PROJECT
MUTATION ANALYSIS OF COVID-19 STRAINS

Aarzoo (2022008)
Akshat Saxena (2022055)
Angadjeet Singh (2022071)
Anushka Srivastava (2022086)
Apaar Garg (2022089)
Arav Amawate (2022091)
Parth Sandeep Rastogi (2022352)

PROBLEM STATEMENT

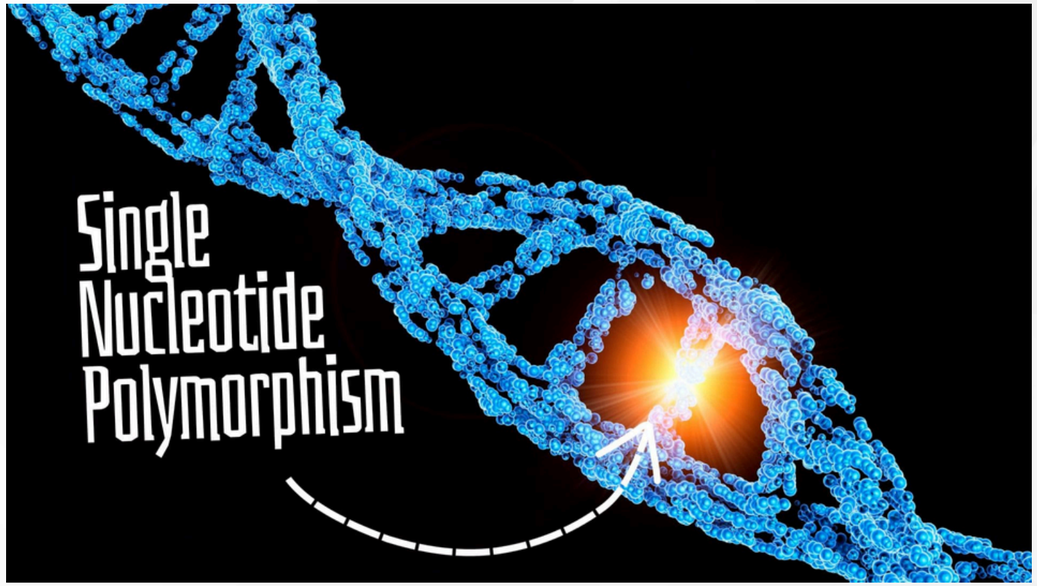
The world has faced the grim reality of the COVID-19 pandemic. India, one of the hardest-hit countries, reported about 40 million cases, around 570,000 severe cases, which constituted approximately 15% of total infections, often led to acute respiratory distress syndrome (ARDS), a severe lung condition.

Our main aim is to analyze the mutation in different COVID-19 strains and study the impact of the mutation on the severity of that particular strain.

SNP Analysis

WHY SNP?

Viruses like SARS-Cov-2 are continuously evolving because of the genetic changes caused by genetic mutations or viral recombination that occurs during the replication phase of the genome.
SARS-Cov-2 has mutated consistently over the 2-3 years span of the pandemic, resulting in genetically different variants from the original Wuhan virus. Throughout the pandemic, many different mutations of viruses were found in other parts of the world.



We are finding SNPs using the NCBI Blast tool to check the changes in the sequences of different strains with the original Wuhan virus.

SOME CHANGES

```
Sbjct  8221  ..........................................................  8280

Query  8281  CTATATGCTCACCTATAACAAAGTTGAAAACATGACACCCCGTGACCTTGGTGCTTGTAT  8340
Sbjct  8281  ..........................................................  8340

Query  8341  TGACTGTAGTGCGCGTCATATTAATGCGCAGGTAGCAAAAAGTCACAACATTGCTTTGAT  8400
Sbjct  8341  ..........................................G...............  8400

Query  8401  ATGGAACGTTAAAGATTTCATGTCATTGTCTGAACAACTACGAAAACAAATACGTAGTGC  8460
Sbjct  8401  ..........................................................  8460

Query  8461  TGCTAAAAAGAATAACTTACCTTTTAAGTTGACATGTGCAACTACTAGACAAGTTGTTAA  8520
Sbjct  8461  ..........................................................  8520

Query  8521  TGTTGTAACAACAAAGATAGCACTTAAGGGTGGTAAAATTGTTAATAATTGGTTGAAGCA  8580
Sbjct  8521  ..........................................................  8580

Query  8581  GTTAATTAAAGTTACACTTGTGTTCCTTTTTGTTGCTGCTATTTTCTATTTAATAACACC  8640
Sbjct  8581  ..........................................................  8640

Query  8641  TGTTCATGTCATGTCTAAACATACTGACTTTTCAAGTGAAATCATAGGATACAAGGCTAT  8700
Sbjct  8641  ..........................................................  8700

Query  8701  TGATGGTGGTGTCACTCGTGACATAGCATCTACAGATACTTGTTTTGCTAACAAACATGC  8760
Sbjct  8701  ..........................................................  8760

Query  8761  TGATTTTGACACATGGTTTAGCCAGCGTGGTGGTAGTTATACTAATGACAAAGCTTGCCC  8820
Sbjct  8761  ..........................................................  8820

Query  8821  ATTGATTGCTGCAGTCATAACAAGAGAAGTGGGTTTTGTCGTGCCTGGTTTGCCTGGCAC  8880
Sbjct  8821  ..........................................................  8880

Query  8881  GATATTACGCACAACTAATGGTGACTTTTTGCATTTCTTACCTAGAGTTTTTAGTGCAGT  8940
Sbjct  8881  ..........................................................  8940

Query  8941  TGGTAACATCTGTTACACACCATCAAAACTTATAGAGTACACTGACTTTGCAACATCAGC  9000
Sbjct  8941  ..........................................A...............  9000

Query  9001  TTGTGTTTTGGCTGCTGAATGTACAATTTTTAAAGATGCTTCTGGTAAGCCAGTACCATA  9060
Sbjct  9001  ..........................................................  9060
```

