Molecular Epidemiology Midterm

Spring Semester 2024

Savannah Hammerton

Table of contents

# Introduction

Genetic drift leads a regular need to re-assess and update influenza vaccine strains from year to year (1). Differences between influenza vaccine strains and circulating strains have been cited as one of the causes of reduced vaccine efficacy (2–5). “Mismatch” between the strains is often defined somewhat arbitrarily and dichotomously, but using continuous genetic distance measures could lead to a better understanding of the impact of genetic drift and vaccine mismatch on changes in vaccine effectiveness in populations (3,5). However, there is not just one circulating strain per season, as genetic drift happens continuously over the flu season. One approach to beginning to understand the impact of genetic drift on vaccine effectiveness is to assess the gradual change and evolution over time and compare genetic distances using phylogenetic tree-based methods. In this project, I plan to do this for influenza A.

The purpose of this project will be to examine the differences in genetic drift and diversity between the H1N1 and H3N2 sub-types of the influenza A virus in humans in the United States over the years 2013-2017.

# Methods

Data were obtained from [BV-VRC.org](https://www.bv-brc.org/view/Taxonomy/11320?and(eq(host_group,Human),eq(isolation_country,%22USA%22),or(eq(collection_year,%222016%22),eq(collection_year,%222015%22),eq(collection_year,%222014%22),eq(collection_year,%222013%22),eq(collection_year,%222017%22)),or(eq(subtype,%22H1N1%22),eq(subtype,%22H3N2%22)),eq(genome_status,%22Complete%22),eq(segment,%224%22))#view_tab=genomes&filter=false), using these search filters(6):

* Genome Status: Complete
* Segment: 4 (HA)
* Subtype: H1N1, H3N2
* Collection Year: 2013-2017
* Isolation Country: USA
* Host Group: Human

The meta-data CSV and sequence FASTA files were downloaded and linked by GenBank Accession numbers in R (FASTA file was loaded using the seqinr package) (7). The final data set was then grouped by sub-type (H1N1 and H3N2) and year (2013-2017). Fifty random samples were taken in each group to reduce the final sample size from 9,853 to 500 (seed was set to 42 for reproducibility).

The H1N1 and H3N2 sequences will be used to generate two separate phylogenetic trees (one for each sub-type) that span 2013-2017, meaning there will be 250 sequences used for each tree. Genetic distances will be used to compare the genetic drift of the two sub-types over the observation period.

There are potential limits with this method and data. To decrease the overall sample size to a number requiring a more reasonable amount of computation time, only 50 sequences were sampled from each sub-type in each year. Additionally, there appears to be too much missing data to examine differences in age and gender ([Table 1](#tbl-h1n1) and [Table 2](#tbl-h3n2)).

# Exploratory Results

[Table 1](#tbl-h1n1) and [Table 2](#tbl-h3n2) show some basic meta-data for the H1N1 and H3N2 sequences, respectively. The tables are stratified by the year the sample was collected, and contain information on what season the sample is associated with and the age and gender of the host (all are human).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Population characteristics for H1N1 strains  **Table** **:** Population characteristics for H1N1 strains   | Var. | **2013**, N = 501 | **2014**, N = 501 | **2015**, N = 501 | **2016**, N = 501 | **2017**, N = 501 | | --- | --- | --- | --- | --- | --- | | season |  |  |  |  |  | | 12-13 | 20 (40%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | | 13-14 | 30 (60%) | 17 (36%) | 0 (0%) | 0 (0%) | 0 (0%) | | 14-15 | 0 (0%) | 30 (64%) | 7 (15%) | 0 (0%) | 0 (0%) | | 15-16 | 0 (0%) | 0 (0%) | 40 (85%) | 43 (91%) | 0 (0%) | | 16-17 | 0 (0%) | 0 (0%) | 0 (0%) | 4 (8.5%) | 17 (41%) | | 17-18 | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 24 (59%) | | Unknown | 0 | 3 | 3 | 3 | 9 | | age | 39 (21, 50) | 32 (22, 45) | NA (NA, NA) | NA (NA, NA) | NA (NA, NA) | | Unknown | 10 | 33 | 50 | 50 | 50 | | host\_gender |  |  |  |  |  | | Female | 16 (39%) | 17 (63%) | 6 (86%) | 1 (25%) | 27 (64%) | | Male | 25 (61%) | 10 (37%) | 1 (14%) | 3 (75%) | 15 (36%) | | Unknown | 9 | 23 | 43 | 46 | 8 | | 1n (%); Median (IQR) | | | | | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Population characteristics for H3N2 strains  **Table** **:** Population characteristics for H3N2 strains   | Var. | **2013**, N = 501 | **2014**, N = 501 | **2015**, N = 501 | **2016**, N = 501 | **2017**, N = 501 | | --- | --- | --- | --- | --- | --- | | season |  |  |  |  |  | | 12-13 | 50 (100%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | | 13-14 | 0 (0%) | 1 (2.2%) | 0 (0%) | 0 (0%) | 0 (0%) | | 14-15 | 0 (0%) | 45 (98%) | 17 (52%) | 0 (0%) | 0 (0%) | | 15-16 | 0 (0%) | 0 (0%) | 16 (48%) | 29 (60%) | 0 (0%) | | 16-17 | 0 (0%) | 0 (0%) | 0 (0%) | 19 (40%) | 23 (64%) | | 17-18 | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 13 (36%) | | Unknown | 0 | 4 | 17 | 2 | 14 | | age | 56 (41, 83) | 70 (52, 75) | 44 (44, 44) | NA (NA, NA) | 72 (72, 72) | | Unknown | 25 | 46 | 49 | 50 | 49 | | host\_gender |  |  |  |  |  | | Female | 15 (58%) | 25 (61%) | 12 (67%) | 9 (56%) | 18 (42%) | | Male | 11 (42%) | 16 (39%) | 6 (33%) | 7 (44%) | 25 (58%) | | Unknown | 24 | 9 | 32 | 34 | 7 | | 1n (%); Median (IQR) | | | | | | |

