Antibody Response Induced by HIV Vaccines and T-cell Suppression Treatments in Rhesus Macaques

Group 3: Kan Luo, Shih-Ni Prim, Frederick Davey, Rizwana Rehman

2020-11-22

1 Introduction

The current report describes multivariate and longitudinal data analyses done on a dataset from an HIV study, in which researchers administered HIV vaccine injections and immuno-suppression treatments to 20 rhesus macaques and measured antibody characteristics. This project aimed to determine whether the vaccine's efficacy, reflected by the variable `Binding`, was increased by vaccine injections and immuno-suppression treatments.

1.1 About the Study

A dominant vaccine development strategy is to induce neutralizing antibodies by immunizing humans with the virus' glycoproteins. However, HIV vaccines that adopted this strategy mostly failed due to the fact that HIV is an RNA virus, which mutates rapidly to escape the inhibition of neutralizing antibodies. By the time the body generates neutralizing antibodies against the glycoproteins of some HIV strains, the RNA virus has already mutated. Thus, the existing neutralizing antibody fails to recognize, bind with, and neutralize the HIV virus. One possible solution is to increase the number of potential neutralizing antibodies that will cycle in the body by releasing a variety of antibodies after glycoprotein immunization. Regulatory T (Treg) cells prevent autoimmune diseases and suppress allergic reactions by inhibiting adaptive antibody immune response in the germinal center. Theoretically, this adaptive response lowers the effectiveness of vaccines. Thus the researchers used T-cell suppression treatments to investigate whether the treatments improve the efficacy of immunization.

During the study, 20 rhesus macaques were given glycoprotein immunization and supplemental antibody doses, as well as one of three treatments (two experimental regulatory T-cell suppression treatments and one control). Blood samples were collected two weeks after vaccine dosing, and antibodies were isolated from those samples. Limited by assay yield, the number of antibodies collected from each blood sample varied. In short, each observation contained information about the antibody isolated post the glycoprotein immunization.

A human antibody is formed heavy chains and light chains. A heavy chain has about 51 V-gene segments, 25 D-gene segments, and 6 J-gene segment. A light chain has about 71 V-gene segments and 9 J-gene segments[ref.5]. Any heavy chain V-D-J combination and light chain V-J combinations

can randomly happen in germinal centers. Theoretically, there can be $51 * 25 * 6 * 71 * 9 = 4.88835 \times 10^6$ combinations of gene segments. Considering the frequent mutations and other factors, each individual can have over **10 billion** different antibodies. Thus, we followed the convention of vaccine studies and viewed each antibody as independent for multivariate data analysis. For longitudinal data analysis, since general linear and linear mixed models allow measurements from the same observational unit to correlate over time, we used different correlation structures to find a best model. This implies that the antibodies from the same macaque can be correlated.

Section 7.1 contains a list of variables with brief descriptions. Note that each antibody contains two sets of heavy chain and light chain, all of which form a Y-shape immunoglobulin. Thus many of the variables start with H or L, indicating the chain from which the information comes. Some variables have missing numbers. We chose to let R functions ignore missing values to keep as much data as possible. As one of the authors for four publications that used the dataset, Kan Luo provided the dataset, which can be found here.

Now we turn to our research questions, after which the Methods section includes exploratory data analysis, multivariate data analysis, and longitudinal data analysis. The statistical results are included in the Methods section and summarized in the Results section. We discuss implications and limitations in the Discussions and some final thoughts in the Conclusions.

1.2 Research questions

The current project focused on understanding whether the number of vaccine injections (`Time_Point`) and the different Treg inhibitor treatments (`Drug`) caused changes in the antibody characteristics and if the changes were related to the immune responses against HIV virus. Our research questions are:

RQ1: Did time points and drugs have effects on the mutation frequency (`HMuFreq` and `LMuFreq`) and the amino acid count in the third complementarity determining region (`H_CDR3` and `L_CDR3`)?

RQ2: How did the binding strength of the antibodies (`Binding` or `logBinding`) develop in response to the number of vaccine dosages (`Time_Point`) and immuno-suppression treatments (`Drug`)?

2 Methods

This section first provides an overview and summaries of the dataset and then uses multivariate and longitudinal data analyses to answer the research questions.

2.1 Data Summaries

2.1.1 An Overview of Antibodies by Time Points, Drug Types, and Isotypes

A total of 2465 antibodies, from 20 rhesus macaques, were collected at four different time points (0, 1, 2, 3) and each macaque was given one of three drugs (1 and 2 are immuno-suppressing drugs and 3 is a mock control). Figure 1 shows the histograms of antibody counts, and Table 1 shows the antibody counts in different combinations of drugs and time points.

2.1.2 Outlier Detection

We included five response variables for the project: `H_CDR3`, `HMuFreq`, `L_CDR3`, `LMuFreq`, and `Binding`. As shown in Figure 2, one data point of `L_CDR3` seems an outlier. The summary statistics of standardized `L_CDR3` in Table 2 show that a maximum value of 30, which is quite unusual. Figure 3 shows the Mahalanobis ditances and Z scores of `L_CDR3`, and the same data point again appears to be an apparent outlier. The value for `L_CDR3`, 47, is quite unlikely. Since we were unable to reexamine the original data, we removed the data point.

2.1.3 Response Variables

Our five response variables were `H_CDR3`, `HMuFreq`, `L_CDR3`, `LMuFreq`, and `Binding`. Figure 4 shows `H_CDR3`'s approximately normal distributions with the center around 13 at different time points. Figure 5 shows the distributions of `H_CDR3` with respect to treatments at different time points, which are again approximately normal. With `L_CDR3`, Figure 6 and Figure 7 show approximately normal distribution, centered around 9, and with a longer right tail. The Q-Q plots in Figure 8 show that `H_CDR3` and `L_CDR3` are both approximately normal.

`HMuFreq` and `LMuFreq` were calculated by dividing `H_Substitution` by `H_VBase` for heavy chains and `L_Substitution` by `L_VBase` light chains. These two variables represent the degree to

which the antibodies mutate. A higher mutation rate usually indicates better virus neutralization. Figure 9, Figure 10, Figure 11, and Figure 12 show that `HMuFreq` and `LMuFreq` are both approximately normal, each with a long right tail. The Q-Q plots in Figure 13 confirm the approximate normality of `HMuFreq` and `LMuFreq`.

Next, a histogram of `Binding` with respect to treatment at different time points and a Q-Q plot are shown in Figure 14 and Figure 15, both of which reveal that `Binding` is not normally distributed. Since the dataset has a large sample size (n = 2464), we can use the Central Limit Theorem and assume normality. However, since many data points have the value 0 for `Binding`, linear models for longitudinal analyses might lead to negative values. To avoid this problem, we transformed `Binding` to $log(Binding + 1)$ and called it `logBinding`. The Q-Q plot of `logBinding` is shown in Figure 16. Lastly, Figure 17 shows that none of the response variables are highly correlated.

2.2 Multivariate Data Analysis

To answer **RQ1** (Did time points and drugs have effects on the mutation frequency (`HMuFreq` and `LMuFreq`) and the amino acid count in the third complementarity determining region (`H_CDR3` and `L_CDR3`)?), we tested whether predictors `Time_Point` and `Drug` had effects on four of our five response variables: `H_CDR3`, `HMuFreq`, `L_CDR3`, and `LMuFreq`. We excluded `Binding` from this section, because it has unequal variances across time points, which violated the equal variance assumption of MANOVA. We used the `manova` function in base R for the MANOVA test and the `emmeans` package for pairwise comparisons.[7].

2.2.1 MANOVA

Since we wanted to compare more than two populations, we used MANOVA to test for effects. We checked that the normality assumption was met due to large sample size (n = 2464). We assumed that antibodies were independent of each other. However, we had some concerns about the equal variance-covariance matrices assumption, which will be addressed in the Discussions. We performed a MANOVA test with the formula **(H_CDR3, HMuFreq, L_CDR3, LMuFreq)^T ~ Time_Point + Drug** and the null hypothesis that the means of the different populations (of `Time_Point` and `Drug`) were equal.

The output in Section 7.2 shows that both main effects of `Time_Point` and `Drug` have very small p-values. Thus we rejected the null hypothesis and concluded that both main effects were significant to the antibodies' four traits, `H_CDR3`, `HMuFreq`, `L_CDR3`, and `LMuFreq`. To further understand the main effects, we proceeded to do pairwise comparisons.

2.2.2 Pairwise Comparison

Table 3 summarizes the pairwise comparison results (see Sections 7.3, 7.4). `Time_Point` have significant pairs for `H_CDR3`, `HMuFreq`, and `LMuFreq`, and time points 0 and 1 appear to have significant differences from time points 2 and 3. `Drug` has significant pairs in `H_CDR3` and `HMuFreq`. For `H_CDR3`, drug 1 has a higher mean than drug 2. For `HMuFreq`, drug 2 has a higher mean than drug 1, which has a higher mean than drug 3 (control group). [Kan, please add/revise the interpretations here.]

2.3 Longitudinal Analysis

We used longitudinal data analyses to answer our **RQ2** (How did the binding strength of the antibodies (`Binding` or `logBinding`) develop in response to the number of vaccine dosages (`Time_Point`) and immuno-suppression treatments (`Drug`)?), including general linear models and linear mixed models. We used the `gls` and `lme` functions from the `nlme` package[8].

2.3.1 One Covariate: Time Point

As shown in Figure 18 and Figure 19, the mean trend by monkey is not linear across all time points. Figure 20 and Figure 21 show that the different time points have different variances. Thus we used piecewise linear models and set variances as unequal over time. We first considered a model with time point as the only covariate:

$$Y_{ij} = \beta_0 + \beta_1 Time_{ij} + e_{ij}$$

We then turned the model into a piecewise linear model and designated different intercepts and

slopes for the line segments. The model includes three indicator variables: $S1, S2, S3$, where

$$S1 = \begin{cases} 1 & \text{if } 0 \leq \text{Timepoint} < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$S2 = \begin{cases} 1 & \text{if } 1 \leq \text{Timepoint} < 2 \\ 0 & \text{otherwise} \end{cases}$$

$$S3 = \begin{cases} 1 & \text{if Timepoint} \geq 2 \\ 0 & \text{otherwise} \end{cases}$$

The new model is

$$Y_{ij} = S1(\beta_0 + \beta_1 Time_{ij}) + S2(\beta_2 + \beta_3 Time_{ij}) + S3(\beta_4 + \beta_5 Time_{ij}) + e_{ij}$$

The trend should be continuous at time points 1 and 2. Our first complete model (`fit.gls`) is

$$Y_{ij} = \beta_0(S1 + 2S2 - S2Time_{ij}) + \beta_1(S1Time_{ij} + 2S2 - S2Time_{ij}) +$$

$$\beta_4(-S2 + S2Time_{ij} + S3) + \beta_5(-2S2 + 2S2Time_{ij} + S3Time_{ij}) + e_{ij}$$

where

$$\mathbf{e}_i \sim N(0, \sigma^2 I)$$

As mentioned earlier, we transformed `Binding` into `logBinding` with the formula

$$logBinding = log(Binding + 1).$$

We first used them as response variables in the above formula in two models (`fit.gls1` and `fit.gls2`, respectively) to find the better response variable. As seen in Table 4, the model using `logBinding` as the response variable has much lower AIC and BIC values. We decided to use `logBinding` as the response variable and build on this model (`gls.fit2`). For the rest of the report, Y_{ij} denotes `logBinding`. Note that some plots below still use `Binding` in the y-axis, for which we

had to plug the fitted values of `logBinding` into the exponential function to find `Binding`.

The model can also be written as

$$Y_{ij} = S1(\beta_0)+S1Time_{ij}(\beta_1)+S2(2\beta_0+2\beta_1-\beta_4-2\beta_5)+S2Time_{ij}(-\beta_0-\beta_1+\beta_4+2\beta_5)$$

$$+S3(\beta_4) + S3Time_{ij}(\beta_5) + e_{ij}$$

We ran the model and plugged in coefficients to find the intercepts and slopes for all three segments of the mean trend:

- S1: Binding = $exp(-0.0921720 + 0.1090234 * time) - 1$
- S2: Binding = $exp((2 * -0.0921720 + 2 * 0.1090234 - 0.3573358 - 2 * -0.1070750) + (0.0921720 - 0.1090234 + 0.3573358 + 2 * -0.1070750) * time) - 1 = exp(-0.109483 + 0.1263344 * time) - 1$
- S3: Binding = $exp(0.3573358 - 0.1070750 * time) - 1$

Figure 22 shows the two segments S1 and S2 have very similar slopes. So we could refit the model with only two line segments between time points 0 and 2 and between time points 2 and 3. We called them S4 and S5. The next model is therefore

$$Y_{ij} = S4(\beta_0 + \beta_1 Time_{ij}) + S5(\beta_2 + \beta_3 Time_{ij}) + e_{ij}$$

where

$$S4 = \begin{cases} 1 & \text{if Timepoint} < 2 \\ 0 & \text{otherwise} \end{cases}$$

$$S5 = \begin{cases} 1 & \text{if Timepoint} \geq 2 \\ 0 & \text{otherwise} \end{cases}$$

Again, the trend should be continuous at Time_Point = 2. Our second complete model (`fit.gls3`) is then

$$Y_{ij} = \beta_1(-2S4 + S4Time_{ij}) + \beta_2(S4 + S5) + \beta_3(2S4 + S5Time_{ij}) + e_{ij}$$

where

$$\mathbf{e}_i \sim N(0, \sigma^2 I)$$

The model could also be written as

$$Y_{ij} = S4(-2\beta_1 + \beta_2 + 2\beta_3) + S4Time_{ij}(\beta_1) +$$

$$S5(\beta_2) + S5Time_{ij}(\beta_3) + e_{ij}$$

After the model was constructed, we used the coefficients to find the mean trends for S4 and S5:

- S4: Binding = $exp((-2 * 0.24395790 + 0.26937508 + 2 * 0.00194581) + 0.2448519 * time) - 1 = exp(-0.2146491 + 0.2448519 * time) - 1$
- S5: Binding = $exp(0.26937508 + 0.00194581 * time) - 1$

As shown in Figure 23, one linear line connects Time_Point 0 and 2 and another connects Time_Point 2 and 3. The two lines are continuous at Time_Point 2. A comparison of AIC And BIC of these two models, shown in Table 5, indicates that the first model (`fit.gls2`) is a better model.

