Correlate Analyses for Acutely Infected HIV Individuals: Progressors vs. Controllers

A Major Qualifying Project submitted to the Faculty of
WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of requirements for the degree of Bachelor of Science in Bioinformatics and Computational Biology

Written By:
Kylie Dickinson (BCB)
April 25, 2019

Approved by:
Elizabeth Ryder, PhD – Advisor

Advised by:
Elizabeth Ryder, PhD – Advisor (WPI)
Jishnu Das, PhD – Advisor (The Ragon Institute)

ABSTRACT

The human immunodeficiency virus (HIV) infects approximately 5,000 new people daily. Of people infected, only a minor population act as disease controllers. The goal of this project was to determine which humoral factors are correlated with HIV controllers and progressors. Through the use of multivariate and univariate statistical and exploratory data analyses, the two sub-populations of HIV-infected individuals were compared using observed HIV-specific IgG subclasses from subjects' plasma over four time points during the acute stage of infection. Results showed the multivariate approach was more effective in differentiating controllers from progressors, suggesting no single IgG subclass is predictive of disease control.

TABLE OF CONTENTS

1. INTRODUCTION

A successful vaccine must be designed and administered in order to terminate the spread of the human immunodeficiency virus (HIV). Although new cases of HIV infections have dropped by approximately thirty six percent over the past eighteen years, the virus was still the cause of death for over 900,000 people in 2017 (WHO 2018). The large-scale infection and mortality rates of this infection have decreased through the use of drug therapies and widespread educational campaigns; however, those statistics could be lowered further through the discovery of a successful vaccination.

Other diseases, such as small pox and polio, have been eradicated through the use of vaccinations, but HIV has unique characteristics which make traditional methods in vaccination production difficult (Banerjee, & Mukhopadhyay, 2016). Since traditional methods will not work, new approaches are being established by research institutions. The traditional methods for vaccine trials involve measuring antibody levels through titers and neutralization, and though these measurements have been previously successful, they are not the best indicators of protection among HIV subjects. Instead of only accounting for these factors, researchers have gone a step further by including the functional components of antibodies as correlates in vaccine trials. These components could include pathogen life cycle and a pathogen's structural components (Arnold & Chung, 2017).

This strategy of examining multiple components of antibody function and humoral immunity features has been termed "Systems Serology" (Ackerman, Barouch, & Alter, 2017). This approach allows for a more extensive array of parameters to be collected about antibody function and structure which can be viewed as a unique profile for each antibody; which also means high dimensional datasets will be compiled. The challenge with the high dimensional dataset is finding the best way to analyze that data. One solution is to further analyze the antibody profiles through machine learning and other statistical tools (Ackerman, Barouch, & Alter, 2017). Research for this project analyzed system serology datasets containing both biophysical assay and functional assay information in order to further study antibody profiles of progressors and controllers. Progressors of HIV are individuals in which the virus can take a typical infectious course and eventually become AIDS, where as non-progressors or controllers have a natural immunity to the infection and are able to live fairly normal lives with low viral loads after initial infection and without any treatment.

Programming tools, specifically MatLab and Python, were used to analyze the high throughput datasets studied in this project. The first dataset, from a non-human primate vaccine study, was used to become familiar with various statistical techniques. It contained 25 features including glycan levels from a biophysical assay (20) and antibody effector functions (5) from functional assays for 34 non-human primates from two different vaccine trials. This dataset was analyzed to try and observe the correlation of antibody glycan features and antibody functions to protection or infection after administering a vaccination. Results for this dataset are still being analyzed as signal for the multivariate analysis was extremely low. The second dataset, which was previously studied in an earlier paper

(Sadanand, et al., 2018), contained measured levels of HIV-specific IgG subclass titers and multiple antibody-dependent effector functions as part of antibody profiles for nineteen acutely infected individuals. The levels for these features were measured at 4, 12, 24, and 48 week time points after initial infection. After the study, ten of the individuals were deemed progressors and nine were deemed natural controllers of the virus. A high-dimensional dataset was then constructed and analyzed using similar machine learning and univariate exploratory analysis techniques as the first analyzed dataset. The second dataset was analyzed with the goal of discovering key antibody differences in progressors and controllers on individuals who have yet to seek treatment in order to better understand controllers' natural "immunity" to the virus.

Results from the different analyses had similar conclusive observations. The first dataset had weak signals for the multivariate approach yet had multiple features observed for univariate Mann Whitney U tests at low p-values. The second dataset, which was previously analyzed picked up on similar features as the original paper (Sadanand, et al., 2018) while adding one other antibody effector function as a correlated feature to predicting progressor versus controller status. A univariate analysis extended the previously published results from the paper by carefully looking at changes in biophysical and functional antibody features over time. This exploratory analysis revealed much less variability in the features of controllers than that of progressors. This may suggest that keeping several antibody features contained within a controlled range of values is important to disease control.

The goal of this project was to examine antibody profiles of subjects infected with HIV through means of 'machine learning' multivariate analyses, univariate analyses, and exploratory analyses, in order to discover how the various humoral factors may correlate to natural disease control, and how those insights may help lead to new discoveries for vaccine design.

2. BACKGROUND

The Immune System

On a daily basis, the human body's natural defense mechanism, the immune system, comes into contact with millions of potential pathogens (Alberts et al, 2002). It battles these pathogens in complex processes to prevent infection of various microbes that could cause deadly illnesses. The two subgroups of the immune system include the innate and adaptive immune systems. Both are utilized in order to protect the body. When potential infection toxin comes into contact with the body, the first step is for the innate immune system to respond quickly to try and slow down the pathogen's infective rate. This initial response then allows other immune responses to become activated (Nicholson, 2016). The adaptive immune system works to fight pathogens in a more targeted approach. Utilizing both B-cells and antibodies for humoral or antibody-mediated responses, and T-cells for cell-mediated responses, the adaptive immune system fights infection in a more specific manner than that of the innate immune system (Clem, 2011). The two subclasses of the immune system are further depicted in Figure 2.1. Research has been, and

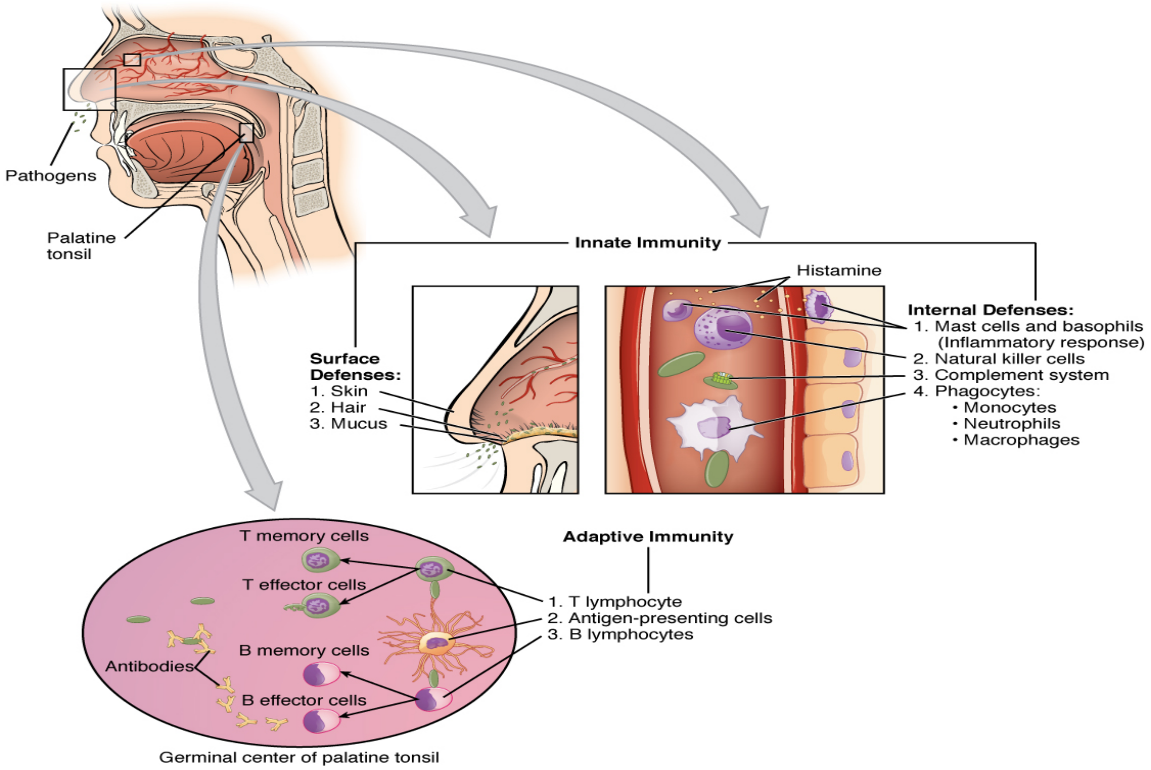


Figure 2.1: This diagram divides the innate and adaptive immunity mechanisms (Betts, 2019).

