Mutation Detection And Analysis in Closely Related Genomes

Ansh Bajpai

Department of Computer Science, Artificial Intelligence Engineering, Amrita Vishwa Vidyapeetham, Kollam, 690525, Kerala, India.

Contributing authors: am.sc.u4aie23108@am.students.amrita.edu;

Abstract

This study presents a comparative genomic analysis of SARS-CoV-2 and its variants using multiple sequence alignment and pairwise assessment techniques. Genome sequences of key variants were aligned using the MUSCLE tool, with subsequent analyses based on the resulting alignment. A confusion matrix and heatmap assessed similarity between the reference genome (MN908947.3-Wuhan) and variants from India, Netherlands, and Japan. Clustermap and phylogenetic analyses revealed hierarchical relationships among the strains.

Genome coverage and abundance were visualized through graphical plots to evaluate sequencing distribution. Detailed mutation analysis involved identifying single nucleotide polymorphisms (SNPs), insertions, and deletions by comparing each variant with the reference. These were organized into structured dataframes using Python and mapped to corresponding genes (e.g., S, ORF1ab, N, M, E, and others).

To visualize mutation patterns, helicoidal genome plots displayed spatial mutation distributions, while stacked bar charts and frequency plots highlighted gene-specific mutation burdens and potential mutation hotspots.

This integrated approach offers insights into variant evolution, genomic divergence, and mutation dynamics in SARS-CoV-2.

Keywords: Multiple Sequence Alignment, Genome Variation, Mutation Analysis, SARS-CoV-2, Viral Genomics, Biopython

1 Introduction

The rapid global spread of COVID-19, caused by the SARS-CoV-2 virus, has led to the emergence of several genetic variants due to ongoing mutations. These genetic changes can modulate viral behavior, impacting its transmission efficiency, immune escape mechanisms, and pathogenicity. As a result, continuous monitoring of genomic changes has become essential, guiding public health strategies, vaccine development, and diagnostic accuracy.

Comparative analysis of viral genomes allows researchers to identify both conserved and variable regions within the viral population. Multiple sequence alignment (MSA) serves as a foundational technique for detecting these genomic differences among variants. While several studies have focused on specific mutations or protein-coding regions, a comprehensive framework that quantifies sequence similarity and mutation distribution across the entire genome remains less explored.

In this study, a structured methodology is used to analyze and compare different SARS-CoV-2 variants using sequence alignment techniques. Through visualization tools such as heatmaps, clustering diagrams, and genome-wide mutation maps, this work offers enhanced insights into the genetic similarities and differences among the variants. The analysis further includes classification and positional mapping of gene-specific mutations, providing a detailed perspective on evolutionary changes within the SARS-CoV-2 genome.

2 Methods

2.1 Datasets

We used the following publicly available datasets:

- **Genome Sequences:** Complete SARS-CoV-2 variant genomes in FASTA format were retrieved from NCBI Virus Portal. Key accessions include PV361325.1 (India), OM287553.1 (Netherlands), OK091006.1 (Japan), and MN908947.3 (reference Wuhan strain).
- **Genomic Annotations:** The GenBank file for MN908947.3 (.gb format) provided gene annotations used to map mutations to coding/non-coding regions.
- **Alignment Data:** Multiple sequence alignments generated by MUSCLE (FASTA format) were used for comparative genomics, identity analysis, and mutation detection.

2.2 Mutation Analysis and Gene Mapping

Aligned sequences were compared to the reference genome to detect SNPs, insertions, and deletions. Using Biopython, mutation coordinates were mapped to genes extracted from the GenBank feature table to assess mutation density per gene.

2.3 Coverage and Abundance

Per-base coverage matrices were computed to evaluate genome completeness. Abundance was visualized using pie charts, while depth histograms illustrated coverage variability across samples.

2.4 Similarity and Clustering

Pairwise identity matrices were derived from aligned genomes to assess conservation. Heatmaps and hierarchical clustering (via Seaborn) grouped genomes based on genetic similarity.

