

*non-synchronized A549 cells*

# **Final Report: Proteome-wide Screen for RNA-dependent Proteins**

Anastasia Möller, Johannes Schadt, Sylviane Verschaeve, Tine Limberg

17.07.2023

## **Contents**

1. Introduction . . . . .	1
2. Methods . . . . .	3
4. Discussion . . . . .	11
5.Outlook . . . . .	14
6. Literature . . . . .	14

Loading the data:

## **1. Introduction**

RNA-binding proteins (RBPs) constitute one of the largest families of proteins in the cell, with over 4000 RBPs identified to date (Gebauer et. al,2020). In addition to the examination of proteins that directly bind RNA, this analysis encompasses RNA-dependent proteins whose interactome relies on RNA, even without direct binding (Caudron-Herger et al., 2019). Unlike proteins with classical RNA-binding domains, these proteins can engage with RNA through intrinsically disordered domains (Corley et al., 2020). For the sake of simplicity, the term “Rdeep” will be used in this report to refer to both RNA-binding and RNA-dependent proteins. These proteins play a crucial role in controlling various aspects of RNA life, function, and efficiency,

thereby acting as essential regulators in numerous cellular processes. Rdeeps are involved in extensive regulatory networks that govern critical processes, including transcription, splicing, RNA modification, intracellular trafficking, translation, and decay (Corley et al., 2020). The significance of Rdeeps in human health is underscored by their involvement in a wide range of diseases. Mutations in genes encoding Rdeeps have been identified as the underlying cause of various disorders, leading to malfunction and tissue-specific defects. Rdeeps are particularly implicated in diseases of the nervous system and cancers, making them promising targets for therapeutic interventions. Remarkably, the prevalence of RBP mutations in diseases is substantial, with nearly one-third of Rdeeps implicated in various pathologies, encompassing over a thousand disease-related Rdeeps identified thus far. In mendelian disorders, Rdeeps outnumber other classes of proteins, including transcription factors, in terms of the prevalence of mutations (Gebauer et. al,2020). To comprehensively elucidate the involvement of Rdeeps in translational control and their roles in disease pathogenesis, further investigations are required. Understanding the mechanistic interplay between Rdeeps and RNA in cellular processes holds great promise for developing targeted therapeutic strategies to rectify RBP-related dysfunctions. Therefore, the goal of this analysis is, to identify which proteins in the given dataset are Rdeeps. The dataset to be analysed contains 3680 different proteins from synchronized mitotic A549 cells. At the end of this project, a linear regression will be developed, which will make it possible to predict RNA dependence for proteins based on their distribution of the control and RNase treated sample.

### **1.1. Importance of RNA-binding proteins**

### **1.2. Experimental Setup**

To collect the mass spectrometry data, a strict protocol was followed. The cells were centrifuged and lysed. One sample was treated with RNase, while the other served as the untreated control. Both samples were divided into 25 fractions using a sucrose gradient. Ultracentrifugation was performed, allowing the proteins to distribute based on their density. For statistical relevance, the protocol included three repetitions of the experiment. Therefore, triplicates are available for each protein for both Control and RNase. The fractions were then subjected to mass spectrometry analysis to determine the protein abundance, measured in arbitrary units (Caudron-Herger et al., 2019). Furthermore, it should be noted that the protein amount is represented by the y-value whilst the number of the fraction is the x-value.

## 2. Methods

### 2.1. Data cleanup

### 2.2. Reproducibility

The reproducibility of the replicates (rep) for the Control and the RNase group is calculated separately by computing the pearson correlation between rep1 and rep2, rep1 and rep3 as well as between rep2 and rep3. There are 2 scenarios where a protein is seen as not reproducible. Firstly, if a replicate of a protein has only zeros, its correlation can not be calculated (NA) and the protein is discarded. 83 proteins are affected. Moreover, if all 3 correlations either in the Control or RNase group of one protein are below 0.9 the protein isn't reliable enough for further analysis. Thus 523 proteins are additionally deleted, resulting 3074 proteins from initially 3680 proteins are left for further analysis. Some proteins contain two replicates similar to each other (correlation < 0.9) and a third one that completely differs. Knowing that these proteins have one high and two smaller correlations, the deviating replicate was set to NA and will be ignored when uniting the replicates per protein. Consequently, important data is preserved without losing too many proteins.