NC_045512.2 VS MT019529.1

```
Sbjct  3754  ..........................................................  3813

Query  3841  AATGAAGAGTGAAAAGCAAGTTGAACAAAAGATCGCTGAGATTCCTAAAGAGGAAGTTAA  3900
Sbjct  3814  ..........................................................  3873

Query  3901  GCCATTTATAACTGAAAGTAAACCTTCAGTTGAACAGAGAAAACAAGATGATAAGAAAAT  3960
Sbjct  3874  ..........................................................  3933

Query  3961  CAAAGCTTGTGTTGAAGAAGTTACAACAACTCTGGAAGAAACTAAGTTCCTCACAGAAAA  4020
Sbjct  3934  ..........................................................  3993

Query  4021  CTTGTTACTTTATATTGACATTAATGGCAATCTTCATCCAGATTCTGCCACTCTTGTTAG  4080
Sbjct  3994  ..........................................................  4053

Query  4081  TGACATTGACATCACTTTCTTAAAGAAAGATGCTCCATATATAGTGGGTGATGTTGTTCA  4140
Sbjct  4054  ..........................................................  4113

Query  4141  AGAGGGTGTTTTAACTGCTGTGGTTATACCTACTAAAAAGGCTGGTGGCACTACTGAAAT  4200
Sbjct  4114  ......................................T...................  4173

Query  4201  GCTAGCGAAAGCTTTGAGAAAAGTGCCAACAGACAATTATATAACCACTTACCCGGGTCA  4260
Sbjct  4174  ..........................................................  4233

Query  4261  GGGTTTAAATGGTTACACTGTAGAGGAGGCAAAGACAGTGCTTAAAAAGTGTAAAAGTGC  4320
Sbjct  4234  ..........................................................  4293

Query  4321  CTTTTACATTCTACCATCTATTATCTCTAATGAGAAGCAAGAAATTCTTGGAACTGTTTC  4380
Sbjct  4294  ..........................................................  4353

Query  4381  TTGGAATTTGCGAGAAATGCTTGCACATGCAGAAGAAACACGCAAATTAATGCCTGTCTG  4440
Sbjct  4354  ..........................................................  4413

Query  4441  TGTGGAAACTAAAGCCATAGTTTCAACTATACAGCGTAAATATAAGGGTATTAAAATACA  4500
Sbjct  4414  ..........................................................  4473

Query  4501  AGAGGGTGTGGTTGATTATGGTGCTAGATTTTACTTTTACACCAGTAAAACAACTGTAGC  4560
Sbjct  4474  ..........................................................  4533

Query  4561  GTCACTTATCAACACACTTAACGATCTAAATGAAACTCTTGTTACAATGCCACTTGGCTA  4620
Sbjct  4534  ..........................................................  4593

Query  4621  TGTAACACATGGCTTAAATTTGGAAGAAGCTGCTCGGTATATGAGATCTCTCAAAGTGCC  4680
Sbjct  4594  ..........................................................  4653

Query  4681  AGCTACAGTTTCTGTTTCTTCACCTGATGCTGTTACAGCGTATAATGGTTATCTTACTTC  4740
Sbjct  4654  ..........................................................  4713
```

NC_045512.2 VS OK091006.1

INFERENCE

- We have analyzed nearly ten strains from different continents with the original Wuhan strain and identified different positions of SNPs.
- Different positions of SNP in each strain infer that the virus has evolved in different mutations, and thus, there is a significant variation in the severity of the virus.
- Certain mutations may confer advantages such as increased infectivity or immune evasion, leading to higher transmission rates or more severe disease outcomes.
- Analysis of SNP positions can reveal patterns of viral spread and circulation across different geographic regions. By tracing the movement of specific strains based on SNP profiles, global spread of the virus can be tracked.

Constant mutations in the SARS-Cov-2 over the course of pandemic resulting in different variants

Gene nucleotide mutations may or may not alter the amino acid residue (aa) sequences of translated proteins. In case of altered protein sequence, the results can be a substitution, deletion, insertion, protein truncation, or shift of open reading frame. Of the SARS-CoV-2 variants analyzed here, non-synonymous nucleotide mutations mostly led to point mutations (substitution, deletion, or insertion). Occasionally, nonsense mutations (causing protein truncations) or shifts of open reading frames were observed, with the latter resulting in either aa substitutions followed by protein truncation or nonstop protein extension beyond the normal stop codon. Below, point mutations of SARS-CoV-2 proteins are presented first. Results on rarer truncations and frameshift nonstop mutations follow toward the end of the section. Figure 3 shows a few examples of the substitution, deletion, or insertion point mutations in the S, N, and NSP6 proteins across the SARS-CoV-2 variants

Source Link

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10222255/

PLOTTING THE MUTATIONS

Github

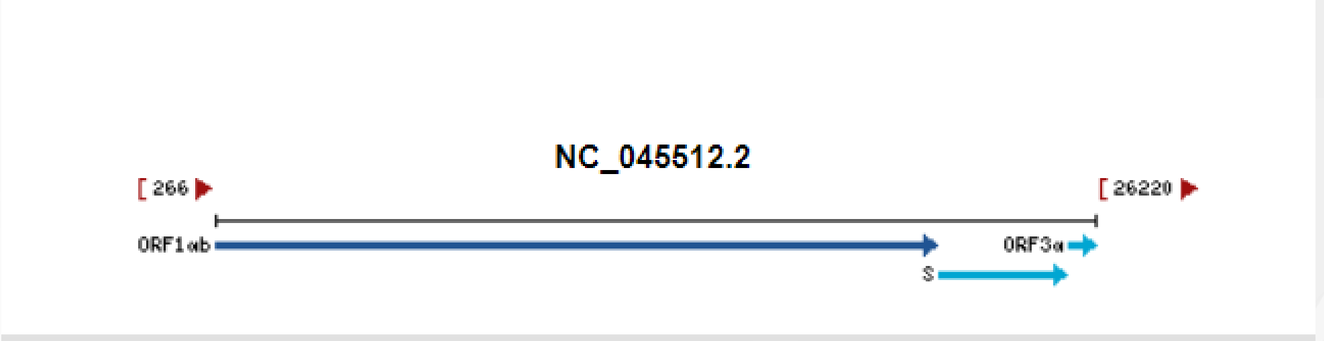
https://github.com/parthrastogicoder/PbProject