continues to be, conducted on all forms of the immune system, in order to move towards treatments and protection against infectious diseases. One consistently studied form of treatment is vaccination. Vaccines work to engage and train the adaptive immune system to respond to potentially fatal pathogens.

The first successfully recorded vaccine was created by Edward Jenner in 1798 when he used cowpox to inoculate and prevent smallpox in humans (Clem A. S., 2011). Since then, consistent research and discoveries have been made on how and why vaccines are both cost-effective and successful in protecting against pathogens (Pulendran & Ahmed, 2011). Today, vaccine trials are being tested in order to find preventative measures against some of the most difficult diseases and illnesses to battle such as the human immunodeficiency virus (HIV).

The Human Immunodeficiency Virus (HIV)

The world suffered an unprecedented HIV outbreak in 1981, which caused scientists to research treatments and vaccinations for the virus (German Advisory, 2016). For the past 30 years, information about the virus has been discovered but there is still no cure due to how the virus affects the immune system. HIV is a retrovirus; it is able to integrate a DNA copy of its RNA genome into its host cell genome. The virus targets the body's CD4 immune cells, which are part of the adaptive immune system. Figure 2.2 shows the structure of HIV as Figure 2.3 displays the process of cellular infection, integration of genetic material, and production of new viral proteins.

The process of HIV infection begins with the viral capsid of the virus entering the host cell. To enter the cell, the virus must first attach through a receptor to the host cell. The HIV protein envelope (spike), comprised of mature surface glycoprotein 120 and transmembrane glcoprotein41 (gp120 and gp41), attaches to the CD4 receptor of the host cell, causing a conformational change in the HIV protein envelope and fusion of the viral and immune cell membranes (German Advisory, 2016). This fusion permits genetic material within a capsid contained in the virus to be transferred into the host cell. Enzymes and genetic material from the capsid are then used to integrate the viral genetic material into the host cell's DNA (German Advisory, 2016). Reverse transcriptase creates viral DNA and then integrase is used to integrate that genetic material into the cell's DNA. The integrated viral DNA will produce viral transcripts and proteins. The viral immature proteins and RNA genome then bud from the cell with the help of viral protease (German Advisory, 2016). Once out of the cell, the viral particle will mature into an infectious virion.

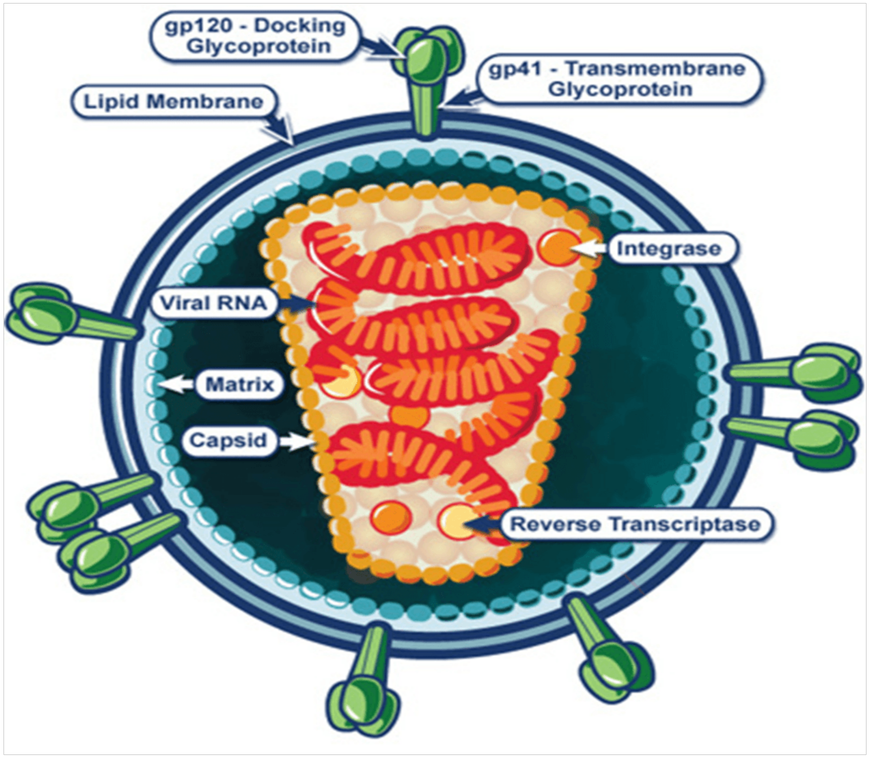


Figure 2.2: Simplified depiction of HIV. The spike shows the glycoproteins 120 and 41 located on the viral membrane which will help with attachment and fusion of the viral membrane to the protein receptors of the target host CD4 cell (Betts, 2019).

Importance and Roles of Glycoproteins

The goal of a virus is to infect host cells. For HIV, infecting the host cell is only possible through the use of its glycoprotein envelope. Glycoproteins surrounding the virus allow for viral and cell binding, fusing, and DNA integration (Checkley, Luttge & Freed, 2011). The specific glycoproteins associated with HIV are gp160, gp120, gp41 and gp140. During infection, the gp160 unit is cleaved into the mature surface gp120 and transmembrane gp41 (Kwong, 1998). The gp140 unit is derived from the gp160 unit by removing the transmembrane and cytoplasmic domains so that the glycoprotein becomes soluble. The gp140 unit is treated similarly to the gp160 unit and can be cleaved into stable trimer forms to produce gp120 and gp41 units (Khattar, Samal, LaBranche, Montefiori, Collins, et al., 2014). Gp120 allows binding to occur between host CD4 cells and the virus. Gp41 supports gp120 in infecting the host cell by aiding in the fusion of the membrane (Banerjee, & Mukhopadhyay, 2016).

HIV and The Body's Response

Once infected by HIV, the body goes through phases at which different symptoms are observed. These stages include acute infection, clinical latency infection, and eventually AIDS. Acute infection is the stage from initial infection up to clinical latency where viral load peaks and CD4 cell counts decrease. This acute infection sees such decreases in immune cells and increases in viral load since the virus is actively infecting the immune cells (Kwong, 1998). Symptoms similar to that of the flu are often associated with the acute infection stage. Clinical latency is defined as the stage at which HIV is latent or dormant and infected individuals are asymptomatic. This stage is often maintained with antiretroviral therapies (ART) (Sadanand et al., 2018). During clinical latency, cells are still being infected but not as rapidly as in the acute infection stage. Eventually, the immune system becomes too compromised to remain at clinical latency and the disease will progress to AIDS. Figure 2.4 shows how levels of HIV RNA and CD4 cells vary at each stage of an HIV infection.

