Name: Anuja Rajendra Godbole

**Title of the project: Characterization of variants in the spike protein of SARS-CoV-2**

**Introduction:**

The ongoing COVID-19 pandemic is caused by a strain of coronavirus called the Severe Acute Respiratory Syndrome Coronavirus-2 (SARS‑CoV‑2) (Magazine et al., 2022). Over the course of the pandemic since the year 2020, the virus has evolved in the human host through different mutations, especially in the spike surface glycoprotein. The SARS-CoV2 spike glycoprotein (S) is the major surface antigen in the virus that aids in binding to the host receptor ACE2 and in mediating membrane fusion for cell entry (Magazine et al., 2022).

Various studies have also demonstrated that the mutations in the spike protein provides enhanced fitness to the evolving lineages of SARS-CoV-2. One of the best examples is seen through the D614G mutation which is observed in over 95% of the SARS-CoV2 lineages at present. D614G mutation was first observed in the year 2020 and spread though the population rapidly (Harvey et al., 2021). The current variant of SAR-CoV-2 omicron contains more than 34 mutations (30 nonsynonymous mutations, 3 deletions, and 1 insertion) in the spike protein compared to the wild type Wuhan strain (*The Latest Coronavirus Variants – Spike Protein Mutants*, n.d.). Therefore, the aim of this study is to understand the mutations in the spike protein in different lineages of SARS-CoV-2 that aid in fitness and evolution of the genomes over time.

**Methods and Data Collection:**

A total of ten SARs-CoV2 isolates from four different lineages (alpha, beta, gamma, omicron, delta) was collected from NCBI. The SRA numbers for the isolates are:

*Alpha lineage* – SRR22671413, DRR413450

*Beta lineage* – DRR413453, DRR413452

*Gamma lineage* – DRR413456, DRR413455

*Omicron* – SRR24146462, SRR24146475

*Delta* – SRR24146443, SRR24146443

The reference genome used was NC\_045512 which is the original Wuhan-Hu-1 strain. The raw paired end reads were downloaded from NCBI and uploaded onto the UTHealth server. All the computation was performed on the UTHealth student server (MobaXterm), Rstudio and ubuntu.

*Data Analysis Pipeline:*

Diagram

Description automatically generated

#The reads are named as follows:

alpha1\_ILMN\_R1.fastq.gz beta1\_ILMN\_R1.fastq.gz gamma1\_ILMN\_R1.fastq.gz delta1\_ILMN\_R1.fastq.gz alpha1\_ILMN\_R2.fastq.gz beta1\_ILMN\_R2.fastq.gz gamma1\_ILMN\_R2.fastq.gz delta1\_ILMN\_R2.fastq.gz alpha2\_ILMN\_R1.fastq.gz beta2\_ILMN\_R1.fastq.gz gamma2\_ILMN\_R1.fastq.gz delta2\_ILMN\_R1.fastq.gz alpha2\_ILMN\_R2.fastq.gz beta2\_ILMN\_R2.fastq.gz gamma2\_ILMN\_R2.fastq.gz delta2\_ILMN\_R2.fastq.gz omicron1\_ILMN\_R1.fastq.gz omicron1\_ILMN\_R2.fastq.gz omicron2\_ILMN\_R1.fastq.gz omicron2\_ILMN\_R2.fastq.gz

#The adapter sequences from the files were removed using the tool fastp (performed in ubuntu):

for i in \*\_R1.fastq.gz;

do

temp1=$(echo ${i} | sed "s/\_ILMN\\_R1\.fastq.gz//")

file1=${temp1}\_ILMN\_R1.fastq.gz

file2=${temp1}\_ILMN\_R2.fastq.gz

fastp -i $file1 -I $file2 -o ${temp1%\_\*}\_ILMN\_R1.fastq.gz -O ${temp1%\_\*}\_ILMN\_R2.fastq.gz -q 30

done

Text

Description automatically generated

After this step, I uploaded the reads to the UTHealth server (MobaXterm) for the following step:

#bwa mem was used to align the reads to the reference sequence

for i in \*\_R1.fastq.gz;

do

temp1=$(echo ${i} | sed "s/\_ILMN\\_R1\.fastq.gz//")

file1=${temp1}\_ILMN\_R1.fastq.gz

file2=${temp1}\_ILMN\_R2.fastq.gz

bwa mem NC\_045512.fasta $file1 $file2 > ${file1%%\_\*}.sam

done

Text

Description automatically generated

#Converting sam files to bam format, sorting and indexing bam files

for files in \*.sam; do samtools -S -b $files > ${files%.\*}.bam ; done

for files in \*.bam; do samtools sort $files > ${files%.\*}\_sorted.bam; done

for files in \*\_sorted.bam ; do samtools index $files ; done







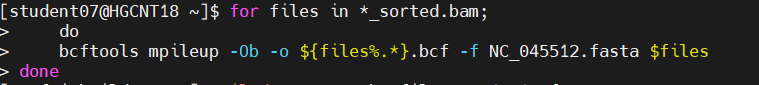
#Creating the bcf and vcf files

for files in \*\_sorted.bam;

do

bcftools mpileup -Ob -o ${files%.\*}.bcf -f NC\_045512.fasta $files

done



#Creating vcf files from bcf files

for files in \*.bcf;

do

bcftools call -vmO z -o ${files%.\*}.vcf.gz $files

done

Text

Description automatically generated

#Subset spike region from vcf files. Create an input.tab file with the spike protein regions.

for files in \*.vcf.gz; do bcftools index ; done

for files in \*.vcf.gz ; do bcftools view -R input.tab $files -o ${files%%\_\*}\_final.vcf.gz ; done

for files in \*\_final.vcf.gz; do bcftools index $files; done

#Unzip the vcf files and annotation

gunzip \*\_final.vcf.gz

A screenshot of a computer

Description automatically generated

**Using the R Package called VariantAnnotator to annotate the variants using a sorted reference gff file.**

#Import packages

pacman::p\_load(

VariantAnnotation,

genbankr,

tidyverse,

GenomicFeatures,

randomcoloR,

gtools,

GenomeInfoDb

)

#Import vcf files

alpha1 <- readVcf("alpha1\_final.vcf")

alpha2 <- readVcf("alpha2\_final.vcf")

beta1 <- readVcf("beta1\_final.vcf")

beta2 <- readVcf("beta2\_final.vcf")

gamma1 <- readVcf("gamma1\_final.vcf")

gamma2 <- readVcf("gamma2\_final.vcf")

delta1 <- readVcf("delta1\_final.vcf")

delta2 <- readVcf("delta2\_final.vcf")

omicron1 <- readVcf("omicron1\_final.vcf")

omicron2 <- readVcf("omicron2\_final.vcf")

#Import reference sorted gff file NC\_045512 and fasta file

txdb = makeTxDbFromGFF(file="ref\_NC\_045512.gff")

fa = open(FaFile("ref\_NC\_045512.fasta"))

#Needs seqlevels to be equal

seqlevels(alpha1) <- seqlevels(txdb)

seqlevels(alpha2) <- seqlevels(txdb)

seqlevels(beta1) <- seqlevels(txdb)

seqlevels(beta2) <- seqlevels(txdb)

seqlevels(gamma1) <- seqlevels(txdb)

seqlevels(gamma2) <- seqlevels(txdb)

seqlevels(delta1) <- seqlevels(txdb)

seqlevels(delta2) <- seqlevels(txdb)

seqlevels(omicron1) <- seqlevels(txdb)

seqlevels(omicron2) <- seqlevels(txdb)

#Predict coding changes for each file as shown below

coding\_alpha1 = predictCoding(alpha1, txdb, fa)

coding\_alpha2 = predictCoding(alpha2, txdb, fa)

coding\_beta1 = predictCoding(beta1, txdb, fa)

coding\_beta2 = predictCoding(beta2, txdb, fa)

coding\_gamma1 = predictCoding(gamma1, txdb, fa)

coding\_gamma2 = predictCoding(gamma2, txdb, fa)

coding\_delta1 = predictCoding(delta1, txdb, fa)

coding\_delta2 = predictCoding(delta2, txdb, fa)

coding\_omicron1 = predictCoding(omicron1, txdb, fa)

coding\_omicron2 = predictCoding(omicron2, txdb, fa)

#converting to a dataframe

alpha1 <- as.data.frame(coding\_alpha1)

alpha2 <- as.data.frame(coding\_alpha2)

beta1 <- as.data.frame(coding\_beta1)

beta2 <- as.data.frame(coding\_beta2)

delta1 <- as.data.frame (coding\_delta1)

delta2 <- as.data.frame(coding\_delta2)

gamma1 <- as.data.frame(coding\_gamma1)

gamma2 <- as.data.frame(coding\_gamma2)

omicron1 <- as.data.frame(coding\_omicron1)

omicron2 <- as.data.frame(coding\_omicron2)

