UNIVERSITY OF CALIFORNIA

Los Angeles

Investigation of Flu Vaccination
by Age and Methylation Profile of Individuals

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Bioinformatics

by

Connor John Razma

2022

ABSTRACT OF THE THESIS

Investigation of Flu Vaccination
by Age and Methylation Profile of Individuals

by

Connor John Razma
Master of Science in Bioinformatics
University of California, Los Angeles, 2022
Professor Alexander Hoffmann, Chair

Influenza affects millions worldwide each year with responses varying from individual to individual. Influenza can be broken down into several subtypes, specifically H1N1, H3N2, Yamagata, and Victoria. One way to measure the immune response to influenza is to measure a person's antibody response to influenza. To measure how many antibodies are present in a sample, a hemagglutination inhibition assay (HAI) is used. DNA methylation is an epigenetic mechanism used to regulate gene expression in cells. Its mechanism of action is the addition of a methyl group to cytosine at a cytosine-guanine pair. DNA methylation has been shown to change in response to stimuli such as viral or bacterial infections. DNA methylation can be measured by bisulfite sequencing. In this study, data was taken from patients who had the flu vaccination. Their antibody data was measured using the HAI assay by the University of Georgia and their methylation data was measured using reduced representation bisulfite sequencing by the Pellegrini and Reed labs at UCLA. Using various statistical learning algorithms, we were able to find methylated sites that were good predictors of vaccine response. Elastic net regression proved to be a particularly good predictor of vaccine response, and after further analysis, it was revealed that the best prediction happened with relatively few methylated sites being used in prediction. Some of these significant sites seem to be involved in regulating immune response and membrane function. Further work will be done to determine the prediction accuracy of these algorithms with just these sites. Ideally, after this future work and other experiments, these methylated sites can be used as biomarkers to predict response to flu vaccination.

The thesis of Connor John Razma is approved.

Aaron Meyer

Matteo Pellegrini

Alexander Hoffmann, Committee Chair

University of California, Los Angeles

2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

The influenza virus, more commonly known as the flu, is a virus that infects humans and is normally responsible for congestion, fever, and a cough. commonly referred to as flu like symptoms. Influenza attacks the body in a similar way to most other viral infections. First, the virus enters a cell and makes its way to the nucleus of a cell. Once at the nucleus, the virus releases its RNA into the nucleus of the cell. The viral RNA inserts itself into the cell's genome and becomes an active infection. The virus uses the cell's machinery to create more viral particles which then spread out of the cell and continue to attack the body [Mog17]. Normally influenza is treated relatively easily and the patients make a full recovery. However, influenza is highly mutable and many strains have led to global pandemics that kill tens of millions of people most notably the 1919 flu pandemic. Additionally, the flu can be much more deadly in older individuals or in individuals with a compromised immune system [MVA20]. In general, it has been shown in this work and others that older individuals have a weaker immune response to infection.

Influenza is a highly mutable virus that is broken down into two subtypes Influenza A and Influenza B. From here they are broken down into multiple strains including H1N1, H3N2 for Influenza A, and Yamagata and Victoria for Influenza B. We will be studying many substrains that come from the 4 strains listed above. While there are many substrains of influenza, the yearly flu vaccine only has 4 strains it protects against. In our study, we have a vaccine with a strain from H1N1, H3N2, Yamagata, and Victoria.

The body fights influenza and other viruses through the immune system. The body uses both the adaptive immune system and the innate immune system to do this. The innate immune system takes advantage of natural killer cells and macrophages to indiscriminately attack cells that could be infected with influenza. In this paper, we will have a stronger focus on the adaptive immune system. The adaptive immune system works by using killer T-cells to target specific cells that have been infected by the virus. The T-cells find infected cells by antibodies which are produced by B-cells. These antibodies attach themselves to the virus and to cells infected by the virus, allowing the T-cells to identify them [BLW16]. For this reason, one way to measure the effectiveness of someone's immune system against a virus is to track their antibody count. There is a seasonal flu vaccine that is designed to prevent infection by influenza by prepping the adaptive immune system. This is done by presenting a weaker virus to the body and allowing the body to create its own antibodies without causing an infection.

