Comparing Global Circulation Patterns of Gene Mutations of Two Key Virulence Proteins of the Influenza A Subtype H3N2

Mari Torii-Karch, Monica Wang, and Anna Kroskrity
Practical Genomics IB 134L, University of California, Berkeley
Berkeley, California

December 6, 2023

Background:

Influenza A is a disease of great interest as it causes seasonal influenza that can, depending on the strain, cause epidemics or pandemics. Seasonal infections are due to rapid variations in two key virulence genes that encode influenza surface glycoproteins: hemagglutinin (HA) and neuraminidase (NA), both of which are key targets to B cell response (Kosik & Yewdell, 2019). HA helps virions attach to host cells, while NA helps cleave the virion from host cells to complete the infectious cycle. Studies have shown that the two genes likely co-evolve due to them functioning as an intimate unit in enhancing viral transmission (Kosik & Yewdell, 2019, Neverov et al., 2015).

Looking into how new mutations in these two genes tend to spread globally, it has previously been seen that HA mutations tend to arise first in Southeast-East Asia and then quickly spread globally. This suggests that "once viruses leave E-SE Asia, they are unlikely to contribute to long-term viral evolution" (Russell et al, 2008; Bedford et al, 2015). This information asserts that mutations in HA correlate more by year than by region, however, this trend hasn't yet been studied in the NA gene. Therefore, we are interested in learning if a similar pattern of evolution and global spread is seen in the NA gene as well. The results of this project can provide more insight on the potential evolutionary relationship between these two key virulence genes of the influenza A subtype H3N2 in how these mutations spread on the global level.

Research Hypothesis and Predictions:

We propose that the spectrum of mutations within the NA gene co-evolve with the pattern in the HA gene within the lineage of the H3N2 viral subtype at the global level, resulting in similar patterns between the phylogenetic trees. In our phylogenetic trees, we predict that this hypothesis will be visualized by sequences grouping more by year than by region for both the HA and the NA genes. This result would further support the finding that HA mutations tend to evolve temporally rather than regionally, in addition elucidating the potential relationship between the global spread of HA and NA genes. Additionally, we predict that the evolutionary rates of the two trees will be similar, which would further suggest that the evolution of these two genes is correlated. If we find instead that there is a difference in the topological structure between the two phylogenetic trees and/or a significant difference in evolutionary rate, it would suggest that while the two genes may co-evolve, there are likely unique selective pressures on both genes that cause the phylogenetic trees or evolutionary rates to diverge.

Methods:

To begin constructing our phylogeny tree of NA and HA genes, we first gathered sequences for the HA and NA genes from the NCBI Viral Database (Bao et al., 2008). We selected five samples from the regions of Europe, the Americas, Africa-Middle East, Oceania, and East-Southeast Asia from the years 2015-2019 plus an outgroup sample from Singapore in 2013, totalling 126 samples per phylogenetic tree. After gathering all our data, we then generated

a multiple sequence alignment (MSA) for each gene using MAFFT in Bash. From there, we used ape to convert our two MSA files into phy objects in Rstudio. With these phy objects, we used phyML in Bash to construct a maximum likelihood tree for each based upon a GTR (generalized time-reversible) model. We then plotted this tree using FigTree to reroot to our selected outgroup sequence from Singapore in 2013 to get a more visually accurate tree. We then used the R package ggTree on this re-rooted tree in order to plot and color-code our phylogenetic trees for the HA and NA genes by region and by year which we then could use to determine trends in HA and NA gene mutation evolution (Gu et al., 2017).

In addition to generating phylogenetic trees of Influenza A HA and NA gene, we also generated evolutionary rates for these two genes, using TreeTime to run a root-to-tip regression plot on our two phylogenetic trees (Sagulenko et al., 2018). To test if the evolutionary rates of HA and NA genes were statistically similar, we used ANCOVA based statistics to generate a p-value to conclude whether the difference between the two genes' evolutionary rates was statistically significant or not.

