Final Report: BIOE 131 Project

Qile Yang, Juno Lee, and Annalea Maeder

December 10, 2024

Our team focused on developing a database installer for the viral family Coronaviridae, specifically targeting SARS-CoV-2, the virus responsible for the COVID-19 pandemic. The thematic focus of this project centers on providing a simple, accessible, and lightweight collection of SARS-Cov-2 sequences from across the world to study the genomic diversity, integrating genomes and annotations from geographically diverse regions. It is paired with an intuitive web-explorer interface to allow non-technical users to explore the data and its implications. This proof of concept project aims to address the urgent need for accessible bioinformatics tools that support global research on viral evolution, variant tracking, and pandemic preparedness.

1 Viral Family Selection: SARS-CoV-2

SARS-CoV-2 was chosen due to its global impact, rapid evolution, and the critical role of genomic research in combating the pandemic. The virus continues to evolve, leading to the emergence of new variants with potentially significant effects on transmissibility, immune evasion, and vaccine efficacy. By focusing on SARS-CoV-2, our project contributes to ongoing efforts to understand these evolutionary dynamics and equips researchers with the tools to study the virus's genetic and functional characteristics in detail. The decision to include genomes from various regions worldwide reflects the virus's global nature and the need for region-specific insights. Variants such as Alpha, Delta, and Omicron have demonstrated how regional mutations can drive significant changes in the virus's behavior and public health impact. By analyzing a curated set of geographically representative genomes, this lightweight platform allows researchers to explore how environmental, demographic, and immunological factors influence viral evolution and spread, without being overwhelmed by extraneous metadata or features.

2 Thematic Focus and Content Principles

The database we developed adheres to several guiding principles to ensure its relevance, utility, and adaptability:

3 Diversity and Evolution

The inclusion of complete genomes from SARS-CoV-2 strains across North America, Europe, Asia, and Africa highlights the virus's genetic diversity. Focusing on a curated selection, the database allows users to pull the data to directly track trajectories and identify regional similarities and differences. Understanding this diversity is essential for analyzing how lineages adapt to different host populations or environments.

4 Annotation and Functional Insights

Each genome is annotated with information on gene structures, regulatory regions, and mutations. Key functional elements such as the spike protein, nucleocapsid protein, and RNA-dependent RNA polymerase are highlighted. These insights are critical for understanding viral interactions with host cells, immune evasion, and resistance to therapies.

5 Geographical Context

The database includes geographic metadata, like the country of origin for each genome. This feature enables researchers to track mutation distribution and variant emergence. For instance, users can study how mutations in the spike protein correlate with the spread of variants in specific regions.

6 Variant Tracking and Research Utility

Researchers can compare genomes across regions and lineages, facilitating the study of mutations linked to transmissibility, vaccine escape, or drug resistance. The database's streamlined approach allows for high-impact correlations that inform public health strategies.

7 Integrative Visualization

Leveraging the JBrowse2 platform, the database offers an interactive visualization of genomic data. Users can explore multiple genomes and gene annotations in one interface, aiding in the connection of genomic features to their functional or clinical implications.

8 Adaptability and Scalability

Designed for future updates, the database can accommodate new sequences, annotations, and metadata. Its lightweight design ensures it remains user-friendly and relevant as the pandemic evolves.

9 Innovation and Impact

Our approach combines diversity, functionality, and accessibility to create a platform tailored to SARS-CoV-2 research. By integrating multiple genomes with rich annotations and extensible metadata, the database supports diverse use cases. This flexibility makes it an invaluable tool for researchers addressing critical questions about viral adaptation, transmission, and pathogenicity. The database's web-based geographic context and integration of jbrowse enables users to explore how regional factors influence viral evolution and public health outcomes, and lowers barrier to entry.

10 Conclusion

This project contributes to the global fight against COVID-19. Through this work, we aim to support long-term efforts to understand and mitigate the impact of emerging viral pathogens, and provide insights for a possible future pandemic within the same family.