2.5 Visualization Tools

We generated various plots using `Matplotlib`, `Seaborn`, and `Pandas`:

- Pairwise identity heatmaps and clustering dendrograms
- Gene-wise and mutation-type distributions (bar charts)
- Coverage depth histograms
- 3D helical plots of genome-wide mutations

2.6 Ethical Statement

This study used only publicly available viral genomes and did not involve human or animal subjects; hence, ethical approval was not applicable.

3 Results, Interpretation and Future Perspectives

3.1 Genomic Conservation and Variant Divergence

To examine genetic similarity across SARS-CoV-2 variants, a pairwise identity matrix was constructed and visualized as a heatmap. As shown in Figure 2a, most variants share above 99.9% sequence identity, indicating a high degree of conservation despite some regional differences. Coverage analysis was performed to assess how completely each variant aligns with the reference genome. Figure 1a shows that most genomic positions are covered by all three variants, with a few positions showing lower alignment, possibly due to insertions or deletions. Relative abundance of aligned genomes is visualized using a pie chart (Figure 1b), where all genomes contribute almost equally, confirming balanced representation during alignment. Lastly, the overall distribution of mutation is shown in Figures 2c and 2d, highlighting the mutation spread along difference genes and locations inside the viral genomes.

3.2 Mutation Landscape Across Genes

Mutation analysis revealed deletions to be the most frequent, followed by SNPs and a smaller portion of insertions (in Table 4). High mutation density was observed in the intergenic, Spike (S), and ORF1ab regions(in Figure 3).

The helicoidal plots showed consistent gene coverage across all variants, with slight differences(in Figures **??**, **??** and **??**).

3.3 Functional Implications of Observed Mutations

Several of the identified mutations were located within the spike (S) gene, which plays a central role in viral entry into host cells via ACE2 receptors. Changes in this region—particularly non-synonymous SNPs—may impact receptor binding affinity and are often associated with increased transmissibility or immune escape potential (in Figure **??** highlighting the general mutation intensity and in Figures **??** and **??** highlighting the mutation region across the reference genome). Additionally, mutations found in accessory genes like ORF8 and ORF3a could influence viral replication efficiency and host immune modulation as shown in Figure 2d and 3.Most of the deletions were located in intergenic regions; however, a small number occurred within coding sequences and may lead to truncated or impaired proteins. These results highlight the significance of ongoing genomic monitoring to identify variants that could affect viral function and inform public health decisions.

3.4 Comparison with Existing Studies

When comparing the results with previous studies, I found a strong overlap in terms of where mutations tend to occur—especially in the spike (S) gene and ORF1ab, which are well known for their role in viral infectivity and replication. The high sequence similarity between most variants also matches what's been observed in global datasets, confirming that much of the SARS-CoV-2 genome remains conserved. That said, we did notice a few differences—like some unique mutations in intergenic regions and ORF8—that aren't commonly reported. These could be due to local variation, sampling differences, or simply underreported changes.

3.5 Limitations of the Study

While the study offers valuable insights into genomic variation across SARS-CoV-2 variants, a few limitations must be acknowledged. First, the dataset used was limited in size, which may not capture the full diversity of circulating variants globally. Second, the alignments were static and based on a snapshot in time, viral genomes continue to evolve rapidly, and ongoing updates would strengthen the findings. Lastly, our analysis focused on nucleotide-level changes and did not extend into structural or protein-level effects, which are essential to fully understand functional impacts. Future work should aim to incorporate 3D protein modeling and broader datasets for a more complete perspective.

3.6 Future Directions

To advance this research, future efforts could incorporate structural modeling techniques to assess the functional impact of specific mutations, particularly within critical regions such as the spike protein. Integrating spatiotemporal metadata would enable analysis of geographic mutation distribution and temporal evolution. Expanding the dataset to include a broader range of globally diverse variants would enhance the robustness of mutation pattern analyses. Additionally, linking genomic alterations

with clinical data may reveal associations between mutation signatures and disease severity or transmission dynamics.