### 2.3. Normalization methods and Reduction/Scaling

Since each normalisation method has advantages and disadvantages, we apply three different methods to the dataframe. The **mean-value method (mvm)** is the first data normalisation we use. The mean protein amount of each protein is subtracted from the protein amount in each fraction. The values that are zero or smaller than the mean become negative through this subtraction. We set these negative values to zero to simplify further analysis. Afterwards, the sum of the protein amount in each row is scaled to 100.

Furthermore we used the **z-transformation**, which transforms our data into a standard normal distribution with a mean of 0 and a standard deviation of 1 by using the formula:  $Z = (X - \mu) / \sigma$ . To avoid negative values and not lose too much information as we did with the mean-value method, the smallest value per protein is added to each fraction of the corresponding protein, meaning the smallest value is now 0. In case of the z-transformation we first normalized then scaled to 100. Afterwards we reduced by calculating the mean between the replicates with a higher correlation than 0.9 and scaled again to 100.

The last used normalization method is **Min-Max scaling (mms)**, which is a very simple scaling-method where the normalized value  $x'$  is calculated from the original value  $x$  as

follows:  $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$ . This means that the highest value is automatically set to one and the lowest values to zero. With this method it is very easy to calculate the global peaks, but on the other hand, the protein amount, so the area beneath the graph cannot be normalized.

## 2.5. Gaussian fit

Goal of the gaussian fit is to fit a gaussian function to the data points, in our case the protein amount in each fraction of one protein. Therefore, we establish a list in which the parameters, which describe the distribution are saved. To fit the parameters to our data we used the `optim()` function implemented in R.

## 2.6. Data description via Parameters

To identify whether a Protein is RNA dependent or not each protein is tested on 4 parameters. The goal of those 4 parameters is the identification of significant differences between the Control and RNase sample. In the following sections the parameters will be elaborated.

**2.6.1. Parameter 1: Significant change of protein amount under global peak** The first parameter identifies a significant change of the protein amount under the global peak for each protein. The global maximum represents the fraction of the sucrose gradient containing the highest protein amount. It is determined by using the *which.max()* function for each protein. If the Protein amount of the global peak fraction is either in the Control or RNase sample 1.7 times higher than the other sample, we defined it as a significant change.

**2.6.2. Parameter 2: Significant change of protein amount under local peaks** The next parameter recognizes a significant change in the total protein amount under the local peaks between the Control and RNase sample for each protein. New local peaks can occur if after RNase treatment the protein either dissociates or gains new interaction partners. To detect a significant change the local peaks have to be identified. To be defined as a local peak four criteria have to be fulfilled. First, the y-value of the local peak fraction has to be higher than the y-value of its neighbor fractions. Afterwards we checked if the sd of the local peak's y-value and its neighbors is higher than the sd of the y-values which contain less than 8 % of the total protein amount (*sd.threshold*). The aim is to sort out small fluctuation between the y-values. The third criteria selects the relevant local peaks by sorting those out which have a smaller protein amount than 3 % of the total protein amount. Most of the already mentioned

