**Changes in protein under Neuropathic Pain Model and local electrical stimulation**

Client: Elena Vikentievna Romanova

STAT 427

Group members:

Yijin Wang

Jue Li

Yongxin Cai

Ziyi Xue

# 

# Table of Content

[**Table of Content**](#_p98v60tvr37i) 2

[**1. Introduction**](#_q3fawmv3lky) 3

[**2. Data Description**](#_97l8w02leykm) 3

[2.1 Introduction to dataset](#_9u4oucieqjrc) 3

[2.2 unbalance & missing value existing](#_r14fm0vbdznn) 4

[2.3. Correlation](#_cuok9l2bu1sz) 5

[**3. Methodology and Analysis**](#_dvrpd3t5r1y9) 6

[3.1 Independent](#_j5keiugpk9wl) 6

[3.2 Normality](#_v08lzx8dxmm4) 6

[3.3 Homogeneity of Variances](#_pw25dzso4w0i) 7

[3.4 ANOVA Analysis](#_834vpohzpnhs) 8

[**4. Result**](#_5ffeijujq9n0) 8

[4.1 Naive vs SNI\_1](#_r9rjgzknbvk9) 9

[4.2 SNI\_1 at the 4 time stage](#_3h7lxzwop4s4) 11

[4.3 SNI\_1 vs SNI\_2](#_n9md3avgpflb) 11

[4.4 different treatments at same time stage](#_txzkrzwk05bv) 12

[4.5 Additional test](#_jgwuuotv3cpn) 13

[**5. Conclusion**](#_tw2igcz3ww0l) 15

[**6. Appendix**](#_2lyo5xlgy5bs) 16

# 1. Introduction

The dorsal root ganglion (DRG) is our body's pain control center. It contains primary sensory neurons that transmit pain signals to the brain. Neuropathic pain is an unpleasant sensation, usually triggered by an event or injury, and sent spontaneously by the body (specifically the DRG) to send pain signals to the brain. Neuropathic pain affects 7% to 8% of the adult population worldwide and can be difficult to treat. Researchers have been working hard to find effective treatments for this type of pain, but there is no perfect solution yet.

Generally, the goal for this project is to determine changes in peptides/proteins in DRGs associated with injury-induced pain and electrical stimulation treatment. In the dataset, we have four different treatment groups Naïve, SNI\_1, SNI\_2 and SNI\_3 and corresponding data associated with different proteins. We want to make group comparisons based on the protein value signal differences between Naïve and SNI treatment groups at 7thday, 21st day, 35th day, and 56th day four time points; with same SNI treatment between the four time points; between different SNI treatments at every time point; and between Naïve and SNI\_1 and / or SNI\_2 and / or SNI\_3. For the above tests, we use the ANOVA method to complete them, and the null hypothesis for each group comparison test is that they are not statistically significantly different in protein signal values.

We will discuss details about the dataset, ANOVA method, and analyzing results in the following sections.

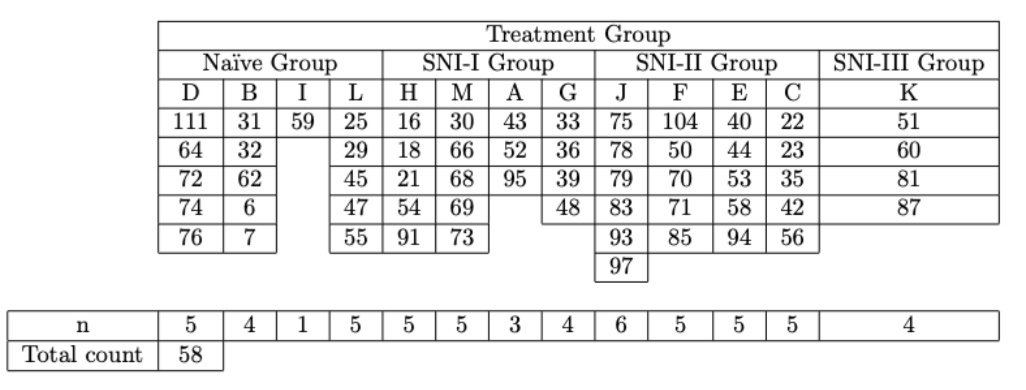
# 2. Data Description

## 2.1 Introduction to dataset

There are three files that include three different Treatment groups, For Naive Treatment Group, SNI-I Group, SNI-II & SNI-III Group, and each small group of them are tested four times: 7th day, 21th day, 35th day, 56th day.

* Naïve Group: D, B, I, L;
* SNI Group: H, M, A, G;
* SNI-I Group: J, F, E, C
* SNI-II Group: K

After removing the useless IDs where it does not contain any value, the following table shows the total number of animals we experimented with in each treatment Group.

