# Proteome-wide Screen for RNA-dependent Proteins in non-synchronized HeLa cells

Sophia Brinkmeier, Sabrina Hämmerle, Sophia Zurmühl
July 17, 2023

## Table of content

# 1. Introduction

In recent decades, it has been discovered that RNA-dependent proteins (RDPs) play a crucial role in cellular processes. RDPs are proteins that either bind directly to the RNA or are part of macromolecular complexes that are associated to the RNA (Gerstberger et al., 2014). They influence the fate of the RNA and serve as critical RNA regulators responsible for modulating various posttranscriptional events in the cell, such as RNA metabolism, stability, mRNA splicing and decay, translation, and more (Liu and Cao, 2023). Over the years an increasing number of RBPs have been associated with various diseases, including autoimmune disorders (Liu and Cao, 2023), neurodegenerative diseases (Wang et al., 2016) and cancer (Liu et al., 2021). For example, RDP signatures can serve as a prognostic biomarker in cancer patients (Li et al., 2021).

Further research and a deeper understanding of RDPs are necessary. More RDPs need to be identified and characterized to utilize this knowledge in the treatment of various diseases. To identify more RNA dependent proteins, a screening method called R-Deep was developed. R-DeeP is a proteome-wide, unbiased, enrichment free screen based on density gradient ultracentrifugation. (Caudron-Herger et al., 2019)
In this experiment, HeLa cells in different stages of the cell cycle were lysed. The cell lysate was split into two groups: one group was treated with RNase to degrade the RNA, thereby preventing RDPs from binding to it, while the other group served as the control without any treatment. For the ultracentrifugation, the cell lysate was transferred onto a sucrose gradient. The gradient was divided into 25 fractions, in which the proteins accumulated after ultracentrifugation depending on their molecular weight. RDPs exhibited a change in molecular weight due to RNA degradation, which affected their migration distance in the sucrose gradient. Heavier molecules traveled further than light molecules, resulting in a right or left shift compared to the control group.

The main objective of this project was to identify RDPs.

## 1.1 Dataset description

The provided dataset consists of 4765 rows, each representing one protein. The 150 columns contain the protein amount in the 25 fractions. For both the control and RNase groups, there are 3 replicates of each fraction to ensure reproducibility and the opportunity to provide statistical evidence. The protein amounts in each fraction are given in arbitrary units.

# 2. Methods

## 2.1 Data clean-up

**Remove replicates without data**

To prepare the data for further analysis, data clean-up is necessary. Firstly, the data frame was split into the control and the RNase groups. These groups were then further divided into the replicates to simplify the following steps of the analysis. We removed all proteins, which had at least one replicate either in the control or RNase group containing only zeros. To identify these replicates the which and apply function were used in combination, iterating over every row, and picking out the indices of those with a total sum of zero.

**Normalization**
Normalization is performed to bring the data into a standardized scale with protein amounts ranging from 0 to 100, and the sum of protein amounts across all fractions equal to 100.
Common reasons as to why normalization is applied, are to allow for a better comparison of variables, to get more accurate results in statistical tests like k-means or t-tests and to help mitigate the impact of outliers.

**Rep-wise normalization**
In the first normalization step, a loop that iterates through the columns of the control or RNase data frame was created, which selected three replicates of the same fraction at a time in order to normalize the replicates against each other.
Inside the loop, the sums for each replicate were calculated. The vector 'diffs' was created to store the absolute difference between the sums and the mean of these sums. The function 'which.min' determined the index of the lowest difference between the sums. The vector 'sum_least_difference' stored the sum with the same index. To compute the normalization-factor the lowest sum is divided by each sum of each replicate. To normalize the values, the normalization factors were multiplied with their corresponding replicates. Since these operations are performed within the loop for each fraction, the code normalizes the dataset by scaling the values based on the lowest sum of the replicates for each fraction.