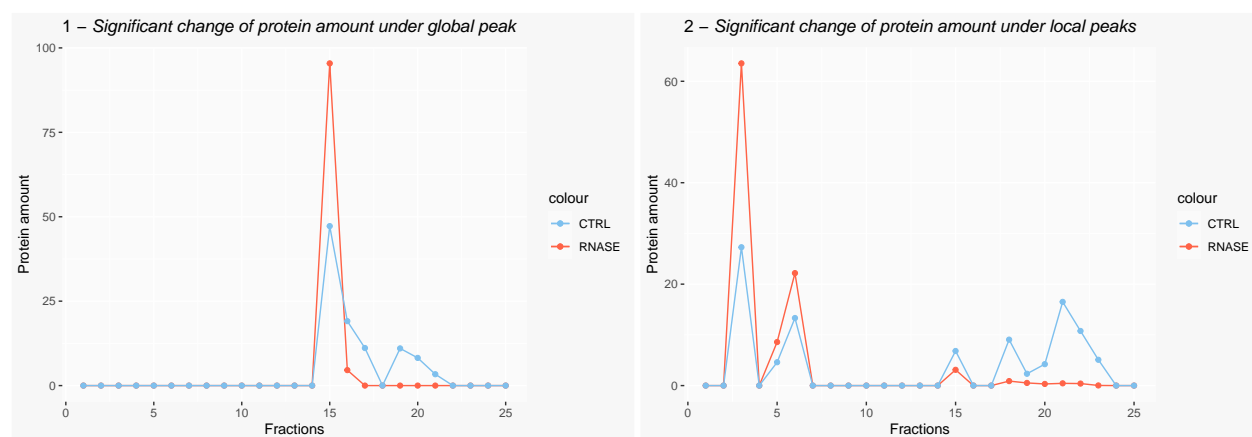
criteria fit to the global peaks as well and thus global maxima have to be removed. The total protein amount under the local peaks for each protein in the Control and RNase sample is of interest. If it differs more than 7.5 from each other a significant change of the protein amount under the local peak is present.

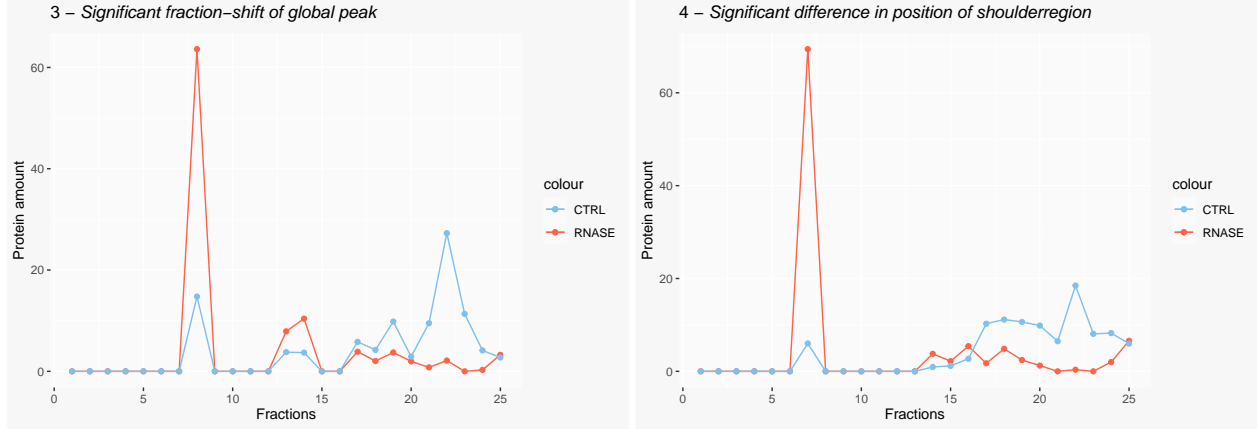
**2.6.3. Parameter 3: Significant fraction-shift of global peak** The third parameter focuses on the x-axis depicting the fractions. If the fraction of the global peak in either the Control or RNase sample differs more than two fractions in the positive or negative x-direction compared to the other sample, the protein is defined as RNA-dependent. It lost or gained an interaction partner due to the RNase treatment.

**2.6.4. Parameter 4: Significant difference in position of shoulderregions** At last it is observed if shoulderregions occur or disappear after the RNase treatment. A shoulderregion contains more than 2 consecutive fractions with a sd less than the *sd.threshold*. On the contrary to the local peak identification the fractions with small fluctuations are sorted. Of interest are those fractions belonging to shoulderregions which either occur in the Control or RNase sample but not in both. Often parts of the shoulderregions are overlapping, resulting in shoulderregions of interest with less than 3 consecutive fractions. So, a shoulderregion of interest is only significant if it has three or more consecutive fractions.

