Report

Khalida Dushimova, Richard Efraim Langi, Madleen Piegsa, Greta Karathanos

2022-07-12

```
library(knitr)
```

1. Introduction

RNA-binding proteins are proteins that bind to the double or single stranded RNA in cells and by this interact with transcripts of RNA-driven processes. This large group of proteins plays the key roles in RNA processing and modification (Alternative splicing, RNA editing and polyadenylation), export, mRNA localization and translation. The malfunction of RBP´s underlies the origin of many diseases from muscular atrophies and neurological disorders to cancer. (Referenz) Therefore, it is necessary to increase the number of recognized RBPs and the understanding of their molecular mechanisms. Since identifying molecular mechanisms is one of the biggest challenge in (nc)RNA (-> non coding), research M.Caudron-Herger et al. created R-Deep with the aim to detect RBPs automatically. In general, R-Deep is a proteome-wide, unbiased, and enrichment-free screen based on density gradient ultracentrifugation. Performing R-Deep, lysates with a previous RNase treatment (RNase group) and without a previous RNase treatment (Ctrl group) were separated by their density. Each fraction was analyzed by mass spectometry or western blot according to their protein amount. True-positive detected RNA-dependent proteins can bind directly to the RNA (RBP) or bind to RBPs.

2. Materials and Methods

2.1 RDeep: R-DeeP is a proteome-wide, unbiased and enrichment-free screen. The principle of R-DeeP is based on cellular lysate fractionation by density gradient ultracentrifugation. (Referenz) The outcome is analyzed by proteome-wide mass spectometry or individual western blotting. In general, R-DeeP is used to determine RNA-dependent protein, that interact directly or indirectly with RNA. Lysates with (RNase group) and without (Ctrl group) RNase treatment are compared. By that differences in molecular weight, hence and the size of the complexes are determined.RBPs are expected to split after RNase treatment and migrate to different fractions in a sucrose density gradient.

2.2 Dataset exploration: Our given dataset consists of mass-spectometry data from non-synchronized A549 cells which are human lung carcinoma cells of a caucasian male. In general, the data show the protein amount of each of our 3680 human proteins per fraction. The RDeep screen has been repeated three times so it comprises three replicates for each sample. All in all, we got 3680 rows with one human proteins per row and 150 columns for our Ctrl and RNase group for 25 fractions and three replicates each.

2.3 Fractionwise Normalization: The total amount of protein of every replicate should be similar. In our given dataset this was not the case. Normalization is the process to account for the bias and make samples more comparable. Our aim was to change the values of columns to a common scale without distorting differences in the range of values and by this reducing the variation between our three technical replicates. Therefore, for our first normalization we compute the sums fractionwise and find the two closest sums. We then define normalization factors for the replicates. The quotients of the mean sums and sums of replicates are the normalization factors.

2.4 Anti-outlier function: For our second normalization, we defined the mean out of the two most similar replicates and by this wrote an own function to exclude outliers. The outliers were found, removed, and replaced by the mean of the other two most similar replicates.

2.5 Peaks identification: Peaks We expect RNA dependent proteins to migrate to migrate to different positions

in the RNase treated sample compared to the Ctrl sample. As a next step we identify the maxima in both samples to characterize a protein as RNA dependent or RNA independent. We detect the local and the global maxima to draw all the biological information we could possibly get from our given values. Our theoretical maximum is a point x whose neighboring values have to be smaller.

The method is based on checking the neighbours of each fraction. For the first fraction, only the right neighbour will be compared because there is no left neighbour. For the second till twenty fourth fraction, both left and right neighbours will be compared. For the last fraction, only the left neighbour will be compared since right neighbour doesn't exist. If the the checked fraction have higher values than the neighbour(s), then the fraction is a maxima. If the the checked fraction have lower values than the neighbour(s), then the fraction is a not a maxima and will be zero-ed in our code.

2.6 Criteria for RNA dependency:

2.6.1 T-Test of absolute maxima: T-Test of global maxima T-Test shows us if the global maxima of each of the proteins (all replicates) between Ctrl and RNase group has a significant difference. We declare a significant difference between both groups as RNA dependent. It will be our first criteria for RNA dependency.

2.6.2 K-Means Clustering (Y-shift and X-shift detection): K-Means Clustering is an unsupervised non-linear algorithm that clusters data based on similarity. Its aim is to partition the observations into a previously definded number of clusters. We define this number by performing the elbow method. Through K-Means Clustering and comparing our t-test results, we define selection criteria (Y-shift and X-shift) to determine RNA dependent proteins. Y-shift is the difference of protein amount between the global maxima of Ctrl and RNase group. X-shift is the difference of locations of global maxima (fraction) between the global maxima of Ctrl and RNase group.

If the X- and Y-shift has positive values (left and down shift), we define the protein as RNA dependent. If the X- and Y-shift has negative values or close to zero (right and up shift), we define the protein as RNA independent. These shifts will be our second criteria for RNA dependency. We also combined our X- and Y-shifts in a dataframe and used the elbow method to determine the optimal number of clusters.

2.6.3 T-Test of local maxima: To find more RNA dependent proteins, we will include a 3rd criteria to classify the protein as RNA dependent. The 3rd criteria is the result of t-test of local maxima.

2.7 Comparison with other databanks: In order to calculate the true positive and true negative rate and see how good our criteria find the RNA dependent proteins we compare our finding with mammalian RNA-binding protein resources (https://r-deep.dkfz.de)

2.8 Linear regression: Finally, we perform a linear regression. This model can generally be used to model the relationship between a dependent variable (regressand) and one or more explanatory variables (predictors). The relationship is represented by a linear function. We can see from the model to what extent the dependent variable can be explained by the other variables. Our regression model predicts the Y-Shift values with the information from correlation between Ctrl and RNase.