**Fraction-wise normalization – scaling**
First, six data frames were created. Each data frame contained one replicate of the rep-wise normalized data. For the scaling, a loop was created that iterates over each row of the data frames and performs the following operations. For each row the sum of values is calculated. To rescale the values, the sum is divided by 100, and each value is then divided by the obtained quotient. Since the scaling is performed row-wise, it is ensured that the protein amount for each protein sums up to 100.

**Smoothing**
To smooth the curve, each value was replaced by the mean of the value itself and its two neighboring values. A function called 'replace_mean' was defined for this purpose. Using the apply function, 'replace_mean' was applied to the normalized data on each row. Thus, background noise was reduced, and important information was enhanced.

## 2.2 Correlation
By performing a correlation analysis, we further wanted to sort out proteins with deviating correlations.
There is a possibility, that during the performance of the experiments, which led to the original data, slight inconsistencies or device errors might have occurred. In this case, the three replicate values of the RNase group, and of the control group respectively, do not correspond with each other. As a result, the reproducibility of the experiment regarding this protein cannot be assured.
To carry out a correlation analysis, we first compared each replicate to one another of both the RNase and the control group.
We chose to use the Pearson-correlation because this method is more fitting for normalized data on a continuous scale.

To perform a correlation test on the replicates of the proteins, we created a for loop. In this loop, two vectors were created, each containing one replicate of one protein. The correlation coefficient was added to an empty vector, which we defined beforehand.
A correlation coefficient of 1 indicates an entirely positive linear correlation between the two data, whereas 0 indicates no linear correlation at all. We chose to use 0.8 as a threshold, since it still indicates a strong positive linear correlation but tolerates slight fluctuations. All values smaller than 0.8 were eventually deleted.

## 2.3 Peak finder

To identify RDPs based on a fraction shift, it is necessary to determine the fractions that contain the maximum protein amount. Firstly, a mean data frame was generated for both the RNase and control groups.

**Identification of the global maxima**

The 'global_peak_finder' function was created, which combined the 'which' and 'max' functions. The 'max' function was, in this case, able to identify the fraction with the highest protein amount for one protein, by iterating through the elements and comparing them to determine the largest value. The 'global_peak_finder' function was applied onto each row of our mean data frames using the 'apply' function.

**Identification of the local maxima**

To find the local peaks, each value was compared with its two neighboring values, on either side, using a for loop. Only if the value was higher than the other four values, it was defined as a local peak and was apended to an empty vector created beforehand.
To reduce peaks from the background noise, the value was not only compared with its direct neighbors, but with the two surrounding values on either side.

## 2.4 Identifying shifts

To identify shifts, we computed the difference between the mean control global peak and the mean RNase global peak. We subtracted the peaks of RNase from the peaks of the control group.

## 2.5 T-test

A t-test is a statistical test utilized for comparing the means of two groups.
It can also be distinguished whether the t-test is two sided or one sided. Which type of t-test should be used, depends on the alternative hypothesis $H_1$. Since in our case, the $H_1$ is not specific and merely states that the means of the two groups are unequal, a two-sided t-test is more appropriate. The p-value, which is computed in such a statistical test, can be seen as the measure of evidence against the null hypothesis $H_0$. To determine whether an effect is significant, a significance level $\alpha$ needs to be defined. If the p-value is below the significance level, the null hypothesis can be rejected, indicating a significant result. (Ludbrook, 2013)
In the t-test we checked if the amount of protein in the fraction of the mean global peaks significantly differed between each replicate of the RNase and control group. We tested for both fractions of the mean control global peaks and mean RNase global peaks. The null hypothesis would be, that there is no significant difference between the protein amount of the RNase and control replicates.
Before performing the t-test on our data, we reviewed if the amount of protein was the same in any replicate between RNase and control group. If so, a '1' was added to the vector of p-

values, which was created beforehand, to show that there is absolutely no difference in these proteins in the control and RNase group. We then performed the two-sided t-test on the rest of the proteins. Afterwards, the 'p.adjust' function with the Benjamini-Hochberg method is used to correct the p-values.