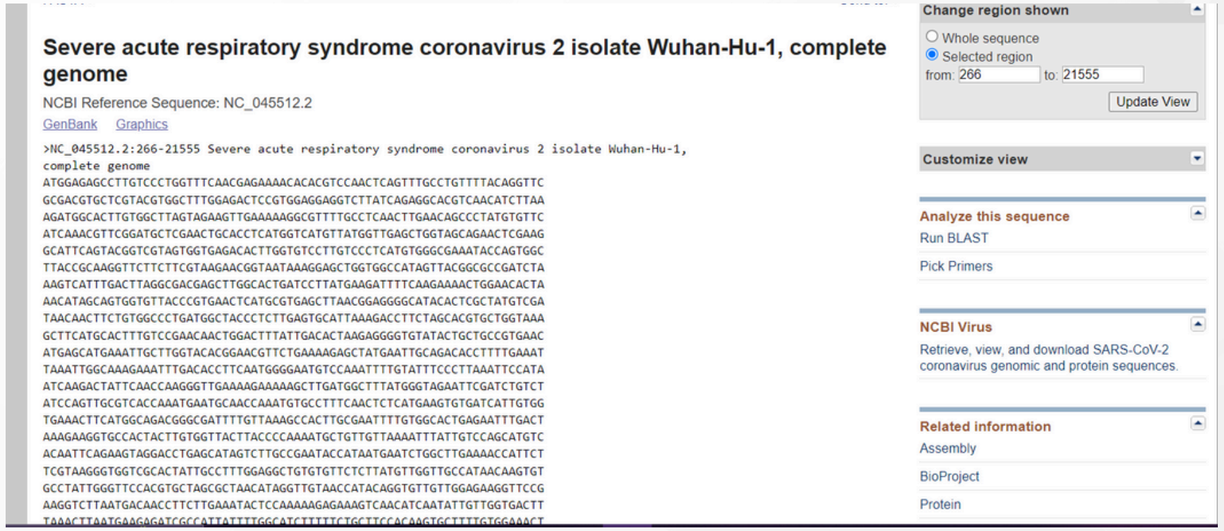
INITIAL IDEA

COVID-19 has 11 main genes: ORF1ab, S, ORF3a, E, M, ORF6. ORF7a. ORF7b, ORF8, N and ORF10. ORF1ab comprises about 75% of the genome while other key genes takes up less than 22%.

We wanted to compare the mutations of the genome of the COVID-19 strains found in China with other countries. We have chosen India, the USA, Great Britain, South Africa and South Korea for our study.

Using this, we also find out the gene protein which mutates the most, which helps in drug development.





WHAT WE HAVE DONE

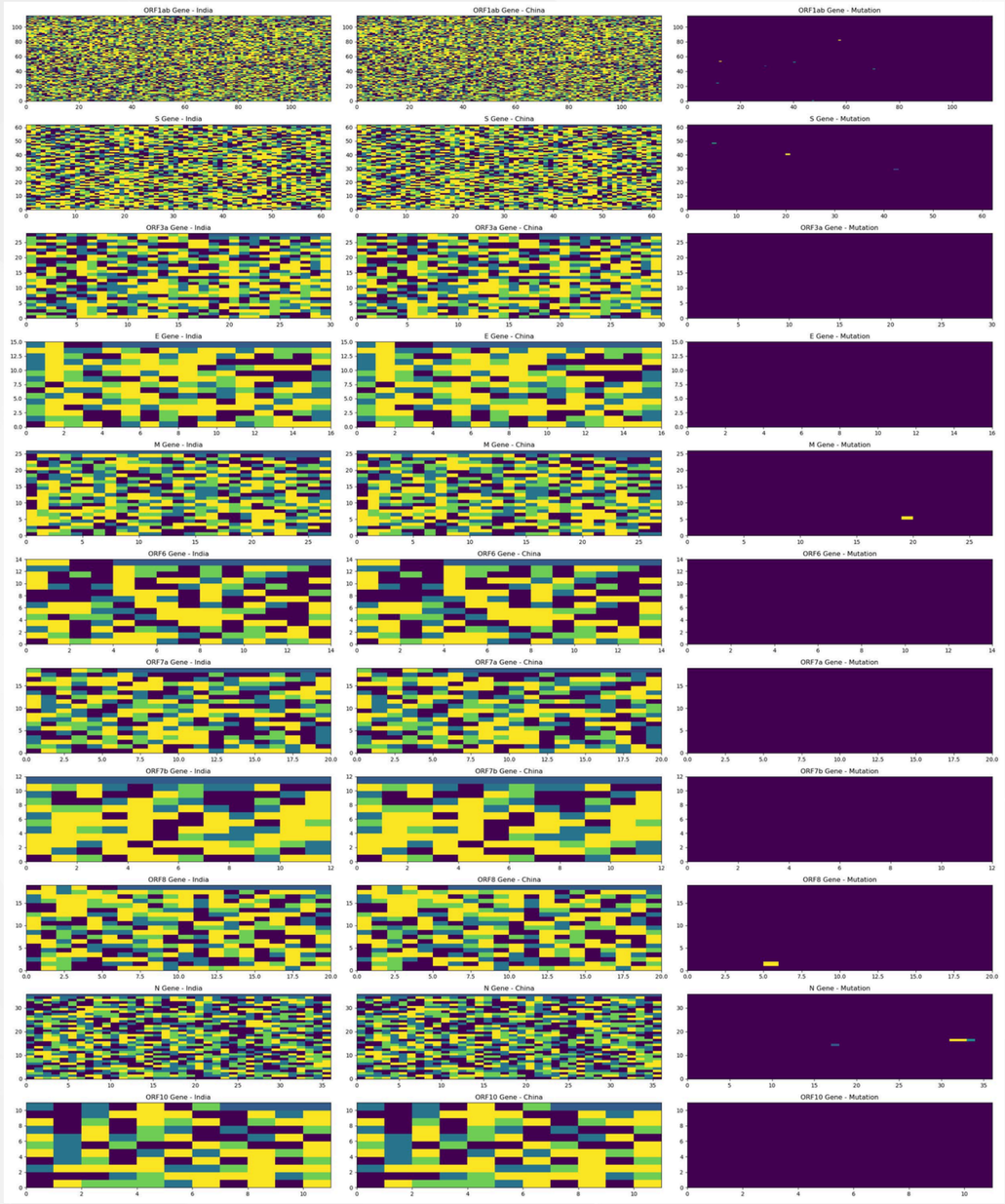
- Using the NCBI database, we found the range in the sequence where the target genes lie.
- These genes were converted into Numpy arrays by assigning arbitrary values to them.
- The Numpy arrays were reshaped so that we could graph the values.
- The difference between the country's Numpy array and China's Numpy array was calculated. Any non-zero value in the resultant array indicated a difference in the nucleotide sequence, showing the mutation.
- All Numpy arrays are plotted.

```
for G in gene_name:
    gene_country = dna(country[G])
    gene_country.transcription()
    numpfy_country = gene_country.numpfy()
    # Reshaping the numpy array for graphing purposes
    numpfy_country = numpfy_country.reshape(numpy_image_dict[G][0])
    ax[row][col].pcolor(numpfy_country)
    ax[row][col].set_title(G + f' Gene - {country_names[country_name]}')
    col+=1
    gene_china = dna(China[G])
    gene_china.transcription()
    numpfy_china = gene_china.numpfy()
    numpfy_china = numpfy_china.reshape(numpy_image_dict[G][0])
    ax[row][col].pcolor(numpfy_china)
    ax[row][col].set_title(G +' Gene - China')
    col+=1
    # Finding difference in both arrays to find mutation
    mut = numpfy_china - numpfy_country
    if mut.any():
        # Non-zero value in numpy array = mutation in the sequence
        mut_nec = np.nonzero(mut)
        x=mut_nec[0]
        y=mut_nec[1]
        l=0
        for i in x:
            country_base = base_codes[numpfy_country[i][y[l]]]
            ch_base = base_codes[numpfy_china[i][y[l]]]
            count_genes[G] += 1
            print("Mutated DNA Base {} in China and Base {} in {} at position {} For the Gene {}".format(ch_base,coun
            l+= 1
    ax[row][col].pcolor(mut)
    ax[row][col].set_title(G + ' Gene - Mutation')
    row+= 1
    col=0
```

