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

11/14/2020

Contents

1 Introduction

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.

During the experiment, 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). For the analysis of mutation frequency and CDR3 count, each antibody within the same treatment was treated as an independent observation.

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 experiment used T-cell suppression treatments to investigate the effect on immunization. These drugs are widely used in post transplant immuno-suppression treatment to prevent rejection.

While we might expect different variance within subject vs between subjects, the number of potential antibodies observed is much higher than the number sampled in a blood draw.

1.2 About the Dataset

Our dataset includes measurements of antibodies measured in 20 rhesus macaques after they were given the same HIV vaccine at three different time-points and one of three randomly selected anti-Treg treatments(drugs). Blood samples were collected two weeks after vaccine dosing, and antibodies were isolated from those samples. A different number of antibodies were collected from each blood sample, limited by assay yield. Each observation contains information about the antibody isolated post the glycoprotein immunization.

A human antibody is formed by a heavy chain and light chain. For heavy chain, a human has about 51 V-gene segments, 25 D-gene segments and 6 J-gene segment. For light chain there are 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 germline center. Theoretically, there can be $51 * 25 * 6 * 71 * 9 = 4.88835 \times 10^6$ combinations of gene segments. Considering the frequently happened mutation and other factors, each individual can have over **10 billion** different antibodies. Thus, we decided to follow the convention of vaccine studies and treat each antibody as independent. [Kan – can you identify a reference article or journal here that uses this convention? We don't need to quote it / change the answer, it's just defending our claim that it's a standard practice]

Below is the list of variables with a brief description from our dataset. Please note that in each antibody, there are two sets of heavy chain and light chain, all of which forming a Y-shape immunoglobulin. Thus many of the variables start with H or L, indicating which chain the information comes from.

1.3 List of variables

- Monkey_id: Lists the identity of monkey
- Treatment(Drug): Treatment A is the mock control, and treatment B and C are two different kinds of Treg inhibitor treatments.
- Time_Points: 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: The category of antibody type; 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 perform an role of primary response. The secondary response IgG appears later in serum with higher binding affinity, and neutralizing potentials against toxins and virus. IgA 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 heany chain.
- HMuFreq: calculated by H_Substitutions / H_VBase
- H_CDR3: the number of amino acid of the heavy chain's third complementarity determining region
- L_VBase: the number of nucleotide of the light chain variable region
- L_Substitutions: the number of relative nucleotide mutations in light chain.
- LMuFreq: calculated by L_Substitutions / L_VBase
- L_CDR3: the number of amino acid of the 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. [Kan, could you check to see if this is correct?] In other words, we want the values to be higher.
- Binding: affinity of antibodies against a selected HIV glycoprotein. The larger value indicates stronger binding.`Binding` indicates the rate of neutralizing, meaning how much the antibodies bind with the virus and thus make the virus ineffective. This is the most important measure of the study.

1.4 Research questions

The main focus of the current project is to understand whether isotypes, the number of vaccine injections, and the different Treg treatments cause changes in the antibody characteristics and if the changes are related to the immune responses against HIV virus. Specifically, we evaluate:

Q1: Do drugs and isotypes have effects on the mutation frequency and/or the amino acid count in the third complementarity determining region (CDR3)?

Q2: How does the binding strength of the antibodies develop in response to the number of vaccine dosages by treatment and drug types?

2 Methods

This section first presents some exploratory data analyses and summaries; then it uses multivariate and longitudinal data analyses to address two 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 monkeys, were collected at four different time points (0, 1, 2, 3) and each monkey was given one of three drugs (1 and 2 are immuno-suppressing drugs and 3 is the control). Figure 1 shows the histograms of antibody counts, and Table 1 and Table 2 show the antibody counts in different combinations of drugs, time points, and isoptypes.

As shown in Figure 2, the histograms of `Isotype`, we observed that IgG and IgM occupied the biggest proportion of antibodies in all time points. Before immunization (time point 0), there were similar weight of IgG and IgM found in blood. After the first immunization (time point 1), primary immune response resulted an increase of IgM, followed by an IgG increase at later time points 2 and 3.

2.1.2 Outlier Detection

Our response variables for the multivariate analysis include five variables: `Binding`, `H_CDR3`, `HMuFreq`, `L_CDR3`, and `LMuFreq`. As shown in Figure 3, `L_CDR3` in one point seems an outlier. As the summary statistics of standardized `L_CDR3` shown in Table 3, the maximum value is greater than 30, which is quite unusal. Figure 4 shows the Mahalanobis ditances and Z scores of `L_CDR3`, and the data point again appears to be an apparent outlier. The value for `L_CDR3` is quite unlikely. Since we can't go back to the original data, we remove the data point and will use the new dataset `Data3`.

2.1.3 Response Variables

We examined our responses: `H_CDR3`, `HMuFreq`, `L_CDR3`, `LMuFreq` and `Binding`. We observed that for `H_CDR3` the distributions were roughly normal with the center around 13 at different time-points (Figure 5) without taking into account different treatments. Figure 6 represents the distribution of H_CDR3 with respect to treatments at different time-points, and slightly centered around 9 for L_CDR3 at different time points. With L_CDR3, Figure 7 and Figure 8 show approximately normal distribution with a longer right tail. The Q-Q plots in Figure 9 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 chain and similarly for light chain. These two variables show how much the antibodies mutate. A higher mutation rate is usually indicative of better virus neutralization. Below we present comparison of mutation rate between heavy chain and light chain. Figure 10, Figure 11, Figure 12, and Figure 13 show that HMuFreq and LMuFreq are both approximately normally distributed, each with a longer right tail. The Q-Q plots in Figure 14 confirm the approximate normality of HMuFreq and LMuFreq.

Next, a histogram of Binding with respect to treatments at different time points and Q-Q plot are shown in Figure 15 and Figure 16. We observe that Binding was not normally distributed. However, since our sample size is larger than 2000, we can use the Central Limit Theorem and assume normality. Lastly, we check whether response variables could be correlated, as shown in Figure 17. In these plots, we observe that none of the response variables were highly correlated.

2.2 Multivariate Data Analysis

To answer **Q1** (Do drugs and isotypes have effects on the mutation frequency and/or the amino acid count in the third complementarity determining region (CDR3)?), we test whether predictors Drug and Isotype had effects on the five responses: H_CDR3, HMuFreq, L_CDR3, LMuFreq, and Binding.

2.2.1 MANOVA

Since there are more than two populations we are comparing, we use MANOVA to test for effects. We check that the normality assumption is met due to large sample size (n = 2464). Each antibody is assumed to be independent. However, when we check for equal variance-covariance structures–by Fligner-Killeen Test of Homogeneity of Variances and by checking the variance-covariance matrices of different populations–many of the response variables do not meet the equal variance-covariance matrix assumption. The only populations that meet the assumption is the response variable vector

(L_CDR3, LMuFreq)^T. Thus we perform a MANOVA test on **(L_CDR3, LMuFreq)^T ~ Time_Point + Drug + Isotype** The output below shows that the main effects of Drug and Isotype have p-values greater than 0.05, suggesting that these main effects are not significant to the two traits, L_CDR3 and LMuFreq, of antibodies. The p-value for Time_Point is quite small. Thus we conclude that Time_Point has significant effect to the two traits, L_CDR3 and LMuFreq, of antibodies. The next section, longitudinal data analysis, will examine the effects of time points and reach a similar conclusion.

```
##             Df   Wilks approx F num Df den Df   Pr(>F)
## drug         2 0.99808   1.1768      4   4906   0.3188
## it           4 0.99462   1.6559      8   4906   0.1039
## tp           3 0.98773   5.0610      6   4906 3.47e-05 ***
## Residuals 2454
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.2.2 Pairwise Comparison