## 2.6 K-means

K-means clustering serves for grouping similar kinds of items into a predetermined number of clusters in a multidimensional space, so that the sum of squares from each point to the assigned cluster center is minimized.
K-means was used to identify proteins with similar characteristics in their relation between the RNase and control peaks.
K-means randomly puts centers into a coordinate system according to the number of clusters we defined. Each point, or in our case each protein, will then be assigned to the nearest cluster according to the RNase and control peak in the 2-dimensional space. For determining the distance between each point and the center, the Euclidean distance is applied by the 'k-means' function.
The centers are then relocated to the focal point of all assigned proteins. The minimal Euclidean distance is calculated again, and each protein is newly assigned to the relocated centers. Those steps are repeated until there is no change in protein assignment to the clusters anymore or until a predetermined number of maximal iterations is reached.
After performing k-means with 7 clusters, we used the elbow method to determine the ideal number of clusters. In the elbow method, the total within-cluster sum of squares of each cluster is plotted against the number of clusters.
The ideal number of clusters is the point, where diminishing returns are no longer worth the additional cost, which is defined as the "elbow point ".

## 2.7 Linear regression

A linear regression is a statistical model. It is used to identify a relationship between an independent and a dependent variable. A requirement to perform a linear regression on data is that the variables need to have a linear relation to one another.
To create a relationship a "regression line" is modeled, which serves as a linear predictor function. The line is defined by the equation $y = b_0 + b_1 x$. Y stands for the dependent variable and x for the independent variable. $B_0$ is the y-intercept and $b_1$ is the slope of the line. The minimum sum of the squared differences between the observed values and the predicted values is used to compute the "regression line". The difference between the predicted and real value is called the residual. If the model works well these residuals are normally distributed around the mean value zero.
To further analyze the model, the F-test can be applied. This statistical test compares the computed model to the null hypothesis. The null hypothesis assumes that the dependent variable can be predicted using the mean of the independent variable. The R2-value is an indicator on how much of the variance of the dependent variable can be explained through the independent variable.
The idea for the linear regression model was to see how well the correlation between the mean values of the control group and the mean values of the RNase group can explain a shift between the global maxima. Theoretically, a high correlation between control and RNase should provide no shift between the global maxima.
We created a data frame which contained the correlation of each protein and the absolute difference between the control global maxima and RNase global maxima, which is the number

of fractions the protein shifted in the experiment. We separated this data frame into two. One with 90% of the rows and one with the remaining 10%. We computed a linear regression model on the data frame containing 90% of the rows. This model was then tested using 'predict.lm' on the data frame containing the last 10%. The 'predict.lm' function created a vector with the predicted amount of fractions in the shift dependent on the given correlation.

### 2.8 Comparison to other data bases

The as RDPs identified proteins were compared with the RBP2GO dataset of the DKFZ (Caudron-Herger et al., 2020), allowing us to determine the true positives, false positives, true negatives, and the false negatives among them.

## 3. Results

### 3.1 Data clean-up

**Remove replicates without data**

In total we removed 61 proteins, because at least one of the replicates of these proteins contained no protein amounts.

**Scaling**

After scaling our data, the bar plots of the protein amount in each replicate are the same.