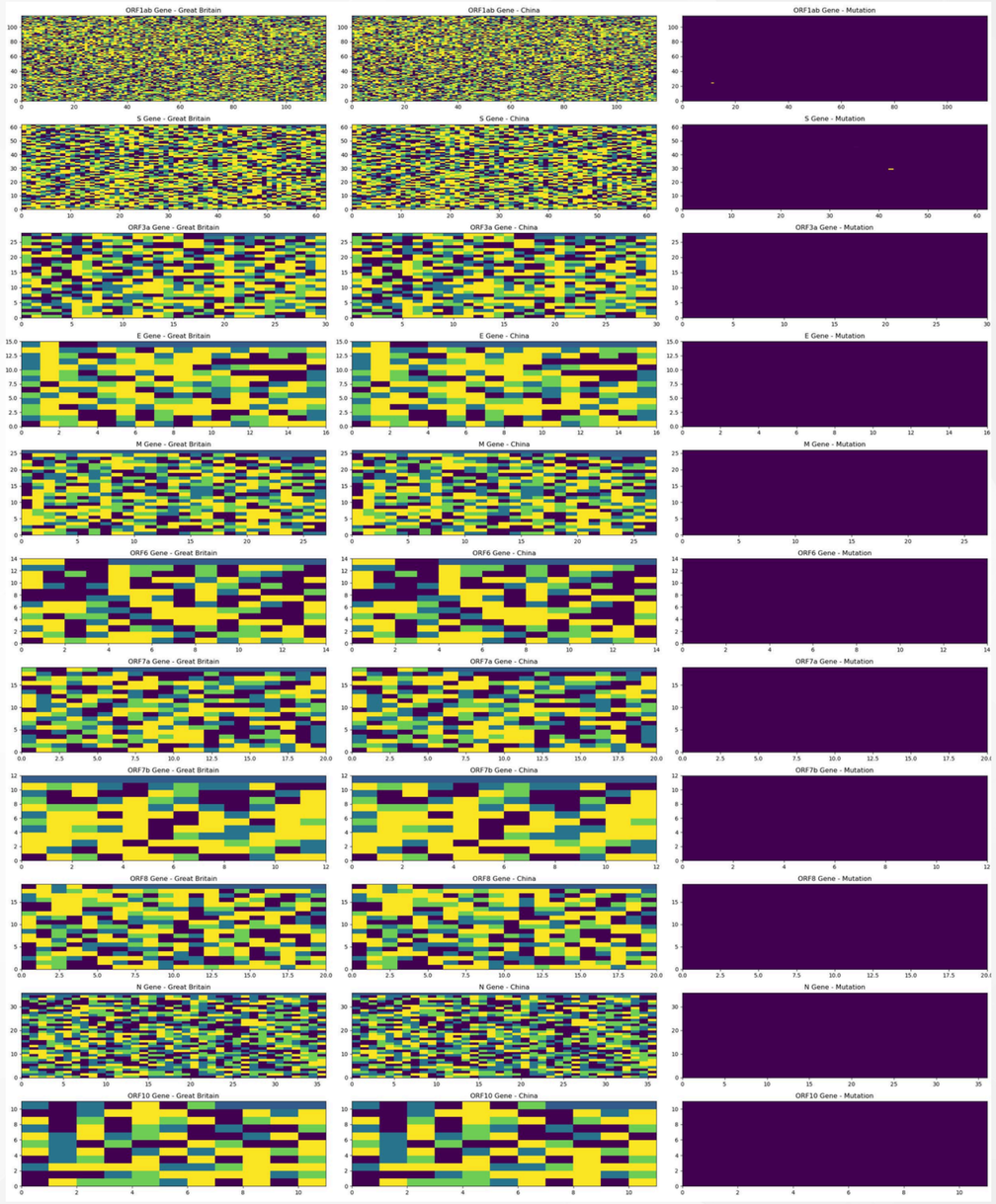
SOURCE

https://github.com/tonygeorge1984/Python-Sars-Cov-2-Mutation-Analysis

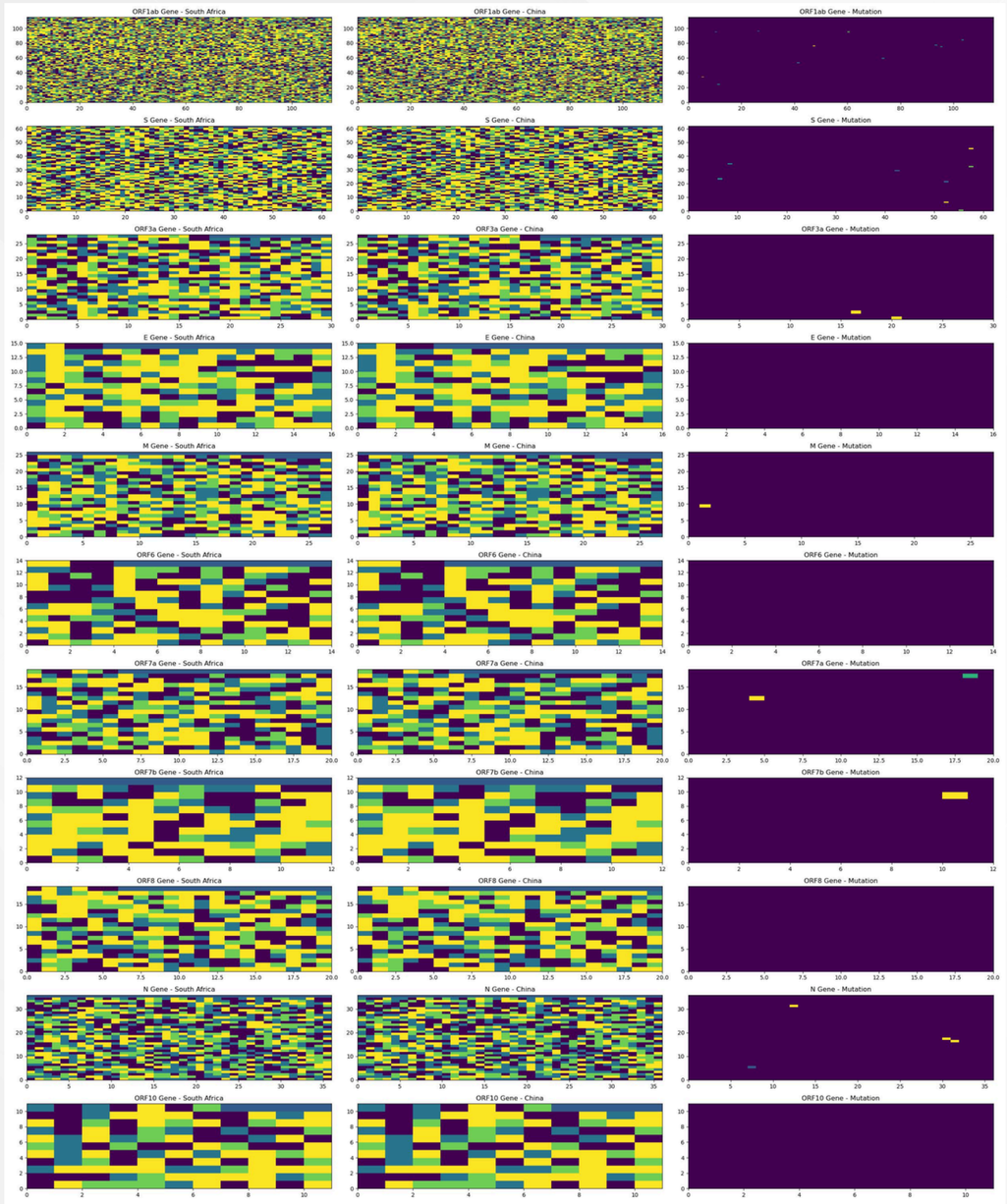
INDIA VS CHINA



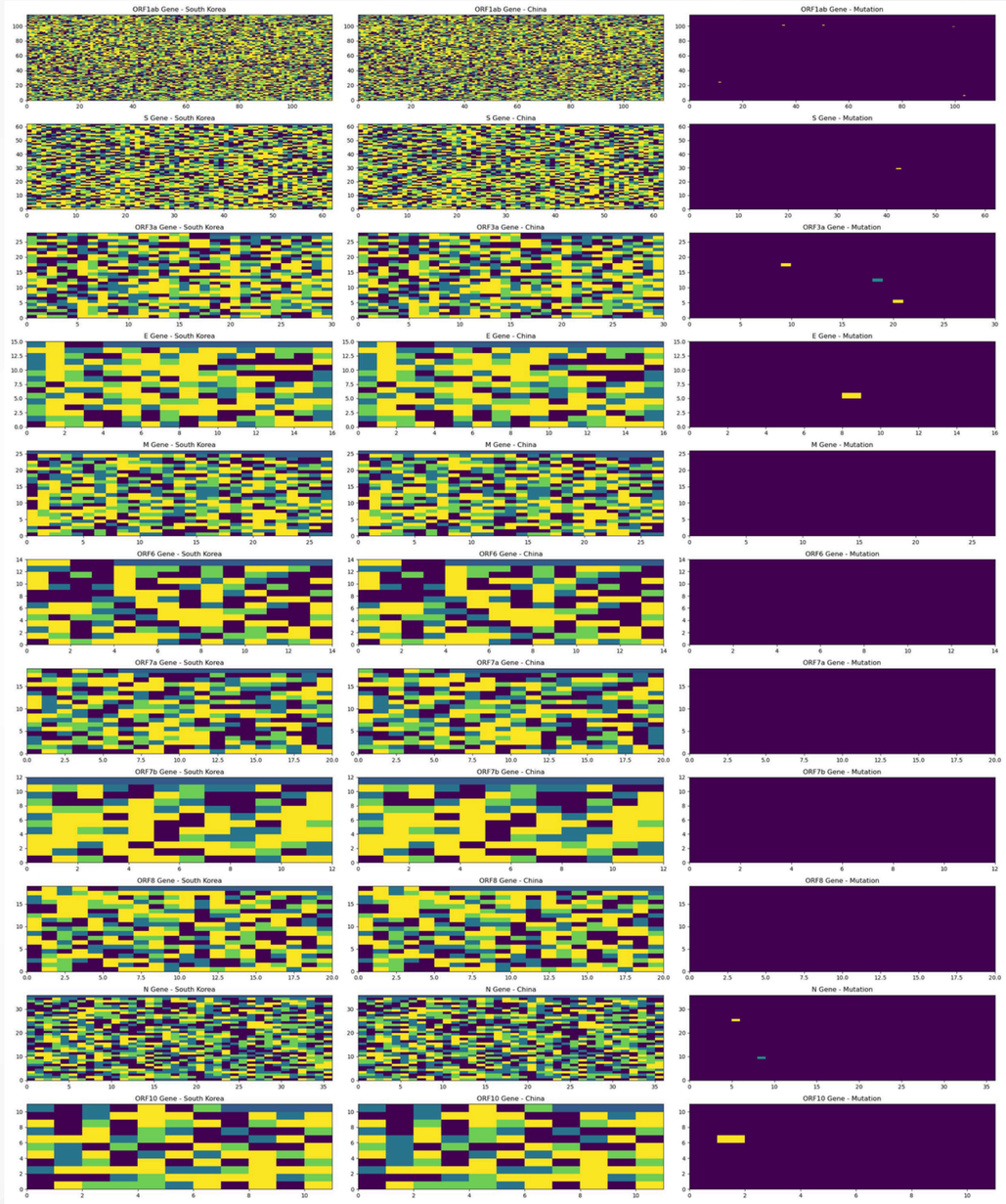
GREAT BRITAIN VS CHINA



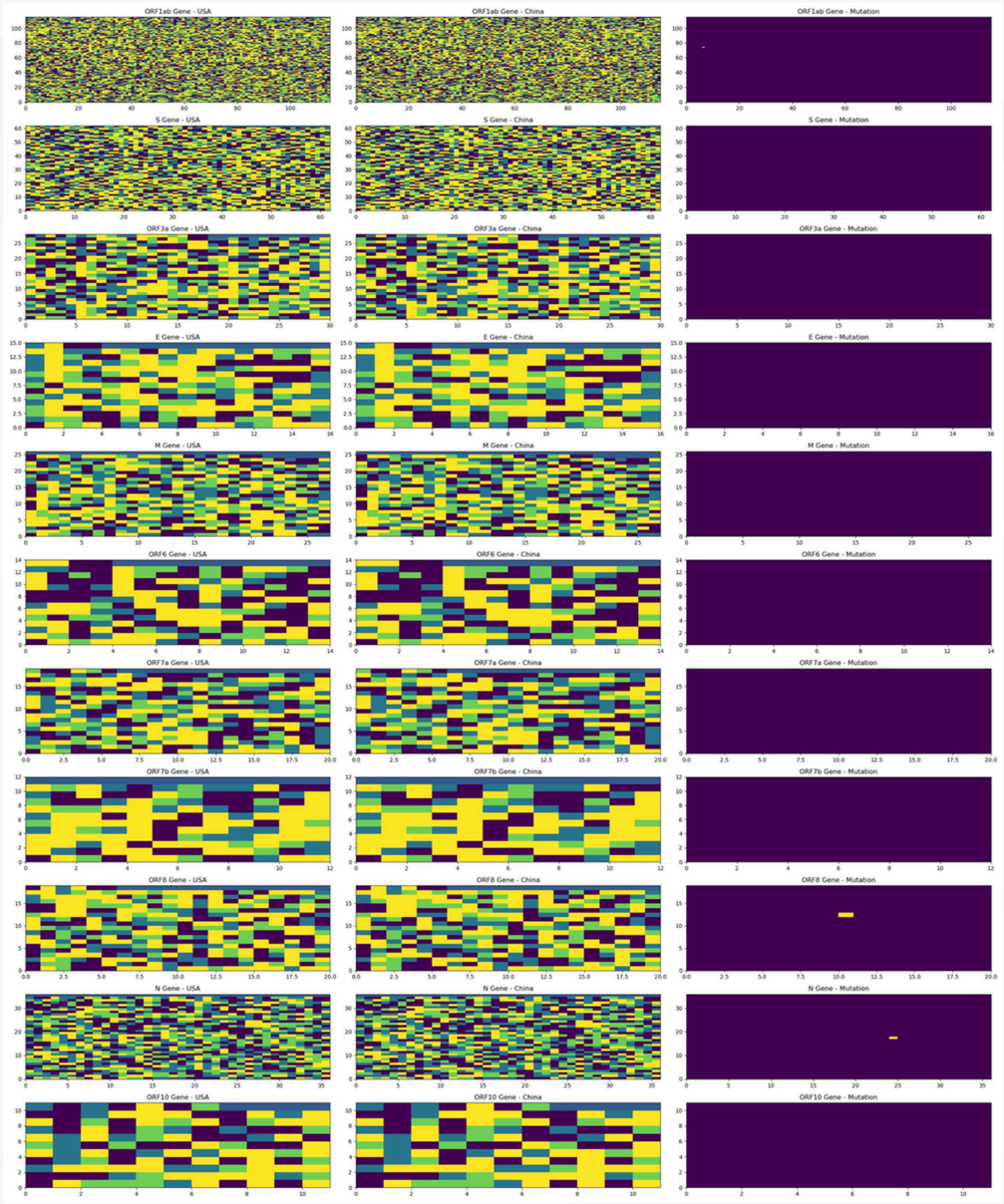
SOUTH AFRICA VS CHINA



SOUTH KOREA VS CHINA



USA VS CHINA

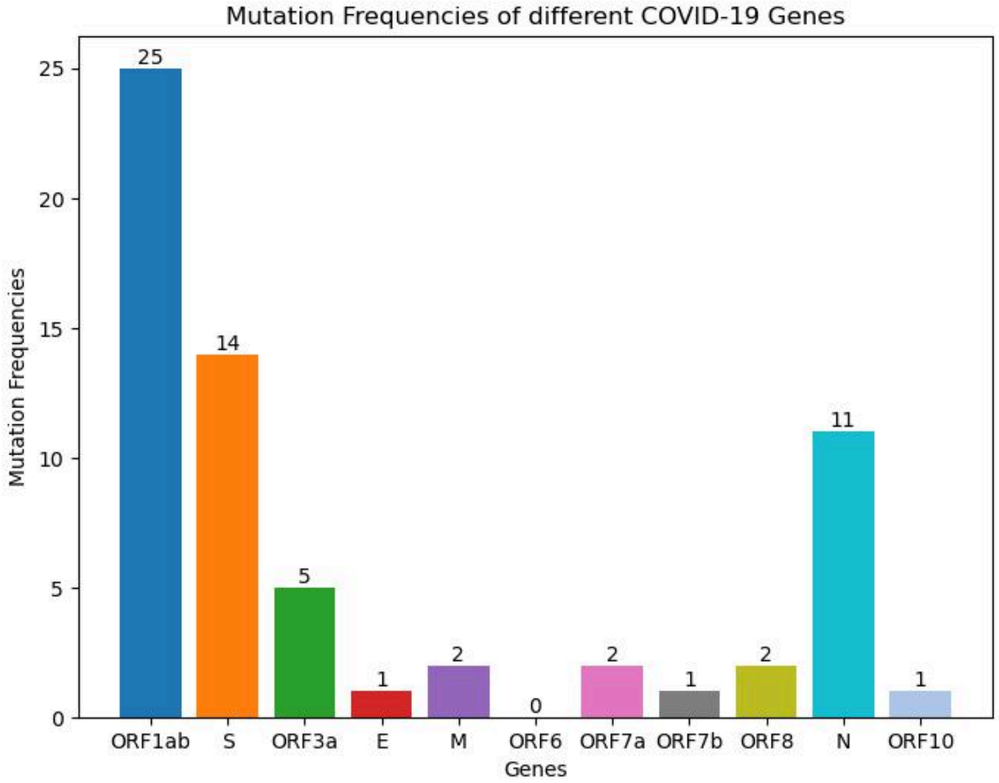

ORF1ab Gene - USA
ORF1ab Gene - China
ORF1ab Gene - Mutation
S Gene - USA
S Gene - China
S Gene - Mutation
ORF3a Gene - USA
ORF3a Gene - China
ORF3a Gene - Mutation
E Gene - USA
E Gene - China
E Gene - Mutation
M Gene - USA
M Gene - China
M Gene - Mutation
ORF6 Gene - USA
ORF6 Gene - China
ORF6 Gene - Mutation
ORF7a Gene - USA
ORF7a Gene - China
ORF7a Gene - Mutation
ORF7b Gene - USA
ORF7b Gene - China
ORF7b Gene - Mutation
ORF8 Gene - USA
ORF8 Gene - China
ORF8 Gene - Mutation
N Gene - USA
N Gene - China
N Gene - Mutation
ORF10 Gene - USA
ORF10 Gene - China
ORF10 Gene - Mutation

WHAT WE FOUND OUT