One interesting initial finding, as shown in [Figure 1](#fig-seq), is that there are some repeating H1N1 sequences across the observation time, but no repeating H3N2 sequences - there are 250 unique H3N2 sequences over the time period. We may expect to see increased genetic diversity in H3N2 when examining genetic distance as well.

|  |
| --- |
| Figure 1: Occurrences of H1N1 (left) and H3N2 (right) strains across 2013-2017 |

There appears to be a large variety of locations from which the sequenced samples arose ([Figure 2](#fig-loc)). However, the proportion of samples that had locations generalized to USA appear to be similar across the sub-types, which implies that, assuming the frequency of specific location is similar across sub-types in general, the random sampling used to generate the final data set did not induce bias (at least in this aspect).

|  |
| --- |
| Figure 2: Geographic location of samples for H1N1 (top) and H3N2 (bottom) strains across 2013-2017 |

H1N1 sample collections seemed to occur around the traditional flu season (late Fall to early Spring), but H3N2 sample collections seem slightly more sporadic ([Figure 3](#fig-date)). For example, the data set does not include any H3N2 samples collected between March of 2013 and April of 2014. However, this could be due to primary circulating strains, as the CDC Vaccine Effectiveness studies reported mostly H1N1 infections in the 2013-2014 season (8). There also appear to be a stronger seasonality in the H1N1 samples compared to the H3N2, which are often more spread out throughout the season, *e.g.* the 2015-2016 season.

There are some samples that have nonspecific collection dates (only the years are reported). The frequency of this by year and sub-type is presented in [Table 3](#tbl-miss).

|  |
| --- |
| Figure 3: Collection date of H1N1 (top) and H3N2 (bottom) samples across 2013-2017 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 3: Number of missing specific collection dates of samples per sub-type per year  **Table** **:** Number of missing specific collection dates of samples per sub-type per year   | Year collected | H1N1 | H3N2 | | --- | --- | --- | | 2015 | 1 | 0 | | 2016 | 1 | 0 | | 2017 | 3 | 6 | |

Overall, H1N1 patterns appear more consistent and regular than H3N2 patterns. Metadata appear similar between the two sub-types, hopefully indicating that random sampling did not induce any bias. I hypothesize that we will observe increased genetic drift and diversity in the H3N2 phylogenetic tree compared to the H1N1 tree.

# References

1. CDC. How Flu Viruses Can Change. 2022;(<https://www.cdc.gov/flu/about/viruses/change.htm>). (Accessed March 18, 2024)

2. Belongia EA, Simpson MD, King JP, et al. Variable influenza vaccine effectiveness by subtype: A systematic review and meta-analysis of test-negative design studies. *The Lancet Infectious Diseases* [electronic article]. 2016;16(8):942–951. (<https://www.sciencedirect.com/science/article/pii/S1473309916001298>). (Accessed September 28, 2023)

3. da Costa JC, Siqueira MM, Brown D, et al. Vaccine Mismatches, Viral Circulation, and Clinical Severity Patterns of Influenza B Victoria and Yamagata Infections in Brazil over the Decade 2010–2020: A Statistical and Phylogeny–Trait Analyses. *Viruses* [electronic article]. 2022;14(7):1477. (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9321334/>). (Accessed February 29, 2024)

4. Tenforde MW, Kondor RJG, Chung JR, et al. Effect of Antigenic Drift on Influenza Vaccine Effectiveness in the United States—2019–2020. *Clinical Infectious Diseases* [electronic article]. 2021;73(11):e4244–e4250. (<https://doi.org/10.1093/cid/ciaa1884>). (Accessed March 13, 2024)

5. Tricco AC, Chit A, Soobiah C, et al. Comparing influenza vaccine efficacy against mismatched and matched strains: A systematic review and meta-analysis. *BMC Med* [electronic article]. 2013;11(1):153. (<https://doi.org/10.1186/1741-7015-11-153>). (Accessed July 14, 2023)

6. BRC B. BV-BRC Viruses::Taxonomy Overview. (<https://www.bv-brc.org/view/Virus/10239>). (Accessed March 18, 2024)

7. Charif D, Lobry JR. SeqinR 1.0-2: A contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U, Porto M, Roman HE, et al., eds. *Structural approaches to sequence evolution: Molecules, networks, populations*. New York: Springer Verlag; 2007:207–232.

8. CDC. Seasonal Influenza Vaccine Effectiveness, 2013-2014 | CDC. 2019;(<https://archive.cdc.gov/www_cdc_gov/flu/vaccines-work/2013-2014.html>). (Accessed March 18, 2024)