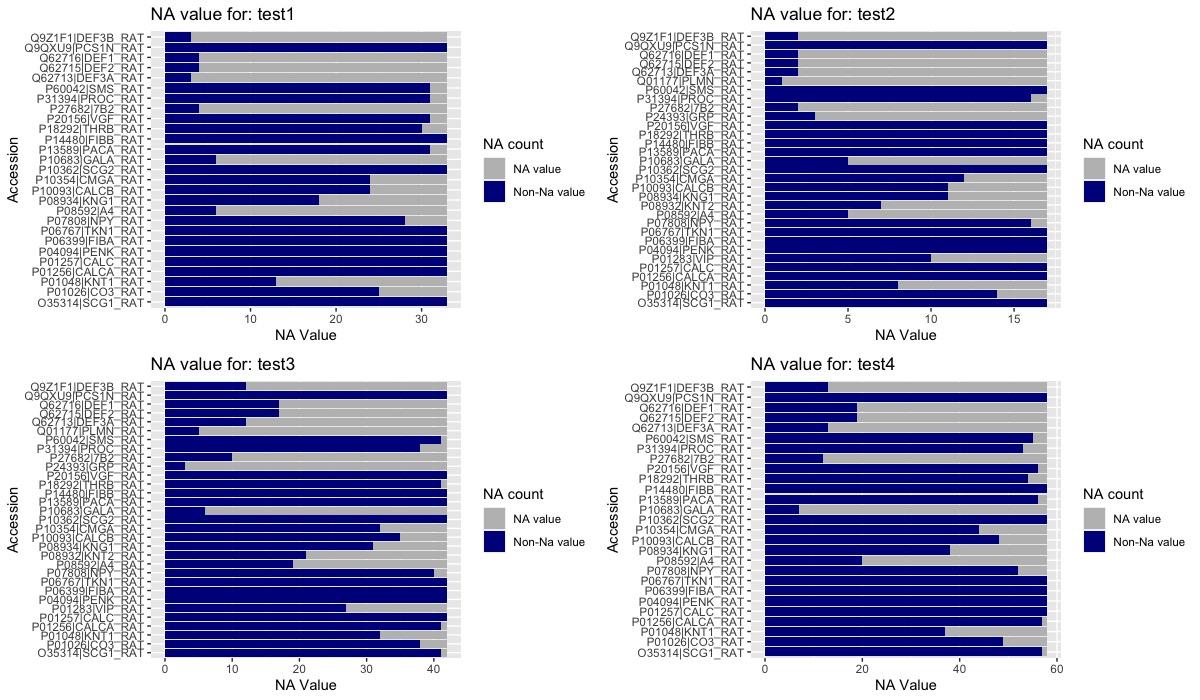


*Table 2.1.1. This table shows the treatment ID used in each experimental group*

Then we process data and save data in four dataset which would be used to test the result.

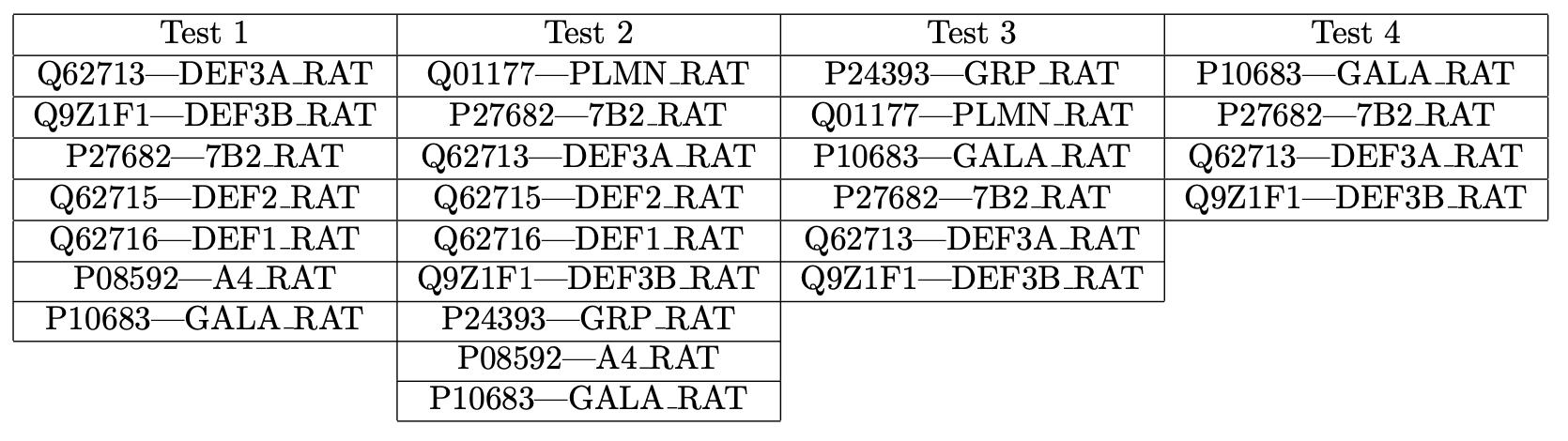
## 2.2 Unbalance Data & Missing Value

After reviewing data uses, we could observe that for each test, there are some Accession containing missing values. The Figure below shows the missing value exist in each test.



*Figure 2.2.1: This graph shows for four tests, each Accession contains missing value and non missing value count. Here, the x-axis is the counting number, y-axis is the Accession.*

By viewing the four graphs, we could observe that there are some Accession which have the most value in NA. Here, we list a table 2.2.1 and shows that for each test, which Accession has 70% or more missing value percent



*Table 2.2.1: This table shows the Accession which contains larger than 70% values are missing values*

We create four new files which already remove all the NA values. And we use these files to test our test.

