

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

11/17/2020

Contents

1	Introduction	2
1.1	About the Study	2
1.2	List of variables	3
1.3	Research questions	4
2	Methods	5
2.1	Data Summaries	5
2.1.1	An Overview of Antibodies by Time Points, Drug Types, and Isotypes . .	5
2.1.2	Outlier Detection	5
2.1.3	Response Variables	6
2.2	Multivariate Data Analysis	6
2.2.1	MANOVA	7
2.2.2	Pairwise Comparison	7
2.3	Longitudinal Analysis	13
2.3.1	One Covariate: Time Point	13
2.3.2	Adding Random Effects	15
2.3.3	Inference about β	16
2.3.4	Two Covariates: Time Point and Drug	17
3	Results	19
4	Discussion	20
4.1	Implications	20
4.2	Limitations	20
5	Conclusions	21
6	References	21

1 Introduction

The current project uses multivariate and longitudinal data analyses on a dataset from an HIV study that gave HIV vaccine injections and immuno-suppression treatments to 20 rhesus macaques. This report's main goal is to show whether the vaccine's efficacy, reflected by the variable Binding, can be increased by the number of vaccine injections and the type of 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 experiment used T-cell suppression treatments, widely used in post transplant immuno-suppression treatment to prevent rejection, to investigate the effect on 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). The resulting dataset includes measurements of antibodies after the 20 macaques were given the same HIV vaccine at three different time points and one of three randomly selected anti-Treg treatments. 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 germinal centers. 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 follow the convention of vaccine studies and treat each antibody as independent for multivariate data analysis. For Longitudinal data analysis, since the models allow measurements from the same observational unit to correlate over time, we will try different correlation structures to find a best model. This essentially implies that the antibodies from the same macaque can be correlated.

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 the chain from which the information comes.

1.2 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. The 7 treatments in the dataset are coded into 3 drugs to simplify the data.
- **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**: 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 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 the heavy 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 the 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.
- Binding: affinity of antibodies against a selected HIV glycoprotein. Larger values indicate 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.

The dataset, which can be found here, was provided by Kan Luo, as he was one of authors for four publications that used the dataset. Now we will turn to our research questions. 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.

1.3 Research questions

The main focus of the current project is to understand whether the number of vaccine injections (Time_Point) and the different Treg inhibitor treatments (Drug) cause changes in the antibody characteristics and if the changes are related to the immune responses against HIV virus. The category of antibody (Isotype) was not controlled by the researchers, but it could be related to the response variables and is thus included as a predictor. We have two research questions:

RQ1: Do time points, drugs, and isotypes 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 does the binding strength of the antibodies (Binding) 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. It then uses multivariate and longitudinal data analyses to address 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 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 isotypes.

Figure 2 shows the histograms of Isotype, in which we observe 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 project include five variables: H_CDR3, HMuFreq, L_CDR3, LMuFreq, and Binding. As shown in Figure 3, L_CDR3 in one point seems an outlier. The summary statistics of standardized L_CDR3 in Table 3 show that the maximum value is greater than 30, which is quite unusual. Figure 4 shows the Mahalanobis distances and Z scores of L_CDR3, and the data point again appears to be an apparent outlier. The value for L_CDR3, 47, is quite unlikely. Since we are not able to go back to the original data, we remove the data point and will use the new dataset Data3.

2.1.3 Response Variables

Here we examine our responses: H_CDR3, HMuFreq, L_CDR3, LMuFreq, and Binding. We observe that for H_CDR3 the distributions are approximately 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 are calculated by dividing H_Substitution by H_VBase for heavy chain and L_Substitution by L_VBase light chain. These two variables show the degree to which 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 normal, each with a long 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 is 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 are highly correlated.

2.2 Multivariate Data Analysis

To answer **RQ1** (Do time points, drugs, and isotypes 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 test whether predictors Time_Point, Drug, and Isotype had effects on four of our five response variables: H_CDR3, HMuFreq, L_CDR3, and LMuFreq. We exclude Binding from this section, because it has unequal variances across time points, which violates the equal variance assumption of MANOVA. We will use it as the main response variable in longitudinal analyses, since longitudinal analysis allows the option of unequal variances over time.

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 $(H_CDR3, HMuFreq, L_CDR3, LMuFreq)^T \sim \text{Time_Point} + \text{Drug} + \text{Iso-type}$. The null hypothesis is that the means of the different populations (of Time_Point, Drug, and Isotype) are different.

```
##              Df    Wilks approx F num Df den Df      Pr(>F)
## tp              3 0.95189   10.1542      12 6477.1 < 2.2e-16 ***
## drug            2 0.97565    7.5907       8 4896.0 3.909e-10 ***
## it              4 0.89584   17.1370      16 7479.4 < 2.2e-16 ***
## Residuals 2451
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output above shows that all of the main effects of Time_Point, Drug, Isotype have very small p-values. Thus we reject the null hypothesis and conclude that all three main effects are significant to the four traits, H_CDR3, HMuFreq, L_CDR3, and LMuFreq, of antibodies. To understand how exactly these three factors have effects, we proceed to do pairwise comparison.

2.2.2 Pairwise Comparison

```
## [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

## [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: 0.9958333333333333
## [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: 0.9958333333333333
## [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: 0.9958333333333333
## [1] "H_CDR3 pairwise CI's"
## contrast estimate      SE   df lower.CL upper.CL
## A - D      0.297 0.530 2456  -1.414    2.009
## A - E     -2.229 1.154 2456  -5.959    1.500
## A - G     -0.728 0.479 2456  -2.274    0.818
## A - M      0.503 0.478 2456  -1.040    2.046
## D - E     -2.527 1.085 2456  -6.031    0.977
## D - G     -1.025 0.270 2456  -1.897   -0.154
## D - M      0.206 0.268 2456  -0.660    1.072
## E - G      1.502 1.060 2456  -1.925    4.928
## E - M      2.733 1.060 2456  -0.692    6.158

```

```

## G - M      1.231 0.142 2456      0.773      1.689
##
## Results are averaged over the levels of: rep.meas
## Note: contrasts are still on the [.: scale
## Confidence level used: 0.99875
## [1] "HMuFreq pairwise CI's"
## contrast estimate      SE    df lower.CL upper.CL
## A - D      0.00523 0.00659 2456  -0.0161  0.02651
## A - E      0.00848 0.01435 2456  -0.0379  0.05485
## A - G     -0.01628 0.00595 2456  -0.0355  0.00294
## A - M      0.00393 0.00594 2456  -0.0153  0.02311
## D - E      0.00325 0.01349 2456  -0.0403  0.04682
## D - G     -0.02151 0.00335 2456  -0.0323 -0.01067
## D - M     -0.00130 0.00333 2456  -0.0121  0.00947
## E - G     -0.02476 0.01319 2456  -0.0674  0.01784
## E - M     -0.00455 0.01318 2456  -0.0471  0.03804
## G - M      0.02021 0.00176 2456   0.0145  0.02591
##
## Results are averaged over the levels of: rep.meas
## Note: contrasts are still on the [.: scale
## Confidence level used: 0.99875
## [1] "L_CDR3 pairwise CI's"
## contrast estimate      SE    df lower.CL upper.CL
## A - D      0.1136 0.1510 2456  -0.3742  0.601
## A - E      0.0667 0.3289 2456  -0.9961  1.129
## A - G      0.0041 0.1364 2456  -0.4365  0.445
## A - M      0.0441 0.1361 2456  -0.3955  0.484
## D - E     -0.0469 0.3091 2456  -1.0455  0.952
## D - G     -0.1095 0.0769 2456  -0.3579  0.139
## D - M     -0.0695 0.0764 2456  -0.3162  0.177

```

