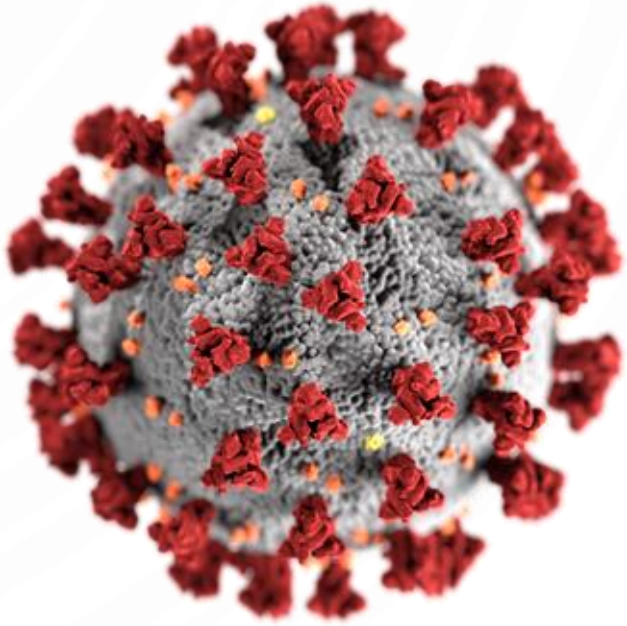
PREDICTING FUTURE MUTATIONS OF COVID

CBIO004
Keith Torpey
East Airport International School
Accra, Ghana



Credit: Wikimedia Commons

BACKGROUND

The Covid-19 pandemic is one that has had its effects seen all over the world, with over 390 million reported cases and over 5 million deaths.

News | Coronavirus pandemic

Thailand reports daily record of more than 20,000 COVID-19 cases

Government has hinted strict curbs could be extended until the end of August amid surging cases.

Credit: Aljazeera

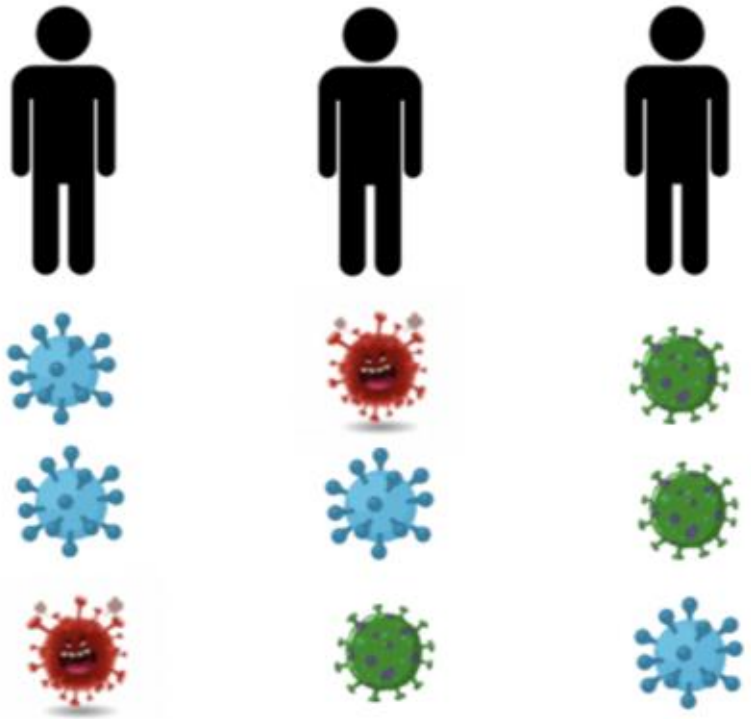
Delta variant, a warning the COVID-19 virus is getting 'fitter and faster'

Credit: UN News

South African variant may 'break through' Pfizer vaccine protection, but vaccine highly effective, Israeli study says

Credit: Reuters.com

Intra-host Mutations



- There are many mutations which can exist within a person.
- We usually take the “most common” variants within each individual in the population -- the “inter-host” sequences.
- But variants which appear within individuals but do not appear in the “inter-host” sequences may eventually be seen there!