2.9 Working with complementary data: We used additional data bank from our tutor, Niklas Engel for our second linear regression. The other one is from RDeep website and is used to compare our results with it.

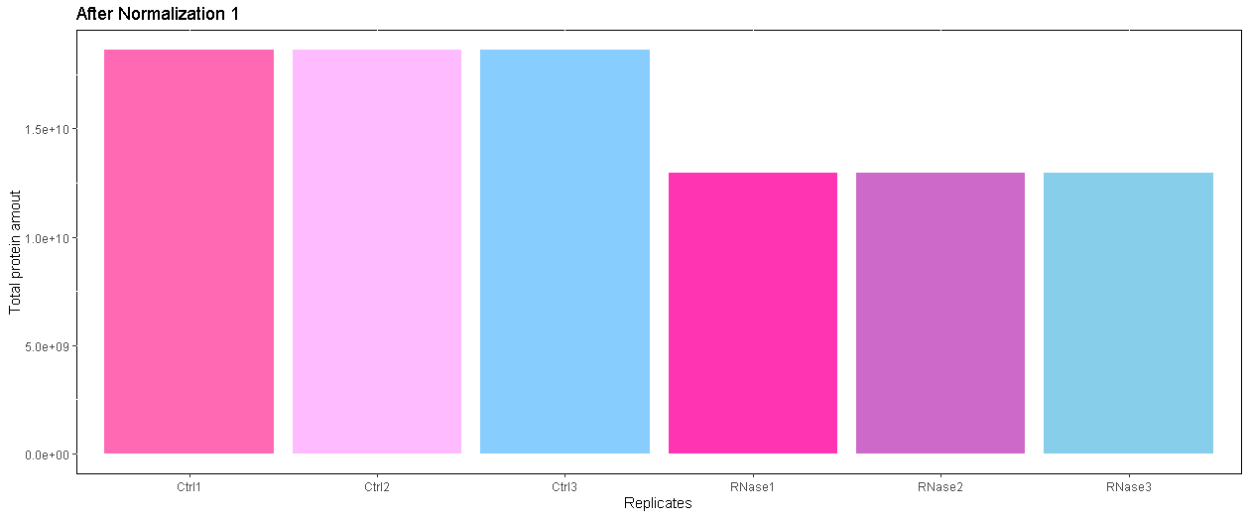
3. Results

3.1 Data Reorganization The raw data was splitted into two data frames, one for Ctrl group **Ctrl** and one for RNase **RNase** group. Each group consists of 25 rows representing the fractions and 11040 columns representing the 3680 proteins including the 3 Reps.

3.2 Data Evaluation For further analysis of the replicates we summed up the total protein amount of all genes per replicate and ploted them side by side in a bar chart. The total protein amount between between each Reps of Ctrl and RNase samples are very variable.

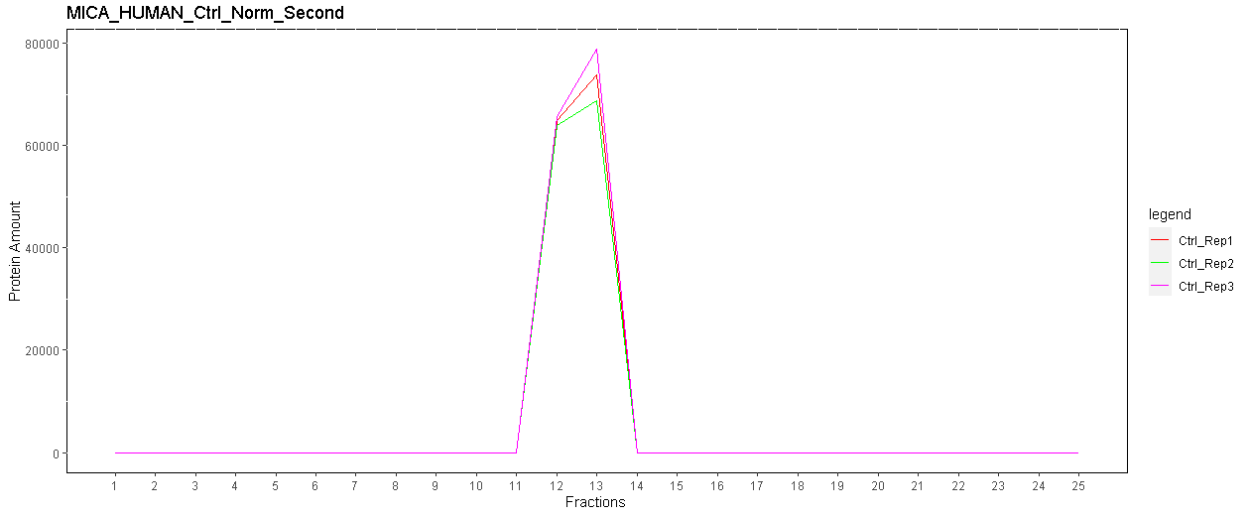
3.2 First Normalization The results of our first normalization can be found in 2 data frames, **Norm_Ctrl** for Ctrl group and **Norm_RNase** for RNase group, each with 11040 rows representing the proteins with 3

Reps and 25 columns for the fractions. To check the results of our normalization, a bar chart of total protein amount in y axis and each Rep of Ctrl and RNase the was created.



The bar chart revealed that the three replicates of Ctrl and RNase have the same total amount. The chart also showed that although the protein amount of all three replicates of Ctrl and RNase is equal and there is a difference between the total amount of Ctrl and RNase.