```

## E - G      -0.0626 0.3022 2456  -1.0389    0.914
## E - M      -0.0225 0.3021 2456  -0.9985    0.953
## G - M       0.0400 0.0404 2456  -0.0905    0.171
##
## Results are averaged over the levels of: rep.meas
## Note: contrasts are still on the [.: scale
## Confidence level used: 0.99875
## [1] "LMuFreq pairwise CI's"
## contrast estimate      SE   df lower.CL upper.CL
## A - D      -0.021206 0.00858 2456 -0.04892  0.00651
## A - E      -0.006829 0.01869 2456 -0.06722  0.05356
## A - G      -0.023476 0.00775 2456 -0.04851  0.00156
## A - M      -0.021669 0.00773 2456 -0.04665  0.00331
## D - E       0.014376 0.01756 2456 -0.04236  0.07111
## D - G      -0.002270 0.00437 2456 -0.01638  0.01184
## D - M      -0.000463 0.00434 2456 -0.01448  0.01356
## E - G      -0.016647 0.01717 2456 -0.07212  0.03883
## E - M      -0.014839 0.01716 2456 -0.07029  0.04061
## G - M       0.001807 0.00230 2456 -0.00561  0.00923
##
## Results are averaged over the levels of: rep.meas
## Note: contrasts are still on the [.: scale
## Confidence level used: 0.99875

```

Table 4 summarizes The pairwise comparison results. 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 and Isotype have significant pairs in H_CDR3 and HMuFreq. For H_CDR3, drug 1 has a higher mean than drug 2, and IgG has a higher mean than IgD and IgM. For HMuFreq, drug 2 has a higher mean than drug 1, which has a higher mean than drug 3 (control group). IgG again has a higher mean than IgD and IgM. [Kan, please add/revise the interpretations here.]

2.3 Longitudinal Analysis

To answer our **RQ2** (How does the binding strength of the antibodies (Binding) develop in response to the number of vaccine dosages (Time_Point) and immuno-suppression treatments (Drug)?), 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

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

