Immunologic Correlates Analysis of RhCMV/SIV Vaccine Efficacy

Applying Machine Learning Techniques

to Model Vaccine Elicited T Cell Responses

Wenjun Song

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington
2018

Committee:
Peter Tarczy-Hornoch
Paul T. Edlefsen

Program Authorized to Offer Degree:
Biomedical and Health Informatics, School of Medicine

University of Washington

Abstract

Immunologic Correlates Analysis of RhCMV/SIV Vaccine Efficacy

Applying Machine Learning Techniques to Model Vaccine Elicited T Cell Responses

Wenjun Song

Chairs of the Supervisory Committee:

Peter Tarczy-Hornoch

School of Medicine

Paul T. Edlefsen

School of Public Health

In the past 30 years, HIV vaccine studies on traditional CD8+ T cell-targeted HIV vaccines were frustrated by the ineffectiveness of mediating immediate vaccinal interception upon infection acquisition prior to the explosive viral amplification. As the most important lesson of past HIV vaccine researches, the first hours to days immediately after viral infection might be the only vulnerable time period for immunologic interceptions.[1, 2] With this regard, immunologists started a novel research on employing Cytomegelovirus (CMV) as vaccine vector in early 2000s, to exploit CMV vectors' unique ability on eliciting and maintaining abundant functional T cell responses at all potential HIV infection sites.[3-6] Recent CMV-based vaccine research, demonstrated by Louis Picker and colleagues, with statistical support by Dr. Edlefsen, manifests a remarkable infection control and clearance on ~50% of HIV-acquired rhesus macaques (RM) vaccinated by Simian immunodeficiency virus (SIV) inserted rhesus cytomegalovirus (CMV) vaccine.[7, 8] This promising protection pattern motivates further immunologic correlates analysis on vaccine efficacy to investigate potential immunological mechanisms of the partial protection. As part of vaccine efficacy analysis, this project was conducted to inspect the relation between CD4+, CD8+ T cell responses elicited by vaccine and the unique protection outcome via interpretability attached machine learning techniques. Interpretability has been regarded as the driven feature of this immunologic correlates machine learning process. After stringent data screening and statistical modeling along with strategic informatics interpretation, I preliminarily identified two immunologic features which correlates with the protection pattern by potentially corresponding to formation of germinal centers that act as HIV virus's shelters.

Keywords: CMV-based SIV Vaccine, T Cell Immune Responses, Vaccine Efficacy, Immunologic Correlates Analysis, Machine Learning, Informatics Interpretation

TABLE OF CONTENTS

Introduction

Backgrounds

HIV/AIDS Interventions

Acquired Immune Deficiency Syndrome (hereafter referred to as AIDS) which challenges immune system has been regarded as incurable infectious disease even under the strongest anti-retroviral therapies (ART). [2, 9-11] Since the first official reporting of the AIDS epidemic in 1981, 77.3 million people have been diagnosed with AIDS, and 35.4 million died of this disease. Although the remarkable decline in new HIV infections and AIDS related deaths from 2010 with development of preventive interventions and ART, the mortality rate remains high. In 2017, 1.8 million people have been diagnosed with AIDS and about a million people died from AIDS-related illnesses. [12] Despite the intensive researches in last 30 years, established AIDS with the human and simian immunodeficiency viruses (HIV, SIV) infection is thought to be only controllable but not reversible. If only standing by current traditional ART and preventive interventions, our fight against this virus is far from over. [9, 10] With this regard, a new research field employing the cytomegalovirus (CMV) for HIV vaccine development, becomes the focus of the Picker Lab at the Vaccine and Gene Therapy Institute (VTGI) at Oregon Health and Science University (OHSU) for the recent decade. [13-15] As recent researches indicated, implementations of current non-vaccine interventions are unsustainable in the long term on a global basis. Although the scope of epidemic can be further reduced in a while, non-vaccine interventions are insufficient to terminate HIV epidemic by themselves in the next 50 years. [16, 17] However, a moderately effective vaccine can greatly enhance HIV prevention. Epidemiologic analysis indicates that, after introducing an HIV/AIDS vaccine with 70% efficacy in 2027, we can decrease new

HIV infections by over 16 million which is 78% of the annual incidence of new HIV infection under best implemented non-vaccine interventions alone.[16]

Vaccine Design

The simian immunodeficiency virus (SIV) inserted rhesus cytomegalovirus (CMV) vaccine (hereafter referred to as RhCMV/SIV) is being developed by Louis Picker's group. As retroviruses originally infecting African non-human primates, SIV is believed to be the original source of HIV-1 and HIV-2, the two human immunodeficiency viruses, by zoonotic transmission across the species barrier. [18] The homology between HIV and SIV reduces the difficulty of transforming to human version vaccine. Rigorous Simian immunodeficiency virus (SIV) vaccine trials on non-human primates provide the most instructive animal models to HIV vaccine development. [19]

From previous researches on HIV infection process, we learned that 1) once the infection been established with explosive systemic replication and diversification, the virus is capable to escape from the most effective immune mechanisms; [2, 10] 2) once host's self-immune responses been elicited after HIV infection, the HIV sanctuary, such as germinal centers, will serve as shelter for HIV virus and sustain the viral infection.[20] Therefore, in-time immune interception immediately upon infection acquisition is critical to the vaccine efficacy. However, the time window prior to the resilient viral reservoir establishment could be the first few hours to days after infection acquisition.[4, 8, 21, 22] This floating time window strengthened the requirement of the immediate interception by vaccine. To address this, Picker et al started a novel exploration of employing a persistent virus like CMV as vaccine vector to elicit immune effector responses in the early 2000s. [2] The CMV vector is able to elicit high frequency T cell responses specially to SIV virus entry while it has unique early-spread capability which ensures immediate interception on newly infected cells. In the last 10 years, researches on RhCMV vaccine vectors, which is a rhesus version CMV, achieved tremendous progress, including breakthroughs on the unique pattern of RhCMV/SIV and RhCMV/TB efficacy in SIV and TB challenge models. Over previous studies on RhCMV-based vaccine, this group, demonstrated an overview that, 1) the immunogenicity and protective capacity of RhCMV-based vaccine is sustainable for many years without decrement; 2) the spread of RhCMV virus can be suppressed in vivo by some modifications without loss of immunogenicity or efficacy;[23] 3) human CMV (HCMV) vector homologues of the spread-deficient RhCMV maintain similar spread inhibition in humanized mice and monkeys and are capable to elicit durable effector responses in such cross-species administration; 4) large parts of the RhCMV genome can be discarded to include at most 6kb of exogenous antigen inserts in 3 different sites; 5) modified RhCMV vectors with multiple inserts can maintain equivalent immunogenicity function by using endogenous promoters; 6) RhCMV vaccine will take effect on all immunological environment no matter the present of immune memory from previous CMV infection. This ubiquitous pathogenic feature ensures effective revaccination with HIV-inserted CMV vector. [8, 21, 24, 25]

Study Innovation

From Picker et al previous study on immune clearance of highly pathogenic SIV infection, they have reported that about 50% (60 out of 113) rhesus macaques (hereafter referred as RM) vaccinated with SIVmac239 inserted strain 68-1 RhCMV vectors manifested durable control and progressive clearance of the highly pathogenic SIVmac239 infection.[7, 8] After 50-70 weeks post infection acquisition, no significant virologic and immunologic distinctions can be distinguished by stringent comparison between protected monkeys and the monkeys that have never been exposed to SIV. Even the ultrasensitive RT-PCR and PCR analysis did not detect singular SIV RNA or DNA in necropsied protected-RM tissues from week 69 to week 172. In addition, no replication-capable SIV was detected in extensive co-culture analysis or adoptive transfer of 60 million hematolymphoid cells to healthy RM. With these promising data, this study made a remarkable conclusion that the pathogenic SIV infection has been functionally cured in RhCMV/SIV vaccine-protected RMs. [8, 24] For the ~50% protected RMs, RhCMV/SIV vaccine provides the first immune-mediated functional cure of a lentivirus leading to AIDS. If this vaccine efficacy were translated to a homogenous HIV inserted human version CMV vaccine, the "control and clear" protection pattern could dramatically enhance HIV prevention. Based on this, a Phase I clinical trial for safety and immunogenicity tests on a prototype spread-deficient (Δpp71) Human (H) CMV/HIVgag vector has been started in early 2017.[23]

To achieve the clinical translation, we need to develop both immunogenicity-optimized and efficacy-optimized HIV vaccine. The blindness of the immunologic basis of protection vs. non-protection in RM seriously hampers the CMV vector platform optimization. In order to strategically modify the translated 2nd generation CMV vector, a better understanding of qualitative and/or quantitative immune correlation of protection, which maps out the immunologic mechanisms corresponding to virus clearance process, is demanded to guide clinical development. According to the biological evidences from previous RhCMV/SIV studies, the ~50% protection outcome is related to the high frequency of differentiated T cell responses at early infection sites.[4, 21, 26-29] Besides, in 2017, Louis Picker et al. demonstrated that the protection outcome is associated to unconventional SIV specific CD8+ T cell responses which recognize unique epitope restricted by Major Histocompatibility

Complex (MHC)-II or MHC-E molecules.[6, 30-34] According to these immunological evidences, T cell responses profiling is potential to illustrate the immunological basis of protection mechanisms. Therefore, one initial task of vaccine efficacy optimization is to trace the cellular T cell immune responses and identify some immunologic correlates corresponding to high/low efficacy outcome. Statistical modeling combined with informatics interpretation can efficiently exploit high dimensional immune responses data and provide immune correlates candidates to assist wet laboratory efficacy analysis. With this regard, my study was designed to assist vaccine efficacy optimization by producing interpretable machine learning conclusion on the immune responses dataset. The primary goal of this project is to address whether the frequency of CD4+ and CD8+ T cell responses specifically elicited by RhCMV/SIV vectors at different time periods, are likely to have specific functionality and correlate with protection outcome. With an immune response dataset recording the frequency magnitudes of T cell CD4+ and CD8+ responses which are specifically elicited by four SIV genes at three time periods, the rationale for the study design is to statistically model the immune response features to the protection outcome and determine whether we can identify significant immunological features predicting protection outcome. The results from this study will strongly help to demonstrate whether the unique protection pattern is mediated by these CD4 and CD8 T cell responses or by other vaccine-associated parameters which we have not identified and measures. Additionally, the potential identification of immunologic correlates will provide immunogenicity target for human vaccine design. In order to provide interpretable model illustrations to immunologists collaborators, the information interpretability has been regarded as an important and driving feature of the machine learning approaches design in this study.

Methods

Data Collection Methods

Flow Cytometric Intracellular Cytokine Analysis

The immune response data used in this project was collected by cytometric intracellular cytokine staining analysis (ICS), as described by Hansen SG, et al. [8, 21], on SIV-specific CD4+ and CD8+ T cell responses measured in blood. The ICS analysis, which marks target cells by intracellularly staining the cytokines of the target cells using anticytokine antibodies, is the most common version of the cytokine flow cytometry (CFC).[35] The method produces incidence percentages of specific T cell responses to each of the 4 SIV genes (GAG, RTN, POL, ENV) as ICS magnitudes. These ICS measurements were taken over a time course beginning prior to vaccination and continuing through two vaccine administrations (prime and boost) and through the day of challenge and beyond. Here, Picker Lab provided the data that were pre-summarized into the three time periods of interest: highest immune response magnitudes post-1st vaccination and prior to the boost; the highest response after the boost and before challenge; and the average of three baseline responses at the moment just prior to challenge. As shown in table 1, the Picker lab also provided some additional monkey covariates including sex, as well some additional combinations of the 24 basic summary measures.

Immune Correlates Data Components	
Immunologic Response Parameters (62 features) **Key Components:** **4 genes:** **GAG, RTN, POL, ENV;** **3 time periods:** **prime, boost, pre-challenge;** **2 T cell types:** **CD4+, CD8+**	1. **Actual Responses: (24 features)** Four genes * three time periods * two T cell types.
	2. **Largest Peak Set: (8 features)** Largest values over 3 time periods for four genes in two T cell types.
	3. **Ratio Set: (8 features)** Two ratios (CD4/CD8, CD8/CD4) for four genes.
	4. **Addition Set: (22 features)** Additive responses over 4 genes of 3 time periods.
Meta-parameters	study: factor variable of 6 study IDs; RM: factor variable showing 113 Rhesus Macaque IDs; sex: binary factor variable of monkey sex (male & female); Vector(s): factor variable of vaccine versions; Insert(s): factor variable of extra vaccine modifications.

Table 1. Immunologic Parameters Matrix.

Unsupervised Screening Methods

Principle Component Analysis

With development of wet laboratory technologies in immunology, we can collect data on more features easily and economically. Although in this project, the 24 features on immunological response are not as much as, for example, genetics data which has over thousands features, this immune correlates database can be regarded as high dimensional data. Abundant features enable comprehensive immunological response modeling. However, the high dimensional data challenges statistical power in modeling analyses. Therefore, identification of an effective way to construct the 24 features to best reveal the structure that explains the variance in the data is demanded.

Principle component analysis (hereafter referred to as PCA) which identifies the variability basis of the data, is the best match corresponding to this data processing motif. [36] The PCA process contains 2 major steps: 1). Standardizing initial data by mean centering and computing Z-scores of the population (or t-statistics of sample observations) to obtain zero empirical mean average and Var(X)=1 of each variable. 2). Eigenvalue decomposition of data correlation (EVD) or singular value decomposition of a data matrix (SVD) by orthogonal transformation to obtain distinct principal components which are linearly uncorrelated.[36, 37] In this project, I performed PCA process simply by using `prcomp` and `princomp` built-in functions from R package `stats`.[38, 39] Both `prcomp` and `princomp` functions performs PCA by singular value decomposition. Whereas, `princomp` operates R-mode PCA which handle the data with at least as many observations as features. While the Q-mode PCA by `prcomp` relaxes the requirement on observation size, which offers an option for analyses on small sample size. [39, 40] Although the immune correlates data is in R-mode with 24 features of 113 monkeys, considering the moderate observation size, I applied both `prcomp` and `princomp` to ensure valid PCA.

ICS Magnitudes Linear Regression

The goal of this linear regression analysis on all ICS magnitudes is to determine whether the 2 variables/variants, time period and T cell types, can explain the variation in log10 ICS magnitudes of the original ICS magnitudes. Achievement of this goal will provide evidence to affirm the importance of time period and T cell type which were primarily reflected by the first 2 PCs as described in main text of results.

As a start, the original data matrix was reconstructed to a new data frame of all ICS magnitude with several ICS variants as labels in column. After data reconstruction, for the whole ICS data, there are 6 ICS variants which are time period (prime, boost and pre-challenge), T cell type (CD4+ or CD8), genes (GAG, POL, RTN and ENV), sex (male and female), study (6 study cohorts) and RM (monkey IDs). Five models are generated from linear regressions on different ICS variant combinations. Then, the model evaluation results of the 5 models could reveal that whether the linear regression model with only time period and T cell response variants could describe the original feature matrix better than the other models.

Hierarchical Cluster Analysis (HCA)

Hierarchical cluster analysis (HCA) is an unsupervised clustering process to group observations according to the internal hierarchy by a certain measure of dissimilarity. [41] In general, HCA has 2 branches: agglomerative HCA and divisive HCA. The first one starts from separate features and converge feature-pairs by hierarchical similarity until all features merged in one cluster. The second one produces a dendrogram starting from one cluster with all features and divisively split until reaching every single feature. In this study, I performed agglomerative HCA by `hclust` function in R package `stats`. [38, 39] In terms of dissimilarity measurements selection, I followed Ward's minimum variance method to avoid resulting in dendrogram with reversals which are hard to interpret. [42] The Ward's minimum variance method can be conducted in `hclust` function by selecting `ward.D2` as method. [39] In this study, HCA is served as a complementary visualization tool to previous PCA results. After acquisition of dendrogram, in order to check the consistency between PCs and hierarchical clusters, I labeled all leaves in dendrogram by time period (prime, boost or pre-challenge), T cell type (CD4+ or CD8) and genes (GAG, POL, RTN or ENV) and inspected whether the upper hierarchy of leaves corresponding with certain labels.

Logistic Regression

Multiple generalized linear regression analysis, with the goal to predict a single binary outcome by multiple independent variables, is commonly used for infectious disease modeling.[43] The greatest challenge of statistical modeling on this comes from that, the binary or categorical outcome do not carry intuitive numerical meanings in themselves. Logistic regression models binary outcomes as odds which are numerical from 0 to 1.[44] In short, logistic modeling on binary outcome is to predict the probability of a random object having or not having the outcome in condition of other independent variables.[44] In this project, I applied `glm` function in R package `stats` to build logistic regression models.[38]

LASSO

LASSO is regarded as the major feature selection methods for statistical modeling preparation.[45] Unlike unsupervised feature selection methods by inspecting on features directly, like PCA and clustering, LASSO is a penalized regression which accomplishes feature selection by identifying best model with a penalty term for the number of independent variables with non-zero coefficients. It might be more intuitive to regard LASSO as a model evaluation process which results in feature selection.

Intuitively speaking, when we try to describe an object, using more features will give a more comprehensive description of the object. However, this is not fairly tenable for statistical modeling practice with moderate sample size. In order to hold certain statistical power, larger sample size with sufficient observations is required as a recompense of taking in more independent variables. With this regard, the goal of model evaluation process can be concluded as identifying certain amount of correlated features which can describe outcome comprehensively and concisely. LASSO is a model fitting evaluation tool by penalizing extra features in the model. There are two most commonly used methods called Elastic Net regularization and least-angle regression (LARS), which overcome shortcomings from original LASSO's stiff penalty.[46, 47] To increase prediction accuracy on samples with small observations, Zou and Hastie introduced elastic net regularization which modifies original penalty by adding an additional ridge regression like penalty in 2005.[48] Another method, LARS, adds least-angle regression like penalty to make LASSO performance more stable.[47] In this project, both LASSO methods have been applied in modeling analyses through glmnet and lars R packages.[49, 50]

Model Evaluation/Selection Methods

Analysis of Variance (ANOVA)

The role of analysis of variance (hereafter referred to as ANOVA) is to evaluate the statistical significance of additional variables added to a statistical model.[51] In detail, with several models with different sets of variables, ANOVA can be used to select the best model through the judgment of whether adding in an extra variable results in significant alteration of variance described by the model. In this project, I conducted ANOVA on linear regression model selection and logistic modeling selection by using `anova` function with likelihood ratio test (LR test) in R package `stats`.[38] In the one-way ANOVA table computed by this `anova` function, the P values of additional variables added to the null hypothesized model can be used to determine the significance of the variable, hence whether the variable should be included.

