## 

**Bioinformatics Project Covid-19**

**Prepared by:**

Abdelrahman Kamal Ibrahim (ID: 4210201)

Mohamedosama (ID:1190234)

Ahmed Gehad Mohamed (ID: 1200387)

Submitted to:

Dr. Ibrahim Mohamed Youssef

**Comparative Genomic Analysis of SARS-CoV-2 Delta and Omicron Variants**

**Introduction**

The emergence of SARS-CoV-2 variants has significantly impacted the course of the COVID-19 pandemic. Among these, the Delta and Omicron variants have garnered particular attention due to their enhanced transmissibility, potential for immune escape, and differences in pathogenicity. Understanding the genetic differences between these variants can provide insights into their behavior and inform public health strategies. This study aims to perform a comparative genomic analysis of the Delta and Omicron variants to elucidate their genetic differences and evolutionary relationships.

**Methods**

**Data Collection**

We retrieved 10 sequences each of the Delta and Omicron variants from Egypt from the GISAID database, a global repository for SARS-CoV-2 sequences. These sequences were chosen to represent the genetic diversity within each variant.

**Software and Tools**

The following software packages and tools were used in this study:

**Data Collection**

* Sequences of the Delta and Omicron variants were downloaded from the GISAID database.
* Ten sequences from each variant were used for the analysis.

**Consensus Sequence Construction**

* **Software:** Biopython
* **Process:**
  + Parsed sequences using SeqIO.
  + Constructed consensus sequences by identifying the most frequent nucleotide at each position across the sequences.

**Multiple Sequence Alignment**

* **Software:** Clustal Omega
* **Process:**
  + Used Clustal Omega for multiple sequence alignment of the Omicron sequences.
  + Used Clustal Omega for multiple sequence alignment of the Delta sequences.
  + Multiple sequence alignment of the Omicron sequences and Delta sequences together for the phylogenetic tree construction.
  + Aligned Delta consensus sequence with Omicron sequences.

**Nucleotide Composition and CG Content**

* **Software:** Biopython
* **Process:**
  + Calculated the average percentage of each nucleotide (A, T, C, G) for both variants.
  + Computed CG content using **gc\_fraction**.

**Phylogenetic Analysis**

* **Software:** Biopython (Phylo module), Matplotlib,MEGAX 11
* **Process:**
  + Constructed a phylogenetic tree using the Neighbor-Joining method.
  + Visualized and saved the tree as an image.
  + Further visualization using megax

**Dissimilar Regions Identification**

* **Software:** Biopython, Counter from collections
* **Process:**
  + Compared the consensus sequence of Delta with Omicron sequences.
  + We find the dominant nucleotide in each column in the Omicron sequences.
  + Identified positions with significant dissimilarities, by comparing the dominant nucleotide with the consensus sequence at their corresponding position.
  + Print the columns of the omicron sequences where there is dissimilarities with their position

**Step-by-Step Procedures**

**Consensus Sequence Construction**

1. **Parsing Sequences**:

Code from Bio import SeqIO

from Bio.SeqRecord import SeqRecord

from Bio.Seq import Seq

1. **Constructing Consensus Sequence**:
   * For each variant, the most frequent nucleotide at each position was identified to construct the consensus sequence.

**Multiple Sequence Alignment**

1. **Aligning Sequences**:
   * Clustal Omega was used to align the sequences. The command executed via subprocess in Python was:

**python**

**Copy code**

**import subprocess**

**clustalo\_path = "path/to/clustalo"**

**cmd = [clustalo\_path, "-i", input\_file, "-o", output\_file, "--outfmt=clustal", "--output-order=tree-order"]**

**subprocess.run(cmd, check=True)**

**Nucleotide Composition and CG Content Analysis**

1. **Calculating Average Nucleotide Percentages**:

**def calculate\_avg\_percentages(fasta\_file):**

1. **Calculating CG Content**:

**from Bio.SeqUtils import gc\_fraction**

**def calculate\_avg\_gc\_content(fasta\_file):**

**# Code to calculate CG content**

**Dissimilarity Analysis**

1. **Identifying Dissimilar Positions**:

**from collections import Counter**

**def extract\_dissimilar\_columns(alignment, positions):**

**# Code to identify dissimilar positions**

**Phylogenetic Analysis**

1. **Constructing Phylogenetic Tree**:

**python**

**Copy code**

**from Bio.Align import MultipleSeqAlignment**

**from Bio.Phylo.TreeConstruction import DistanceCalculator, DistanceTreeConstructor**

**from Bio import Phylo**

**import matplotlib.pyplot as plt**

**# Code to construct and visualize phylogenetic tree**

**Results and Discussion**

**Consensus Sequence Analysis**

The consensus sequences for the Delta and Omicron variants were constructed successfully. These sequences represent the most frequent nucleotide at each position across the respective variant sequences.