4 Tables

To better understand the genomic similarity and structural consistency among different SARS-CoV-2 variants, we analyzed both the pairwise sequence identity and gene-wise segment lengths. Table 1 presents the pairwise sequence identity values among the first four genome samples considered in this study. This identity matrix helps quantify the extent of similarity at the nucleotide level across full-genome alignments.

Table 1: Sample Pairwise Sequence Identity Matrix

	PV361325.1	OM287553.1	OK091006.1	MN908947.3
PV361325.1	1.000	0.992	0.993	0.992
OM287553.1	0.992	1.000	0.993	0.993
OK091006.1	0.993	0.993	1.000	0.996
MN908947.3	0.992	0.993	0.996	1.000

Additionally, Tables 2, 3 and 4 provide an overview of gene lengths, mutation counts, and mutation types (SNPs, insertions, and deletions) across three SARS-CoV-2 variants compared to the reference genome. Among all genes, ORF1ab is the longest, spanning over 21,000 base pairs. In contrast, the spike (S) gene, though significantly shorter, exhibits a much higher mutation density. Specifically, 173 mutations were identified in the S gene, while ORF1ab showed only 126, despite being nearly five times longer.

This elevated mutation rate in the S gene may be attributed to selective pressures linked to its role in host cell entry and immune system interactions. Meanwhile, structural genes such as E, M, and ORF6 remain highly conserved across variants.

Table 2: Gene lengths for different COVID-19 genome variants

Variant ID	**orf1ab**	**S**	**ORF3a**	**E**	**M**	**ORF6**	**ORF7a**	**ORF8**	**N**	**ORF10**
PV361325.1	21290	3810	828	228	669	186	366	360	1260	117
OK091006.1	21290	3822	828	228	669	186	366	366	1260	117
OM287553.1	21278	3804	828	228	669	186	366	366	1251	117
MN908947.3	21290	3822	828	228	669	186	366	366	1260	117

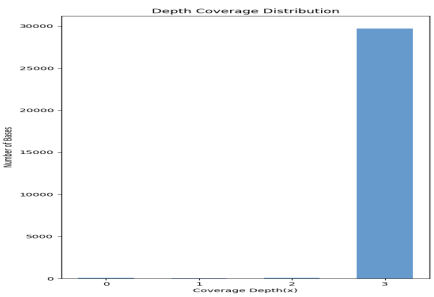
Table 3: Mutation counts per gene for each SARS-CoV-2 genome variant.

Variant_ID	E	M	N	ORF3a	ORF6	ORF7a	ORF8	S	Intergenic	orf1ab	Total
OK091006.1	0	1	5	1	0	1	0	14	72	48	142
OM287553.1	1	3	14	1	1	0	0	57	132	25	234
PV361325.1	0	1	4	1	0	1	3	102	110	53	275

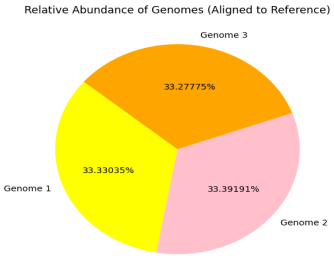
Table 4: Mutation types per gene

Gene	SNPs	Deletions	Insertions
Spike (S)	134	30	9
ORF1ab	114	12	0
ORF3a	3	0	0
Envelope (E)	1	0	0
Membrane (M)	5	0	0
Nucleocapsid (N)	13	10	0
ORF6	1	0	0
ORF7a	2	0	0
ORF8	0	3	0

5 Figures

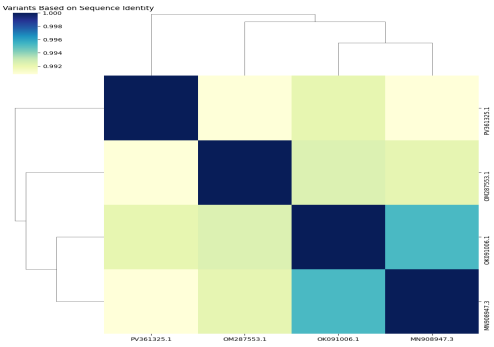


(a) Distribution of read depth across the genome. Most bases have a coverage depth of 3, indicating consistent sequencing and suggesting high confidence in the observed genome with relatively few ambiguous or low-confidence regions.