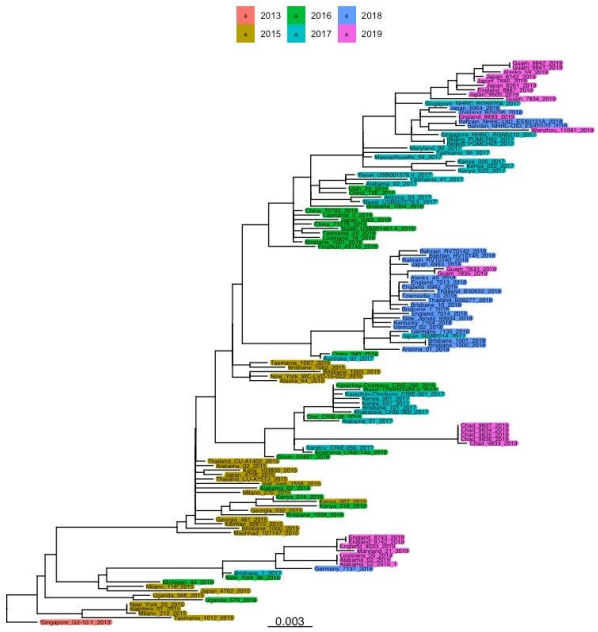
Results:

When looking at the phylogenetic trees color-coded by region and time, it appears that sequences from both HA and NA gene trees tend to be more closely related to other samples from the same time period rather than from the same region. An example of this is seen in Figure 1a, where we see the closely related samples at the top of the tree all originating from 2019 despite being from four different regions of the world (Oceania, East-Southeast Asia, Americas, and Europe). Visualizing the NA phylogeny tree, a similar trend was observed, where we see more closely related strains tending to be from the same year rather than from the same region. An example of this seen in Figure 2a is once again at the top portion of the tree where all the closely related NA samples originate from 2018 despite being from multiple different regions (Americas, East-Southeast Asia, Europe, Africa-Middle East).

Looking at the root-to-tip regression plots, it's seen that the evolutionary rates of the NA and HA genes were $3.13e^{-3}$ and $3.39e^{-3}$ respectively. After performing an ANCOVA statistical test to quantify the significance of the difference between these two evolutionary rates, it was calculated that the p-value was less than 0.05 (Table 1) indicating that the difference in the evolutionary rate between the NA and HA phylogenetic trees is statistically significant. Thus, mutations in NA do not follow the evolutionary trajectory of HA.

Figures and Tables:

a)



b)

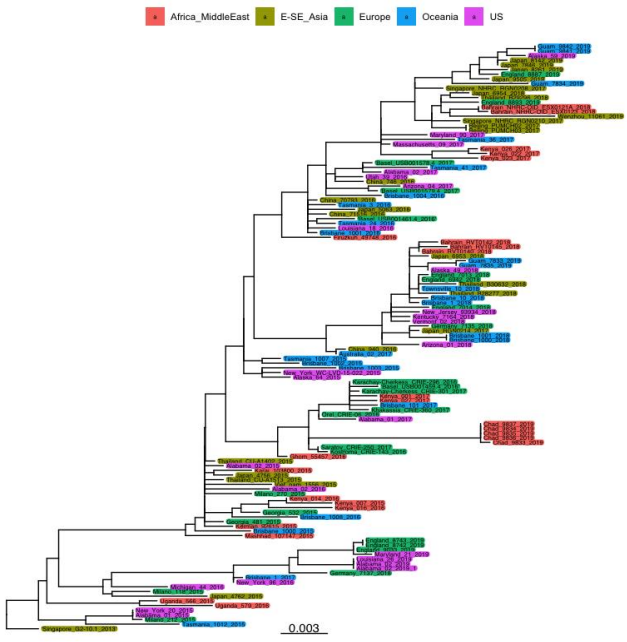
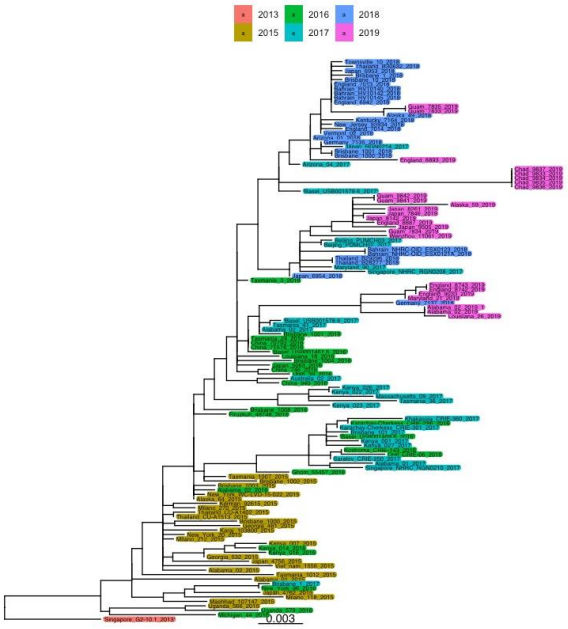