The following code creates two plots in one diagram, so that the Control and RNase group can be compared.

```
## Warning: Paket 'ggplot2' wurde unter R Version 4.2.3 erstellt
```





**2.6.5. Boundaries and Precipitated proteins** The local peakfinder can't find local maxima at the boundaries (fraction 1 and fraction 25) because they have only one neighbor fraction. A local peak at the boundary has to fulfill following requirements. Firstly, it has to have a higher y-value than the two succeeding fractions. Secondly, the protein amount of those three fractions has to be bigger than 10. Smaller protein amounts are not relevant enough. Also precipitated proteins are not described by the parameters and have to be identified separately. They have a global peak in fraction 25 and their total protein amount of 100 has to be split between fraction 23, 24 and 25.

## 2.7. K-means clustering

Kmeans is used to group a set of data points of a  $d$  - dimensional space into a certain number  $k$  of clusters. Each cluster has a center, the so called centroid. At the beginning of the clustering process  $k$  number of clusters are set randomly. Then, the following steps have to be repeated again and again ( $n$ -times), until no further change can be observed: 1. The points are assigned to the cluster to whose centroid they are closest to. The distance is commonly the euclidean distance. 2. Thus, the centers of the clusters change and the centroids move. In our case we grouped our proteins in a two-dimensional space (fraction of control-peak and fraction of rnase) into  $k = 4$  clusters. We performed kmeans to have an alternative Method for identification of RNA-dependent Proteins, besides our Parameters.

## 2.8. Regression analysis

Linear regression can predict the value of a variable (dependent variable) based on the value of another value (independent variable). The beginning 2/3 of the dataset is used to train the

regression model and the remaining third is tested by the model. As independent variable the correlation between the Control and RNase group over all fractions per protein is used to predict the global shift amount [fractions] of the tested data.

The goal of linear regression is to model the mathematical relationship between a dependent variable and an independent variable. The objective is to create a linear equation, typically represented as  $y = mx + n$ , which best fits the observed data. The residuals, calculated as the difference between the actual y-values and the predicted y-values, provide an indication of the model's accuracy. We use the correlation of the protein amount in Control and RNase as the independent variable and the amount of the global peak shift in fractions as the dependent variable. This relationship is crucial for understanding the effects of equality of protein levels in Control and RNase on global peak shifts. Using linear regression, we can quantify and analyse this relationship. As a second model we use the same independent variable for each protein but used the information of our parameters as the dependent variable. Both models are trained with 2500 proteins, so that a prediction can be made for 574 proteins.