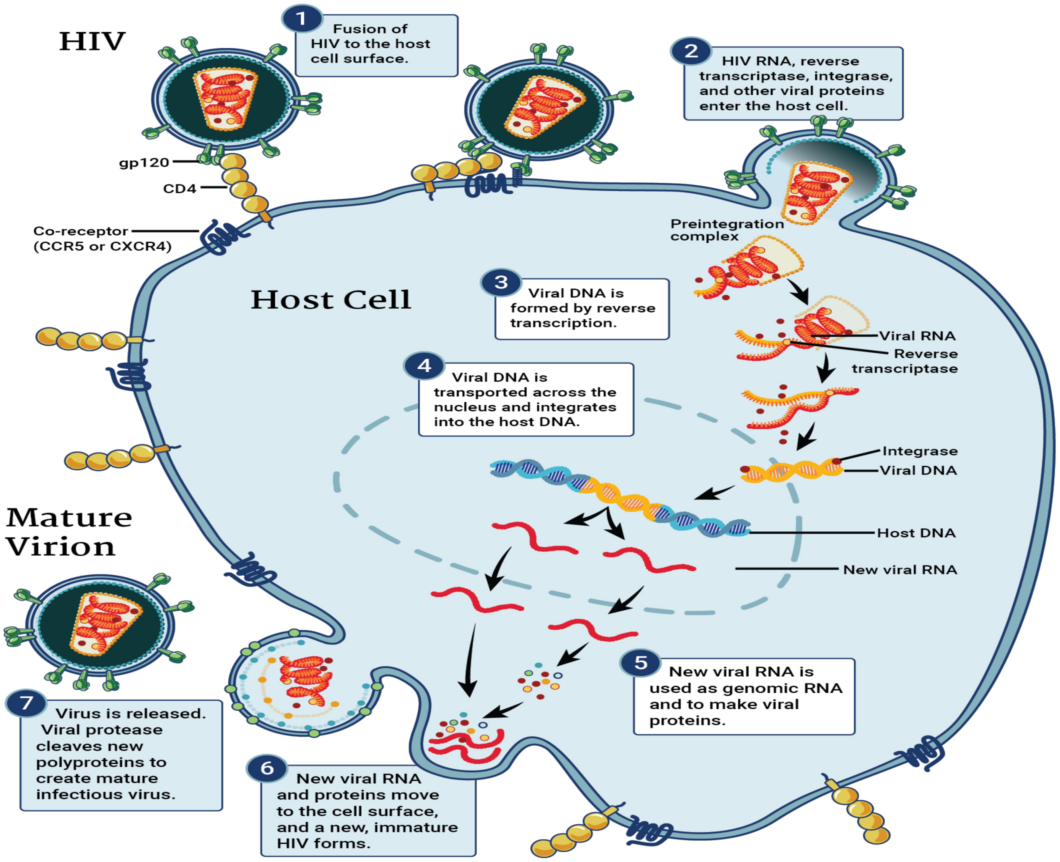

HIV
1 Fusion of HIV to the host cell surface.
2 HIV RNA, reverse transcriptase, integrase, and other viral proteins enter the host cell.
gp120
CD4
Co-receptor (CCR5 or CXCR4)
Preintegration complex
Host Cell
3 Viral DNA is formed by reverse transcription.
Viral RNA
Reverse transcriptase
4 Viral DNA is transported across the nucleus and integrates into the host DNA.
Integrase
Viral DNA
Host DNA
New viral RNA
Mature Virion
5 New viral RNA is used as genomic RNA and to make viral proteins.
7 Virus is released. Viral protease cleaves new polyproteins to create mature infectious virus.
6 New viral RNA and proteins move to the cell surface, and a new, immature HIV forms.

Figure 2.3: The diagram shows the process in which HIV infects CD4 cells within the body. (1) Binding, (2) Penetration, (3) Reverse transcriptase creates viral DNA, (4) Integrase allows for the viral DNA to integrate into the host cell's genetic material, (5) Viral RNA is now used to create viral proteins, (6) Immature viral proteins reach the surface of the host cell, (7) The virus is released from the host cell and begins to mature into an infectious virus (Betts, 2019).

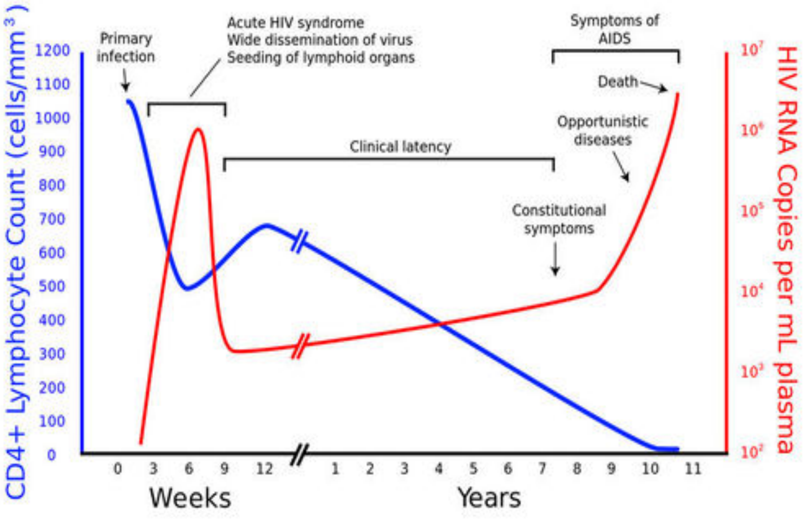


Figure 2.4: The stages of HIV. The blue lines show levels of the body's CD4 cells and the red shows levels of HIV RNA copies within the blood (Bhatti, Usman, & Kandi,

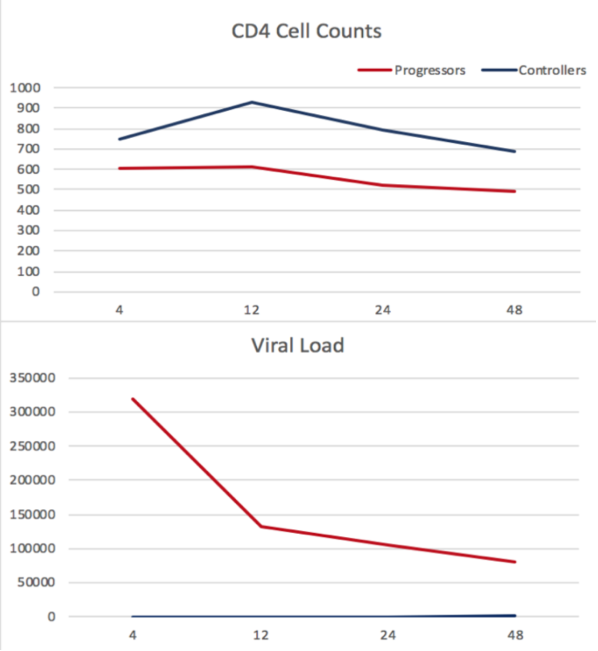


Figure 2.5: CD4 counts and viral loads from the second and main dataset analyzed throughout this project.

Typically, infected individuals will follow the previously described stages and these people are known as disease progressors. Interestingly, a small subpopulation of infected individuals are known as non-progressors or controllers. These people naturally control the infection and live for many years with low viral loads (the red line in Figure 2.4) even without antiretroviral therapy(ART) treatments (Sadanand et al., 2018). Figure 2.5 shows the viral loads and CD4 T Cell counts for subjects from the dataset used for this project. Researchers are using various techniques in order to better understand what determines controller and progressor status. Researchers have also found new insights that non-neutralizing antibodies may be factors in controlling the virus (Sadanand et al., 2018).