```
## [1] "L_CDR3  pairwise CI's"
##  contrast estimate     SE   df lower.CL upper.CL
##  0 - 1     -0.0237 0.0649 2460   -0.210    0.162
##  0 - 2     -0.0355 0.0664 2460   -0.226    0.155
##  0 - 3     -0.0102 0.0761 2460   -0.228    0.208
##  1 - 2     -0.0117 0.0447 2460   -0.140    0.117
##  1 - 3      0.0135 0.0581 2460   -0.153    0.180
##  2 - 3      0.0253 0.0598 2460   -0.146    0.197
##
## 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
##  0 - 1     0.00912 0.00368 2460 -0.001421   0.0197
##  0 - 2     0.01524 0.00376 2460  0.004447   0.0260
##  0 - 3     0.01948 0.00431 2460  0.007122   0.0318
##  1 - 2     0.00611 0.00253 2460 -0.001153   0.0134
##  1 - 3     0.01036 0.00329 2460  0.000916   0.0198
##  2 - 3     0.00424 0.00339 2460 -0.005473   0.0140
##
## Results are averaged over the levels of: rep.meas
## Note: contrasts are still on the [.: scale
## Confidence level used: 0.995833333333333
```

The pairwise comparison results show that the changes in `LMuFreq` are significant between time points 0 and 2, 0 and 3, and 1 and 3.

2.3 Longitudinal Analysis

To answer our **Q2** (How does the binding strength of the antibodies develop in response to the number of vaccine dosages by treatment and drug types?), we use longitudinal data analysis, including general linear models and linear mixed models. For the longitudinal analysis of binding strength vs number of vaccine doses, we use the `gls` and `lme` functions from the nlme package[7].

2.3.1 One Covariate: Time Point

We first take a look at the data over time. As seen in Figure 18 and Figure 19, the mean trend is not linear, and the different time points have different variances. This information suggests that we should use piecewise linear models and set variances as unequal over time.

We first consider a model with time point as the only covariate:

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

We then turn the model above into a piecewise linear model, in which each segment has different intercepts and slopes. We use three indicator variables: $S1, S2, S3$ as the indicator variables, where

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

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

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

