Introduction

Human influenza virus is a widely circulating pathogen, with pandemic strains potentially infecting 20-40% of the global population in a single year. ­­It’s presence is made known yearly, with a decent number of people seeking medical attention. Within the last decade, there has been an increasing effort to sequence flu strains from sick patients. As this has become technologically more feasible in recent years, there has been a great increase in the number of flu sequence available each year. This allows influenza’s evolution to be analyzed.

The most common type of flu in humans is influenza A H3N2. This nomenclature identifies the types of hemagglutinin and neuraminidase proteins, both of which are antigenically important in influenza mechanisms. This project examines the evolution of the hemagglutinin protein. The hemagglutinin protein is a surface glycoprotein which allows the virus to bind and enter host cells. It is composed of two subunits: a globular head domain (HA1 ectodomain) and a stem domain (HA2 ectodomain). The globular head contains epitopes recognized by the immune system, and is thus under constant evolutionary pressure in order to evade detection by antibodies while maintaining function. This has led to a high mutational tolerance in the HA1 domain, **insert stat here.** This mutational tolerance is incredibly useful in the process of evading the immune system and exhibit epistatic effects. These substitutions have shown patterns of genetic interaction; immune-escape and drug resistant phenotypes often appear through a combination of several mutations that have epistatic effects on pathogen fitness.

This project identified the largest clusters forming through evolutionary selection in sequences submitted to GISAID Epiflu between January of 2010 and June of 2018. The rates of non-synonymous and synonymous mutation were examined within these clusters in order to identify periods which experience higher than expected rates of non-synonymous mutation. **If was found that these periods exhibited higher than normal rates of nonsynonymous mutation.** Deep mutational scanning data was utilized to identify unexpected yet recurrent mutations. **We found these mutations to show these effects.**

Methods and Results

This project utilized surveillance sequence data made publicly available through the Global Initiative for Sharing All Influenza Data (GISAID) EpiFlu database. The amino acid sequences, nucleotide sequences, unique identifiers, and dates of sample collection for all influenza A H3 samples from January of 2010 to June of 2018 were obtained. Only sequences with valid amino acids in positions one to five hundred and sixty-six were considered in this analysis. These positions represent the signal peptide, HA1 and HA2 ectodomains, transmembrane, and cytoplasmic tail domains. In addition to the GISAID EpiFlu data, deep mutational scanning data of A/Perth/16/2009 (H3N2) was utilized to identify unlikely mutations.

Sequences were evaluated at the nucleotide level to create a distance matrix using the hamming.neighbors() function from the ‘vwr’ package available for R 3.0.1. All sequences belonging to clusters of at least three members were considered, however the following analysis used a subset of data that consisted of the seqeuences which made up the twenty largest clusters. The resulting networks were plotted using a weighted multidimensional scaling layout derived from the distance matrix as calculated through the wmds() function from the ‘vegan’ package available for R 3.2.0.

These sequences with the most one nucleotide neighbor within each of the top twenty clusters were identified. These sequences were ordered by date and a distance matrix was created for these sequences was created at the amino acid level. Several of these sequences were either synonymous sequences, or within one to three amino acids of each other. Several paths through sequence space and time were constructed. The ratio of nonsynonymous to synonymous neighbors was calculated for each other top twenty sequences. This was used to identify single amino acid changes which preceded changes in the nonsynonyomus to synonymous mutation ratio. **Why were we looking for this.** We were able to identify one mutation which caused an increase in the nonsynonymous to synonymous ratio and whose reversion caused a decrease in the nonsynonymous to synonymous ratio. This mutation was seen several times between the top twenty sequences, however we were unable to identify one amino acid linkages between clusters in order to further evaluate it.

In addition to identifying discrepancies in ratio of non-synonymous to synonymous mutation, this project also worked to identify uncommon yet recurrent mutations informed by deep mutational scanning data. To do this, all non-synonymous mutations which deviated from the **insert strain name** wildtype were identified at the amino acid level. Deep mutational scanning data was used to select for mutations which were seen in less than one percent of all deep mutational scanning trials. The nucleotide sequences for these mutations were evaluated to identify temporal periods when these mutations were observed, accessible via one nucleotide substitution, or inaccessible via one nucleotide substitution (**Figure XXX**). It was expected that some mutations would be seen more frequently toward the end of the sampling period due to increased sampling density. Mutations which exhibited patterns discrepant with this assumption were identified as: **XXX.**

Discussion

References

Outline:

* Created cluster network:
* Subset data to the top twenty clusters
* Created network among the top twenty sequences
* Several which were one or two nucleotide neighbors away from each other
  + Goal was to look for times when the NS:S ratio changes dramatically after one nucleotide substitution
* After identifying interesting trends, locate the one nucleotide change that caused it and identify its structural role
* From a one nucleotide substitution perspective the other thing that was done was looking at DMS data and identifying recurrent substitutions that shouldn’t be allowed.
  + Caveat: had to be cognizant of sampling density and that we could see ones that were just very popular toward the end of the sampling period
* Choose a few mutations that didn’t exhibit this trend and looked at the accessibility via the amino acid code
* Looked at these positions relevance in functionality.