Nonneutralizing Antibodies and Antibody Effector Functions

Researchers previously have focused on neutralizing antibodies as key determinants in HIV protection. These neutralizing antibodies begin to develop months after initial infection, which could mean their role in initial control of the disease is limited (Sadanand et al., 2018). Studying non-neutralizing antibodies may help better understand early stages regarding early viral replication. These HIV-specific non-neutralizing antibodies have been found to induce activation of macrophages and natural killer cells (Margolis, Koup, and Ferrari, 2017). In addition, nonneutralizing antibodies are also capable of protecting against viral infection by processes such as phagocytosis and mediating virolysis. This is due to the fact that nonneutralizing antibodies are able to activate different immune complement factors and engage Fc-receptor cells, or cells that are capable of antibody reception (Mayr, Su, & Moog, 2017). For HIV, these cells are induced by the Fc regions of various antibody subclasses, where some subclasses hold higher affinities for the receptors such as IgG1 and IgG3 (Overbaugh, & Morris, 2012). These subclasses are more capable of inducing responses of antibody-dependent (AD) effector functions such as antibody-dependent cellular viral inhibition (ADCVI) and antibody-dependent cellular cytotoxicity (ADCC) (Margolis, Koup, and Ferrari, 2017). These responses vary from other infection responses where IgG2 is integral as a controlling factor (Overbaugh, & Morris, 2012). The HIV-specific IgG non-neutralizing antibodies used in this analysis can be seen in Table 2.1. In order to find a more effective treatment and potentially a cure for HIV, it is important to understand how HIV-envelope specific IgGs and AD effector functions correlate to HIV infected controllers natural immunity.

Table 2.1: Four subclasses of IgG specific to three glycoproteins associated with HIV infection were measured to create the antibody profiles for the dataset. This is in addition to the six antibody effector functions (Sadanand et al., 2018).

Gp120 IgG1	**Gp120 IgG2**	**Gp120 IgG3**	**Gp120 IgG4**
Gp41 IgG1	**Gp41 IgG2**	**Gp41 IgG3**	**Gp41 IgG4**
Gp140 IgG1	**Gp140 IgG2**	**Gp140 IgG3**	**Gp140 IgG4**

Systems Serology Approaches

Researchers have created a new approach to studying HIV infection, vaccines, and antibody responses called 'systems serology' (Ackerman, Barouch, & Alter, 2017). This new approach to understanding humoral immune responses and their correlation to protection against viral infections helps to analyze high-throughput datasets containing unique antibody profiles composed of both biophysical and functional antibody assay measurements (Arnold, & Chung, 2017). The goal of systems serology is to quantitatively understand correlations between the various antibody features both biophysical and functional, and the clinical outcomes of a study. In order to research and understand these high-dimensional datasets, or datasets which contain more features than it has subjects, of antibody profiles, different statistical and computational approaches must also be implemented.

Computational Components

Large datasets require data driven computational techniques in order to be properly analyzed. Systems serology utilizes "machine learning" computational approaches in order to better understand correlations between antibody profiles (Ackerman, Barouch, & Alter, 2017), antibody dependent functions, and outcomes such as protection due to vaccine or, as in the case of this study, a controller's ability to have natural immunity to the virus. For this study, supervised multivariate approaches were used, meaning the outcomes of the data, in this case, viral progressor versus viral controller status, was used to create a model that helped select features that may correlate to viral control [17]. Various supervised learning approaches can be used to help determine associations of features and outcomes of large datasets such as LASSO (least-absolute square shrinkage operator) or PLSDA (partial least squares discriminant analysis). The methodology of this report explains how the LASSO model was utilized as part of the machine learning model analysis for this project. These machine learning techniques are able to highlight features from high-dimensional datasets that are most associated with the dataset outcomes. Due to the advancements made with systems serology and its associated machine learning computational techniques, deeper understandings of the humoral immune system continue to be found.

3. METHODOLOGY

Various statistical approaches and tools were used in order to analyze two datasets. The first dataset, the nonhuman primate vaccine dataset, held information about 34 nonhuman primates, each with 25 associated glycan counts. This dataset showed little to no signal in the data when trying to analyze using various correlate analyses and is currently in the process of being reevaluated with other approaches. Due to the low signal in the analyses, the first dataset was not further analyzed over the duration of this project. The second dataset, referred to as the natural immunity dataset, containing information from four time points throughout the acute stages of infection for 19 subjects (10 of which were progressors of HIV and 9 which were deemed controllers), was analyzed using a variety of analyses both multivariate, univariate, and exploratory, in order to observe antibody features that correlate to disease controllers and their ability to have natural immunity.

Data Organization and Preprocessing

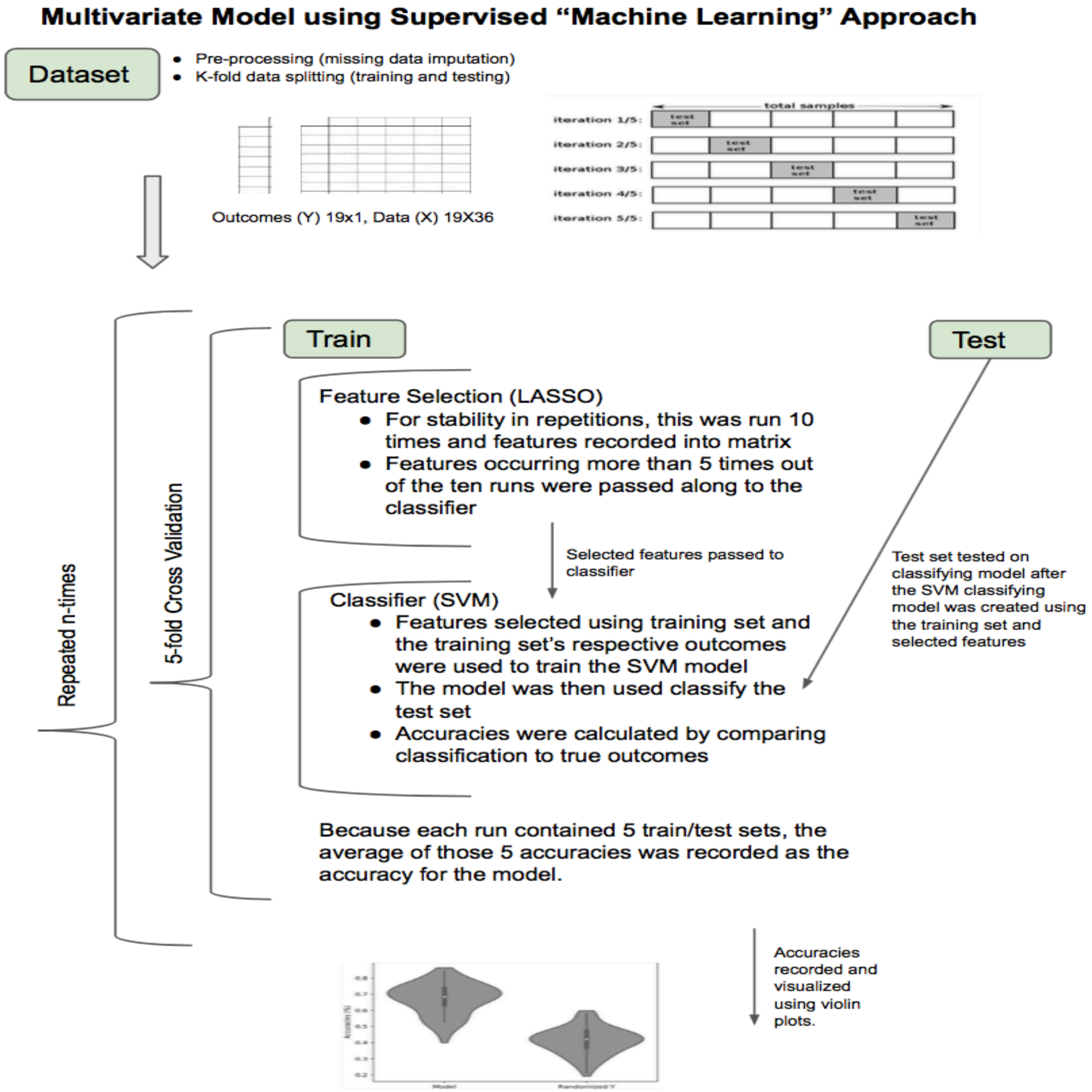
The natural immunity dataset was first retrieved as an excel sheet of size 76x20 (19 subjects with features at four different time points). Viral load and CD4 cell counts were then excluded from the feature dataset to avoid losing signal in later analyses of other antibody features that could correlate to disease progression and natural viral control. The dataset was then transformed so that each subject had features at the four time points, or 76 features. This created a dataset with dimensions 19x76, making the dataset high dimensional and thus requiring analysis by a technique suited for high dimensional data such as LASSO. The dataset was further pre-processed with MatLab to impute any missing data in order to avoid any run-time errors that could arise when coding statistical models and tests in both MatLab and Python. Missing data was imputed using the KNN-impute built-in MatLab method using the Euclidean distance setting. The dataset was then analyzed using the multivariate model and univariate exploratory analyses techniques.