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

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

The new model is

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

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

$$Y_{ij} = S1(\beta_0) + S1Time_{ij}(\beta_1) + S2(2\beta_0 + 2\beta_1 - \beta_4 - 2\beta_5) + S2Time_{ij}(-\beta_0 - \beta_1 + \beta_4 + 2\beta_5) + S3(\beta_4) + S3Time_{ij}(\beta_5) + e_{ij}$$

We run the model and plug in coefficients to 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 line segments 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

$$Y_{ij} = S4(\beta_0 + \beta_1Time_{ij}) + S5(\beta_2 + \beta_3Time_{ij}) + e_{ij}$$

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

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

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

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

After the model is constructed, we use the coefficients to find the mean trends 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 continuous at Time_Point 2. A comparison of AIC And BIC of these two models, shown in Table 5, indicates that the second model (`fit.gls2`) is indeed a better model.

2.3.2 Adding Random Effects

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:

$$Y_{ij} = \beta_1(-2S4 + S4Time_{ij}) + \beta_2(S4 + S5) + \beta_3(2S4 + S5Time_{ij}) + b_{0i} + b_{1i}Time_{ij} + e_{ij}$$

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 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.

2.3.3 Inference about β

Now we would like to know if the slopes between Time_Point 0 and 2 and between Time_Point 2 and 3 equal zero.

$$H_0 : \text{slope of } S4 = 0 \text{ and slope of } S5 = 0$$

which means

$$H_0 : \beta_1 = 0 \text{ and } \beta_3 = 0$$

Thus, we can check for two tests:

$$\mathbf{L}_1 \beta = 0$$