Figure 1. Influenza A H3N2 HA gene phylogenetic tree color-coded by region and time. The phylogeny has a total of 126 samples of the HA gene: 5 samples for each year (2015-2019) across 5 regions (Americas, Europe, Oceania, SouthEast-East Asia, Africa-Middle East) and an outgroup sample. The outgroup is a sample of Influenza A HA H3N2 from Singapore in 2013. (a) HA gene phylogenetic tree color-coded by time across the time period of 2015-2019. (b) HA gene phylogenetic tree color-coded by region.

a)



b)

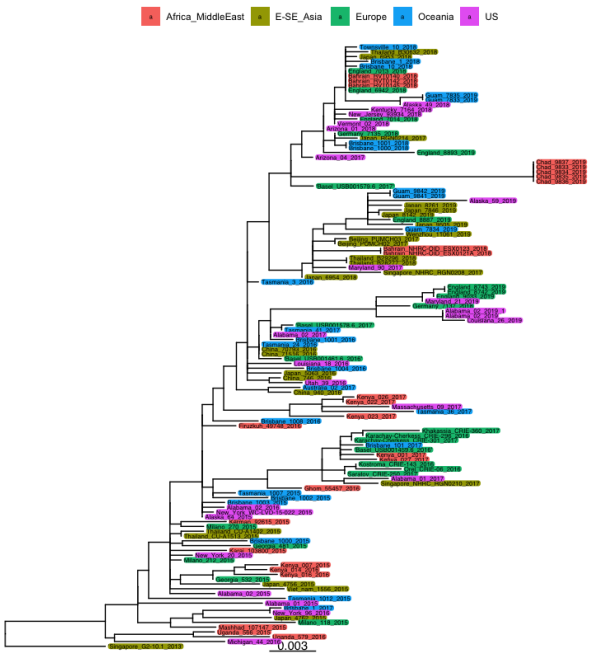
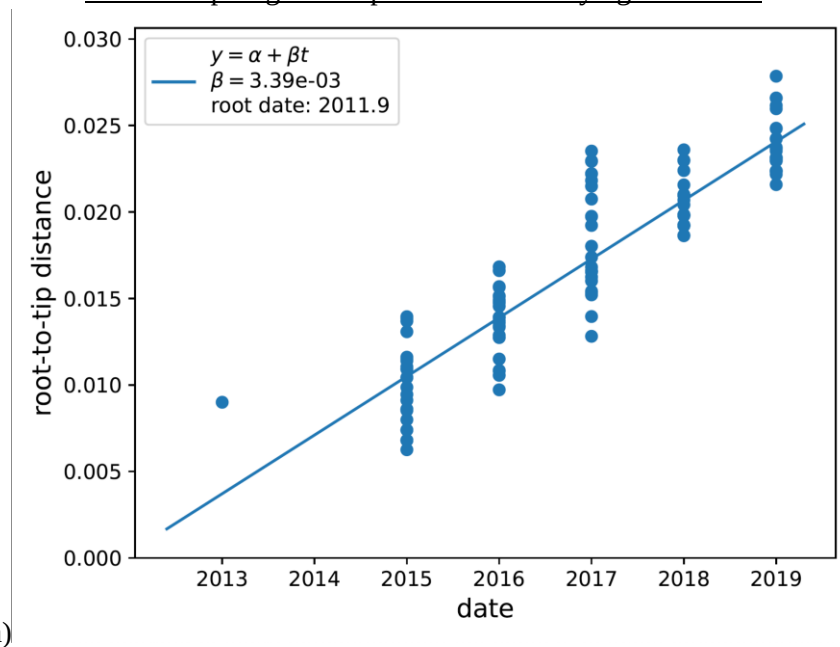