The new model is thus

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

We also want to make sure that the trend is continuous at time points = 1 and 2. Our first complete model (`fit.gls`) is now $Y_{ij} = \beta_0(S1 + 2S2 - S2Time_{ij}) + \beta_1(S1Time_{ij} + 2S2 - S2Time_{ij}) + \beta_4(-S2 + S2Time_{ij} + S3) + \beta_5(-2S2 + 2S2Time_{ij} + S3Time_{ij}) + e_{ij}$ where

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

The model can also be written as

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

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

From the model above, we can find the intercepts and slopes for all three segments of the mean trend:

- S1: $-0.2221651 + 0.2432183 * time$
- S2: $(2 * -0.2221651 + 2 * 0.2432183 - 0.7699600 + 2 * 0.2432756) + (0.2221651 - 0.2432183 + 0.7699600 - 2 * 0.2432756) * time = -0.2413024 + 0.2623556 * time$

- S3: $0.7699600 - 0.2432756 * time$

We make a plot of the line segments in Figure 20, which shows the two segments S1 and S2 have very similar slopes. So we can refit the model with only two piecewise sections between time points 0 and 2 and between time points 2 and 3. We'll call them S4 and S5. The new model is therefore

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

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

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

Again, we want to make sure that the trend is continuous at Time_Point = 2. Our second complete model (`fit.gls2`) 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 can also be written as $Y_{ij} = S4(-2\beta_1 + \beta_2 + 2\beta_3) + S4Time_{ij}(\beta_1) + S5(\beta_2) + S5Time_{ij}(\beta_3) + e_{ij}$

We first find the mean trend for S4 and S5:

- S4: $(-2{*}0.5310975{+}0.5720853{+}2{*}{-}0.0000723){+}0.5310975{*}time = -0.4902543 + 0.5310975 * time$
- S5: $0.5720853 - 0.0000723 * time$

We can make the plot again to see if the model is reasonable, as shown in Figure 21. Indeed, there is a linear line between Time_Point 0 and 2 and one between Time_Point 2 and 3. The two lines are contiuous at Time_Point 2. A comparison of AIC And BIC of these two models, shown in Table 4, indicates that the second model (`fit.gls2`) is indeed a better model.

