Cyrill Castro
Phung Le
Melinda Luo
Peter Nguyen

BioE 231 Final Project

For our database, we chose to focus on influenza, a common virus that affects millions of people each year. While a majority of infected people fully recover from the flu, young children and older adults are much more vulnerable to the virus and can sometimes die from it. This makes finding treatments for the disease crucial for such populations, and finding highly mutated genomic regions is often a plausible target for therapeutic treatments. Our database aims to display information related to highly mutated influenza regions which may give us insights into the evolution of the influenza virus. Equipped with this understanding, we will be able to identify genes of interest coupled with their protein structures, construct viral lineages, and explore potential areas for further research questions in the field of bioinformatics.

One aspect of our database involves displaying the influenza genome. While there are various versions of influenza, from influenza A to D, we chose to specifically focus on influenza A, as it is the most common type responsible for the seasonal flu and epidemics. We display the known NCBI reference genome sequences from 1934 and 2009, where the H1N1 influenza strain was responsible for the flu pandemics totaling around 151,700-575,400 deaths worldwide. These sequences will allow us to view how the influenza genome has evolved, as well as view how specific sequences have mutated. There is an abundance of genes located across the influenza A viral genome which include but are not limited to PB1, PB2, PB1-F2, PA, PA-X, HA, NP, NA, M2, M1, NS1, and NEP as well as their protein products which include polymerases, matrix proteins, and other nonstructural proteins. In the influenza A virus, we have chosen to focus on the hemagglutinin (HA) and neuraminidase (NA) genes which produce two glycoproteins essential for viral infection and release.

Our database also features a 3D protein viewer, which will allow us to not only examine hemagglutinin and neuraminidase protein structure but also discuss their function. Examining the protein structure of these two genes on our genome reveals that hemagglutinin is a trimeric glycoprotein composed of a globular head (HA1) and a long alpha helix stem (HA2). As seen in the protein viewer, the three parts of the protein have a head region responsible for recognizing sialic acid which is ubiquitous on cell-surface glycoproteins and glycolipids, while the stem region assists in fusion with host cells. In contrast, neuraminidase is a tetrameric transmembrane protein containing four identical protomers of the larger four-part protein identified in orange, green, blue, and magenta. Each of these four protomers contains a head, stem, and transmembrane anchor, each serving a different purpose. The head is responsible for cleaving sialic sites, the stem is responsible for increased surface area for substrate interaction, and the anchor will attach the protein to the viral envelope due to its nature as a transmembrane protein. Both of these proteins together as hemagglutinin essentially targets sialic acid residues on the cell surface and neuraminidase will destroy the sialic acid for newly made viruses so that the new viruses do not bind to already infected host cells. This mechanism is the basis for influenza

vaccination where antibodies block infection, but subsequent repeated infections require updated antibodies due to the change in the glycoprotein properties of a cell surface post- immune system response. Antibodies will subsequently target this glycoprotein to prevent the spread of the virus, where neuraminidase will be inhibited either via direct binding at the active site or steric hindrance of the antibodies binding to the site. Over time, mutations in different influenza strains will drive the development of vaccines, and this process has been visualized through time with the different influenza epidemics across the centuries.

Another feature of our database is a viral influenza A lineage tree from 1918 to 2009 constructed using the hemagglutinin gene's protein sequences obtained from NCBI. Sorting the data order by year and performing multiple sequence alignments using MAFFT, we generated a phylogenetic tree using this data to visualize and better understand the pattern. This tree allows us to examine how the gene has evolved and its potential relationship to the 2009 pandemic. Most of the strains found during the 2009 period are clustered together in their own clade in the tree, separate from the strains found in other years. While there were many factors that contributed to the 2009 pandemic, this clustering suggests that changes in the HA gene may have affected the virus' transmissibility and adaptations during that time.

In addition to constructing a mutation lineage tree, we wanted to observe whether the 2009 variant of the H1N1 virus and the range of host response were attributed to differences in human genomic makeup via a genome-wide association study (GWAS) featured in our database. With data retrieved from a total of 150 mild to severe influenza cases in Europe with immune responses, we were able to construct a Manhattan plot based on the most significant SNPs (single nucleotide polymorphisms) found on each gene, thus yielding several sparse but strongly correlated data values represented as lines on our GWAS track. Although there are a variety of genes with a high log-transformed p-value, the highest were LOC286114 and LZTS1 with a value of 31.522. The former gene has not been thoroughly studied yet and its function is unknown. However, LZTS1 is a Leucine Zipper Tumor Suppressor protein that does have a huge role in the regulation of the cell cycle, although it is mainly implicated in suppressing cancerous tumors. As viruses make use of the host cell's native duplicating processes, LZTS1 may have a role in affecting influenza infection for this European population.

Future work – not only limited to these two genes but on diverse populations – can be made using GWAS. As we know, hemagglutinin facilitates viral entry into host cells while neuraminidase enhances the viral release from cells post-infection. Therefore, human genome datasets can be retrieved from diverse population samples to infer a relationship between the influenza A glycoproteins and the host response. For example, a mutation or variant in a human gene may be responsible for changing the sialic acid host cell makeup to affect influenza A detection. In doing so, GWAS may help to develop supporting research into drug development and protection against an evolving virus.