## 2.3. Correlation

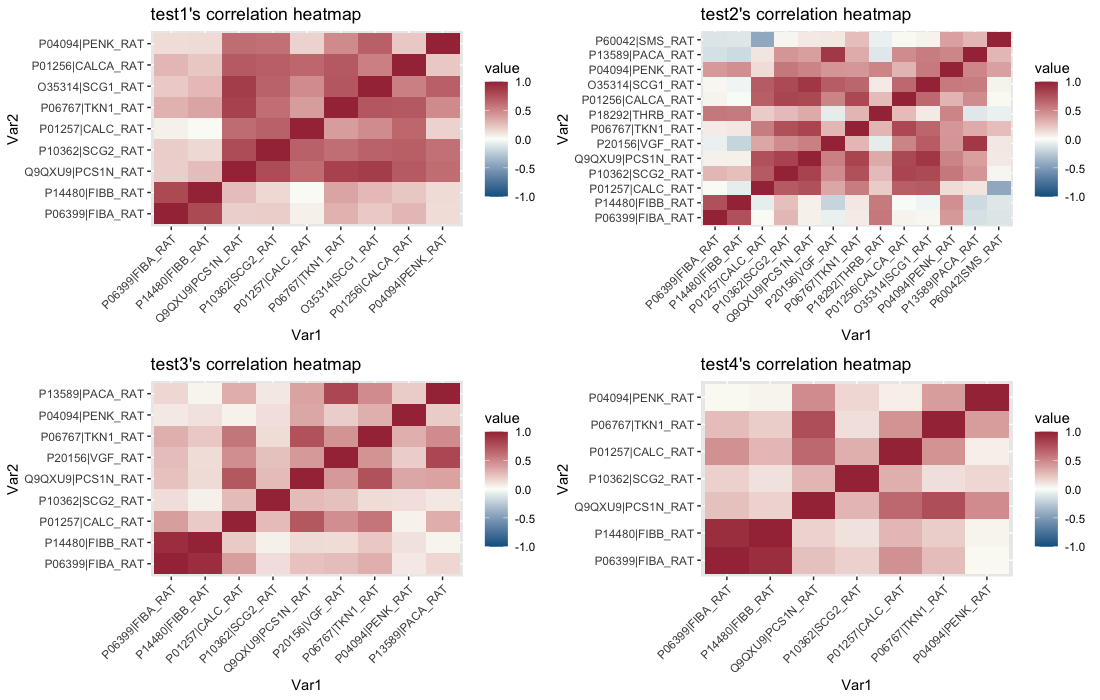
Correlation is a statistical measure that describes the degree of association or linear relationship between two variables. It measures how closely related two variables are and the direction and strength of that relationship.

In this part we are interested in the correlation between two different Accession. Since if the value used for correlation test contained a missing value, the result would return a NA value, so that we remove the Accession which contains the missing value,

The correlation function we used is Pearson correlation test which function is:

and use this method to calculate the correlation value.

Then we created a correlation heatmap , in figure 2.3.1, which shows every Accesion’s correlation result.



*Figure 2.3.1, This plot shows the correlation heatmap for four test, where Var1 , Var 2 shows the two Accession name, The color shows the correlation value, if it become a visualize dark red, it would be a strong correlation, if it tend to be blue, it would be a negative correlation, Since the color become deeper, its absolute correlation value would be higher, if the color be white, the correlation would be zero*

By the graph, there are some deep red squares in the plot, which means that there are some Accession in each test which contains high correlation.

For test 1, we conducted a correlation test between two different Accession and found that out of 36 correlation tests, 7 pairs showed strong correlation. For test 2, there are 14 strong correlation results. For test 3, there are 4 strong correlation results. And for test 4, there are 2 strong correlation results.

There are 2 Accession pairs has a strong correlation in every test, these Accession pairs are:

* P14480|FIBB\_RAT P06399|FIBA\_RAT
* P06767|TKN1\_RAT Q9QXU9|PCS1N\_RAT

# 3. Methodology and Analysis

# 

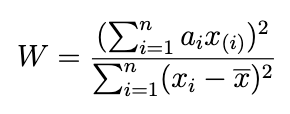
To ensure the reliability and accuracy of our results, several assumptions of ANOVA need to be met.

## 3.1 Independent

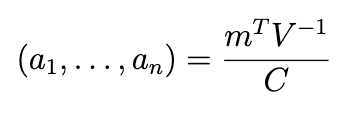
The first assumption is independence, which means that the observations within each group must be independent of each other. In our project, since the mice are independent, we satisfy this assumption.

## 3.2 Normality

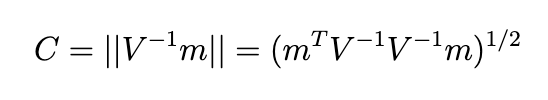
The second assumption is normality, which states that the data within each group should follow a normal distribution. We can verify this by using a histogram or conducting a Shapiro-Wilk test. For our project, we utilize the Shapiro-Wilk test, which is a test of normality. The test statistic is calculated using the following formula:



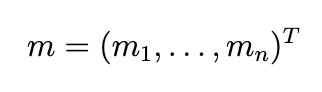
In the formula, represents the th order statistic (i.e., the th-smallest number in the sample), denotes the sample mean. The coefficients are given by:



where is a vector norm:



and the vector ,

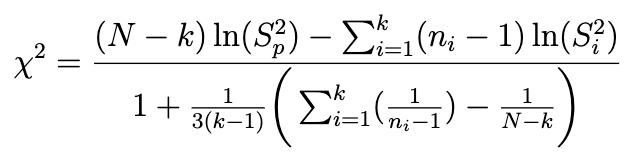


is made of the [expected values](https://en.wikipedia.org/wiki/Expected_value) of the [order statistics](https://en.wikipedia.org/wiki/Order_statistic) of [independent and identically distributed random variables](https://en.wikipedia.org/wiki/Independent_and_identically_distributed_random_variables) sampled from the standard normal distribution, and is the [covariance matrix](https://en.wikipedia.org/wiki/Covariance_matrix) of those normal order statistics. The distribution of does not have a specific name. The cutoff values for the statistics are calculated through Monte Carlo simulations.

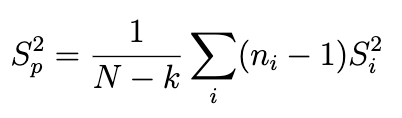
The null hypothesis of this test is that the population is normally distributed. If the p-value is lower than the chosen alpha level (0.05 in our project), we reject the null hypothesis, indicating that the tested data are not normally distributed. Conversely, if the p-value is greater than the chosen alpha level, we fail to reject the null hypothesis, suggesting that the data may come from a normally distributed population. (Wikipedia)

## 3.3 Homogeneity of Variances

Lastly, ANOVA assumes homogeneity of variances, which means that the variances of the different groups should be equal or approximately equal. We can assess this assumption by using a boxplot or conducting a statistical test like Bartlett's test. In our project, we employ Bartlett's test to check this assumption. The test statistic is calculated using the following formula:



where is the variance of the th group, is the total sample size, is the sample size of the th group, is the number of groups, and is the pooled variance. The pooled variance is a weighted average of the group variances and is defined as,



The test statistic follows an approximate distribution. Therefore, the null hypothesis is rejected when the calculated test statistic exceeds the upper tail critical value for the distribution. (Wikipedia)

By adhering to these assumptions, we ensure the robustness and validity of our ANOVA analysis.