Our study found that maximum mutations occur in the ORF1ab protein, and the most conserved protein is the ORF6 protein.

So, according to this, we can see that ORF1ab is a bad candidate protein for drug discovery as it mutates frequently. Meanwhile, ORF6 and other more conserved proteins like ORF7b and ORF10 are a better candidate.

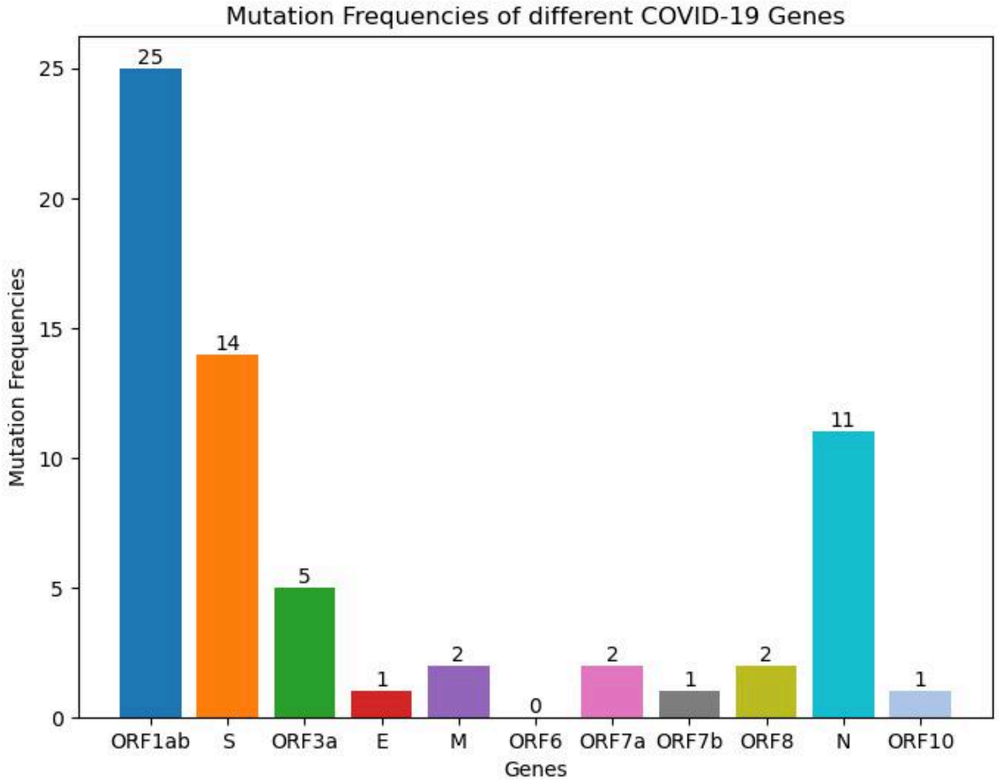
In the further slides, we can see the relation between different COVID-19 strains in the given countries in a phylogenetic tree.



WHAT WE FOUND OUT

However, since our dataset is small, this is not a reliable result. Similar methods can be used to find results for larger datasets. Finding conserved proteins in a genome sequence while analyzing the mutations help us find good candidates for drug discovery and treatment of COVID-19.

Mutations in the S protein are particularly important because the S protein is key for the first step of viral transmission.



Different genes show different levels of mutations. Analysing this is important for drug development.