Works Cited:

1) Creytens, S., Pascha, M. N., Ballegeer, M., Saelens, X., & de Haan, C. A. M. (2021, October 29). *Influenza neuraminidase characteristics and potential as a vaccine target*. Frontiers. https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2021.786617/full
2) Garcia-Etxebarria K, Bracho MA, Galán JC, Pumarola T, Castilla J, Ortiz de Lejarazu R. Rodríguez-Dominguez M, Quintela I, Bonet N, Garcia-Garcerà M, Domínguez A, González-Candelas F, Calafell F; CIBERESP Cases and Controls in Pandemic Influenza Working Group. No Major Host Genetic Risk Factor Contributed to A(H1N1)2009 Influenza Severity. PLoS One. 2015 Sep 17;10(9):e0135983. doi: 10.1371/journal.pone.0135983. Erratum in: PLoS One. 2015 Oct 30;10(10):e0141661. doi: 10.1371/journal.pone.0141661. PMID: 26379185; PMCID: PMC4574704.
3) Gamblin SJ, Skehel JJ. Influenza hemagglutinin and neuraminidase membrane glycoproteins. J Biol Chem. 2010 Sep 10;285(37):28403-9. doi: 10.1074/jbc.R110.129809. Epub 2010 Jun 10. PMID: 20538598; PMCID: PMC2937864.
4) Hansen L, McMahon M, Turner HL, Zhu X, Turner JS, Ozorowski G, Stadlbauer D, Vahokoski J, Schmitz AJ, Rizk AA, Alsoussi WB, Strohmeier S, Yu W, Choreño-Parra JA, Jiménez-Alvarez L, Cruz-Lagunas A, Zúñiga J, Mudd PA, Cox RJ, Wilson IA, Ward AB, Ellebedy AH, Krammer F. Human anti-N1 monoclonal antibodies elicited by pandemic H1N1 virus infection broadly inhibit HxN1 viruses in vitro and in vivo. Immunity. 2023 Aug 8;56(8):1927-1938.e8. doi: 10.1016/j.immuni.2023.07.004. Epub 2023 Jul 27. PMID: 37506693; PMCID: PMC10529248.
5) Li Y, Wang L, Si H, Yu Z, Tian S, Xiang R, Deng X, Liang R, Jiang S, Yu F. Influenza virus glycoprotein-reactive human monoclonal antibodies. Microbes Infect. 2020 Jul-Aug;22(6-7):263-271. doi: 10.1016/j.micinf.2020.06.003. Epub 2020 Jun 19. PMID: 32569735; PMCID: PMC7303604.
6) Lu IN, Kirsteina A, Farinelle S, Willieme S, Tars K, Muller CP, Kazaks A. Structure and applications of novel influenza HA tri-stalk protein for evaluation of HA stem-specific immunity. PLoS One. 2018 Sep 27;13(9):e0204776. doi: 10.1371/journal.pone.0204776. PMID: 30261065; PMCID: PMC6160157.
7) *phylo.io*. (n.d.). https://mafft.cbrc.jp/alignment/server/spool/_phyloio.241206193929469.html

Contributions:

Phung Le contributed analysis on the phylogenetic tree and the evolution of influenza.
Peter Nguyen and Cyrill Castro contributed analysis on 3D protein structure, function, mechanism, and GWAS Manhattan plotting.
Melinda Luo contributed an annotational overview of influenza genome.
Melinda Luo and Cyrill Castro contributed to the GitHub genome browser compilation with instructions.

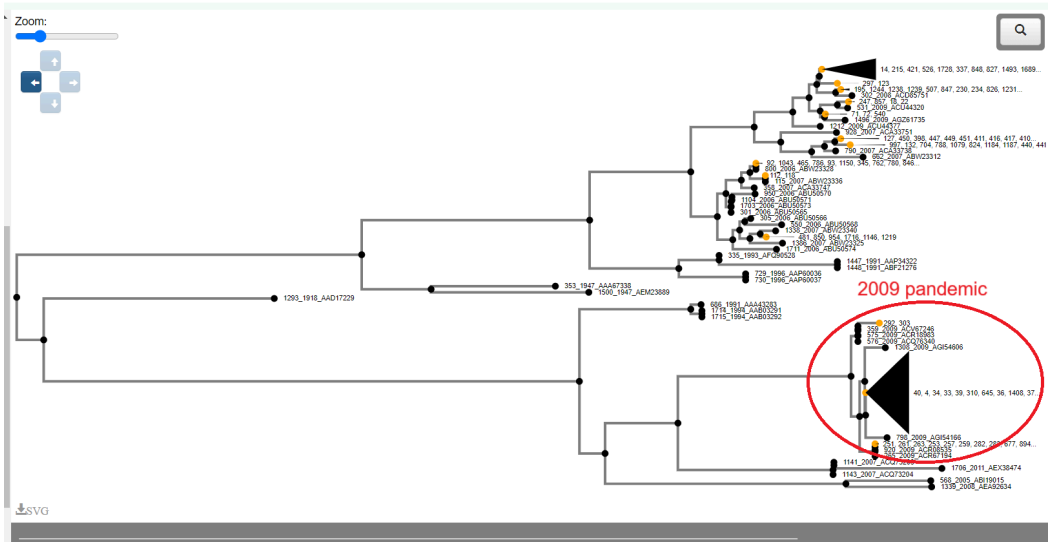

Figure 1.

Phylogenetic tree for HA 1918-2012. (All the 2009 strains are clustered together under similar nodes, suggesting that that specific version of hemagglutinin plays an important role in the 2009 pandemic.)