# Basic Project Information

## Project Title: *Influenza Genome Sequence Trajectories*

## Project Repository

<https://github.com/sukrut56/Data_Wrangling_Project>

## Members

|  |  |  |
| --- | --- | --- |
| **Name** | **email** | **UID** |
| Hannah Van Hollebeke | vanhollebeke.hannah@gmail.com | u0697848 |
| Kolton Hauck | kolton.hauck@utah.edu | u1019364 |
| Sukrut Shishupal | sukrut.shishupal@utah.edu | u1401898 |
| Nick Patterson | Nicholas.Patterson@utah.edu | u6020304 |
| Vladimir Vinnik | v.vinnik@utah.edu | u6028096 |

# Background and Motivation

Influenza viruses are constantly evolving by undergoing genetic changes over time through genetic drift, selection, or recombination events. The CDC conducts year-round surveillance of circulating influenza strains to monitor changes in the genome of these viruses. The information from studying influenza genome trajectory plays an important role in public health by helping to determine whether current vaccines and antiviral drugs are effective against circulating influenza strains or if there is a need to develop new treatments.

Influenza viruses have been responsible for some of the deadliest pandemic events in human history and the nature of its genome plasticity and abundance of natural hosts make it a perpetual threat. Genomic sequencing has led to profound insights into the global pattern identification of circulating and evolving Influenza virus. These data are available online and can be used to visualize global sequence trajectories which can help us understand the spread of the Influenza virus and the changes in its genome over time and geographic location. Given the global impact that viruses have had in recent years, government agencies and corporations worldwide have invested enormous resources to study the origins of pandemics and mitigate their effects on society. Therefore, learning how to work with viral genomic data is an important skill for a modern data scientist interested in this type of research.

# Project Description

## Objective

The primary goal is to prepare data that show incidence of human-associated influenza A strain subtypes with a temporal component and provide the geographic location of countries across the globe where those subtype isolates were documented. The data will contain information regarding all H and N subtypes, as some of the most severe antigenic responses in humans are due to subtypes for which the natural reservoirs are other animals. The dataset would allow for connections between avian, swine, and human influenza subtypes. Additionally, the data would include a linkage to the isolate assignments for the relevant subtypes and the origin of the sequence data. These goals will be accomplished using publicly available influenza data from GISAID and NCBI.

These data could subsequently be used for a wide variety of research purposes dealing with analysis of the influenza genome. Potential useful information that could be gleaned includes mutational frequency, genomic location, and type (SNPs, duplications, recombination). These parameters could also be evaluated in a temporal and geographic context. Throughout the course of this project, we hope to become familiar with the fundamentals of data practices in this field and how to prepare these data to facilitate downstream analysis.

## Dataset Information

Influenza A genomes and associated information will be obtained from GISAID or NCBI Influenza Database. GISAID has a total of 348,530 Influenza A viruses and 1,627,175 sequences due to the segmented genome. The GISAID database organization runs analyses on the dataset that can be found here: <https://gisaid.org/database-features/influenza-genomic-epidemiology/>. By creating an account and logging in, the sequence and associated data can be directly downloaded. The download limit is 20,000 viruses and will therefore need to be executed in batches.

An alternate data source is the NCBI Influenza virus database which has a similar dataset: <https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?go=database>

Supplementary data may be added to the dataset as needed to enrich the information and/or provide more confidence in the data quality. An example dataset could be WHO data such as found here: <https://www.who.int/news-room/events/detail/2023/02/20/default-calendar/who-consultation-on-the-composition-of-influenza-virus-vaccines-for-use-in-the-2023-2024-northern-hemisphere-influenza-season>. These data estimate global infection rates, among other information.

## Data Processing

GISAID and NCBI are both highly curated databases which should allow us to retrieve a substantial dataset without considerable data cleaning. However, there may be some isolates that are missing values or have incomplete sequence data that we will need to address. Furthermore, it may be useful to generate some basic quality metrics on the sequence and virus data which would require some functions and wrangling. For example, it may be useful to a user to have an attribute with the percentage of IUPAC ambiguities (N, R, Y, M, K, B, D, H, V). This way the consumer could filter data based on some user defined threshold.

Data manipulation and formatting (e.g., consistent date formats) may be required to get all of the features that are required for the research question as well as filter out features not relevant to the intended use. For example, each isolate in GISAID only specifies what subtype the isolate is. We may need to consult other resources to also determine the specific variant each subtype is in so that we can provide further resolution in the data.