Figure 2. Influenza A H3N2 NA gene phylogenetic tree color-coded by region and time. The phylogeny has a total of 126 samples of the NA gene: 5 samples for each year (2015-2019) across 5 regions (Americas, Europe, Oceania, SouthEast-East Asia, Africa-Middle East) and an outgroup sample. The outgroup is a sample of Influenza A NA H3N2 from Singapore in 2013. (a) NA gene phylogenetic tree color-coded by time across the time period of 2015-2019. (b) NA gene phylogenetic tree color-coded by region.

a)



b)

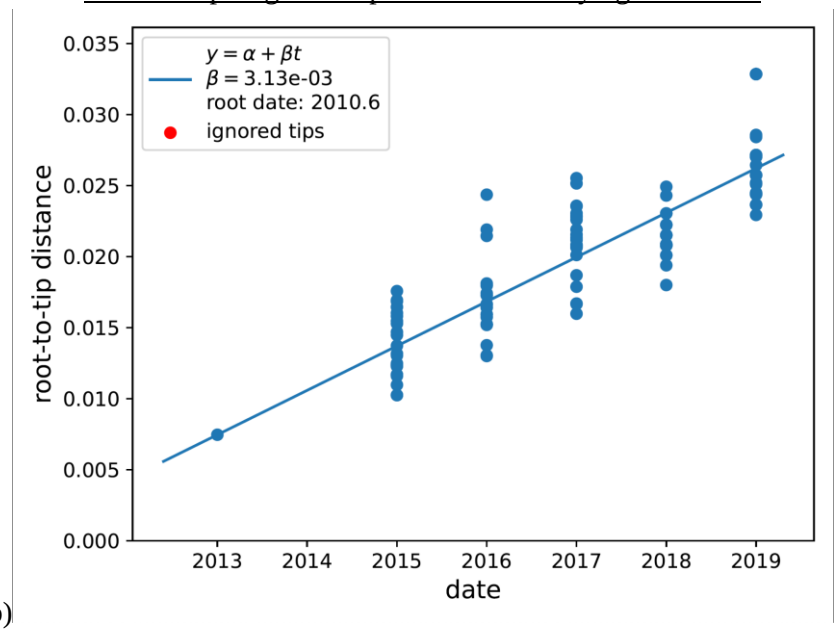


Figure 3. Evolutionary Rates of HA and NA genes via Root-Tip-Regression. Each dot represents a sample for each clarified gene. The line represents the line of best fit. β value represents the evolutionary rate. (a) Regression plot for HA gene (b) Regression plot for NA gene.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
date	1	0.015289	0.015289	5101.593	<2e-16	***
gene	1	0.000975	0.000975	325.335	<2e-16	***
date:gene	1	0.000026	0.000026	8.661	0.0034	**

Table 1. Statistical Significance test of root-to-tip distances for NA and HA gene based on ANCOVA statistical testing. All the ANCOVA calculated p-values are below the threshold value of 0.05 ($p < 0.05$) which rejects the null hypothesis that there is no significant difference in evolutionary rates between the two genes, aka there is a statistical difference between the evolutionary rates.

Discussion:

In answering our initial question asking if, on a global scale, mutations within the NA gene co-evolve in a similar pattern to the HA gene within the influenza A viral subtype H3N2, we have concluded that while these two genes have similar phylogenetic structure and grouping patterns, they follow different evolutionary trajectories. Through evaluation of our phylogenetic trees, we found that the known importance of temporality over region as a factor of viral evolution for the HA gene is apparent in the NA gene as well. However, through running an ANCOVA statistical analysis, we found the evolutionary rates of HA and NA to be dissimilar, with NA evolving at a slower rate. This difference suggests that while mutations in the nucleotide sequences of both the HA and NA genes are correlated with the same factor of viral evolution, time, there are likely differing selective pressures on these two genes or differences in gene properties causing their evolutionary rates to deviate.

This study could have been furthered by including more regions and selecting from a more diverse grouping of countries within each region. However, due to limitations in data availability we were unable to do so. Additionally, due to time constraints we selected only five individuals per region and studied only five years. Broadening the number of individuals and years we examined could have increased the accuracy of our phylogenetic trees and could have resolved some of the polytomies evident in our trees. Nevertheless, our research allowed a furthering of the understanding of the evolutionary dynamics of influenza virulence genes, which is crucial in predicting the antigenic evolution of the virus and in the design of effective vaccines.