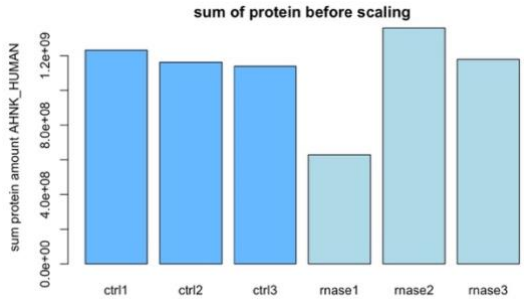


Figure 1: Protein AHNK_HUMAN before scaling in each rep of control and RNase

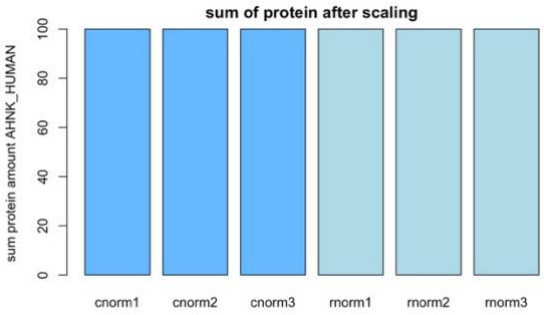


Figure 2: Protein AHNK_HUMAN after scaling in each rep of control and RNase

**Smoothing**

As shown in the following figures the background noise of the plots was reduced and important information was enhanced.

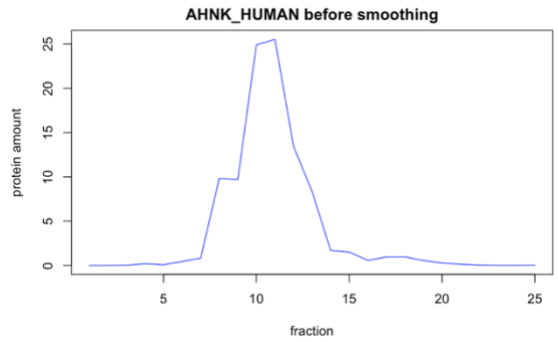


Figure 3: AHNK_HUMAN before smoothing

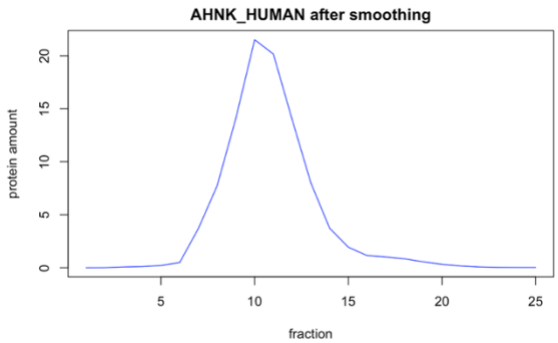


Figure 4: AHNK_HUMAN after smoothing

## 3.2 Correlation

We identified the proteins, that had a correlation coefficient of <0.8. In total, 144 proteins were eliminated.

Having performed this step, we were left with only those proteins, that show a strong correlation between all three replicates and eliminated the proteins, in which 1 or more replicates showed a significant deviation to the other corresponding replicates.

## 3.3 Peak finder

**Global peak finder**

After applying the previously explained function on our data frames, we created two vectors which contained the indices of the fractions of the global maxima. One with the global maxima of the control group and one with the RNase group.

**Local peak finder**

After applying the for loop, the indices of the fractions, in which the local peaks were located, were saved in two vectors. One vector for the control and one for the RNase group.

## 3.4 Identifying shifts

Because of the subtraction of RNase peaks from control peaks we concluded that the proteins that shifted to a lower sucrose density level, are those which have a positive difference in peaks. We saved all these proteins in the vector 'left_shift'. A total of 1029 proteins were found to exhibit a left shift. All the proteins which traveled further along the sucrose density gradient after the RNase treatment have a negative difference in peaks. 392 proteins were identified as right shifting proteins and were saved in the vector 'right_shift'. We set the threshold of more than one fraction to tolerate slight fluctuations. All the proteins which have no difference in peaks or a difference of 1 fraction were defined in the vector 'no_shift'. 3167 proteins did not show a significant shift. Proteins which accumulated in the last fraction in both RNase and control group were categorized as precipitated proteins. A total of 18 proteins were defined as precipitated.