where $\mathbf{L}_1 = (1, 0, 0)$ and $\beta = (\beta_1, \beta_2, \beta_3)^T$ and

$$\mathbf{L}_2 \beta = 0$$

where $\mathbf{L}_2 = (0, 0, 1)$ and $\beta = (\beta_1, \beta_2, \beta_3)^T$

As shown in Table 7, the slope 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.

2.3.4 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

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

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

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

$$\begin{aligned} Y_{ij} = & \beta_1(-2S4 + S4Time_{ij}) + \beta_2(S4 + S5) + \beta_3(2S4 + S5Time_{ij}) + \\ & \beta_4D2(-2S4 + S4Time_{ij}) + \beta_5D2(S4 + S5) + \beta_6D2(2S4 + S5Time_{ij}) + \\ & \beta_7D3(-2S4 + S4Time_{ij}) + \beta_8D3(S4 + S5) + \beta_9D3(2S4 + S5Time_{ij}) + b_{0i} + b_{1i}Time_{ij} + e_{ij} \end{aligned}$$

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

$$\mathbf{L}_3\beta = 0$$

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

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

$$\mathbf{L}_4\beta = 0$$

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

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

$$\mathbf{L}_5\beta = 0$$

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

We found that, as shown in Table 8, Drug 1 and Drug 2 do not have significantly different effects on Binding rates. As shown in Table 9, Drug 1 and Drug 3 do not have significantly different

effects on Binding rates. Also, as shown in Table 10, Drug 2 and Drug 3 do not have significantly different effects on Binding rates. In other words, drug groups do not have significant effects on our longitudinal model. Thus `fit.a2` is our best model for longitudinal data analysis.

3 Results