3.3 Second Normalization The results of our second normalization from our **df_combi_function** can be found in 2 data frames, **tCtrl_combi_df** for Ctrl group and **tRNase_combi_df** for RNase group, each with 11040 columns representing the proteins with 3 Reps and 25 rows for the fractions. To check the results of our normalization, 3 graphs of 3 reps of a Protein **MICA_Human_Ctrl** before and after normalization with fraction in x axis and the protein amount in y axis were produced.



The 3 grpahs of normalized replicates revealed that the replicates have similar amount of protein and there are no more significant outliers left.

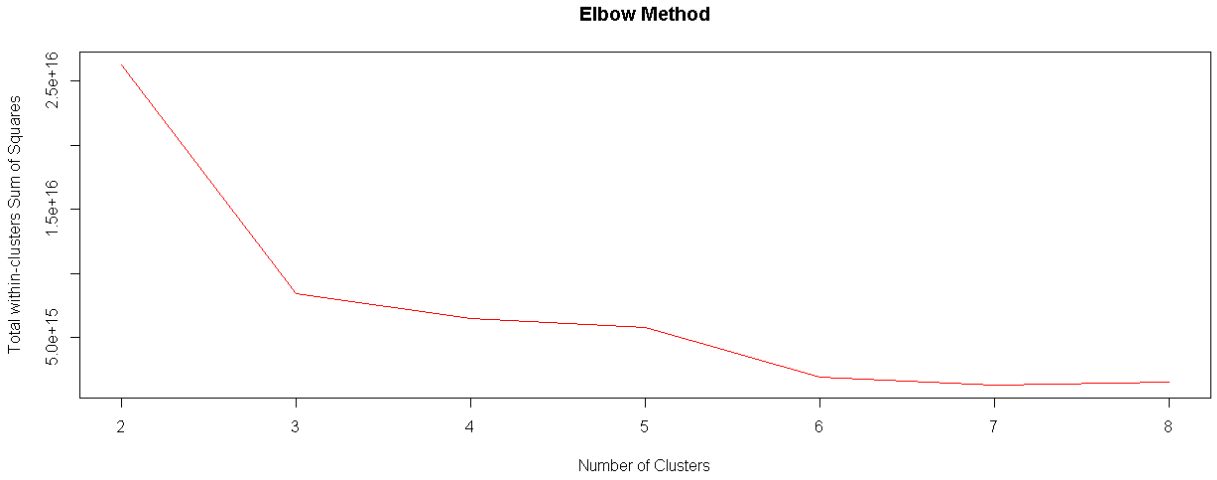
3.4 Peaks Identification: Using our self-written **maximafunction**, 9 data frames were received, each of them had a different threshold. **maxima_Ctrl_i** is the result for Ctrl group and **maxima_RNase_i** for RNase group, where **i** represents the percentage of global maxima (threshold). **i** has values between 0.1 - 0.9. The data frames consisted of 11040 columns represented proteins inlcuding reps and 25 columns for fractions. The values of protein amount of global (absolute) and local maxima were presented whereas not maxima are zero-ed. Our self-written **maxnum_plot_col** produced a plot, which ploted a random protein with threshold in x-axis and number of maxima in y-axis.

3.5 Criteria for RNA dependency

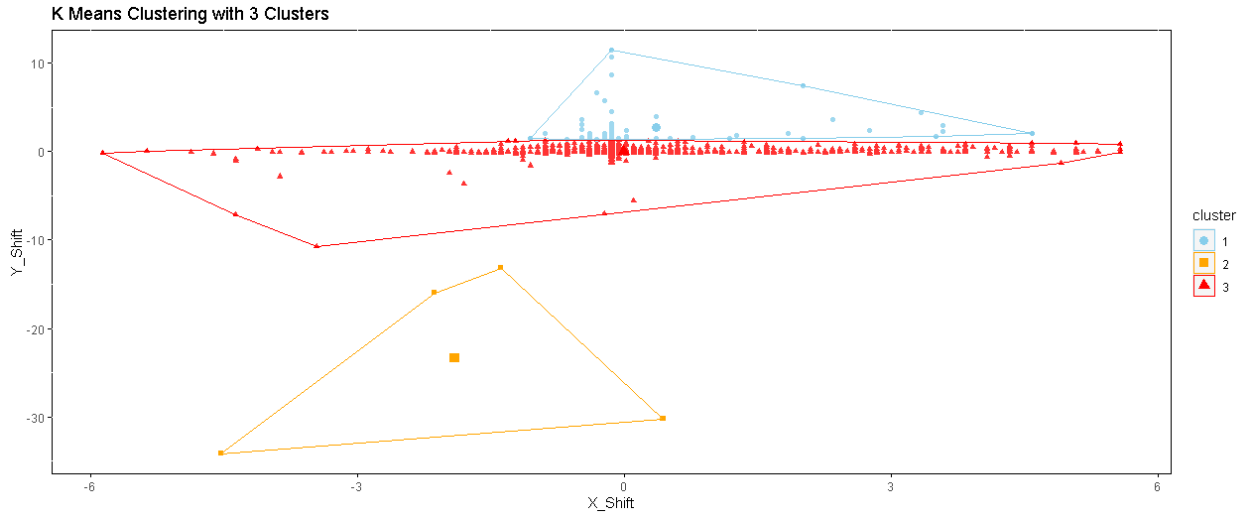
3.5.1 T-Test of Global and Local maxima: A data frame **test** with 3680 rows and one column was produced after performing t-test of global maxima between Ctrl and RNase group. The rows represented the names of the proteins and were alphabetically ordered. The column represented the RNA dependency with two kinds of results, TRUE or FALSE.