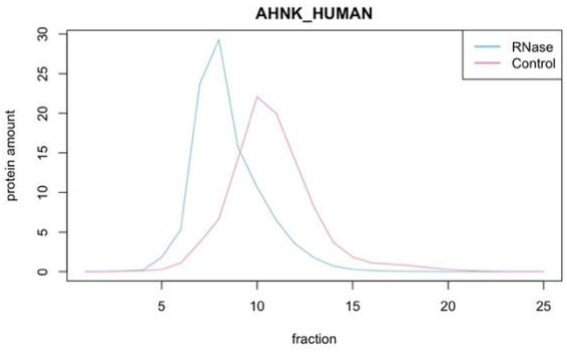


Figure 5: left shift

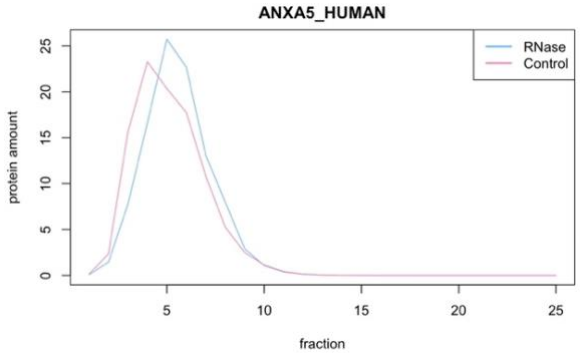


Figure 6: right shift

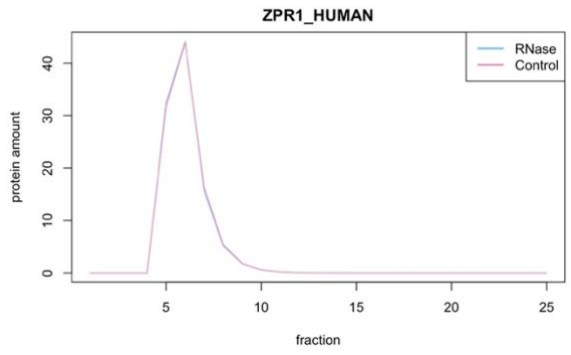


Figure 7: no shift

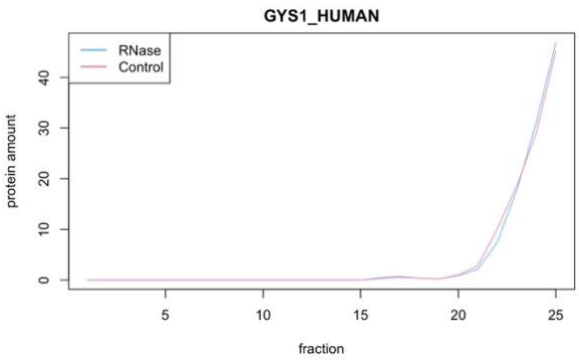


Figure 8: precipitated

## 3.5 T-Test

We identified 77 proteins with at least one replicate of control and RNase with the same amount of protein in the global peak fraction.
We determined our α as 0.05. This left us with 2644 proteins, which had a p-value under that significance level. After adjusting the p-values, we were left with 2214 significant proteins in the mean control global peak fractions.
We performed the t-test on the amount of proteins in the mean RNase global peak fractions as well. Here we also identified 77 proteins with at least one replicate of RNase and control with the same amount of protein in the peak fraction. 2359 proteins were identified as significant after adjustment of the p-value.
The real significant proteins are proteins, which are significant in both t-tests. Therefore, we created a vector with all the proteins which had significant p-value in both tests, leaving us with 1987 significant proteins.

## 3.6 K-means

We chose four as the point, where the diminishing returns are no longer worth the additional cost. As shown in the figure below, the total within-cluster sum of squares does not change significantly from five clusters up.