For multivariate analyses, we performed a MANOVA test on the main effects of `Time_Point`, `Drug`, and `Isotype` on the response variable vector $(H_CDR3, HMuFreq, L_CDR3, LMuFreq)^T$. We find that all three main effects have very small p-values. We also performed pairwise comparison to see where the effects were, as shown in Table 4. [need revisions]

For longitudinal analyses, we find that the best linear mixed model is $Y_{ij} = \beta_1(-2S4 + S4Time_{ij}) + \beta_2(S4 + S5) + \beta_3(2S4 + S5Time_{ij}) + b_{0i} + b_{1i}Time_{ij} + e_{ij}$ 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 inferences on β for the two line segments, We reject the hypothesis that the slope of the first line segment (between time points 0 and 2) is zero, but we fail to reject the hypothesis that the slope of the 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 groups 1 and 2, 2 and 3, and 1 and 3), which led to three p-values greater than 0.05. Thus We fail to reject the null hypotheses that the mean trend of either of the two drug groups was equal.

4 Discussion

4.1 Implications

Our findings show that the number of vaccine injections do contribute to higher binding rates, although we do not determine whether the increase can be translated into immunity against HIV. The lack of evidence for effects of Treg inhibitor treatments, on the other hand, suggests that the theory of using immuno-suppressing drugs to enhance the efficacy of HIV vaccines has not been proven.

4.2 Limitations

While our analyses reached some findings, some further investigations could improve our analyses. In our multivariate analyses, we follow the common method of treating antibodies (rows in our data) as independent from 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 perform statistical analyses to choose the most reasonable option, but it would require more biological knowledge and the investigation would most likely be beyond the scope of a final report. This remains an interesting topic that could be explored.