Annotated Code:

1. Editing file names: using the sed command
    a. This command was used to edit the names of the sequences retrieved from the NCBI viral database so that they were in a format where we could successfully run MAFFT.
        i. We got rid of spaces, parentheses, and backslashes and replaced them either with an underscore or got rid of the character completely.

b. Down below is the code run on the HA sequence file, and the same set of code was used on the NA sequence file so that the names of each strain matched across both files (eg. both sequences from an individual in singapore from 2013 were named: Singapore_StrainID_2013)

```
HA_clockresults2/root_to_tip_regression.pdf
jovyan@jupyter-maritoriikarch:~/fall2023/Final$ sed "s/ /_/g" HA_sequences.fasta > HA_edits1.fa
jovyan@jupyter-maritoriikarch:~/fall2023/Final$ sed "s/(//g" HA_edits1.fa > HA_edits2.fa
jovyan@jupyter-maritoriikarch:~/fall2023/Final$ sed "s/)//g" HA_edits2.fa > HA_edits3.fa
jovyan@jupyter-maritoriikarch:~/fall2023/Final$ sed "s/\//_/g" HA_edits3.fa > final_HA.fa
```

2. Running a multiple sequence alignment: running MAFFT
    a. This command was used twice, once on the HA fasta file with all the HA sequences and once on the NA sequences. Using this command, we retrieved a multiple sequence alignment for each gene which could be used later when constructing a phylogenetic tree.

```
jovyan@jupyter-maritoriikarch:~/fall2023/Final$ mafft final_HA.fa > final_HA_align.fa
jovyan@jupyter-maritoriikarch:~/fall2023/Final$ mafft final_NA.fa > final_NA_align.fa
```

3. Creating the phy object: using the APE package in R
    a. We used the APE package in R to convert the multiple sequence alignment file into a phy object. This phy object was created to later run in phyMLin order to build our phylogenetic tree.

```
library('ape')
library('phytools')
library('ggtree')

nHA <- read.dna("final_HA_align.fa", format="fasta")
write.dna(nHA, "shortHA.phy", format="interleaved", nbcol=-1,colsep="")

NA_tree <- read.dna("final_NA_align.fa", format = "fasta")
write.dna(NA_tree, "finalNA.phy", format = "interleaved", nbcol =-1, colsep = "")
```

4. Creating the phylogenetic tree: Running phyML
    a. We ran phyML on the phy object we created in the previous step in order to get our maximum likelihood phylogenetic tree for both of our virulence genes. The output of this command was a phylogenetic tree in the format of a newick file.

```
jovyan@jupyter-anna-2dkrosk:~/fall2023$ phyml -i shortHA.phy -d nt -b -1 -m GTR
```

5. Getting of Phylogenies: Plotting with ggtree
    a. After running the phyML and getting our tree in the newick format, we used the application FigTree in order to reroot our tree with our designated outgroup: Singapore 2013 (this didn't require any code)
    b. We then exported the tree as a newick file, and imported it into an R file in order to plot our tree using the ggtree package (Gu et al., 2017).