For t test of local maxima, the results were presented in a data frame **test__0.4** that consists of 3680 rows and 1 column. The name of the proteins were shown in the rows and the only column explained the RNA dependency. Three outcomes were generated, TRUE, FALSE or NA.

3.5.2 K-Means Clustering To find the optimal number of clusters, a plot of elbow method was used, where the x axis represented the number of clusters and y axis the total within clusters sum of squares. The elbow method revealed a hard kink for three clusters.



The result of k means clustering was shown in **km** with 72.1% as a proprotion of between and total sum squared. A plot of y shift in y axis and x shift in x axis was created to better visualize the cluster. The second cluster had the most data points, then the first cluster and lastly the third cluster with interestingly only 4 data points. What is noticeable is that the second cluster has mostly positive x and y shift values.



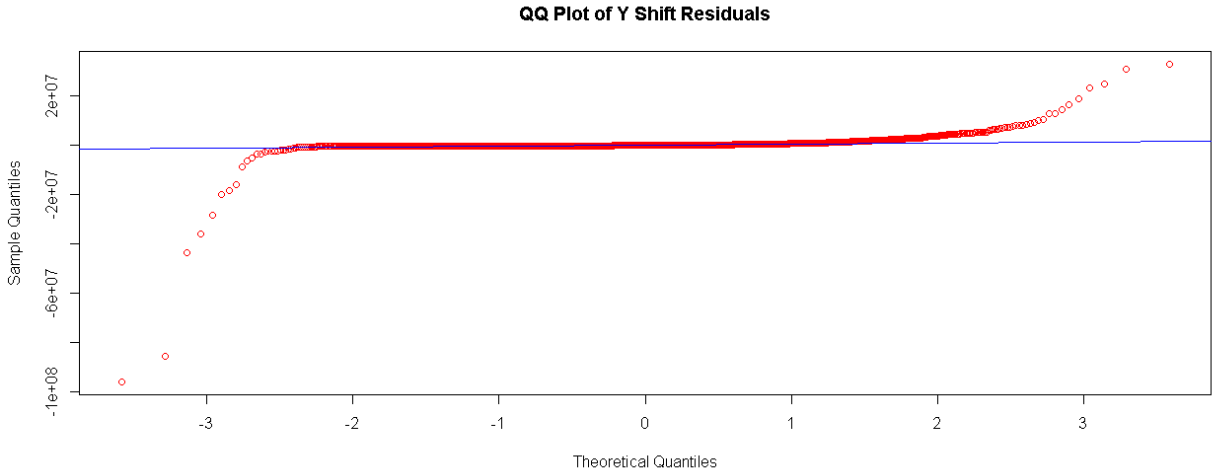
3.5.3 Comparison with Data Bank with two Criteria The results of the RNA dependent proteins (2 Criteria) were presented in 2 data frames. 305 proteins **Ctrl__Dependent__Abmax__1** were identified when one

of the two criteria was fulfilled and 63 proteins **Ctrl_Dependent_Abmax_2** when both criteria were fulfilled.

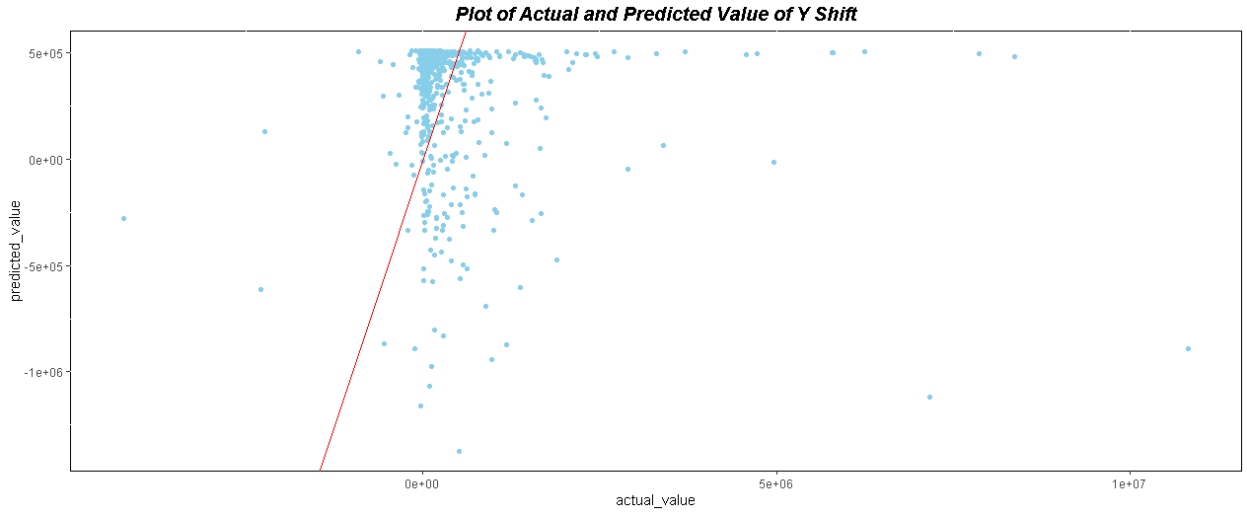
To check how many RNA dependent proteins were "correctly" identified, our results were compared with a data bank **table_RBP_lists.csv** from RDeep website. Out of 63 identified RNA-dependent proteins that fulfill both criteria, 57 had a match with the comparable databank (**Rdeep2**) and for one criteria fulfilled out of 305 identified proteins 228 havd match (**Rdeep1**). Unfortunately our applied criteria could not identify 1860 RNA dependent proteins (**Not_identified_RDeep_1**).