## 3.4 ANOVA Analysis

In our current research, our primary objective is to investigate the existence of significant differences among different groups. In order to achieve this objective, we have opted to utilize the ANOVA (Analysis of Variance) test. This statistical method enables us to compare the means of two or more groups, and determine whether there is a statistically significant difference between them.

This test relies on several assumptions, including normality of data distribution, similarity of variances across groups, and independence of observations. To ensure the validity of our ANOVA test, we have carefully selected only those Accessions that have passed the normality test and the test for Homogeneity of Variance.

The ANOVA test calculates an F-statistic, which is the ratio of the variance between the groups to the variance within the groups. This enables us to estimate the p-value. In formulating our hypotheses for this test, we have developed a null hypothesis stating that there are no significant differences between the different groups, while our alternative hypothesis posits that significant differences do exist between the groups.

In the event that the F-statistic is large and the associated p-value is small (usually less than 0.05), we are able to reject the null hypothesis and conclude that there is a statistically significant difference between the means of the groups.

We have also noted that our test data are unbalanced, and have hypothesized that the sample size of the treatment group is independent of the sample size. Thus, we have decided to employ the ANOVA test with Type III error. This particular type of test allows for the testing of a main effect after the other main effect and interaction. This approach is valid when significant interactions are present.

# 4. Result

Based on our test goals, we firstly estimated whether there exist group differences in each comparision. In these four tests, as shown in table 4.1, we could observe that group differences are significant in all of the subtests, with the small p-values (less than 0.05), shown as table 4.1 p-value. This means that the detected signals of proteins have differences between the Naive and SNI\_1 on 7th day, 21th day, and 56th day. The detected signals change across time. The detected signals perform differently with different treatments at the same time stage . And there exist differences between SNI\_3 on the 56th day and SNI\_2 on the 35th day. After that, we figured out whether the group differences exist for each protein and each sub-test.

| Test | Sub-test | p-value |
| --- | --- | --- |
| Naive vs SNI\_1 | DH | 2.23E-10 |
| BM | 6.31E-08 |
| LG | 1.06E-05 |
| SNI\_1 for 4 time stages | HMAG | 1.91E-08 |
| SNI\_1 vs SNI\_2 | HJ | 4.57E-07 |
| MF | 0.000808537 |
| AE | 7.59E-06 |
| GC | 1.95E-05 |
| Naive vs SNI\_1 vs SNI\_2  (vs SNI\_3) | DHJ | 2.20E-09 |
| BMF | 7.76E-05 |
| LGCK | 2.12E-07 |
| Additional test | KEC | 1.64E-05 |

*table 4.1: This table shows the subtests of each tests and p-values for ANOVA analysis*