2.3.2 Adding Random Effects

Next we checked whether adding random effects could improve our model (`fit.gls2`). We assumed that random effects existed in the intercept and slope. Our linear mixed model is then:

$$Y_{ij} = \beta_0(S1 + 2S2 - S2Time_{ij}) + \beta_1(S1Time_{ij} + 2S2 - S2Time_{ij}) +$$

$$\beta_4(-S2 + S2Time_{ij} + S3) + \beta_5(-2S2 + 2S2Time_{ij} + S3Time_{ij}) + b_{0i} + b_{1i}Time_{ij} + e_{ij}$$

where

$$\mathbf{b}_i \sim N\left(0, \mathbf{D} = \begin{bmatrix} D_{11} & D_{12} \\ & D_{22} \end{bmatrix}\right)$$

and

$$\mathbf{e}_i \sim N(0, \sigma^2 I)$$

We fit two models with random effects: `fit.a1` assumes random intercept and slope for time

point, compound symmetric correlation structure, and unequal variances over time; and `fit.a2` assumes random intercept and slope for time point, AR1 correlation structure, and unequal variances over time. As shown in Table 6, the model `fit.a2` has the lowest AIC and BIC, so it is the best model. We checked residuals for three models: `fit.gls2`, `fit.a1`, `fit.a2`, as shown in Figure 24. All three Q-Q plots show approximate normality.

2.3.3 Inference about β

Note the meanform for our three models are:

$$E(Y_{ij}) = S1(\beta_0)+S1Time_{ij}(\beta_1)+S2(2\beta_0+2\beta_1-\beta_4-2\beta_5)+S2Time_{ij}(-\beta_0-\beta_1+\beta_4+2\beta_5)$$

$$+S3(\beta_4) + S3Time_{ij}(\beta_5)$$

We would like to know if the slopes between time points 0 and 1, 1 and 2, and 2 and 3 equal zero, which means

$$H_0 : \beta_1 = 0, -\beta_0 - \beta_1 + \beta_4 + 2\beta_5 = 0, \beta_5 = 0$$

Thus, we performed three tests:

$$\mathbf{L_1}\beta = 0$$

where $\mathbf{L_1} = (0, 1, 0, 0)$ and $= (\beta_0, \beta_1, \beta_4, \beta_5)^T$

$$\mathbf{L_2}\beta = 0$$

where $\mathbf{L_2} = (-1, -1, 1, 2)$ and $= (\beta_0, \beta_1, \beta_4, \beta_5)^T$

$$\mathbf{L_3}\beta = 0$$

where $\mathbf{L_3} = (0, 0, 0, 1)$ and $= (\beta_0, \beta_1, \beta_4, \beta_5)^T$

As shown in Table 7, all three slopes have very small p-values, which means the rates of change for `logBinding` in all three segments are significant. As shown in Figure 22, binding rates increase in S1 and S2 and decrease in S3. The use of `logBinding` as Y_{ij} made interpretation more difficult,

but, with the very small p-values of β about `logBinding` and the direction of slopes for `Binding`, we could tentatively conclude that time point 2, when the monkeys had received two vaccines, had the highest binding rates, while the third vaccine injection at time point 3 failed to increase binding rates.

2.3.4 Two Covariates: Time Point and Drug

Next we added `Drug` as a covariate to the model `gls.a2` to see if it had effects on `logBinding`. We used two indicator variables: `D2` and `D3`, where

$$D2 = \begin{cases} 1 & \text{if Drug = 2} \\ 0 & \text{otherwise} \end{cases}$$

$$D3 = \begin{cases} 1 & \text{if Drug = 3} \\ 0 & \text{otherwise} \end{cases}$$

Building on the model `gls.a2` and assuming that the random effects were the same for each drug, our model (`fit.a3`) with the extra covariate `Drug` is:

$$Y_{ij} = \beta_0(S1 + 2S2 - S2Time_{ij}) + \beta_1(S1Time_{ij} + 2S2 - S2Time_{ij}) +$$

$$\beta_2(-S2 + S2Time_{ij} + S3) + \beta_3(-2S2 + 2S2Time_{ij} + S3Time_{ij}) +$$

$$\beta_4 D2(S1 + 2S2 - S2Time_{ij}) + \beta_5 D2(S1Time_{ij} + 2S2 - S2Time_{ij}) +$$

$$\beta_6 D2(-S2 + S2Time_{ij} + S3) + \beta_7 D2(-2S2 + 2S2Time_{ij} + S3Time_{ij}) +$$

$$\beta_8 D3(S1 + 2S2 - S2Time_{ij}) + \beta_9 D3(S1Time_{ij} + 2S2 - S2Time_{ij}) +$$

$$\beta_{10} D3(-S2 + S2Time_{ij} + S3) + \beta_{11} D3(-2S2 + 2S2Time_{ij} + S3Time_{ij}) +$$

$$b_{0i} + b_{1i}Time_{ij} + e_{ij}$$

where

$$\mathbf{b}_i \sim N\left(0, \mathbf{D} = \begin{bmatrix} D_{11} & D_{12} \\ & D_{22} \end{bmatrix}\right)$$

and

$$\mathbf{e}_i \sim N(0, \sigma^2 I)$$

Table 8 shows that the model with drug as the second covariate has the lowest AIC but not the lowest BIC. Considering the added complexity of the current model and the slight improvement with respect to AIC, we selected `fit.a2` as our best model. Next we made inference about β to find whether the drugs had different effects. To see whether Drug 1 and Drug 2 had different effects, we performed a hypothesis test on $H_0 : \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$ by testing

$$\mathbf{L_4}\beta = 0$$

where

$$\mathbf{L_4} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9, \beta_{10}, \beta_{11})^T$

To see whether Drug 1 and Drug 3 had different effects, we performed a hypothesis test on $H_0 : \beta_8 = \beta_9 = \beta_1 0 = \beta_1 1 = 0$ by testing

$$\mathbf{L_5}\beta = 0$$

where

$$\mathbf{L_5} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9, \beta_{10}, \beta_{11})^T$

To see whether Drug 2 and Drug 3 had different effects, we performed a hypothesis test on $H_0 : \beta_4 = \beta_8, \beta_5 = \beta_9, \beta_6 = \beta_{10}, \beta_7 = \beta_{11}$ by testing

$$\mathbf{L_6}\beta = 0$$

where

$$\mathbf{L_6} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{bmatrix}$$

and $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9, \beta_{10}, \beta_{11})^T$

As shown in Table 9 and Table 11, the p-values are smaller than 0.05. We rejected the null hypothesis and concluded that drugs 1 and 2 and drugs 2 and 3 had different effects on logbinding rates. Table 10, on the other hand, shows a p-value slightly greater than 0.05, suggesting that drugs 1 and 3 did not have different effects on logbinding rates. Drug 3 was in fact a mock control, which should not have had any effects. We concluded that drug 2 acted differently from drug 1 and the control and could be used in future research for further study.

3 Results

For multivariate analyses, we performed a MANOVA test on the main effects of `Time_Point` and `Drug` on the response variable vector (`H_CDR3`, `HMuFreq`, `L_CDR3`, `LMuFreq`)$^{\mathrm{T}}$. We found that both main effects had very small p-values. We also performed pairwise comparison to see where the effects were, as shown in Table 3. [Kan, do you have anything to add?]

For longitudinal analyses, we found that the transformed variable `logBinding` was a better option and the linear mixed model `fit.a2` had the lowest BIC:

$$Y_{ij} = \beta_0(S1 + 2S2 - S2Time_{ij}) + \beta_1(S1Time_{ij} + 2S2 - S2Time_{ij}) +$$

$$\beta_4(-S2 + S2Time_{ij} + S3) + \beta_5(-2S2 + 2S2Time_{ij} + S3Time_{ij}) + b_{0i} + b_{1i}Time_{ij} + e_{ij}$$

where

$$\mathbf{b}_i \sim N\left(0, \mathbf{D} = \begin{bmatrix} D_{11} & D_{12} \\ & D_{22} \end{bmatrix}\right)$$

and

$$\mathbf{e}_i \sim N(0, \sigma^2 I)$$

We performed F-tests to make inference about β for the three line segments. We rejected the hypothesis that the slopes of all three line segments were zero. We then added `Drug` as another covariate to the above model and made inference about β. The comparison between three drug groups was done with three F-tests (between Drug groups 1 and 2, 2 and 3, and 1 and 3). The p-values for the tests between drugs 1 and 2 and between drugs 2 and 3 were smaller than 0.05, while the p-value for the test between drugs 1 and 3 were slightly greater than 0.05. Thus We rejected the null hypotheses that effects of drugs 1 and 2 as well as 2 and 3 were equal but failed to reject the null hypotheses that the effects of drugs 1 and 3 were equal.

4 Discussion

4.1 Implications

Our findings show that the number of vaccine injections did contribute to higher binding rates, although we did not determine whether the increase could be translated into immunity against HIV. Drug 2 appeared to have different effects from drug 1 and control, but our analyses did not determine whether using immuno-suppressing drugs could enhance the efficacy of HIV vaccines.

4.2 Limitations

4.2.1 Independence of antibodies

In our multivariate analyses, we followed the common method of treating antibodies (rows in our data) as independent of each other. In our longitudinal analyses, the models allow measurements from the same observational unit (macaque) to correlate over time. It would be beneficial to formally determine whether the antibodies were correlated or independent, but it would require

more biological knowledge and the investigation would go beyond the scope of the current report. This remains an interesting topic that could be explored.

4.2.2 Equal variance-covariance matrices assumption for MANOVA

For multivariate data analyses, the use of MANOVA was restricted by the assumptions of equal variance-covariance matrices among different populations. We ran the Fligner-Killeen Test of Homogeneity of Variances on all the four response variables `H_CDR3`, `HMuFreq`, `L_CDR3`, and `LMuFreq`. As shown in the output in Section 7.5, most of the p-values are very small, meaning the null hypothesis of equal variance is rejected. In most cases, `H_CDR3` and `HMuFreq` did not meet the equal variance-covariance matrices assumption.

Comparisons of the variance-covariance matrix of the response variables in different groups using ratios (one matrix divided by another matrix) also reveals that the variance-covariance matrices might not be equal. Some matrices seem quite different. For example, as shown in Table 12, the ratio between variance-covariance matrices of Drug 2 and Drug 3 has values as large as 26. Furthermore, the sample sizes of each populations, as shown in Table 1, are unequal. Thus, the results of the MANOVA test should be viewed with caution.

4.2.3 Longitudinal Models

For longitudinal data analyses, we did not try out more combinations of models. For example, we only tried two correlation structures (compound symmetry and AR1); other structures might have achieved better results. When we added drug as another covariate, we did not go back to test which correlation structure performed better and whether the piecewise model should include two or three line segments. Further, we did not assume different random effects for different line segments. These additional steps could lead to better models.

5 Conclusions

In our project, we performed multivariate data analyses and longitudinal data analyses to understand whether time points and drugs had effects on characteristics of antibodies and enhanced the efficacy of HIV vaccines. Our statistical analyses provided answers to our two research questions.

We performed a MANOVA test to answer our first research question, "Did time points and drugs have effects on the mutation frequency (`HMuFreq` and `LMuFreq`) and the amino acid count in the third complementarity determining region (`H_CDR3` and `L_CDR3`)?" and found significant main effects for time points and drugs. [Kan: anything to add here?]

To answer our second research question, "How did the binding strength of the antibodies (`Binding` or `logBinding`) develop in response to the number of vaccine dosages (`Time_Point`) and immuno-suppression treatments (`Drug`)?", we constructed longitudinal models. We found the model with the transformed response variable `logBinding`, three line segments, random effects of intercept and slope of time point, AR1 correlation structure, and unequal variances over time performed best. F-tests for inference about β revealed that time point 2 had the highest `logBinding` value, suggesting that two vaccine injections induced the most antibody response. We also found that adding drug as a covariate did not greatly improve the model and drug 2 appeared to have different effects from drug 1 and the mock control.

To date, no HIV vaccines have been found effective in creating immunity against the HIV virus. Our analyses showed that the vaccine injections did increase binding rates, but perhaps the change was not enough. The second type of Treg inhibitor treatments might be worth further examination. We hope that, even if the result did not reveal a solution, more analyses could find promising directions for future research.

