Analysing UK SARS-CoV-2 spike protein mutations via t-SNE paired with K-means clustering

Sam Aldous

Context & Aims

COVID-19 pandemic:

- 770,000,000 infections[1]
- Estimated death toll: 18,000,000 – 32,000,000[2]

Spike glycoprotein:

- Receptor binding motif (RBM) residues bind directly to ACE2[3]
- Receptor binding domain (RBD), two regions S1 & S2[4]

Infectivity:

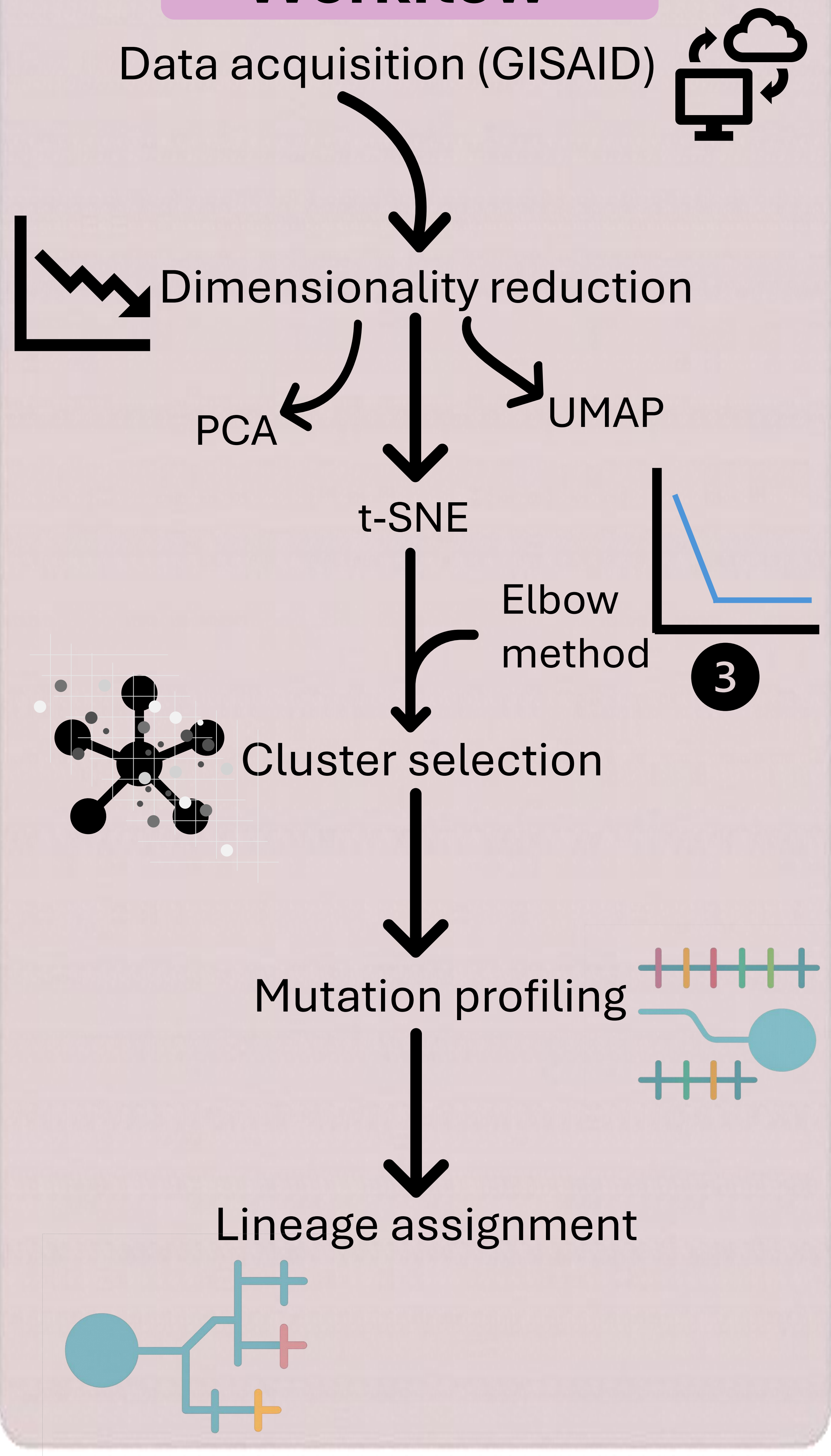
- RBD/RBM mutations modulate infectivity[5]
- Examples: N501Y & D614G enhance affinity and stability changes[6]
- How many key mutations occur in these regions may infer infectivity?