Variate Testing

The original paper laid out many approaches in order to analyze the dataset. The first step to this study was to replicate and try to reproduce the multivariate model results. Then, from those results, further questions arose and were analyzed using univariate exploratory approaches.

Least Absolute Shrinkage and Selection Operator and Support Vector Machine Model

A regression model was created to predict which features were most correlated to an infected individual’ s ability to obtain natural viral control. The dataset was manipulated similarly to that of the original paper (Sadanand et al., 2018) by averaging early (weeks 4 and 12) and late (weeks 24 and 48) time points, so that the dataset had dimensions 19x36. The model consisted of a



- Pre-processing (missing data imputation)
- K-fold data splitting (training and testing)

Feature Selection (LASSO)

- For stability in repetitions, this was run 10 times and features recorded into matrix
- Features occurring more than 5 times out of the ten runs were passed along to the classifier

Selected features passed to classifier

Classifier (SVM)

- Features selected using training set and the training set's respective outcomes were used to train the SVM model
- The model was then used classify the test set
- Accuracies were calculated by comparing classification to true outcomes

Test set tested on classifying model after the SVM classifying model was created using the training set and selected features

Because each run contained 5 train/test sets, the average of those 5 accuracies was recorded as the accuracy for the model.

Accuracies recorded and visualized using violin plots.

Figure 3.1: The graphic displays how the multivariate model was designed to be used to help determine features that correlated to viral controller versus progressor status of infected individuals.

LASSO regression feature selector and an SVM classifier. The model was programmed using MatLab. The built-in LASSO function was optimized and used to select features from the dataset. LASSO works to find a subset of features which can accurately predict outcomes (natural control versus progression) for the entire dataset. LASSO reduces the dimensionality of the dataset by adding a constraint to all features which forces a majority of features to become zero. The remaining non-zero features are then used in a classification model that will try to predict the outcomes of the data without having to use the entire dataset. The LASSO feature selector also utilized a stability component in order to make sure a similar number of features were selected with each LASSO iteration. This was done by running the LASSO model ten times and recording which features were selected with each run, and then using only the features selected five or more times as the LASSOs final selected features that would be used in a model for classification. The LASSO feature selector was paired with an SVM classifier. The SVM classifier was programmed using the built-in SVM fit function in Matlab. This was customized using a linear kernel function. The results from the classifier were then compared to that of the true outcomes during each fold of the cross validation to produce a series of percentages that represented accuracies. The five accuracies were then averaged to produce the model's accuracy. This process was repeated 100 times and plotted using a violin plot.

To understand if this model's accuracy was more successful than that of a random model, a permutated model was created. This used the same dataset; however, the outcomes (progressors versus controller status) were randomized so that they did not match the correct HIV infected subject. Figure 3.1 illustrates how the model was designed.

Mann-Whitney U Testing

Non-parametric univariate statistical testing was also used to analyze the natural immunity dataset. A univariate Mann-Whitney U Test was performed using MatLab . This approach was used to observe variations between progressors and controllers for each feature at the averaged early and late time points as in the multivariate analysis. Because this was an exploratory analysis, the Mann-Whitney U test threshold for the p-value was set to 0.1, and no correction was made for multiple comparisons.

Visualizations

Visualizations helped provide further understanding of the multivariate and univariate analyses. The visuals used in this analysis included heat maps, violin plots, line graphs, boxplots, and paired boxplots. These were all important to observing patterns in the changes of antibody features over the four time points. For example, violin plots were used to observe the accuracies in the multivariate model, as boxplots were used to better understand variations of features among progressors and controllers at early and late time points. All of the visuals were constructed using Python plotting and graphing tools.

4. RESULTS

Using multiple statistical and computational approaches, the natural immunity dataset was analyzed. This dataset's analyses helped support information from the original paper (Sadanand et al., 2018). as well as bring insight on other correlates potentially involved with the determination of controller versus progressor status for infected individuals.

Initial Data Analysis

The dataset analyzed contained antibody profiles consisting of HIV specific antibody IgG subclass counts and antibody effector function counts recorded at 4, 12, 24, and 48 week time points after initial infection for 19 subjects for which ten were determined to be progressors and nine were determined to be controllers. The dataset was z-scored and visualized in a heat map as seen in Figure 4.1. This heat map visualizes how levels of the