6 References

1. Luo K, Liao HX, Zhang R, et al. Tissue memory B cell repertoire analysis after ALVAC/AIDSVAX B/E gp120 immunization of rhesus macaques. *JCI Insight*. 2016;1(20):e88522. Published 2016 Dec 8. doi:10.1172/jci.insight.88522
2. Bradley, T., Kuraoka, M., Yeh, C.-H., Tian, M., Chen, H., Cain, D. W., . . . Haynes, B. F. (2020). Immune checkpoint modulation enhances HIV-1 antibody induction. *Nature Communications*, 11(1), 948. doi:10.1038/s41467-020-14670-w
3. Easterhoff, D., Pollara, J., Luo, K., Tolbert, W. D., Young, B., Mielke, D., . . . Ferrari, G. (2020). Boosting with AIDSVAX B/E Enhances Env Constant Region 1 and 2 Antibody-Dependent Cellular Cytotoxicity Breadth and Potency. *Journal of Virology*, 94(4), e01120-

01119. doi:10.1128/jvi.01120-19

4. Wiehe, K., Easterhoff, D., Luo, K., Nicely, N. I., Bradley, T., Jaeger, F. H., Dennison, S. M., Zhang, R., Lloyd, K. E., Stolarchuk, C., Parks, R., Sutherland, L. L., Scearce, R. M., Morris, L., Kaewkungwal, J., Nitayaphan, S., Pitisuttithum, P., Rerks-Ngarm, S., Sinangil, F., Phogat, S., . Haynes, B. F. (2014). Antibody light-chain-restricted recognition of the site of immune pressure in the RV144 HIV-1 vaccine trial is phylogenetically conserved. *Immunity*, 41(6), 909-918. https://doi.org/10.1016/j.immuni.2014.11.014
5. Lefranc MP, Giudicelli V, Ginestoux C, Bodmer J, Muller W, Bontrop R, Lemaitre M, Malik A, Barbie V, Chaume D. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res*. 1999;27:209-212. doi: 10.1093/nar/27.1.209.
6. Jenny M Woof , Dennis R Burton,Human antibody-Fc receptor interactions illuminated by crystal structures.*Nat Rev Immunol*. 2004 Feb;4(2):89-99. doi: 10.1038/nri1266.
7. Russell Lenth (2020). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.5.0. https://CRAN.R-project.org/package=emmeans
8. Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team (2020). *nlme: Linear and Nonlinear Mixed Effects Models*. R package

7 Appendix

7.1 List of variables

- Monkey_id: The identity of monkey
- Treatment (Drug): The 7 treatments in the dataset are coded into 3 drugs to simplify the data. Drug 1 and 2 are two different kinds of Treg inhibitor treatments, and drug 3 is the mock control.
- Time_Point: 0 represents before immunization; 1 represents 2 weeks post 1st immunization; 2 represents 2 weeks post 2nd immunization; and 3 represents 2 weeks post 3rd immunization, respectively.
- Isotype: There are 5 kinds of immunoglobulin isotypes: IgG, IgA, IgM, IgE, IgD. The two most important kinds are IgG and IgM. IgM occurs in the acute stage of infection and plays a role in the primary response. The secondary response IgG appears later in serum with higher binding affinity and neutralizing potentials against toxins and virus. IgA is mostly found in mucosal tissues such as Nasal mucosa. Non-dominant IgD and IgE are typically lower than 1% in blood.
- H_ID and L_ID: heavy chain and light chain IDs for the particular observation
- H_VBase: the number of nucleotide of the heavy chain variable region
- H_Substitutions: the number of relative nucleotide mutations in heavy chains
- HMuFreq: calculated by H_Substitutions / H_VBase
- H_CDR3: the number of amino acid of a heavy chain's third complementarity determining region
- L_VBase: the number of nucleotide of a light chain's variable region
- L_Substitutions: the number of relative nucleotide mutations in the light chain
- LMuFreq: calculated by L_Substitutions / L_VBase
- L_CDR3: the number of amino acid of a light chain's third complementarity determining region.`H_CDR3` and `L_CDR3` indicates the length of the third complementarity-determining region on the variable heavy chain and light chain. The longer they are, the more potential there is to produce diverse antibodies.
- Binding: affinity of antibodies against a selected HIV glycoprotein. `Binding` indicates the

rate of neutralizing, meaning how much the antibodies bind with the virus and thus make the virus ineffective. Larger values indicate stronger binding.

7.2 Output for the MANOVA test

```
##             Df   Wilks approx F num Df den Df    Pr(>F)
## tp           3 0.95348   9.8224     12 6487.7 < 2.2e-16 ***
## drug         2 0.97728   7.0843      8 4904.0 2.375e-09 ***
## Residuals 2455
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7.3 Pairwise comparison by time point

```
## [1] "H_CDR3  pairwise CI's"
##  contrast estimate    SE   df lower.CL upper.CL
##  0 - 1       0.489 0.231 2457   -0.223    1.202
##  0 - 2       0.194 0.237 2457   -0.535    0.923
##  0 - 3      -0.608 0.271 2457   -1.442    0.226
##  1 - 2      -0.295 0.159 2457   -0.784    0.194
##  1 - 3      -1.097 0.206 2457   -1.732   -0.462
##  2 - 3      -0.802 0.212 2457   -1.455   -0.148
##
## Results are averaged over the levels of: rep.meas
## Note: contrasts are still on the [.: scale
## Confidence level used: 0.997916666666667
## [1] "HMuFreq  pairwise CI's"
##  contrast estimate      SE   df  lower.CL upper.CL
##  0 - 1     0.00870 0.00288 2457 -0.000188   0.0176
##  0 - 2     0.01872 0.00295 2457  0.009629   0.0278
##  0 - 3     0.02356 0.00338 2457  0.013153   0.0340
##  1 - 2     0.01002 0.00198 2457  0.003926   0.0161
##  1 - 3     0.01486 0.00257 2457  0.006938   0.0228
##  2 - 3     0.00484 0.00265 2457 -0.003316   0.0130
##
## Results are averaged over the levels of: rep.meas
## Note: contrasts are still on the [.: scale
## Confidence level used: 0.997916666666667
## [1] "L_CDR3  pairwise CI's"
##  contrast estimate     SE   df lower.CL upper.CL
##  0 - 1     -0.0169 0.0652 2457   -0.218    0.184
##  0 - 2     -0.0287 0.0667 2457   -0.234    0.177
```

```
##  0 - 3     -0.0034 0.0764 2457   -0.239    0.232
##  1 - 2     -0.0117 0.0447 2457   -0.150    0.126
##  1 - 3      0.0135 0.0581 2457   -0.166    0.193
##  2 - 3      0.0253 0.0598 2457   -0.159    0.210
##
## Results are averaged over the levels of: rep.meas
## Note: contrasts are still on the [.: scale
## Confidence level used: 0.997916666666667
## [1] "LMuFreq  pairwise CI's"
##  contrast estimate      SE   df  lower.CL upper.CL
##  0 - 1     0.00853 0.00369 2457 -0.002847   0.0199
##  0 - 2     0.01465 0.00378 2457  0.003004   0.0263
##  0 - 3     0.01889 0.00432 2457  0.005567   0.0322
##  1 - 2     0.00611 0.00253 2457 -0.001694   0.0139
##  1 - 3     0.01036 0.00329 2457  0.000213   0.0205
##  2 - 3     0.00424 0.00339 2457 -0.006196   0.0147
##
## Results are averaged over the levels of: rep.meas
## Note: contrasts are still on the [.: scale
## Confidence level used: 0.997916666666667
```

7.4 Pairwise comparison by drug

```
## [1] "H_CDR3  pairwise CI's"
##  contrast estimate    SE   df lower.CL upper.CL
##  1 - 2      0.4907 0.160 2458   0.0316    0.950
##  1 - 3      0.4438 0.175 2458  -0.0589    0.946
##  2 - 3     -0.0469 0.171 2458  -0.5372    0.444
##
## Results are averaged over the levels of: rep.meas
## Note: contrasts are still on the [.: scale
## Confidence level used: 0.995833333333333
## [1] "HMuFreq  pairwise CI's"
##  contrast estimate      SE   df lower.CL  upper.CL
##  1 - 2      -0.006 0.00199 2458 -0.01172 -0.000285
##  1 - 3       0.011 0.00218 2458  0.00476  0.017279
##  2 - 3       0.017 0.00213 2458  0.01091  0.023127
##
## Results are averaged over the levels of: rep.meas
## Note: contrasts are still on the [.: scale
## Confidence level used: 0.995833333333333
## [1] "L_CDR3  pairwise CI's"
##  contrast estimate     SE   df lower.CL upper.CL
##  1 - 2      0.0451 0.0450 2458  -0.0839    0.174
##  1 - 3      0.1018 0.0492 2458  -0.0394    0.243
##  2 - 3      0.0567 0.0480 2458  -0.0810    0.194
##
## Results are averaged over the levels of: rep.meas
## Note: contrasts are still on the [.: scale
## Confidence level used: 0.995833333333333
## [1] "LMuFreq  pairwise CI's"
```

```
##  contrast  estimate      SE   df lower.CL upper.CL
##  1 - 2    -0.000176 0.00256 2458 -0.00752  0.00717
##  1 - 3     0.002320 0.00280 2458 -0.00572  0.01036
##  2 - 3     0.002496 0.00273 2458 -0.00535  0.01034
##
## Results are averaged over the levels of: rep.meas
## Note: contrasts are still on the [.: scale
## Confidence level used: 0.995833333333333
```

7.5 Fligner-Killeen Test of Homogeneity of Variances

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  H_CDR3 by Time_Point
## Fligner-Killeen:med chi-squared = 18.036, df = 3, p-value = 0.0004323

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  HMuFreq by Time_Point
## Fligner-Killeen:med chi-squared = 50.662, df = 3, p-value = 5.775e-11

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  L_CDR3 by Time_Point
## Fligner-Killeen:med chi-squared = 0.56825, df = 3, p-value = 0.9037

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  LMuFreq by Time_Point
## Fligner-Killeen:med chi-squared = 16.585, df = 3, p-value = 0.0008601

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  H_CDR3 by Drug
## Fligner-Killeen:med chi-squared = 13.463, df = 2, p-value = 0.001193

##
```

```
##  Fligner-Killeen test of homogeneity of variances
##
## data:  HMuFreq by Drug
## Fligner-Killeen:med chi-squared = 48.86, df = 2, p-value = 2.456e-11

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  L_CDR3 by Drug
## Fligner-Killeen:med chi-squared = 2.4113, df = 2, p-value = 0.2995

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  LMuFreq by Drug
## Fligner-Killeen:med chi-squared = 0.17372, df = 2, p-value = 0.9168
```