**Nucleotide Composition and CG Content**

The nucleotide composition and CG content for both Delta and Omicron variants were calculated as follows:

**Delta Variant**:

* C: 18.30%
* A: 29.94%
* T: 32.18%
* G: 19.59%
* CG Content: 37.88%

**Omicron Variant**:

* C: 17.89%
* A: 29.28%
* T: 31.60%
* G: 19.19%
* CG Content: 37.85%

The slight differences in nucleotide composition and CG content between the two variants could influence their replication dynamics and interaction with the host immune system.

Similar CG Content and Thermostability:

Genetic material with similar CG content tends to exhibit comparable thermostability. This means that DNA or RNA sequences with similar CG percentages are more likely to have similar melting temperatures (**the temperature at which they denature**).

The relationship between CG content and thermostability is relevant because it can impact the survival and adaptation of organisms in different environments, including geographical regions with varying temperatures.

**Dissimilar Regions**

Significant dissimilar positions between the Delta consensus sequence and the Omicron sequences were identified, indicating regions of high variability. These positions include:

[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 209, 669, 883, 1058, 1626, ...]

dissimilarity region length =618 base avg omicron length=30794 base,

Average dissimilarity percentage 2%.

Sequence dissimilarity refers to the degree of difference or divergence between two genetic sequences. The difference between the consensus sequence (representative sequence) and the 10 Omicron variant sequences.

Researchers use dissimilarity measures to: Analyze genetic variation: By quantifying how much sequences differ, scientists can understand the genetic diversity within a population.

Identify conserved regions: Conserved regions are parts of the genome that remain relatively unchanged across different individuals or species. These regions often play critical roles (e.g., encoding functional proteins). These regions may correspond to changes in viral proteins that affect transmissibility and immune evasion.

Conserved regions are crucial for vaccine development and drug design. Why?

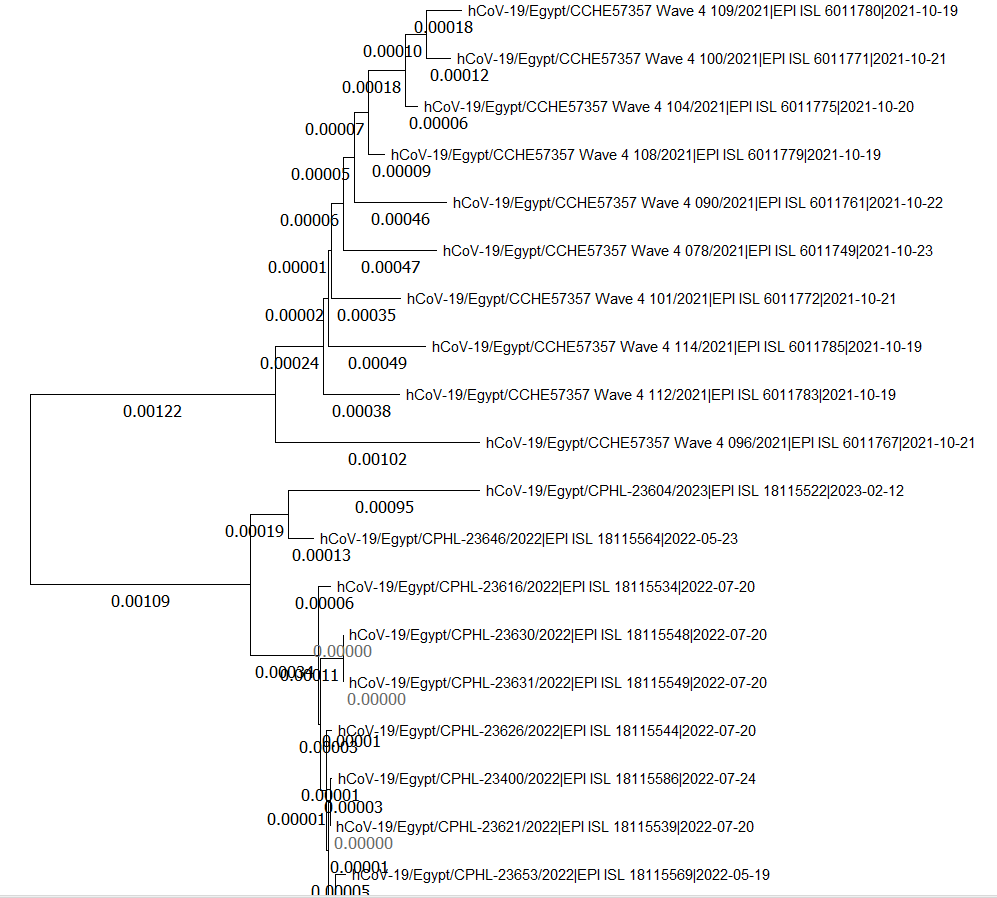
Vaccines often target specific conserved regions (e.g spike proteins) to generate an immune response.