DNA methylation is an epigenetic mechanism that regulates gene expression in cells. DNA methylation works by attaching a methyl group to a cytosine nucleic acid. This methyl group alters gene expression in the cell and is important in cell differentiation and cell response to stimuli. DNA methylation can change in response to stimuli to either up-regulate or down-regulate certain genes [SG99]. We are interested in seeing if DNA methylation can act as a predictor of immune response and if so what specific regions seem to be important in a strong immune response. Additionally, we are interested in seeing if DNA methylation changes as a result of a flu vaccination.

CHAPTER 2

Methods

Vaccines were administered at the University of Georgia to 4 different cohorts. Cohort one consisted of 149 patients during 2016, cohort two consisted of 256 patients during 2017, cohort three consisted of 260 patients during 2018, and cohort four consisted of 461 patients during 2019. Flu vaccines were administered to the patients and demographic data was recorded on age, biological sex, BMI, and ethnicity. The vaccine administered had one vaccine strain for H1N1, H3N2, Yamagata, and Victoria. Samples were also taken the day of vaccination to be used for a hemagglutination inhibition assay [CIC19] [NCA17]. Similar samples were later collected 21 days after vaccination for cohort 1 and 2 and 28 days after vaccination for cohort 3 and 4 to be used for a hemagglutination assay as well.

For each subtype of H1N1, H3N2, Victoria, and Yamagata there was HAI data on the vaccine strain used as well as additional data for other strains under each subtype. The number of substrains in each subtype varied but was generally between 5-10.

For cohort 4, blood samples were also taken at day 0 and day 28 from 62 of the individuals to be used for methylation profiling. These individuals were profiled using reduced representation bisulfite sequencing at both day 0 and day 28 [MFM22].

All demographic data and HAI data was placed in an Excel sheet and analyzed by our group. All methylation data was similarly put in an excel sheet.

All statistical work was done in R, Python, or occasionally Excel. In R, prepossessing and data transformation was performed on the original UGA 1-4 datasets. The main data transformation process was to add seroprotection and seroconversion indicators to each strain.

These indicators labeled whether a sample was seroprotected or seroconverted for a certain influenza substrain. Seroprotection is defined as an HAI score of above 40 and indicates strong immunity to a certain substrain of the virus. Seroconversion is defined as an HAI score that has quadrupled 21 days after vaccination or infection and is above 40. It indicates a strong immune response to a vaccine as well as a immunity to the virus. The seroprocetion and seorconversion indicators were added through a function that added an extra column to the data set for each substrain at day 0 and day 21/28 to indicate seroprotection status. As well as an extra column at day 21/28 for each substrain to indicate seroconversion status. The seroprotection columns were then populated with 1's and 0's to indicate seroprotection or no seroprotection respectively. These 1's and 0's were determined by going back to the original data set and doing a comparison with each sample. If the sample had an HAI score of 40 or above a, 1 was placed in the column. If it was below 40, a zero was placed. The seroconversion columns were populated with 1's and 0's to indicate seroconversion or no seroconversion status, respectively. Population was done by going to the original day 21/28 column and making sure the HAI score was above 40, and that the day 21/28 column value had a quadruple increase from the day 0 if both of these requirements were met then a 1 was placed, otherwise a 0 was placed. The addition of seroprotection and seroconversion columns allowed for more binary and classification analysis instead of just regression analysis.