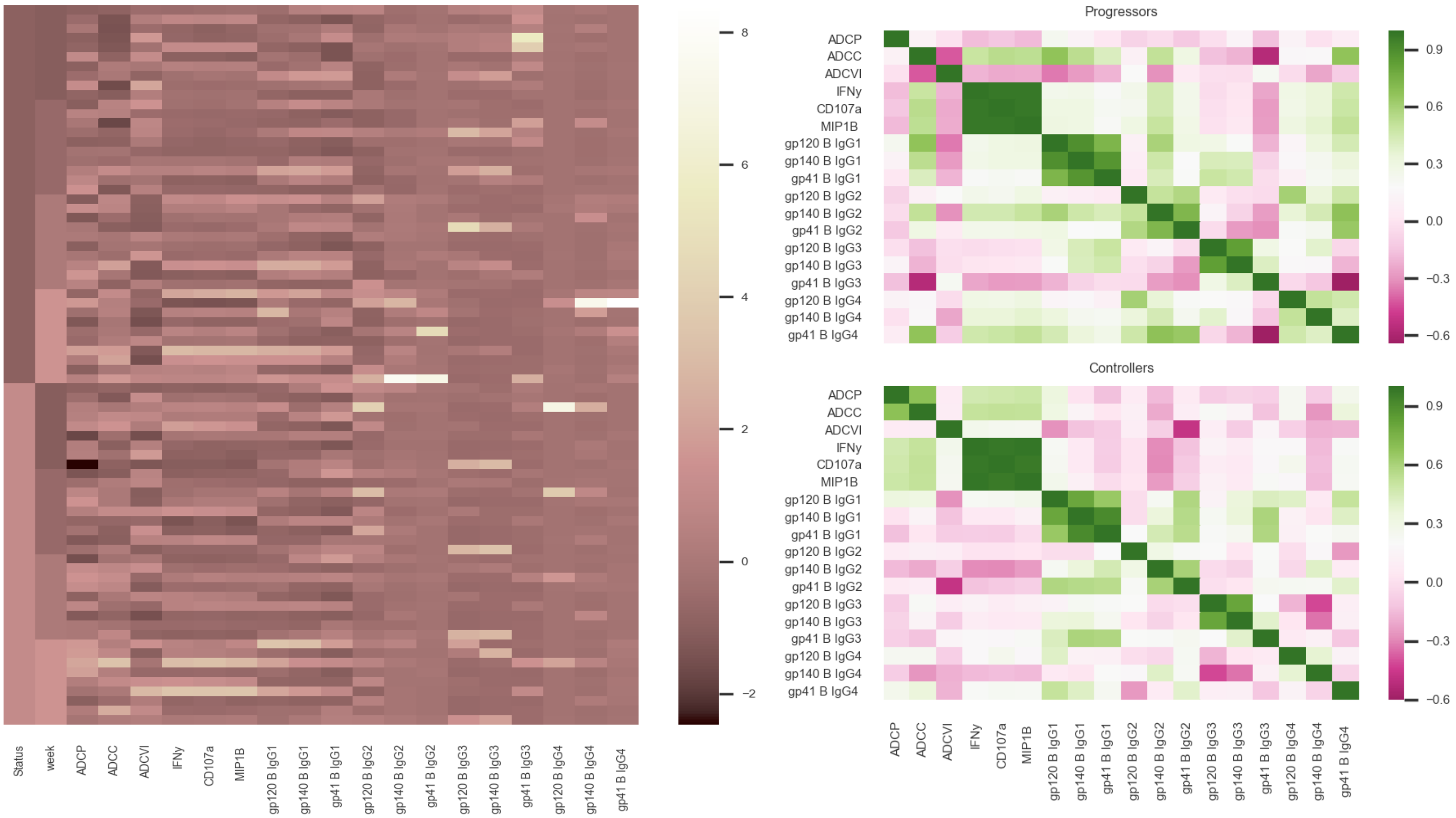


Figure 4.1: The heat map on the left shows the breakdown of the z-scored dataset for the two groups of HIV infected individuals. The two groups are seen in the first column labeled status with the top block representing progressors, the bottom controllers. For each group, four different time points were observed. These are labeled in column two (week) with the darkest square representing week 4 and the lightest representing week 48. The remaining columns represent the various antibody features. The heat maps on the right show Pearson correlation matrices for progressor features (top) and controller features (bottom). The correlation matrix shows the different IgG subclasses as well as effector functions being clustered.

features changed over the time points; however, the counts of different features showed no obvious patterns of change. Variability is seen more often in the progressor region than that of the controller region. A Spearman's correlation matrix was used for each group of infected individuals to better understand how features may correlate for each group. Figure 4.1 shows this matrix in heat maps (right heat maps) and, as would be expected, many features show high correlation to other features, but the most correlated features are those within the same HIV-specific IgG subclass. This was done as part of an exploratory analysis to better understand the features of the high-dimensional data.

Lasso/SVM Model Analysis

Further analysis was done using the averaged early time points (weeks 4 and 12) and the averaged late time points (weeks 24 and 48) in a multivariate statistical analysis which used a least-absolute square shrinkage operator (LASSO) as a variable selector and a support vector machine (SVM) for classification. The machine-learning model approach was used try and reduce the dimensionality of the dataset. LASSO was used in order to reduce the high dimensional dataset by selecting the features that when used in a classifying model could accurately predict controller and progressor outcomes without using the same dataset. Ideally, the goal of the analysis is to identify which features can support the model with the least amount of data but provide a similar amount of information about the dataset in order to predict outcomes. LASSO was also used as the feature selector in the original paper's multivariate correlate analysis.

The selected features in the original paper included three features; gp120 B IgG2, gp41 B IgG3, and gp140 B IgG2. The new

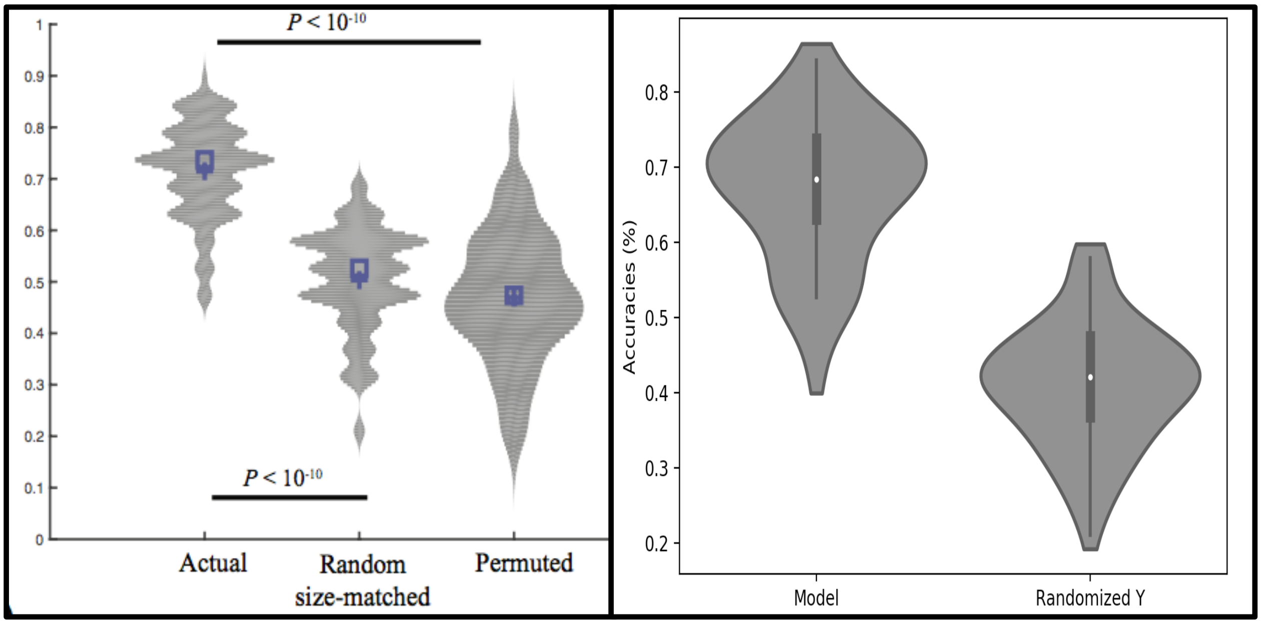


Figure 4.2: Multivariate models were used to analyze the data; the model's accuracies are displayed above in violin plots. The left plot is from the original paper (Sadanand et al., 2018) and the right plot was a replicated model using LASSO/SVM. The accuracies were not as high yet feature selection was reproduced. Each model ran 100 rounds using a 5-fold cross-validation.