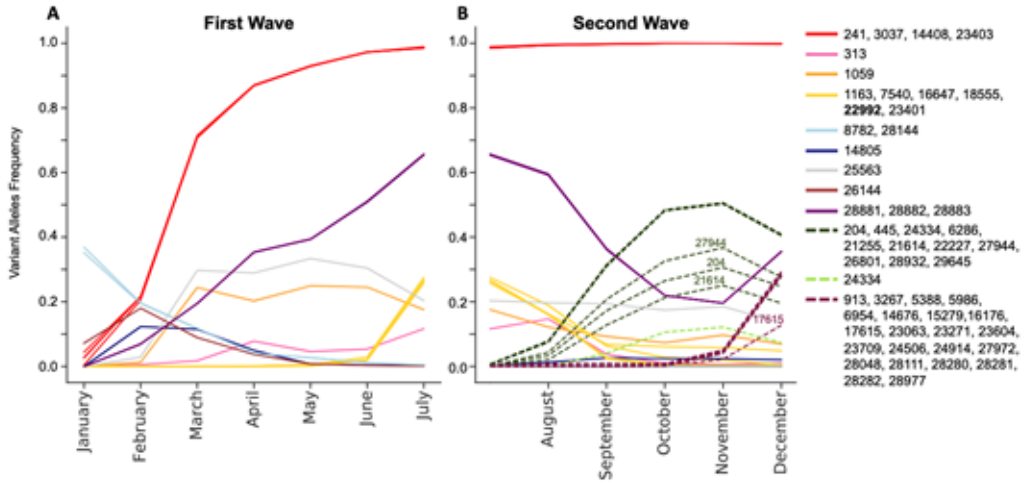
Image provided by mentor, Matthew Scicluna

Variant Allele Frequency (VAF)

- The Variant Allele Frequency can be used to quantify the mutations seen in the intra-host sequences.
- It is the number of Variant Reads divided by the number of total reads.
- Our project uses the VAF data for each base in the SARS-CoV-2 RNA sequence to explore the question - Can VAF data predict mutaions.
- We hypothesize VAF data can be used to predict future mutations.

Methods

- We used the difference in allele frequency between the first and second waves of Covid as our targets.
- The following formula was used *Abs_(Base in {A,C,U,G}(|rate(base, first wave) – rate(base, second wave)|))*



- We then set thresholds and binarized our targets.

Figure credit: Fatima Mostefai
Image provided by mentor

Neural Network

- We used Google Colab and PyTorch for the project.
- We trained a neural network to predict whether a base position would mutate or not, using the VAF data.
- Binary Cross Entropy was used as a loss function
- We used the Adam optimizer
- The VAF and target data were split into training, validation and test sets. (validation to check for overfitting)