Akaike Information Criterion (AIC)

Besides ANOVA, Akaike information criterion is a specialized indicator of model quality from information representation aspect. When estimating model quality, AIC considers the model simplicity and information lost rate at the same time. With AIC score of each model, we can determine the relative information lost rate for model A by calculating $\exp(AIC_{min}-AIC_{A})/2)$.[52] In this project, AIC was conducted on linear regression model selection and logistic modeling selection by using `AIC` function in R package `stats`.[38]

Results

1. Data Preparation

1.1 Data Separation

From laboratory RhCMV/SIV challenge studies, Louis Picker and his colleges have created an immune correlates matrix which served as the database for statistical analysis in this project. By flow cytometric intracellular cytokine staining analysis (hereafter referred as ICS), the CD4+/CD8+ immune responses were measured at 3 time periods on 4 SIV-specific antigens, GAG, RTN, POL and ENV. In detail, the 3 time periods are prime (upon first immunization), boost (upon second immunization) and pre-challenge (right before week 59 after post-initial vaccination). Interpretation of the ICS magnitudes collected in data is the incidence percentage of specific CD4+/CD8+ immune responses. Typically, the ICS magnitudes of prime and boost time periods are the monkey specific peak response by 4 antigens. The ICS magnitudes of pre-challenge are the monkey specific average response by 4 antigens. In total, the data recorded 62 immunologic response parameters, 5 meta-parameters and challenge outcome data of 113 monkey samples (Table 1.).

Immune Correlates Data Components	
Immunologic Response Parameters (62 features) **Key Components:** **4 genes:** **GAG, RTN, POL, ENV;** **3 time periods:** **prime, boost, pre-challenge;** **2 T cell types:** **CD4+, CD8+**	5. **Actual Responses: (24 features)** Four genes * three time periods * two T cell types.
	6. **Largest Peak Set: (8 features)** Largest values over 3 time periods for four genes in two T cell types.
	7. **Ratio Set: (8 features)** Two ratios (CD4/CD8, CD8/CD4) for four genes.
	8. **Addition Set: (22 features)** Additive responses over 4 genes of 3 time periods.
Meta-parameters	study: factor variable of 6 study IDs; RM: factor variable showing 113 Rhesus Macaque IDs; sex: binary factor variable of monkey sex (male & female); Vector(s): factor variable of vaccine versions; Insert(s): factor variable of extra vaccine modifications.

Table 1. Immunologic Parameters Matrix.

The immunologic correlates response matrix originally has 62 immunologic response features. However, only the first 24 features are actual T cell responses which contains CD4+ and CD8+ responses of 4 genes at 3 time periods. The rest features are computed features from the first 24 actual T cell response features. The data providers clarified that, the last 3 sets which record some combinations of the actual responses, are some alternative forms of immunologic responses for their tentative analyses. According to definitions of the last 3 computed sets in Table 1, the addition set and the largest peak set are sum and maximum of the actual response set respectively. The computed features in the last 3 sets are perfectly multi-collinear to the actual response features. Features with these collinearities are statistical covariates which will potentially break statistical rules if including them together in later machine learning tools. Given this, correlates analyses should be conducted only on one set at a time to secure not breaking any statistic rules.

For this study, as a preliminary modeling analysis on these immunologic responses, later machine learning would start on the 24 actual response features to check whether these basic response values can reveal statistical significant correlates to the unique protection outcome. Table 3 summarized all features used for later immune correlates analysis in this study. From this standpoint, intuitive collinearities from definitions in raw data have been cleaned by data separation.

Study Data Components	
24 Immunologic Response Features Dimension: Four genes * three time periods * two T cell types.	GAG_Prime_CD4+, RTN_Prime_CD4+, POL_Prime_CD4+, ENV_Prime_CD4+ GAG_Prime_CD8+, RTN_Prime_CD8+, POL_Prime_CD8+, ENV_Prime_CD8+ GAG_Boost_CD4+, RTN_Boost_CD4+, POL_Boost_CD4+, ENV_Boost_CD4+ GAG_Boost_CD8+, RTN_Boost_CD8+, POL_Boost_CD8+, ENV_Boost_CD8+ GAG_PreC_CD4+, RTN_PreC_CD4+, POL_PreC_CD4+, ENV_PreC_CD4+ GAG_PreC_CD8+, RTN_PreC_CD8+, POL_PreC_CD8+, ENV_PreC_CD8+ (PreC: pre-challenge time period)
Meta-parameters	study: factor variable of 6 study IDs; RM: factor variable showing 113 Rhesus Macaque IDs; sex: binary factor variable of monkey sex (male & female); Vector(s): factor variable of vaccine versions; Insert(s): factor variable of extra vaccine modifications.

Table 3. Study Data Summary

1.2 Balanced Data Partition

In preparation for later machine learning and according to relatively small sample size, the immune correlates matrix was split to training and test sets in 80% and 20% divisions. Although impacts from the 4 meta-parameters (sex, study IDs, vaccine insert and challenge route) are not clear, statistical modeling in this study was designed to target on the quantitative immunologic response parameters and determine whether they are likely to specifically correlated with protection. Thus, to ensure consistency of training and test sets, I balanced the 5 meta-parameters by conducting stratified random sampling in R. As a result, training and test datasets have consistent distribution as original data; hence, have balanced the 5 meta-parameters and protection outcome. After this step, the test set data was reserved and all statistical analyses were conducted on training data set until model validation.

1.3 Data Quality Check

Data quality considerations originate from the nature of sampling and go beyond data cleaning and transformation.[53] In regard of this, I looked back to check on biological laboratory settings which determined the nature of data sampling and screened the immunologic correlates matrix to check whether the outcome condition and observation distribution are good for later statistical tools to produce qualified results.

When looking at the ~50% protection outcome condition, correlates identification on this database would attain the greatest statistical power because of the ideal balance between protected and non-protected outcomes. In terms of some separated analysis on extreme study cases in which vaccination was perfectly, or nearly, efficacious (like study 179), the statistical power will diminish but potentially stay high to recognize correlates under the hypothesis that there is a measured immunologic response value that differentiate protected monkeys from non-protected monkeys.

In terms of data distribution, many statistical tools prefer data normally distributed. First, I checked the original distribution of each feature across all observations. Since pre-challenge data does not have the same ICS range as prime and boost, I plot pre-challenge data separately. (Figure 1,2) In general, the original distribution is left skewed and has a long right tail. I determined to transform the data to log10 scale. After log10 transformation, the data produced normal distributions. (Figure 3,4)

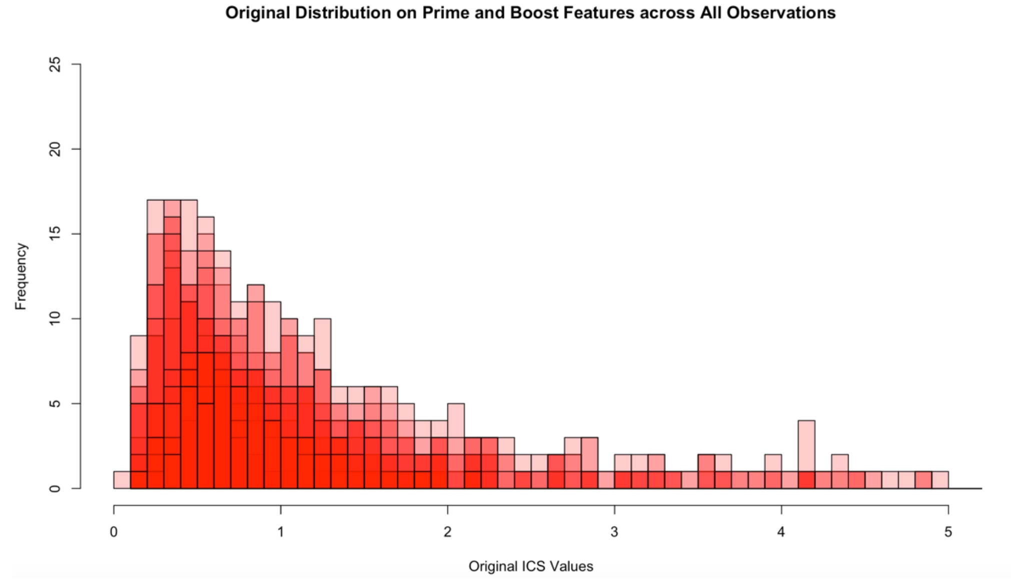


Figure 1. Original Distribution on Prime and Boost Features across All Observations. On this plot, I overlapped 12 histograms for 6 prime features and 6 boost features. Each histogram shows distribution of the 113 observed ICS values of one feature.

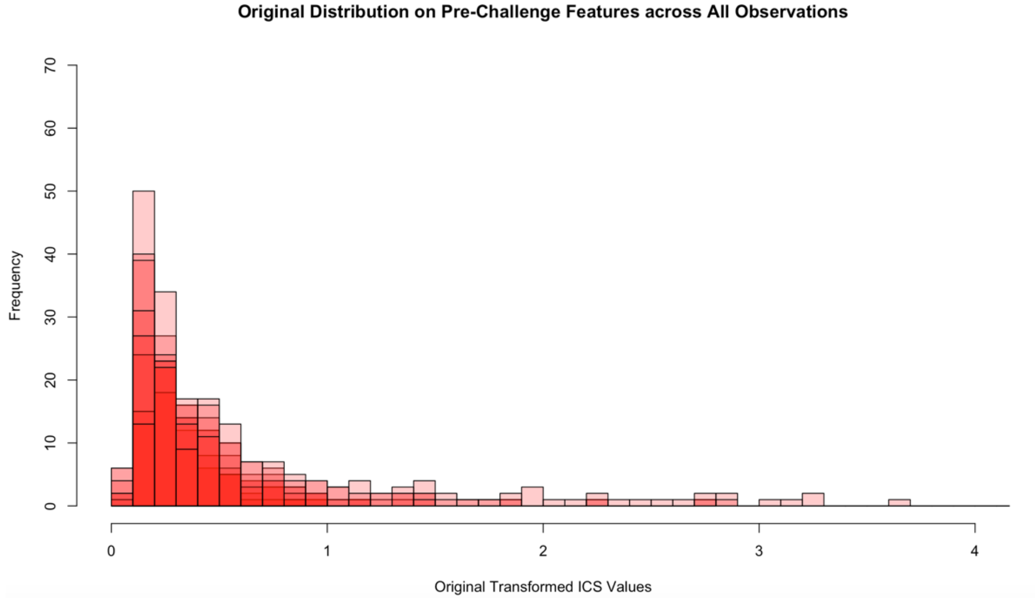


Figure 2. Original Distribution on Pre-challenge Features across All Observations On this plot, I overlapped 6 histograms for 6 pre-challenge features. Each histogram shows distribution of the 113 observed ICS values of one feature.

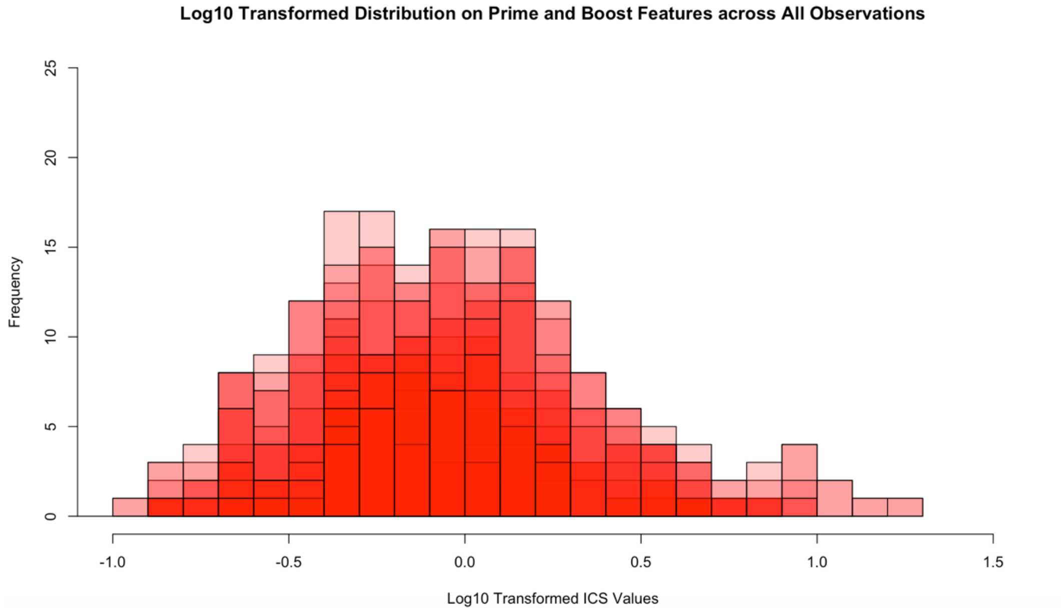


Figure 3. Log10 Transformed Distribution on Prime and Boost Features across All Observations

On this plot, I overlapped 12 histograms for 6 prime features and 6 boost features. Each histogram shows distribution of the 113 observed log10 transformed ICS values of one feature.

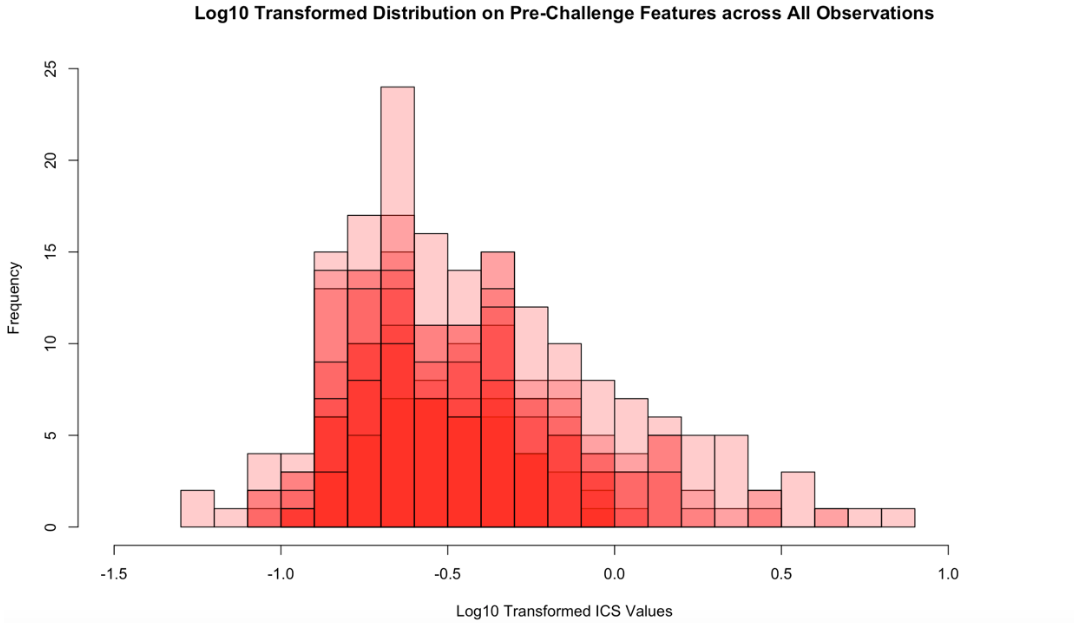


Figure 4. Log10 Transformed Distribution on Pre-challenge Features across All Observations

On this plot, I overlapped 6 histograms for 6 pre-challenge features. Each histogram shows distribution of the 113 observed log10 transformed ICS values of one feature.

1.4 Pairwise Correlation Visualization

In this visualization process, I traced the correlations between each pair of the 24 features by calculating pairwise Pearson correlation coefficients, and visualizing linear/non-linear correlation type in pairwise scatterplots.

After calculating Pearson scores for each feature pair, I visualized correlations by correlation heat map. (Figure 5.) The heat map intuitively suggested 2 correlated regions: features in right up corners and features in left down region. These 2 regions actually corresponding to CD4+ and CD8+ segregation. With this suggestion, I plot Trellis plots for these 2 regions to further visualize the correlation scores and linear/non-linear correlation types. (Figure 6,7)

First, inspection on Pearson correlation scores could serve as a further collinearity check inside the 24 features. In result, correlation scores in right upper panel suggested that feature pairs with high correlation scores should not be included in one model for later modeling analyses. In general, considering the inevitable biological relationship among these T cell immunologic responses, although some feature pairs' Pearson correlation scores are relatively high, it is tolerable to try transformation and later linear modeling coefficients interpretation on the 24 features.

Second, I used the scatter plots in left down panel to inspect the linear/non-linear correlation types which could suggest whether this data need other transformations. As a comparison, I also plot same trellis plots for original data without log10 transformation. (Figure 8,9) In result, pairwise correlations are linear in general. Given this result, this data does not need other transformations. Starting from this point, the data with 24 immunologic response features is statistically ready for actual correlates analysis.