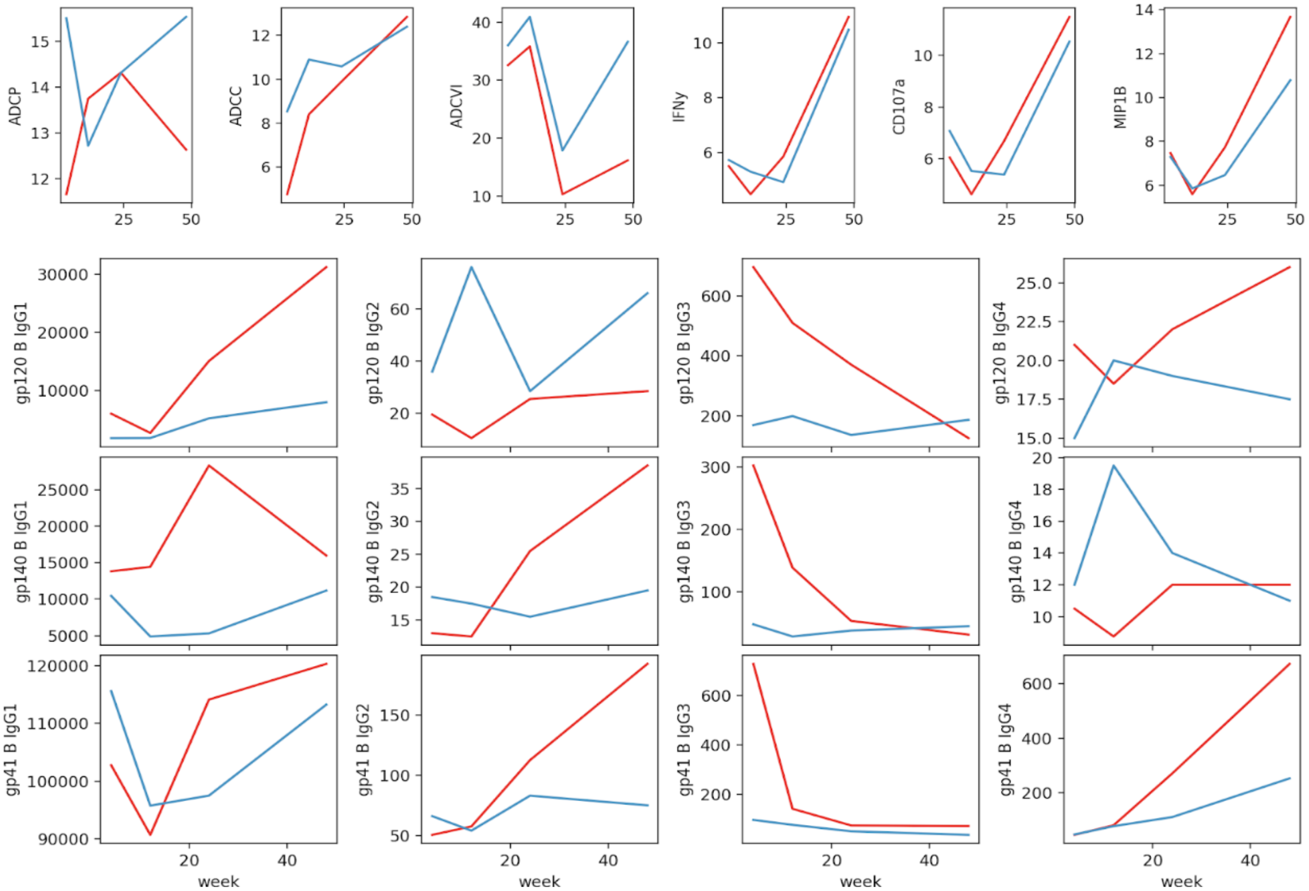


Figure 4.3: Exploratory analysis of features over time. Line graphs were plotted for each feature using medians for progressors (red) and medians for controllers (blue) at each recorded time point (4, 12, 24, and 48 weeks).

model was able to reproduce these features and due to differences in optimization also selected ADCVI. The new model did not work at the same accuracy as the original model as seen in Figure 4.2. The figure shows accuracies from the original paper (on the left) and the new model (on the right). The new model was only able to predict outcomes slightly better than a randomized model, so further research and model optimization would be necessary for generating more accurate and conclusive results.

Exploratory Analysis: Univariate

Exploratory analyses were conducted with univariate approaches to better understand each feature and how those features changed over time. Figure 4.3 show feature by feature line plots of the median values at each time point for progressors and controllers. From this, it is observed that controllers often seem to maintain values over time while progressors have large variations in median values, especially in the IgG2's and IgG3's. The features that visually had the greatest variations were among specific IgG2 and IgG3 antibodies. Gp120 IgG1 also showed variance at the late time points but was not selected as a feature in either the univariate or multivariate analyses. Gp 41 IgG4 also showed changes at later time points but was not selected as a feature in the multivariate analysis. For effector functions, ADCVI was the only function selected. It was selected in both univariate and multivariate analyses.

A Mann Whitney U test was performed for all features comparing progressor and controller values at both the early and late time points (averaged as in the multivariate analysis). Because this was an exploratory analysis, p-value was set to $p<0.1$, with the understanding that for the number of features being analyzed the statistical significance is diminished. The non-parametric U test found several features with p-values within the threshold that varied between progressors and controllers, either at the early or late time points. These features included the four features from the multivariate analysis as well as gp41 IgG2, gp41 IgG4, and gp140 IgG3. The statistically significant features from the U-Test were visualized using boxplots (removing outliers). These graphs can be seen in Figure 4.4.

It was also observed for the features selected from the model and the significant features from the Mann-Whitney U test that there were significant decreases between the averaged early time points and late time points. A series of bar plots for each of these features

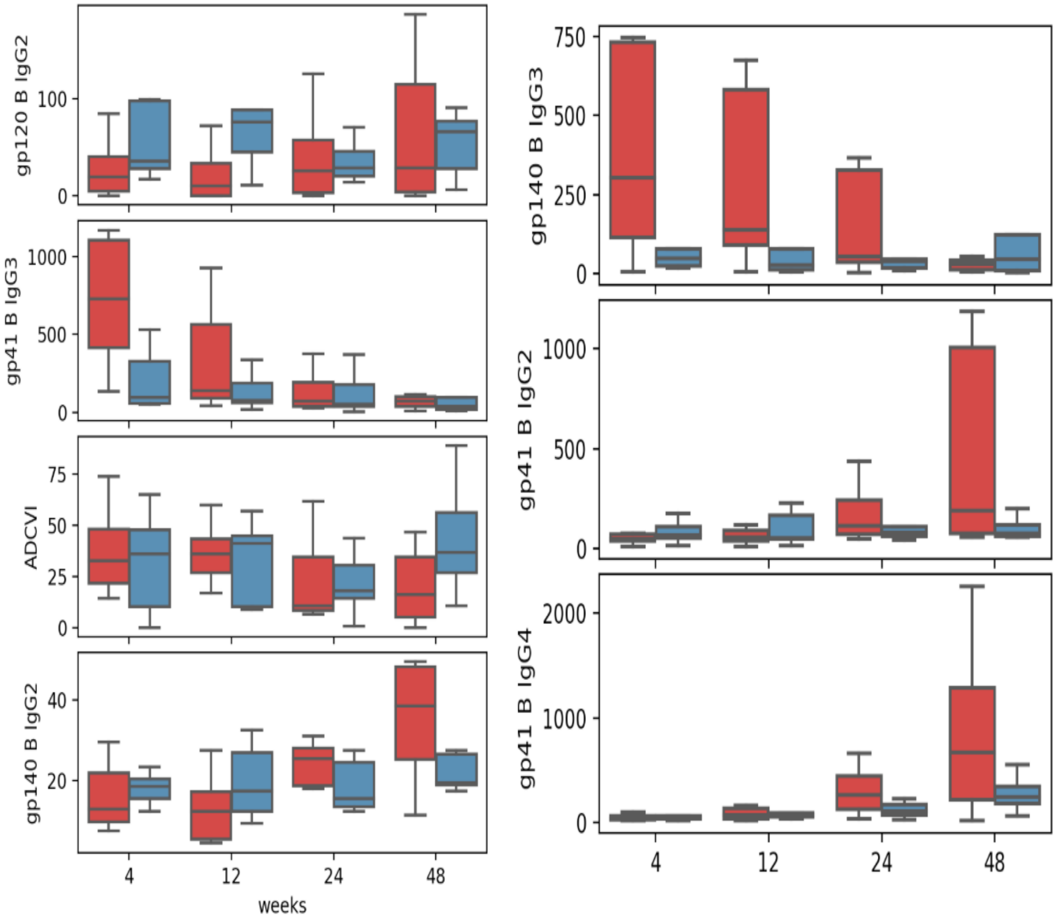


Figure 4.4: The boxplots show features selected from both the univariate and multivariate analyses. The multivariate features are displayed in the left column and the additional univariate on the right. Red, progressors; blue, controllers.