For multivariate data analyses, the use of MANOVA was restricted by the assumptions of equal variance-covariance matrices among different populations. Furthermore, the pairwise comparison results are difficult to interpret. [need revisions]

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. Further, we did not assume different random effects on different line segments. These additional steps might lead to a better model.

5 Conclusions

[need revisions] 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 and enhance the efficacy of HIV vaccines. Our statistical analyses provide answers to our two research questions.

We performed a MANOVA test to our first research question, “Do time points, drugs, and isotypes 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, 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: please revise interpretations here.]

To answer our second research question, “How does the binding strength of the antibodies (Binding) develop in response to the number of vaccine dosages (Time_Point) and immunosuppression treatments (Drug)?”, 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, indicating that the Treg inhibitor treatments have not been proven effective in producing immunity against HIV.

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., Searce, 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

1	Histograms of Antibodies	24
2	Histograms of Isotypes	25
3	Histogram of Response Variables	26
4	Mahalanobis distances and Z scores	27
5	Histogram H_CDR3	28
6	Histograms of H_CDR3 vs Treatment and Timepoint	29
7	Histogram L_CDR3	30
8	Histograms of L_CDR3 vs Treatment and Timepoint	31
9	Q-Q Plots of H_CDR3 and L_CDR3	32
10	Histogram HMuFreq	33
11	Histograms of HMuFreq vs Treatment and Timepoint	34
12	Histogram LMuFreq	35
13	Histograms of LMuFreq vs Treatment and Timepoint	36
14	Q-Q Plot of HMuFreq and LMuFreq	37