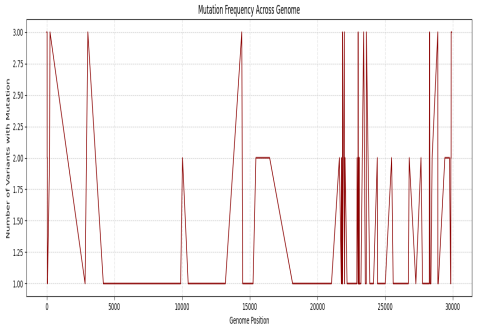


(b) Relative abundance of aligned genomes. All genomes contribute nearly equally to the analysis.

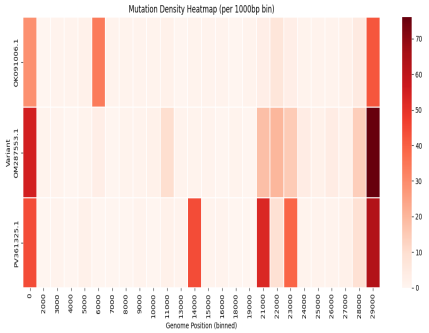
Fig. 1: Sequencing quality and genome representation in the dataset.



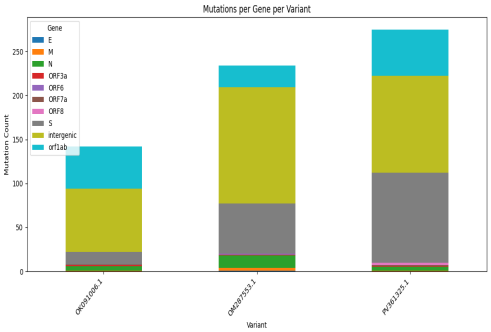
(a) Hierarchical clustering of genome variants based on pairwise identity. Closer branches indicate higher sequence similarity.



(b) Line plot showing mutation frequency along the genome. Peaks represent mutation hotspots.

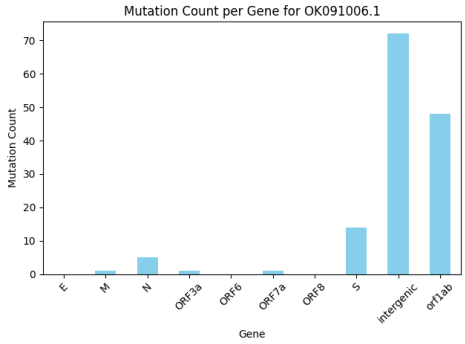


(c) Heatmap depicting mutation frequency along the genome. Darker regions indicate higher mutation density.

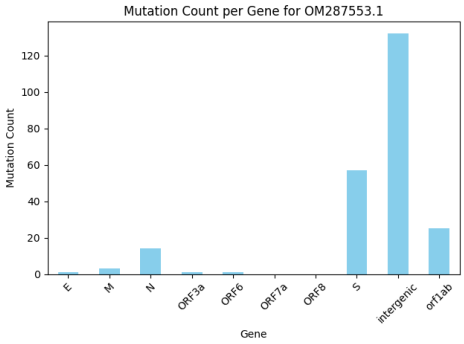


(d) Stacked bar plot of mutation counts per gene categorized by mutation type.

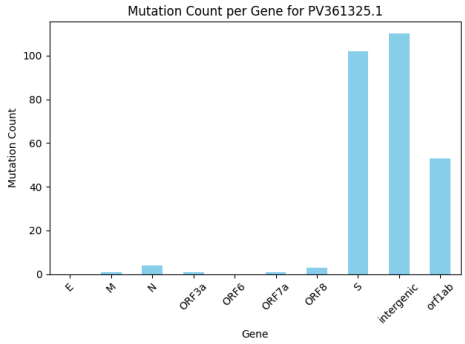
Fig. 2: Visualizations showing clustering and mutation patterns across the SARS-CoV-2 genome variants.