#saving as csv files

write\_csv(alpha1, "alpha\_1.csv")

write\_csv(alpha2, "alpha\_2.csv")

write\_csv(beta1, "beta\_1.csv")

write\_csv(beta2, "beta\_2.csv")

write\_csv(gamma1, "gamma\_1.csv")

write\_csv(gamma2, "gamma\_2.csv")

write\_csv(delta1, "delta\_1.csv")

write\_csv(delta2, "delta\_2.csv")

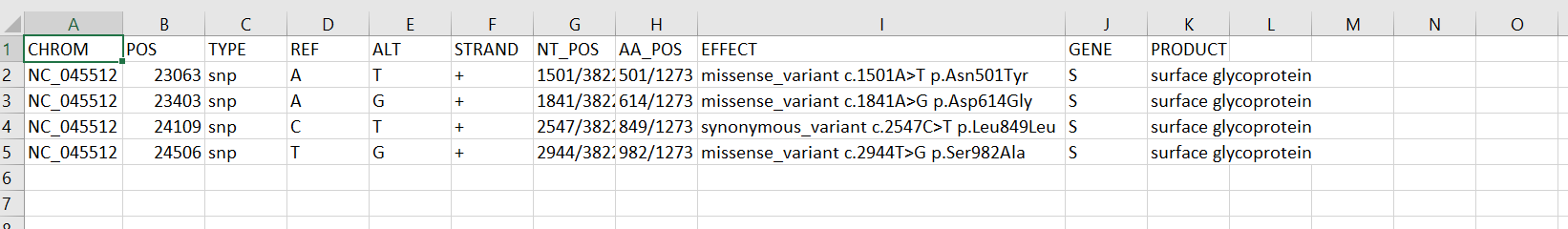
write\_csv(omicron1, "omicron\_1.csv")

write\_csv(omicron2, "omicron\_2.csv")

**Results and discussion:**

1. *Alpha lineage variants:*

Table No.1 The annotated variants in alpha lineage SARs-Cov2 are shown below:



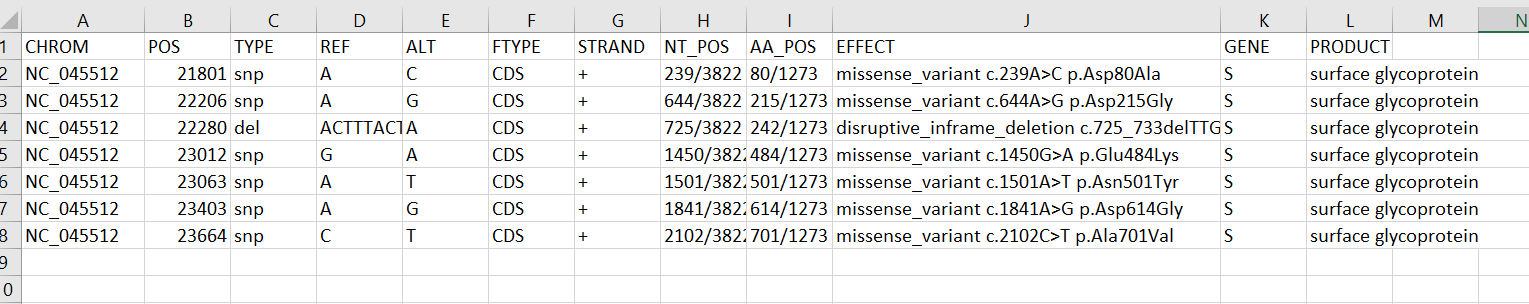
The alpha lineage was first discovered in December 2020 in UK and has found to be 50% more infectious and deadly than other lineages. As shown in the table above Asn501Tyr (N501Y) results in increased binding to ACE-2.