3.5.4 Comparison with Data Bank with three Criteria The results of the RNA dependent proteins (3 Criterias) were presented in a data frame **dependent_3** where 472 proteins were classified as RNA-dependent. Out of 472 identified RNA-dependent proteins that fulfill either criteria, 350 had a match (**RDeep3**). Unfortunately our applied criteria could not identify 1738 RNA dependent (**Not_identified_RDeep_3**).

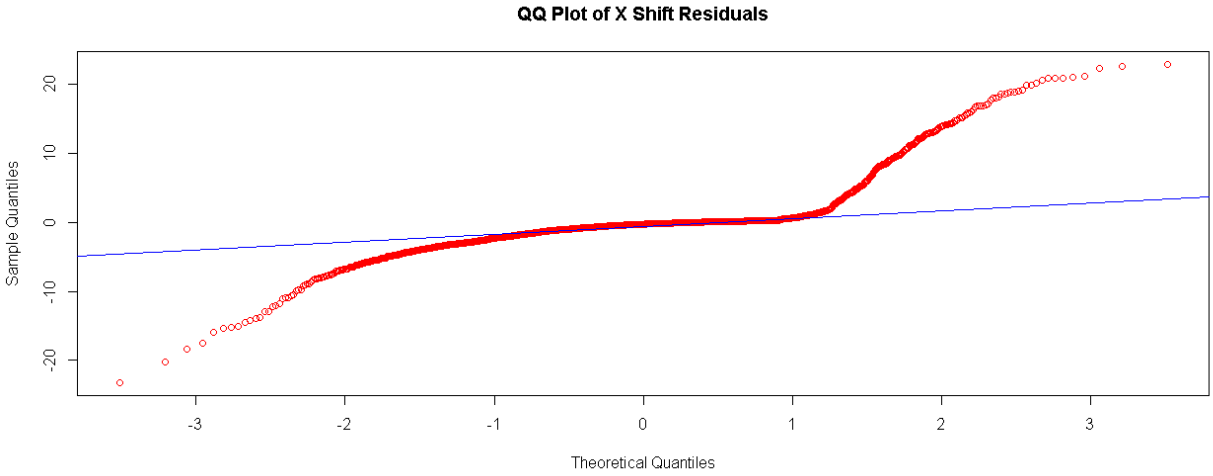
3.6 Linear Regression The first linear regression was performed between correlation of Ctrl and RNase absolute maxima for x axis and y shift for y axis. The linear equation from the regression was $y = 1690907x - 1184175$. The p value for the predictor variable y was $7.98e^{-06}$ and for x is $5.96e^{-09}$. The p value of our model was $5.962e^{-09}$ (**lm**). The QQ plot of the theoretical and sample residues aligned with the QQ line except when the theoretical quantiles were below -2.5 and above 2.5.



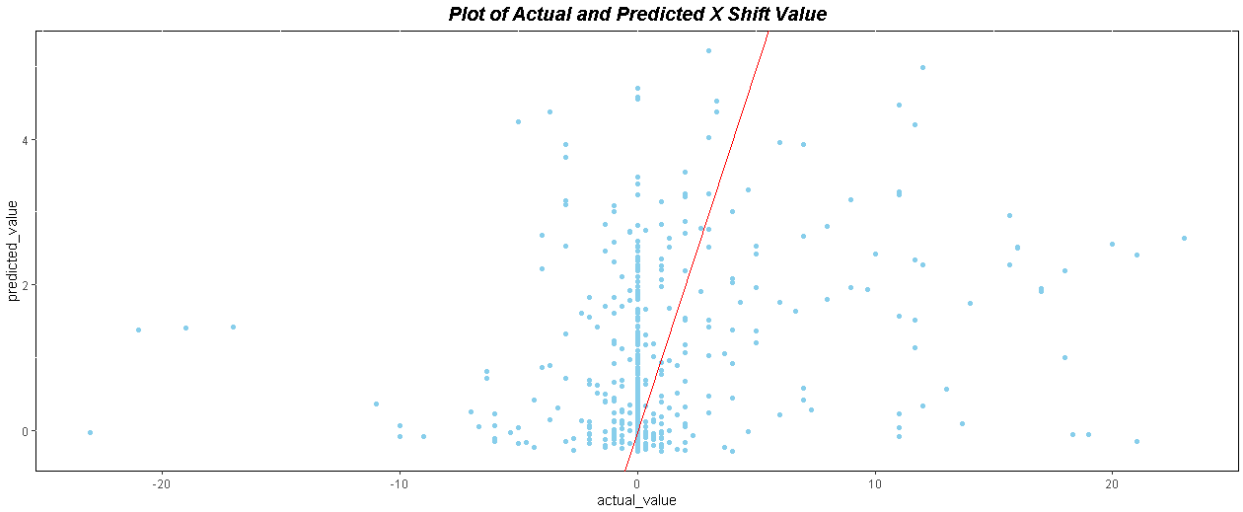
The last plot was a plot with actual value in x axis, predicted value in y axis and a linear line with intercept 0 and slope 1. The data points gather mostly on the place where the predicted and actual values are above 0. Below 0 there were only few points that are near the linear line. Nonetheless, there were also some points that are very far from the linear line.