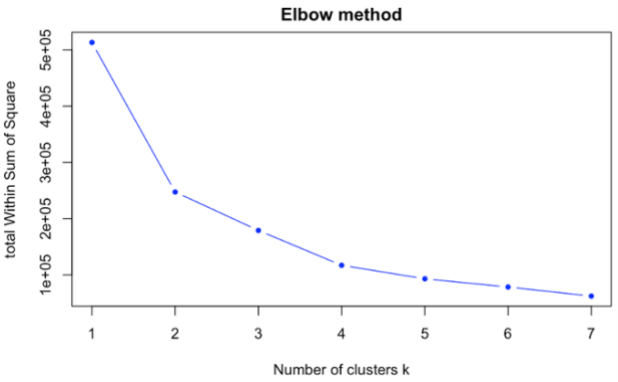


Figure 9: K-means elbow-method

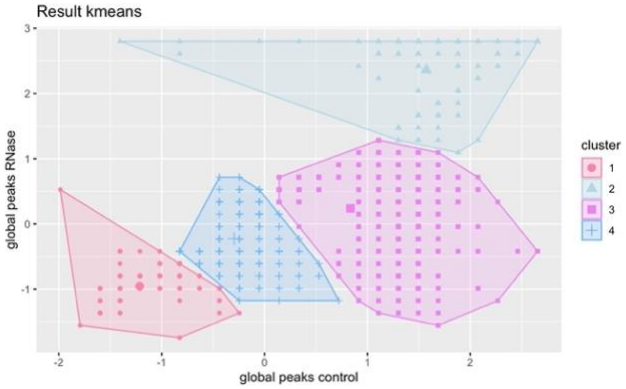


Figure 10: K-means cluster

Figure 10 shows the different clusters. As shown above, dividing our protein into 4 clusters was the most efficient. These clusters group proteins with similar characteristics, regarding their global peak behavior, into 1 group.

## 3.7 Linear Regression

After performing a linear regression on our first 90% of our data frame we received a $R^2$-value of this model of 0.7815, which means that 78.15% of the variance in the shifts can be explained by the correlation between the RNase and control group. The p-value of the F-test is $2.2*10^{-16}$, which means, that the null hypothesis can be rejected. By looking at the data, we concluded that this model predicts shifts with a lower number of fractions between the RNase and control global peaks better than the shifts with a higher number of fractions in shifts.

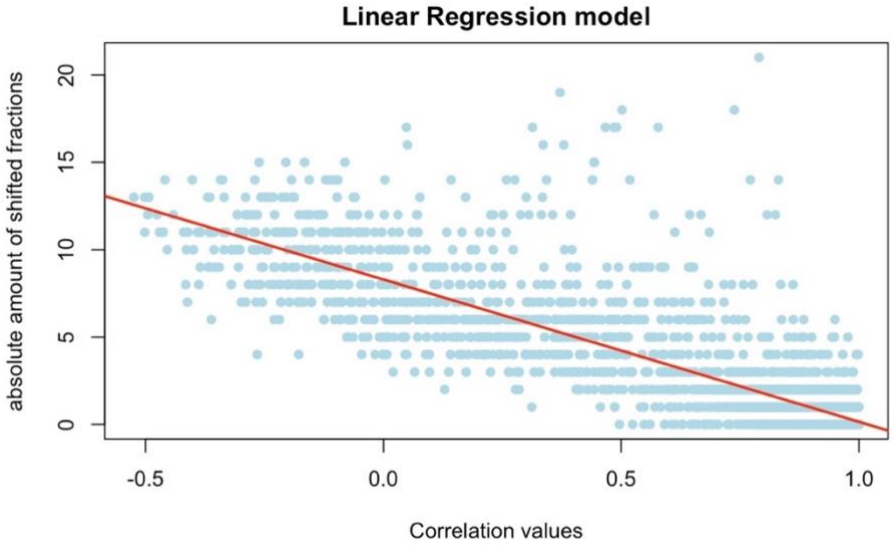


Figure 11: Linear Regression model

To analyze the model, we looked at the residuals. To determine the distribution, we analyzed the residuals graphically through a histogram and a Q-Q plot.