trends as those of the gene nucleotides. The S, N, NSP6, and ORF8 proteins showed relatively high mean aa mut% (counting internal substitutions, deletions, and insertions) of 0.909% (SD, 0.683%), 0.739% (SD, 0.355%), 0.738% (SD, 0.472%), and 0.716% (SD, 0.995%; n = 187 clones of 13 variants), respectively. The ORF3a, E, and ORF7a proteins had intermediate mean aa mut% ranging from 0.278% to 0.371%. On the other hand, the ORF1a and ORF1ab polyproteins, M, ORF6, ORF7b, ORF10, and all other NSP proteins (except NSP6) exhibited no or relatively low mean aa mut% ranging from 0.000% to 0.226% (Figure 4a). Single-factor Anova and Student's *t*-tests revealed significantly higher mean aa mut% in the S, ORF8, N, and NSP6 proteins, as compared with those of the ORF1a, ORF1ab, M, ORF6, ORF7a, ORF7b, ORF3a, E, and all other NSP proteins ($p < 0.01$) except NSP6. It should be noted that owing to the small sizes of the NSP11, E, ORF6, ORF7b, and ORF10 proteins (13, 75, 61, 43, and 38 aa residues, respectively), their mean aa mut% might statistically still be subject to significant random deviations.

Source Link

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10222255/

Different genes show different levels of mutations. Analysing this is important for drug development.

There are also discrepancies between the present and previous data. In contrast to the low-level mutations in ORF1ab shown here, the viral gene/polyprotein has previously been reported as one of the most mutative in SARS-CoV-2 isolates from India [25]. Some disagreements could be due to using different data collection methods, analysis, or quantification. The current study analyzed only the variants of concern/interest that are of more clinical relevance. In contrast, many previous reports used genome sequencing data encompassing hundreds of variants from diverse geographic origins [5]. Other analyses considered only specific variants identified from specific regions [25,26].