c. This code helped us graph our phylogenetic trees for each gene and additionally color-code the trees either by region or by year. This was done within the code by downloading and opening ggtree, reading in all our necessary files (the newick trees as well as the information regarding year and region), and then plotting our NA/HA trees color-coding by either region/year.

```
library('ggplot2')
install.packages('ggtree')
library('phytools')

if (!require("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install(version = "3.18")
library('ggtree')
browseVignettes("ggtree")


fig_tree = read.tree('~/Desktop/trialtree2')
fig_tree


#ggtree(fig_tree) + geom_tiplab(align = TRUE, linesize = 0.5, size = 2) + coord_cartesian(clip = 'off')
test_region <- read.csv("~/Desktop/test_region.csv")
test_year <- read.csv("~/Desktop/year.csv")
print(test_year)

base_tree <- ggtree(fig_tree) + xlim(0, 0.04) + geom_treescale()
HA_tree <- base_tree %<+% test_year +
  geom_tiplab(aes(fill = factor(year)), color = "black", geom = "label", label.padding = unit(0.04, "lines"), label.size = 0, size = 2) +
  theme(legend.position = "top", legend.title = element_blank(), legend.key = element_blank()) +
  ggtitle('Influenza HA Tree')

NAtree = read.tree('~/Desktop/natrialtree')
NA_region <- read.csv('~/Desktop/NAregion.csv')
NA_year <- read.csv('~/Desktop/NAyear.csv')

base_na_tree <- ggtree(NAtree) + xlim(0, 0.04) + geom_treescale()
NA_tree <- base_na_tree %<+% NA_year +
  geom_tiplab(aes(fill = factor(year)), color = "black", geom = "label", label.padding = unit(0.04, "lines"), label.size = 0, size = 2) +
  theme(legend.position = "top", legend.title = element_blank(), legend.key = element_blank()) +
  ggtitle('Influenza NA Tree')
```

```
base_tree <- ggtree(fig_tree) + xlim(0, 0.04) + geom_treescale()
HA_tree <- base_tree %<+% test_region +
  geom_tiplab(aes(fill = factor(region)), color = "black", geom = "label", label.padding = unit(0.04, "lines"), label.size = 0, size = 2) +
  theme(legend.position = "top", legend.title = element_blank(), legend.key = element_blank()) +
  ggtitle('Influenza HA Tree')

NAtree = read.tree('~/Desktop/natrialtree')
NA_region <- read.csv('~/Desktop/NAregion.csv')
NA_year <- read.csv('~/Desktop/NAyear.csv')

base_na_tree <- ggtree(NAtree) + xlim(0, 0.04) + geom_treescale()
NA_tree <- base_na_tree %<+% NA_region +
  geom_tiplab(aes(fill = factor(region)), color = "black", geom = "label", label.padding = unit(0.04, "lines"), label.size = 0, size = 2) +
  theme(legend.position = "top", legend.title = element_blank(), legend.key = element_blank()) +
  ggtitle('Influenza NA Tree')
```

6. Getting evolutionary rates: Running TreeTime
   a. We used the program TreeTime to run a root-to-tip regression in order to calculate the evolutionary rates of each gene (Sagulenko et al., 2018). This was done by inputting in our re-rooted newick tree for each gene, a file of each strain with the corresponding year, and the length of the gene sequence. As an output, we expected to get a root-to-tip regression plot along with the root-to-tip distances for each strain.

```
jovyan@jupyter-anna-2dkrosk:~/fall2023/project$ treetime clock --tree natrialtree --dates HAyear.csv --sequence-length 1410
--outdir clock_resultsna
```

   a) Command to run treetime on our NA tree and obtain our root-to-tip regression plot and an rtt.csv file to run in ANCOVA

```
jovyan@jupyter-anna-2dkrosk:~/fall2023/project$ treetime clock --tree trialtree2 --dates HAyear.csv --sequence-length 1701
--outdir clock_resultsha
```

b) Command to run treetime on our HA tree and obtain our second root-to-tip

7. Calculating p-value to test if evolutionary rate of HA and NA are significant using ANCOVA from RStudio
    a. Loaded .csv files for both HA and NA and calculated their 95% confidence level for the regression slope
    b. Compared the two gene's 95% confidence intervals to see if there was any overlap
    c. Used aov() to generate p-values where response variable was root-to-tip distance is a function of the date, the gene, the effect of date depending on which gene

```r
#loading library necessary for p-val test
library(ggplot2)

# Read the data from the HA file
HAdata <- read.csv("HArtt.csv", skip = 1)
HAdata$date <- as.numeric(HAdata$date)
HAdata$gene <- 'HA'  # Add a gene column to the HA dataset

# Perform linear regression on the HA dataset
HAmodel <- lm(root.to.tip.distance ~ date, data = HAdata)

# Calculate 95% confidence interval for the slope of the HA dataset
confintHA <- confint(HAmodel, "date", level = 0.95)
confintHA

# Read the data from the NA file
NAdata <- read.csv("NArtt.csv", skip = 1)
NAdata$date <- as.numeric(NAdata$date)
NAdata$gene <- 'NA'  # Add a gene column to the NA dataset

# Perform linear regression on the NA dataset
NAmodel <- lm(root.to.tip.distance ~ date, data = NAdata)

# Calculate 95% confidence interval for the slope of the NA dataset
confintNA <- confint(NAmodel, "date", level = 0.95)

# Print the confidence intervals
print(confintHA)
print(confintNA)

# Compare the intervals
overlap <- !(max(confintHA[1,1], confintNA[1,1]) > min(confintHA[1,2], confintNA[1,2]))
overlap
print(paste("Is there an overlap in the confidence intervals? ", overlap))

# Combine the datasets
combined_data <- rbind(HAdata, NAdata)

# Perform ANCOVA
ancova_result <- aov(root.to.tip.distance ~ date + gene + date:gene, data = combined_data)
summary(ancova_result)

#perform GLM to confirm ANCOVA
model <- glm(root.to.tip.distance ~ date + gene + date*gene, data = combined_data)
summary(model)