Next we check whether adding random effects improve our second complete model (`fit.gls2`). We assume that random effects exist in the intercept and slope. Our linear mixed model is then:

$Y_{ij} = \beta_1(-2S4+S4Time_{ij})+\beta_2(S4+S5)+\beta_3(2S4+S5Time_{ij})+b_{0i}+b1iTime_{ij}+e_{ij}$
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 uneuqal variances over time. As shown in Table 5, the model `fit.a2` has the lowest AIC And BIC, so it seems the best model. We now check residuals for three models: `fit.gls2`, `fit.a1`, `fit.a2`, as shown in Figure 22. All three Q-Q plots show approximate normality. To further investigate the effects of drugs, We now use `fit.a2` for further analysis.

Now we would like to know if the slopes between Time_Point 0 and 2 and between Time_Point 2 and 3 equal zero. H_0 : slope of $S4 = 0$ and slope of $S5 = 0$, which means $H_0 : \beta_1 = 0$ and $\beta_3 = 0$

Thus, we can check for two tests:

$$\mathbf{L_1} = 0$$

where $\mathbf{L_1} = (1, 0, 0)$ and $= (\beta_1, \beta_2, \beta_3)^T$ and

$$\mathbf{L_2} = 0$$

where $\mathbf{L_2} = (0, 0, 1)$ and $= (\beta_1, \beta_2, \beta_3)^T$

As shown in Table 6, the slop of S4 has a very small p-value, while the slope of S5 is quite large, indicating that the change in Binding rate between Time_Point 0 and Time_Point 2 is significant while the change between Time_Point 2 and Time_Point 3 is not significant. We conclude that Time_Point 2, when the monkeys had received two vaccines, had the highest Binding rate, while the last vaccine shot at Time_Point 3 did not make a difference to the Binding rate.

2.3.2 Two Covariates: Time Point and Drug

Next we add `Drug` as a covariate to the model `gls.a2` to see if it has effects on Binding. We use two indicator variables: `D2` and `D3`, where

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

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

Assuming that the random effects are the same for each drug, our model (`fit.a3`) with the extra covariate `Drug` is:

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

$$\beta_4 D2(-2S4 + S4Time_{ij}) + \beta_5 D2(S4 + S5) + \beta_6 D2(2S4 + S5Time_{ij}) +$$

$$\beta_7 D3(-2S4+S4Time_{ij})+\beta_8 D3(S4+S5)+\beta_9 D3(2S4+S5Time_{ij})+b_{0i}+b1iTime_{ij}+e_{ij}$$

where