15	Histograms of Binding Strength vs Treatment and Timepoint	38
16	Q-Q Plot of Binding	39
17	Plots of response variables	40
18	Mean trend by monkey	41
19	Variances over time by monkey	42
20	Piecewise Linear Function—three segments	43
21	Piecewise Linear Function—two segments	44
22	Q-Q plots of models: GLS, compound symmetry, AR1	45

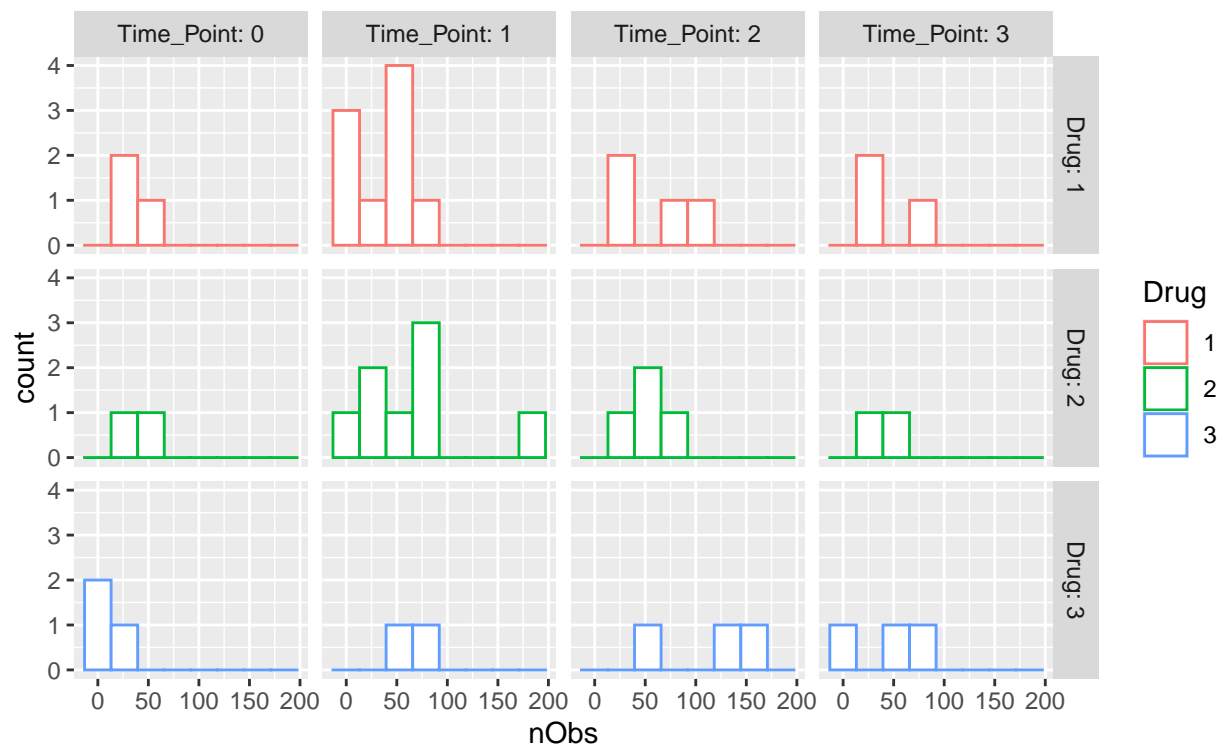


Figure 1: Histograms of Antibodies

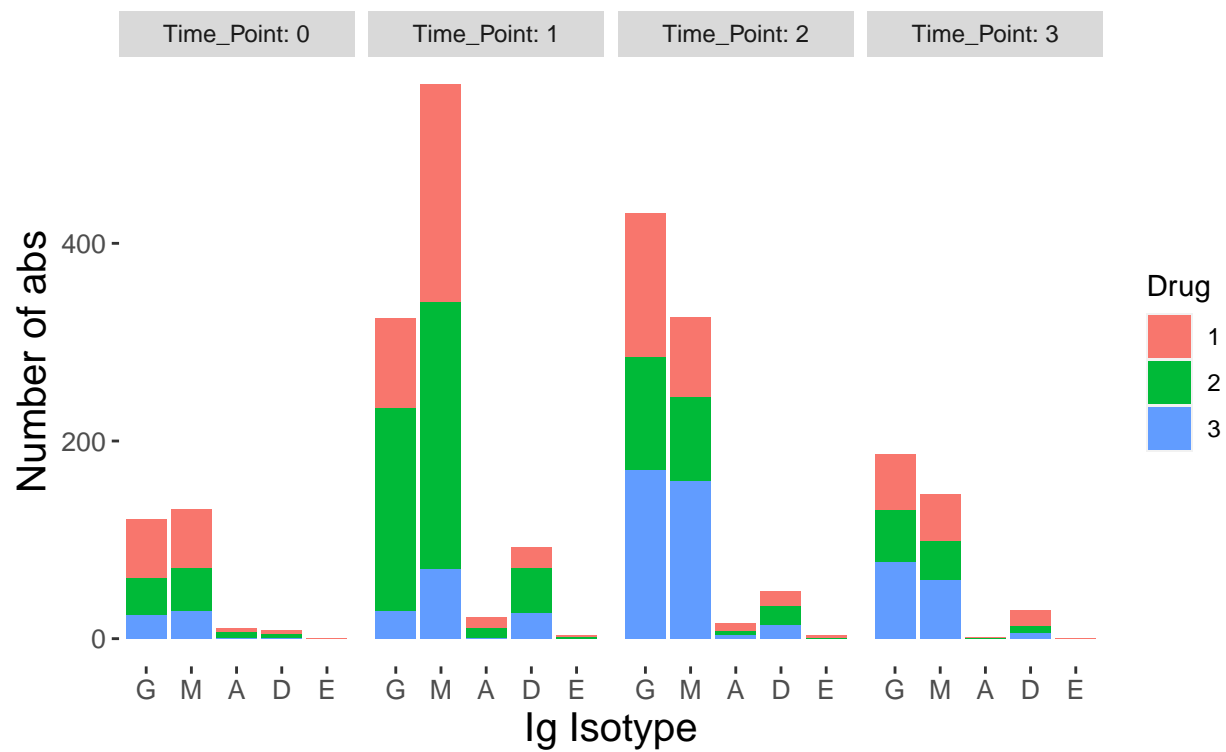


Figure 2: Histograms of Isotypes

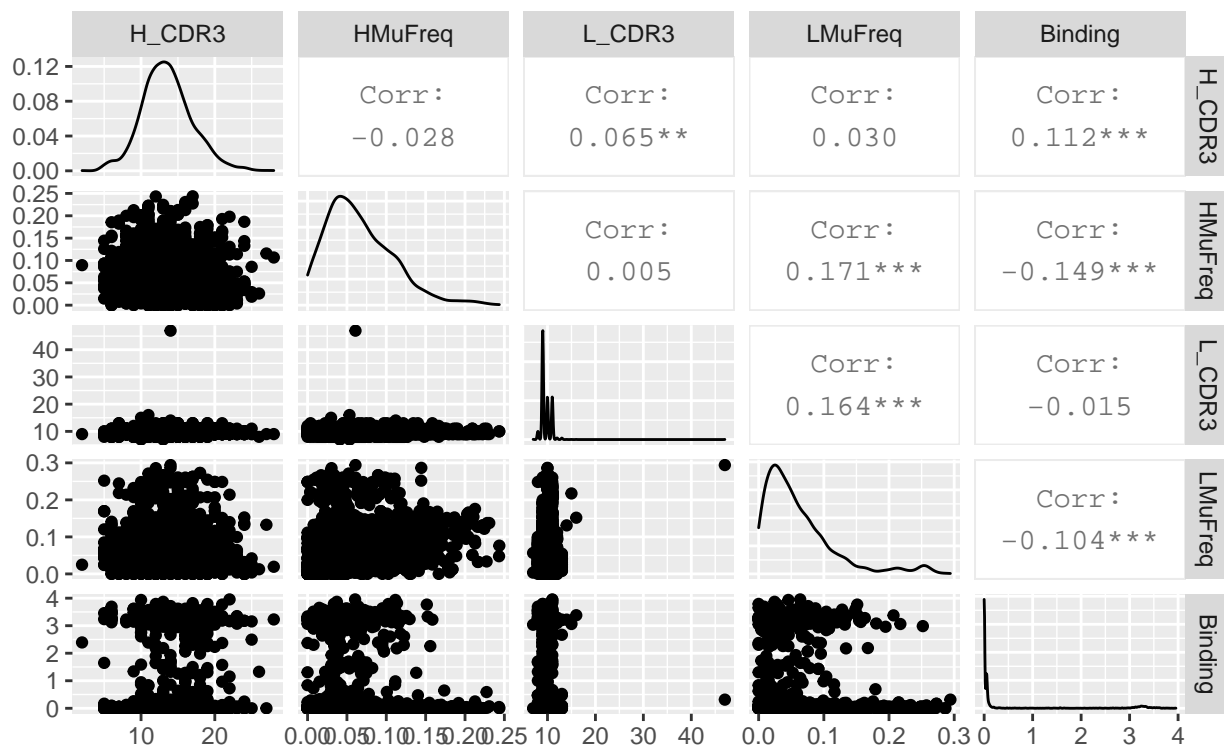


Figure 3: Histogram of Response Variables

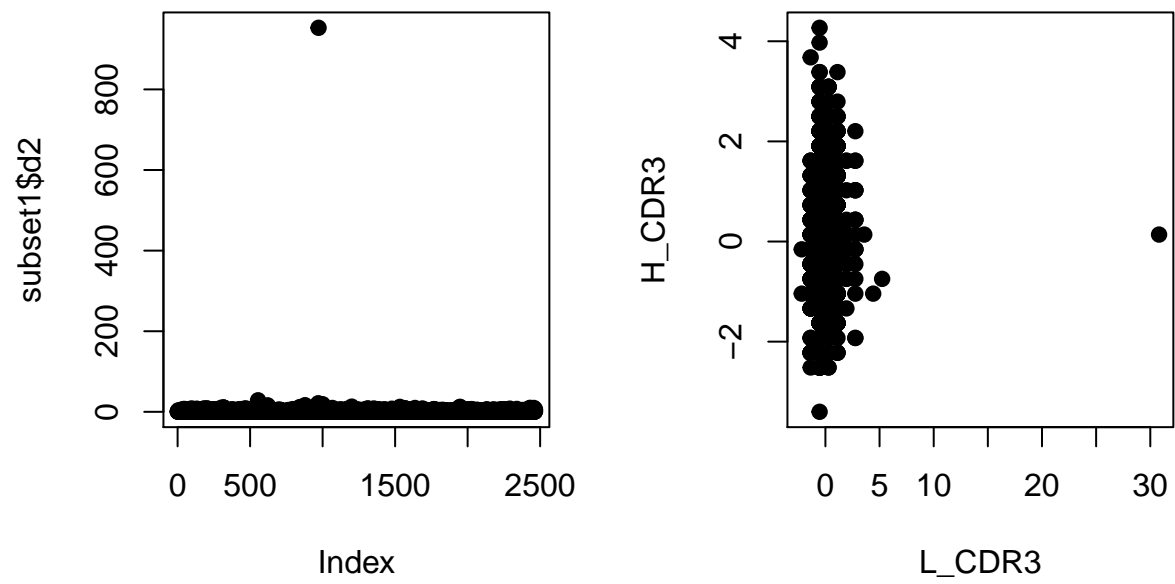


Figure 4: Mahalanobis distances and Z scores

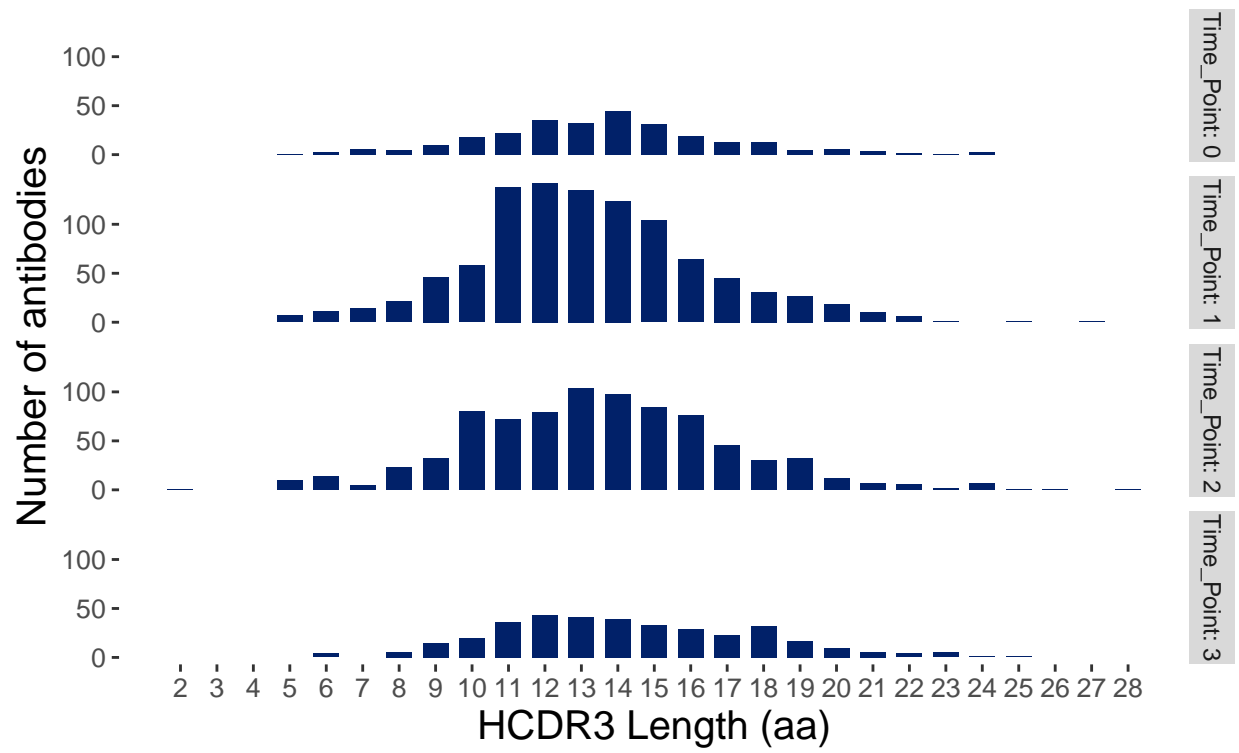


Figure 5: Histogram H_CDR3

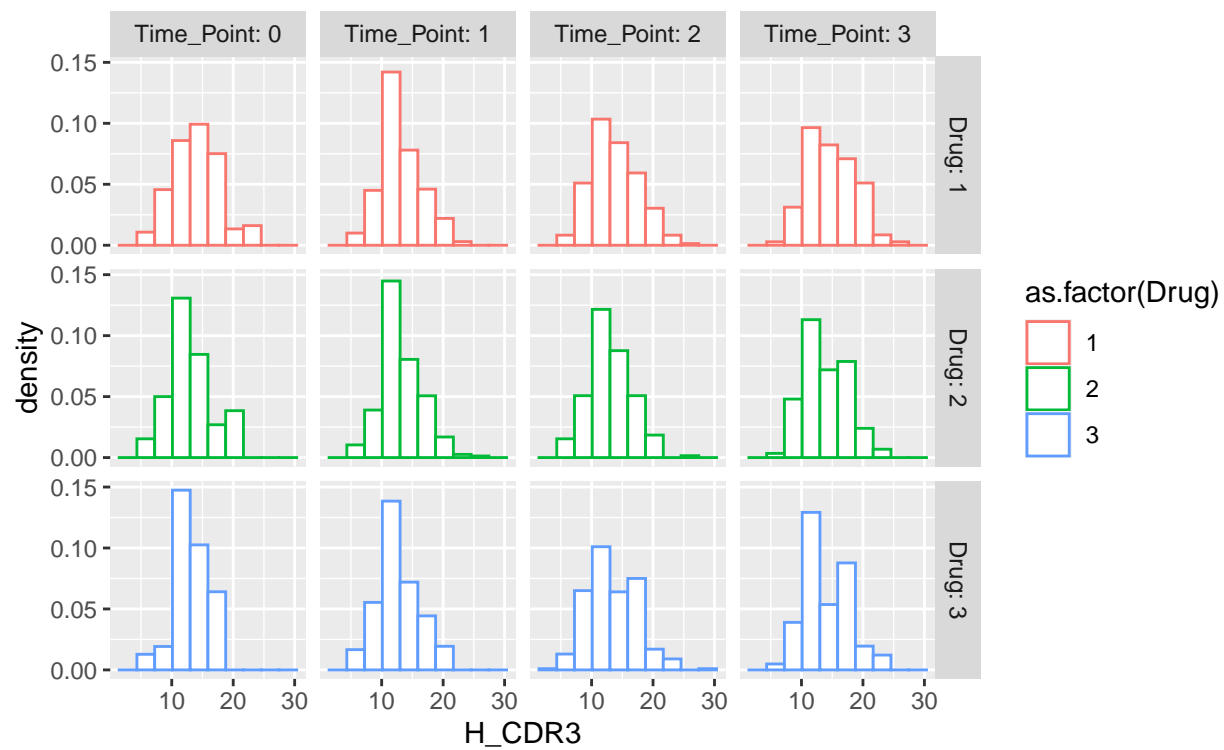


Figure 6: Histograms of H_CDR3 vs Treatment and Timepoint

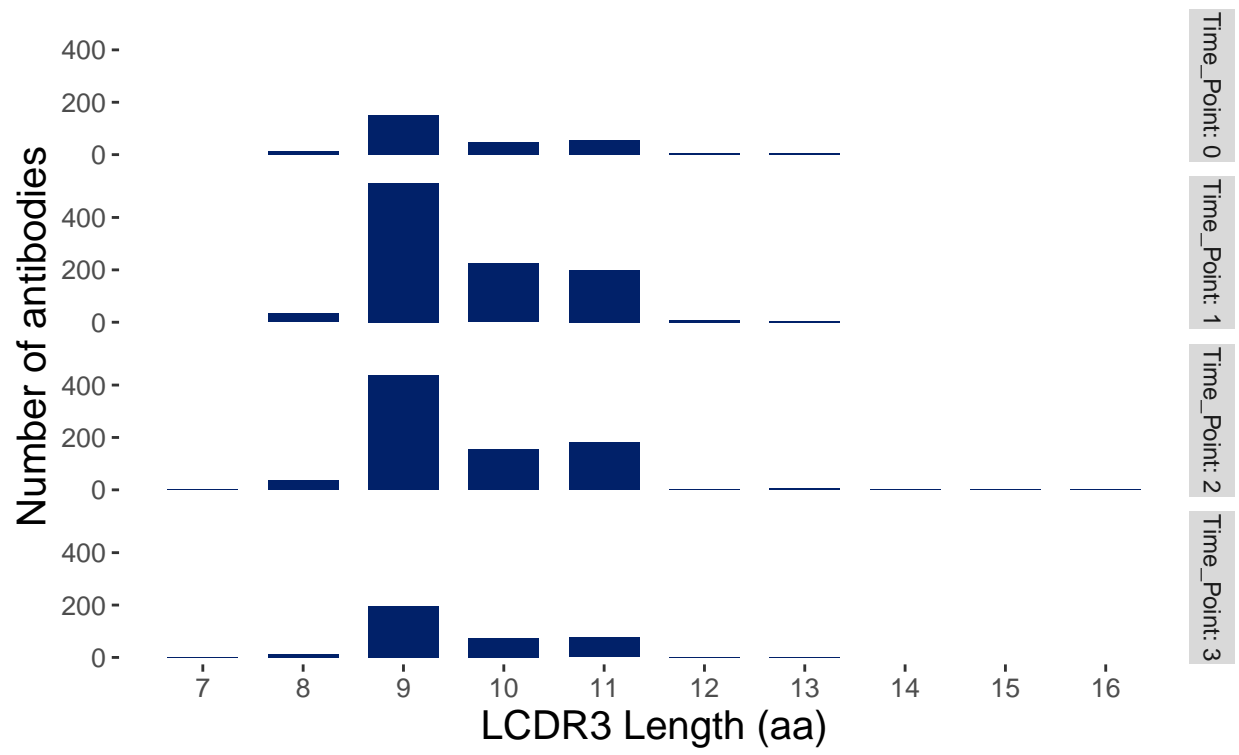


Figure 7: Histogram L_CDR3

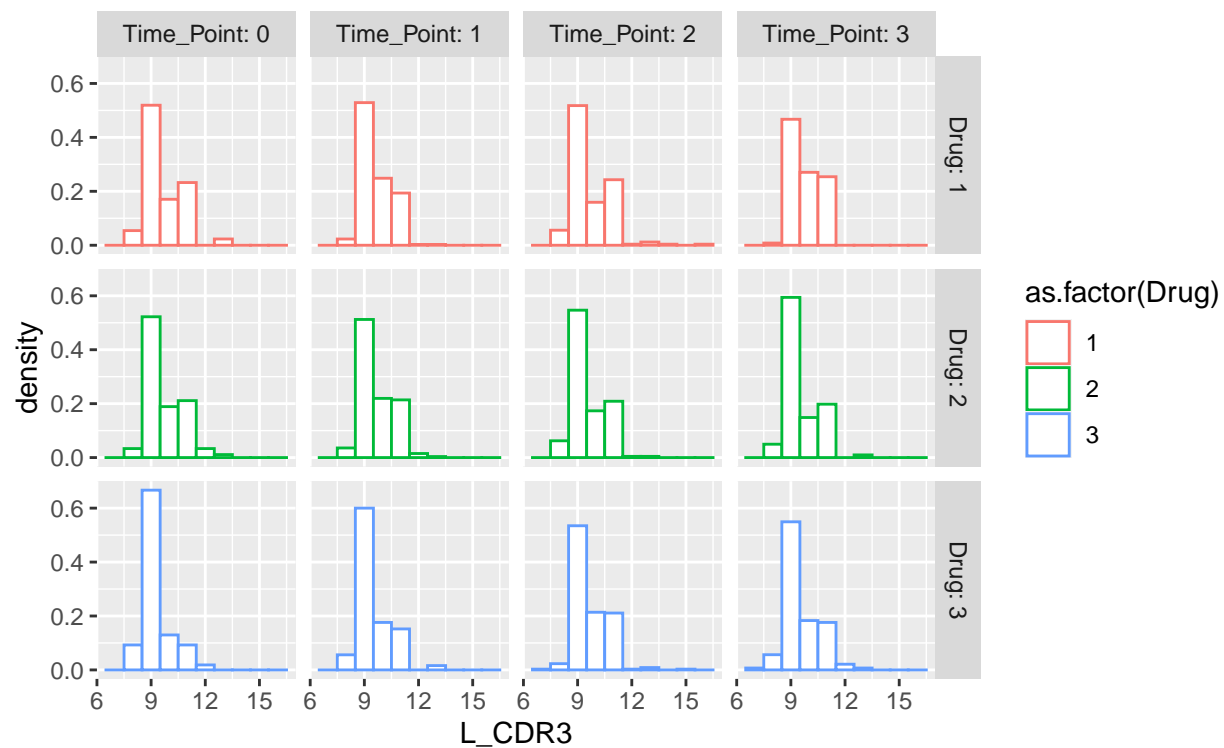


Figure 8: Histograms of L_CDR3 vs Treatment and Timepoint

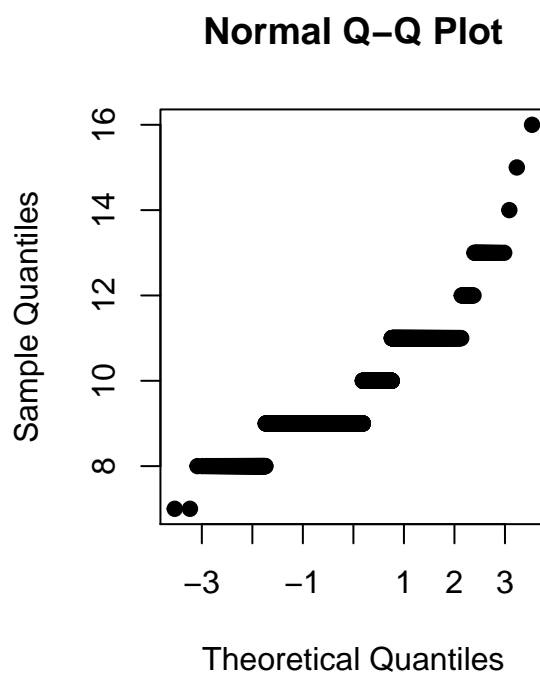
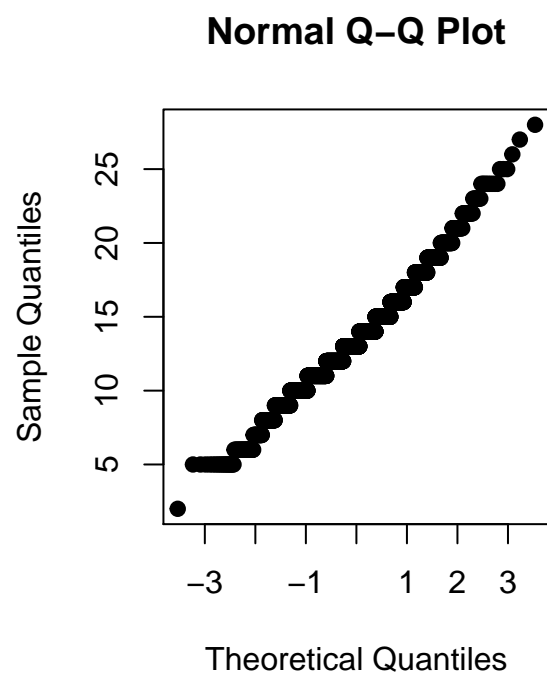


Figure 9: Q-Q Plots of H_CDR3 and L_CDR3

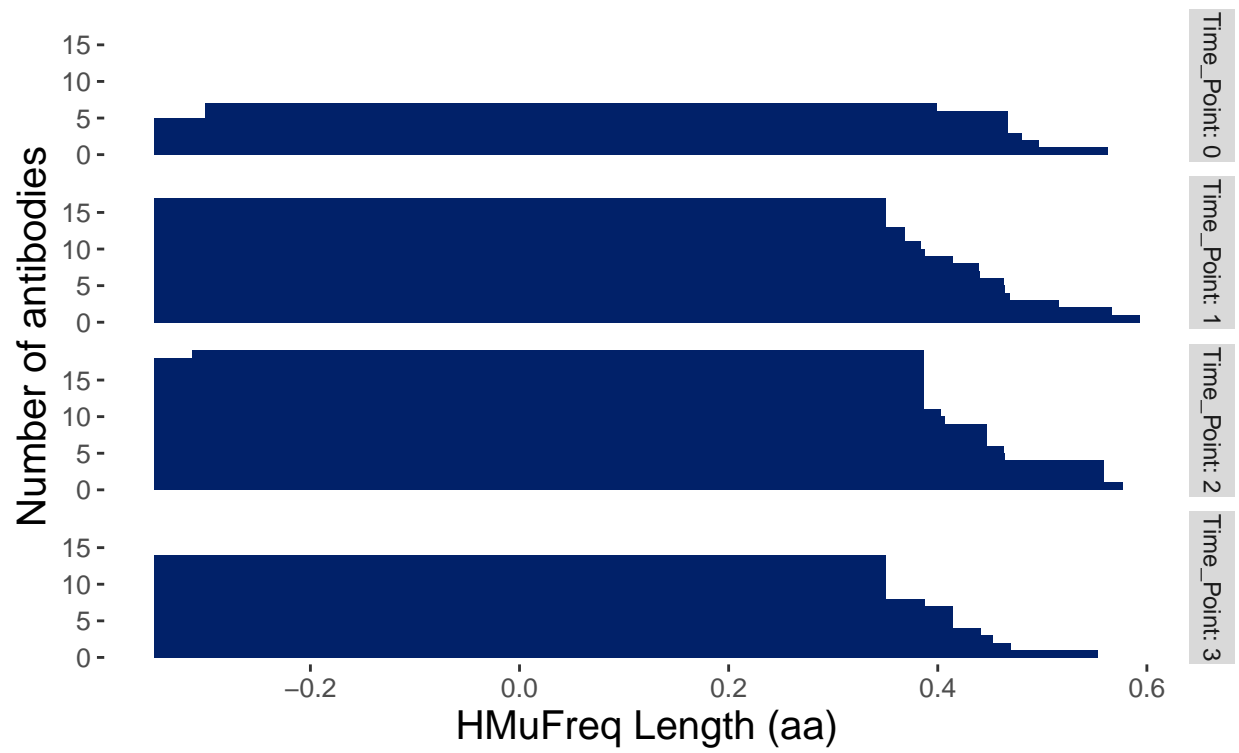


Figure 10: Histogram HMuFreq

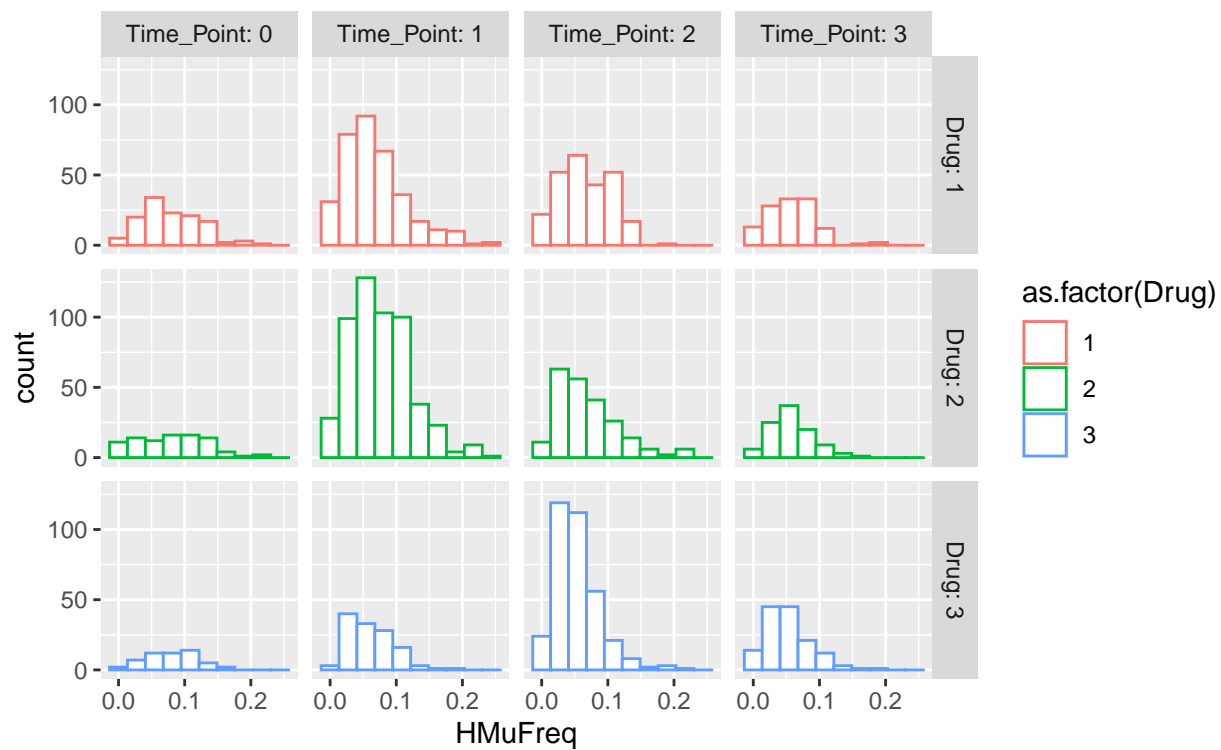


Figure 11: Histograms of HMuFreq vs Treatment and Timepoint

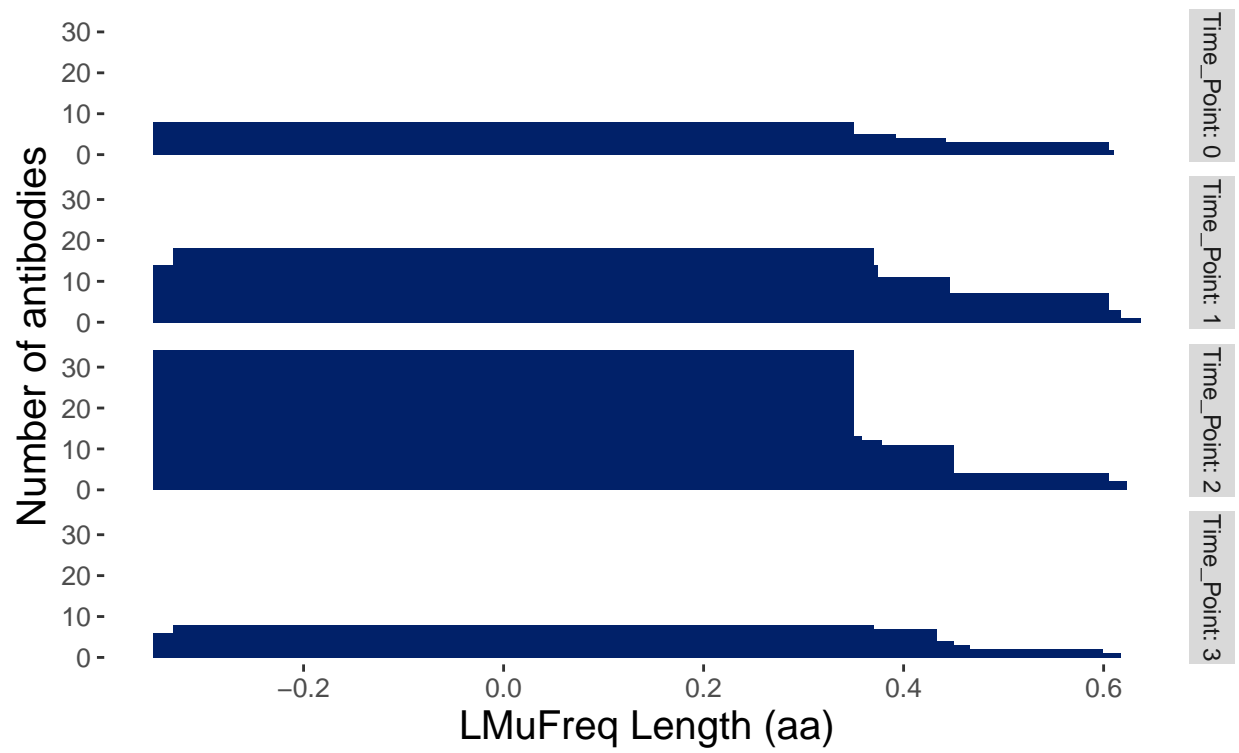


Figure 12: Histogram LMuFreq

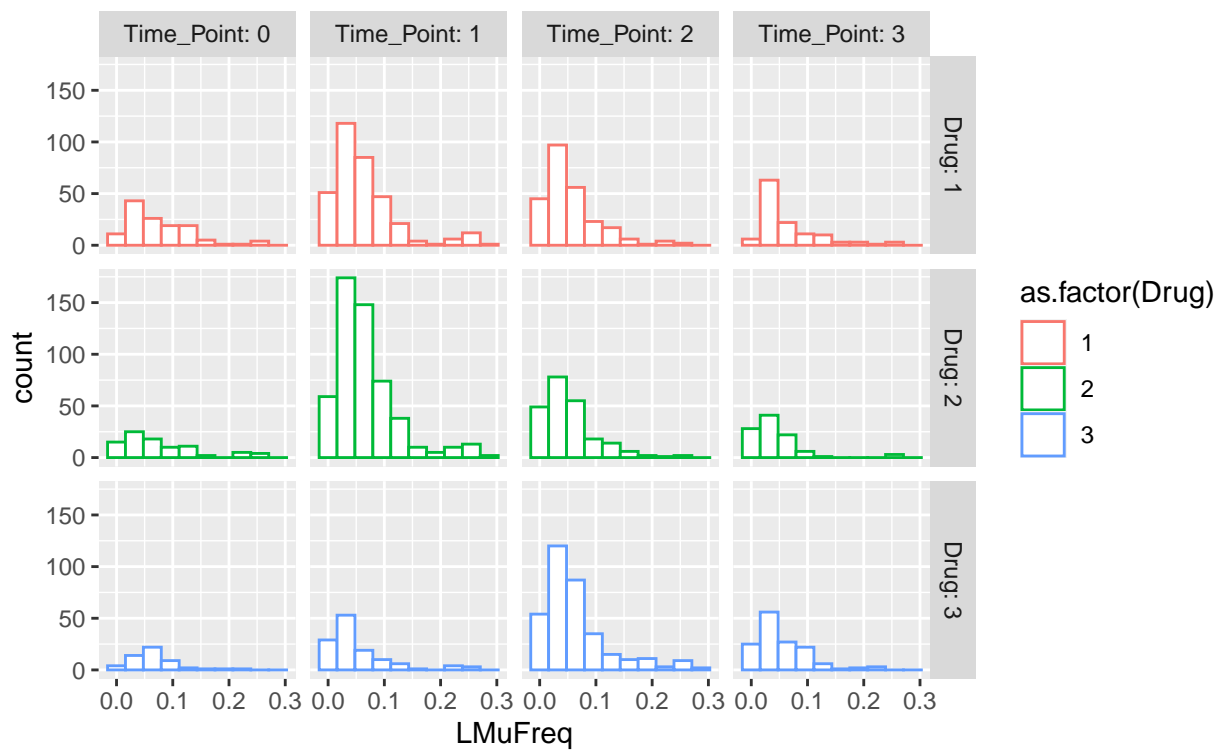


Figure 13: Histograms of LMuFreq vs Treatment and Timepoint

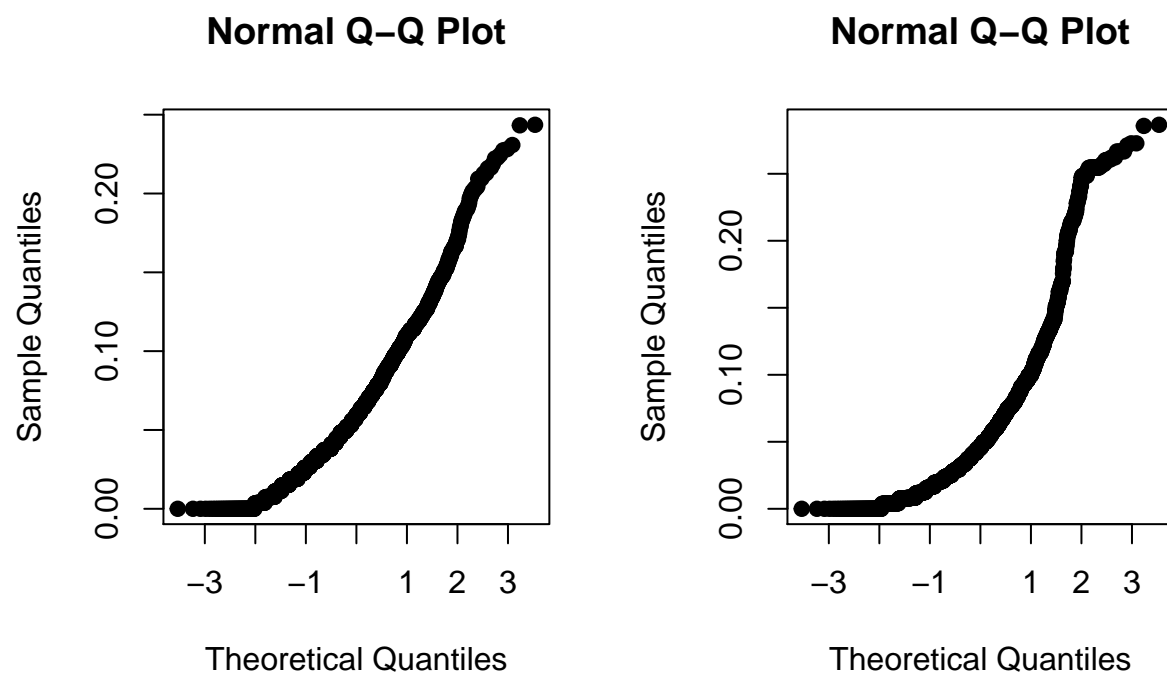


Figure 14: Q-Q Plot of HMuFreq and LMuFreq

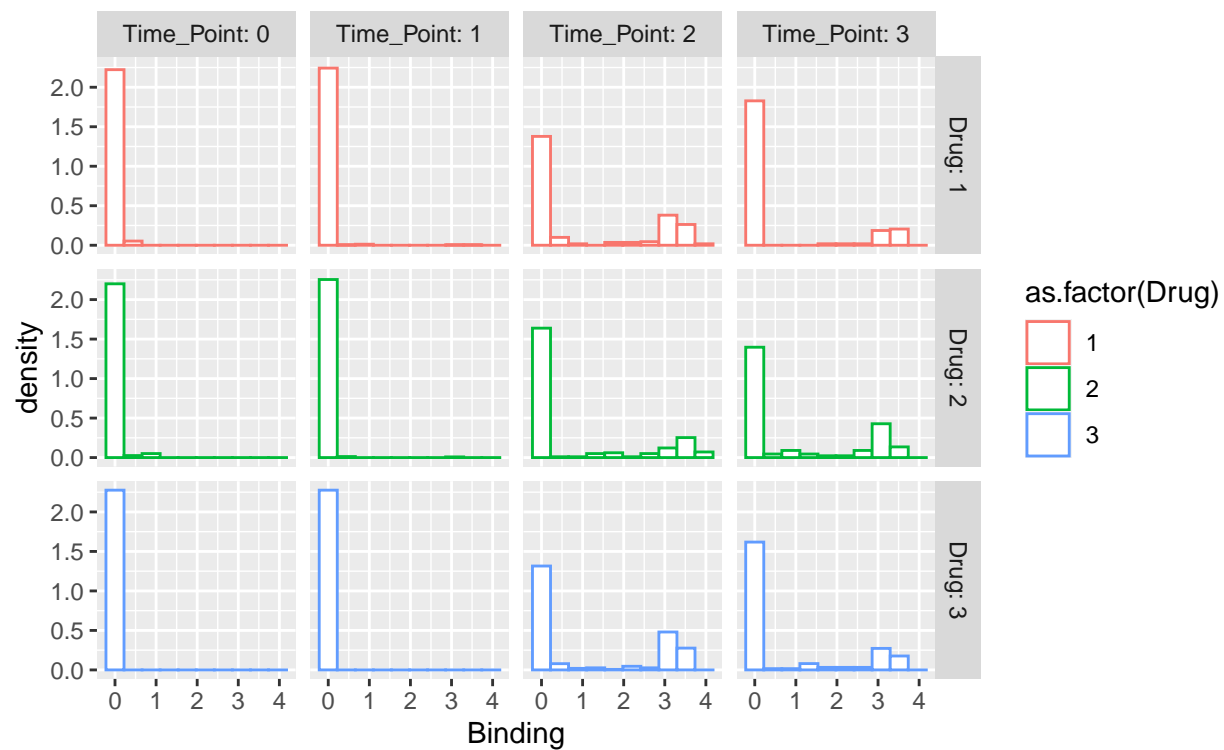


Figure 15: Histograms of Binding Strength vs Treatment and Timepoint

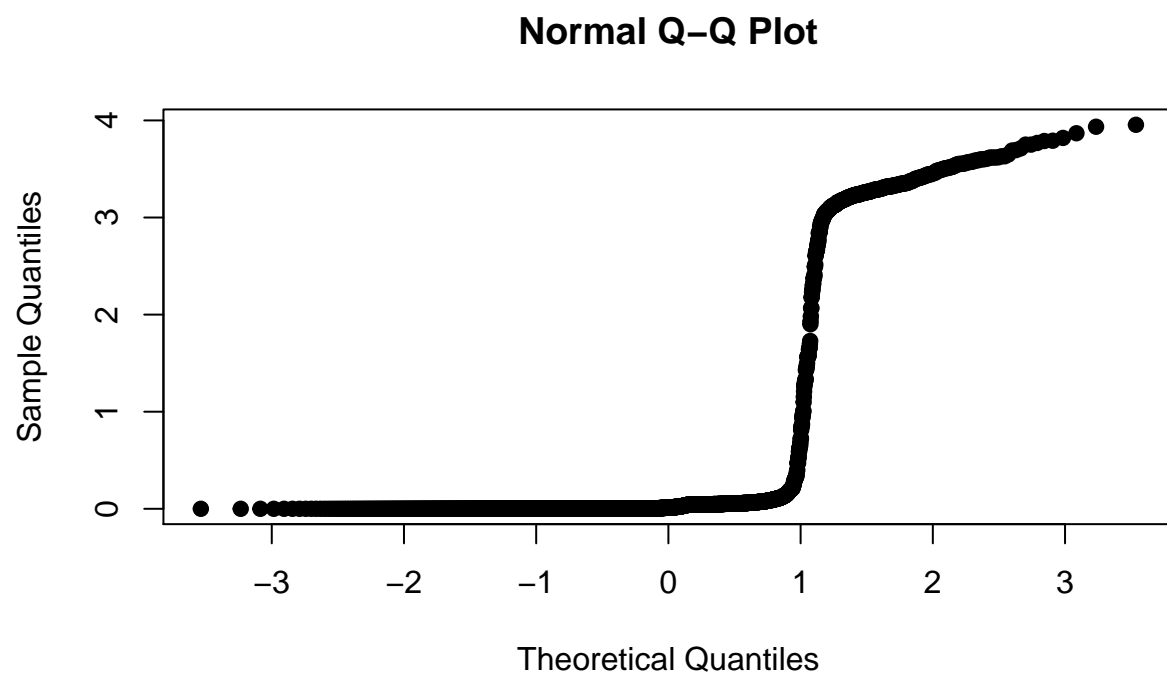


Figure 16: Q-Q Plot of Binding

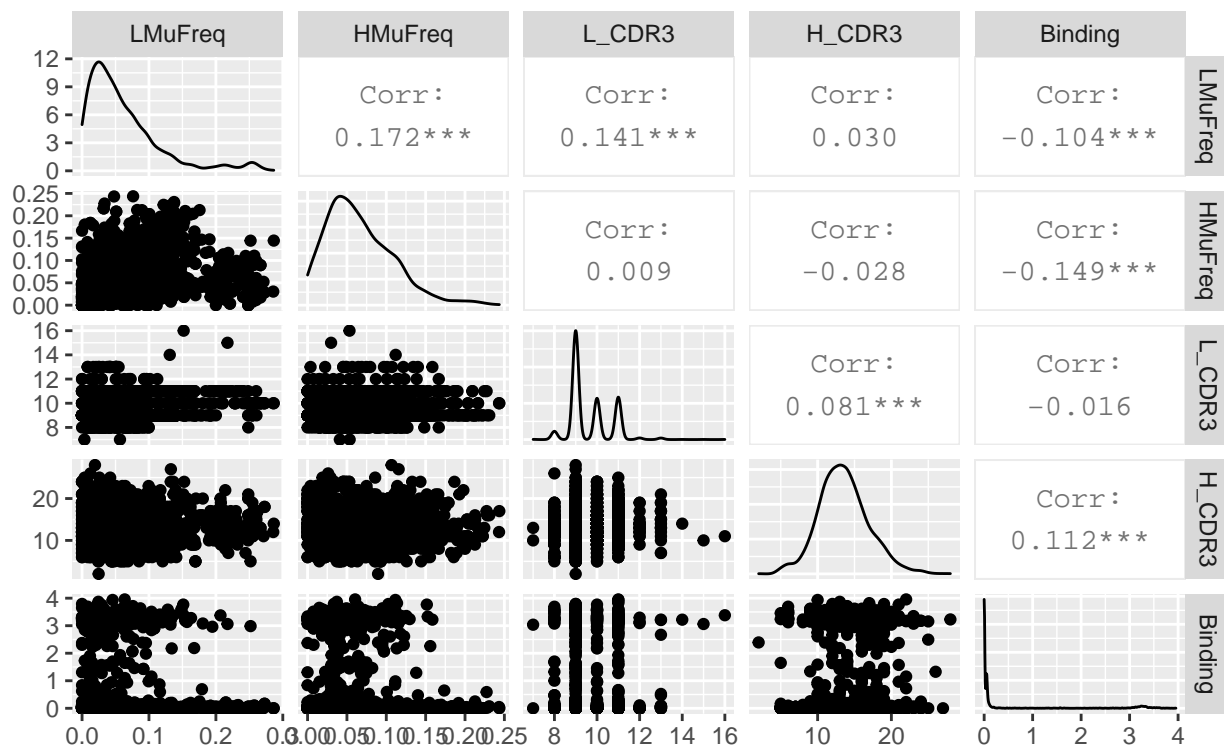


Figure 17: Plots of response variables

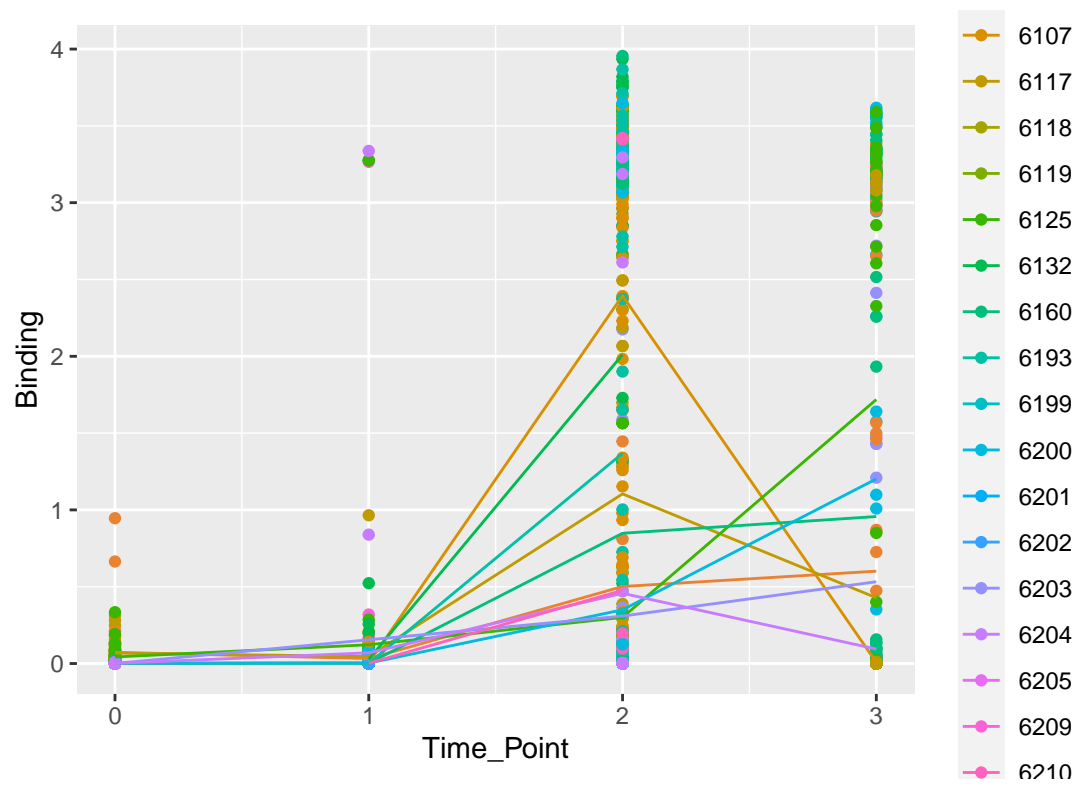