```

Paragraph of contributions:

Mari gathered sequences for the HA and NA genes from the NCBI Viral Database and formatted the names of the sequences. All members ran multiple sequence alignment (MSA) for each gene using MAFFT in Bash and used ape to convert our two MSA files into phy objects in Rstudio. With these phy objects, we used phyML in Bash to construct a maximum likelihood tree for each based upon a GTR (generalized time-reversible) model. Mari then plotted this tree using FigTree in order to reroot our tree to a sample from Singapore 2013. She also used the ggTree on this re-

rooted tree in order to build and color-code our phylogenetic trees for the HA and NA genes by region and by year which we then could use to determine trends in HA and NA gene mutation evolution (Gu et al., 2017). Anna then generated evolutionary rates of these two genes using TreeTime that ran a root-to-tip regression plot on our two phylogenetic trees (Sagulenko et al., 2018). Monica used ANCOVA based statistics on the data generated from the root-to-tip regression plots to generate a p-value to test whether or not there was statistically significance between the two gene's evolutionary rate. Throughout the whole process, we all talked with the GSIs from this course about our process, gained insight on how/what to troubleshoot, and determined which packages/softwares to use to get the results we wanted, such as FigTree and ggtree.

References:

1. Bedford, T., Riley, S., Barr, I. G., Broor, S., Chadha, M., Cox, N. J., Daniels, R. S., Gunasekaran, C. P., Hurt, A. C., Kelso, A., Klimov, A., Lewis, N. S., Li, X., McCauley, J. W., Odagiri, T., Potdar, V., Rambaut, A., Shu, Y., Skepner, E., & Smith, D. J. (2015). Global circulation patterns of seasonal influenza viruses vary with antigenic drift. Nature, 523(7559), 217–220. https://doi.org/10.1038/nature14460.
2. Bao Y., P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman. J. Virol. 2008 Jan;82(2):596-601. https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?go=database.
3. G Yu, DK Smith, H Zhu, Y Guan, TTY Lam*. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*. 2017, 8(1):28-36. doi: 10.1111/2041-210X.12628.
4. Kosik, I., & Yewdell, J. W. (2019). Influenza Hemagglutinin and Neuraminidase: Yin–Yang Proteins Coevolving to Thwart Immunity. Viruses, 11(4), 346. https://doi.org/10.3390/v11040346.
5. Neverov, A. D., Kryazhimskiy, S., Plotkin, J. B., & Bazykin, G. A. (2015). Coordinated Evolution of Influenza A Surface Proteins. PLOS Genetics, 11(8), e1005404. https://doi.org/10.1371/journal.pgen.1005404.
6. Russell, C. A., Jones, T. C., Barr, I. G., Cox, N. J., Garten, R. J., Gregory, V., Gust, I. D., Hampson, A. W., Hay, A. J., Hurt, A. C., de Jong, J. C., Kelso, A., Klimov, A. I., Kageyama, T., Komadina, N., Lapedes, A. S., Lin, Y. P., Mosterin, A., Obuchi, M., & Odagiri, T. (2008). The Global Circulation of Seasonal Influenza A (H3N2) Viruses. Science, 320(5874), 340–346. https://doi.org/10.1126/science.1154137.
7. Sagulenko, P., Puller, V., & Neher, R. A. (2018). TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution*, *4*(1). https://doi.org/10.1093/ve/vex042.