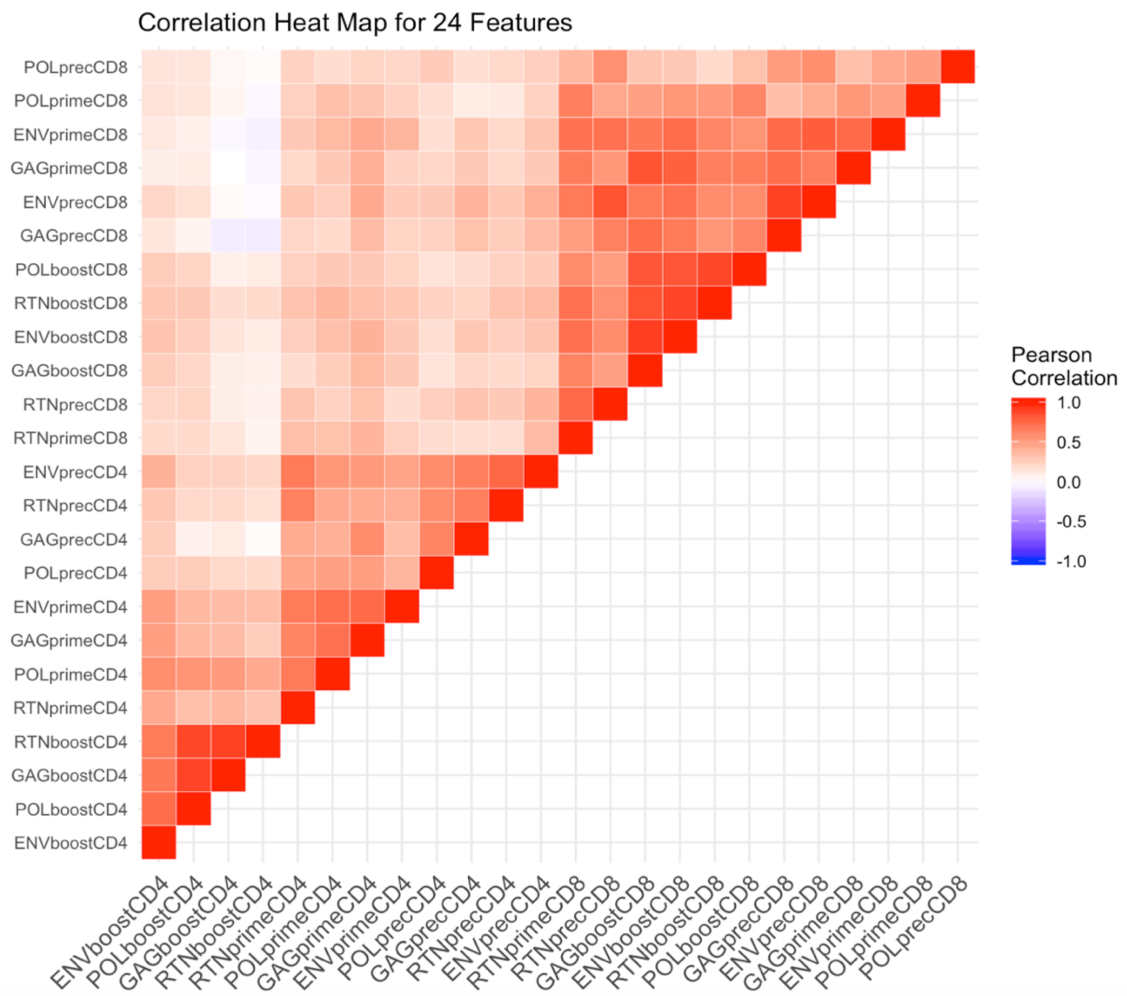


Figure 5. Correlation Heat Map for 24 Features.

This figure shows pairwise correlation of 24 features according to Pearson correlation scores. The 24 features were reordered by correlation hierarchy. This heat map intuitively suggested 2 correlated regions: right up corner and left down corner. (note: abbreviation prec refers to pre-challenge time period.)

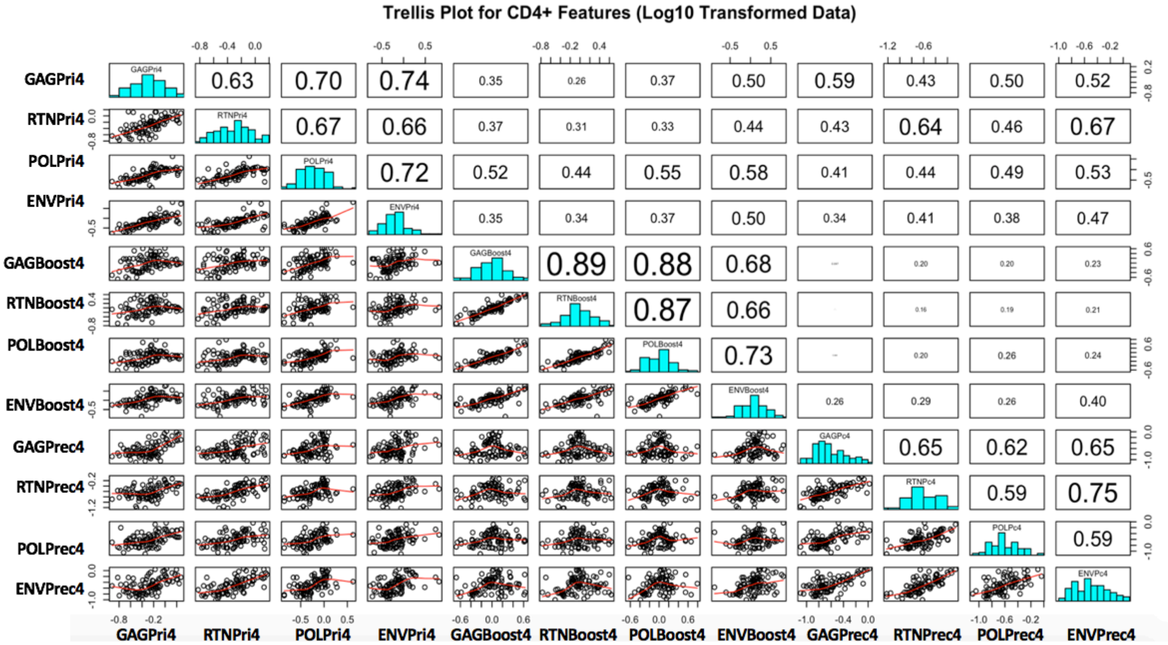


Figure 6. Trellis Plot for Log10 Transformed CD4+ Features

This trellis plot gives names of Cd4+ feature in diagonal, along with histogram of each feature. The right upper panel of this plot shows pairwise Pearson correlation scores. The increase of score's font size indicates increase of correlation strength. Pairwise scatter plots with LOESS lines in red were plotted in left down panel, which show linear/non-linear correlation types.

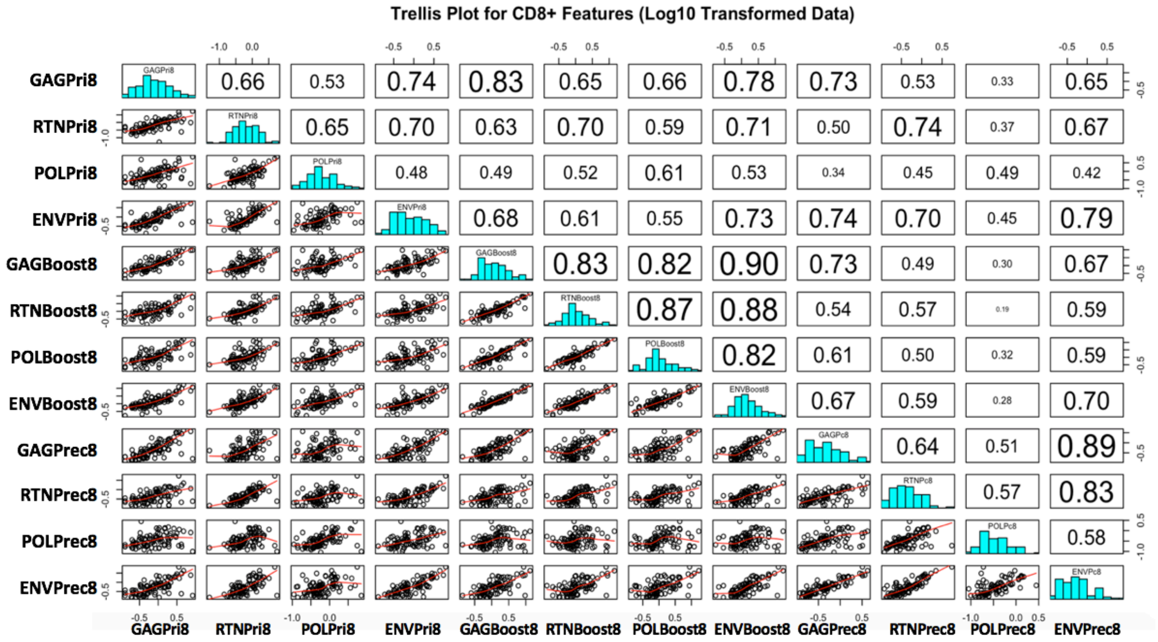


Figure 7. Trellis Plot for Log10 Transformed CD8+ Features

This trellis plot gives names of Cd8+ feature in diagonal, along with histogram of each feature. The right upper panel of this plot shows pairwise Pearson correlation scores. The increase of score's font size indicates increase of correlation strength. Pairwise scatter plots with LOESS lines in red were plotted in left down panel, which show linear/non-linear correlation types.

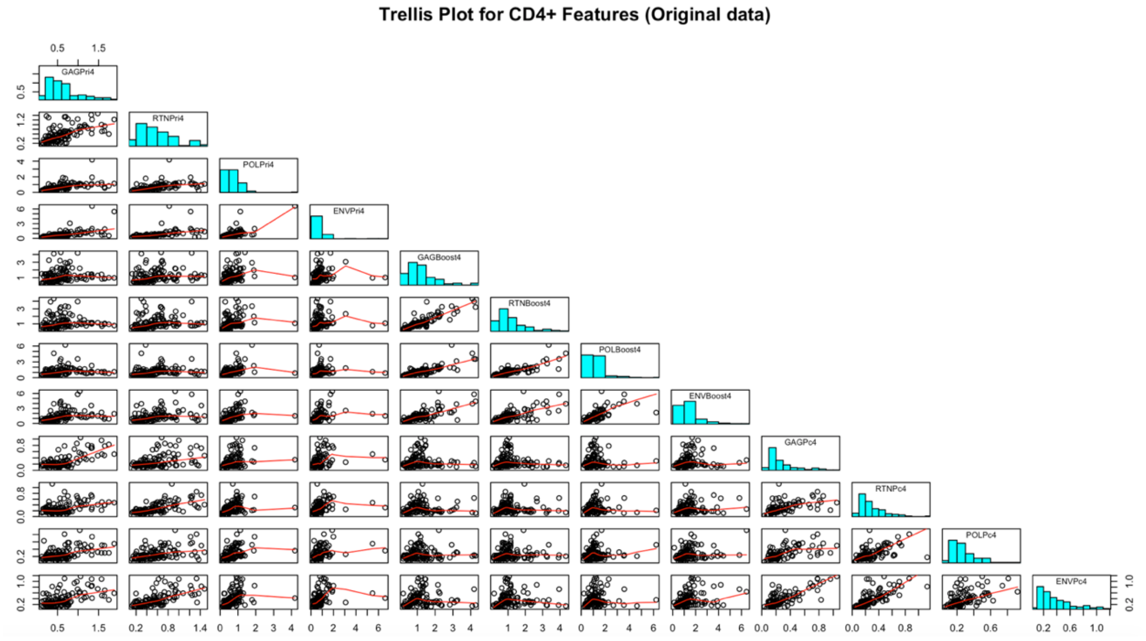


Figure 8. Trellis Plot for CD4+ Features (no transformation)

This trellis plot gives names of Cd4+ feature in diagonal, along with histogram of each feature. The right upper panel of this plot shows pairwise Pearson correlation scores. The increase of score's font size indicates increase of correlation strength. Pairwise scatter plots with LOESS lines in red were plotted in left down panel, which show linear/non-linear correlation types.

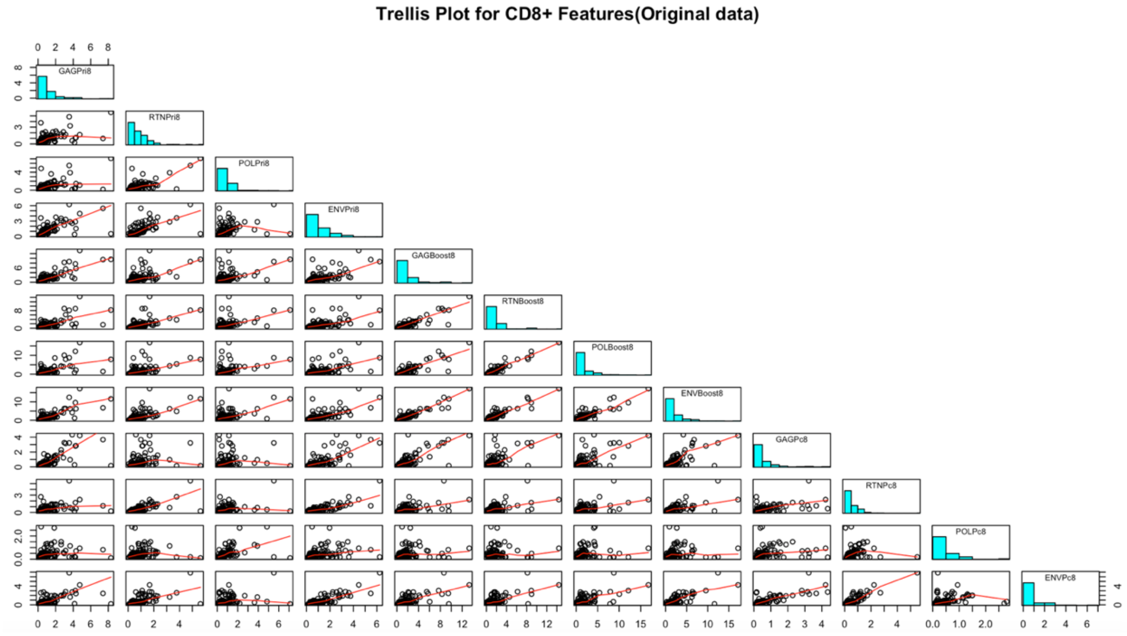


Figure 9. Trellis Plot for CD8+ Features (no transformation)

This trellis plot gives names of Cd8+ feature in diagonal, along with histogram of each feature. The right upper panel of this plot shows pairwise Pearson correlation scores. The increase of score's font size indicates increase of correlation strength. Pairwise scatter plots with LOESS lines in red were plotted in left down panel, which show linear/non-linear correlation types.

2. Unsupervised Feature Selection and Feature Reconstruction

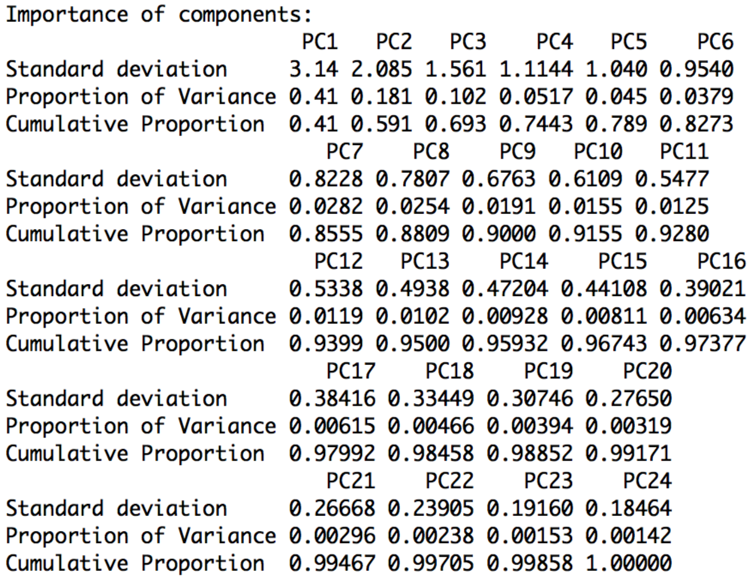
2.1 Unsupervised Screening Overview

With the known RhCMV vector anti-viral mechanisms from previous studies by Louis Picker et al [7], T cell responses magnitudes collected in the data used in this project might not be sufficient or symptomatic to identify correlates to the outcome. With regard of the potential hardship of statistically identify immunologic correlates from this database, a better understanding of the features in data is demanded to help decrease noises, detect useful hidden information, exclude potential factors to cause models to fail and thus, increase the possibility of identifying correlates by later modeling. Such an understanding was expected to achieve four-fold goals: 1) to determine latent structures inside these features; 2) to check whether there are redundant features which can be deducted to reduce dimension; 3) to check whether there are new forms of these features to better reflect the biological meaning of the known RhCMV vector anti-viral mechanisms; 4) in the whole process, to prevent introducing subjective errors by including outcomes.

Given these expected goals, a thorough unsupervised data screening on features matrix would be the best match with following considerations. First, unsupervised learning allows objective analysis purely on features without influences of the outcome, hence accomplishes the fourth goal. Besides, from its exploratory nature, unsupervised learning usually does not have specific single goal, which fits the expectation to understand the feature matrix broadly. I conducted data variability basis check by principle component analysis followed by principle components interpretation. The intuition of doing principle components interpretation is to identify the original features in the first few principle components and conduct feature pre-selection on original features scale. Although using the first few principle components is another way to pre-select features, using original features to conduct a PCA-guided dimension deduction is a better way to attach biological meaning to the statistical results.

2.2 Principle Component Analysis

After performing principle component analysis (hereafter referred as PCA) on the original strata by `princomp` function in R, the 24 original immunologic features were converted into 24 linearly uncorrelated principle components (hereafter referred to as PCs) by orthogonal transformation algorism.[54] The PCs, sorted by large to small proportion of variance explained by reach PC, were generated by an orthogonal basis set which is a linear combination of original features across all monkey sample (Figure 10.). It was promising that more than 59.1% of the data's variability came from the first 2 PCs.



```
Importance of components:
                        PC1   PC2   PC3    PC4   PC5    PC6
Standard deviation     3.14 2.085 1.561 1.1144 1.040 0.9540
Proportion of Variance 0.41 0.181 0.102 0.0517 0.045 0.0379
Cumulative Proportion  0.41 0.591 0.693 0.7443 0.789 0.8273
                          PC7    PC8    PC9   PC10   PC11
Standard deviation     0.8228 0.7807 0.6763 0.6109 0.5477
Proportion of Variance 0.0282 0.0254 0.0191 0.0155 0.0125
Cumulative Proportion  0.8555 0.8809 0.9000 0.9155 0.9280
                         PC12   PC13    PC14    PC15    PC16
Standard deviation     0.5338 0.4938 0.47204 0.44108 0.39021
Proportion of Variance 0.0119 0.0102 0.00928 0.00811 0.00634
Cumulative Proportion  0.9399 0.9500 0.95932 0.96743 0.97377
                          PC17    PC18    PC19    PC20
Standard deviation     0.38416 0.33449 0.30746 0.27650
Proportion of Variance 0.00615 0.00466 0.00394 0.00319
Cumulative Proportion  0.97992 0.98458 0.98852 0.99171
                          PC21    PC22    PC23    PC24
Standard deviation     0.26668 0.23905 0.19160 0.18464
Proportion of Variance 0.00296 0.00238 0.00153 0.00142
Cumulative Proportion  0.99467 0.99705 0.99858 1.00000
```

Figure 10. Statistics of Principle Components

This figure is the R output using `princomp` function. PC refers to principle component here which are sorted by large to small proportion of variability explained. The cumulative proportion shows cumulative proportion of variance explained by previous PCs.