Figure 18: Mean trend by monkey

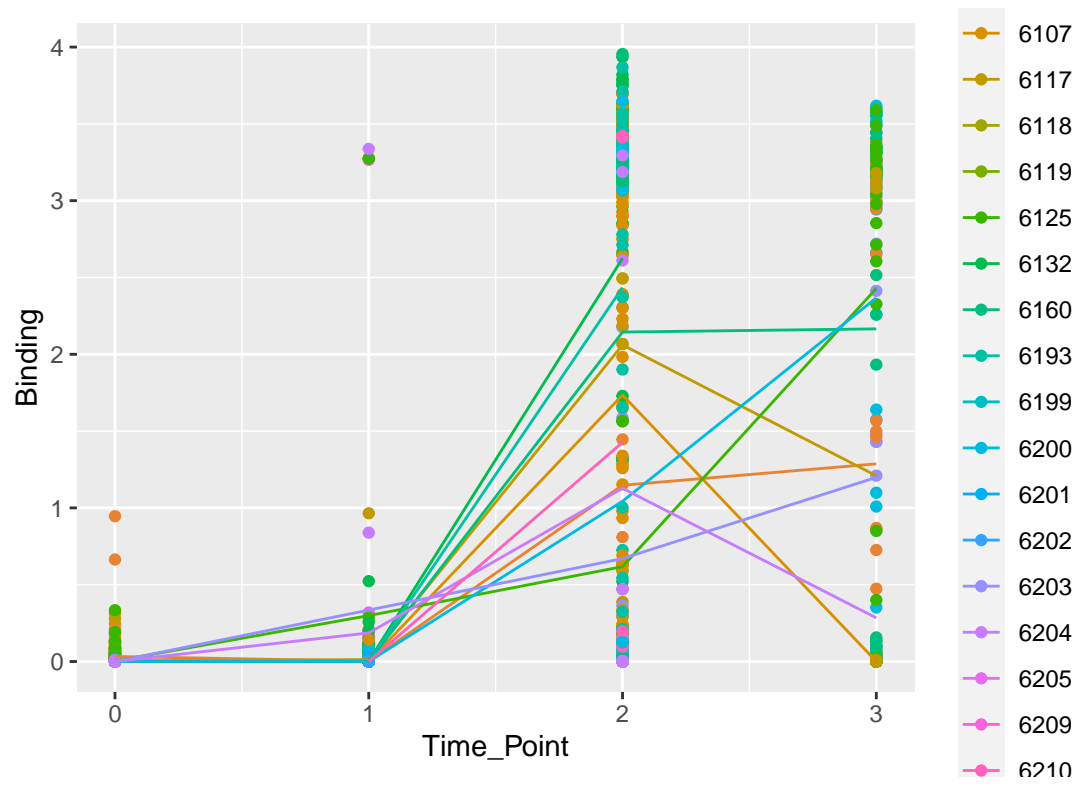


Figure 19: Variances over time by monkey

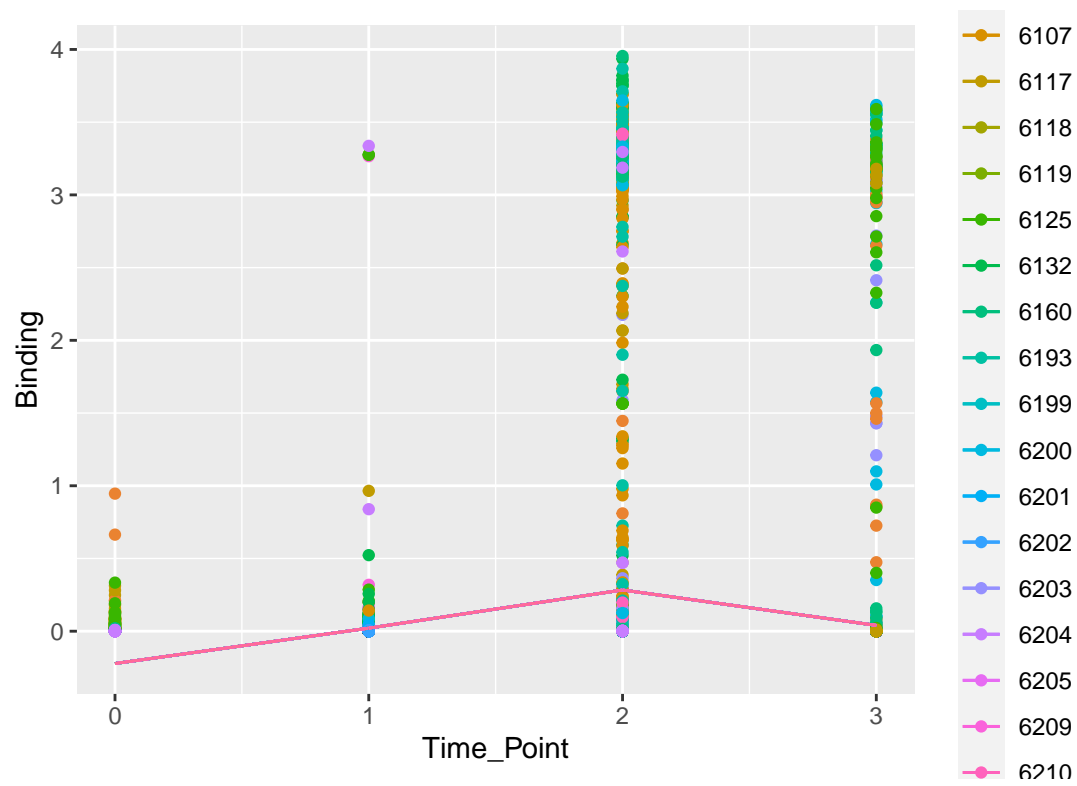


Figure 20: Piecewise Linear Function—three segments

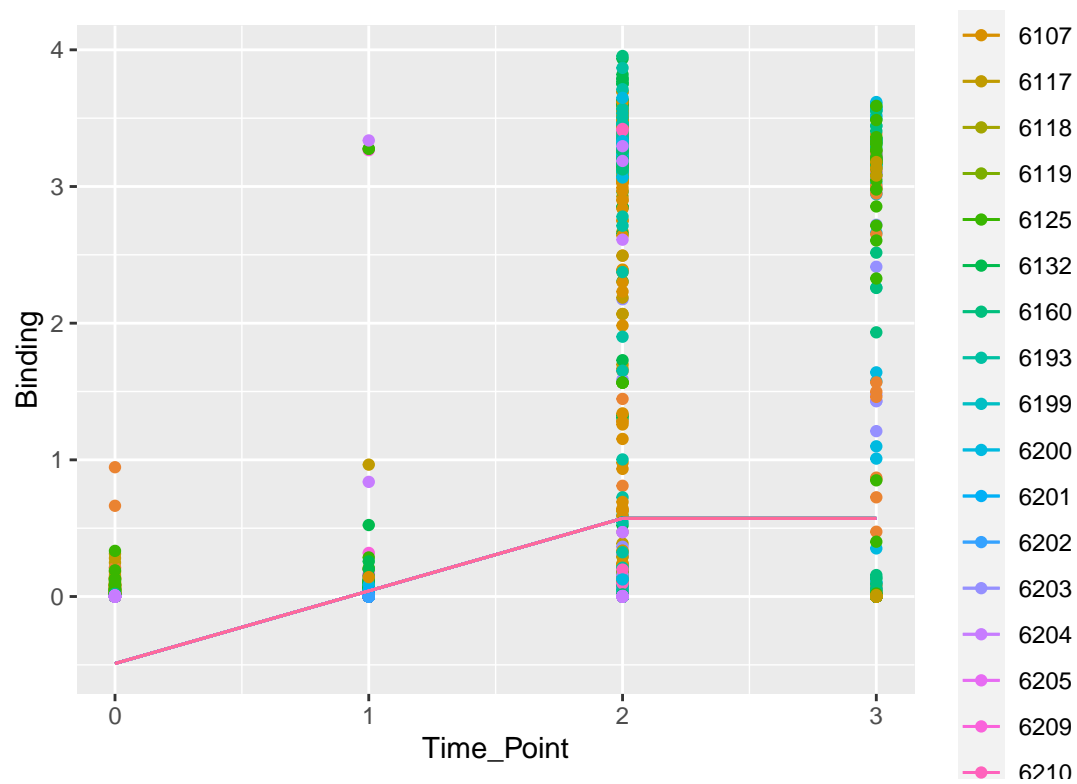


Figure 21: Piecewise Linear Function—two segments

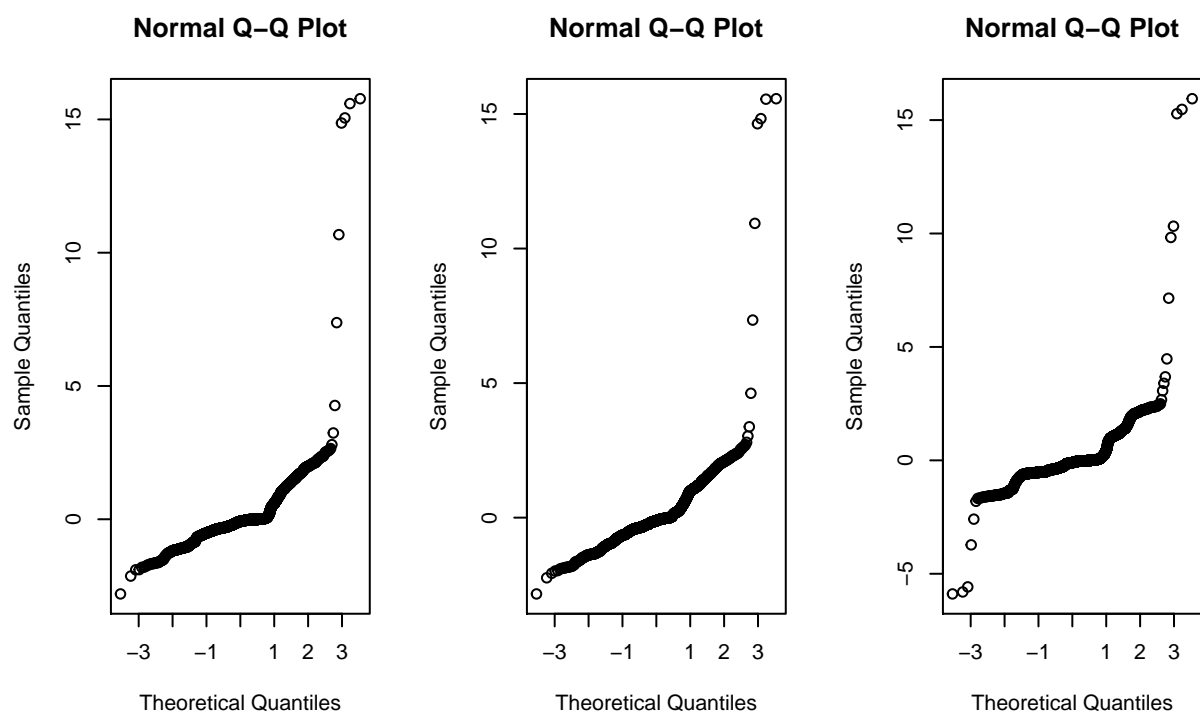


Figure 22: Q-Q plots of models: GLS, compound symmetry, AR1

List of Tables

1	Frequency tables of drug vs. timepoints	47
2	Frequency tables of timepoints vs. isotypes for drug = 1 (left), 2 (middle), 3 (right)	48
3	Summaries of standardized LCDR3	49
4	Significant Pairs in terms of Time Point, Drug, and Isotype: Significant Pairs . .	50
5	AIC and BIC between two gls models	51
6	AIC and BIC for three models	52
7	Inference of S4 ad S5 slopes	53
8	Test whether drug 1 = drug 2	54
9	Test whether drug 1 = drug 3	55
10	Test whether drug 2 = drug 3	56

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		A	D	E	G	M		A	D	E	G	M
0	4	4	1	60	60	0	6	4	0	37	43	0	1	1	0	24	28
1	11	22	2	91	220	1	10	45	2	205	271	1	1	26	0	28	70
2	8	15	3	145	80	2	4	19	1	115	86	2	4	14	0	170	159
3	1	16	1	57	47	3	1	7	0	53	40	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: Significant Pairs in terms of Time Point, Drug, and Isotype: Significant Pairs

	Time_Point	Drug	Isotype
H_CDR3	3 > 1, 3 > 2	1 > 2	IgD < IgG, IgG > IgM
HMufreq	0 > 2, 0 > 3, 1 > 2, 1 > 3	2 > 1 > 3	IgD < IgG, IgG > IgM
L_CDR3	none	none	none
LMufreq	0 > 2, 0 > 3, 1 > 3	none	none

Table 5: 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 6: 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 7: 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 8: Test whether drug 1 = drug 2

Fstat	p_value
1.065151	0.3626666

Table 9: Test whether drug 1 = drug 3

Fstat	p_value
1.231968	0.2964737

Table 10: Test whether drug 2 = drug 3

Fstat	p_value
1.255448	0.288075