Methylation analysis began with loading the data sets into Python and into dataframes through Pandas. Along with the methylation datasets, datasets with the UGA HAI data were loaded along with the seroprotection and seroconversion indicators attached. A dataset containing metadata for each individual (age, sex, BMI, ethnicity, and smoking status) was also attached. The datasets had a full inner merge performed between them to only keep the samples that had methylation data and to combine the datasets. This left only 62 samples that had methylation and the other samples from UGA 4 had to be removed. The final dataset contained metadata, HAI data, seroprotection/seroconversion status, and methylation data. The Methylation data consisted of 10,000 methylated sites. The methylation data was then scaled using the Sckit-Learn Package. [PVG11] Scaling is done by calculating the z-statistic for each site. After this, principal component analysis (PCA) was performed using the Sckit-Learn package and leaving 2 principal components which retain 12 percent of the original variance. We kept 2 principal components as that allowed for us to graph the data on a two dimensional axis. PCA is a data reduction tool for visualization that maximizes the amount of original covariance conserved in as many principal components that the user specifies. After the top 2 principal components were calculated, each sample was graphed against the 2 component's. These samples were then overlaid with their seroprotection status as well as their age status. The overlay with seroprotection status allowed us to look for separability in each strain and the overlay with age allowed me to compare seroprotection trends with age trends in the top 2 principal components. Along with overlay from actual age, I overlaid the PCA graph with age predicted by the methylation data. Methylation data can be used to predict someone's age and this was done by members of Matteo Pellegrini's group who shared the methylation age data with me. After this, the top loadings of the PCA were looked at to see if any one particular chromosome or chromosome region contributed heavily to either principal component.

After early PCA exploratory analysis, I performed a prediction analysis using the methylated sites as features and one of the influenza strains HAI score as the response variable. Prediction analysis allows us to see how much a set of independent variables (features) contribute to a dependent variable (response variable). It also allows for feature selection from the independent variables or the selection of the most important features to the response variable. In our case, we are interested in seeing how much DNA methylation contributes to the HAI score of patients. Prediction analysis was performed in Rstudio and took advantage of the dplyr and caret packages [WFH22] [Kuh08]. In order to perform prediction analysis, I chose a substrain HAI score as the predictor variable and used all the methylation sites as predictors and ran these datasets through a prediction model. The models used were elastic net Gaussian regression, elastic net Poisson regression, partial least squares regression, and random forests. Elastic net Gaussian regression is similar to normal linear regression except for the addition of regularization terms that are designed to simplify the model. Elastic net uses a structure where a L1 regularization and L2 regularization are performed and there is a tuning parameter to choose more heavily towards L1 or L2 regularization. L1 regularization removes variable's from the model where L2 regularization simply has features contribute less. Poisson elastic net is similar to Gaussian elastic net except a link function is used to convert the output from the features into a Poisson distribution. A Poisson link function was chosen because the Poisson distribution is a discrete data distribution, and since the HAI data is a count (discrete) dataset a Poisson link function is a natural choice. PLS regression is a regression that reduces the original feature space to a reduced space based on feature correlation and then performs normal linear regression on the reduced feature space. Finally, random forests is a bagging technique where many different individual decision trees are trained by the model each one trained on a randomly selected subset of the original feature space. Each of these models have hyperparameters to be tuned and they were tuned and selected for by the train function in caret through grid search using leave one out cross validation.

After performing prediction analysis, the models and hyperparamaters were observed. I then looked at the features that had the most importance in predicting a certain strain across the 4 model types. I paid particular to the most accurate models according to R^2 and mean squared error metrics. I took the top most important features from the best performing model and created a new model for predicting the strain HAI score using just those most important features. The reasoning for creating models with less features is to create simpler models that focus only on the most influential features. This allows us to see what methylation sites seem to contribute the most to the HAI score of each strain. We then can interpret what is happening biologically and make recommendations on wet lab experiments that can be performed to confirm our hypothesis. Biological interpretation was done by looking at the most important features from the variable reduced model through the UCSC genome browser. The browser gives us insight into what gene or genomic region each methylated site affects and allows us to create a biological hypothesis as well as further experiment ideas that can be implemented to confirm our hypothesis.

CHAPTER 3

Results

Principal Component Analysis was performed on the DNA methylation and overlaid with seroprotection status for each vaccine strain in accordance with the methods stated above.