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

and

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

With the model constructed, we want to make inference on β to find whether the drugs have any effects. To see whether Drug 1 and Drug 2 have any difference, we want to perform a hypothesis test on $H_0 : \beta_4 = \beta_5 = \beta_6 = 0$, thus we can do the test

$$\mathbf{L_3} = 0$$

where

$$\mathbf{L_3} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

and $= (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9)^T$

To see whether Drug 1 and Drug 3 have any difference, we want to perform a hypothesis test on $H_0 : \beta_7 = \beta_8 = \beta_9 = 0$, thus we can do the test

$$\mathbf{L_4} = 0$$

where

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

and $= (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9)^T$

To see whether Drug 2 and Drug 3 have any difference, we want to perform a hypothesis test on $H_0 : \beta_4 = \beta_7, \beta_5 = \beta_8, \beta_6 = \beta_9$, thus we can do the test

$$\mathbf{L_5} = 0$$

where

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

and $= (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9)^T$

We found that, as shown in Table 7, Drug 1 and Drug 2 do not have significantly different effects on Binding rates. As shown in Table 8, Drug 1 and Drug 3 do not have significantly different effects on Binding rates. Also, as shown in Table 9, Drug 2 and Drug 3 do not have significantly different effects on Binding rates. In other words, drug groups do not have signifant effects on our longitudinal model. Thus we will retain `fit.a2` as our best model.

3 Results

For multivariate analyses, we performed a MANOVA test on the main effects of `Isotype` and `Drug` on the response variable vector (`H_CDR3`, `HMuFreq`, `L_CDR3`, `LMuFreq`, `Binding`)$^{\mathrm{T}}$. We found that both `Isotype` and `Drug` have very small p-values. We also performed pairwise comparison to see where the effects were.

For longitudinal analyses, we found that the best linear mixed model is $Y_{ij} = \beta_1(-2S4 + S4Time_{ij}) + \beta_2(S4 + S5) + \beta_3(2S4 + S5Time_{ij}) + b_{0i} + b1iTime_{ij} + e_{ij}$
where

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

and

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