Aims:

Gaps remain in our understanding of SARS-CoV-2 clustering patterns in the UK and extracting meaningful information from large-scale datasets

1. **Divulge temporal and general trends**
2. **Identify which dimensionality reduction technique best pairs with K-means clustering regarding efficiency and cluster clarity**
3. **Reveal dominant clusters and explore their infectivity & potential lineage from their centroids**

Workflow



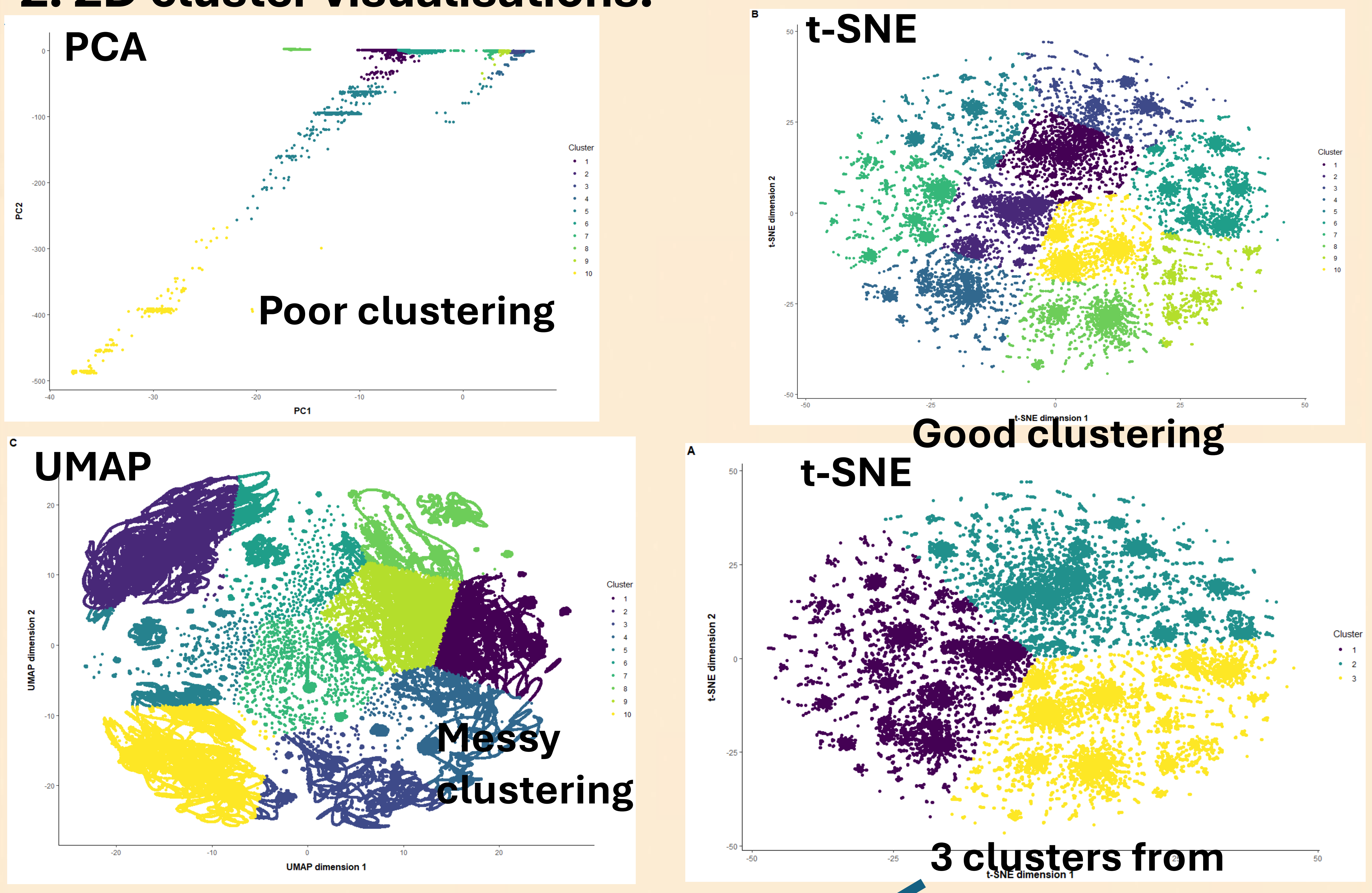
Key results

Over 6 hours

1. Efficiency comparison:

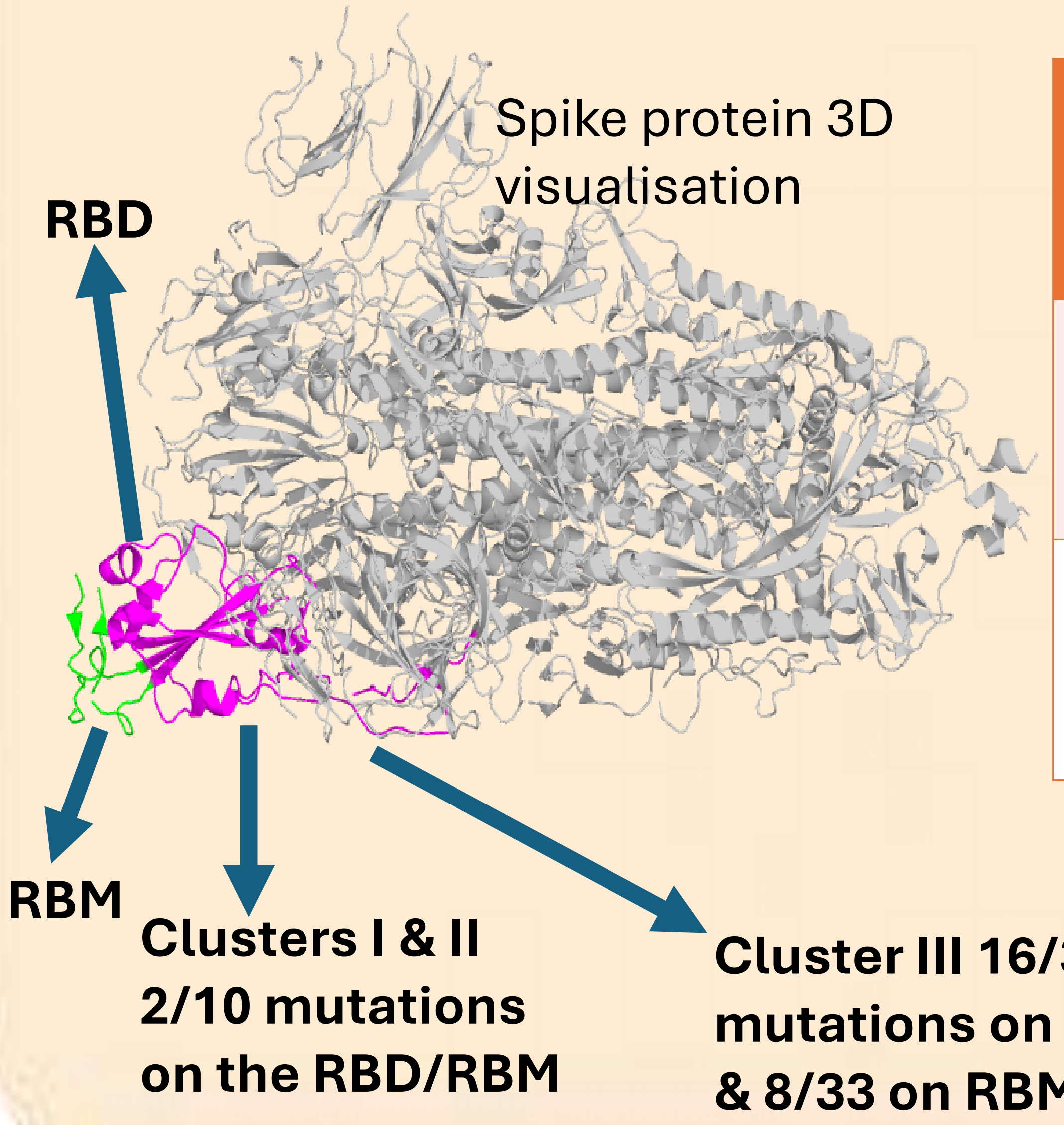
Techniques	Reduction time	Clustering time (seconds)	Efficiency increase compared to K-means alone
Principal Component Analysis (PCA)	1 hour 28 minutes	69.98	x4
t-distributed stochastic neighbour embedding (t-SNE)	**4 minutes**	**3.60**	**X1000**
Uniform Manifold Approximation and Projection (UMAP)	37 minutes	36	X9