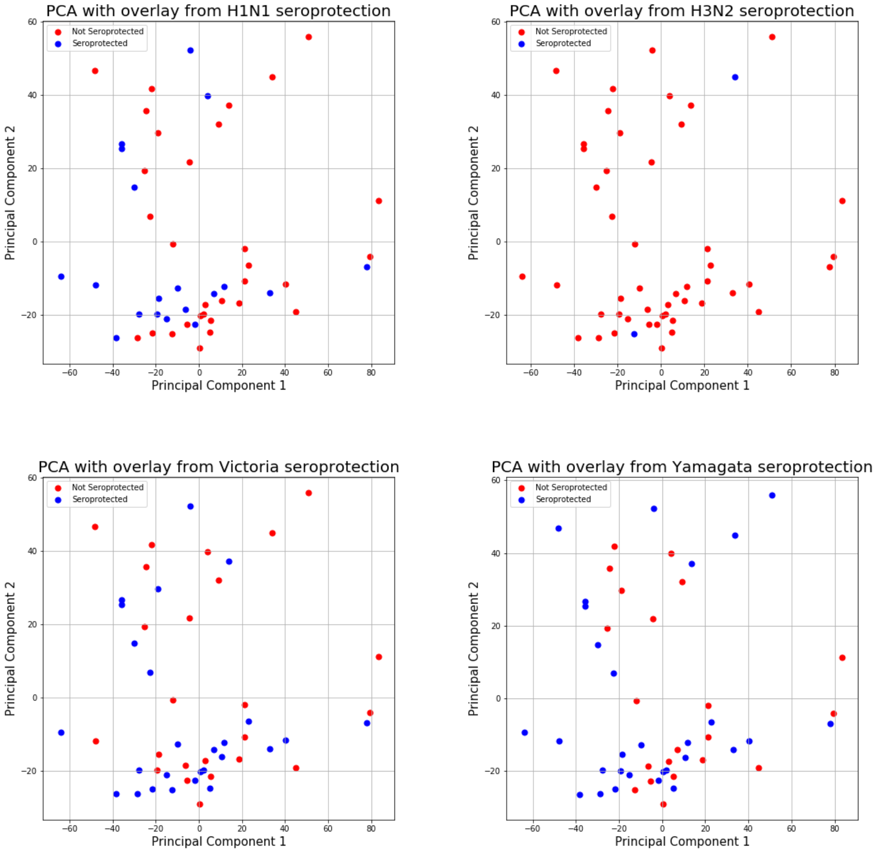

Figure 3.1: PCA Graphs

From the figure, we can see the seroprotection status across each principal component for each viral type. As can be seen there appears to be a split in the H1N1 vaccine strain between seroprotected and not seroprotected individuals, and this split occurs along Principal component 1. There are not enough seroprotected individuals in the H3N2 graph to determine if there is any split along seroprotection. In the Victoria as well as Yamagata strains there doesn't seem to be a split between seroprotected and not seroprotected individuals. This split in H1N1 implies that there may be more separability in the original dataset between seroprotected individuals and not seroprotected individuals. The separability may imply that there is a stronger effect that DNA methylation has on HAI.

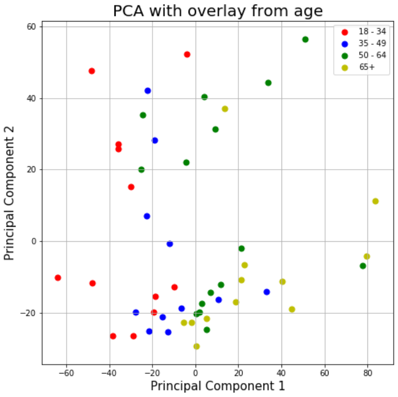

Figure 3.2: PCA Age Graph

Looking at the figure, we see that principal component 1 seems to follow an age related trend, with younger individuals on the left side of principal component 1 and older individuals towards the right of principal component 1. By comparing this to the PCA graph with H1N1 overlaid, we see that both the trends are similar implying that as individuals get older they have a weaker immunity to H1N1.

After this initial exploratory analysis, I fit regression models to the data as detailed in the methods section with methylation data acting as the independent features and HAI score acting as the dependent feature.