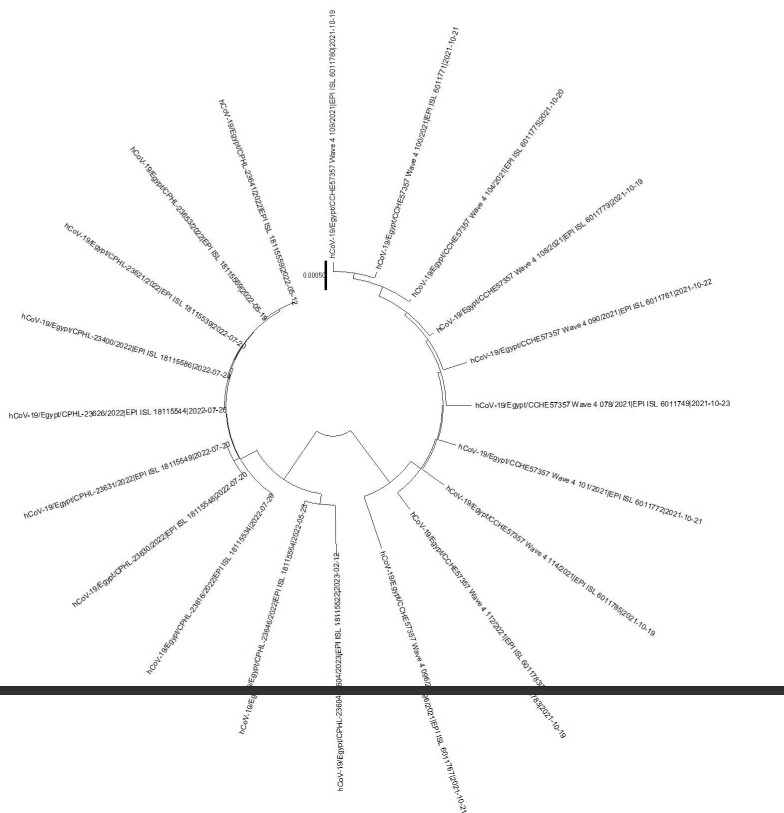
Less conserved regions (high dissimilarity) are challenging for vaccine development because they may mutate rapidly, making it harder to create a universal vaccine effective against all variants.

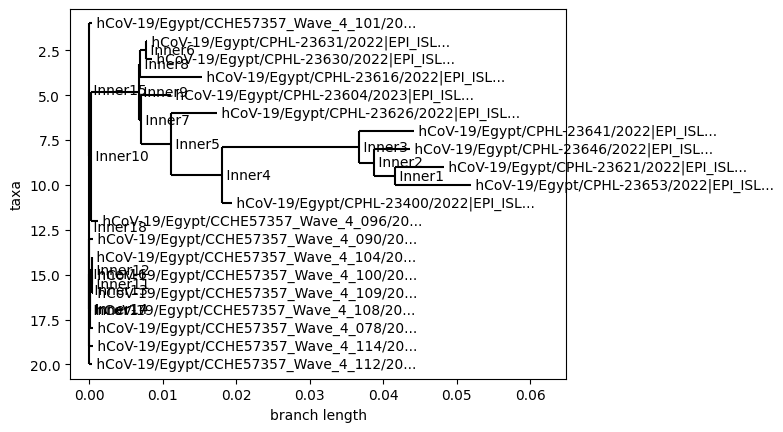
Omicron’s low similarity percentage (2%) suggests significant divergence from the consensus sequence, which complicates vaccine design.

**Phylogenetic Analysis**

The phylogenetic tree constructed using the Neighbor-Joining method revealed the evolutionary relationships between the Delta and Omicron sequences. The tree showed distinct clustering of sequences from each variant, reflecting their evolutionary divergence. The tree was visualized and saved as **output\_tree.png.**

****

****

****

Co-circulation of variants and sub-lineages in close evolutionary environments suggests convergent and directional evolution.

This analysis provides insights into prospective vaccine efficacy against different SARS-CoV-2 strains (Omicron).

Phylogenetic analysis plays a crucial role in understanding the evolution, spread, and potential vaccine efficacy against SARS-CoV-2 variants, including Omicron. Researchers continue to explore these aspects to inform public health strategies and vaccine development.

**Conclusion**

This study provides a comprehensive comparative genomic analysis of the SARS-CoV-2 Delta and Omicron variants. Key findings include:

* Construction of consensus sequences for Delta and Omicron variants.
* Differences in nucleotide composition and CG content between the variants.
* Identification of dissimilar regions that may contribute to the distinct properties of each variant.
* Phylogenetic analysis revealing evolutionary relationships.

These findings enhance our understanding of the genetic differences between these significant variants and can inform future research and public health strategies. Further functional studies on the identified dissimilar regions are warranted to understand their impact on viral behavior and interaction with the host. The low similarity percentage in the Omicron variant highlights the challenge of developing broadly effective vaccines due to its highly dissimilar regions.