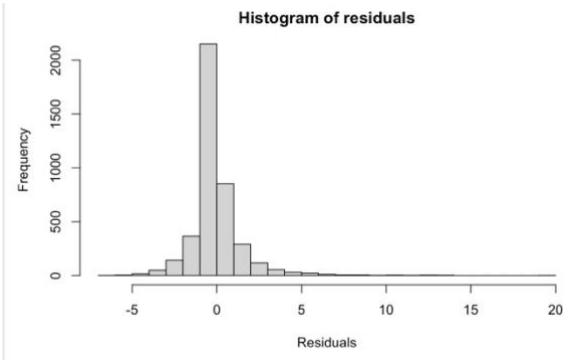


Figure 12: Histogram of the residuals

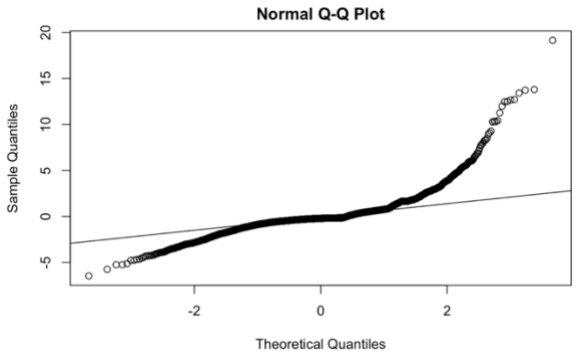


Figure 13: Q-Q plot of the Residuals

After looking at the graph we came to the conclusion, that the residuals are normally distributed with outliers in the shifts with a higher amount of fractions between RNase and control global peak.

We plotted the predicted values from the 'predict.lm' function against the real values of the shifts to judge our model in respect to its ability to predict the shifts dependent on the correlation of unknown data.

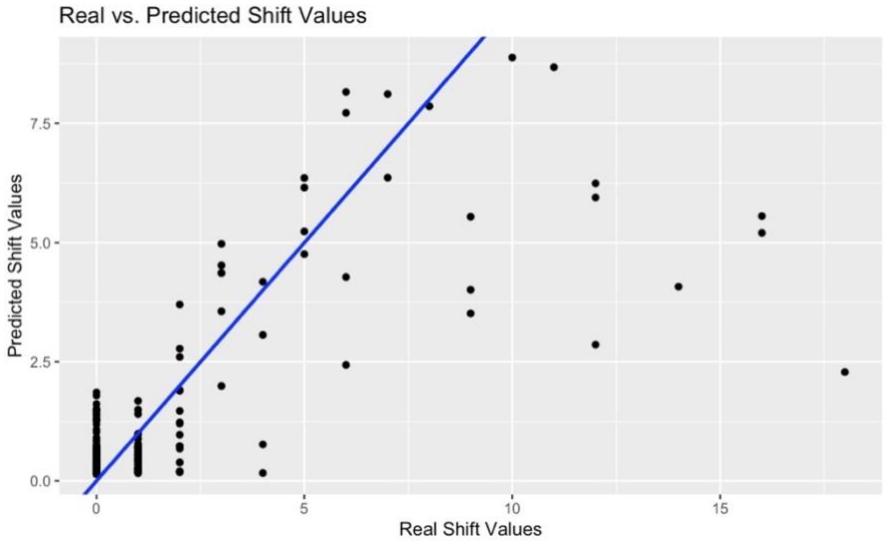

Figure 14: predicted vs. real shifts

Looking at this plot we concluded that the model can predict low shifts relatively well but shifts with a higher amount of fractions are not very well predicted.