The second linear regression is performed between RBP2GOScore in x axis and x shift in y axis. The linear equation from the regression is $y = 0.069630x - 0.368093$. The p value for the predictor variable y is 0.00285 and for x is $2e^{-16}$. The p value of our model is $2.2e^{-16}$ (**lm_4**). The QQ plot of the theoretical and sample residues aligned with the QQ line except when the theoretical quantiles were below -1.5 and above 1.



The last plot was a plot with actual value in x axis, predicted value in y axis and a linear line with intercept 0 and slope 1. The data points gathered mostly near the linear line when the actual value wass between -2 and 2 and the predicted value is around 0.5.



4. Discussion

4.1 Normalization (Fractionwise Normalization and Anti Outlier Function): The normalization worked as expected. The total amount of protein per replicate was equal for all Ctrl replicates and equal for all RNase replicates, which made the replicates good comparable. The **df_combi_function** function did exactly what we wanted. The outliers were reduced and did not have a significant effect on our data after the normalization.

4.2 Peaks Identification: The self-written **maximafunction** detected the global and local maxima for different thresholds. In order to decide which threshold is the best for the local maxima we ran our **maxnum_plot_col** function, which ploted a random protein with threshold in x-axis and number of maxima in y-axis. Having ran the function several times, we decided to use a threshold value of 40% of the global maxima. Our chosen threshold of 40% was enough to get only significant maxima and reduce the effect of noisy data.

4.3 Criteria for RNA dependency:

4.3.1 Criteria for RNA dependency, both T-Test (global maximum) and K-Means (X- and Y-shift for global maxima): The elbow method revealed that the optimal number of clusters was three: RNA-dependent proteins cluster (X-shift and Y-shift are positive), RNA-independet proteins cluster (almost no Y-shift or negative) and another RNA-independet proteins cluster with high outliers (negative Y-shift).

We firstly defined two criteria for RNA-dependency. The k-means clustering of the Y-Shift and X-Shift and the T-Test significance for global maxima. Although our defined two criteria could find RNA-dependent proteins with high true positive rate, the false negative rate was still too high.

4.3.2 Comparison with Data Bank with two Criteria: In order to find the proteins that fulfill the criteria for RNA-dependency, we compared the T-test results for significant difference of the global maxima for Ctrl and RNAse and the K means clustering (Y-shift and X-shift detection) with the databank. The comparison with the table of mammalian RNA-binding protein resources has shown that our applied criteria finds the RNA-dependent proteins with very low false positive rate, but very high false negative rate.

4.3.3 Comparison with Data Bank with three Criteria: To find more RNA dependent proteins, we included a 3rd criteria, T-test results for local maxima to decrease the false negative and increase the true positive rates. As can be seen from the table the false negative rate could be decreased but unfortunately our criteria were still not exact enough and the false negative rates are still way too high.

1. +2. criteria

true positive	75%
true negative	11%
false positive	25%
false negative	89%

1. +2. +3. criteria

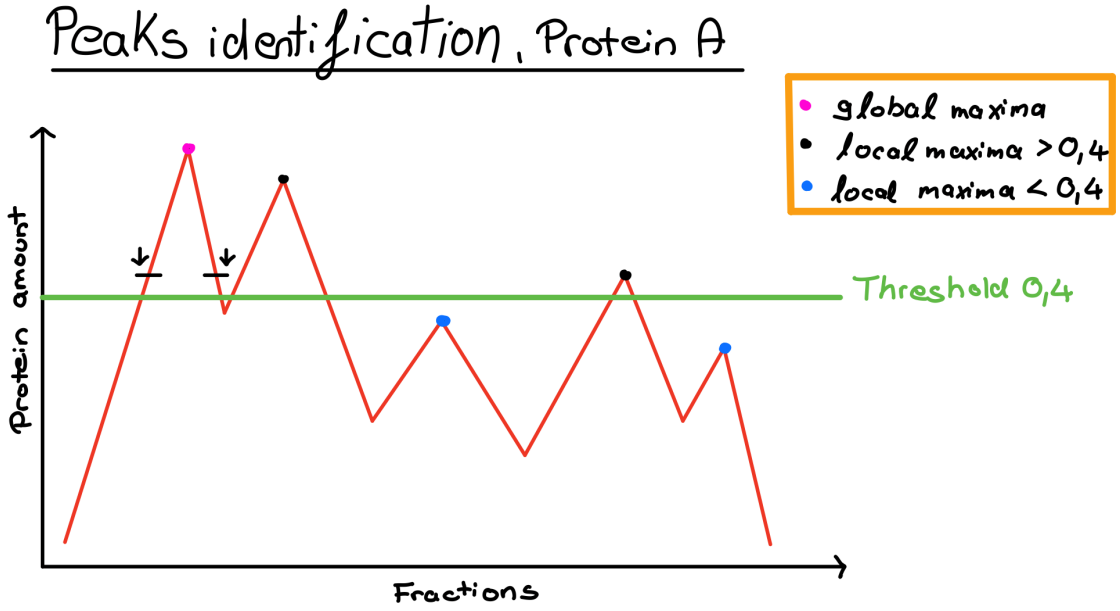
true positive	74,2%
true negative	16,2%
false positive	25,8%
false negative	83,2%