Using inferences on β for the two line segments, We rejected the hypothesis that the slope of the first line segment (between time points 0 and 2) is zero, but we failed to reject the hypothesis that the slope of hte second line segment (between time points 2 and 3) is zero. We then added `Drug` as another covariate to the above model and made inference on β. The comparison between three drug groups was done with three F-tests (between Drug grups 1 and 2, 2 and 3, and 1 and 3), which led to three p-values greater than 0.05. Thus We failed to reject the null hypotheses that the mean trend of either of the two drug groups was equal.

4 Discussion

While our analyses reached some findings, some further investigations could improve our analyses. A common method to analyze antibody traits is to treat antibodies (rows in our data) as independent from each other. We considered perform statistical analyses to prove or disprove such a convention, but it requires more biological knowledge and the investigation would most likely be beyond the scope of a final report. However, this remains an interesting topic that could be explored.

For longitudinal data analyses, we did not try out more combinations for models. For example, we only tried two correlation structures (compound symmetry and AR1); other structures might achieve better results. When we added drug as another covariate, we did not go back to test which correlation structure might perform better or whether the piecewise model should include two or three line segments.

5 Conclusions

In this report, we performed multivariate data analyses and longitudinal data analyses to understand whether time points, drugs, and isotypes have effects on characteristics of antibodies. Our statistical analyses provide answers to our two research questions. We performed a MANOVA test to our first research question, "Do drugs and isotypes have effects on the mutation frequency and/or the amino acid count in the third complementarity determining region (CDR3)?" and found significant main effects for both drugs and isotypes. However, pairwise comparisons reveal that drugs increased mutation rates for heavy chain but did not increase binding rates, which are the key to better efficacy of vaccines. [Kan: how to interpret IgG's higher mutation rates and binding rates?]

