domain bound with ACE2

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.