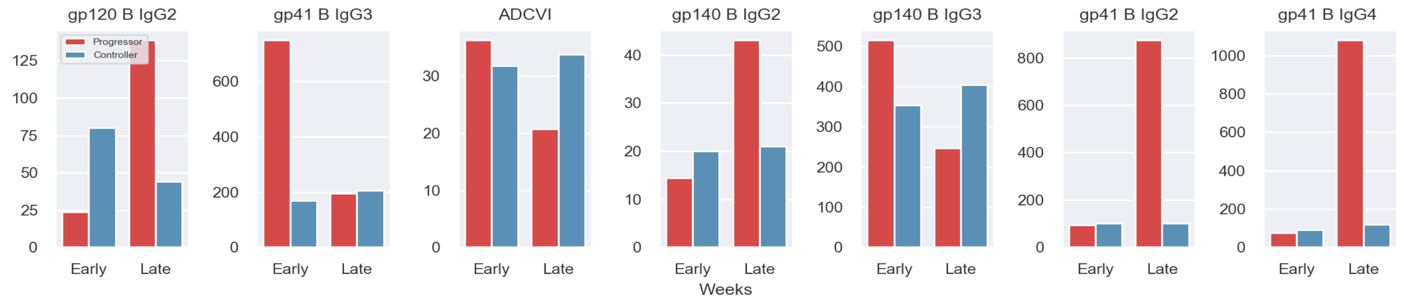


Figure 4.5: The bar plots show how levels of selected features from both the univariate analyses and the multivariate analyses changed over time. Early and late week medians of the averaged early and late time points were taken from each group for each feature and graphed. The variations of controllers and progressors are observed.

is shown in figure 4.5. Each plot reveals the median value for the early and late time points for progressors and controllers. The comparison seen in the bar graphs helps visualize how controller values stay constant through the study and progressors change drastically. The graphs also help to see how ADCVI variations over time seem to correlate with the increase of IgG2s and the decrease of IgG3s.

5. CONCLUSION

Researchers continue to study HIV infection due to the fact that it is still affecting people all around the world (Bhatti, Usman, & Kandi, 2016). Due to a rise in the interest in nonneutralizing antibodies, this analysis focused on HIV-specific IgG subclasses and antibody-dependent effector functions. The original paper focused on understanding the multivariate quantitative results obtained by using machine learning techniques. The original study concluded that maintenance of gp-120 specific and gp140 specific IgG 3 may be associated to disease control if obtainable in the acute stage of infection (Sadanand, Das, Chung, Schoen, Lane, Suscovich, …Alter, 2018). After further analysis, this conclusion was supported and extended.

The multivariate analysis selected similar features as to that of the original analysis of the data. The selected features included both Env-specific IgG2 and IgG3 antibody subclasses. The combination of these antibodies being selected together show correlations and new insights on potentially significant changes of the antibody profile, and how those changes can predict natural immunity outcomes. The results also show that not only could the antibody subclass counts play a role in predicting disease outcomes, but also antibody-dependent functions may contribute as well, specifically ADCVI.

Patterns observed about specific features from the univariate testing and visualizations also lead to new insights about the dataset and what contributes to viral control. Patterns were observed with HIV-specific IgG2s, IgG3s, and ADCVI. The changes in these HIV-specific IgGs could correlate to changes in the effector function ADCVI. Decreases in ADCVI occurred in progressors but not as significantly in controllers. Decreases were recognized with other trends in the subclasses. IgG 3 antibody counts were seen to decrease over time similarly to that of the ADCVI, where IgG2 antibody counts increased from the early to later time points. The boxplots show how, for some IgGs and effector functions, variance among levels of these features is much greater in progressors than in controllers. This imbalance of IgGs may also correlate to progressor or controller outcomes. Overall, trends were not seen in any one subclass or Env-specific HIV antibody; however, multiple trends were observed supporting the idea that there are multiple correlates that contribute to controller functionality. More specifically, stable levels of IgG2 and IgG3 antibody subclasses could be a determinant of preserving control over the disease.

In order to best understand nonneutralizing antibodies, antibody-dependent effector functions, and how they work together to contribute to disease control, other approaches may want to be considered. In the future, genomic data for each patient and their specific antibodies may also be an important factor maintaining IgG subclass levels which could ultimately lead to natural disease control. Other HIV studies have also found significant correlation between glycosylation and disease control outcomes. These studies found that different glycan combinations and glycan shield formations correlate with antibody neutralization breadth (Wagh et al., 2018). Adding more depth to the antibody profiles for each subject may help in generating more insightful and conclusive results regarding HIV disease control.

REFERENCES

Ackerman, M. E., Barouch, D. H., & Alter, G. (2017). Systems Serology for evaluation of HIV vaccine trials. Immunological Reviews, 275(1), 262–270. http://doi.org/10.1111/imr.12503

Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. Innate Immunity. Available from: https://www.ncbi.nlm.nih.gov/books/NBK26846/

Arnold, K. B., & Chung, A. W. (2017). Prospects from systems serology research. Immunology, 153(3), 279–289. doi:10.1111/imm.12861

Banerjee, N., & Mukhopadhyay, S. (2016). Viral glycoproteins: biological role and application in diagnosis. Virusdisease, 27(1), 1–11. doi:10.1007/s13337-015-0293-5

Bhatti, A. B., Usman, M., & Kandi, V. (2016). Current Scenario of HIV/AIDS, Treatment Options, and Major Challenges with Compliance to Antiretroviral Therapy. Cureus, 8(3), e515. doi:10.7759/cureus.515

Checkley, B. G. Luttge & E. O. Freed (2011) HIV-1 envelope glycoprotein biosynthesis, trafficking and incorporation. Journal of Molecular Biology 410, 582- 608.

Clem A. S. (2011). Fundamentals of vaccine immunology. Journal of global infectious diseases, 3(1), 73-8. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3068582/

German Advisory Committee Blood (Arbeitskreis Blut), Subgroup ‘Assessment of Pathogens Transmissible by Blood’ (2016). Human Immunodeficiency Virus (HIV). Transfusion medicine and hemotherapy : offizielles Organ der Deutschen

Gesellschaft fur Transfusionsmedizin und Immunhamatologie, 43(3), 203–222. doi:10.1159/000445852

Informed Health Online [Internet]. Cologne, Germany: Institute for Quality and Efficiency in Health Care (IQWiG); 2006-. How does the immune system work? 2010 Nov 24 [Updated 2016 Sep 21]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK279364/

Khattar SK, Samal S, LaBranche CC, Montefiori DC, Collins PL, et al. (2014) Correction: Comparative Immunogenicity of HIV-1 gp160, gp140 and gp120 Expressed by Live Attenuated Newcastle Disease Virus Vector. PLOS ONE 9(1): 10.1371/annotation/670112f8-b4fd-4622-8a8c-203dc647f807.

Kwong PD, Wyatt R, Robinson J, Sweet RW, Sodroski J, Hendrickson WA 1998. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. Nature 393: 648–659

Margolis, D. M., Koup, R. A. and Ferrari, G. (2017), HIV antibodies for treatment of HIV infection. Immunol Rev, 275: 313-323. doi:10.1111/imr.12506

Mayr, L. M., Su, B., & Moog, C. (2017). Non-Neutralizing Antibodies Directed against HIV and Their Functions. Frontiers in immunology, 8, 1590. doi:10.3389/fimmu.2017.01590

Nicholson L. B. (2016). The immune system. Essays in biochemistry, 60(3), 275-301.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5091071/

Overbaugh, J., & Morris, L. (2012). The Antibody Response against HIV-1. Cold Spring Harbor perspectives in medicine, 2(1), a007039. doi:10.1101/cshperspect.a007039

Pulendran, B., & Ahmed, R. (2011). Immunological mechanisms of vaccination. Nature immunology, 12(6), 509-17.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3253344/

Sadanand, S., Das, J., Chung, A. W., Schoen, M. K., Lane, S., Suscovich, T. J., … Alter, G. (2018). Temporal variation in HIV-specific IgG subclass antibodies during acute infection differentiates spontaneous controllers from chronic progressors. AIDS (London, England), 32(4), 443–450. doi:10.1097/QAD.0000000000001716

Wagh, K., Kreider, E. F., Li, Y., Barbian, H. J., Learn, G. H., Giorgi, E., … Korber, B. (2018). Completeness of HIV-1 Envelope Glycan Shield at Transmission Determines Neutralization Breadth. *Cell reports*, *25*(4), 893–908.e7. doi:10.1016/j.celrep.2018.09.087