Data processing will primarily be done via Python using the Pandas Library. The data will be compiled into various csv files, including a data dictionary, and an associated relational schema will be developed to aid in wrangling and manipulation.

## Design

The primary way the data will be displayed is via a choropleth map, a plot that can give a lot of information in a simple plot. Choropleth maps can show various sub types of Influenza spatially simultaneously, magnitudes of infection in a location can be given, and a temporal component of the map will be included to show the spread of the subtypes across time.

Other plots and graphs will be included that will show other information not depicted in the choropleth map. These may include histograms to show the distributions of the dataset, or distributions of subsets of the dataset. Time series plots may also be included to show trends and relationships within specific subtypes.

Additionally, some basic charts or plots describing the dataset composition as well as basic descriptive statistics and quality metrics will be included.

An alternate to the choropleth map can be a dot-density map which uses dots or other symbols on the map to show values of one or more numeric data fields. Areas with many dots indicate a high concentration of values for the chosen field, while fewer dots indicate a lower concentration.

## Essential Features

We consider the following features essential, or without which the project would not be able to be completed:

* Subtype (H and N) and lineage if applicable
* Isolate name and ID
* Segment id
* Host
* Geographical region of sample collection
* Sample collection date
* Linkage to nucleotide and protein sequence data

## Desired Features

The following features are deemed desired, as they may contribute to the quality of the results of the project, improve the data quality of the dataset, and extend the usefulness of the data for further genomic analysis:

* Sequence quality metrics (e.g., percent ambiguity, sequence length, segment completeness)
* Association with recorded infections and mortality (WHO data)

# Project Schedule

The following schedule outlines general tasks and milestones we would like to meet. Specific tasks will be organized in weekly meetings and divided amongst team members accordingly.

## Week 1 (2023-02-16): Project Proposal

**Tasks**:

* Decide on project topic
* Organize project documents and shared resources
* Find datasets
* Submit project proposal
* Self-assessment

## Week 2 (2023-02-23): Data Exploration and Project Plan

**Tasks**:

* Collect all relevant data from GISAID/NCBI
* Data harmonization
* Data cleaning
* Create specific tasks and divide amongst team

## Week 3 (2023-03-02): Meet with Instructors

**Tasks:**

* Receive feedback
* Make changes accordingly
* Data exploration and descriptive statistics

## Week 4 (2023-03-09):

**Tasks:**

* Data exploration and descriptive statistics
* Data cleaning as needed
* Data structure mostly set and data model drafted

## Week 5 (2023-03-16): Project Update Submission

**Tasks:**

* Self-assessment
* Make sure GitHub is current

## Week 6 (2023-03-23):

**Tasks:**

* Start Presentation for peer feedback (focus on [Project Peer Feedback Document](https://docs.google.com/document/d/1fMsWBsujsYmT630IOOPagjONxB-hgb-raCtnbWl6hjg/edit))
* Main figures and tables drafted
* Potential biases identified
* Basic quality assessment complete

## Week 7 (2023-03-30): Present Intermediate Work

**Tasks:**

* Finish Presentation for peer feedback (focus on [Project Peer Feedback Document](https://docs.google.com/document/d/1fMsWBsujsYmT630IOOPagjONxB-hgb-raCtnbWl6hjg/edit))
* Peer feedback
* Review feedback and make changes accordingly

## Week 8 (2023-04-06): Meet with Instructors

**Tasks:**

* Receive feedback
* Make changes accordingly
* Discussing outcomes based on statistics

## Week 9 (2023-04-13):

**Tasks:**

* Start Presentation for peer feedback (focus on [Project Peer Feedback Document](https://docs.google.com/document/d/1fMsWBsujsYmT630IOOPagjONxB-hgb-raCtnbWl6hjg/edit))
* Discussing outcomes based on statistics

## Week 10 (2023-04-20):

**Tasks:**

* Presentation for peer feedback (focus on [Project Peer Feedback Document](https://docs.google.com/document/d/1fMsWBsujsYmT630IOOPagjONxB-hgb-raCtnbWl6hjg/edit))
* Discussing outcomes based on statistics

## Week 11 (2023-04-27): Presentation and Peer Feedback

**Tasks:**

* Finish Presentation for peer feedback (focus on [Project Peer Feedback Document](https://docs.google.com/document/d/1fMsWBsujsYmT630IOOPagjONxB-hgb-raCtnbWl6hjg/edit))
* Make final changes according to feedback
* Finalizing data outcomes and preparing report

## Week 12 (2023-05-04): Final Submission

**Tasks:**

* Wrangled Data in final format
* Git is current
* Presentation complete
* Documentation current