*with p-values < 0.05, the group effect for each subtest is significant*

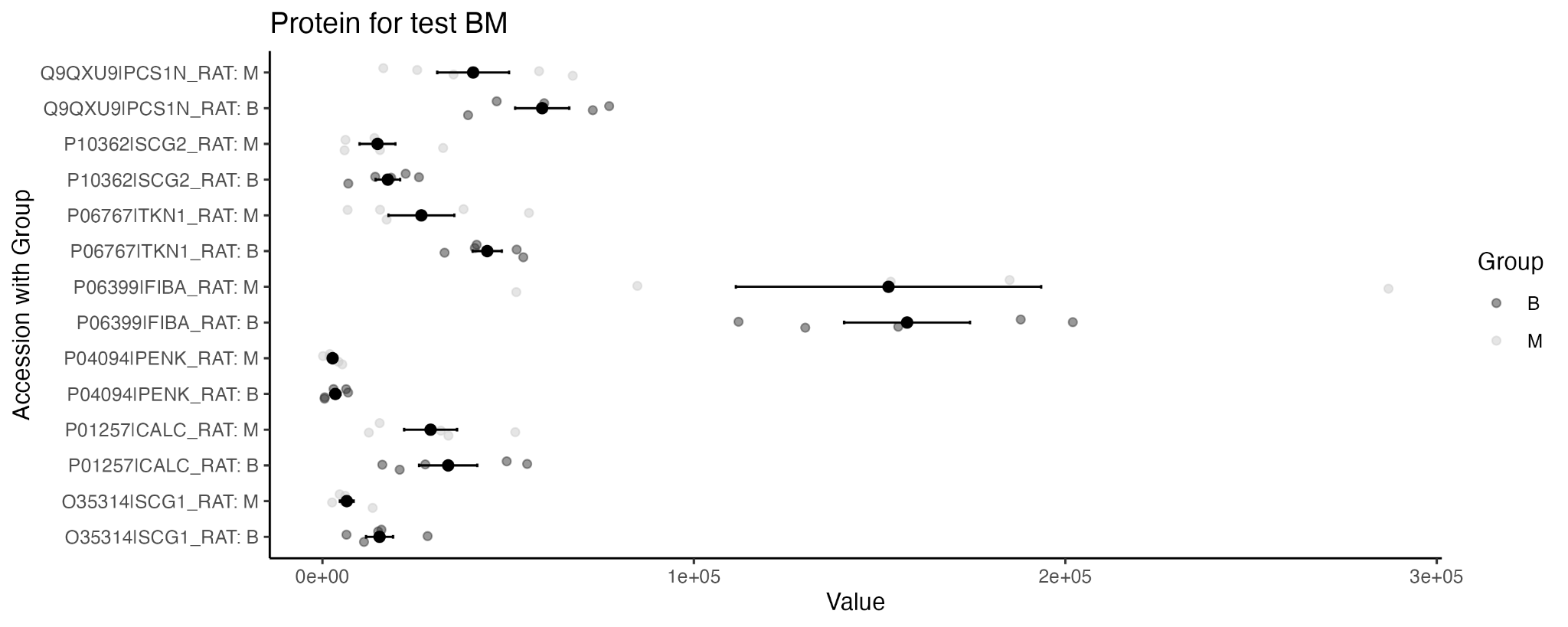
## 4.1 Naive vs SNI\_1

In the comparison between group D and H (which are on the 7th day), we figure out there are seven proteins satisfying the normal and constant variance assumption of ANOVA test. And based on the small p values (less than 0.05), shown as table 4.1.1, we could conclude that the differences of detected signals of these seven proteins are significant between group D and H. By comparing group B and M (which are on the 21th day), we figure out there are seven proteins (shown as table 4.1.1) satisfying the assumptions of ANOVA test. And the differences of detected signals of these seven proteins are statistically significant for test B vs M. For the third sub-test, L vs G, there are only five proteins satisfying the ANOVA assumptions. And based on the p-values (shown as table 4.1.1), we could figure out except the protein, “O35314|SCG1\_RAT”, the differences of other proteins are statistically significant. The similar result could also be observed from the figure 4.1.1, which shows that the values of the detected signals of “O35314|SCG1\_RAT” are pretty close.

|  | p-value for D vs H | p-value for B vs M | p-value for L vs G |
| --- | --- | --- | --- |
| P06399|FIBA\_RAT | 0.000640194 | 0.000400038 | 8.77E-05 |
| P14480|FIBB\_RAT | 0.001058232 |  | 0.00013245 |
| Q9QXU9|PCS1N\_RAT | 0.001815973 | 0.000109691 | 0.011695974 |
| P10362|SCG2\_RAT | 0.002858316 | 0.001781726 | 0.019981099 |
| P01257|CALC\_RAT | 0.001137975 |  |  |
| P06767|TKN1\_RAT | 0.001175647 | 0.000231123 |  |
| P01256|CALCA\_RAT | 4.59E-05 |  |  |
| P01257|CALC\_RAT |  | 0.001126552 |  |
| O35314|SCG1\_RAT |  | 0.001364093 | 0.063544552 |
| P04094|PENK\_RAT |  | 0.016346394 |  |

*Table 4.1.1 This table shows the subtests of Naive vs SNI\_1 and p-values for ANOVA analysis*

*with p-values < 0.05, the group effect for the protein is significant*

**

*Figure 4.1.1: this figure shows the distribution of the values of detected signals for each protein of group B and group M*

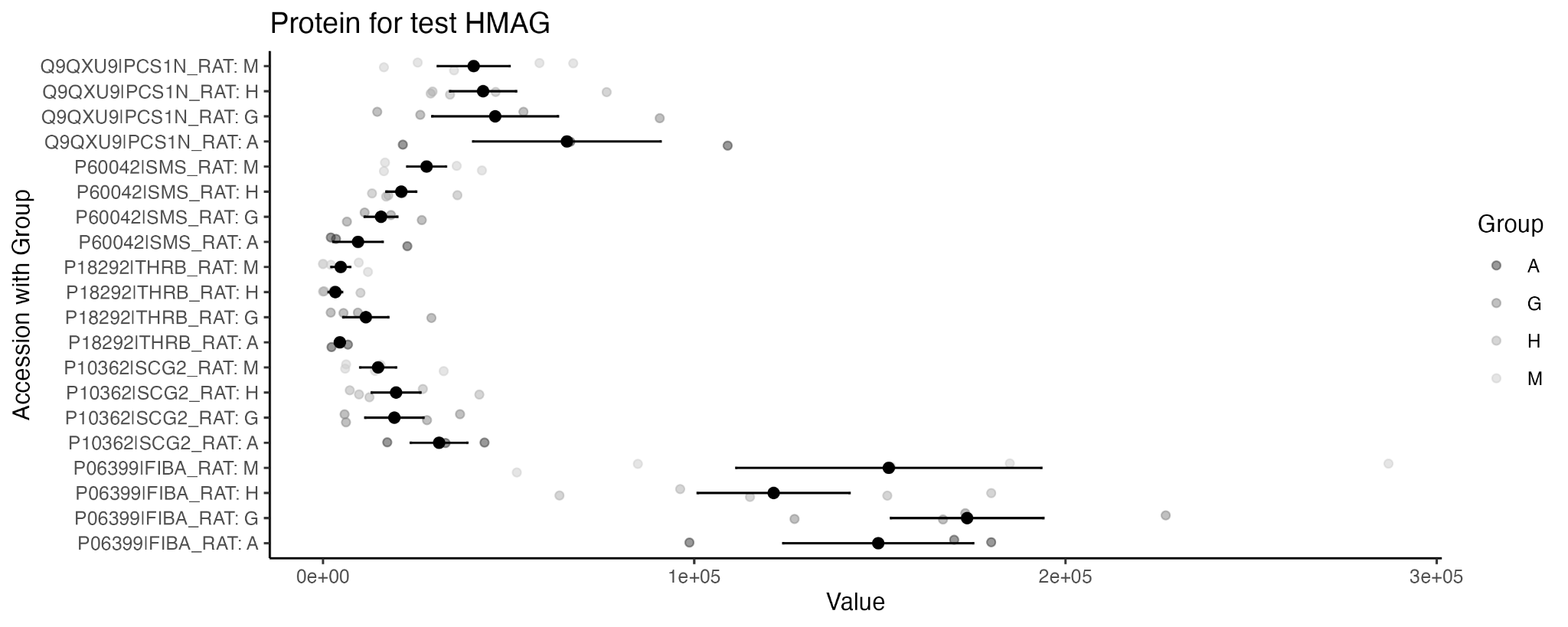
Additionally, based on the mean comparison plots (in appendix), we observed that for all six proteins on the 7th day, the detected signals are larger for SNI\_1 (H) than the Naive group (D). And for all six proteins on the 21th day, the detected signals are larger for the Naive (B) than the SNI\_1 group (M). And for proteins, "9QXU9|PCS1N\_RAT", "P10362|SCG2\_RAT", and "O35314|SCG1\_RAT" on the 56th day, the detected signals are larger for the Naive (L) than the SNI\_1 group (G). And "P14480|FIBB\_RAT" and "P06399|FIBA\_RAT" are smaller for the Naive (L) than the SNI\_1 group (G).