### 3.8 Comparison to other data bases

Through the significance testing 1163 RDPs were identified. A comparison with the RBP2GO database revealed that 1093 proteins were correctly identified as RNA-dependent proteins and are therefore true positives. This leaves us with 70 false positives. 2332 proteins were falsely identified as not RNA dependent by us but are RDPs according to the RBP2GO database. Those are the false negatives. Since 3425 proteins were found in our data base as well as in the RBP2GO database the false positive rate equals 7.8%, which is relatively low. However, the false negative rate reaches 65.9%, which is higher than we would have liked to.
The test sensitivity, or also called the true positive rate is rather low with 34.0%. The test specificity, or also called true negative rate equals 92.2%

## 4. Discussion

The goal of the project was to identify RDPs as accurately as possible. To achieve this, the false positive and false negative rate should be minimized. Comparing our results to the RBP2GO database, we observed that the false positive rate was low, indicating a good performance. However, the false negative rate was relatively high, indicating room for improvement.
The normalization and scaling part of our analysis went as expected and as discussed in the project proposal.
The false negative rate could be adjusted by including local peaks in the analysis to verify more shifts between RNase and control group. Furthermore, the analysis could be expanded through a gaussian fit model. This could reduce background noise even further and would allow the identification of more accurate global peaks and therefore, would allow a deeper knowledge about the shifting behavior.
Also, choosing a higher significance level for the t-test might have led to the identification of more RDPs. However, this would have also increased the false positive rate and have had a negative impact on our model. Therefore, a higher significance level is not beneficial overall.
To improve our linear regression model, a multiple linear regression could be performed. This would help to explain more of the variance in the shifting behavior.

To identify these variables a dimension reduction analysis such as PCA could be performed to gain more insight on the explanatory variables.
After all these modifications, a standardized and easy-to-use algorithm could be developed. This would be beneficial to make the identification of RDPs easier in the future.
The identification of RDPs is crucial for future research in various diseases and therefore could help to save the lives of many patients.

# References


Caudron-Herger, M., Jansen, R.E., Wassmer, E., and Diederichs, S. (2020). RBP2GO: a comprehensive pan-species database on RNA-binding proteins, their interactions and functions. Nucleic Acids Research *49*, D425-D436. 10.1093/nar/gkaa1040.

Caudron-Herger, M., Rusin, S.F., Adamo, M.E., Seiler, J., Schmid, V.K., Barreau, E., Kettenbach, A.N., and Diederichs, S. (2019). R-DeeP: Proteome-wide and Quantitative Identification of RNA-Dependent Proteins by Density Gradient Ultracentrifugation. Molecular Cell *75*, 184-199.e110. https://doi.org/10.1016/j.molcel.2019.04.018.

Gerstberger, S., Hafner, M., and Tuschl, T. (2014). A census of human RNA-binding proteins. Nature Reviews Genetics *15*, 829-845. 10.1038/nrg3813.

Kim, T.K. (2015). T test as a parametric statistic. kja *68*, 540-546. 10.4097/kjae.2015.68.6.540.

Li, T., Hui, W., Halike, H., and Gao, F. (2021). RNA Binding Protein-Based Model for Prognostic Prediction of Colorectal Cancer. Technology in Cancer Research & Treatment *20*, 15330338211019504. 10.1177/15330338211019504.

Liu, J., and Cao, X. (2023). RBP–RNA interactions in the control of autoimmunity and autoinflammation. Cell Research *33*, 97-115. 10.1038/s41422-022-00752-5.

Liu, M.-j., Guo, H., Jiang, L.-l., Jiao, M., Wang, S.-h., Tian, T., Fu, X., and Wang, W.-j. (2021). Elevated RBP-Jκ and CXCL11 Expression in Colon Cancer is Associated with an Unfavorable Clinical Outcome. Cancer Management and Research *13*, 3651-3661. 10.2147/CMAR.S298580.

Ludbrook, J. (2013). Should we use one-sided or two-sided P values in tests of significance? Clinical and Experimental Pharmacology and Physiology *40*, 357-361. https://doi.org/10.1111/1440-1681.12086.

Wang, E.T., Taliaferro, J.M., Lee, J.A., Sudhakaran, I.P., Rossoll, W., Gross, C., Moss, K.R., and Bassell, G.J. (2016). Dysregulation of mRNA Localization and Translation in Genetic Disease. J Neurosci *36*, 11418-11426. 10.1523/JNEUROSCI.2352-16.2016.