4.4.1 Linear Regression between Y-shift and Correlation

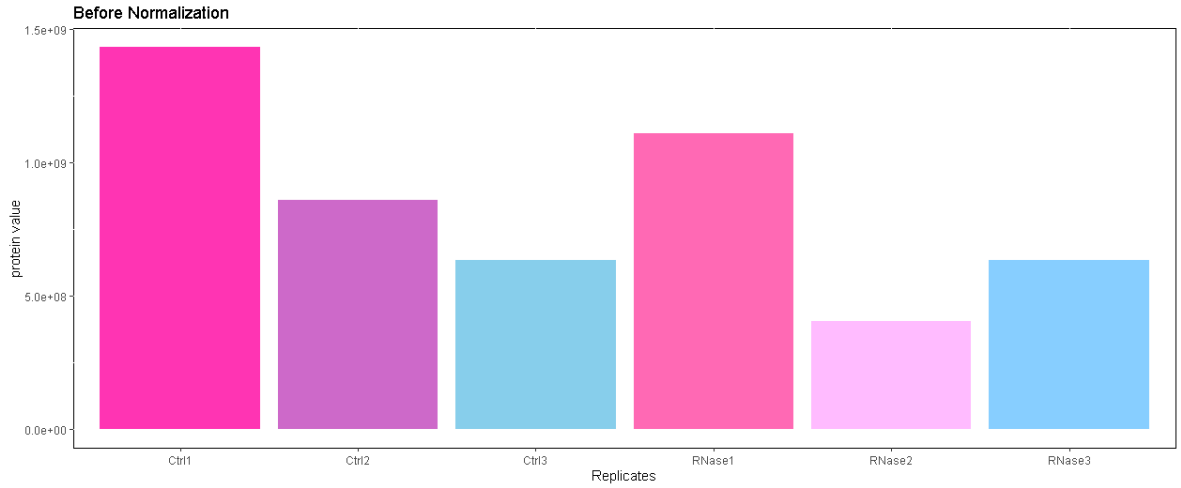
4.4.2 Linear Regression between X-shift and RBP2GO_Score

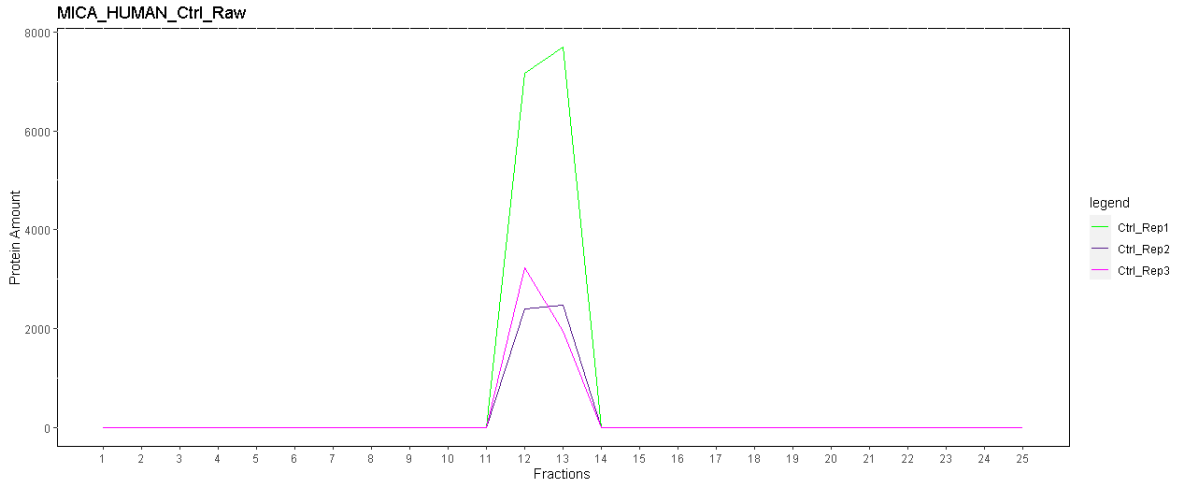
5. Conclusion One possible approach for better results could be using more replicates, define different criteria or maybe try different databanks, since we only compared our findings with one databank.