List of Figures

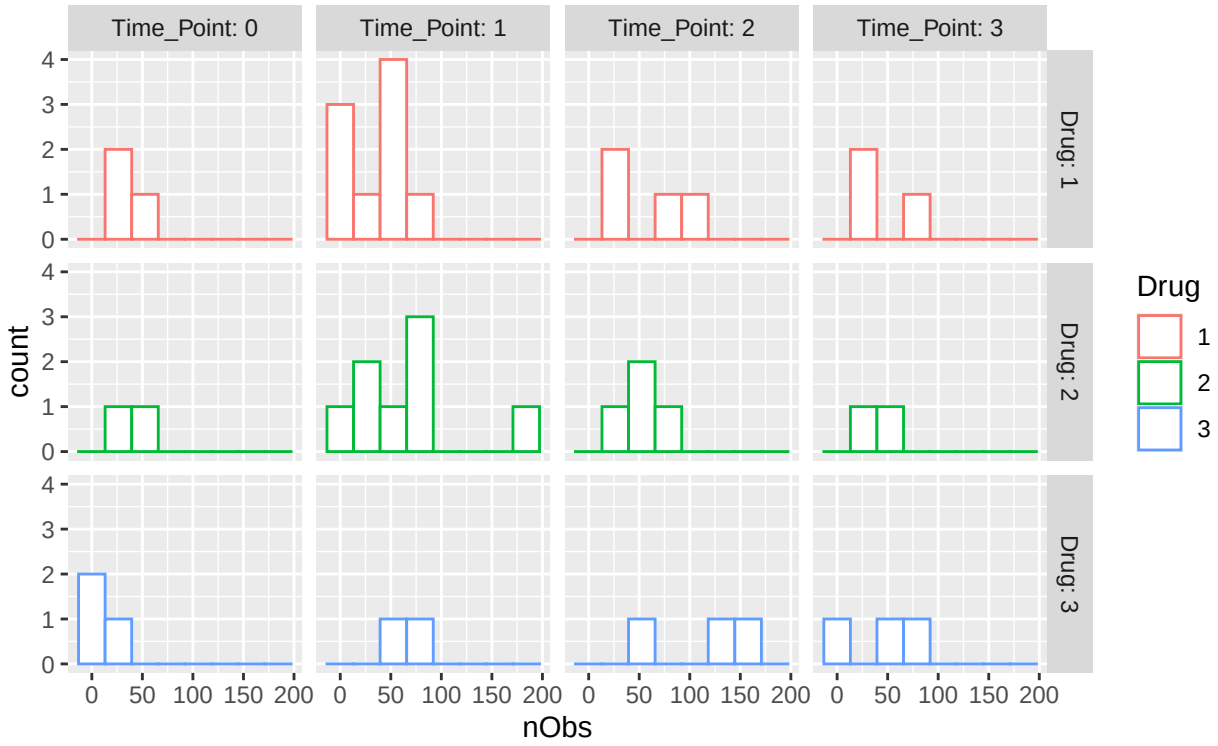

Figure 1: Histograms of Antibodies

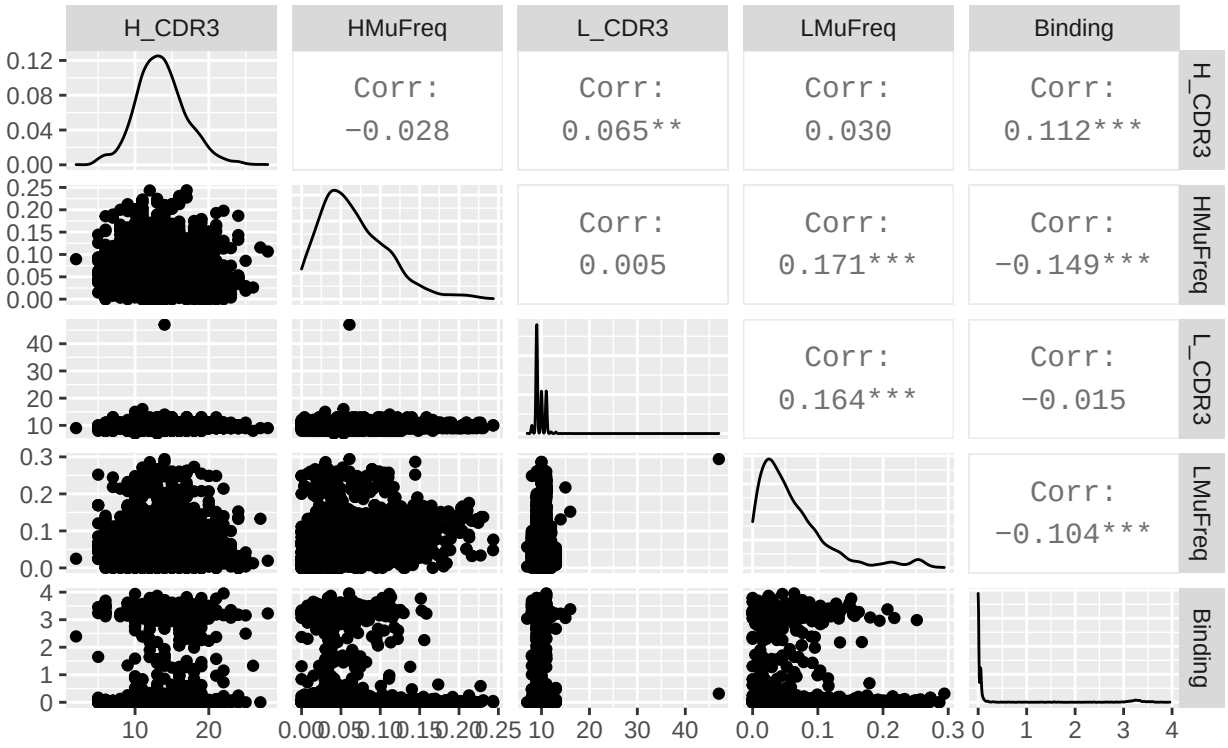

Figure 2: Scatterplot Matrix of Response Variables

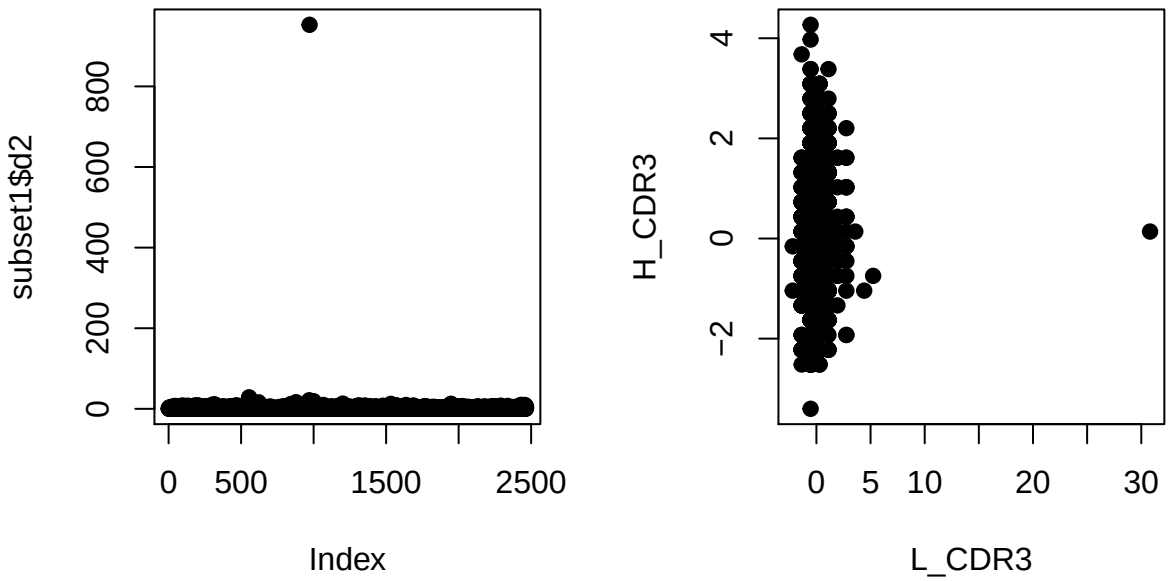

Figure 3: Mahalanobis Distances and Z Scores

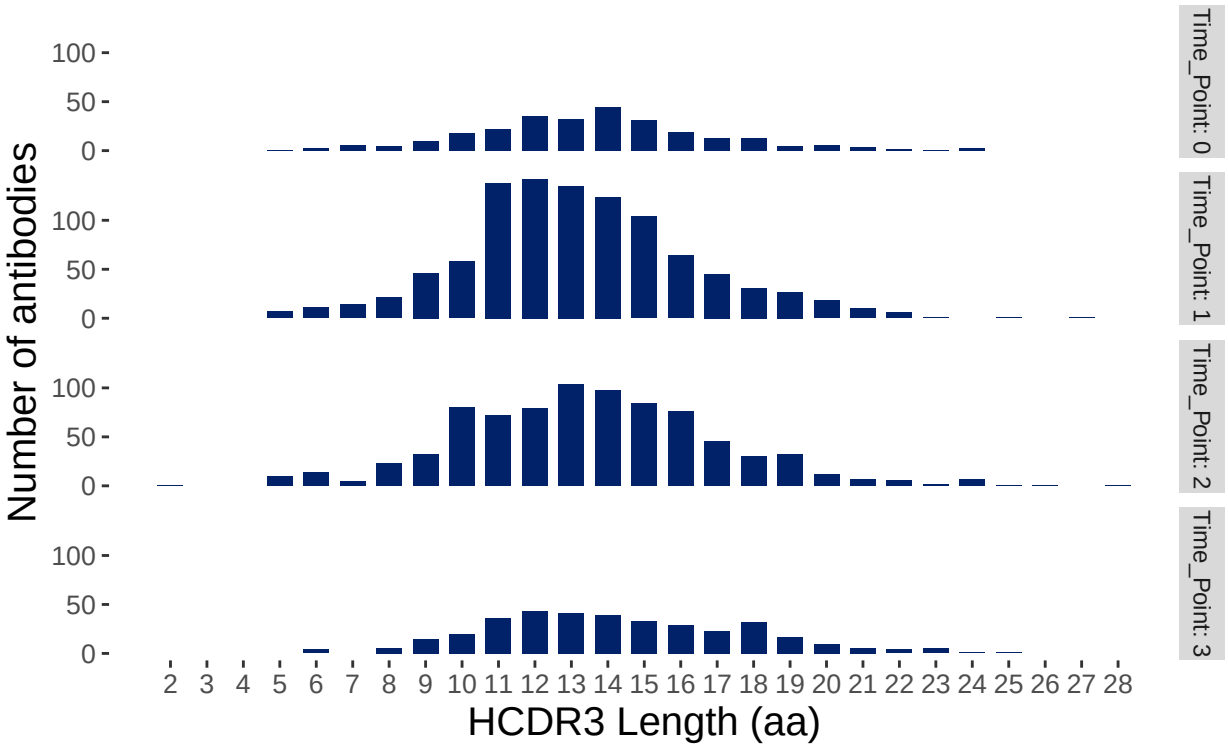

Figure 4: Histogram H_CDR3 by Time Point

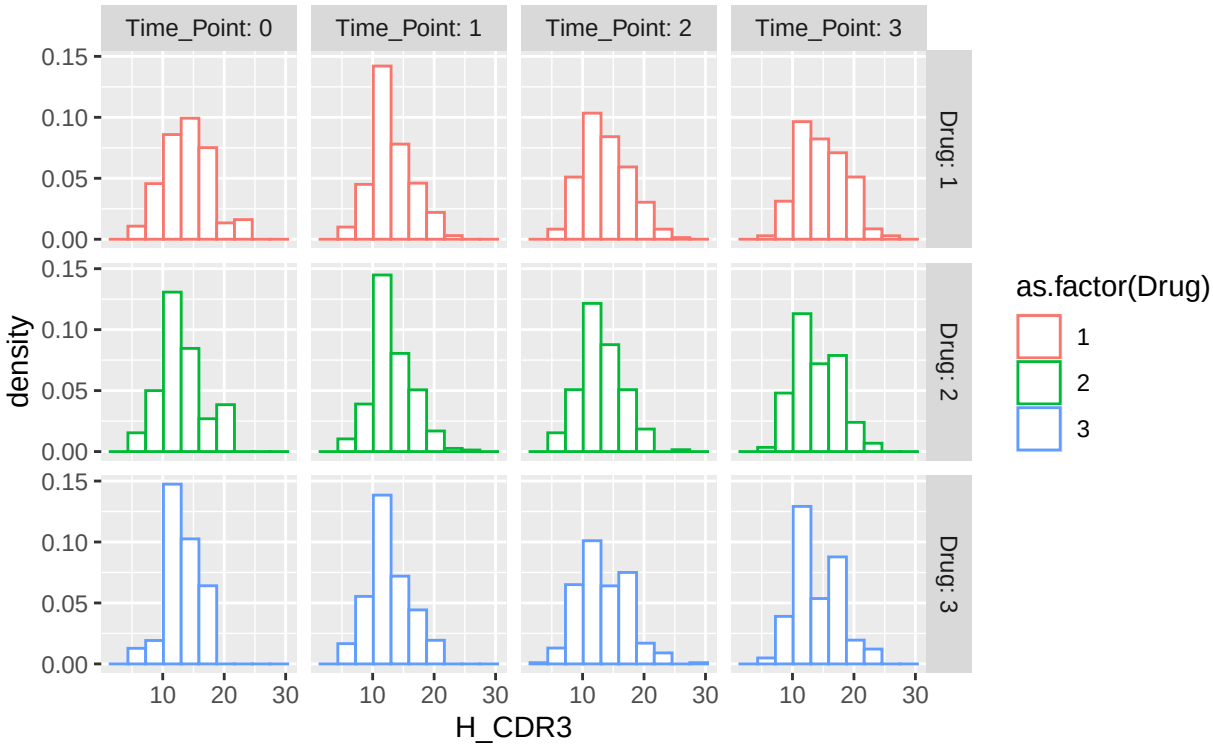

Figure 5: Histograms of H_CDR3 by Drug and Time Point

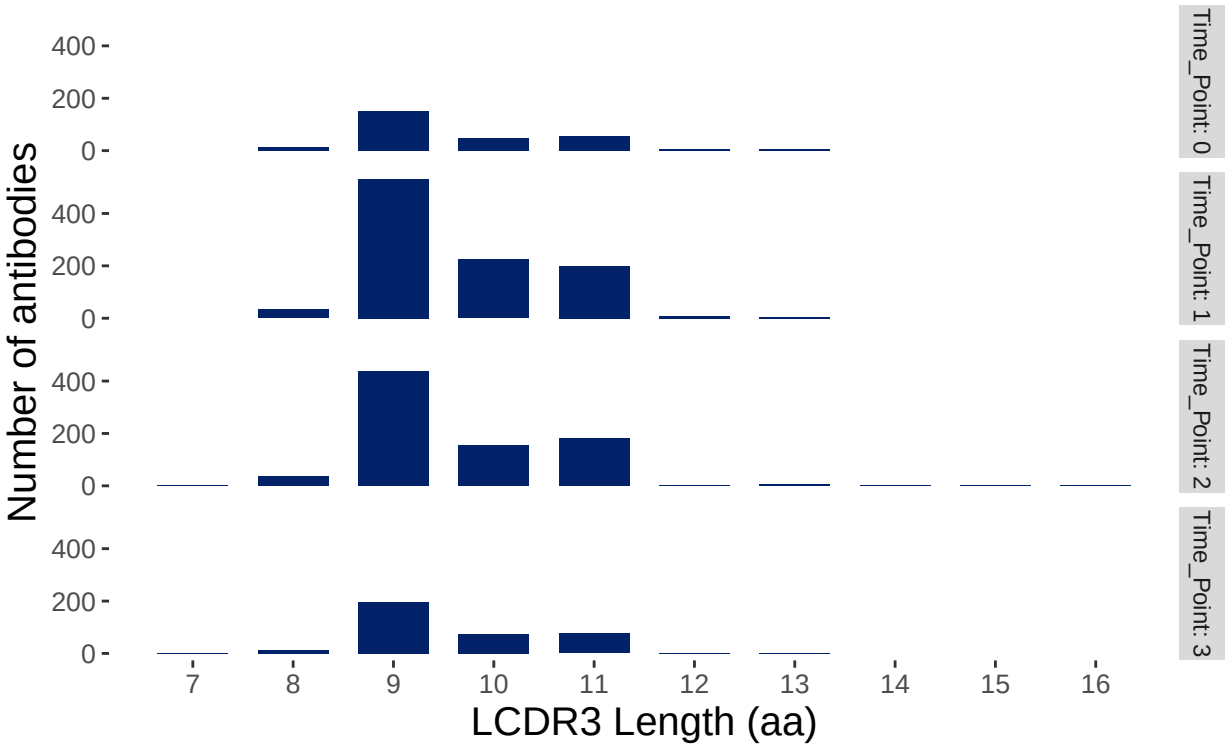

Figure 6: Histogram L_CDR3 by Time Point

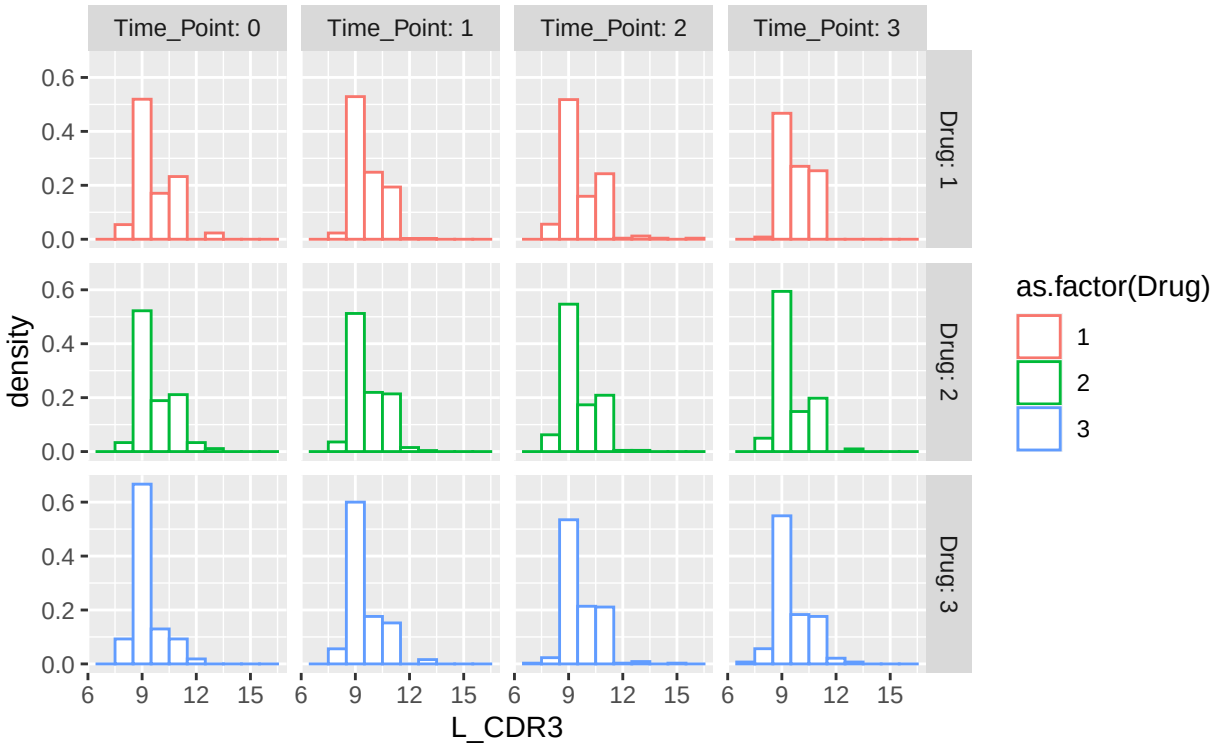

Figure 7: Histograms of L_CDR3 by Drug and Time Point

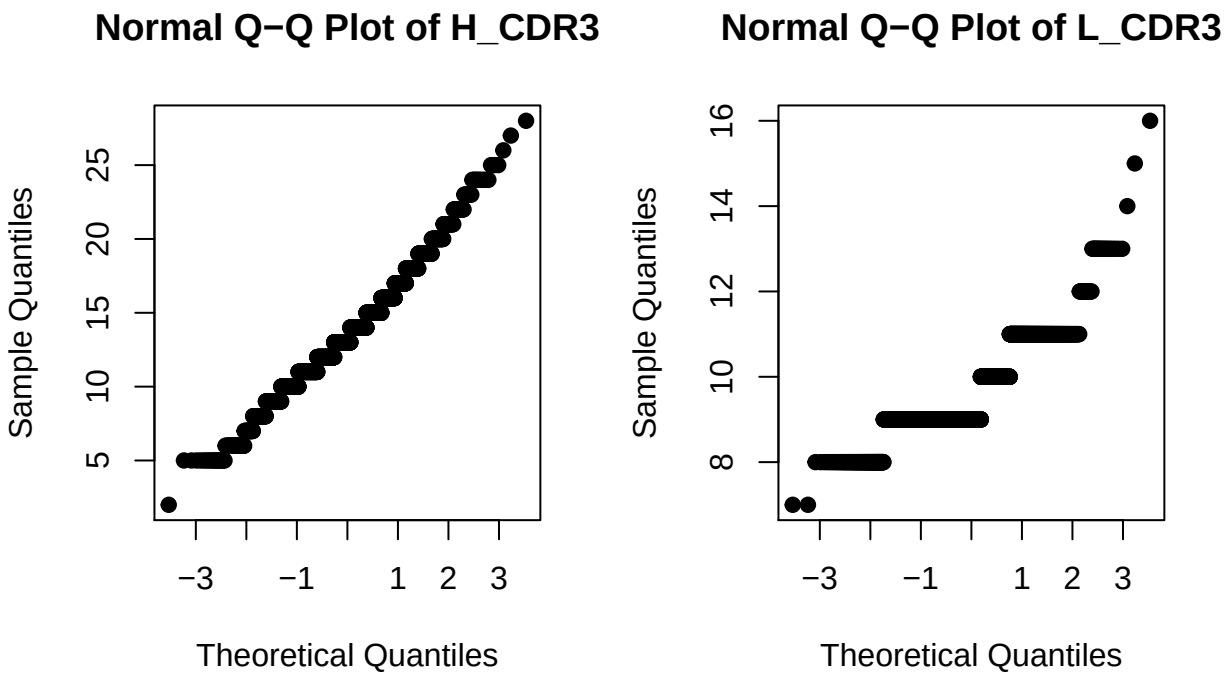

Figure 8: Q-Q Plots of H_CDR3 and L_CDR3

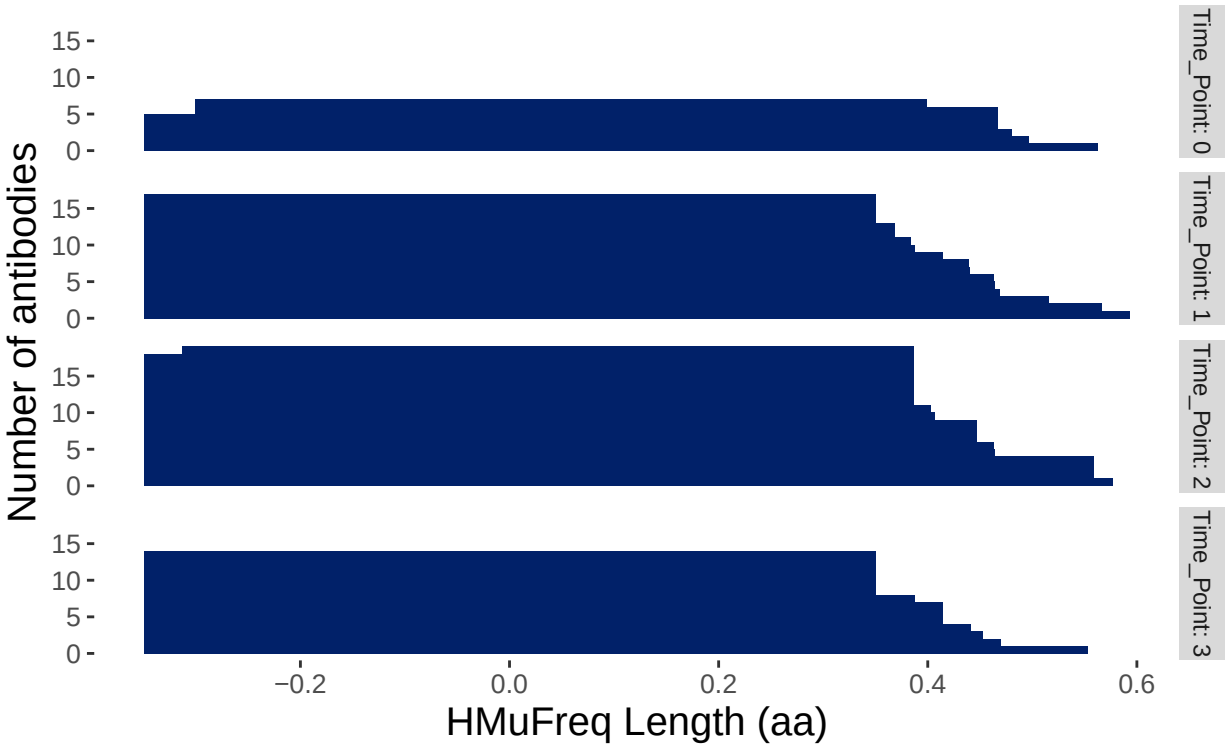

Figure 9: Histogram HMuFreq by Time Point