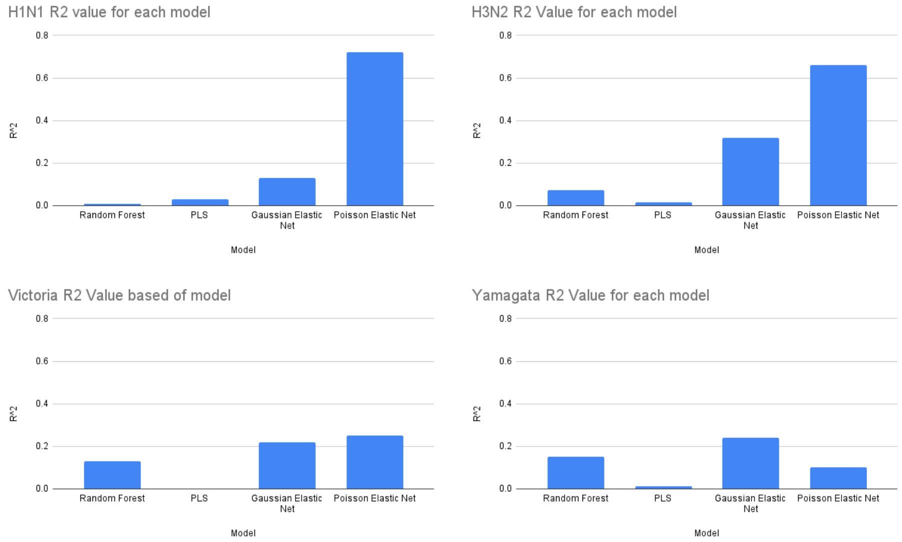

Figure 3.3: Performance Metrics

The figures above show the best top performing models validation R^2 score for each flu type.

As can be seen from the figures above, the best performing model was Poisson Elastic Net for all subtypes except Yamagata. Poisson Elastic net having the most accurate performance can be expected because the original HAI data follows a count distribution class and Poisson Elastic Net accounts for its dependent variable following a Poisson distribution which is a discrete probability distribution. We also see that generally the highest performance in the models comes from H1N1 and this can be explained by H1N1's higher separability discussed above.

Now we can look at some of the most important features that were extracted from each model.