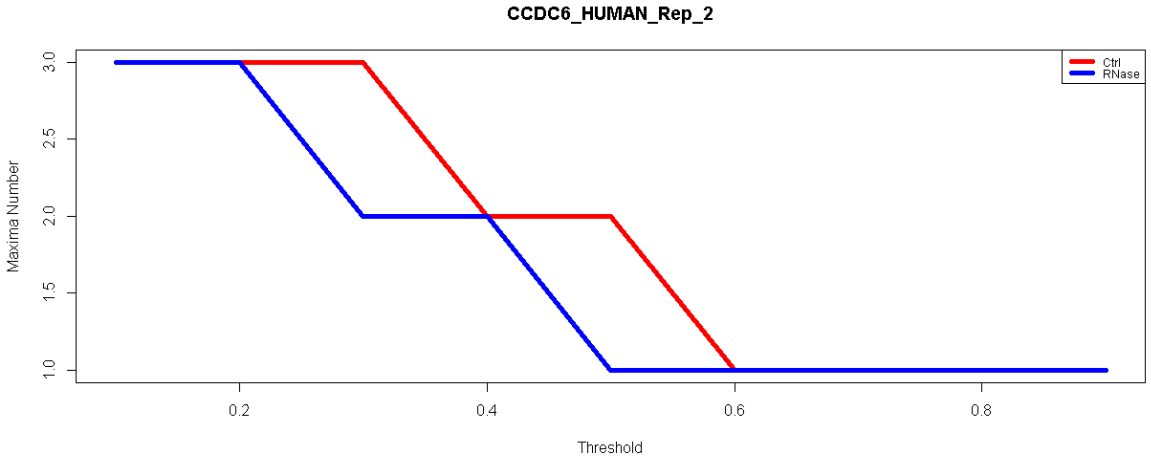
6. References

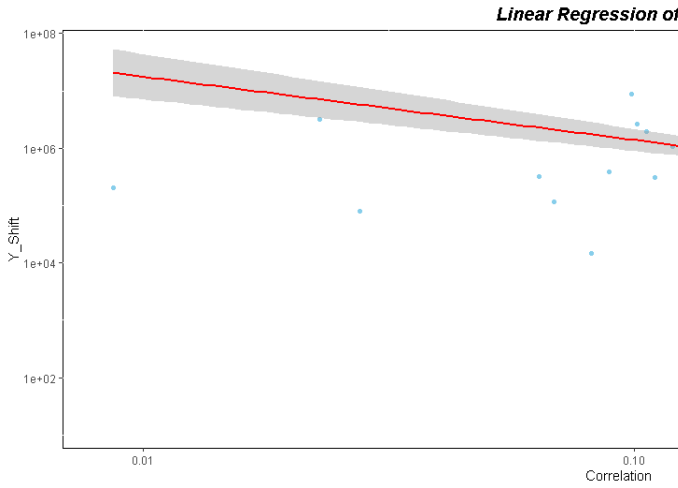


7. Anhang

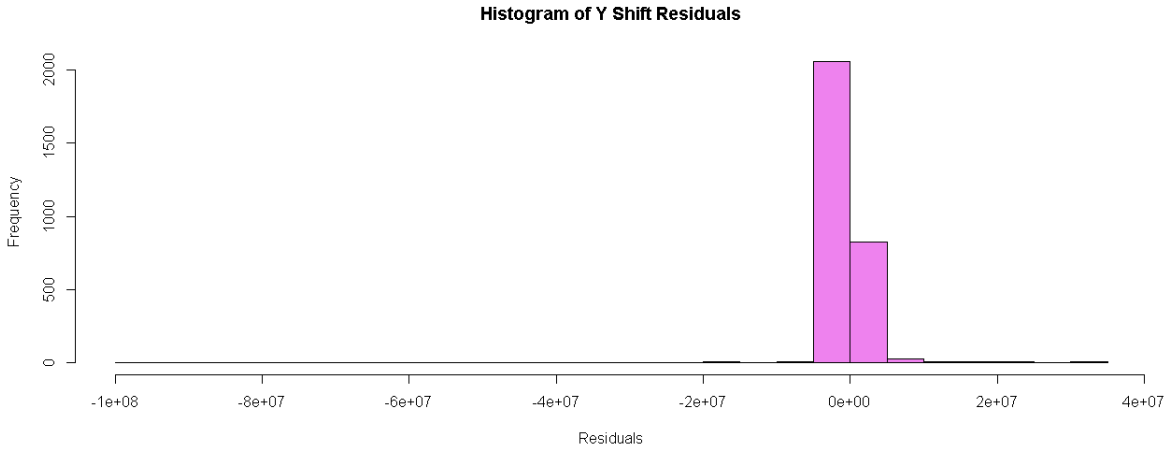

Before Normalization
protein value
1.5e+09
1.0e+09
5.0e+08
0.0e+00
Ctrl1
Ctrl2
Ctrl3
RNase1
RNase2
RNase3
Replicates

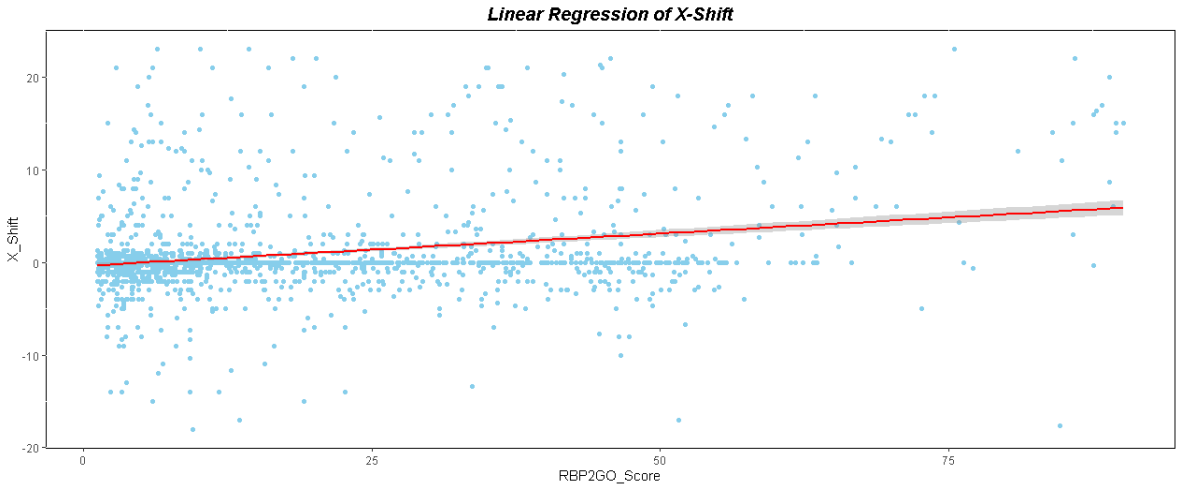

MICA_HUMAN_Ctrl_Raw
Protein Amount
8000
6000
4000
2000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
Fractions
legend
Ctrl_Rep1
Ctrl_Rep2
Ctrl_Rep3


CCDC6_HUMAN_Rep_2
Maxima Number
3.0
2.5
2.0
1.5
1.0
0.2
0.4
0.6
0.8
Threshold
Ctrl
RNase




```
Within cluster sum of squares by cluster:
[1] 2.275488e+15 2.637920e+15 3.487322e+15
 (between_SS / total_SS =  72.1 %)
```




Linear Regression of X-Shift
X_Shift
RBP2GO_Score
20
10
0
-10
-20
0
25
50
75