2. 2D cluster visualisations:



3 clusters from elbow method

3. Mutation profiling & lineages:



Spike protein 3D visualisation
RBD
RBM
Clusters I & II 2/10 mutations on the RBD/RBM
Cluster III 16/33 mutations on RBD & 8/33 on RBM

Cluster	Key mutations	Infectivity	Lineage
I & II	L452R, P681R	Moderate	Delta
III	N501Y, D614G, E484A	High	Omicron

4. Temporal & General trends:

- 15,213 unique mutations
- Sequences accumulated mutations over time, 3 (2020) to 65 (2024) → Muller's ratchet[7]
- Strong positive correlation, $\rho=0.895$, $p \ll 0.001$

Conclusions and future work

- **t-SNE paired with K-means most efficient & most clear cluster visualisation**
- **3 major cluster groups, I & II likely share a common ancestor**
- **Cluster III highest proportion of RBD/RBM mutations, potentially the most infectious**
- **Cluster & mutation profiling can guide future vaccine development**
- **t-SNE assisted K-means clustering has the potential method to deal with large-scale SARS-CoV-2 datasets**

Acknowledgements & References

(1) Estimated cumulative excess deaths during COVID-19 [Internet]. Our World in Data. [cited 2025 Feb 4]. Available from: https://ourworldindata.org/grapher/excess-deaths-cumulative-economist-single-entity?focus=~Confirmed+deaths

(2) COVID-19 cases [Internet]. datadot. [cited 2025 Feb 21]. Available from: https://data.who.int/dashboards/covid19/cases?n=c

(3) Chen J, Wang R, Wang M, Wei G-W. Mutations strengthened SARS-CoV-2 infectivity. J Mol Biol [Internet]. 2020 Sep 4;432(19):5212–26. Available from: http://dx.doi.org/10.1016/j.jmb.2020.07.009

(4) McCallum M, Walls AC, Bowen JE, Corti D, Veesler D. Structure-guided covalent stabilization of coronavirus spike glycoprotein trimers in the closed conformation. Nat Struct Mol Biol [Internet]. 2020 Oct 4 [cited 2025 Feb 23];27(10):942–9. Available from: https://www.nature.com/articles/s41594-020-0483-8?fromPaywallRec=false

(5) (5) Wang R, Chen J, Gao K, Hozumi Y, Yin C, Wei G-W. Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants. Commun Biol [Internet]. 2021 Feb 15 [cited 2024 Oct 27];4(1):228. Available from: https://www.nature.com/articles/s42003-021-01754-6

(6) Liu Y, Liu J, Plante KS, Plante JA, Xie X, Zhang X, et al. The N501Y spike substitution enhances SARS-CoV-2 infection and transmission. Nature [Internet]. 2022 Feb;602(7896):294–9. Available from: http://dx.doi.org/10.1038/s41586-021-04245-0

(7) Metzger JJ, Eule S. Distribution of the fittest individuals and the rate of Muller's ratchet in a model with overlapping generations. PLoS Comput Biol [Internet]. 2013 Nov 7;9(11):e1003303. Available from: http://dx.doi.org/10.1371/journal.pcbi.1003303