Graphical user interface

Description automatically generated with medium confidence

1. *Beta Lineage variants*:

Table No.2 The annotated variants in beta lineages are shown below:



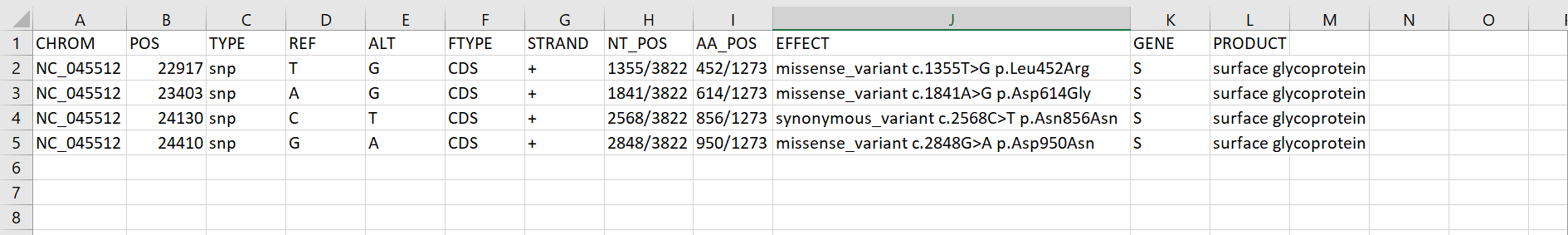
First identified in South Africa in December 2020, it has now reached at least 48 countries and was detected in the United States in January, and today has spread to at least 25 states. One of the variants of concern (VOC) observed in the beta lineages in Glu484Lys (E484K) which decreases the antibodies efficiency in the vaccine to recognize spike proteins.

Graphical user interface

Description automatically generated with medium confidence

1. *Delta Lineage variants:*

Table No.3 The annotated variants in delta lineages are shown below:



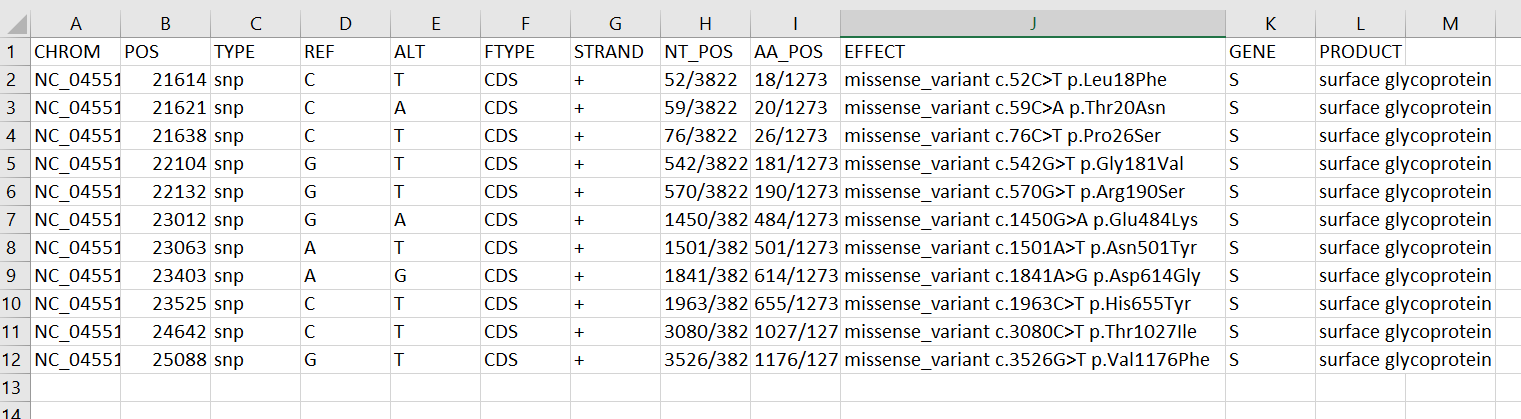
The delta variant was first identified circulating in India in October 2020. One of the defining variants of concern observed in the delta lineage is Leu452Arg (L452R) that disrupts a hydrophobic interaction on the Spike RBD surface potentially affecting antibody neutralization and ACE-2 binding.