Moreover, according to the bi-plot of PC1 and PC2 (Figure 11.), two groups of eigenvectors segregated by T cell types (CD4+ and CD8+) are almost orthogonal to each other, which indicates that T cell type acts as a major impact from the original 24 features on the first 2 PCs. Besides, as the black boxes circled in Figure 11, eigenvectors of one time period gather together for CD4+ features. This result suggests that most variability of the 24 features is from T cell type and time period. With this suggestion, I attempted to confirm this association by following PC interpretation analysis.

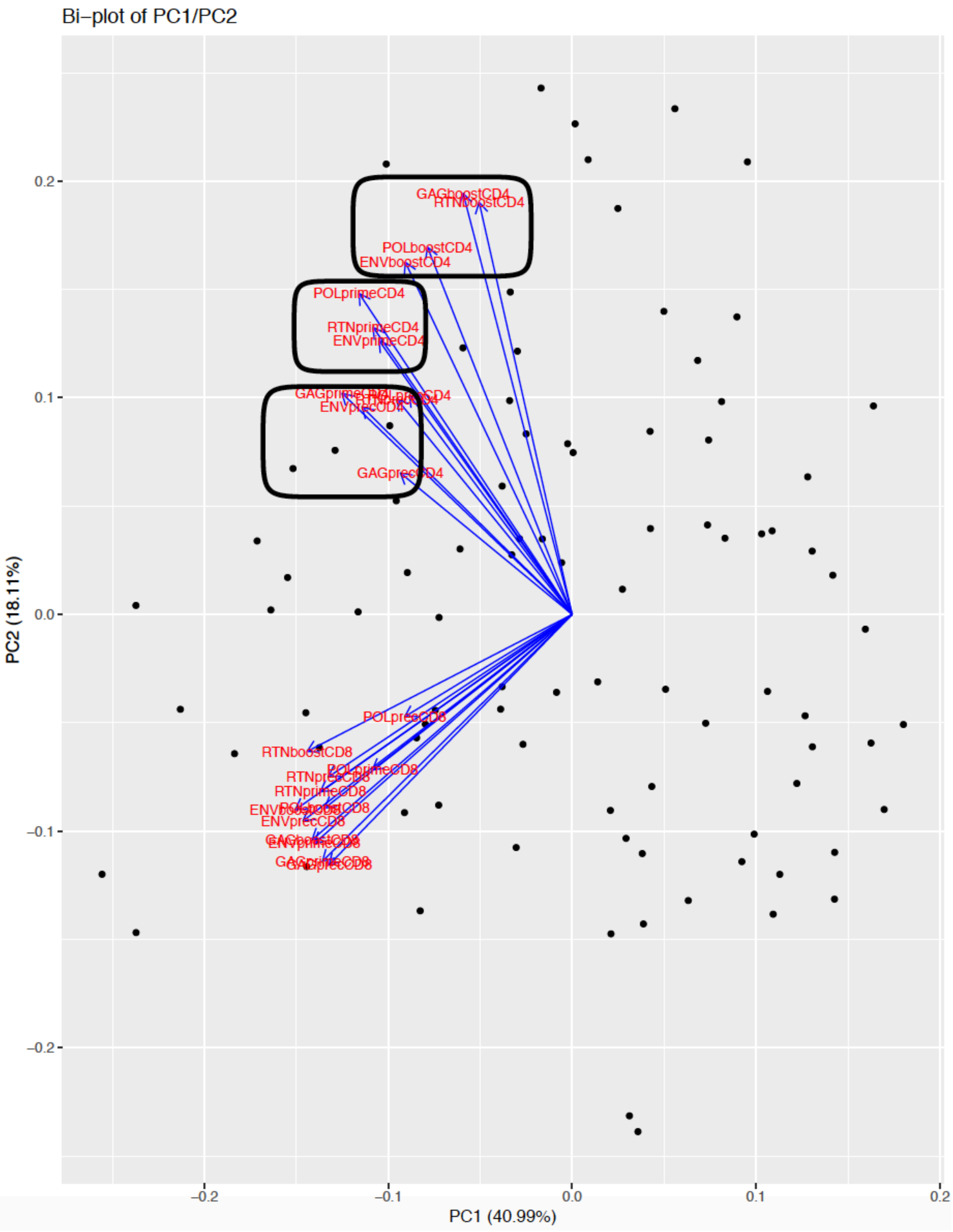


Figure 11. PC1, PC2 Biplot

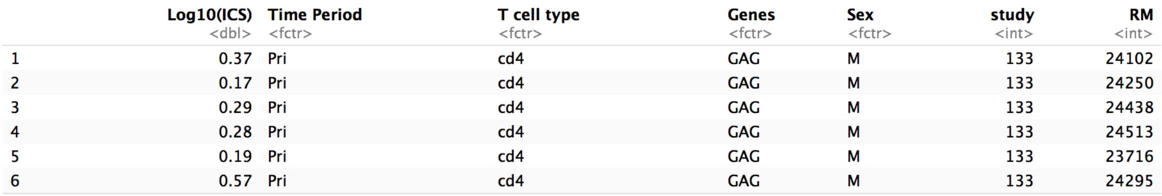
This bi-plot displayed all 113 observations as black dots with eigenvectors of the 24 features as blue arrows, on PC1 and PC2 scale. The eigenvectors are gathered in two main directions, which segregate eigenvectors of 2 T cell types. The arrows pointing down left corner are all CD8+ eigenvectors and the arrows pointing up left corner are all CD4+eigenvectors. Moreover, circled by 3 black round boxes, CD4+ eigenvectors are segregated according to the 3 time points.

2.3 Principle Components Interpretation

With the suggestion from PCA, I determined to further inspect whether the 3 time periods and 2 T cell types are corresponding to the first 2 principle components, thereby are able to act as new features which deduct genes in the original features. Inspection on this is worthwhile since, once determined such correspondence, the data would not only reduce dimension largely and likely to gain more power on later statistical modeling analysis with less features, but also would reserve the same interpretability as original features. One way to transcribe this inspection to statistical language is to determine whether the time period and T cell type as 2 factor variables, can estimate the whole variability of the original ICS magnitudes with certain statistical power. The statistical method I selected here to check the ability of estimation is to linear regress all ICS magnitudes which contain all original variance, to 1). the two 2 factor variables (time period and T cell type); 2). the two factor variables plus factor variable of genes; 3). the two factor variables, factor variable of genes plus other meta-features which might contribute to ICS variance. Then, by model selection process based on fitting precision, this method would check whether the linear regression model with only time period and T cell type could describe the original feature matrix better than the other models including other variants like genes (GAG, POL, RTN and ENV), sex (male and female), study variable (6 study cohorts) and RM variable (113 monkey IDs). Similar to that PCs are principle eigenvalues of the whole data, the features included in the best model are the principle features which best describes ICS magnitudes variability. However, these principle variables nominated by linear regression are on original feature's scale but not on orthogonal transformed scale. In brief, the new principle component analysis on original feature's scale by linear regression is a good attempt to enable interpretations.

The dataset used in this analysis unfolds all log10 transformed ICS magnitudes (2713 rows) of 7 factor variables (7 columns). (Table 4) This dataset was reshaped from the original 24 features data in order to extract factor variables of time period, T cell type and genes. Considering that both study (factor variable referring 6 different studies) and RM (monkey IDs) express the individual specificity, they are perfectly collinear which would break the linear regression algorism. Since the effect from studies is more problematic than monkey effect, I included only study but excluded RM in later linear regression process. I built 5 linear models after linear regressed the ICS magnitudes on different combinations of variables. Then, I conducted model selection by AIC assessment and ANOVA F-test evaluation.

During this process, I noticed that there is a trend in log10 ICS values across 6 studies (study variable). However, from the laboratory standpoint, there should not be any trend across studies since all studies are expected on the same scale. This suggested that the trend in primate cohort need to be normalized by some normalization methods. Since this normalization is not expected in the study mainline, I discussed and attempted study trend normalization in discussion part of this paper. Due to this uncertainty in study variable, for following linear regression analysis, I decided to exclude the Model 5 which involves study variable.



	Log10(ICS)	Time Period	T cell type	Genes	Sex	study	RM
	<dbl>	<fctr>	<fctr>	<fctr>	<fctr>	<int>	<int>
1	0.37	Pri	cd4	GAG	M	133	24102
2	0.17	Pri	cd4	GAG	M	133	24250
3	0.29	Pri	cd4	GAG	M	133	24438
4	0.28	Pri	cd4	GAG	M	133	24513
5	0.19	Pri	cd4	GAG	M	133	23716
6	0.57	Pri	cd4	GAG	M	133	24295

Table 4. Linear Regression Data Matrix

Descriptions on columns: Log10(ICS) is a continues quantitative variable including all log10 transformed ICS magnitudes (2713 rows); Time Period is a factor variable with 3 levels (prime, boost and pre-challenge); T cell type is a factor variable with 2 levels (CD4+ and CD8+); Genes is a factor variable with 4 levels (GAG, RTN, POL, ENV); Sex is a factor variable with 2 levels (male, female); study is a factor variable referring to different studies with 6 levels (6 studies); RM is a factor variable referring monkey IDs with 113 levels (113 monkeys).

First, from the AIC score in Table 5, excluding model 5, the model 4 with minimum AIC is the best model to minimize the information loss. Besides direct inspection on AIC scores, I assessed the information loss minimization ability of the models by calculating their AIC statistics using formula exp((AIC_{min}-AIC_i)/2).[52] All of AIC results were recorded in Table 5. From AIC statistics which suggest the probability that model X can minimize information loss as well as the best model, none of the other four models are close to the best model.[52] However, the big AIC jumps occurred at adding the first 3 variants which are time period, T cell type and genes, indicated by a line chart of AIC distinction between the models pairs. (Figure 12.) In result, indicated by AIC assessment, 1). the best model should keep all 4 variables in sake of less ICS information loss; 2). among the 6 variants, the time period, T cell types and genes bring in most variances. In other words, from AIC assessment, not only time period and T cell type, but also genes, stand out from the original features, which is inconsistent with PCA analysis where genes were not condensed by eigenvectors.

	Model 1	Model 2	Model 3	Model 4	Model 5
Features Included	+time period	Model 1 + T cell Type	Model 2 + gene	Model 3 +sex	Model 4 +study
*AIC score	1774.0	1654.0	1608.8	1604.5	---
**AIC statistic (Comparing all models)	2.44e-74	3.18e-22	1.36e-2	1	---
** pair-wise AIC statistic distinction	7.67e-53				---
		2.34e-20			
			1.36e-2		

Table 5. Test Models Descriptions and AIC Summary

Note: The Model 5 has been excluded from this analysis since the normalization issue in study variable. This table collected AIC score and computed AIC statistics. The model with minimum AIC score is the best model with least information lost under penalty of the amount of variables included in the model. AIC statistics suggest the probability that model X can minimize information loss as well as the best model. The pair-wise AIC statistics showed the degree of changes between model pairs.

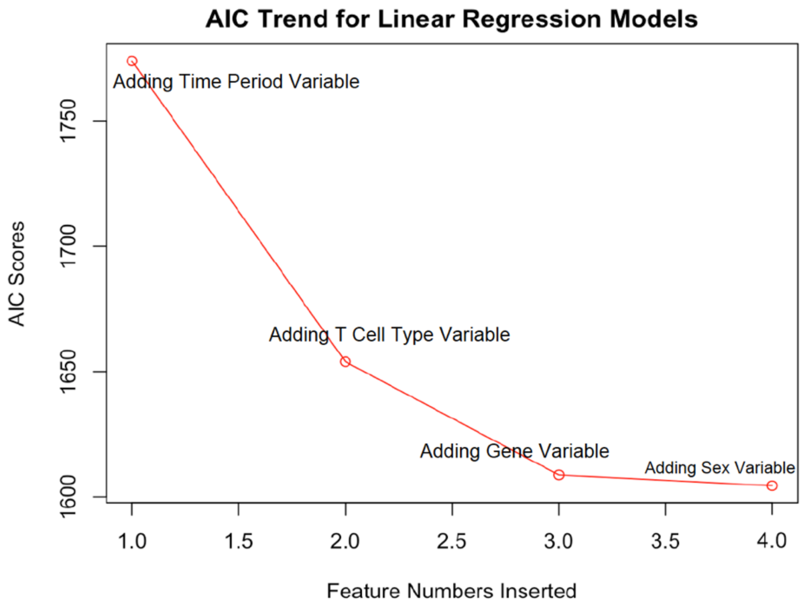


Figure 12. AIC Trend

This plot intuitively shows big AIC drops happens at adding time period variable, adding T cell type variable and adding gene variable.

Second, according to ANOVA table **(Figure 10.)**, adding of all the 4 variables can significantly optimize the model, although sex showed relative modest significance. This result indicates that all of 4 variables represent important ICS variability.

	Zero Model	Model 1	Model 2	Model 3	Model 4
Features Included	**+ constant**	**Zero Model + time period**	**Model 1 + T cell Type**	**Model 2 + gene**	**Model 3 + sex**
Degree of Freedom (Df)	2711	2709	2708	2705	2704
Degree of Freedom Loss	2				
		1			
			3		
				1	
F values	578.69				
		127.09			
			17.21		
				6.31	
P values	<2e-16***				
		<2e-16***			
			4.546e-11***		
				0.012*	

Table 6. ANOVA Table.

This ANOVA table records results of sequential ANOVA F tests on Model 1 to Model 4. The NULL model for each ANOVA F test is the previous model in sequence. The P values shows the probability of seeing current model is not better than previous model.

Overall, the linear regression analysis pointed out that to best describe variability in ICS magnitudes, all of the 3 time period, 2 T cell type and 4 genes which are actually all 24 features, should be included into the model. Nonetheless, time period and T cell type are the two features strongly suggested by the first 2 PCs. Although this linear regression analysis introduced genes as another significant feature, the result of this linear regression validate again that time period and T cell types are important. In purpose of feature pre-selection for later modeling, I will try both including all 24 features to model and including only time period/T cell types to model.

3. Statistical Modeling

3.1 Modeling Description

I involved in protection outcome for later analyses start from this point. The overall goal of statistical modeling is to determine whether the quantitative immune response parameters, pre-selected by preceding unsupervised analysis, are likely to specifically correlate with protection outcome, and if possible, to interpret the correlation to help biologists determine the kinetics of CMV/SIV vector-mediated protection and validate the sequential attenuate vaccine design. To accomplish this goal, I conducted 2 modeling analyses on different variable sets. The first modeling analysis includes all 24 features, as indicated from previous ICS magnitudes linear regression analysis. The second modeling analysis was conducted on less redundant variables to better detect hidden correlates. With the suggestion on time period and T cell type from PCA, this second modeling analysis excluded genes but only included 6 response features on 2 T cell types across 3 time periods . After determined the candidate model, preliminary model inspection and model validation were conducted. At the stage of model inspection, biological meanings were attached to each correlate to evoke further discussion on interaction analysis and model interpretation.

3.2 First Modeling Analysis

3.2.1 Feature Selection

Regarding to that our ICS data has binary protection outcome and multi-independent variables, logistic regression is the first modeling option. The traditional logistic regression method has a widely recognized limitation on sample size, which is called the "one in ten rule".[55] This rule demonstrates that a minimum of 10 events per explanatory variable (EPV) is required to stabilize prediction accuracy of logistic regression models. The one in ten rule is the primary limitation to avoid risk of overfitting in logistic modeling, which can be upgraded to "one in 20 rule" or "one in 50 rules" under stricter statistical power restriction.[55] Although recent studies attempt to prove relaxing this primary rule, breakaway from it will bring in more limitations on research questions. For our 1st modeling trial, with 24 explanatory variables under 50% protection proportion, logistic regression requires at least 10*24/0.5=480 observations to limit over-fitting risks in the sense of one in ten rule while we only have 91 monkeys in training data. With this regard, prior to logistic regression, I conducted least absolute shrinkage and selection operator (LASSO) on the original 24 features for feature deduction with the aim to select at most 3 features. In terms of LASSO methods selection, both least-angle regression (LARS) and Elastic-Net method (glmnet package in R) are capable to our data since the variables do not need specialized LASSO penalty. Moreover, this modeling analysis incorporates semi-supervised clustering as a complementary visualizing procedure to validate logistic regression modeling result.

3.2.2 Model Fitting Results

To conduct LASSO analysis, I first used lars package in R to perform least-angle regression function with internal cross validation (hereafter referred to as CV-LARS). For each step adding in one variable, the LARS function provides a corresponding Mallows's Cp statistic. The smaller Mallows's Cp indicates relatively the more precise model. From CV-LARS Cp plot (Figure 13.), Cp value increases from the beginning. In other words, the smallest Cp is obtained at adding in the first variable. This extreme trend of Cp values indicates that none of the 24 features are necessary for a good prediction of the outcome under LARS function. As a complementary attempt, I conducted LASSO again by Elastic-Net method using R's glmnet package. Similar to LARS, the glmnet plot (Figure 14.) suggested to include 0 feature into the model. Drawn from these results, it can be concluded that none of variables are worth to logistic regression modeling. There are two potential reasons behind this LASSO result: first, there is no actual relations between outcome and time period/T cell type/genes; second, there is actual relations but the relations could not be detected due to unknown internal structure pitfalls such as confounding, of the original 24 features. Detecting the internal structure of the original 24 features is hard without sufficient biological evidences. Therefore, to validate the second potential reason, using simplified features could be a good attempt, and if possible, constructing the features in a better shape. Based on this, the proposed second modeling analysis on time period and T cell type, deducting genes, will be a good subsequent attempt.