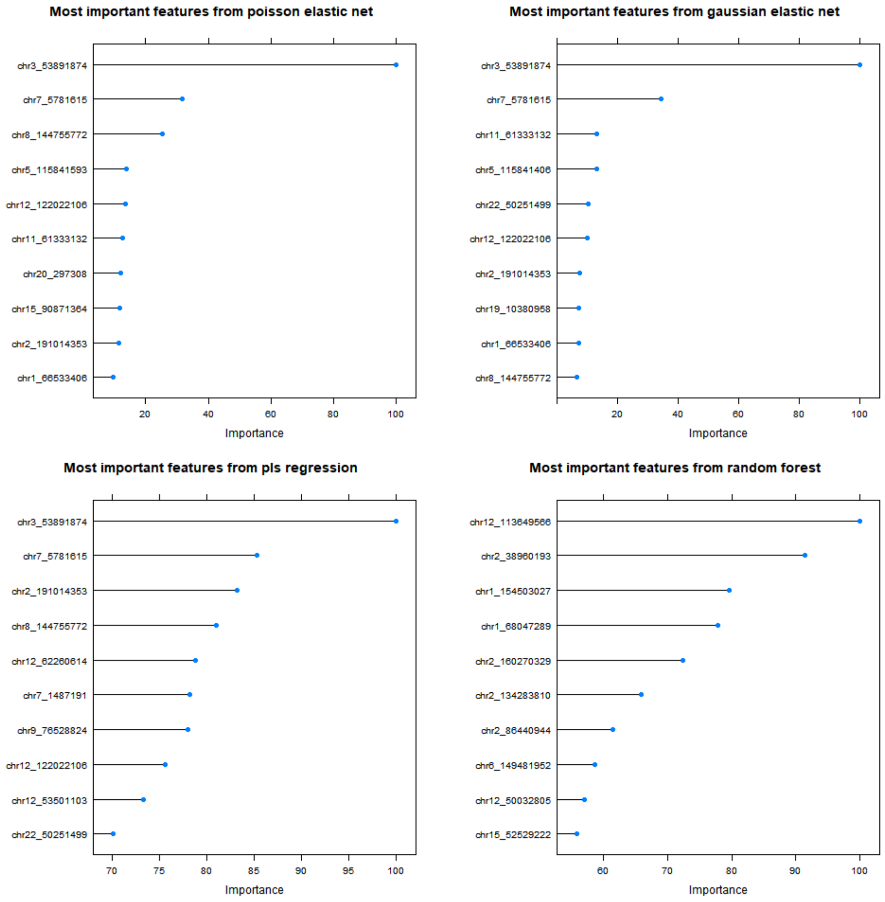

Figure 3.4: Feature Importance

Above are the graphs for the feature importance for each model applied on methylation data predicting H1N1 HAI scores. Feature importance for Poisson elastic net, Gaussian elastic net, and PLS was calculated by normalizing the absolute value of the t-statistic for each model. Feature importance for the random forest was calculated by going to each tree and calculating the accuracy of the tree, then calculating the accuracy after permuting each predictor variable. The difference between the two is calculated for each tree, averaged out, and finally normalized to get the feature importance value.

As can be seen, the two most important chromosome sites according to the Poisson elastic net, Gaussian elastic net, and PLS regression are Chr3-53891847 and Chr7-5781615.

According to the UCSC genome browser [KBD03], the site on chromosome 3 is in an exon region of the human gene Selenok. Selenok is a gene that encodes a transmembrane protein that has been shown to be involved in promoting Ca(2+) flux in immune cells. This implies that methylation on this gene affects people's innate immunity to H1N1 and the gene should be studied further. Again, according to the UCSC genome browser, the site on chromosome 7 is part of an exon region on human gene RNF216. This gene encodes for a protein that interacts with the serine/threonine protein kinase and can also act as a E3 ubiquitin ligase. Again, further analysis should be done on this site and chromosome.

Now since Poisson elastic net seemed to perform the best we can take a look into the structure of Poisson elastic net.

$L(\hat{\beta}) = \frac{\sum_{i=1}^{n}(y_i - y_i log(x_i^T \hat{\beta}))}{2n} + \lambda \frac{1-\alpha}{2} \sum_{j=1}^{m}(\hat{\beta}^2) + \alpha \sum_{j=1}^{m} |\beta|$

Above is the loss function for an Poisson elastic net this function is the normal Poisson regression loss function $\frac{\sum_{i=1}^{n}(y_i - y_i log(x_i^T \hat{\beta}))}{2n}$ plus a regularization parameter $\lambda \frac{1-\alpha}{2} \sum_{j=1}^{m}(\hat{\beta}^2) + \alpha \sum_{j=1}^{m} |\beta|$. This regularization parameter and its hyperparameters are important for interpretation of our model. The two hyperparameters of interest are α and λ. λ controls over all how much regularization there is i.e a higher value means more regularization. α controls the type of regularization with a value of 1 being full L1 regularization and a value of 0 being full L2 regularization. A high amount of L1 regularization will lead to the elimination of many individual features and this implies strong correlation between features.

Alpha Value	Lambda Value	R^2 Value
0.1	1.18	0.041
0.8	1.18	0.028
0.8	33.5	0.73

Table 3.1: Model Comparison

Above is a table showing a selected list of Poisson elastic net models predicting H1N1

HAI scores R^2 values and their hyperparameters from the hyperparameter tuning stage. As can be seen, the most accurate model had a high lambda value (33.5) and a high alpha value (0.8). This implies that our model performs best when there is a lot of regularization and when that regularization is using a L1 regularization method. Taken together, this implies large correlation between prediction features and only a couple of true features contributing to the HAI score at day 0 for H1N1. Further work should be done to narrow down our features and find the ones contributing to day 0 immunity.