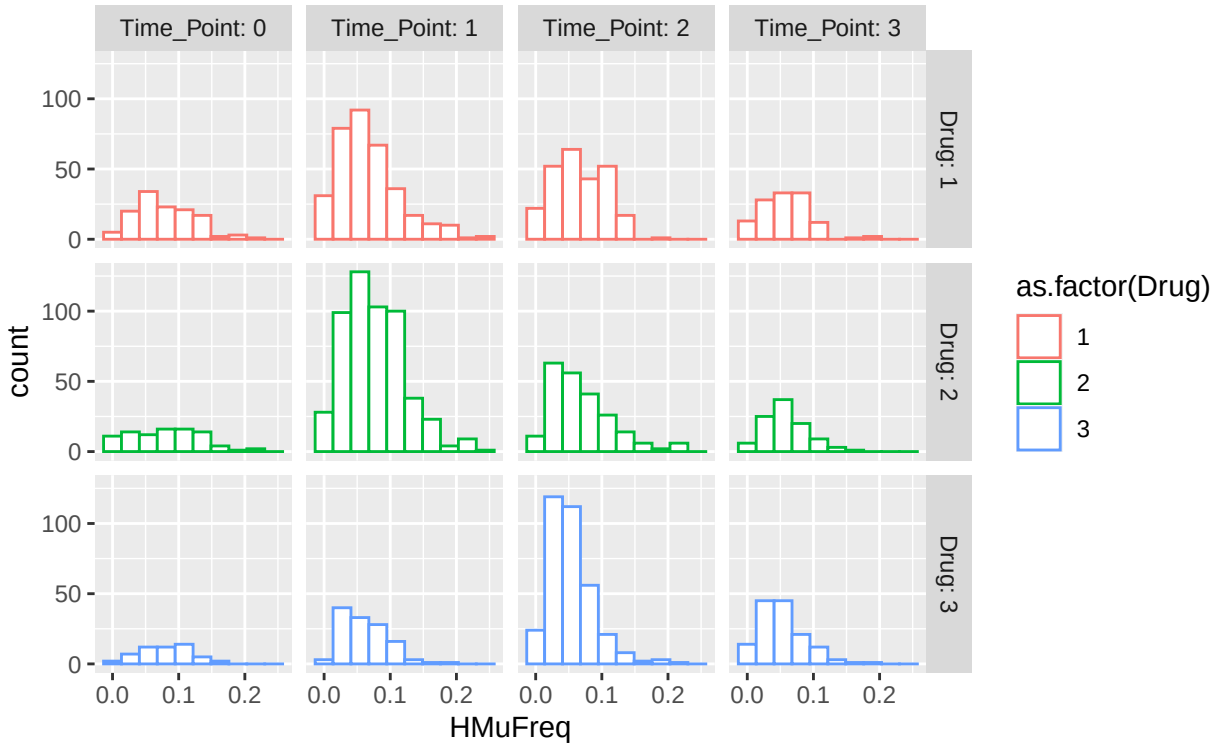

Figure 10: Histograms of HMuFreq by Drug and Time Point

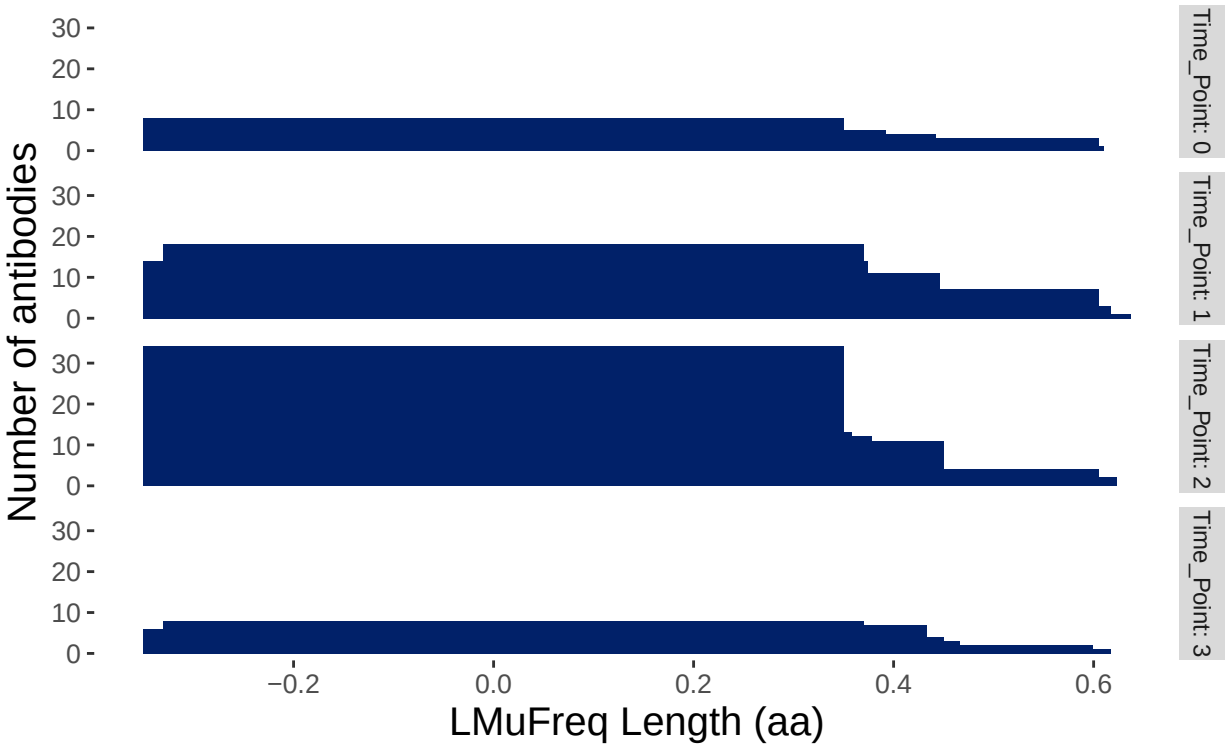

Figure 11: Histogram LMuFreq by Time Point

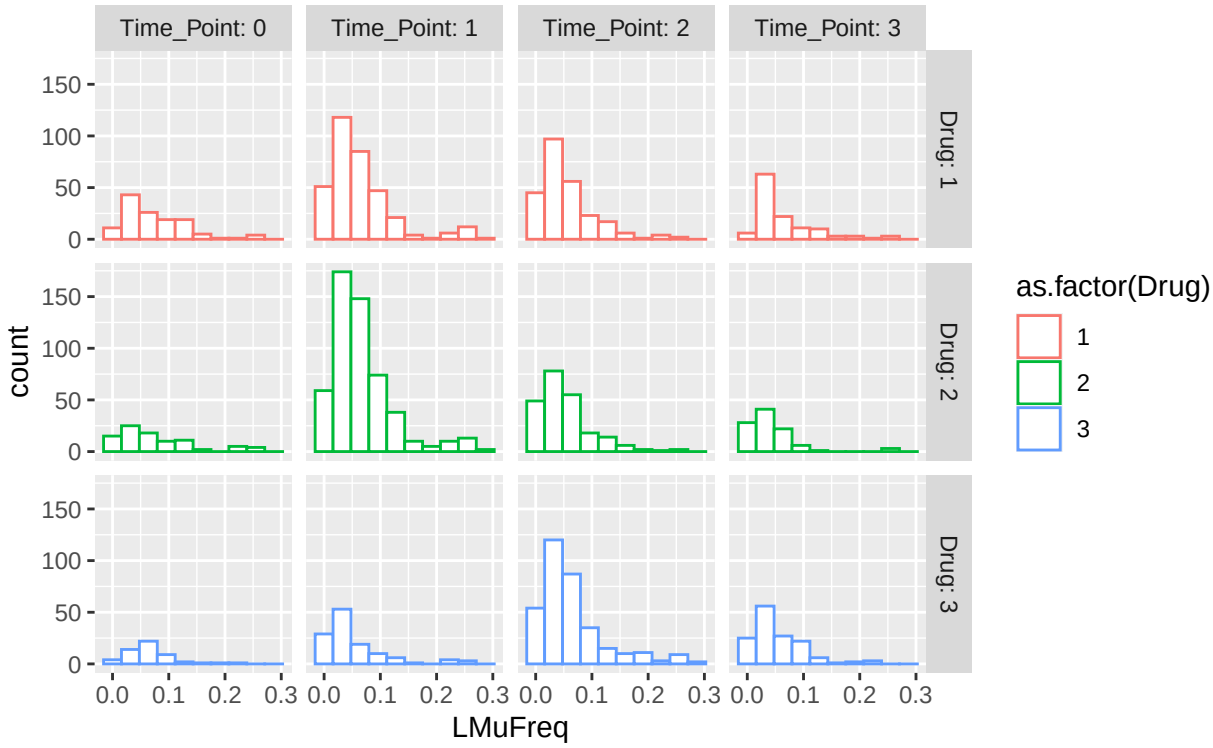

Figure 12: Histograms of LMuFreq by Drug and Time Point

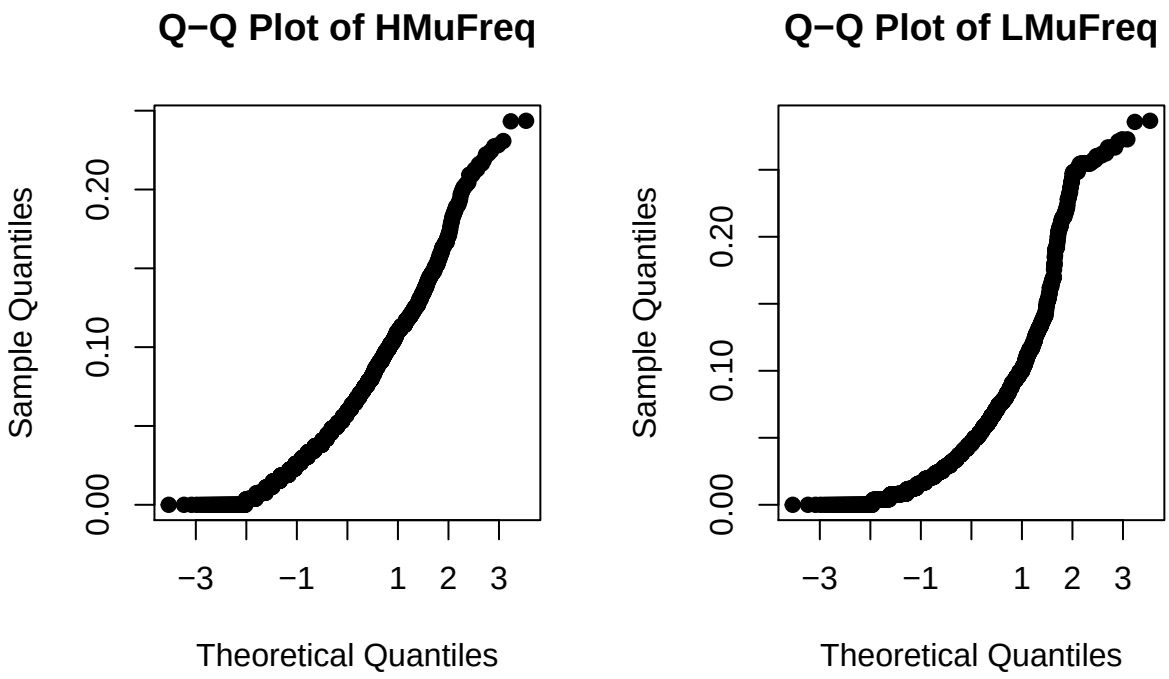

Figure 13: Q-Q Plot of HMuFreq and LMuFreq

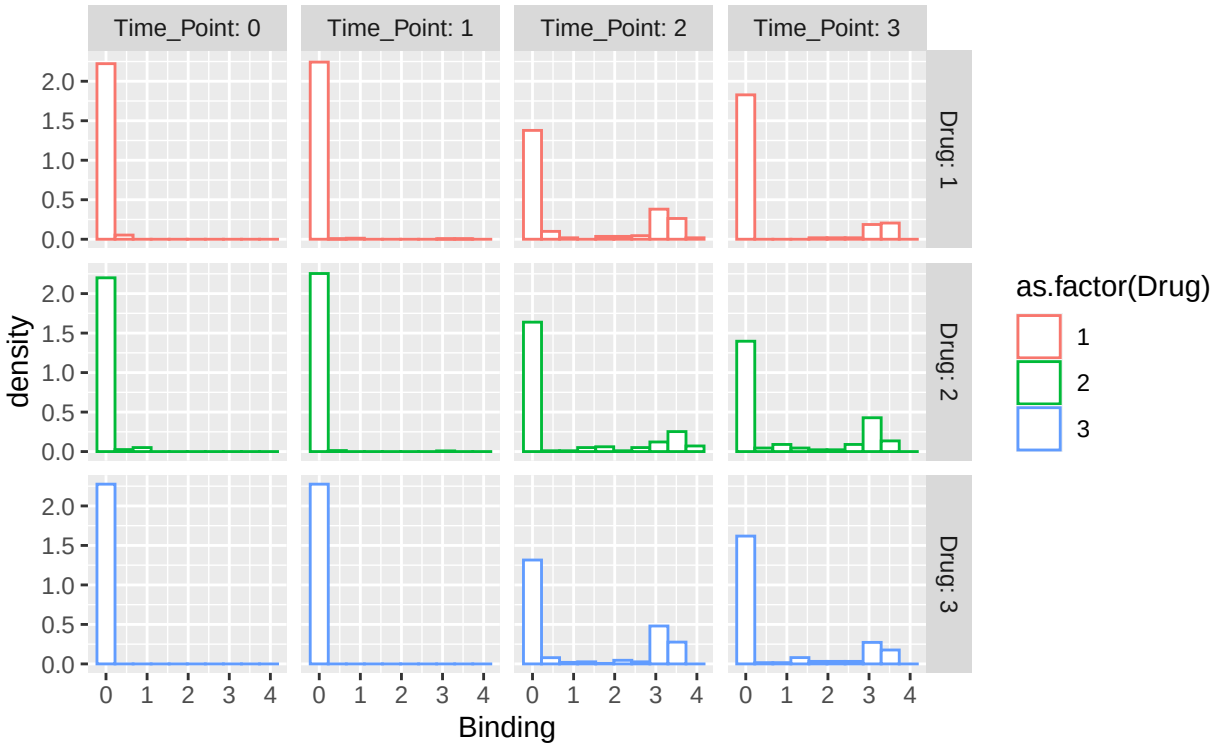

Figure 14: Histograms of Binding Strength by Drug and Time Point

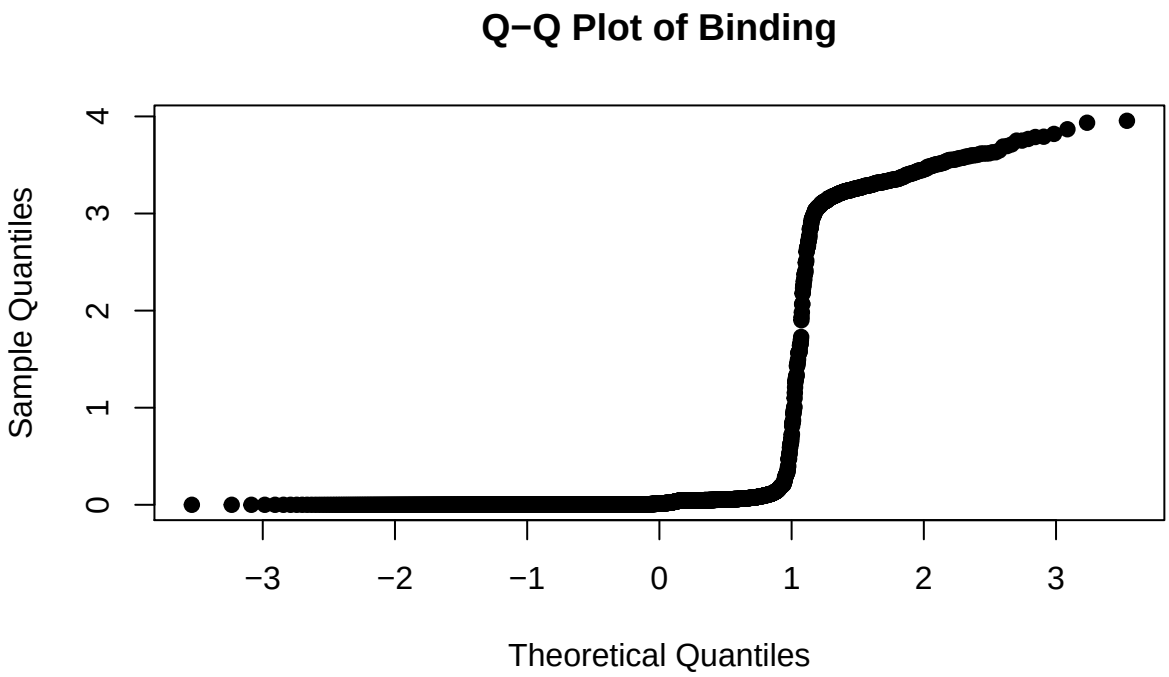

Figure 15: Q-Q Plot of Binding

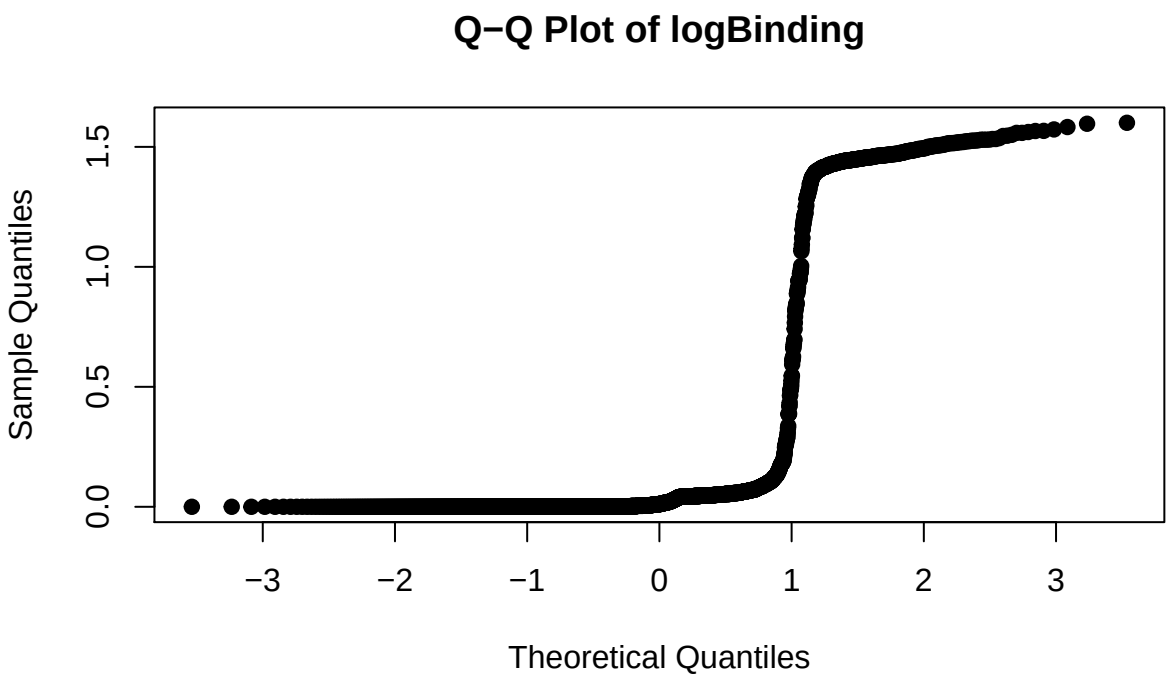

Figure 16: Q-Q plot of logBinding

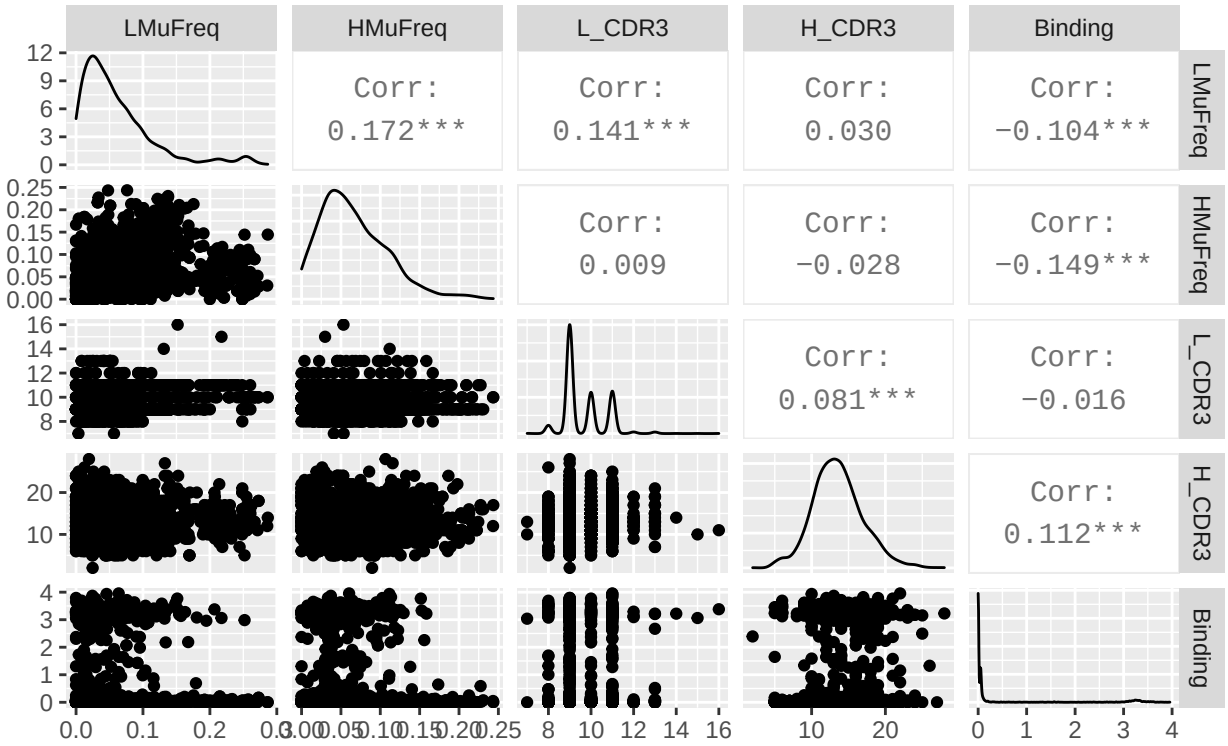

Figure 17: Scatterplot Matrix of Response Variables without Outlier

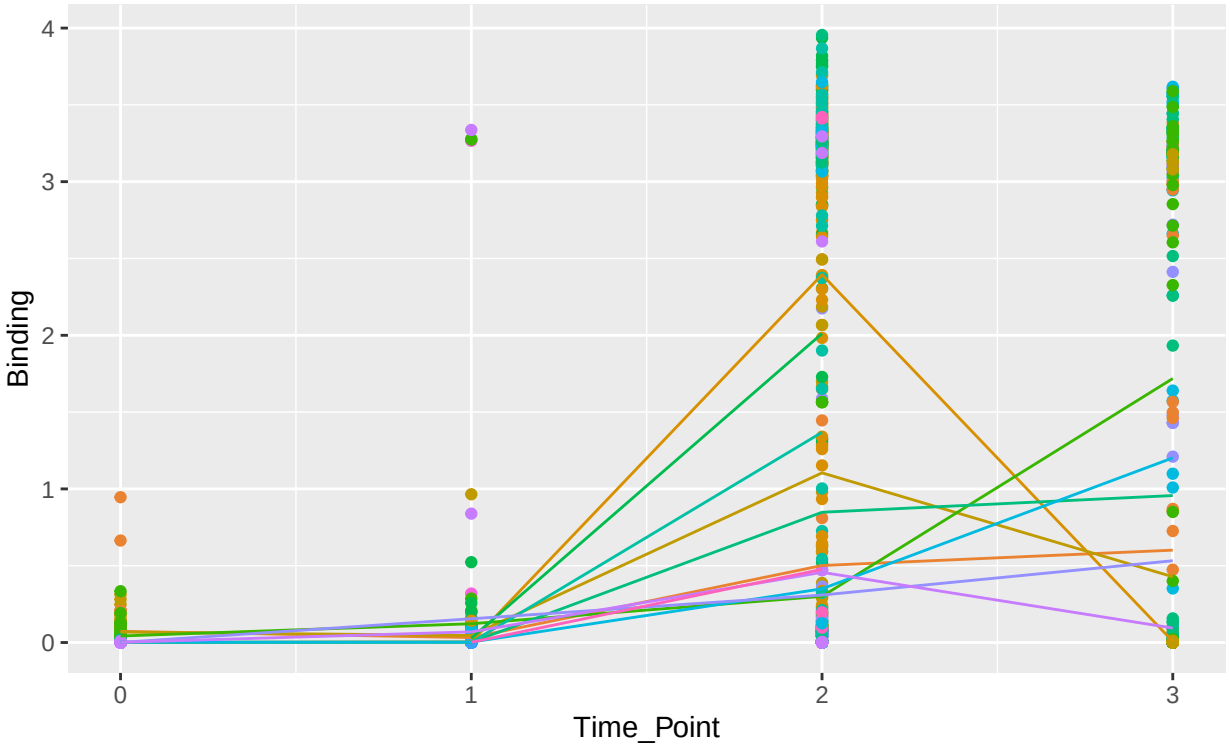