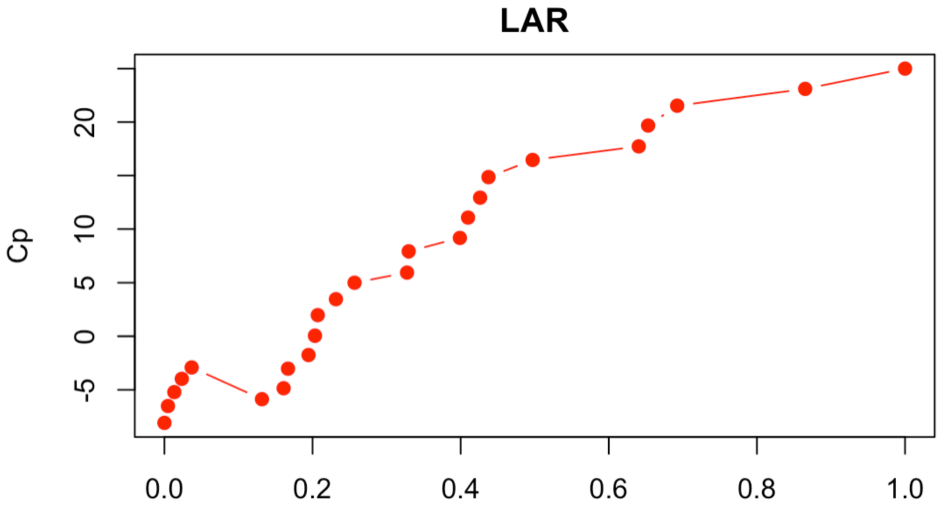


Figure 13. Mallows's Cp plot from CV-LARS LASSO analysis for the first modeling analysis. Each red dot in this plot represents the Mallows's Cp value for adding in each of the 24 features. Basically, the best model should include all features before the step (inclusive) with the least Mallow's Cp value. Based on this, this plot suggested 0 feature should be included.

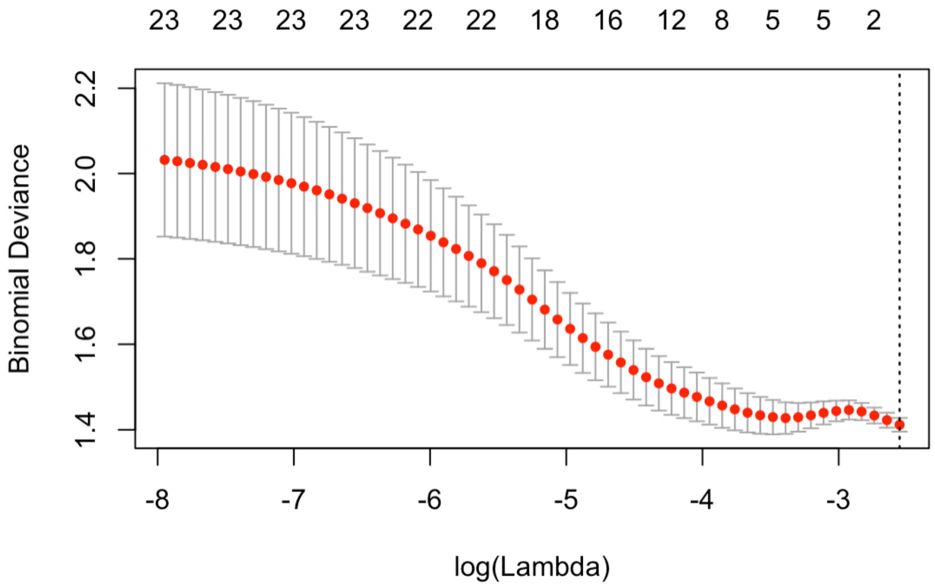
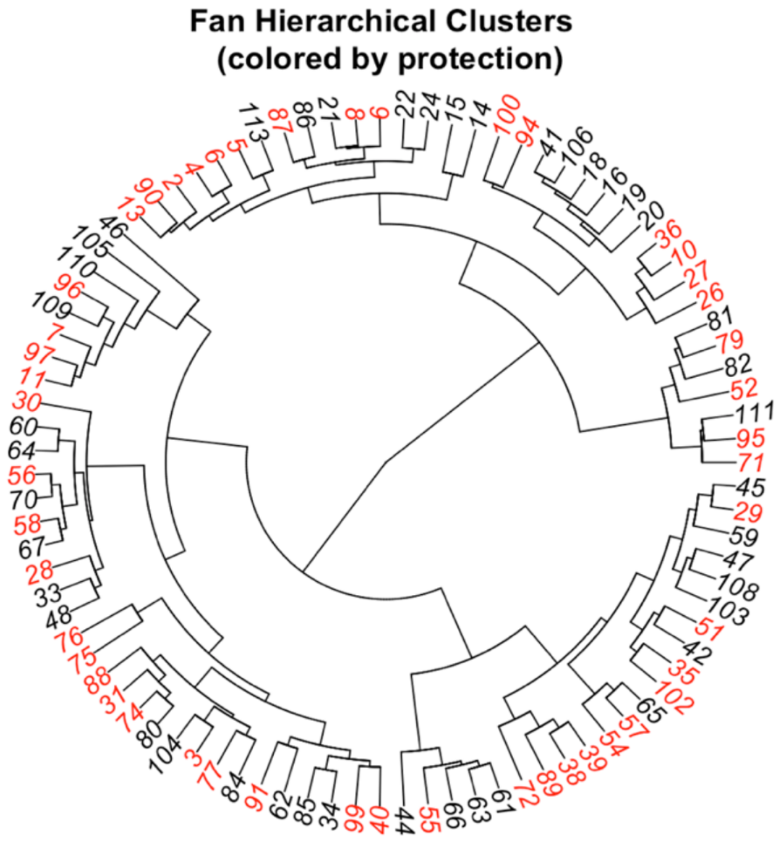


Figure 14. Elastic-Net LASSO (glmnet) summary plot for the first modeling analysis This summary plot gives a dash line which indicates the amount of features that should be included in modeling analysis, suggested by Elastic-Net LASSO model.

3.2.3 Semi-Supervised Clustering Visualization

This semi-supervised clustering analysis added protection outcome as labels to the unsupervised hierarchical clustering. Though this clustering analysis included outcome variable, the clustering process is still unsupervised and restricted to feature variables. The outcome variable was simply added to the unsupervised clusters, hence the name semi-supervised clustering.

Although this semi-supervised clustering visualization was originally designed to assist logistic regression modeling, it could also be a qualitative test to the extreme negative LASSO result at current stage. Unlike LASSO which is based on complex mathematical calculus, semi-supervised clustering directly uses the initial data to visualize the relation between protection and the 24 features. The goal of this semi-supervised clustering analysis is to validate LASSO result by another crude method. The consistency between clustering result and LASSO result will provide convincing evidence to prove that the 24 original features are not competent to model the protection outcome. From the visualization of semi-supervised hierarchical clustering (Figure 15.), the outcome values are dispersedly distributed in hierarchical tree. To sum up, result of semi-supervised clustering visualization validated LASSO result that there is no significant relation between the 24 original features and protection outcome.



red: Protected
black: Unprotected

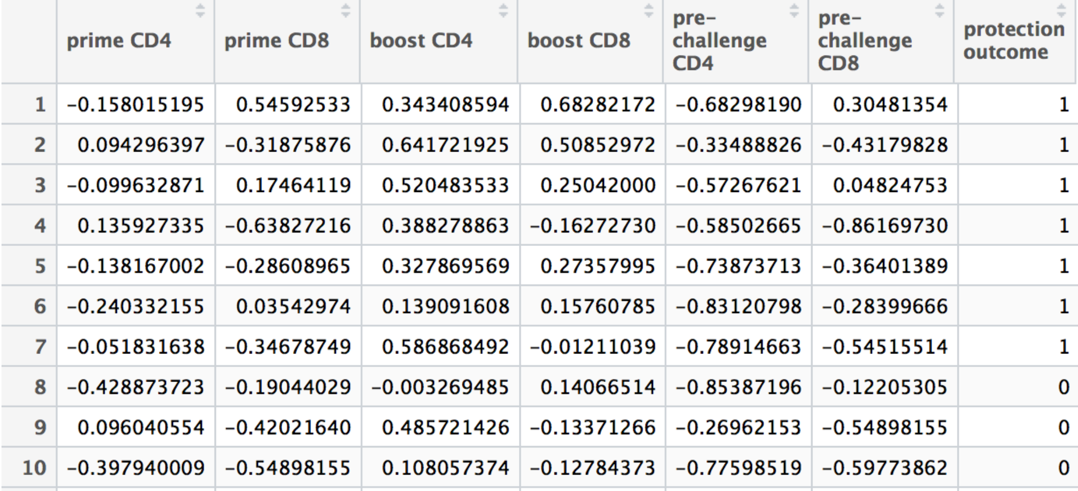
Figure 15. Fan Hierarchical Dendrogram Labeled by Protection Outcome

This plot intuitively shows the protection outcomes distribution in the log10 ICS hierarchy. In general, there are no significant protection segregation in log10 ICS hierarchy.

3.3 Second Modeling Analysis

3.3.1. Data Preparation

The second modeling analysis picked only time period and T cell type as independent variables to model the protection outcome, directed by the first modeling analysis results and along with the strong suggestion on time period and T cell type from PCA. Revisiting Table 3., the data we used for the first modeling analysis has 24 features with the structure of 4 genes * 2 T cell types * 3 time periods. To deduct genes dimension from this original strata, the most concise approach is to average ICS magnitude of the 4 genes. At this point, the sequence of averaging and log10 transformation should be considered. From mathematical view, averaging the log10 transformed ICS data do not hold the same biological meaning as to averaging the raw ICS data. In sake of reserving the original interpretation ability of the data, I decided to average the raw ICS magnitude first then do log10 transformation. The prepared data has 6 features as Table 7 recorded.



	prime CD4	prime CD8	boost CD4	boost CD8	pre-challenge CD4	pre-challenge CD8	protection outcome
1	-0.158015195	0.54592533	0.343408594	0.68282172	-0.68298190	0.30481354	1
2	0.094296397	-0.31875876	0.641721925	0.50852972	-0.33488826	-0.43179828	1
3	-0.099632871	0.17464119	0.520483533	0.25042000	-0.57267621	0.04824753	1
4	0.135927335	-0.63827216	0.388278863	-0.16272730	-0.58502665	-0.86169730	1
5	-0.138167002	-0.28608965	0.327869569	0.27357995	-0.73873713	-0.36401389	1
6	-0.240332155	0.03542974	0.139091608	0.15760785	-0.83120798	-0.28399666	1
7	-0.051831638	-0.34678749	0.586868492	-0.01211039	-0.78914663	-0.54515514	1
8	-0.428873723	-0.19044029	-0.003269485	0.14066514	-0.85387196	-0.12205305	0
9	0.096040554	-0.42021640	0.485721426	-0.13371266	-0.26962153	-0.54898155	0
10	-0.397940009	-0.54898155	0.108057374	-0.12784373	-0.77598519	-0.59773862	0

Table 7. Second Modeling Analysis Data Matrix

This table displays the first 10 observations of the 6 features included in second modeling analysis, with their binary protection outcomes. The protected and unprotected monkeys are indicated by 1s and 0s respectively.

3.3.2. Model Fitting Results

In consideration of one in ten rule[55], to conduct logistic regression without risks of over-fitting, the second modeling analysis on 6 features with 50% protection rate needs at least 10*6/0.5=120 observations. Given that there are only 91 observations in the training data set, feature selection by LASSO was performed previous to logistic regression. Similar to the LASSO analysis for the first modeling analysis, two LASSO methods, LARS and Elastic-

Net, have been performed on the 6 features data matrix. From LARS Cp plot (Figure 16.), Cp value has a remarkable drop when adding the 3rd feature. However, this drop did not diminish the Cp value to the minimum Cp. Although the Cp value of adding the 3rd feature is very close to the minimum Cp, the minimum Cp is held by the first step, which indicates adding 0 features into model. Due to the hardship of assessing this special tie of minimum Cp, it would be better to refer to LASSO results by Elastic-Net. As the summary plot showing (Figure 17.), the model suggests to select 3 features. As a result, by looking up the coefficients, the 3 features are prime CD4+, boost CD4+ and pre-challenge CD4+. Referring back to one of ten rule, our data with 91 observations satisfied the rule that, 3 features with 50% protection rate needs at least 10*3/0.5=60 observations.

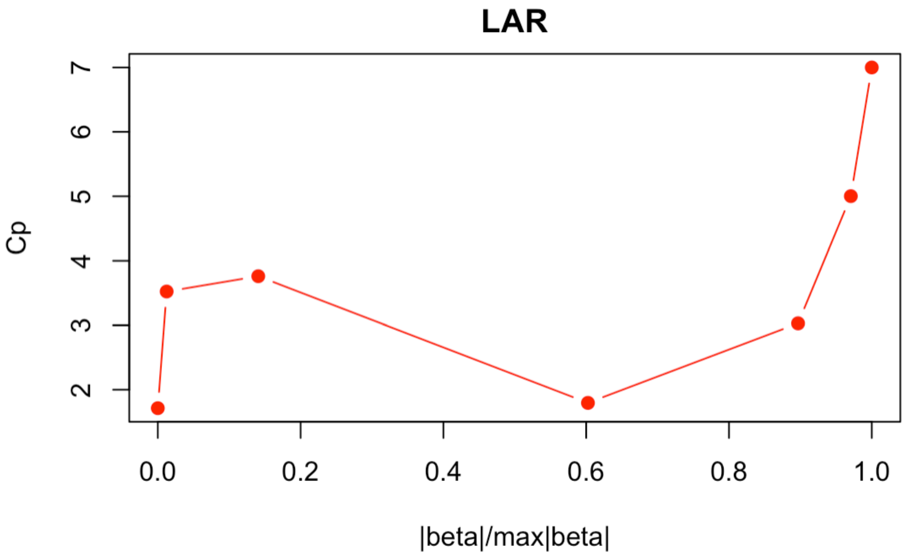


Figure 16. Mallows's Cp plot from CV-LARS LASSO analysis for the second modeling analyses The spread of red dots in this plot shows the Mallows's Cp value for adding in each of the 6 features. Basically, the best model should include all features before the step (inclusive) with the least Mallow's Cp value. Based on this, the least Cp value is held by step 0 which suggests to include 0 feature. However, the big drop at the 4th dot invokes further inspections using Elastic-Net LASSO.

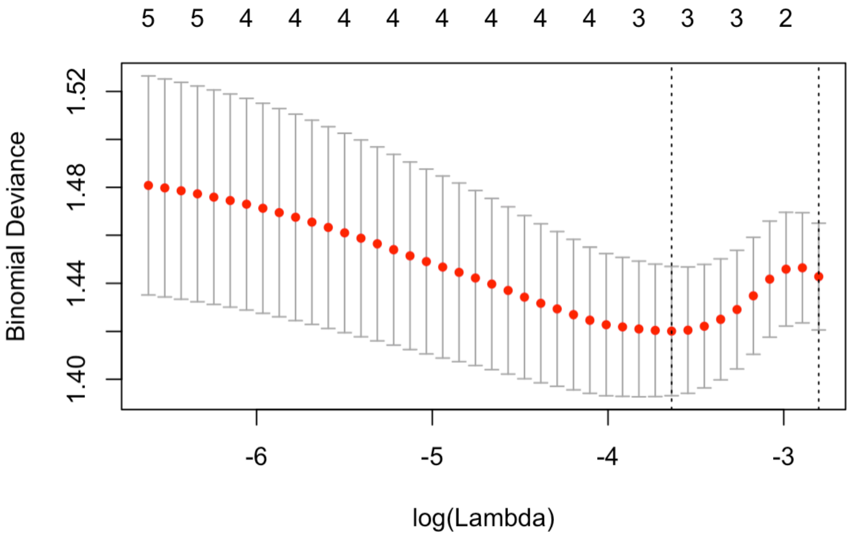
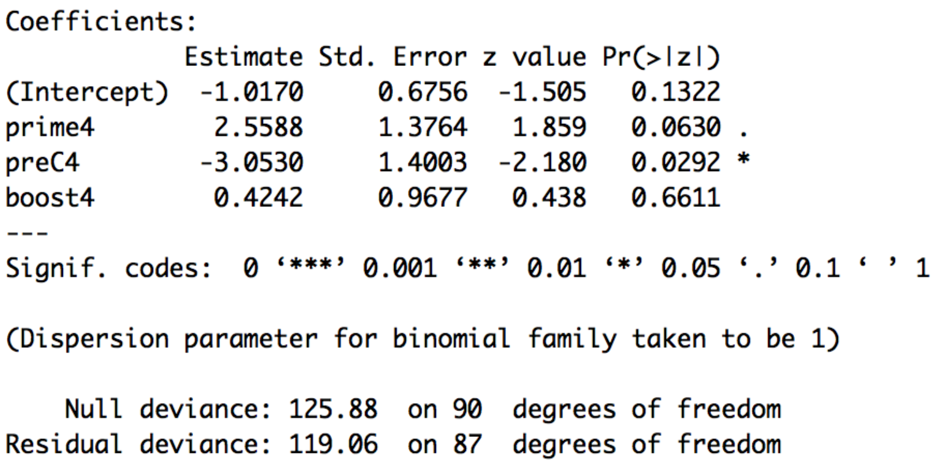


Figure 17. Elastic-Net LASSO Summary Plot for the Second Modeling Analysis

This summary plot gives a dash line which indicates the amount of features that should be included in modeling analysis, suggested by Elastic-Net LASSO model. The Elastic-Net LASSO model clearly suggested to include 3 features into modeling analysis.