In conclusion, the differences of the detected signals of almost all the proteins between the Naive and SNI\_1 at the three time stages are statistically significant. And we only found the difference of the detected signal of “O35314|SCG1\_RAT” is not significant (p-value > 0.05) in the sub-test, L vs G. This means the detected signal of “O35314|SCG1\_RAT” may keep the same in the Naive and SNI\_1 group on the 56th day.

## 4.2 SNI\_1 at the 4 time stage

We detected five proteins satisfying the normality and constant variance assumption. And they are “P06399|FIBA\_RAT”, “P10362|SCG2\_RAT”, “Q9QXU9|PCS1N\_RAT”, “P18292|THRB\_RAT” and “P60042|SMS\_RAT”. And because the p-values for each protein are very small (less than 0.05), the difference of the detected signal of the five proteins are statistically significant. This means there exist group differences for the detected signals of these five proteins with SNI\_1 treatment across time.

Additionally, based on the mean comparison plot (figure 4.2.1), we could only observed the values of detected signals of the protein "P18292|THRB\_RAT" become larger with increase of time (from H to M to A to G). The other change of values does not correspond to the time.

*Figure 4.2.1: the mean comparison plot of H vs M vs A vs G*

## 4.3 SNI\_1 vs SNI\_2

In the sub-test, H vs J, there are six proteins satisfying the assumption of ANOVA test. And p-values of the six proteins, shown as table 4.3.1, are very small (less than 0.05). The small p-values means the differences of the detected signals of these proteins between group H and J are statistically significant. By comparing the group M and F, which are on the 21th day, five proteins are satisfying the assumptions. And all differences are significant based on the small p-values (less than 0.05), shown as table 4.3.1. For sub-test A vs E, six proteins satisfying the ANOVA assumptions are statistically significant with small p-values (less than 0.05). And for the comparison between group G and C, there are only four proteins satisfying the ANOVA assumptions. The detected signals of these proteins are statistically significant with small p-values (less than 0.05), shown as Table 4.3.1.

|  | p-value for  H vs J | p-value for  M vs F | p-value for  A vs E | p-value for  G vs C |
| --- | --- | --- | --- | --- |
| P06399|FIBA\_RAT | 0.000137247 | 9.67E-05 |  | 0.000201091 |
| P01257|CALC\_RAT | 0.00118793 |  | 0.016334871 | 0.011801214 |
| P10362|SCG2\_RAT | 0.00040209 |  | 0.001928897 | 0.005590257 |
| Q9QXU9|PCS1N\_RAT | 0.000374361 |  | 0.002184541 | 0.002235801 |
| P06767|TKN1\_RAT | 0.001325137 | 0.003737754 | 0.005997704 |  |
| P13589|PACA\_RAT | 0.002805251 | 0.004626002 | 0.011327008 |  |
| P20156|VGF\_RAT |  | 0.000281004 | 0.02161828 |  |
| P04094|PENK\_RAT |  | 0.001060294 |  |  |

*Table 4.3.1 This table shows the subtests of SNI\_1 vs SNI\_2 and p-values for ANOVA analysis*

*with p-values < 0.05, the group effect for the protein is significant*

Additionally, based on the mean comparison plots (in appendix), we observed that for the proteins satisfying the ANOVA assumptions, the values of detected signals of each protein for the SNI\_1 group are larger than the SNI\_2 group for all four time courses.

In conclusion, the differences of detected signals for all the proteins satisfying the ANOVA assumptions are statistically significant for SNI\_1 and SNI\_2.

## 4.4 Different treatments at same time stage

In the test of different treatments at the same time course, there are three subtests, D vs H vs J (which are groups on the 7th day), B vs M vs F (on the 21th day), L vs G vs C vs K (on the 56th day). In the sub-test D vs H vs J, five proteins satisfying the ANOVA assumptions have small p-values (less than 0.05), shown as table 4.4.1, which means the differences of the detected signals across treatments on the 7th day of these five proteins are significant. In the sub-test B vs M vs F , there are only three proteins satisfying the ANOVA assumptions. And with small p-values (less than 0.05), shown as table 4.4.1, the differences across treatments for each protein are significant. The last sub-test is L vs G vs C vs K, which is to compare the Naive, SNI\_1, SNI\_2, and SNI\_3 groups on the 56th day. As shown on the table 4.4.1, there are only three proteins satisfying the ANOVA assumptions. And with small p-values (less than 0.05), the differences of each protein are statistically significant for the different treatments.

|  | p-value for  D vs H vs J | p-value for  B vs M vs F | p-value for  L vs G vs C vs K |
| --- | --- | --- | --- |
| P06399|FIBA\_RAT | 1.27E-05 | 1.53E-06 | 6.42E-07 |
| Q9QXU9|PCS1N\_RAT | 0.000104932 |  | 4.27E-05 |
| P10362|SCG2\_RAT | 1.98E-05 |  | 0.000743571 |
| P01257|CALC\_RAT | 0.00016599 |  |  |
| P06767|TKN1\_RAT | 0.000154344 | 2.44E-05 |  |
| P04094|PENK\_RAT |  | 0.000471582 |  |

*Table 4.4.1 This table shows the subtests of different treatment at the same time courses*

*and p-values for ANOVA analysis*

*with p-values < 0.05, the group effect for the protein is significant*

Additionally, based on the mean comparison plots (in appendix), we observed the proteins, "P01257|CALC\_RAT" and "P06767|TKN1\_RAT" on the 7th day, show the very clear pattern that values of the signals are largest for the SNI\_2 group, smaller for the SNI\_1, and smallest for the Naive. And the protein "P06399|FIBA\_RAT" has the largest value for SNI\_2 on the 21th day and has the largest value for SNI\_3, smaller value for SNI\_2, smaller value for SNI\_1 and smallest value for Naive on the 56th day.