Training

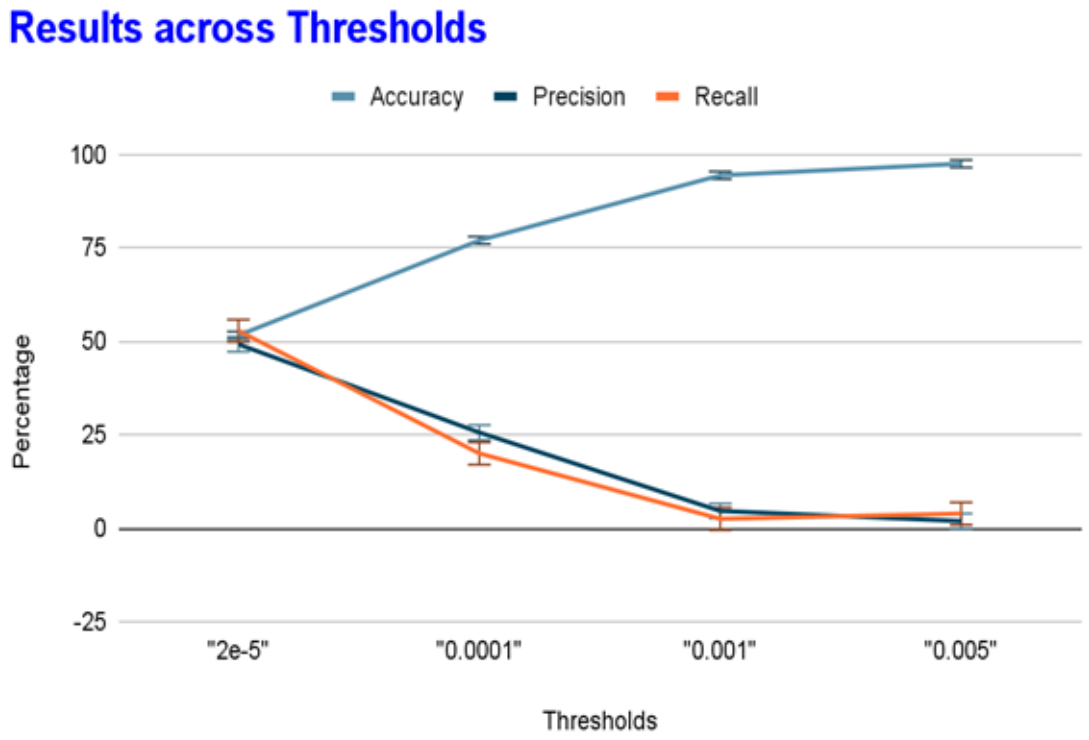
Experimental Set-Up

- We trained the network with 4 different thresholds for what is sufficient mutation.
- The network was trained by taking the VAF in as an input and comparing it with the targets for each base position
- Each threshold was trained 5 times with different seeds, and the mean values were calculated.

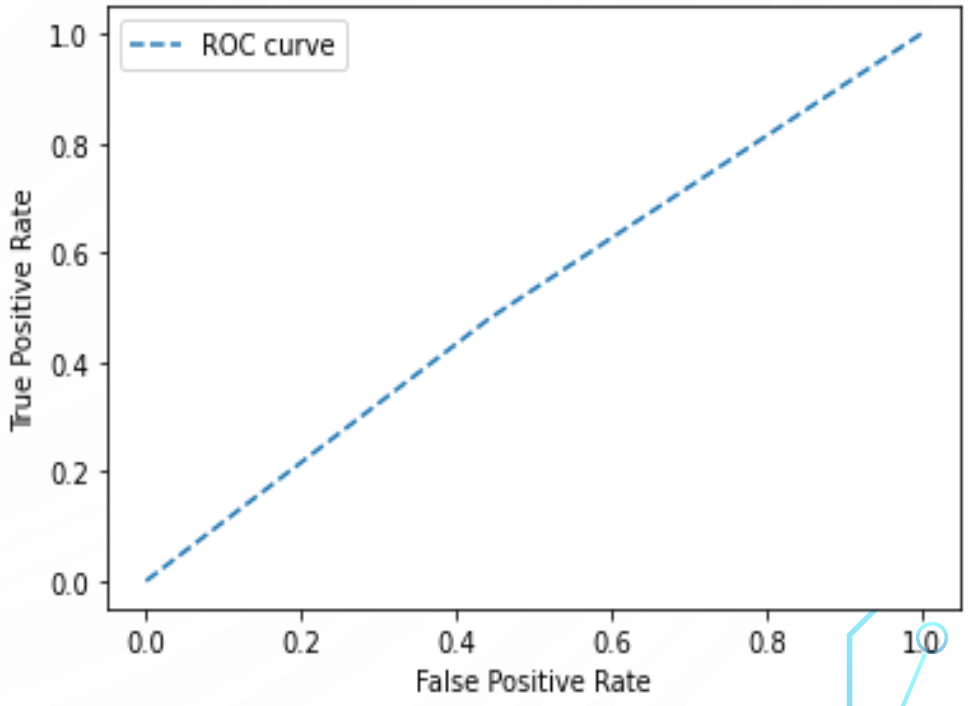
If we can predict the future mutations, then we have evidence towards our hypothesis.

Results

Accuracy, Precision, Recall:



Example ROC Curve



Images from Keith Torpey

Discussion

The results of the tests were not significant towards the research hypothesis.

There was a low precision and recall of the system, and the ROC curve lay close to the diagonal, with an AUC value near 0.5.