At this point, feature dimension was deducted to 3 which satisfied one in ten rule for logistic regression. Hence, I conducted logistic regression on the protection outcome and the three features, prime CD4+/boost CD4+/pre-challenge CD4+. As result, with P<0.05 as threshold, both prime CD4+ (P=0.063) and pre-challenge CD4+ (P=0.029) are significant to be included in the model to describe the protection outcome. The boost CD4+ with P value 0.661 is not significant enough to be included in the model. (Figure 18) According to this result, I excluded boost CD4+ and conduct modeling on prime CD4+ and pre-challenge CD4+. This two-features model performed well in regard of significant P values for both features. (Figure 19)



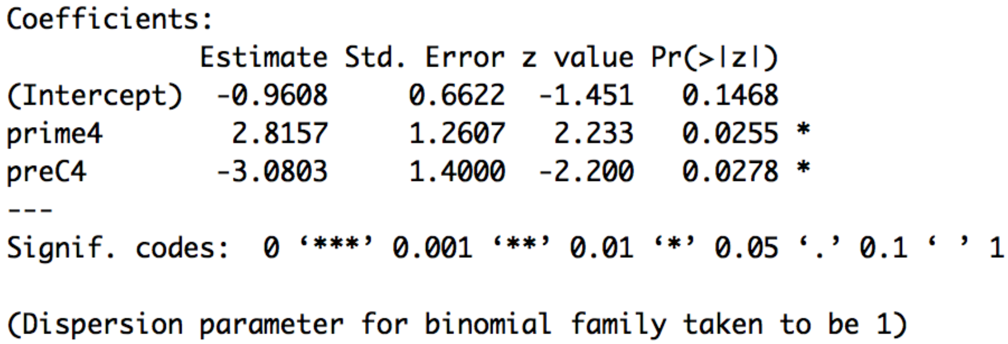
```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.0170     0.6756  -1.505   0.1322
prime4        2.5588     1.3764   1.859   0.0630 .
preC4        -3.0530     1.4003  -2.180   0.0292 *
boost4        0.4242     0.9677   0.438   0.6611
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 125.88  on 90  degrees of freedom
Residual deviance: 119.06  on 87  degrees of freedom
```

Figure 18. Model Summary of the Logistic Model with 3 Features

In this figure, prime4 refers to prime CD4+, boost4 refers to boost CD4+ and preC4 refers to pre-challenge CD4+.



```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.9608     0.6622  -1.451   0.1468
prime4        2.8157     1.2607   2.233   0.0255 *
preC4        -3.0803     1.4000  -2.200   0.0278 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

Figure 19. Model Summary of the Logistic Model with prime CD4+ and pre-challenge CD4+

In this figure, prime4 refers to prime CD4+ and preC4 refers to pre-challenge CD4+.

To further check the estimating stability of prime CD4+ and pre-challenge CD4+, I recorded model coefficients and calculated their confidence intervals (hereafter referred as CIs) by `confint` function in R package `stats`.[38] This `confint` function based on likelihood-ratio statistic by likelihood profiling is more accurate than traditional Wald method.[38] The 95% CI of the correlation provides estimate of where the estimated correlation, between pre-challenge CD4+, prime CD4+, interaction and outcome, could lie, in 95% of replicate experiments, if the true correlation were the estimated correlation. In order to interpret logistic model's coefficients and CIs, which are not as straightforward as those from linear model, I recorded both original and exponential coefficients and CIs in Table 8. According to the logit setting of logistic model, 1). Given another feature fixed, one feature's correlation direction can be indicated by the sign of original coefficient, 2). Given another feature fixed, one feature's correlation strength can be interpreted by the odd ratio of the feature which equals to the exponential coefficient of the feature.[43] Therefore, the results in Table 8 give the following 2 suggestions. First, indicated by the opposite signs of coefficients, given each other feature fixed, the pre-challenge CD4+ has negative effect on protection outcome and prime CD4+ has positive impact on protection outcome. Second, in terms of two features' odds ratios, the odds of being a protected animal would be 16.705 times (95% CI: 1.588 to 233.334) as likely to occur with one unit log10 ICS magnitude increase in prime CD4+. One unit log10 ICS magnitude increase in pre-challenge CD4+ would make the odds of being a protected animal 21.739 (95% CI: (1.548,333.333) times less likely to occur.

Model: outcome ~ prime CD4+ + pre-challenge CD4+				
	Coefficients	**95% CIs of Coefficients**	**Odds Ratios**	**95% CIs of Odds Ratios**
Prime CD4+	2.816	(0.462, 5.452)	16.705	(1.588, 233.334)
Pre-challenge CD4+	-3.080	(-5.979, -0.438)	21.739	(1.548,333.333)

Table 8. Model Summary on Coefficients/CIs

To sum up, the second modeling analysis manifested that the 2 immunologic responses, prime CD4+ and pre-challenge CD4+, might correlate with the protection outcome, hence, they might be capable to describe the protection outcome in a logistic model. The logistic model performs well under wide CI ranges and the correlation between prime CD4+, pre-challenge CD4+ and outcome is significant with qualified p-values. Besides checking on p-values and CIs, further assessments on the model quality and feature interactions were conducted in the model inspection section below.

3.3.3. Model Inspection

To inspect the quality of the logistic model, I adopted model evaluation tools to compare correlation significances of the logistic model on training data set. The two linear model evaluation methods, ANOVA and AIC, used in 2.3 are capable for logistic models evaluation as well. To begin with, I performed AIC on the 5 test models, in which Model 3 is the previous logistic model. Similar to the model selection process in 2.3, I constructed 5 test models to assess the 3 candidate features and recorded AIC summary in Table 9. The AIC scores indicates that test model 3 is the best model to minimize the information loss. Same as previous AIC assessment on log10 ICS magnitudes linear regression in section 2.3, AIC statistic is calculated by exp((AIC_{min}-AIC_i)/2) and recorded in Table 9. From the AIC statistics which show the probability that model X can minimize information loss as well as the best model, none of the other four models are close to the best model. As the result, the best model with least information lost under penalty of features amount, is Model 3. AIC confirmed good performance of the logistic model by model comparison.

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
Features Included	**pre-challenge CD4+**	**prime CD4+**	**pre-challenge CD4+ and prime CD4+**	**pre-challenge CD4+, prime CD4+ and boost CD4+**	**Constant (no feature included)**
***AIC score**	128.8	128.5	125.3	127.1	127.9
****AIC statistic (Comparing all models)**	0.03	0.04	1 (minimum AIC)	0.17	0.07

Table 9. Five Test Models Descriptions and AIC Summary

This table collected AIC score and computed AIC statistics. The model with minimum AIC score is the best model with least information lost under penalty of the amount of variables included in the model. AIC statistics suggest the probability that model X can minimize information loss as well as the best model.

Besides AIC, I conducted ANOVA to check the significance of 1). including both two features in the logistic model; 2). including pre-challenge CD4+ in the logistic model; 2). including pre-challenge CD4+ in the logistic model. The preliminary logistic model with the 2 features was regarded as the NULL hypothesis model for all comparisons. ANOVA results of these comparisons were recorded in Table 10. First, I compared the NULL model which only includes a constant, to the two-features logistic model. The p value (0.0364) from this comparison suggested that the preliminary logistic model, is significant better than the NULL model. Second, I compared the NULL model, which include only Prime CD4+, to the two-features logistic model. The p value (0.0218) suggested that the two-features logistic model including one more feature, pre-challenge CD4+, is significant better than NULL model with only prime CD4+. Third, similarly, I compare the NULL model which include only pre-challenge CD4+, to the two-features logistic model. The p value (0.0181) suggested that the two-features logistic model including one more feature, prime CD4+, is significant better than NULL model with only pre-challenge CD4+. To sum up, both prime CD4+ and pre-challenge CD4+ in the preliminary logistic model are significant to describe the protection outcome. This preliminary logistic model is worth being analyzed by further model validation process.

Comparision 1: Checking on both features	NULL Hypothesis Model	model
Features Included	~ 1	~ PrimeCD4+ + Pre-ChallengeCD4+
Degree of Freedom (Df)	90	88
Degree of Freedom Change	2	
P value	0.0364*	

Comparision 2 Checking on PrimeCD4+	NULL Hypothesis Model	model
Features Included	~ PrimeCD4+	~PrimeCD4+ + Pre-ChallengeCD4+
Degree of Freedom (Df)	89	88
Degree of Freedom Change	1	
P value	0.0218*	

Comparision 3 Checking on PrimeCD4+	NULL Hypothesis Model	model
Features Included	~ Pre-challengeCD4+	~ PrimeCD4+ + Pre-ChallengeCD4+
Degree of Freedom (Df)	89	88
Degree of Freedom Change	1	
P value	0.0181*	

Table 10. Three Comparisons ANOVA Table (by LR)

This table recorded results from 3 ANOVA likelihood ratio tests comparing 1). the significance of including both two features in the logistic model; 2). the significance of including pre-challenge CD4+ in the logistic model; 2). the significance of including pre-challenge CD4+ in the logistic model. The P values show the probability of seeing pre-challengeCD4+ (or primeCD4+, or both pre-challengeCD4+ and primeCD4+) is not important in the logistic model.

3.3.4 Features Interaction Analysis

During ANOVA comparison, I noticed that models including either prime CD4+ or pre-challenge CD4+ is not significant better comparing to the zero model. (Table 11) In other words, the single feature by its own cannot describe the protection outcome. However, the two features together can describe the protection outcome on some levels. This result motivated further analysis on the interaction between prime CD4+ and pre-challenge CD4+.

Comparation 1:	**zero model (NULL Hypotesis Model)**	**single feature model 1**
Features Included	**~ 1**	**~ PrimeCD4+**
Degree of Freedom (Df)	90	89
Degree of Freedom Change	1	
P value	0.2437	
Comparation 2:	**zero model (NULL Hypotesis Model)**	**Single feature model 2**
Features Included	**~ 1**	**~ Pre-ChallengeCD4+**
Degree of Freedom (Df)	90	89
Degree of Freedom Change	1	
P value	0.3063	

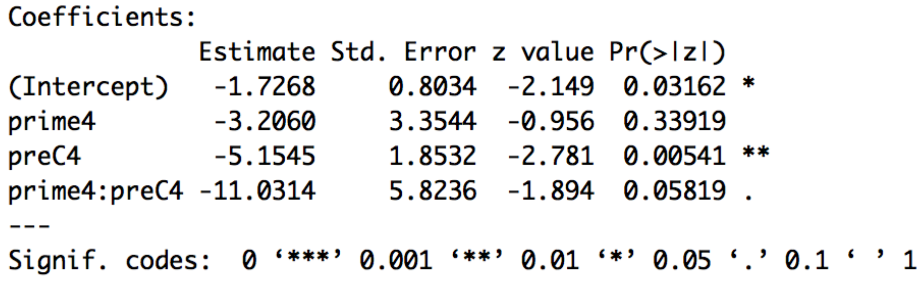
Table 11. ANOVA table for Single Feature Exploration.

This table recorded results of 2 ANOVA comparison on 2 single feature model. In result, neither prime CD4+ or pre-challenge CD4+ is not significant comparing to the zero model.

Based on the single feature model exploration result, I inspected the interaction between pre-challenge CD4+ and prime CD4+ by interaction-involved logistic regression. After adding in interaction, p-values of the model suggest a stronger significance on pre-challenge CD4+ and rejects prime CD4+. The interaction between the two features is considered as closing to significant. (Figure 20.)

To further validate this interaction pattern, ANOVA likelihood ratio tests were conducted to accomplish 2 comparisons. First, I compared the 2-features model with the interaction-involved 2-features model. This comparison was designed to validate the significance of the interaction. As a result, p value (0.0524) confirmed weak significance of the interaction. (Figure 21) Second, considering that the strong significance of pre-challenge CD4+ in the interaction-involved two-features model, indicated by p value (0.0054) in Figure 20, the second comparison was conducted to validate that whether pre-challenge CD4+ itself is sufficient and even better than including interaction and both prime CD4+ and pre-challenge CD4+. I compared the model with only pre-challenge CD4+ to the interaction-involved 2-features model. In result, the model with both features and interaction works much better than the model only with pre-challenge CD4+. (Figure 22.)

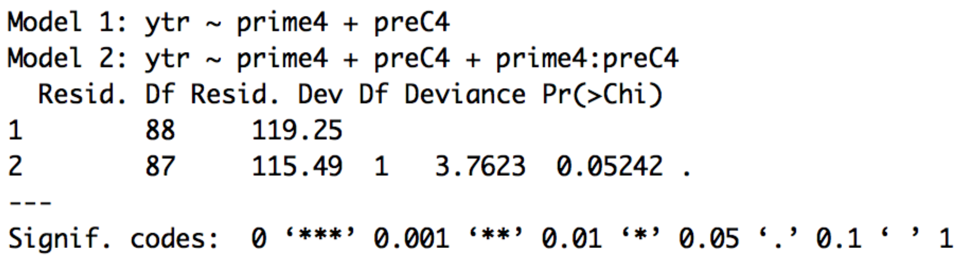
In Figure 19-21, prime4 refers to prime CD4+ and preC4 refers to pre-challenge CD4+. The term prime4:preC4 refers to interaction between prime CD4+ and pre-challenge CD4+.



```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.7268     0.8034  -2.149  0.03162 *
prime4        -3.2060     3.3544  -0.956  0.33919
preC4         -5.1545     1.8532  -2.781  0.00541 **
prime4:preC4 -11.0314     5.8236  -1.894  0.05819 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

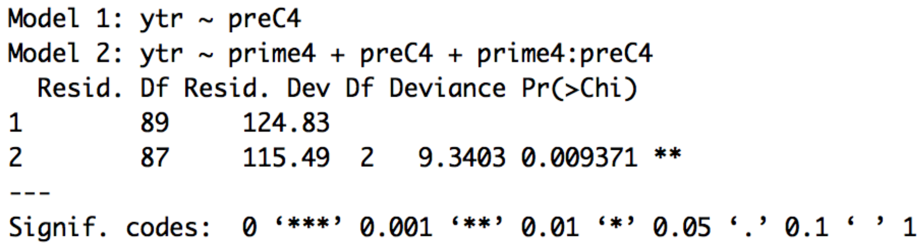
Figure 20. Interaction-involved Logistic Model Summary

After adding in interaction, the model has a stronger significance on pre-challenge CD4+ but rejects prime CD4+. The interaction pattern in this model is very weak and closing to significant.



```
Model 1: ytr ~ prime4 + preC4
Model 2: ytr ~ prime4 + preC4 + prime4:preC4
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        88     119.25
2        87     115.49  1   3.7623  0.05242 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 21. LR ANOVA Summary for comparison between interaction-involved model and the original two-features model. The original two-features model (Model 1 on the top of this figure) acted as the NULL hypothesis model in this ANOVA comparison. This comparison confirmed that interaction pattern is weak.



```
Model 1: ytr ~ preC4
Model 2: ytr ~ prime4 + preC4 + prime4:preC4
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        89     124.83
2        87     115.49  2   9.3403 0.009371 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 22. LR ANOVA Summary for comparison between the model with only pre-challenge CD4+ and the interaction-involved model. Model 1 on the top of this figure acted as the NULL hypothesis model in this ANOVA comparison. This comparison manifested that although pre-challenge CD4+ is significant in interaction involved logistic model, having both features and interaction is significant better than just having pre-challenge CD4+.

In short, the interaction results indicate that 1). the significance patterns of the two features were altered by adding-in interaction which weakly support the prediction model; 2). Pre-challenge CD4+ with smaller p-values is more significant after adding in interaction; 3). although the strong significance on pre-challenge CD4+, it is not sufficient to only include pre-challenge CD4+ for the prediction model. To sum up, the interaction has certain effect on the original patterns of the two features. Inspection on the interaction mechanism requires advanced interaction analyses which could be conducted as another project in the future. For this project, I decided to continue model validation on the original logistic model without interaction. These indications elicit further advanced exploration on the interaction decomposition to identify the certain impact from the weakly supported interaction on the two features.

3.4. Final Model Validation

At this stage of model development, I have identified 2 potential immunologic correlates, pre-challenge CD4+, prime CD4+, and a logistic model to describe their relations to the protection outcome. To further explore this model, I conducted preliminary model validation test to check the model performance on test data set.

The model predictability test exploited the logistic model fit on the test data set by plotting receiver operating characteristic curves (hereafter referred as ROC curve). In machine learning process, ROC curve is widely used to provide the model predictability by comparison of the true positive rate (TPR) against the false positive rate (FPR) as the discrimination changes.[56] The ROC curve which locates at up-left corner indicates better model prediction. With this regard, the area under the ROC curve (hereafter referred as AUC) can be used as a reference that if the AUC value is larger than 0.7, then the model has a good prediction rate. [57]

After fitting the logistic model to the test data set, I used `roc` function in R package `pROC` to plot the ROC curve with AUC value and max AUC polygon.[58] From Figure 23, this logistic model did not give good predictions on the test data set.

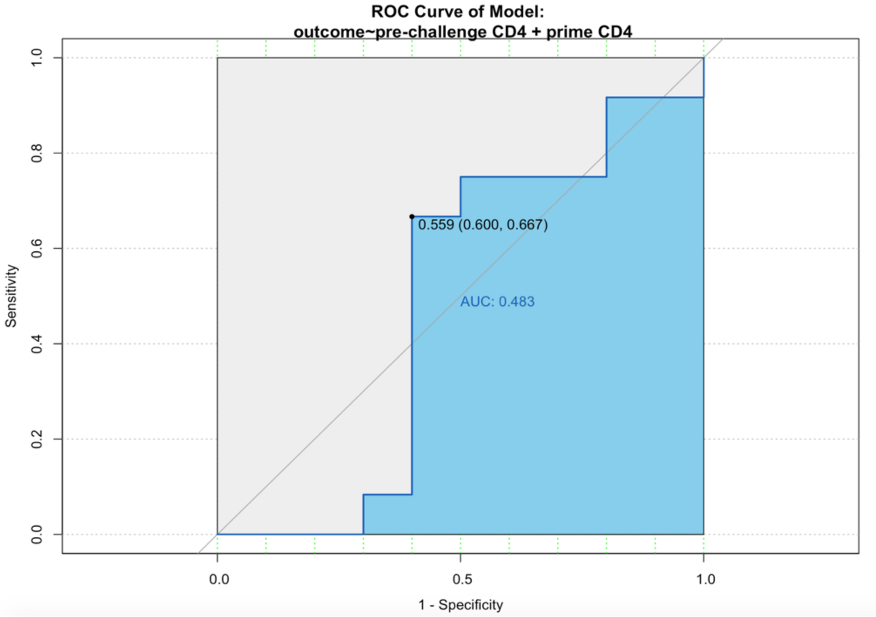


Figure 23. ROC curve of the logistic model: outcome ~ pre-challenge CD4+ + prime CD4+

This figure shows the overall AUC value (0.483) in blue and the maximum AUC angle in black. The predictions for the first several observations are very bad.