In conclusion, the differences of detected signals for all the proteins satisfying the ANOVA assumptions are statistically significant for different treatments at the same time course.

## 4.5 Additional test

The additional test is to compare the SNI\_2 and SNI\_3 on the 56th day (K vs C), and the SNI\_2 on the 35th day and SNI\_3 on the 56th day (K vs E). For the two subtests, the same six proteins satisfy the ANOVA assumptions. And the p-values in both the two subtests are small (less than 0.05), shown as table 4.5.1, which illustrates the difference of the detected signals are significant.

|  | p-value for  K vs E | p-value for  K vs C |
| --- | --- | --- |
| P06399|FIBA\_RAT | 0.004351112 | 0.00038339 |
| P14480|FIBB\_RAT | 0.016796998 | 0.002252158 |
| Q9QXU9|PCS1N\_RAT | 0.000281777 | 0.000344055 |
| P10362|SCG2\_RAT | 0.003699229 | 0.008274146 |
| P01257|CALC\_RAT | 0.005805802 | 0.002747393 |
| P06767|TKN1\_RAT | 0.001485025 | 0.009215514 |

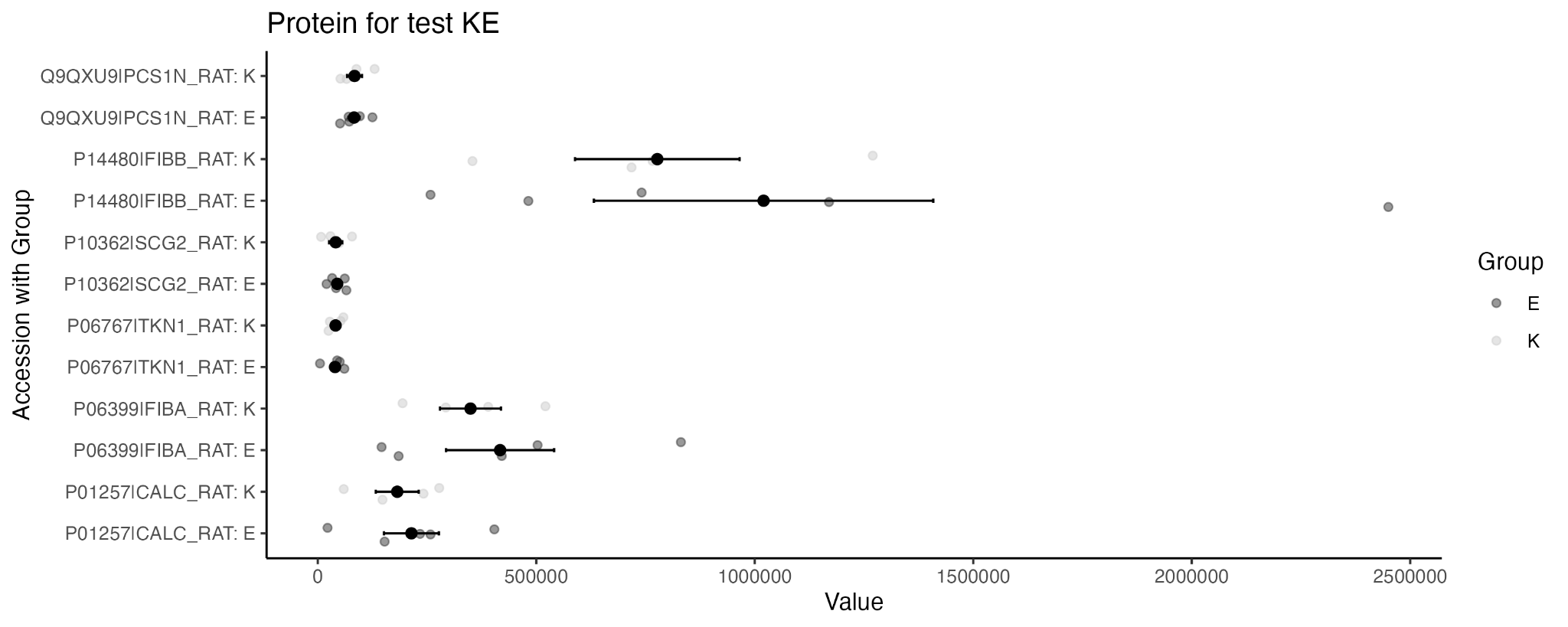
*Table 4.5.1 This table shows the test SNI\_2 and SNI\_3 on the 56th day*