To answer our second research question, How does the binding strength of the antibodies develop in response to the number of vaccine dosages by treatment and drug types?", we first used time point as the only covariate and constructed with two general linear models with two and three line segments. We then added random effects for intercept and slope for time point as well as different correlation structure–compound symmetry and AR1–and found the model with two line segments (time points 0 to 2 and time points 2 to 3), random effects of intercept and slope of time point, AR1 correlation structure, and unequal variances over time performs best. F-tests for inferences for β reveals that time point 2 have the highest binding rates, suggesting that two vaccine injections improved the binding rates while the third injection did not make a difference. We also found that adding drug as a covariate did not improve the model.

6 References

The dataset, which can be found here, was provided by Kan Luo, as he was one of authors for the following four publications that used the dataset:

1. Luo K, Liao HX, Zhang R, et al. Tissue memory B cell repertoire analysis af-

ter 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. Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team (2020). *nlme: Linear and Nonlinear Mixed Effects Models*. R package

List of Figures

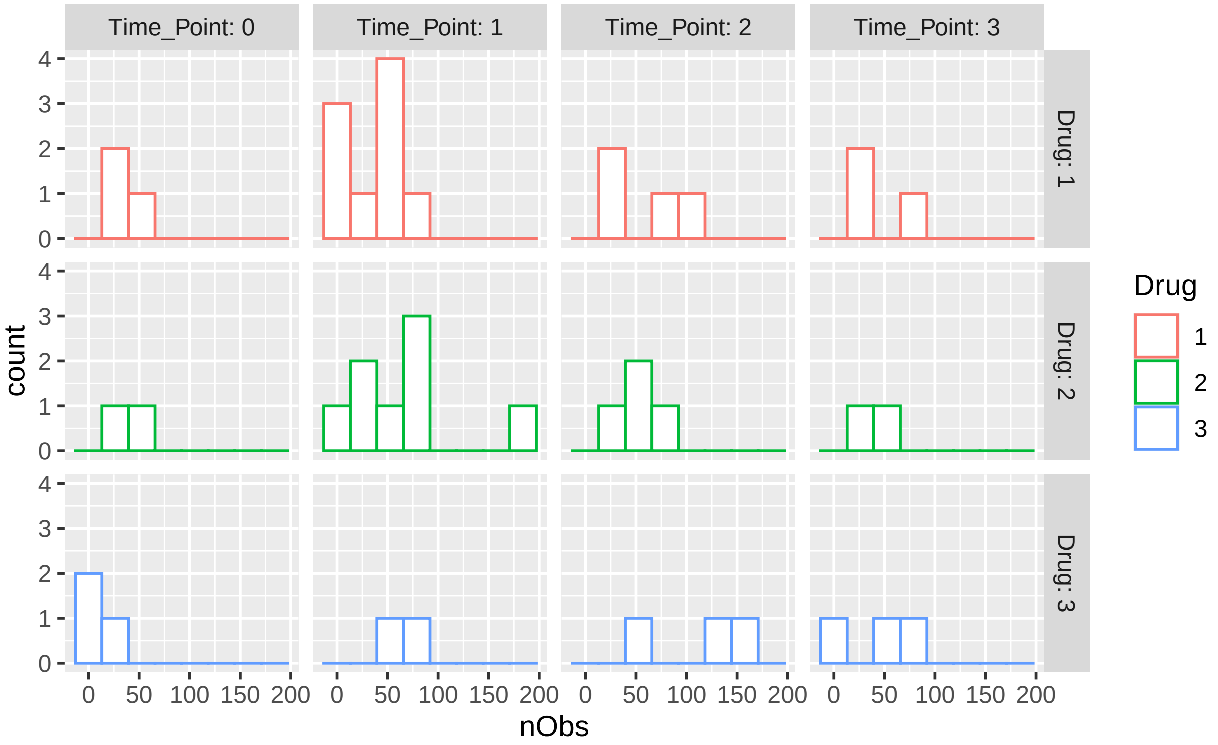

Figure 1: Histograms of Antibodies

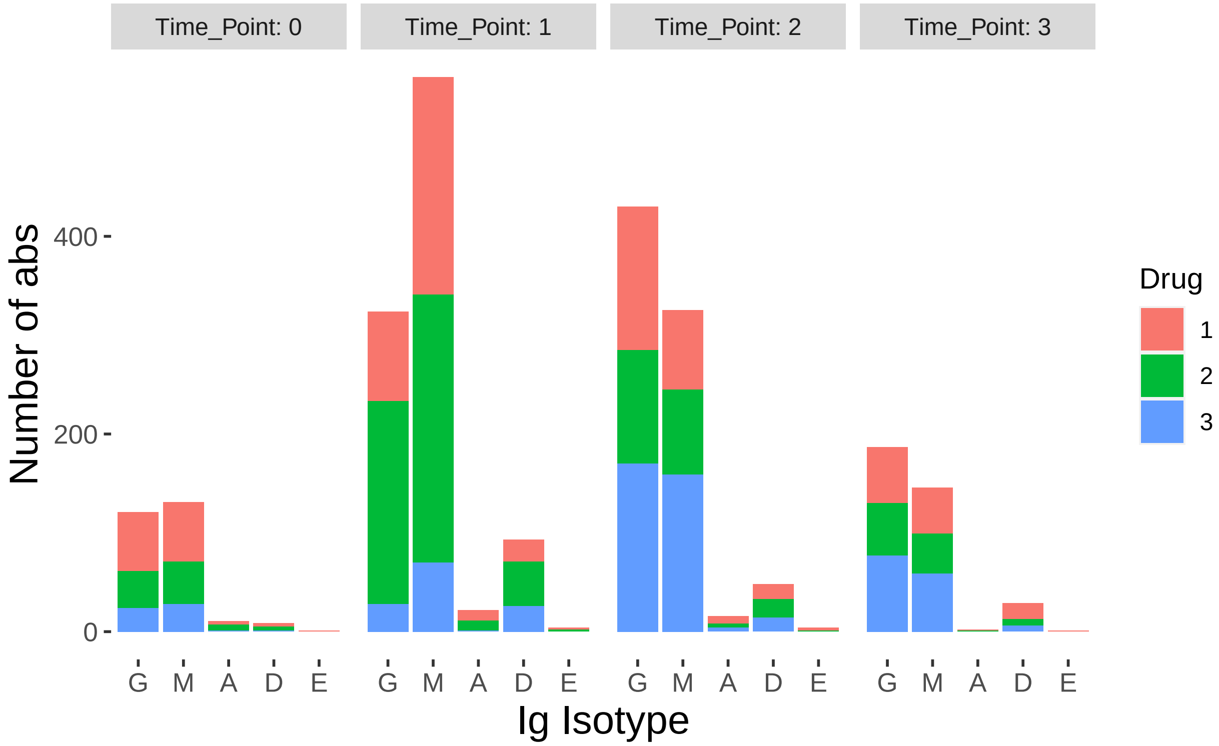

Figure 2: Histograms of Isotypes

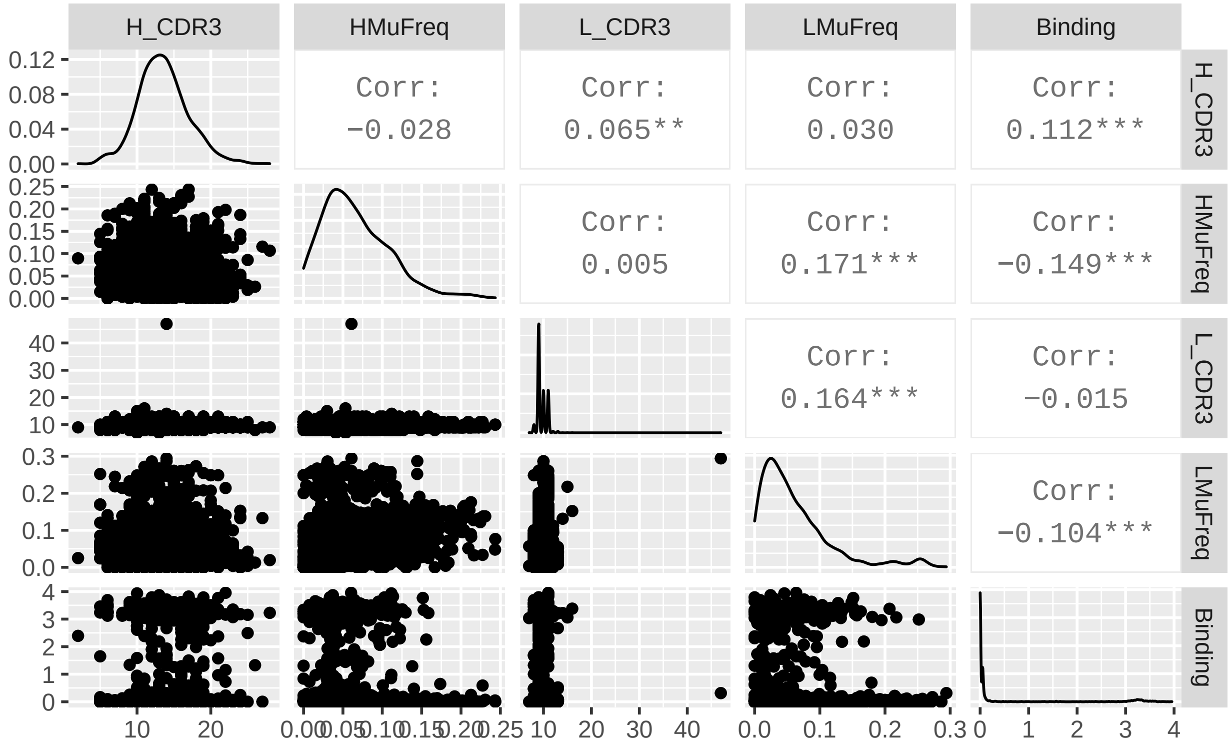

Figure 3: Histogram of Response Variables

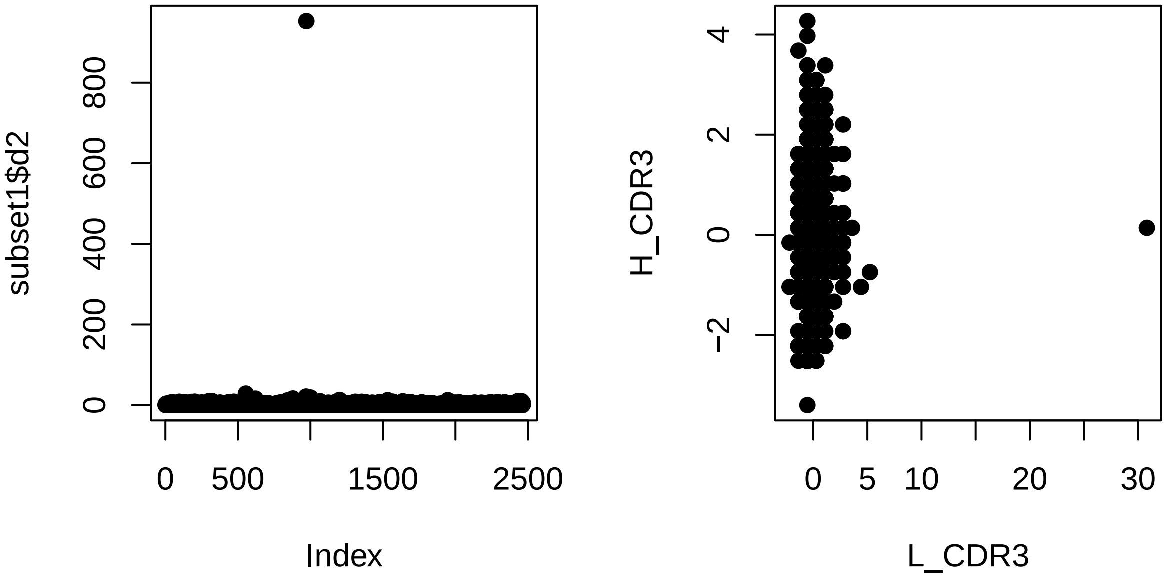

Figure 4: Mahalanobis distances and Z scores

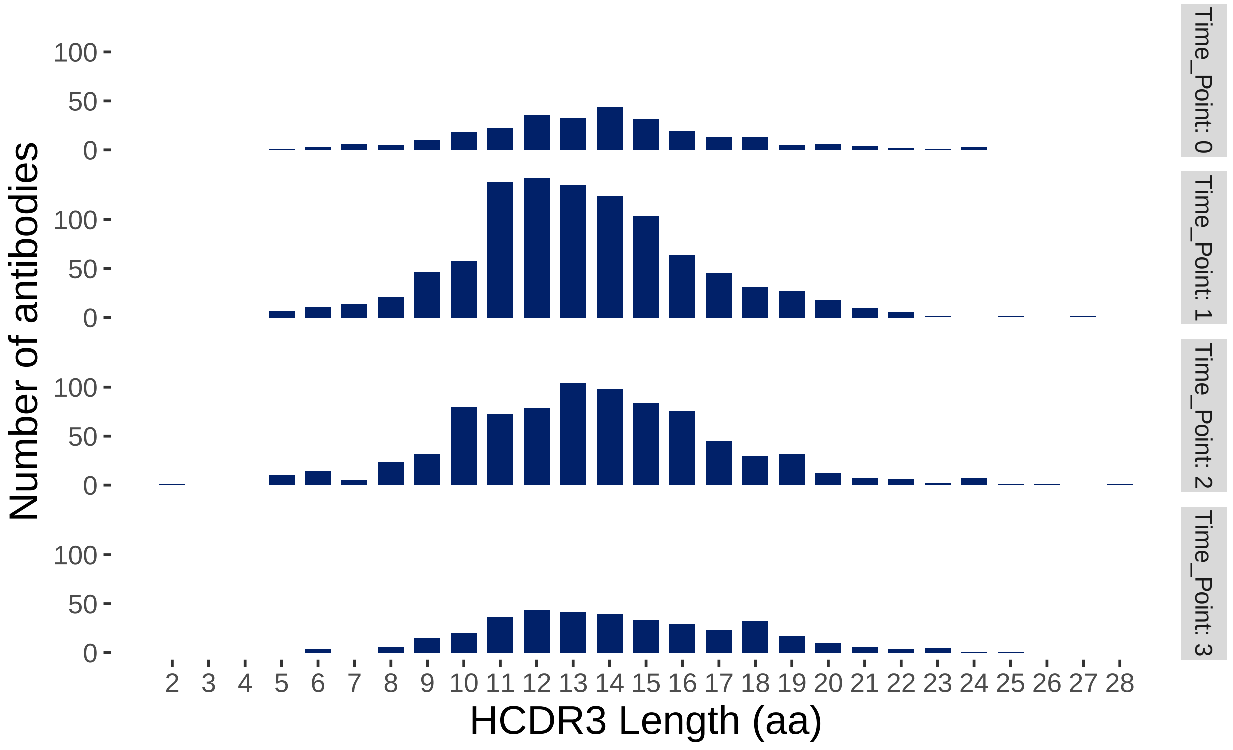

Figure 5: Histogram H_CDR3

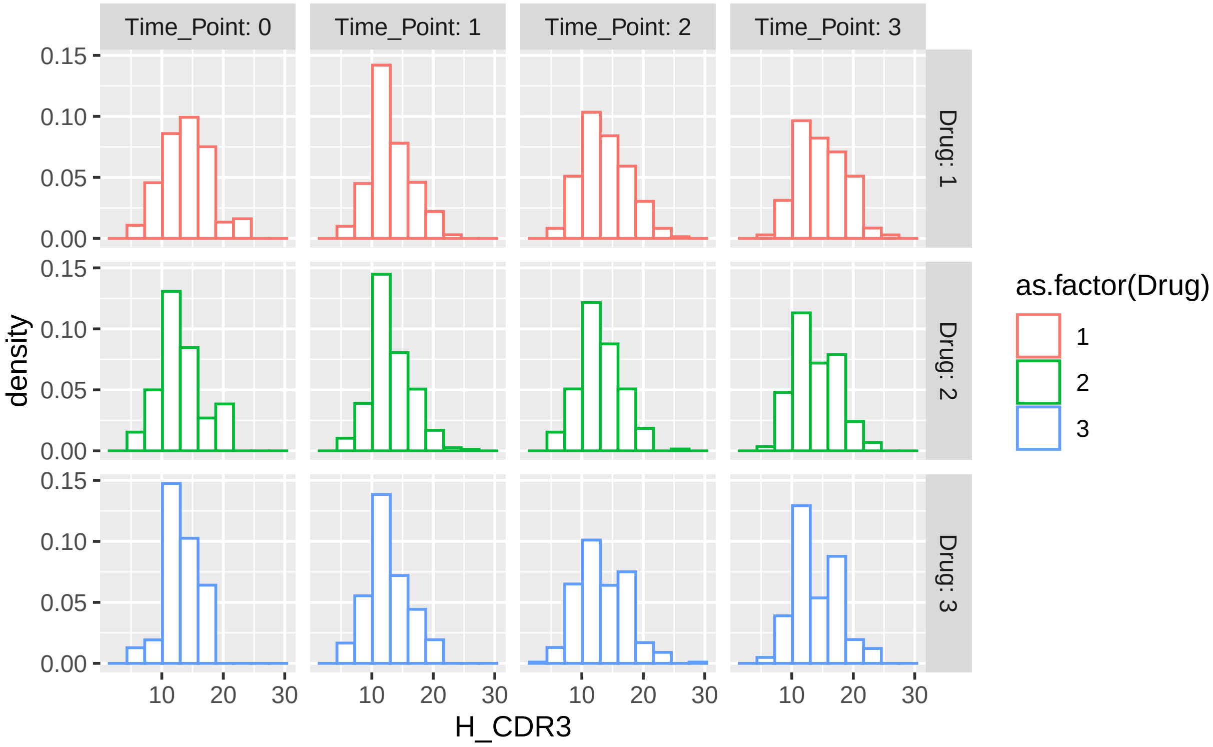

Figure 6: Histograms of H_CDR3 vs Treatment and Timepoint

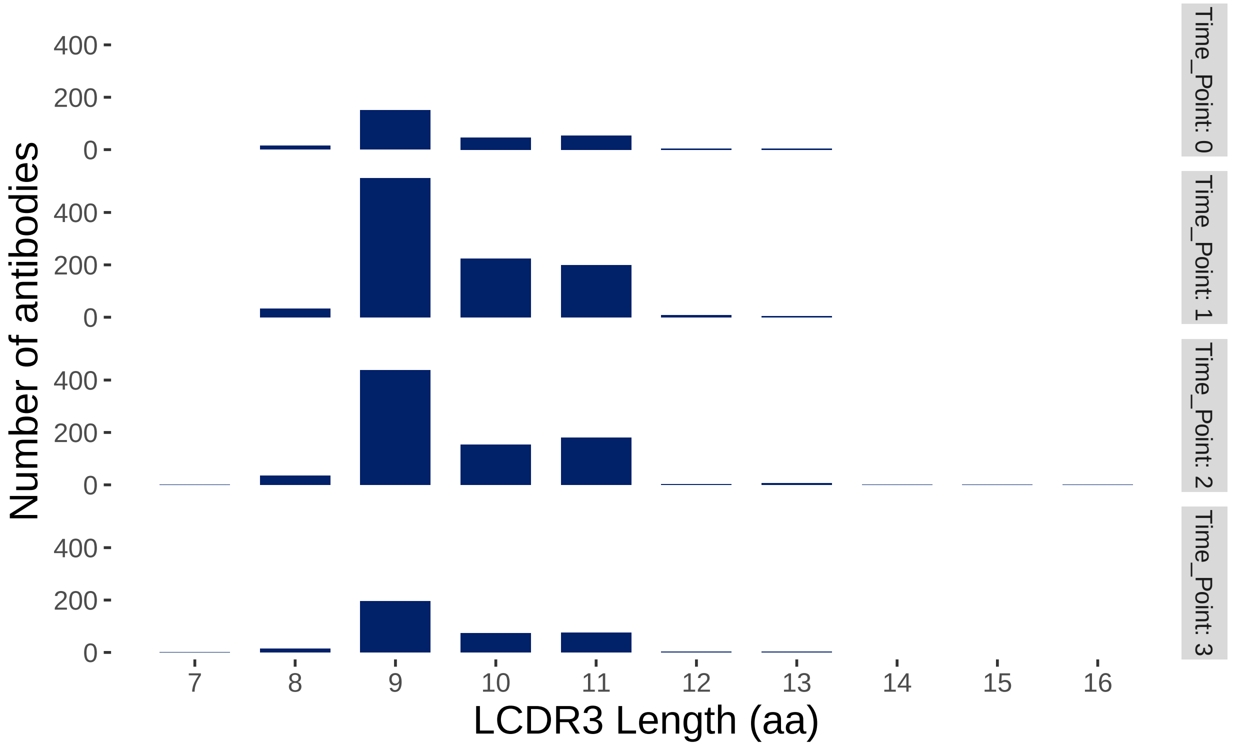

Figure 7: Histogram L_CDR3

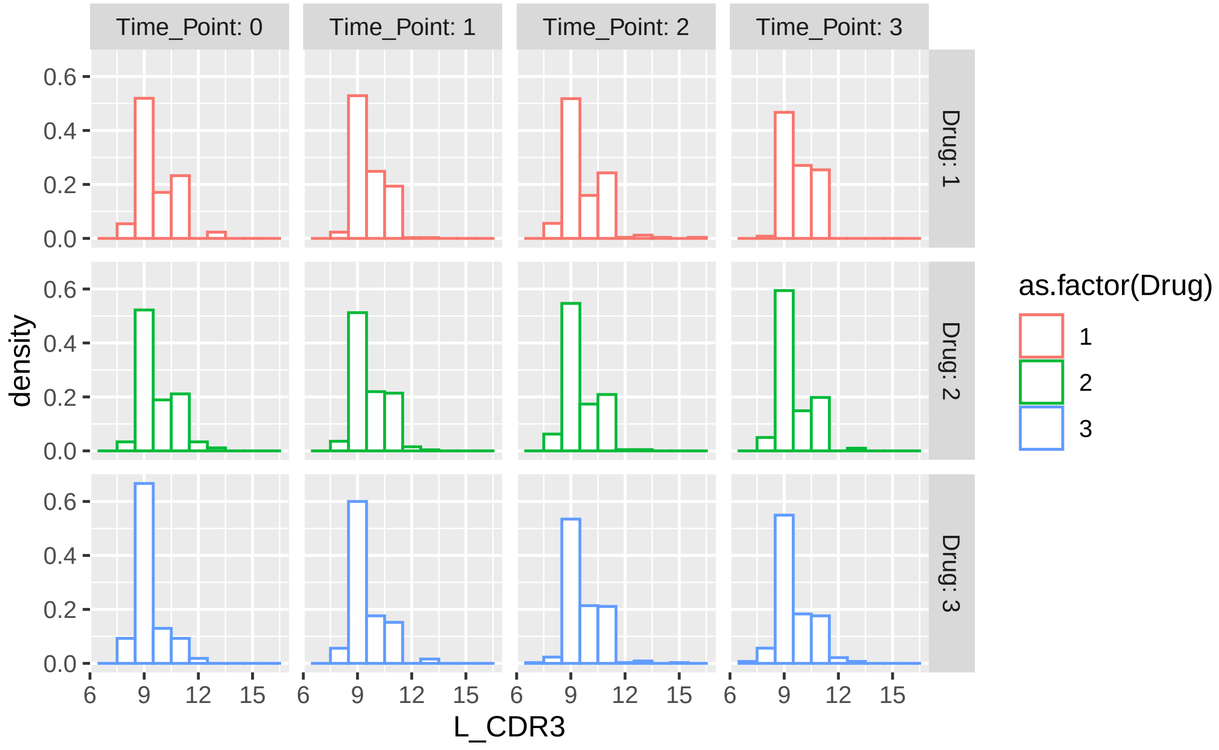

Figure 8: Histograms of L_CDR3 vs Treatment and Timepoint

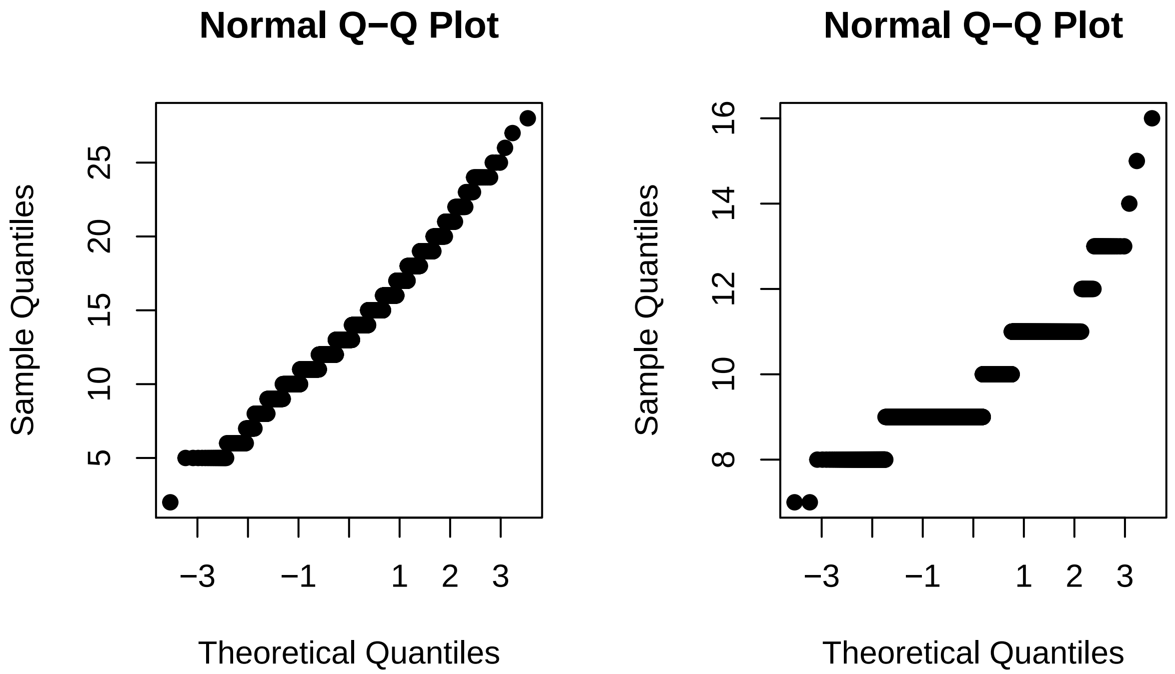

Figure 9: Q-Q Plots of H_CDR3 and L_CDR3

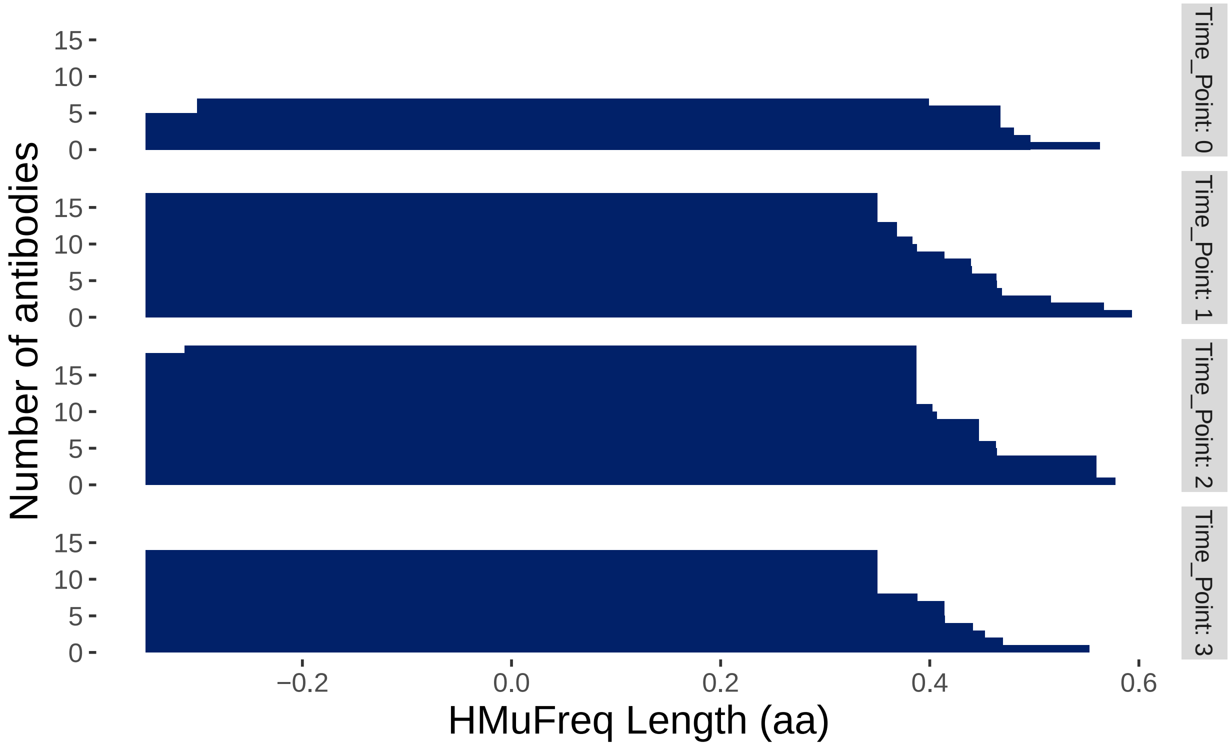

Figure 10: Histogram HMuFreq