As an attempt to explain the poor performance, I plotted the model's ROC curve on training data set. (Figure 24.) From Figure 24, although the model gave a better ROC curve on training data set, the predictive ability of the model is weak indicated by the AUC with value 0.645. Looking back to the logistic model summary in Figure 19, I identified a potential reason of this modest predictive performance. Although the p-values satisfied the p<0.05 threshold, the significances are relatively weak. The modest predictive performance of the model might be attribute to the modest significance level of the two features from the logistic model summary.

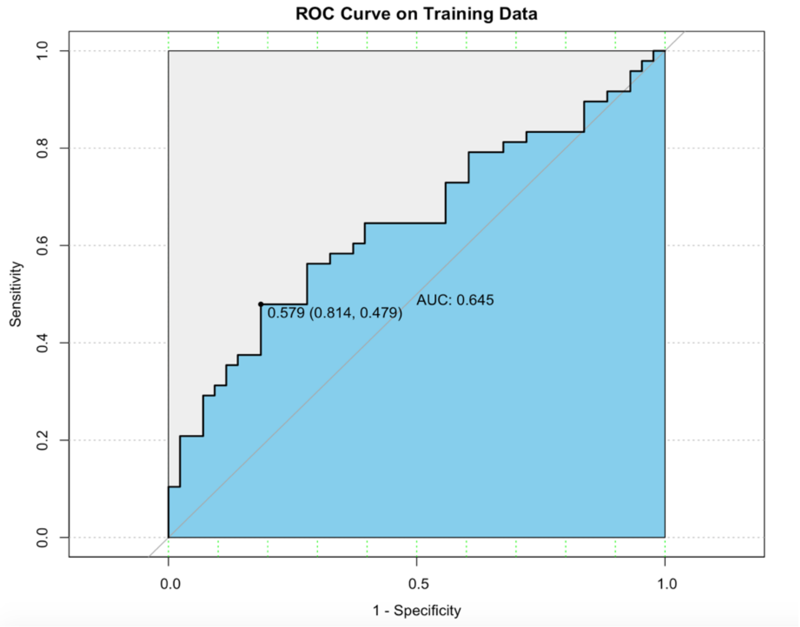


Figure 24. ROC curve of the interaction-involved logistic model

This figure shows the overall AUC value (0.645) in the middle and the maximum AUC angle. In general, the ROC curve is slightly above the right down triangle.

With the negative results from ROC curves, I attempted to further confirm the model predictability by directly visualize the model. The primary aim of this visualization process is to intuitively identify the protection outcome differentiation with/without the two features. To begin with, I visualized the central distributions of the conditional outcomes by plotting a boxplot of outcome labeled pre-challenge CD4+ and prime CD4+ (Figure 25.). Indicated by the median ICS distributions (the black lines inside the boxes), the sample medians are slightly differentiated across protection outcomes for pre-challenge CD4+. However, for prime CD4+, sample medians are almost the same for both outcomes. Moreover, the median itself is insensitive to observations in the tails. If the important correlation patterns were latent in the tails, the boxplot by median distribution would not be capable to capture and visualize the correlation. With this regard, I compared mean values of log10 ICS for each group. In result, the means are even closer across outcomes than the medians. To sum up, the central distributions, such as mean and median, of the two features are incapable to distinguish the protection outcome. This central distribution pattern corresponds to the model's poor predictive performance. Same as the ROC curves, this central distribution pattern might be attributed to the weak correlation significance of the two features.

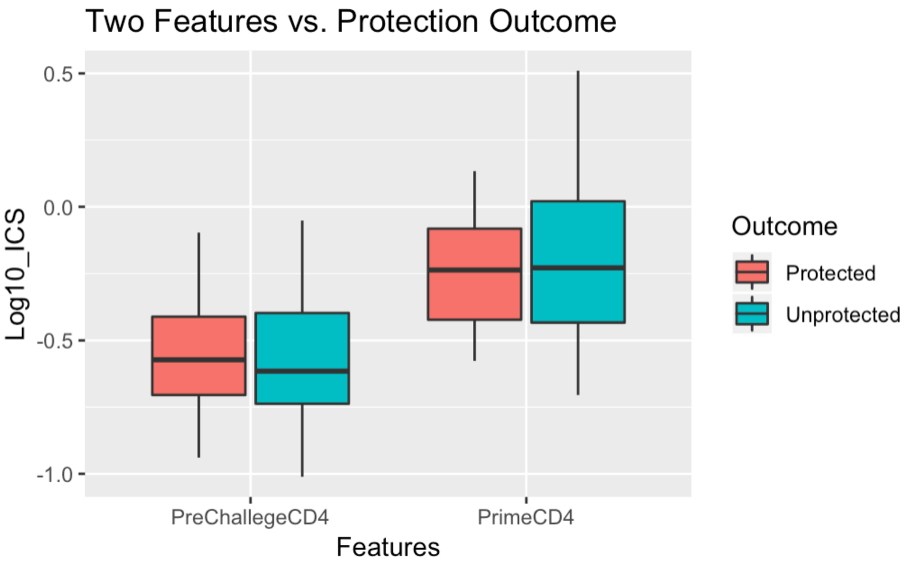


Figure 25. Boxplot of Outcome Labeled Features

Black lines inside each box are the medians of log10 ICS magnitudes for each group. The edge of box indicates 1st ~ 3rd quartile range of log10 ICS magnitudes for each group. The black bar outside of the box indicates the maximum and minimum log10 ICS magnitudes for each group.

In addition to visualize the distribution, I visualized the correlated distribution of pre-challenge CD4+ and prime CD4+ colored by protection outcomes. First, the two features are collective in one direction which can be summarized by both linear regression and LOESS non-linear regression. (Figure 26.) However, if separately inspect the two outcome groups, the non-linear LOESS captured more data trends than the linear regression. (Figure 27.) This result might suggest that there could be some non-linear components inside the correlation between outcome and the two immunologic features. Further suggested by this, another trigger of the poor model predictability could be that the linear logistic model cannot capture the non-linear relations latent in the immunological mechanism.

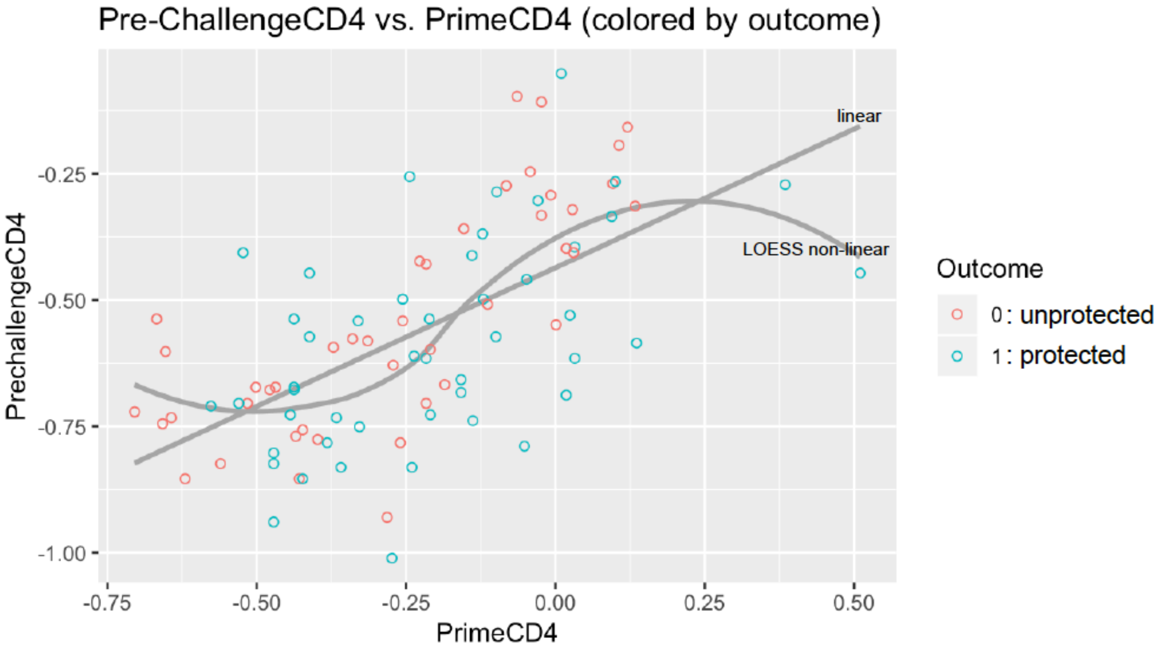


Figure 26. Correlated Distribution of the 2 Features with Linear and Non-linear Regression Lines

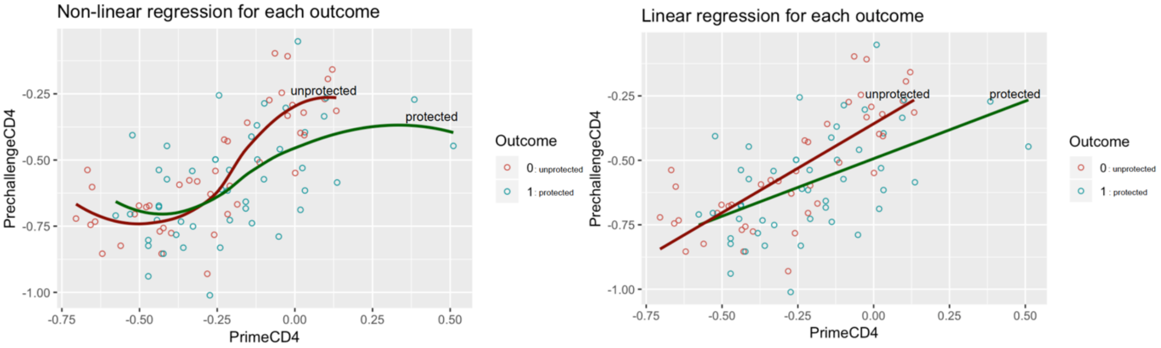


Figure 27. Linear vs. LOESS Non-linear Regression Lines for Each Outcome

The LOESS lines in left plot better describe cytomegalovirus dots trend than linear regression lines in right plot.

Second, as for the outcome differentiation, the outcome cannot be distinguished on the overall scale. However, there might be some patterns in two particular regions. One example could be that, when looking at the data in right region in Figure 28, the unprotected monkeys tend to have higher pre-challenge CD4+ log10 ICS magnitude when present-percentage of prime CD4+ is over -0.20. Besides, it might be worthwhile to inspect those unprotected monkeys with low prime CD4+ present-percentage in right region at the down left corner. If bring in previous suggestion about non-linear correlation, analysis on a typical region can be regarded as a process to extract the linear components from the non-linear correlation.

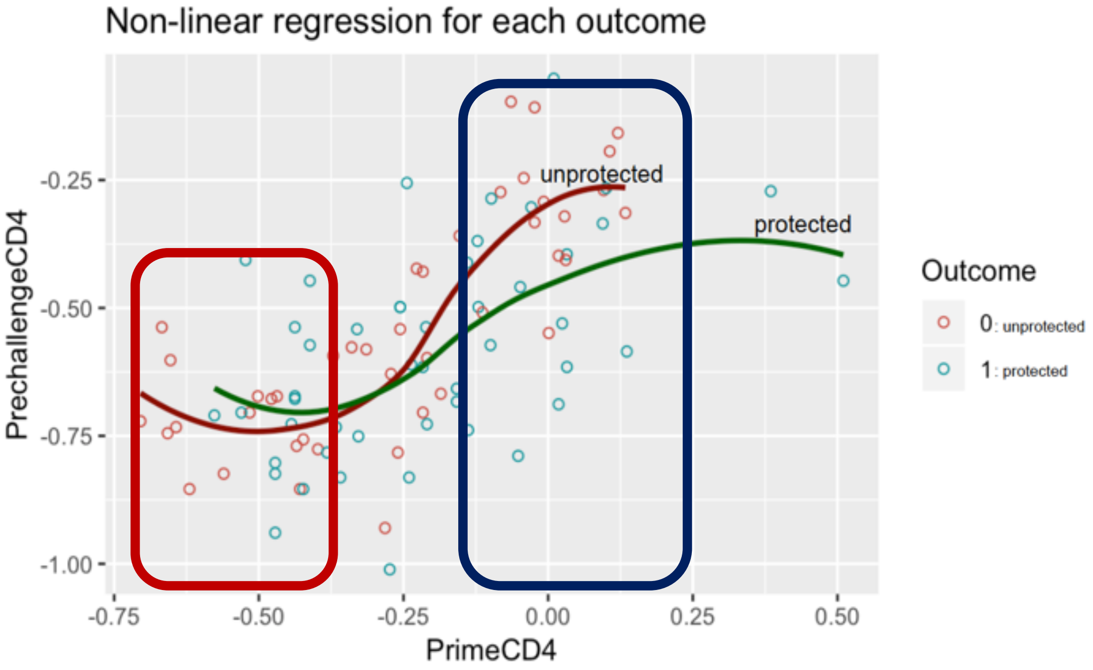


Figure 28. Two potential regions of interests

The two regions circled by red and blue round boxes have typical patterns corresponding to protection outcome. Observations in the left region with low prime CD4+ log10 ICS magnitude are mainly unprotected. In the right region, unprotected monkeys have higher pre-challenge CD4+ log10 ICS magnitude than protected monkeys.

To sum up, the two model visualizations further validated the poor predictive performance of the logistic model. Besides, the visualization patterns also gave suggestions on the non-linear possibility and data stratification which could potentially provide explanation to current model's poor predictability. Moreover, the non-linear possibility might relate to the special interaction patterns from previous interaction analysis. Non-linear modeling and stratification could be potential approaches to decompose and inspect the interaction. Even though the model validation provided negative results, these results also initiate valuable future researches to optimize the model by interaction inspection and non-linear modeling. Up to this point, according to model inspection and model validation results, the current logistic model has the potential to describe the protection outcome but is not sufficient to accurately predict the outcome.

Discussion

Data Normalization

In this study, except for a z scale standardization was done for PCA, I did not conduct normalization for other analyses in consideration of: 1). features having same unit and similar ICS magnitude ranges; 2). limited understanding of the original data structure; 3). potential loss of information latent in the distance due to crude normalization scalar; 4). potential interpretation lost caused by unit change. However, without normalization, some meaningless data discrepancy introduces bias to statistical analyses.

When computing the AIC scores for linear model selection in section 2.3, the big AIC score drop by adding in study variable indicated a remarkable data discrepancy across the 6 studies. To further inspection this discrepancy, I visualized the distributions for the 6 study cohorts (Figure 29.) As a result, the boxplots showed a remarkable trend of median ICS magnitudes across different studies. This discrepant median trend implied the potential inconsistency in wet laboratory settings across different studies, which challenged our assumption that all monkeys have the same start point. The variance among different study cohorts might disturb statistical analysis on the real meaningful variance or even hinder the identification of real correlates.

With a clear understanding that this variance across study cohorts is not expected and could potentially bring in confounding factor with variation unrelated to what we are studying, I shared this information with the data providers, Louis Picker's group, to search for their authentication on this trend from immunological standpoint. As a result, since such trend across study cohorts have never been seen and managed before, they suggested to keep working on original data without normalization. However, from statistics standpoint, normalization on such discrepancy which could potentially improve model performance, is still worth trying. Based on these, I regard the data normalization on study cohorts as one potential next step and put some preliminary efforts on normalization in the study discussion.

I tried to identify a good normalization method to accomplish two aims: 1). scale the ICS values in different study cohorts to the same level; 2). at the same time prevent loss of other variances during normalization process. With the two-folded aims, one normalization method I have attempted is to extract residuals of the linear regression model which regressed all ICS magnitudes to the 6 study cohorts, and use the residuals as normalized ICS magnitudes. The regression residuals, by statistical definition, reflects the rest variance except the variance in mean trend. Taking advantage of regression process, I could easily use residuals to exclude the mean variance which was isolated and concentrated along the regression line. In result, corresponding to the first normalization aim, this normalization process effectively adjusted the trend of ICS magnitudes when comparing the 6 studies before and after normalization (Figure 29.).

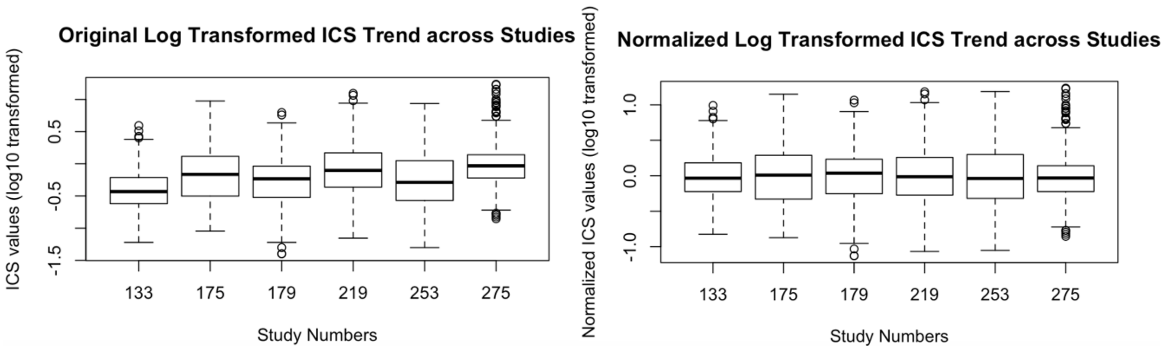


Figure 29. Log10 Transformed ICS magnitudes trend before (left) and after (right) normalization.