*and the test SNI\_2 on the 35th day and SNI\_3 on the 56th*

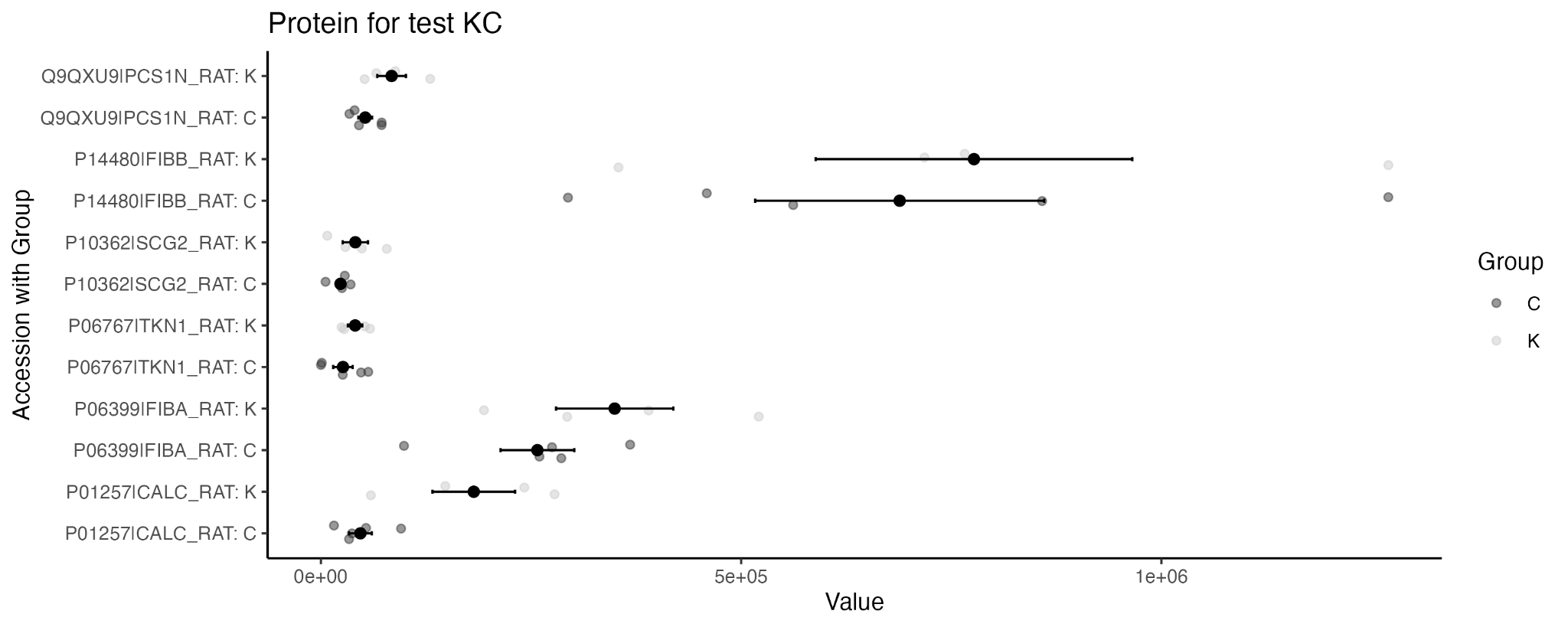
*p-values for ANOVA analysis*

*with p-values < 0.05, the group effect for the protein is significant*

Additionally, according to the figure and figure , we observed the values of signals of proteins, "P06399|FIBA\_RAT","P14480|FIBB\_RAT", "P01257|CALC\_RAT" have large differences across the groups. The group C and group E tend to have larger signals of these three proteins than the group K. This illustrates the SNI\_2 group tends to have larger values of signals of "P06399|FIBA\_RAT","P14480|FIBB\_RAT", "P01257|CALC\_RAT" than the SNI\_3 group.



*Figure 4.5.1: the mean comparison plots for each protein of group E and K*



*Figure 4.5.2: the mean comparison plots for each protein of group C and K*

In conclusion, the differences of detected signals for all the proteins satisfying the ANOVA assumptions are statistically significant for the SNI\_2 and SNI\_3 on the 56th day and for the SNI\_2 on the 35th day and SNI\_3 on the 56th day.

# 

# 5. Conclusion

After our data cleaning process such as removing missing values and some proteins as previously mentioned, all the group comparisons had statistically significant differences, as shown in *table 4.1*. It means that: there are protein value signal differences between Naïve and SNI1 treatment groups at 7th day, 21st day, 35th day, and 56th day four time points; with same SNI treatment between the four time points; at the same time point between different SNI treatments; and additional tests between Naïve and SNI\_1 and/or SNI\_2 and/or SNI\_3. Generally, there are statistically significant changes in proteins in DRGs that are associated with injury-induced or neuropathic pain and various treatments by local electrical stimulation.

For the group comparison tests for each specific protein, there is only one comparison test - the comparison test between group L and G for protein name O35314|SCG1\_RAT - has p-value greater than 0.05 (0.063544552), as shown in *table 4.1.1*. It means that there is no significant signal value difference for this particular protein between the Naïve and SNI1 treatments at the 56th day of the treatment time course. (i.e., after 56 days, the rat under Naïve and SNI1 two treatments having similar O35314|SCG1\_RAT signal values.)

Finally, we hope that in the future this project will involve more methods to improve the accuracy of the results, and we believe the researchers and scientists will find effective treatments for the Neuropathic pain as the technology advances.

# 

# 6. Appendix

Box folder, STAT 427 Report Appendix, with all the information:

<https://uofi.app.box.com/folder/206093302717?s=233wn05aa4jrc7krqdmuq45k412lj5s0>

STAT 427 Report Appendix/Code:

Five rmd files with R code and one document including code descriptions.

STAT 427 Report Appendix/data for each test (used for code only):

Eight csv files for the four tests with two versions (with and without missing data), used only for code.

STAT 427 Report Appendix/plots with two versions:

Mean comparison plots for all subtests with two versions.

STAT 427 Report Appendix/data satisfying the assumption:

csv files for each sub-tests including proteins satisfying the ANOVA assumption.

STAT 427 Report Appendix/Result:

csv files including p-values obtained from ANOVA test for each protein for each sub-tests.

a csv file including p-values obtained from ANOVA test summarizing all protein performance.

STAT 427 Report Appendix/Correlation:

contains four file :  
 Test 1 Group, Test 3 Group, Test 4 Group: The sub Group correlation test

test1 - test4: Whole Test correlation test

in these four file it each contain three file:

correlation plot: the correlation heatmap graph

strong correlation value & table: correlation value between two Accession and its absolute value larger than 0.7

full correlation value: All the correlation value exists for each Accession pairs