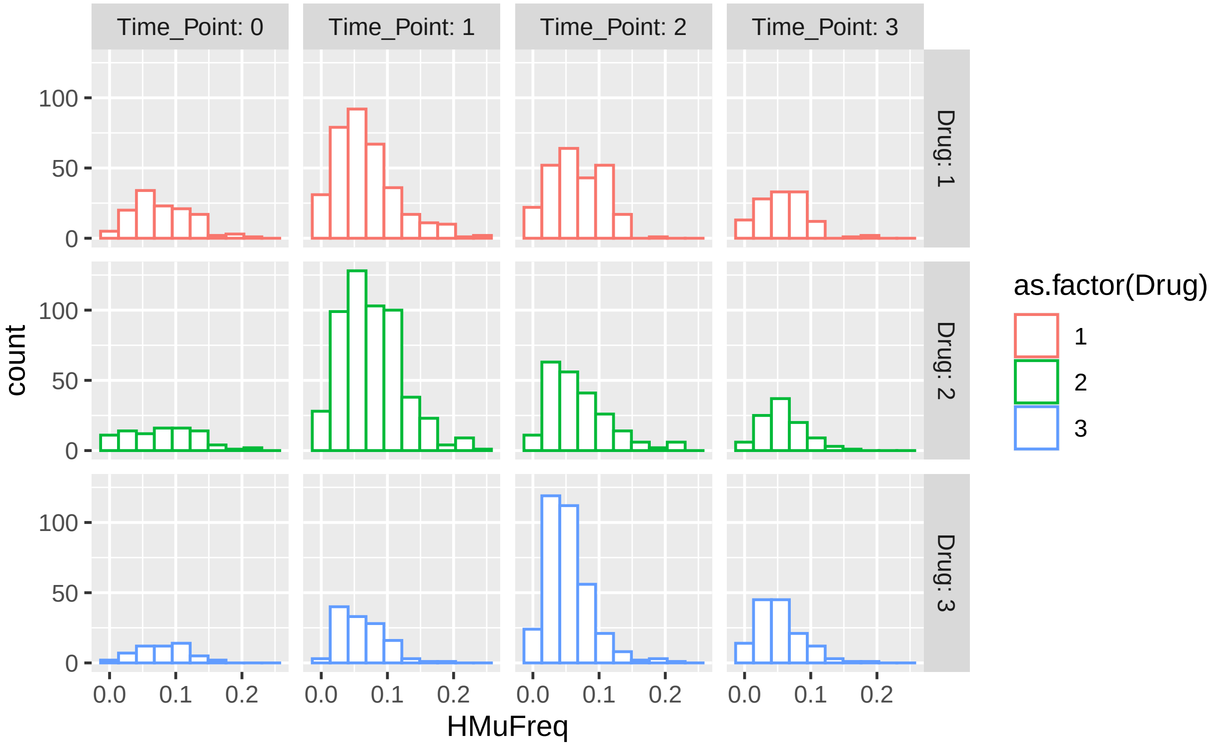

Figure 11: Histograms of HMuFreq vs Treatment and Timepoint

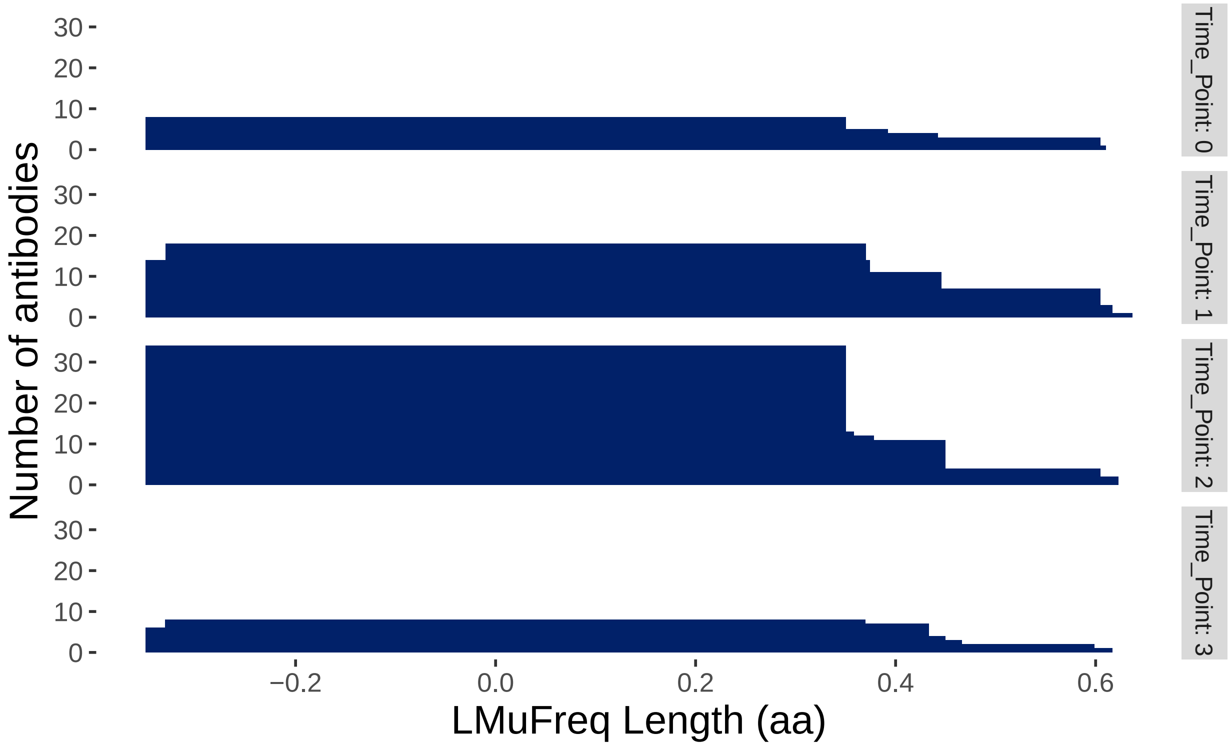

Figure 12: Histogram LMuFreq

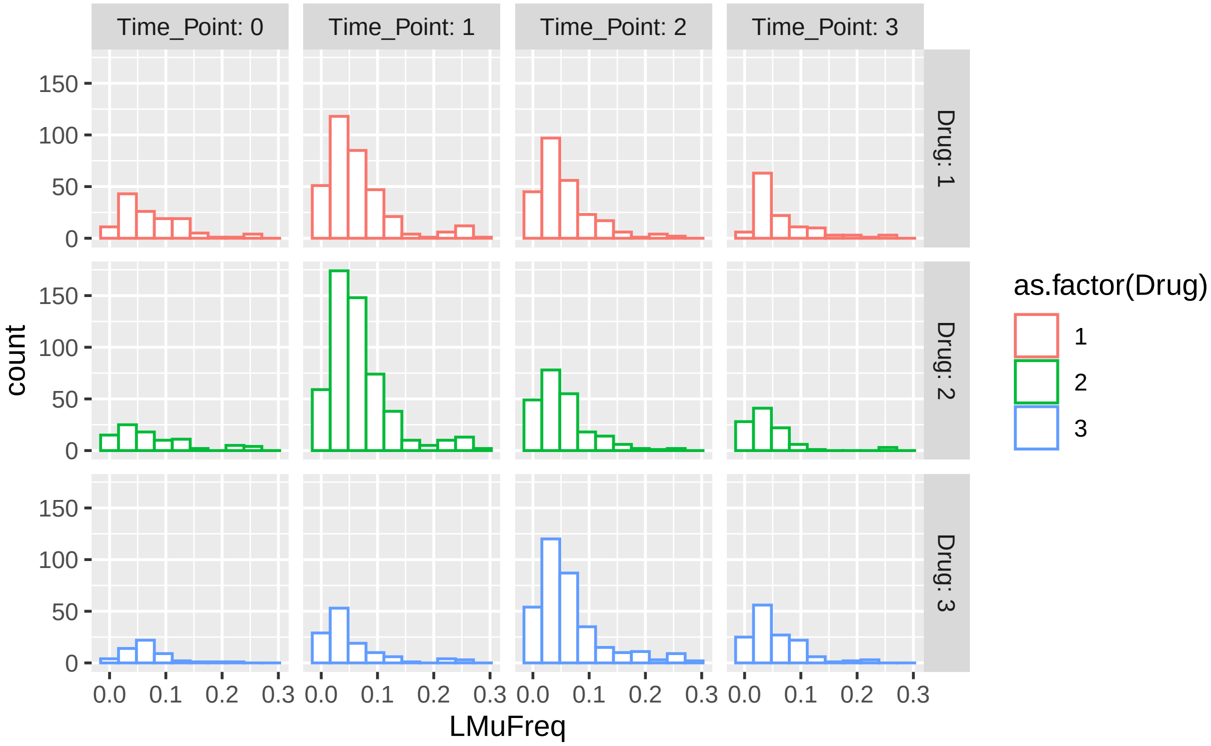

Figure 13: Histograms of LMuFreq vs Treatment and Timepoint

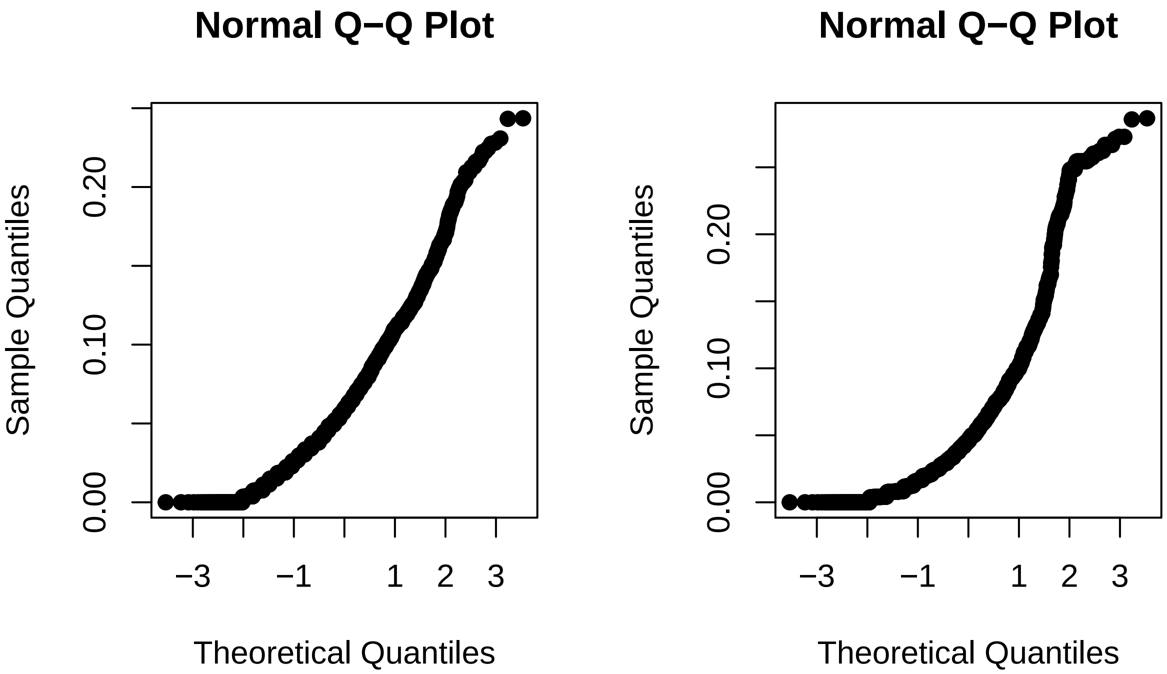

Figure 14: Q-Q Plot of HMuFreq and LMuFreq

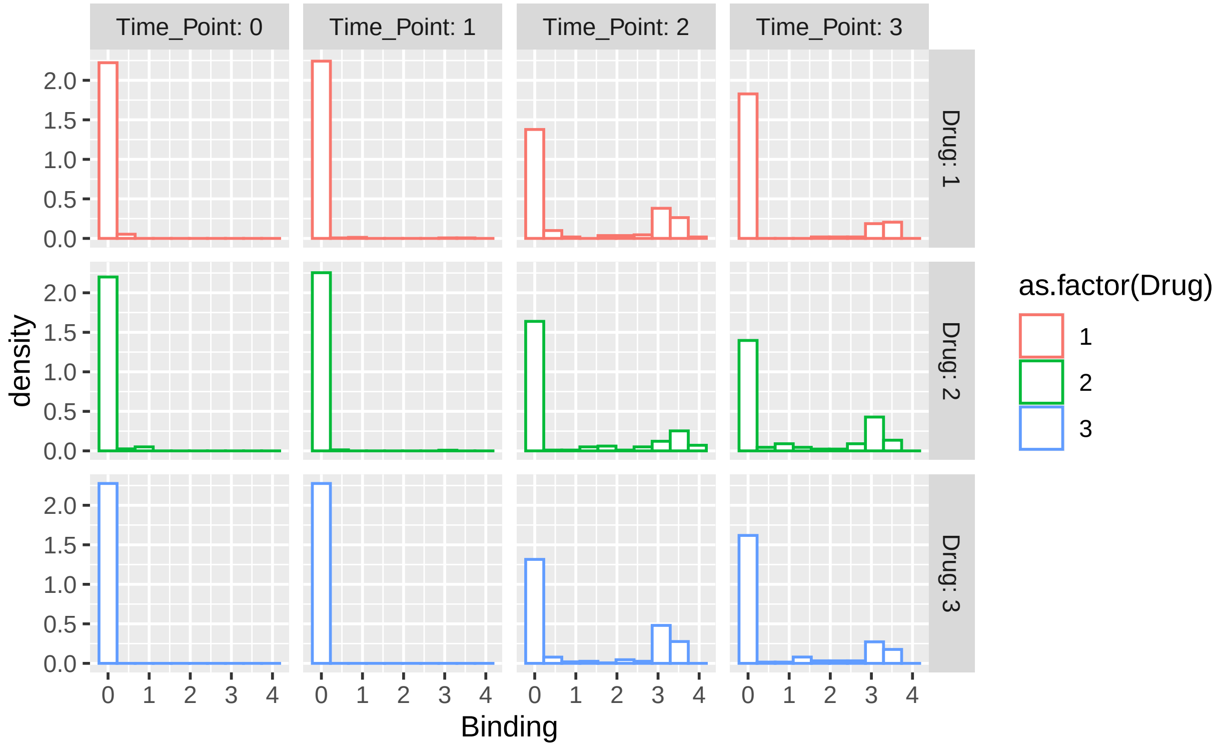

Figure 15: Histograms of Binding Strength vs Treatment and Timepoint

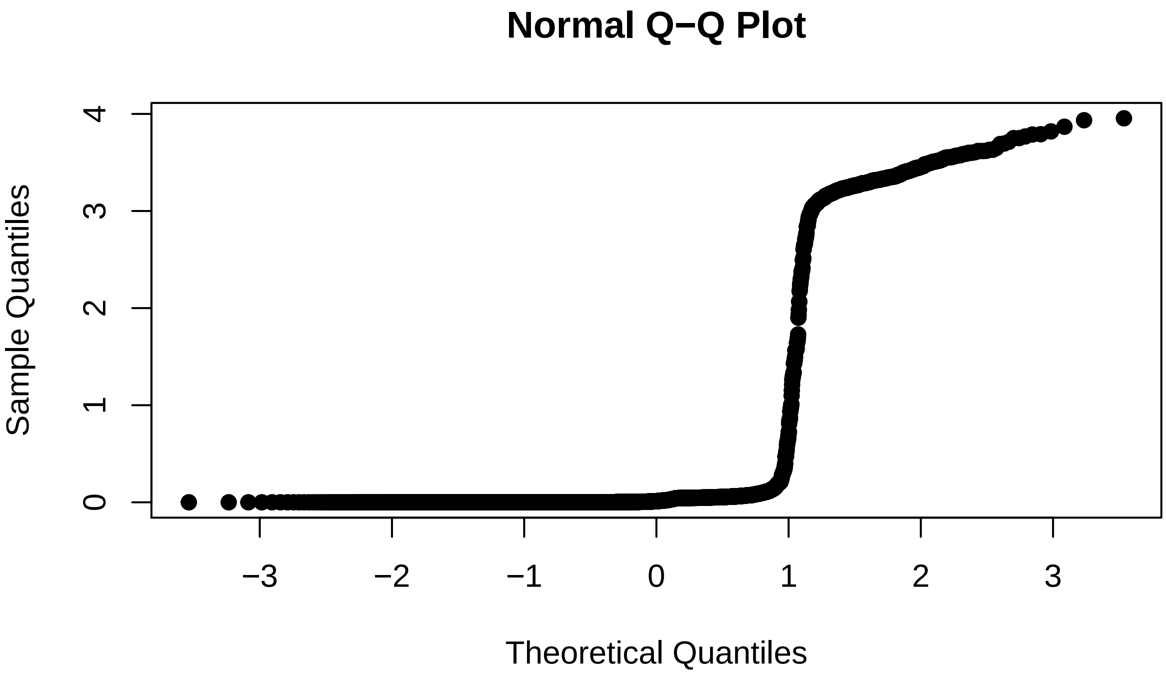

Figure 16: Q-Q Plot of Binding

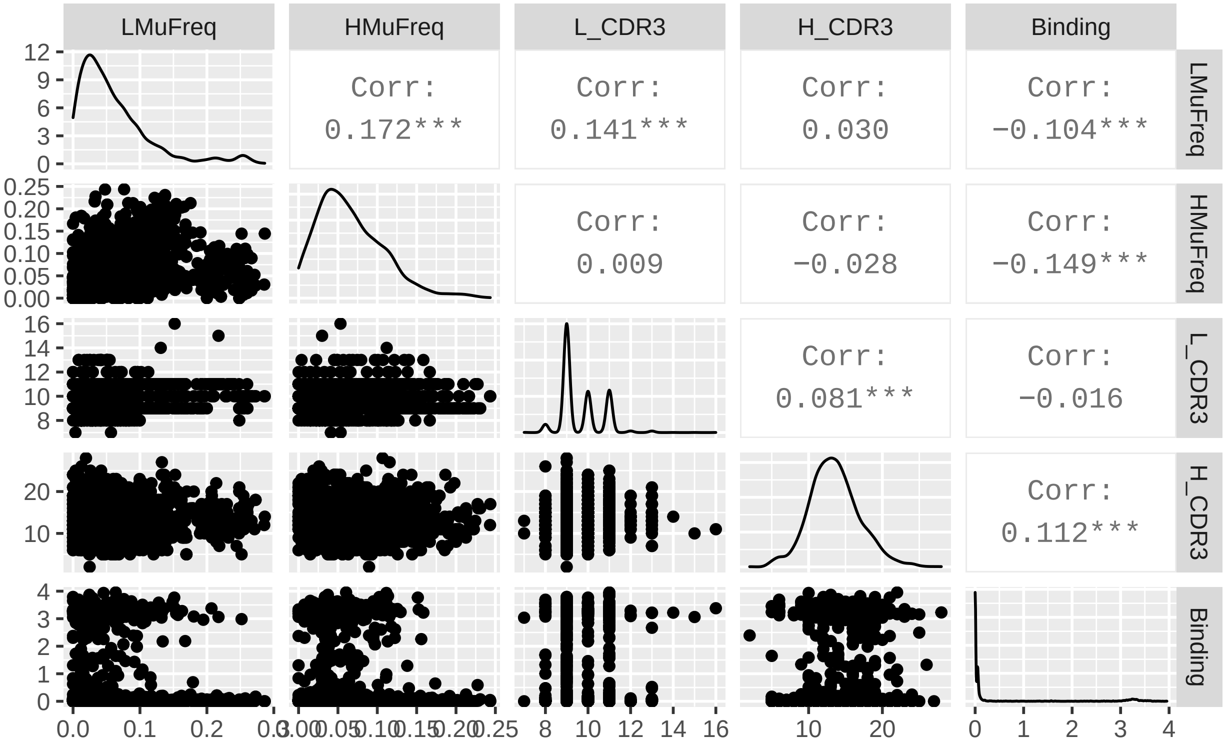

Figure 17: Plots of response variables

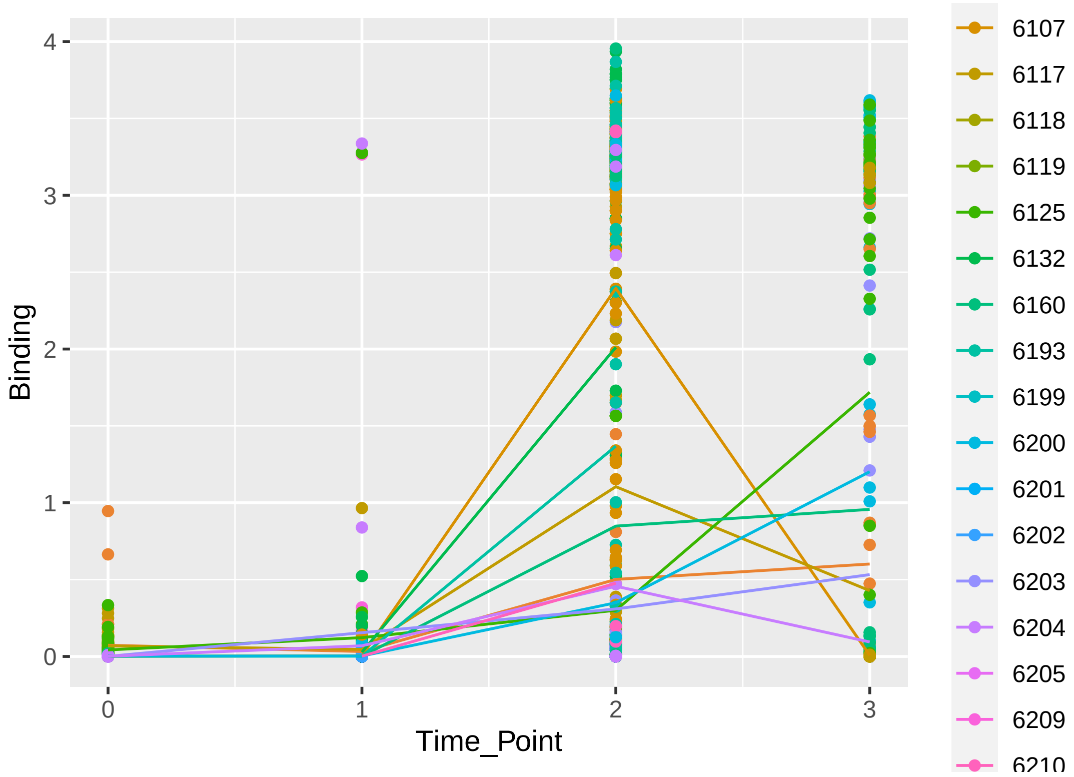

Figure 18: Mean trend by monkey

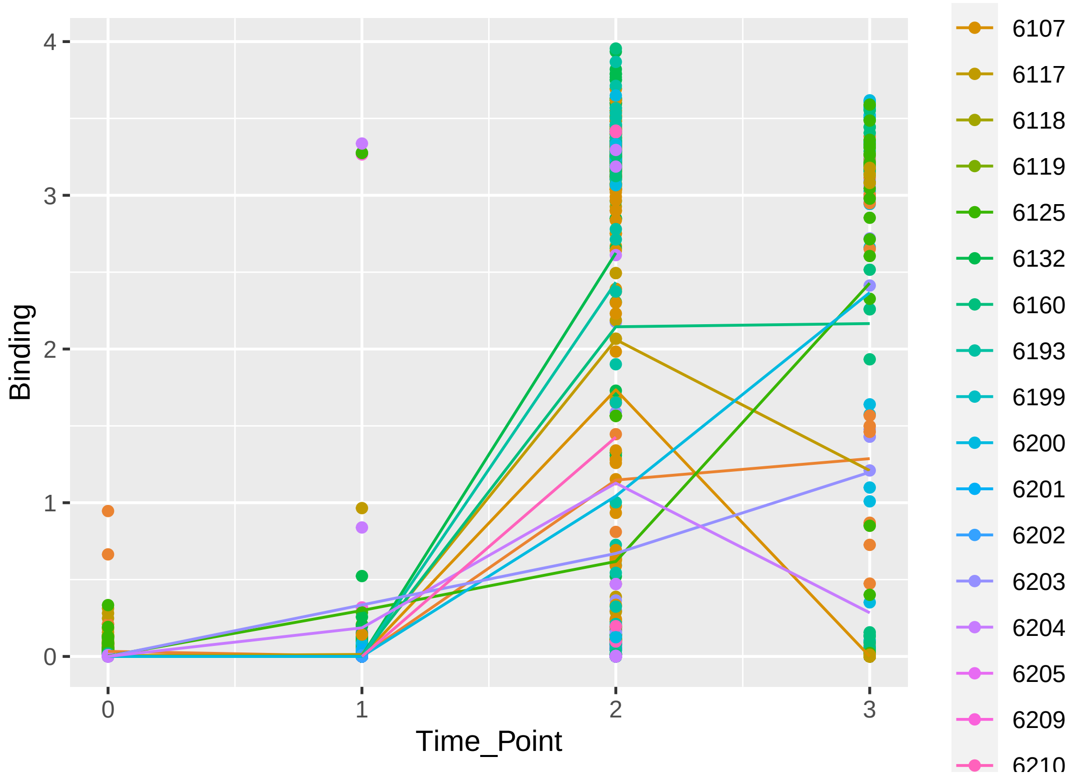

Figure 19: Variances over time by monkey

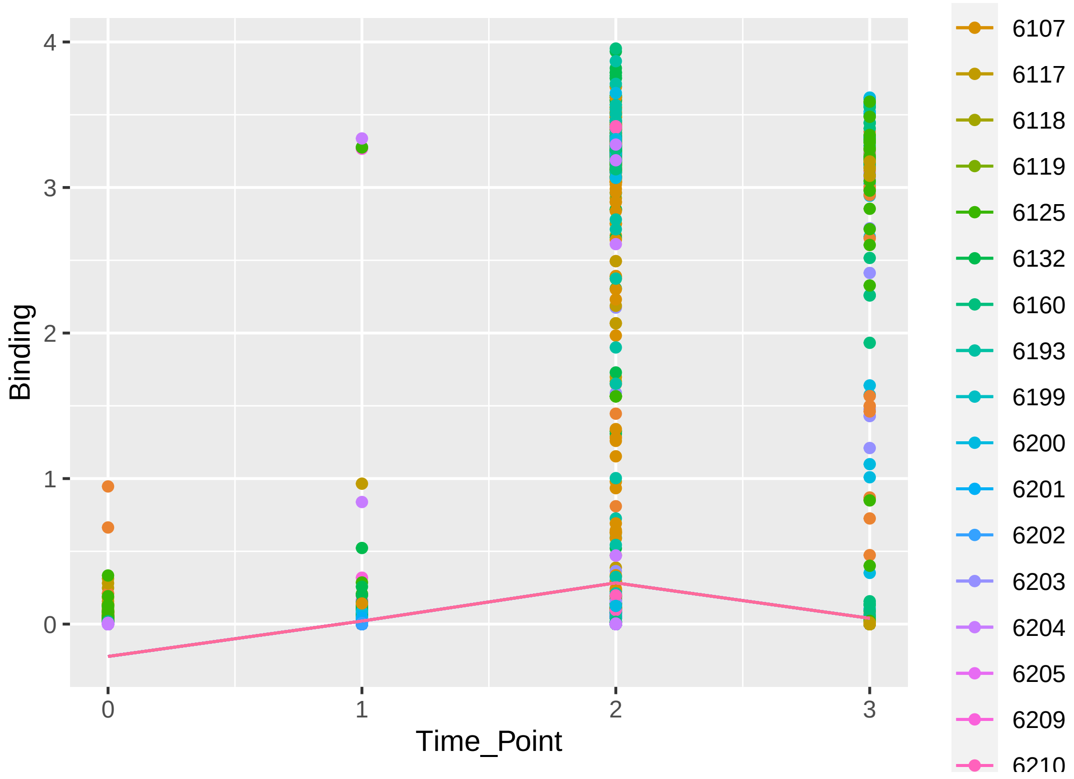

Figure 20: Piecewise Linear Function–three segments

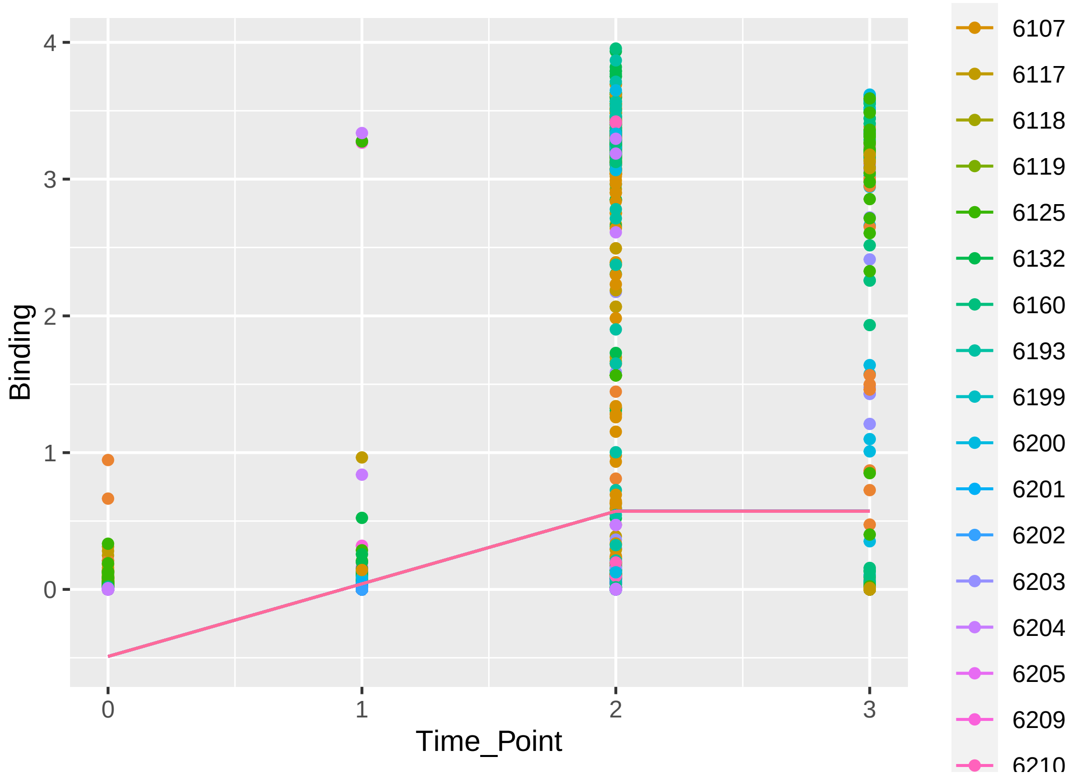

Figure 21: Piecewise Linear Function–two segments

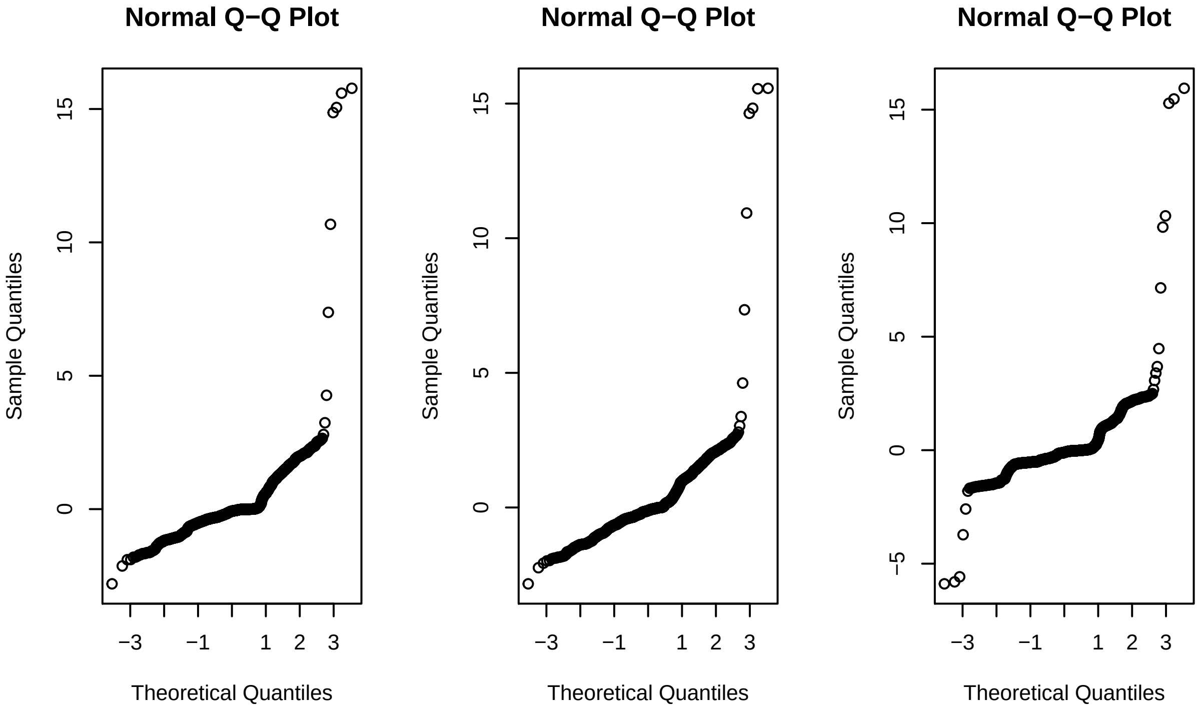

Figure 22: Q-Q plots of models: GLS, compound symmetry, AR1

List of Tables

Table 1: Frequency tables of drug vs. timepoints

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

Table 2: Frequency tables of timepoints vs. isotypes for drug = 1 (left), 2 (middle), 3 (right)

	A	D	E	G	M
0	4	4	1	60	60
1	11	22	2	91	220
2	8	15	3	145	80
3	1	16	1	57	47

	A	D	E	G	M
0	6	4	0	37	43
1	10	45	2	205	271
2	4	19	1	115	86
3	1	7	0	53	40

	A	D	E	G	M
0	1	1	0	24	28
1	1	26	0	28	70
2	4	14	0	170	159
3	0	6	0	77	59

Table 3: Summaries of standardized LCDR3

	V1
	Min. :-2.1860
	1st Qu.:-0.5361
	Median :-0.5361
	Mean : 0.0000
	3rd Qu.: 0.2888
	Max. :30.8110

Table 4: AIC and BIC between two gls models

	df	AIC	df.1	BIC
fit.gls	9	3323.050	9	3375.322
fit.gls2	8	3315.264	8	3361.730

Table 5: AIC and BIC for three models

	df	AIC	df.1	BIC
fit.gls2	8	3315.264	8	3361.730
fit.a1	11	3234.628	11	3298.520
fit.a2	11	3063.290	11	3127.182

Table 6: Inference of S4 ad S5 slopes

numDF	denDF	F.value	p.value
1	2442	244.2324506	0.0000000
1	2442	0.0317192	0.8586602

Table 7: Test whether drug 1 = drug 2

Fstat	p_value
1.065151	0.3626666

Table 8: Test whether drug 1 = drug 3

Fstat	p_value
1.231968	0.2964737

Table 9: Test whether drug 2 = drug 3

Fstat	p_value
1.255448	0.288075