Graphical user interface, application

Description automatically generated

4. *Gamma Lineage Variants:*

Table No.4 The annotated variants in gamma lineages are shown below:



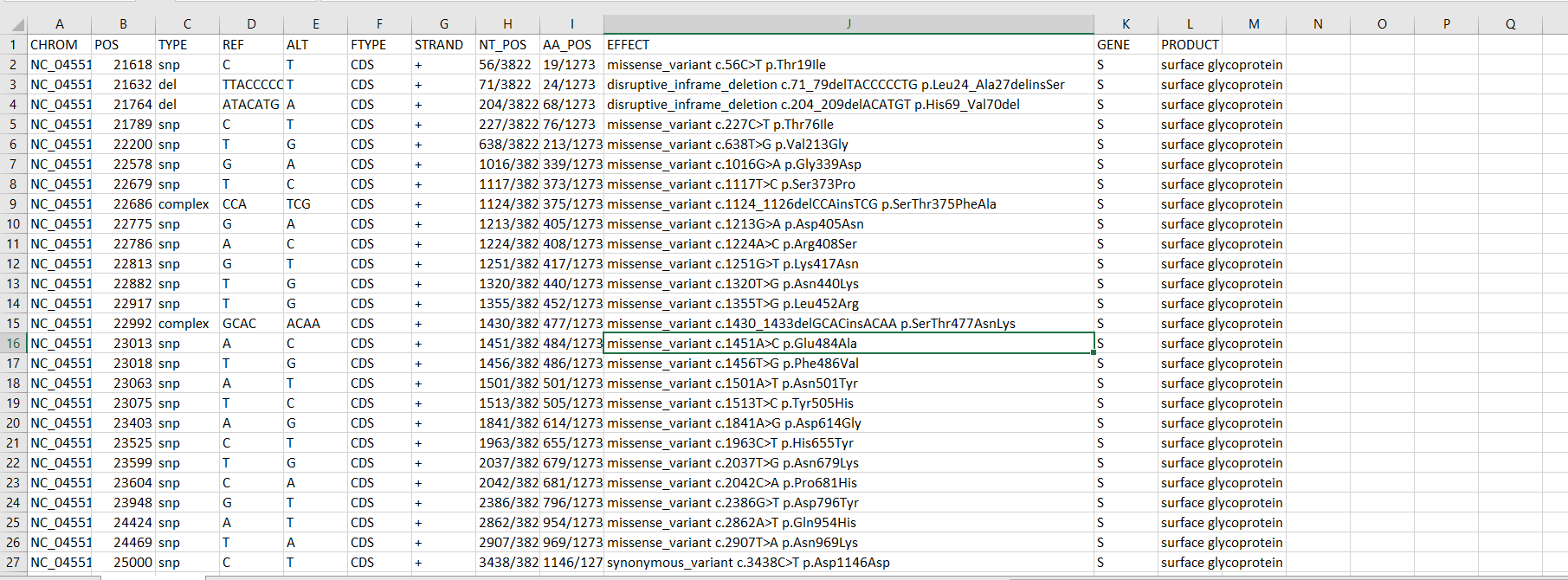
The gamma lineage was first discovered in Brazil 2020. As shown from the table above the gamma lineage has accumulated a lot of variants and mutations in the spike protein compared to the lineages described above contributing to enhanced fitness.

A picture containing application

Description automatically generated

1. *Omicron Lineage Variants:*

Table No.5 The annotated variants in the omicron lineages are shown :



WHO reported omicron variant in November 2021. It is expected to be the most dominant variant due to its high transmission rate. The omicron variant has over 30 mutations in the spike protein alone as shown above.

Timeline

Description automatically generated with medium confidence

For all the results and discussion: (*The Latest Coronavirus Variants – Spike Protein Mutants*, n.d.)

**Conclusion:**

From the results section shown above, we can see that the number of mutations in the spike protein increase significantly as we go from alpha to omicron lineages. This suggests the role of computation in identifying and visualizing the mutations in the COVID spike protein region. Additionally, it can be employed in public health research and surveillance.

**References:**

Harvey, W. T., Carabelli, A. M., Jackson, B., Gupta, R. K., Thomson, E. C., Harrison, E. M., Ludden, C., Reeve, R., Rambaut, A., Peacock, S. J., & Robertson, D. L. (2021). SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology*, *19*(7), Article 7. https://doi.org/10.1038/s41579-021-00573-0

Magazine, N., Zhang, T., Wu, Y., McGee, M. C., Veggiani, G., & Huang, W. (2022). Mutations and Evolution of the SARS-CoV-2 Spike Protein. *Viruses*, *14*(3), 640. https://doi.org/10.3390/v14030640

*The Latest Coronavirus Variants – Spike Protein Mutants*. (n.d.). Bio-Techne. Retrieved May 4, 2023, from https://www.bio-techne.com/