Figure 18: Mean Trends for Binding over Time by Macaque

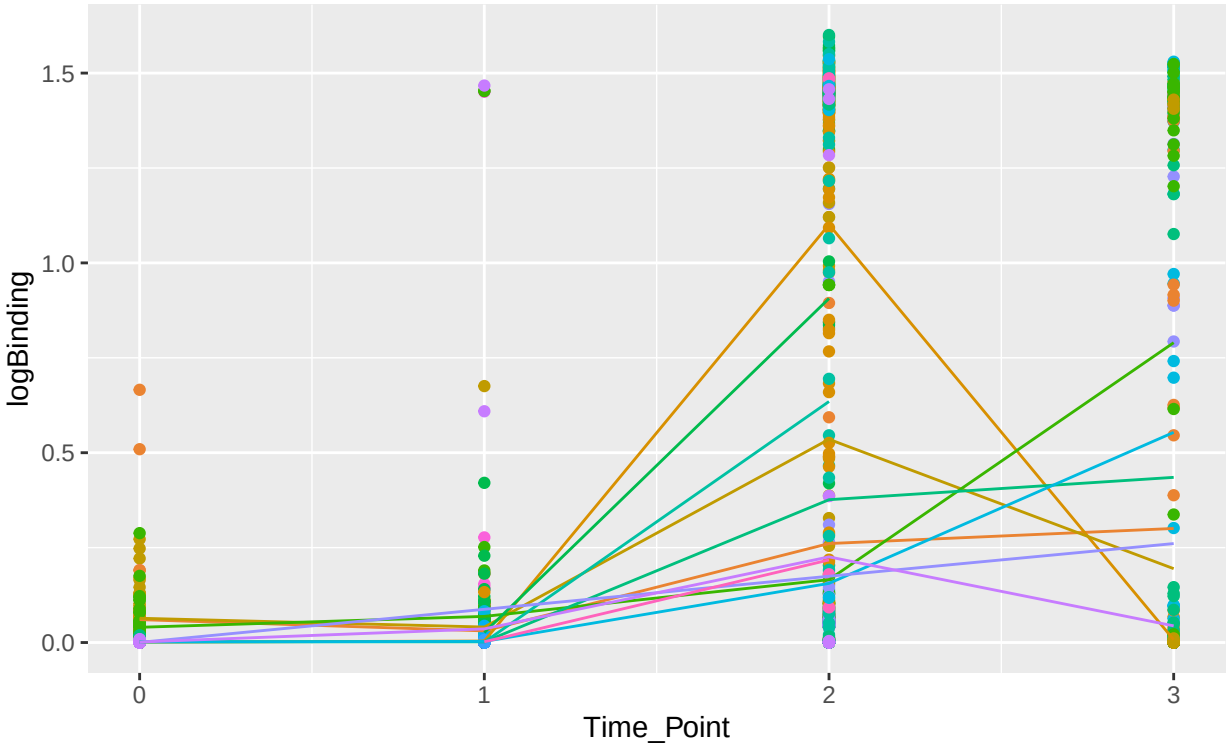

Figure 19: Mean Trends for logBinding over Time by Macaque

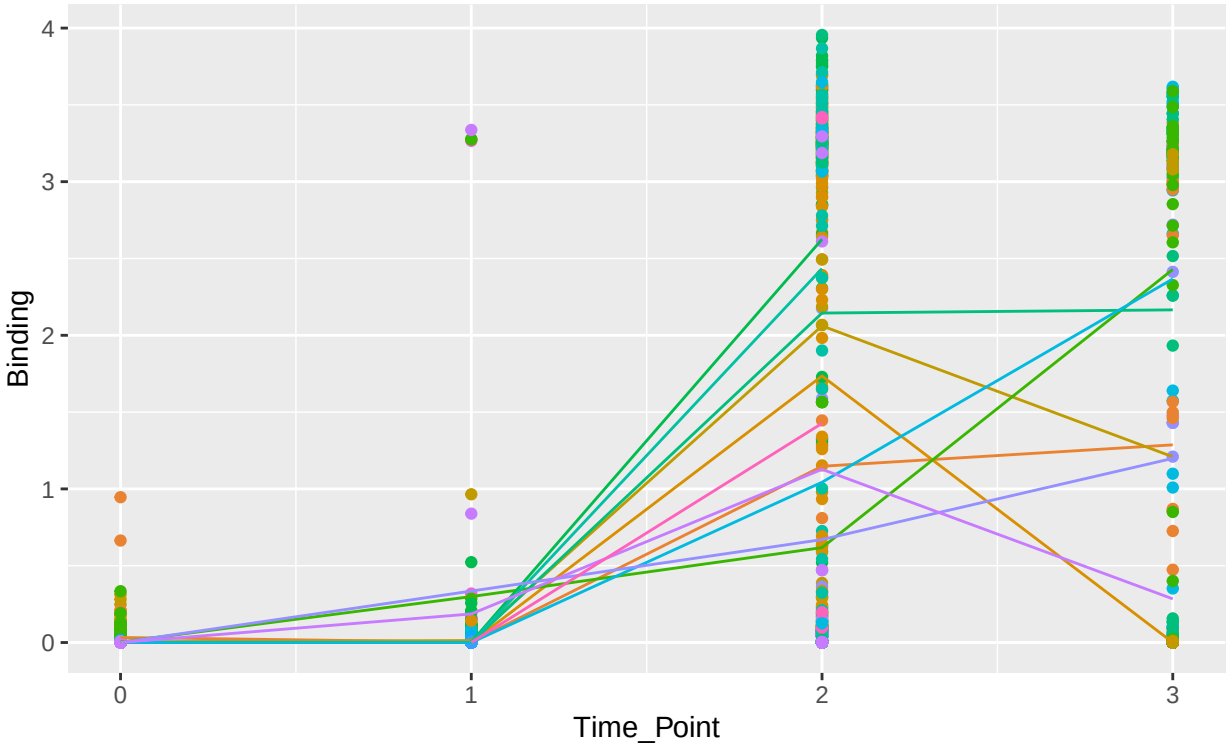

Figure 20: Variance Trends for Binding over Time by Macaque

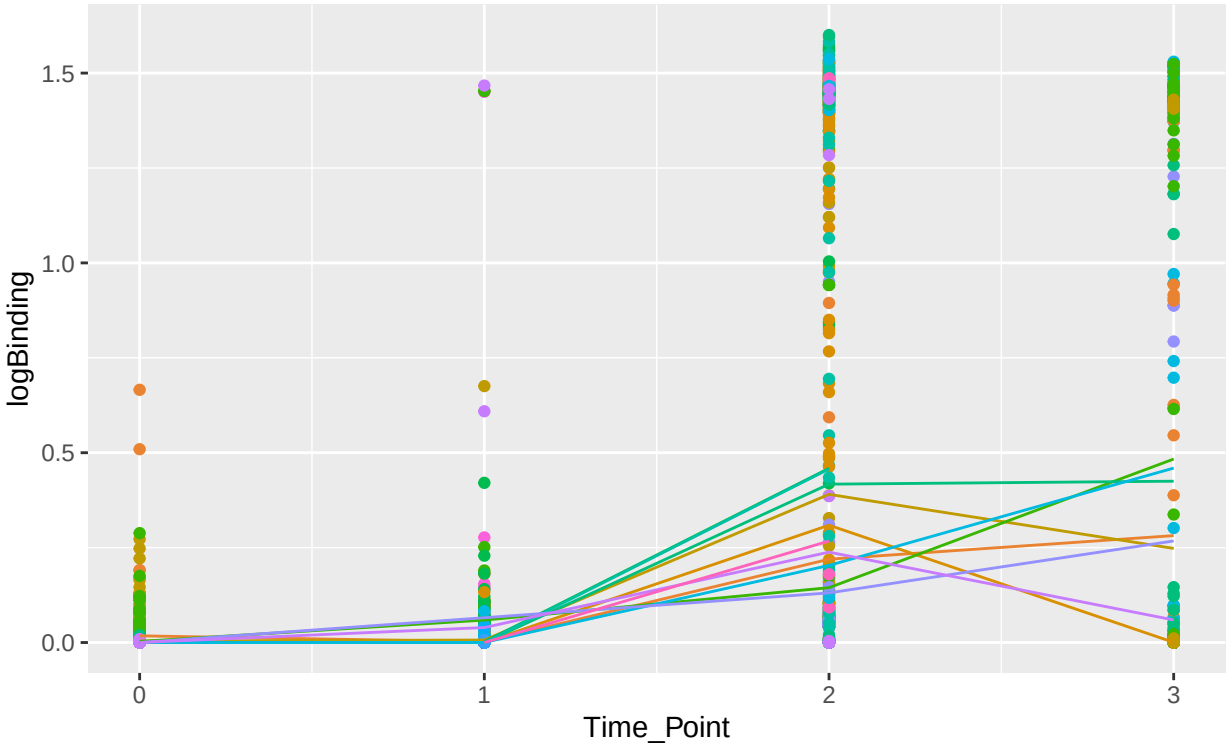

Figure 21: Variance Trends for logBinding over Time by Macaque

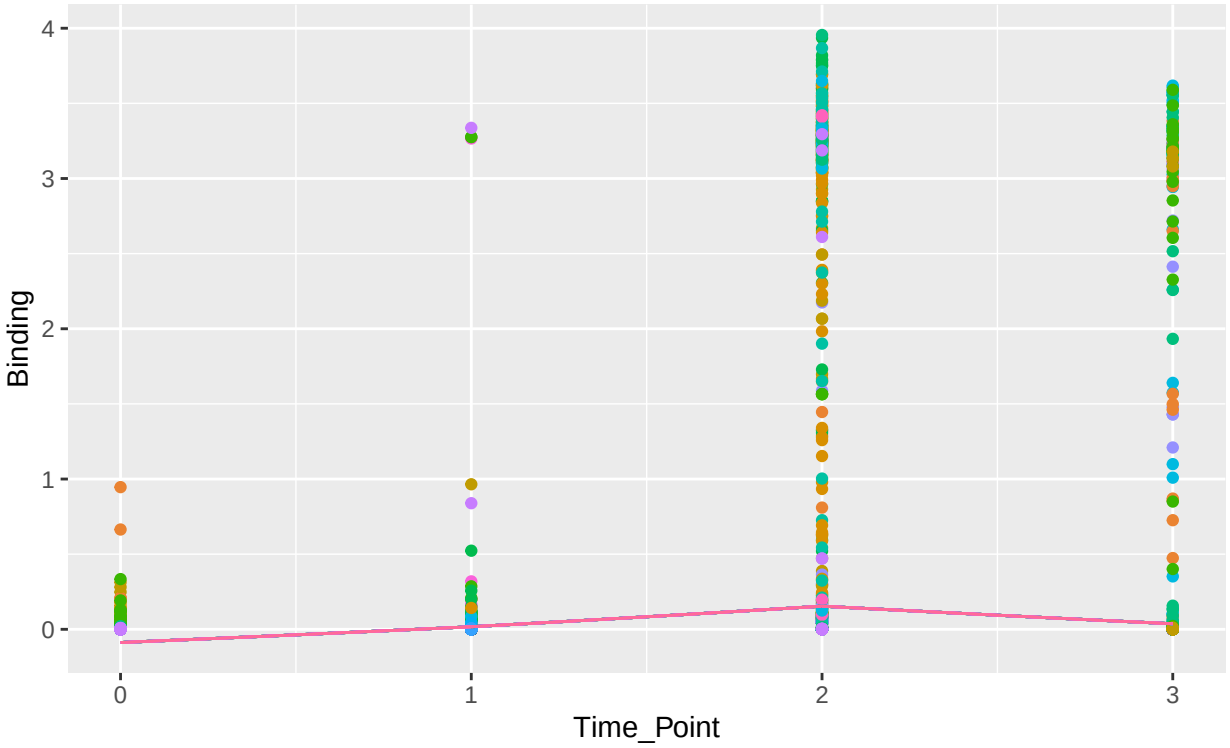

Figure 22: Piecewise Linear Function–Three Segments

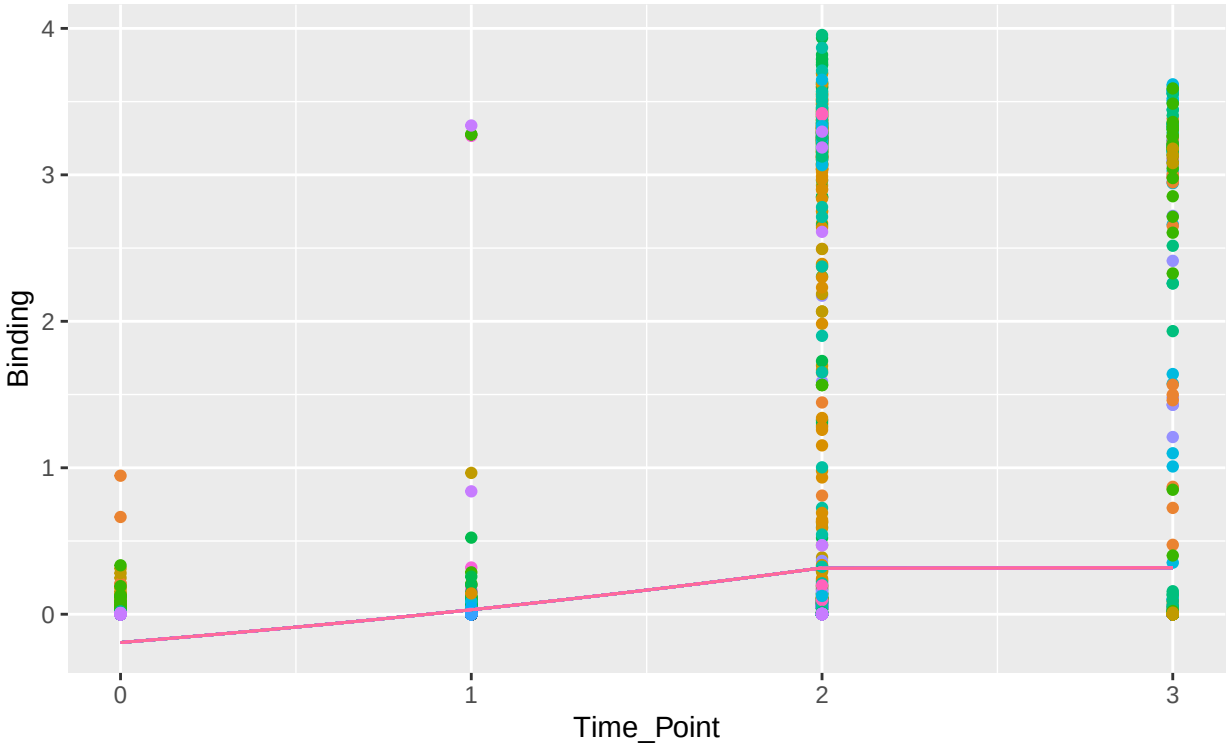

Figure 23: Piecewise Linear Function–Two Segments

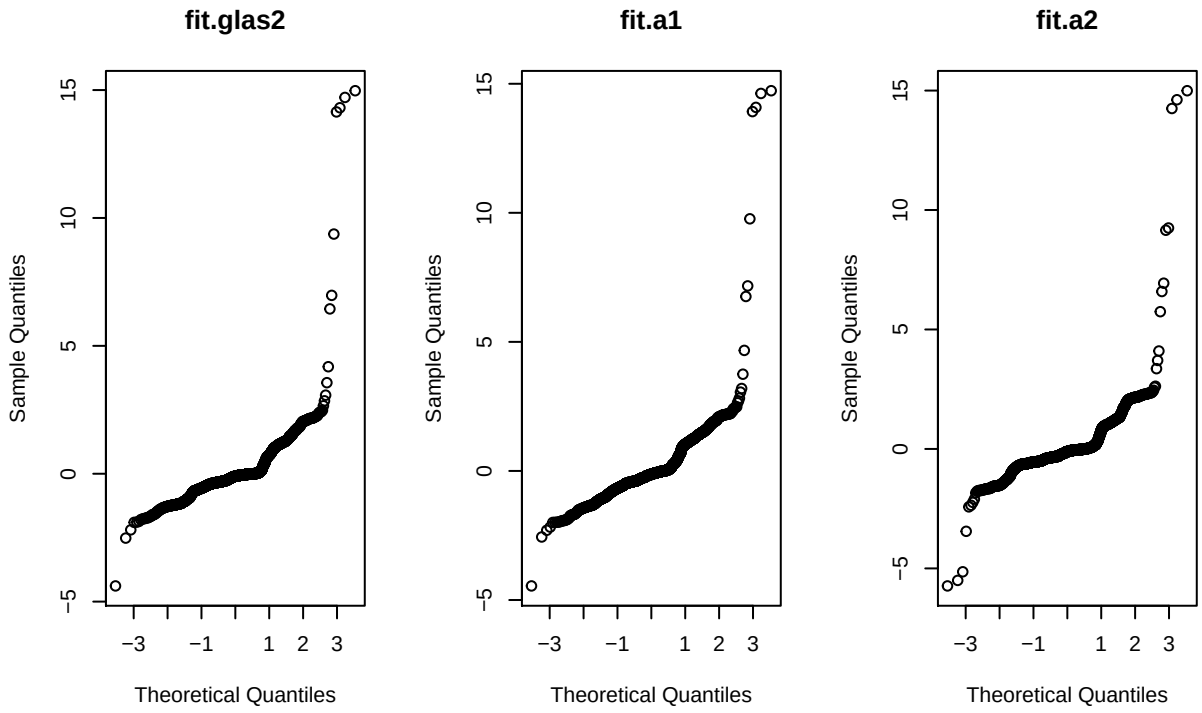

Figure 24: Q-Q Plots of Normalized Residuals

List of Tables

Table 1: Frequency Table of Drug and Time Point

	0	1	2	3
1	129	346	251	122
2	90	533	225	101
3	54	125	347	142

Table 2: Summaries of Standardized LCDR3

	Summaries