This means that the model was not effectively predicting the cases

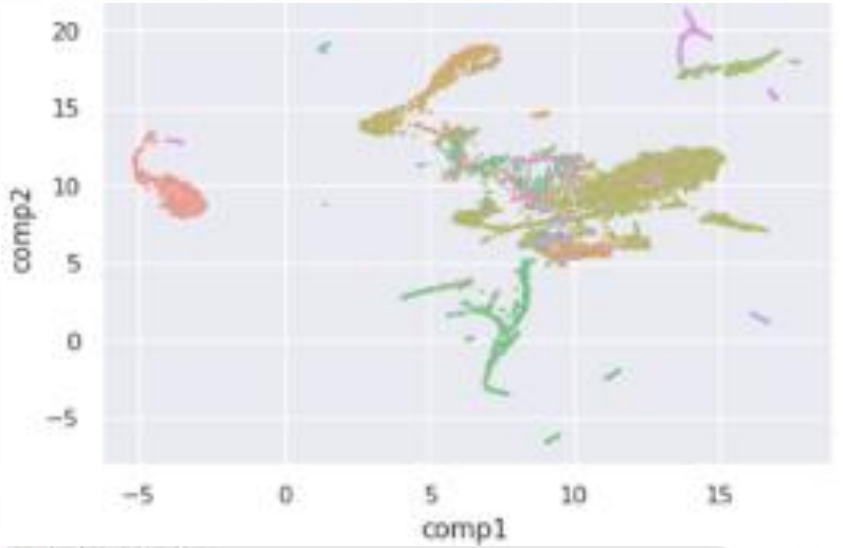
The accuracy was high when the threshold was high, but this was only due to target imbalance.

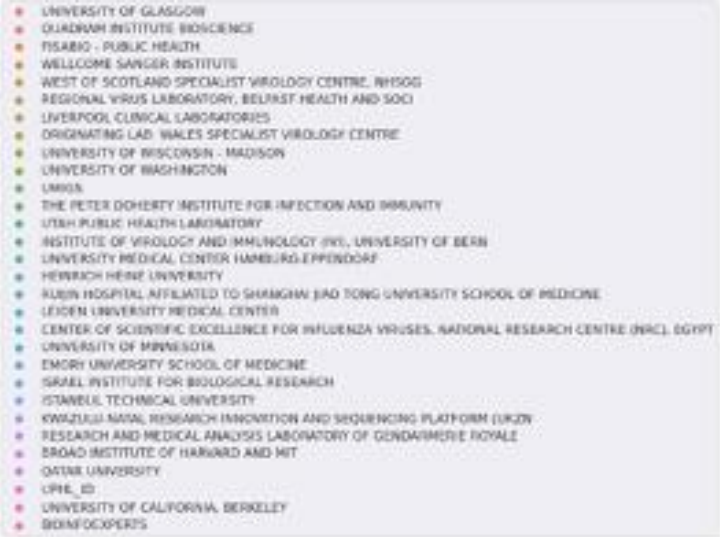
Limitations

One limitation is the batch effects found in VAFs below 5%, shown by the UMAP

There may be sequencing errors, based on the sequencing center

Another limitation could be that the Research hypothesis is false





Images from Keith Torpey

Conclusion

We trained a neural network with the aim of testing the hypothesis that VAFs could be used to predict future mutations of Covid.

The network could predict with high accuracy under some thresholds but did not give evidence that it predicted positive cases accurately

The results did not give much support to the hypothesis, but further improvements and tests could be made to explore the claim.

References

1. Worldometer, Coronavirus Cases, https://www.worldometers.info/coronavirus/,
2. Johns Hopkins Medicine, New Variants of Coronavirus: What you should know, https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/a-new-strain-of-coronavirus-what-you-should-know,
3. Center for Disease control and Prevention, What is Genomic Surveillance? https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-surveillance.html
4. Wiley Online Library, The prognostic impact of variant allele frequency (VAF) in TP53 mutant patients with MDS: A systematic review and meta-analysis, https://onlinelibrary.wiley.com/doi/10.1111/ejh.13483
5. IBM, What are Neural Networks? https://www.ibm.com/cloud/learn/neural-networks
6. David R. Kelley, Jasper Snoek, and John L. Rinn, Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4937568/pdf/990.pdf
7. ArXiv.org, PyTorch: An Imperative Style, High-Performance Deep Learning Library, https://arxiv.org/abs/1912.01703, 3/12/2019
8. Center for Disease control and Prevention, About Variants of the Virus that Causes COVID-19, https://www.cdc.gov/coronavirus/2019-ncov/variants/variant.html
9. Leland McInnes, John Healy, James Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.” *ArXiv.org*, 18 Sept. 2020, arxiv.org/abs/1802.03426.

Mutations of the Virus

- Mutations often occur in the sequence of the virus that causes Covid-19, creating nev
- New variants can cause increased vaccine resistance,

  For example- Early data shows that the Oxford-AstraZeneca Covid-19 vaccine provi protection from the B.1.351 coronavirus variant.

- Due to this, there is a need for genomic surveillance to predict mutations