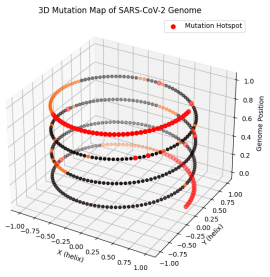
(a) Mutations per gene — Japan strain (OK091006.1)



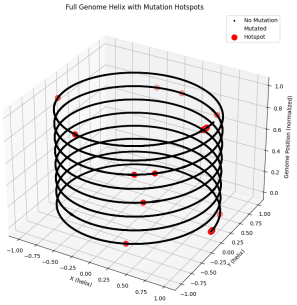
(b) Mutations per gene — Netherlands strain (OM287553.1)



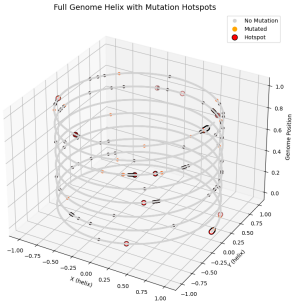
(c) Mutations per gene — Indian strain (PV361325.1)



(d) 3D mutation map of genome.



(e) Helical view 1 with mutation positions.



(f) Helical view 2 showing mutation density.

Fig. 3: Visualizations of gene-level mutation counts and structural perspectives across SARS-CoV-2 variants.

6 Conclusion

This study investigated the mutational dynamics of multiple SARS-CoV-2 variants through a detailed alignment-based approach, supplemented by multi-dimensional visualization techniques.

Point mutations, insertions, and deletions were identified by aligning each variant against the reference genome (*MN908947.3*) and annotating the changes. Gene-level mutation profiling indicated that regions such as `S` and `ORF1ab` exhibited high variability, reflecting their functional significance in viral replication and host interactions.

A 3D helical representation of the genome provided an intuitive spatial view of mutation clustering, helping to pinpoint regions with dense variation. Complementary analyses, including heatmaps and clustering, highlighted inter-variant similarities and implied possible evolutionary relationships.

These insights underscore the ongoing evolution of SARS-CoV-2 and the need for robust genomic surveillance to detect emerging variants of concern. Although the study presents a versatile visual and computational framework, it is constrained by the availability of sequence data and static genome annotations. As new variants arise, interpretations may need to be updated accordingly.

Future directions could involve integrating protein structure prediction tools to evaluate functional impacts of mutations. Broadening the dataset and incorporating host interaction profiles would further strengthen its relevance for epidemiological tracking and vaccine strategy development.

Supplementary Information. Supplementary figures, plots, and code used for mutation detection, gene annotation, and visualization are available upon request. No electrophoretic gels or blots were used in this computational study.

Acknowledgements. The authors would like to thank the NCBI Virus portal and GenBank for providing access to publicly available SARS-CoV-2 genome sequences. We also acknowledge the developers of Biopython, Matplotlib and Seaborn libraries used in this research.

Declarations

- **Funding** :
  Not applicable.
- **Conflict of interest/Competing interests** :
  The authors declare no competing interests.
- **Ethics approval and consent to participate** :
  Not applicable. This study does not involve human or animal subjects.
- **Consent for publication** :
  Not applicable.
- **Data availability** :
  All genomic data analyzed in this study are publicly available from the NCBI Virus portal and GenBank. Accession numbers include: `MN908947.3`, `PV361325.1`, `OM287553.1`, and `OK091006.1`.

- **Materials availability** :
  Not applicable.
- **Code availability** :
  All source code used in this study, including scripts for mutation extraction, visualization, and alignment-based analysis, is available at: https://github.com/AnshBajpai05/Quantification-and-Mutation-Analysis-of-closely-related-genomes

References

[1] World Health Organization. (2020). Coronavirus disease (COVID-19) pandemic. Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019

[2] Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., Garry, R. F. (2020). The proximal origin of SARS-CoV-2. *Nature Medicine*, 26(4), 450–452. https://doi.org/10.1038/s41591-020-0820-9

[3] Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

[4] Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., ... de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163