	Min. :-2.1860
	1st Qu.:-0.5361
	Median :-0.5361
	Mean : 0.0000
	3rd Qu.: 0.2888
	Max. :30.8110

Table 3: Significant Pairs by Time Point and Drug

	Time_Point	Drug
H_CDR3	3 > 1, 3 > 2	1 > 2
HMuFreq	0 > 2, 0 > 3, 1 > 2, 1 > 3	2 > 1 > 3
L_CDR3	none	none
LMuFreq	0 > 2, 0 > 3, 1 > 3	none

Table 4: AIC and BIC Comparison for Binding and logBinding

	df	AIC	df.1	BIC
fit.gls1	9	3293.6959	9	3345.9818
fit.gls2	9	-317.1771	9	-264.8913

Table 5: AIC and BIC Comparison between Two and Three Segments

	df	AIC	df.1	BIC
fit.gls2	9	-317.1771	9	-264.8913
fit.gls3	8	-310.8418	8	-264.3654

Table 6: AIC and BIC Comparison among GLS and LME Models

	df	AIC	df.1	BIC
fit.gls2	9	-317.1771	9	-264.8913
fit.a1	12	-407.9577	12	-338.2432
fit.a2	12	-591.4157	12	-521.7012

Table 7: Inference about S1, S2, and S3 Slopes

numDF	denDF	F.value	p.value
1	2441	58.83591	0
1	2441	137.96033	0
1	2441	56.69165	0

Table 8: AIC and BIC Comparison among GLS and LME Models with Time Point and Drug

	df	AIC	df.1	BIC
fit.gls2	9	-317.1771	9	-264.8913
fit.a1	12	-407.9577	12	-338.2432
fit.a2	12	-591.4157	12	-521.7012
fit.a3	20	-599.1739	20	-482.9831

Table 9: H0: drug 1 = drug 2

Fstat	p_value
2.483881	0.0418059

Table 10: H0: drug 1 = drug 3

Fstat	p_value
2.111719	0.0768574

Table 11: H0: drug 2 = drug 3

Fstat	p_value
5.251142	0.0003279

Table 12: Variance-covariance Matrices Comparison of Drug 2 and 3

	H_CDR3	HMuFreq	L_CDR3	LMuFreq
H_CDR3	0.8233467	0.0940387	1.7311285	-8.1369695
HMuFreq	0.0940387	1.6459865	-1.2232736	26.2072877
L_CDR3	1.7311285	-1.2232736	0.9656927	0.7827240
LMuFreq	-8.1369695	26.2072877	0.7827240	0.9973451