Whereas, in terms of the second aim to keep other potentially useful variances unaffected thereby, this normalization process left some uncertainties, which is the main reason for not including this normalization process into the main line of this study. The accuracy of normalizing and only normalizing on study variable depends on the fitness of the linear regression model. In this regard, I checked the linear regression model fitting by QQ-plot. The linear regression fitting turned out to be satisfactory. (Figure 30)

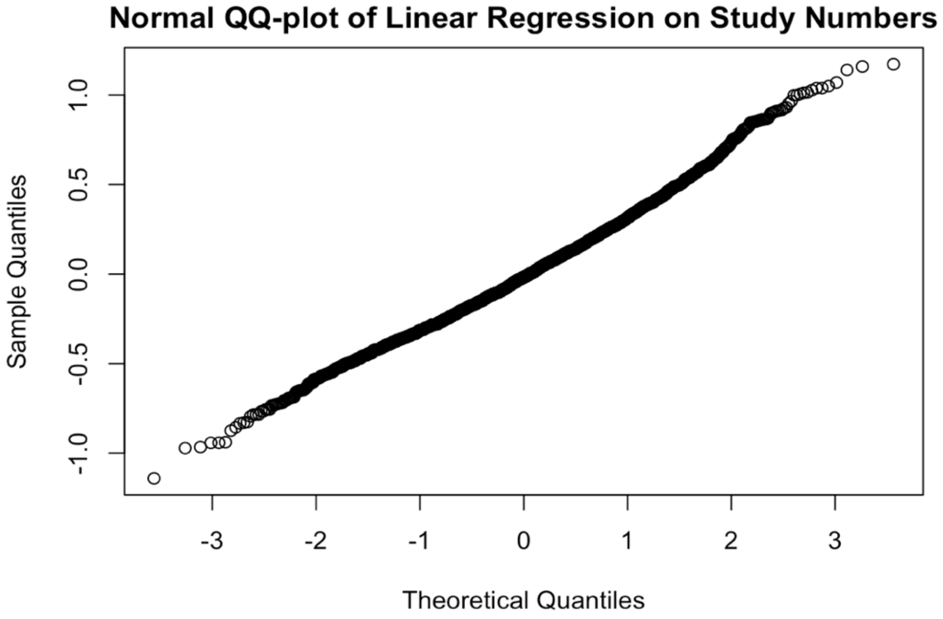


Figure30. Normal QQ-plot of Linear Regression on Study Cohorts

After several attempts on using normalized data to redo modeling analyses, I found some inconsistent modeling results that, using normalized data, the second modeling analysis failed to produce any significant correlate, when the first modeling analysis gave that prime CD4+ and pre-challenge CD8+ are significant (data not shown). Given that even the redundant data in first modeling analysis could produce two significant features, the failure in second modeling analysis with less redundant data is hard to explain. To be clear, I redo modeling analyses after all unsupervised analyses, which ensures this check did not introduce any fishing concerns. After tradeoff between the uncertain normalization method and the ICS trend in original data and based on the uncertainty in choice of normalization method and the inconsistent results, I decided to stick on using the original data to statistical analyses in this study. After this study, one potential research direction after this study should be identifying an advanced normalization method and normalizing the ICS trend on study cohorts properly, and hence, optimizing the whole statistic modeling.

Potential Biological Illustration of the Logistic Model

From the model coefficients direction in Table 8, the logistic model conveyed a remarkable information that pre-challenge CD4+ and prime CD4+, given each other, are oppositely correlated to the protection outcome. In detail, given fixed pre-challenge CD4+, one-unit log10 ICS magnitude increase of prime CD4+ will make the odds of being a protected animal 16.705 times as likely to occur (95% CI: 1.588 to 233.334). However, given fixed prime CD4+, one-unit increase of pre-challenge CD4+ percentage will decrease the protection possibility by 21.739 times (95% CI: 1.548 to 333.333). According to the timeline of immunological response, prime CD4+ percentages were recorded at the start period, while pre-challenge CD4+ percentages were recorded at the later maturing period. Therefore, the primary interpretation of this opposite correlation pattern could be that, 1). the immediate immune response upon first immunization will promote later protection; 2). given positive prime CD4+ response, early matured immune CD4+ responses in pre-challenge time period will arrest the protection occurrence.

The second model interpretation corresponds to the B cell follicle sanctuary mechanism, which is a SIV viral immune escape mechanism demonstrated by Louis J. Picker and his colleges.[59] Louis Picker and his colleges demonstrated that the formation of the B cell follicles, which is called germinal centers alternatively, served as sanctuaries for productive SIV infected cells even under potent SIV-specific immune responses. Moreover, the formation of the B cell follicles is corresponding to the immune response maturing exactly at pre-challenge period. Based on this, the higher CD4+ percentage at pre-challenge time period indicates earlier mature of immune responses, and potentially brings greater formation of the B cell follicles.[59] This could be one way in which the higher pre-challenge CD4+ percentage negatively influence the protection outcome, if this turned out to be supported by subsequent analysis.

These potential biological illustrations upon prime CD4+ and pre-challenge CD4+ attaches further meanings to the logistic model, which might assist future model optimization. Especially, viral immune escape mechanisms will help to track the paths of the 2 features in future model interaction inspection. At the meantime, the immunologic correlates modeling results from this study will assist future biological researches to identify more details in these immunological mechanisms.

Limitation

Although the overall T cell responses data used in this study provided a comprehensive source for preliminary immunologic correlates identification, the study was limited as the data is too crude to explore the conventional and unconventional restricted responses and targeted epitopes. This limits exploration on unconventional CD8+ responses and further exploration on the non-linear effects of the two CD4+ significant responses. To address this limitation, a complementary data was expected to profile restricted T cell responses for targeted epitopes under more continues time periods. Such complimentary data would assist to identify significance in unconventional CD8+ responses and to perform advanced interaction inspections on the nonlinear relations between the two CD4+ responses.

Conclusion

In conclusion, I identified pre-challenge CD4+ and prime CD4+ as 2 potential immunologic correlates of the RhCMV/SIV vaccine efficacy. The 2 immunologic responses have the potential to describe the unique ~50% protection pattern. Although these 2 immunological responses could not provide satisfactory predictability in a preliminary logistic model, the results of model validation provided concrete future research directions on non-linear modeling and interaction analysis to better characterize the 2 immunologic responses in the statistical model. The identification of pre-challenge CD4+ and its interaction with prime CD4+ will promote some related vaccine immunological mechanism researches, such as researches on germinal centers formation and HIV immune escape mechanism at pre-challenge time period.

One success of this study is the interpretability in machine learning process. In this study, the high dimensional immunologic response data with 24 features has been properly and carefully processed to preserve original interpretability while deducting redundant components. From informatics prospective, since this study was established upon the union of immunologists and statisticians, the interpretability ensured efficient information sharing. Besides, the interpretability in the machine learning process assisted illustration of model inspection on coefficients, interactions, predictability and visualizations. Moreover, such interpretability contributed to demonstrate interpretable and concrete future research directions. Giving credit to the interpretability, this study provided future research directions on interaction mechanisms of prime CD4+ and pre-challenge CD4+ and non-linear modeling on immunologic correlates.

References

1. Goulder PJ, Watkins DI: **Impact of MHC class I diversity on immune control of immunodeficiency virus replication**. *Nat Rev Immunol* 2008, **8**(8):619-630.
2. Picker LJ, Hansen SG, Lifson JD: **New paradigms for HIV/AIDS vaccine development**. *Annual review of medicine* 2012, **63**:95-111.
3. Fruh K, Picker L: **CD8+ T cell programming by cytomegalovirus vectors: applications in prophylactic and therapeutic vaccination**. *Current opinion in immunology* 2017, **47**:52-56.
4. Jarvis MA, Hansen SG, Nelson JA, Picker LJ, Früh K: **Vaccine Vectors Using the Unique Biology and Immunology of Cytomegalovirus**. In: *Cytomegaloviruses: From Molecular Pathogenesis to Intervention* Edited by Reddehase MJ, vol. 2, 2nd Edition of CYTOMEGALOVIRUSES: Molecular Biology and Immunology (Caister Academic Press, 2006) edn: Caister Academic Press; 2013.
5. Lodoen MB, Lanier LL: **Viral modulation of NK cell immunity**. *Nat Rev Microbiol* 2005, **3**(1):59-69.
6. Farrell H, Degli-Esposti M, Densley E, Cretney E, Smyth M, Davis-Poynter N: **Cytomegalovirus MHC class I homologues and natural killer cells: an overview**. *Microbes Infect* 2000, **2**(5):521-532.
7. Hansen SG, Sacha JB, Hughes CM, Ford JC, Burwitz BJ, Scholz I, Gilbride RM, Lewis MS, Gilliam AN, Ventura AB *et al*: **Cytomegalovirus vectors violate CD8+ T cell epitope recognition paradigms**. *Science* 2013, **340**(6135):1237874.
8. Hansen SG, Ford JC, Lewis MS, Ventura AB, Hughes CM, Coyne-Johnson L, Whizin N, Oswald K, Shoemaker R, Swanson T *et al*: **Profound early control of highly pathogenic SIV by an effector memory T-cell vaccine**. *Nature* 2011, **473**(7348):523-527.
9. Lewin SR, Evans VA, Elliott JH, Spire B, Chomont N: **Finding a cure for HIV: will it ever be achievable?** *Journal of the International AIDS Society* 2011, **14**:4.
10. Chun TW, Fauci AS: **HIV reservoirs: pathogenesis and obstacles to viral eradication and cure**. *AIDS (London, England)* 2012, **26**(10):1261-1268.
11. Deeks SG, Walker BD: **Human immunodeficiency virus controllers: mechanisms of durable virus control in the absence of antiretroviral therapy**. *Immunity* 2007, **27**(3):406-416.
12. UNAIDS: **Global HIV & AIDS statistics - 2018 fact sheet**. In: *UNAIDS Fact Sheet.* 2018.
13. Barouch DH, Picker LJ: **Novel vaccine vectors for HIV-1**. *Nat Rev Microbiol* 2014, **12**(11):765-771.
14. McMichael AJ, Picker LJ: **Unusual antigen presentation offers new insight into HIV vaccine design**. *Current opinion in immunology* 2017, **46**:75-81.
15. Parks CL, Picker LJ, King CR: **Development of replication-competent viral vectors for HIV vaccine delivery**. *Current opinion in HIV and AIDS* 2013, **8**(5):402-411.
16. Harmon TM, Fisher KA, McGlynn MG, Stover J, Warren MJ, Teng Y, Naveke A: **Exploring the Potential Health Impact and Cost-Effectiveness of AIDS Vaccine within a Comprehensive HIV/AIDS Response in Low- and Middle-Income Countries**. *PLoS One* 2016, **11**(1):e0146387.
17. Stover J, Hallett TB, Wu Z, Warren M, Gopalappa C, Pretorius C, Ghys PD, Montaner J, Schwartlander B, New Prevention Technology Study G: **How can we get close to zero?**

The potential contribution of biomedical prevention and the investment framework towards an effective response to HIV. *PLoS One* 2014, **9**(11):e111956.

18. Filippone C, de Oliveira F, Betsem E, Schaeffer L, Fontanet A, Lemee V, Gessain A, Plantier JC: **Simian Immunodeficiency Virus seroreactivity in inhabitants from rural Cameroon frequently in contact with non-human primates**. *Virology* 2017, **503**:76-82.
19. Chen Z: **Monkey Models and HIV Vaccine Research**. *Advances in experimental medicine and biology* 2018, **1075**:97-124.
20. Fukazawa Y, Lum R, Okoye AA, Park H, Matsuda K, Bae JY, Hagen SI, Shoemaker R, Deleage C, Lucero C *et al*: **B cell follicle sanctuary permits persistent productive simian immunodeficiency virus infection in elite controllers**. *Nature medicine* 2015, **21**(2):132-139.
21. Hansen SG, Vieville C, Whizin N, Coyne-Johnson L, Siess DC, Drummond DD, Legasse AW, Axthelm MK, Oswald K, Trubey CM *et al*: **Effector memory T cell responses are associated with protection of rhesus monkeys from mucosal simian immunodeficiency virus challenge**. *Nature medicine* 2009, **15**(3):293-299.
22. Okoye A, Park H, Rohankhedkar M, Coyne-Johnson L, Lum R, Walker JM, Planer SL, Legasse AW, Sylwester AW, Piatak M, Jr. *et al*: **Profound CD4+/CCR5+ T cell expansion is induced by CD8+ lymphocyte depletion but does not account for accelerated SIV pathogenesis**. *J Exp Med* 2009, **206**(7):1575-1588.
23. **Development of Attenuated CMV Vectors for an HIV/AIDS Vaccine**
24. Hansen SG, Piatak M, Jr., Ventura AB, Hughes CM, Gilbride RM, Ford JC, Oswald K, Shoemaker R, Li Y, Lewis MS *et al*: **Immune clearance of highly pathogenic SIV infection**. *Nature* 2013, **502**(7469):100-104.
25. Powers C, DeFilippis V, Malouli D, Fruh K: **Cytomegalovirus immune evasion**. *Curr Top Microbiol Immunol* 2008, **325**:333-359.
26. Fukazawa Y, Park H, Cameron MJ, Lefebvre F, Lum R, Coombes N, Mahyari E, Hagen SI, Bae JY, Reyes III MD: **Lymph node T cell responses predict the efficacy of live attenuated SIV vaccines**. *Nature medicine* 2012, **18**(11).
27. Thome JJ, Farber DL: **Emerging concepts in tissue-resident T cells: lessons from humans**. *Trends Immunol* 2015, **36**(7):428-435.
28. Cicin-Sain L, Sylwester AW, Hagen SI, Siess DC, Currier N, Legasse AW, Fischer MB, Koudelka CW, Axthelm MK, Nikolich-Zugich J *et al*: **Cytomegalovirus-specific T cell immunity is maintained in immunosenescent rhesus macaques**. *J Immunol* 2011, **187**(4):1722-1732.
29. Sylwester AW, Mitchell BL, Edgar JB, Taormina C, Pelte C, Ruchti F, Sleath PR, Grabstein KH, Hosken NA, Kern F *et al*: **Broadly targeted human cytomegalovirus-specific CD4+ and CD8+ T cells dominate the memory compartments of exposed subjects**. *J Exp Med* 2005, **202**(5):673-685.
30. Price DA, Sewell AK, Dong T, Tan R, Goulder PJ, Rowland-Jones SL, Phillips RE: **Antigen-specific release of beta-chemokines by anti-HIV-1 cytotoxic T lymphocytes**. *Curr Biol* 1998, **8**(6):355-358.
31. Holling TM, Schooten E, van Den Elsen PJ: **Function and regulation of MHC class II molecules in T-lymphocytes: of mice and men**. *Hum Immunol* 2004, **65**(4):282-290.
32. Braud VM, Allan DS, McMichael AJ: **Functions of nonclassical MHC and non-MHC-encoded class I molecules**. *Current opinion in immunology* 1999, **11**(1):100-108.

33. Li Q, Duan L, Estes JD, Ma ZM, Rourke T, Wang Y, Reilly C, Carlis J, Miller CJ, Haase AT: **Peak SIV replication in resting memory CD4+ T cells depletes gut lamina propria CD4+ T cells**. *Nature* 2005, **434**(7037):1148-1152.
34. Haase AT: **Perils at mucosal front lines for HIV and SIV and their hosts**. *Nat Rev Immunol* 2005, **5**(10):783-792.
35. Lauer FT, Denson JL, Beswick E, Burchiel SW: **Intracellular Cytokine Detection by Flow Cytometry in Surface Marker-Defined Human Peripheral Blood Mononuclear T Cells**. *Current protocols in toxicology* 2017, **73**:18.19.11-18.19.14.
36. Wikipedia: **Principal component analysis**. In: *Wikipedia, The Free Encyclopedia.* Wikipedia, The Free Encyclopedia.; 2018.
37. Wikipedia: **Eigendecomposition of a matrix**. In: *Wikipedia, The Free Encyclopedia.* Wikipedia, The Free Encyclopedia.; 2018.
38. R Development Core Team: **R: A language and environment for statistical computing.** In. Vienna, Austria: R Foundation for Statistical Computing; 2017.
39. **R: A Language and Environment for Statistical Computing** [http://www.R-project.org/]
40. Becker R, Chambers J, Wilks AJBNSL: **The New S Language Pacific Grove CA: Wadsworth & Brooks/Cole**. 1988.
41. Corpet FJNar: **Multiple sequence alignment with hierarchical clustering**. 1988, **16**(22):10881-10890.
42. Hartigan JAJJoc: **Statistical theory in clustering**. 1985, **2**(1):63-76.
43. Harrell FE: **Ordinal logistic regression**. In: *Regression modeling strategies.* Springer; 2015: 311-325.
44. Wikipedia: **Logistic regression**. In: *Wikipedia, The Free Encyclopedia.* Wikipedia, The Free Encyclopedia; 2018.
45. Tibshirani RJJotRSSSB: **Regression shrinkage and selection via the lasso**. 1996:267-288.
46. Ogutu JO, Schulz-Streeck T, Piepho H-P: **Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions**. In: *BMC proceedings: 2012*. BioMed Central: S10.
47. Usai MG, Goddard ME, Hayes BJJGr: **LASSO with cross-validation for genomic selection**. 2009, **91**(6):427-436.
48. Zou H, Hastie T: **Regularization and Variable Selection via the Elastic Net**. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 2005, **67**(2):301-320.
49. Friedman J, Hastie T, Tibshirani RJJoss: **Regularization paths for generalized linear models via coordinate descent**. 2010, **33**(1):1.
50. Trevor Hastie BE: **LARS: Least Angle Regression, Lasso and Forward Stagewise**. In.: R package version 1.2; 2013.
51. Sokal RR, Rohlf FJ: **The principles and practice of statistics in biological research**: WH Freeman and company San Francisco:; 1969.
52. Akaike H: **Factor analysis and AIC**. In: *Selected Papers of Hirotugu Akaike.* Springer; 1987: 371-386.
53. Gudivada V, Apon A, Ding J: **Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations**, vol. 10; 2017.
54. Abdi H, Williams LJ: **Principal component analysis**. 2010, **2**(4):433-459.

55. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR: **A simulation study of the number of events per variable in logistic regression analysis**. *Journal of Clinical Epidemiology* 1996, **49**(12):1373-1379.

56. Hanley JA, McNeil BJJR: **The meaning and use of the area under a receiver operating characteristic (ROC) curve**. 1982, **143**(1):29-36.

57. Jin H, Ling CX: **Using AUC and accuracy in evaluating learning algorithms**. *IEEE Transactions on Knowledge and Data Engineering* 2005, **17**(3):299-310.

58. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller MJBB: **pROC: an open-source package for R and S+ to analyze and compare ROC curves**. 2011, **12**(1):77.

59. Fukazawa Y, Lum R, Okoye AA, Park H, Matsuda K, Bae JY, Hagen SI, Shoemaker R, Deleage C, Lucero CJNm: **B cell follicle sanctuary permits persistent productive simian immunodeficiency virus infection in elite controllers**. 2015, **21**(2):132.