Source Link

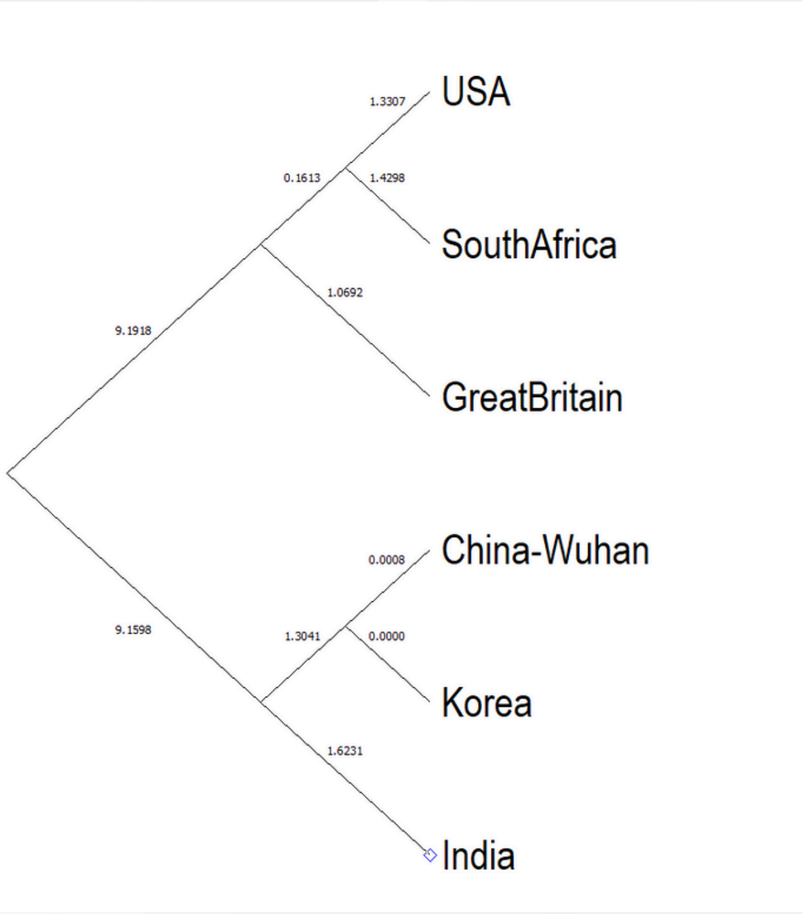
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10222255/

PHYLOGENETIC ANALYSIS

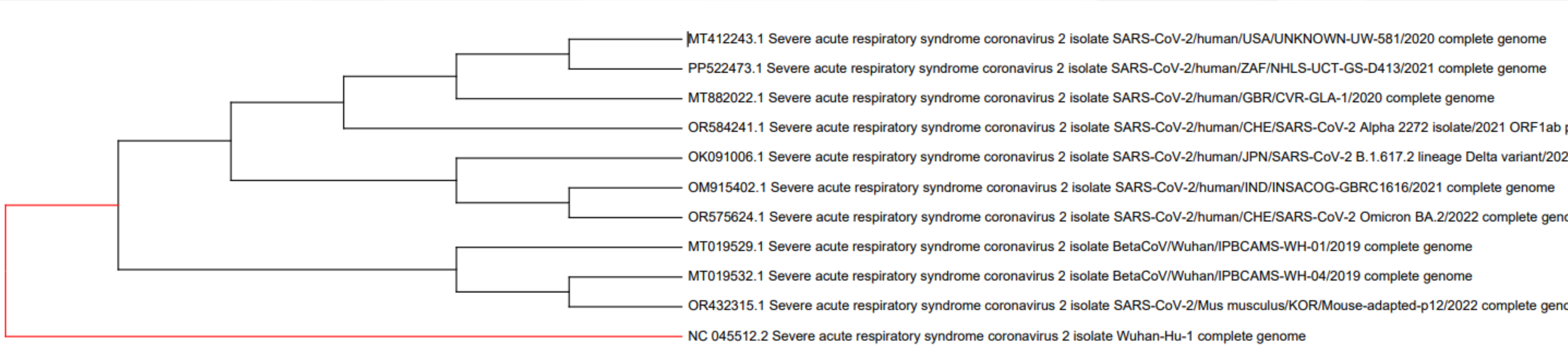
COUNTRY-WISE ANALYSIS

- Tool We Used :-Mega 11.10.0
- Steps:- Firstly found SNP in all sequence combined then constructed Phylogenetic Tree.

The evolutionary analysis of SARS-CoV-2 reveals the ancestral strain from China-Wuhan as the root of the tree. Branching off from this root, the USA strain appears first, possibly due to early flights, followed by India and Korea. This suggests a closer genetic link between the Wuhan strain and those from Korea and India, compared to strains from the USA, South Africa, and Great Britain. Moreover, a closer relation between the USA and South Africa hints at possible transmission routes influenced by geographical proximity. These findings underscore the intricate evolutionary paths of the virus, offering insights into its global spread and transmission dynamics.



STRAIN-WISE
ANALYSIS



STRAIN-WISE INFERENCE

- It is important to note that this is just a tiny sample of SARS-CoV-2 strains, and there are many other strains circulating around the world. The relationships between these strains can change over time as the virus continues to evolve.
- Phylogenetic trees are a useful tool for understanding the spread of viruses, but they should not be interpreted as definitive maps of viral transmission. Other factors, such as travel patterns, can also play a role in the spread of viruses.
- Example of the 3 Wuhan strains from given phylogenetic tree Clearly visible that Wuhan is the source of the spread.
- Finding the source of infection is one of the most important tasks in the prevention and treatment of viruses, and tracing the virus genome is an important method to find the source and path of infection.

REFERENCES

Research Papers

https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2022.851323/full

https://www.sciencedirect.com/science/article/pii/S2452014421000492

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10222255/

https:/www.journalofinfection.com/article/S0163-4453(20)30159-6/fulltext

https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html

Github References

https://github.com/tonygeorge1984/Python-Sars-Cov-2-Mutation-Analysis

https:/github.com/ai-covariants/analysis-mutations

CONTRIBUTIONS

SNP ANALYSIS AND DATASET IDENTIFICATION : APAAR , ANGADJEET SINGH

MUTATIONS ANALYSIS AND GRAPH PLOTTING : ANUSHKA SRIVASTAVA , AARZOO

PHYLOGENETIC ANALYSIS AND RESEARCH WORK : PARTH SANDEEP RASTOGI , ARAV AMAWATE

PRESENTATION : AKSHAT SAXENA

THANK YOU