CHAPTER 4

Conclusion

These results show a general trend of higher separability among H1N1 patients that is not seen in the other influenza subtypes. This separability also follows a similar trend for age. From this, we see that day 0 immunity for H1N1 is affected by age, but age does not affect immunity for the other subtypes. After running regression analysis on the different subtypes, we see that the most accurate method is Poisson elastic net. Specifically, it is Poisson elastic net with a high alpha and lambda value. Taken together, this shows our predicted variable (HAI score) follows a count distribution and our features have correlation among them. More work should be done to narrow down the features space to find the features actually contributing to day 0 immunity biologically. Finally, some of the most important features seem to be involved in genes that have been shown to be active in the immune system from previous literature. After narrowing down the feature space, experiments can be done on these sites to determine the exact contribution they have to day 0 immunity. There is more analysis that can be done on these datasets with the generation of new data. Methylation data from day 28 of our cohort could allow for a similar regression analysis on day 28 patients HAI scores. Additionally, longitudinal analysis could be done between day 0 methylation and day 28 methylation. This would give insight into how methylation is changing in response to the influenza vaccine, and if patients who have a stronger immune response have a different methylation trajectory. Finally with a larger cohort, epigenome wide association analysis (EWAS) could be done on this type of data to confirm the function of important sites found earlier, and discover new sites

REFERENCES

[BLW16] Azadeh Bahadoran, Sau H Lee, Seok M Wang, Rishya Manikam, Jayakumar Rajarajeswaran, Chandramathi S Raju, and Shamala D Sekaran. "Immune responses to influenza virus and its correlation to age and inherited factors." *Frontiers in microbiology*, **7**:1841, 2016.

[CIC19] Michael A Carlock, John G Ingram, Emily F Clutter, Noah C Cecil, Moti Ramgopal, Richard K Zimmerman, William Warren, Harry Kleanthous, and Ted M Ross. "Impact of age and pre-existing immunity on the induction of human antibody responses against influenza B viruses." *Human Vaccines & Immunotherapeutics*, **15**(9):2030–2043, 2019.

[KBD03] Donna Karolchik, Robert Baertsch, Mark Diekhans, Terrence S Furey, Angie Hinrichs, YT Lu, Krishna M Roskin, Matthias Schwartz, Charles W Sugnet, Daryl J Thomas, et al. "The UCSC genome browser database." *Nucleic acids research*, **31**(1):51–54, 2003.

[Kuh08] Max Kuhn. "Building Predictive Models in R Using the caret Package." *Journal of Statistical Software, Articles*, **28**(5):1–26, 2008.

[MFM22] Marco Morselli, Colin Farrell, Dennis Montoya, Tarık Gören, Ramazan Sabırlı, İbrahim Türkçüer, Özgür Kurt, Aylin Köseler, and Matteo Pellegrini. "DNA methylation profiles in pneumonia patients reflect changes in cell types and pneumonia severity." *Epigenetics*, pp. 1–15, 2022.

[Mog17] Mohsen Moghadami. "A narrative review of influenza: a seasonal and pandemic disease." *Iranian journal of medical sciences*, **42**(1):2, 2017.

[MVA20] Janet E McElhaney, Chris P Verschoor, Melissa K Andrew, Laura Haynes, George A Kuchel, and Graham Pawelec. "The immune response to influenza in older humans: beyond immune senescence." *Immunity & Ageing*, **17**(1):1–10, 2020.

[NCA17] Ivette A Nuñez, Michael A Carlock, James D Allen, Simon O Owino, Krissy K Moehling, Patricia Nowalk, Michael Susick, Kensington Diagle, Kristen Sweeney, Sophia Mundle, et al. "Impact of age and pre-existing influenza immune responses in humans receiving split inactivated influenza vaccine on the induction of the breadth of antibodies to influenza A strains." *PLoS One*, **12**(11):e0185666, 2017.

[PVG11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, **12**:2825–2830, 2011.

[SG99] Rakesh Singal and Gordon D Ginder. "DNA methylation." *Blood, The Journal of the American Society of Hematology*, **93**(12):4059–4070, 1999.

[WFH22] Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2022. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.