### **3. Results**

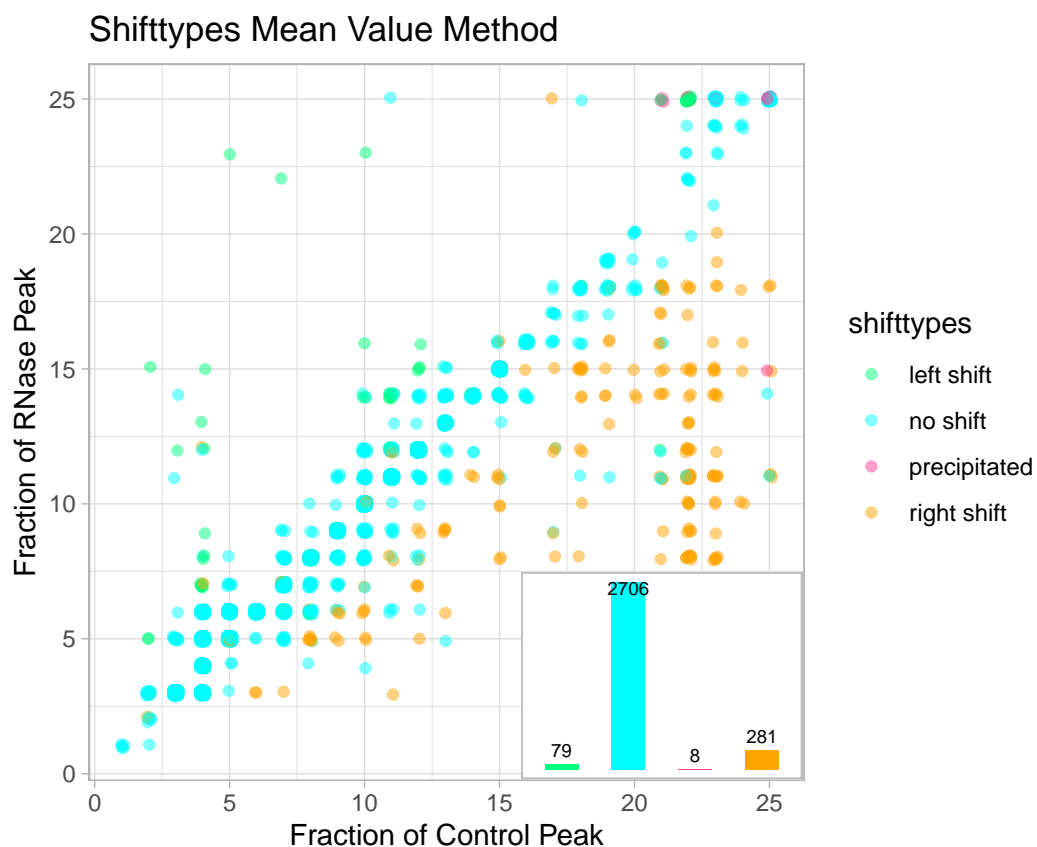
#### **3.1. Cleaned Dataset**

#### **3.2 Gaussian fit**

On the grounds of our lists, containing the parameters for the gaussian distribution for every protein, it is possible to plot the gaussian curves. The control and RNase curves can be plotted separately or together in one plot for better comparison. Because we use the `optim()` function and do not implement further parameters ourselves, the distribution does not show local peaks or shoulder regions. Therefore, the gaussian fit is only used for visualisation not for further analysis. **Graph ???**

#### **3.3. RNA-dependent Proteins**

Using just parameter 1 (global peaks), we were able to characterize the shift-type, so whether the peaks show no shift, a right shift, a left shift, or are precipitated. The following graph visualizes our results for MVM:

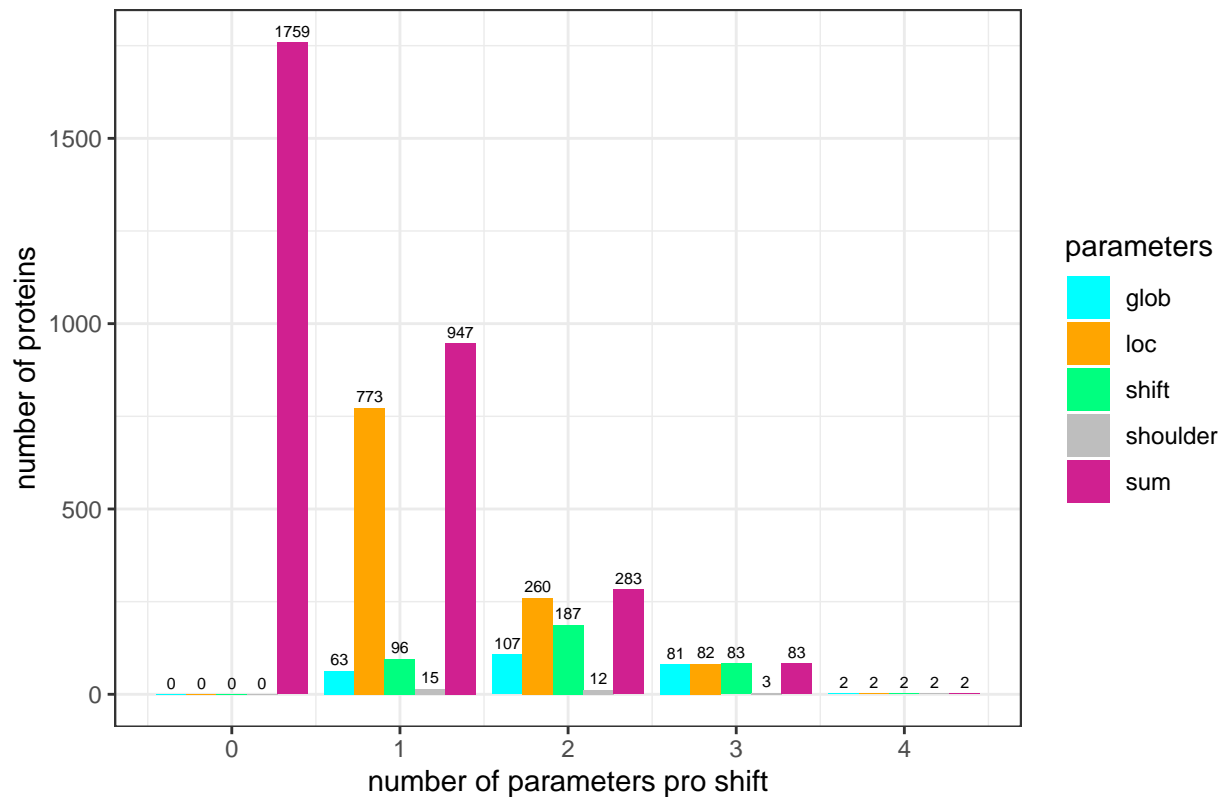


Using only parameter 1 we were able to identify 368 RNA-dependent Proteins with MVM, 349 with z-transformation and 320 with MMS.

Using our parameters, we were able to identify 464 RNA-dependent proteins with MVM, 468 with z-transformation and 396 with MMS. The following plot shows the significance of the different parameters and how they they contributed to the characterization for MVM. All proteins that have two positive parameters and all proteins that have solely a global shift are classified as RNA-dependent.



### How many parameters support the proteins? MVM



### 3.4. Comparison of the normalization methods

### 3.5. K-means clustering

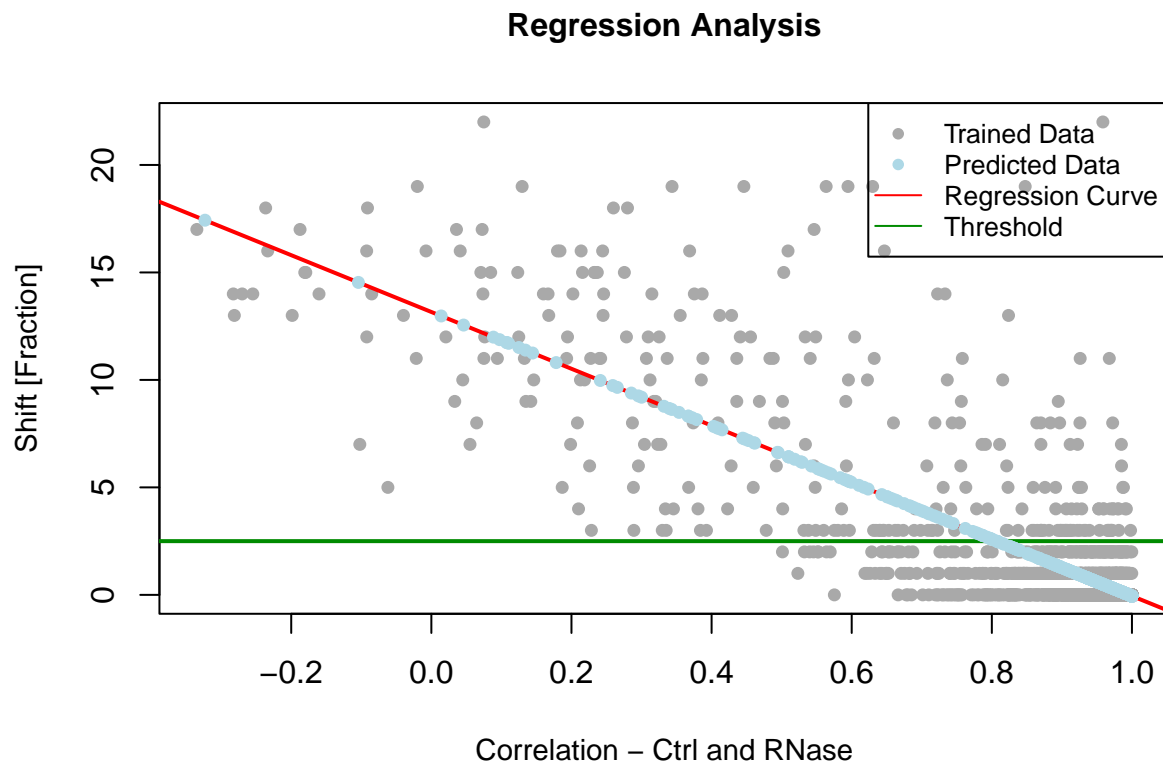
We want to cluster the Proteins depending on their global control peaks and their global RNase Peaks. To find out how many proteins would be optimal in theory we used the elbow method that showed us, that two clusters would be optimal. But looking at the biological background, two clusters would not help us to identify RNA-dependent proteins, but rather group them into heavy and light proteins. To gain useful results from kmeans, we forced it to create four clusters. Kmeans clustered the proteins as followed:

We chose the cluster at the bottom right for every normalization method, to be RNA-dependent (for MVM this would be cluster two). The other clusters can't be identified as RNA-dependent or not. There are no visible differences between the RNA-dependent proteins clusters of the different normalization methods. Using kmeans we were able to identify 160 RNA-associated Proteins for MVM, 155 for z-transformation and 159 for Min-Max-Scaling.

	mvm	zt	mms
True Positives	142	108	115
False Positives	17	6	13
True Negatives	114	125	118
False Negatives	300	334	327

### 3.6. Regression analysis

The linear regression was performed as previous mentioned under the topic methods. Each protein is depicted as a dot in the plot below. As expected the majority of the trained data, marked in grey shows a high correlation with a shift under 2 fractions. On the other hand those with a small correlation have shift upwards of 2 fractions. As already decided for the parameter **Significant fraction-shift of global peak**: a protein is classified as RNA-dependent protein if it has a shift higher than two fractions. The threshold is marked as green line in the plot.



The tested data, marked in blue could now be grouped in RNA-dependent protein or not RNA-dependent protein. AUSWERTUNG RATEN

	mvm	zt	mms
FNR	0.6787330	0.7556561	0.7398190
FPR	0.1297710	0.0458015	0.0992366
Precision	0.8930818	0.9473684	0.8984375

	mvm	zt	mms
True Positives	422	417	356
False Positives	39	47	37
True Negatives	695	687	697
False Negatives	1905	1910	1971

### 3.7. Comparison with Database

We compared our results with the RDeep Database and a table containing non-RNA-associated Proteins from another paper (Quelle?). The following table shows how well the Method via our parameters worked. The table on the left shows the number of true positives, false positives, true negatives and false negatives. The table on the right shows the false negative rate (FNR), the false positive rate (FPR) and the precision.

The following tables do the same for our kmeans results:

The following tables do the same for our results that only depended on our global shift:

There are thirteen proteins, that are not present neither in the RDeep data set, nor the table containing non-RNA-associated Proteins from the paper. The following tables show the names of these proteins and whether they are considered as RNA-dependent by our analysis or not. On the left side are the results using our parameters, on the right side the ones obtained by kmeans:

## 4. Discussion

Our goal was to identify RNA-dependent proteins using the data provided by mass-spectrometry. To do this, we used several methods to find out which variant of our analysis was the best. Overall, we had a very high false negative rate, independent of the methods we used. The lowest was 81.9%. This could be, because our criteria were too strict (e.g. **w**).

	mvm	zt	mms
FNR	0.8187	0.8208	0.8470
FPR	0.0531	0.0640	0.0504
Precision	0.9154	0.8987	0.9059

	mvm	zt	mms
True Positives	152	145	151
False Positives	7	9	7
True Negatives	727	725	727
False Negatives	2175	2182	2176

	mvm	zt	mms
FNR	0.9347	0.9377	0.9351
FPR	0.0095	0.0123	0.0095
Precision	0.9560	0.9416	0.9557

	mvm	zt	mms
True Positives	339	315	336
False Positives	26	31	31
True Negatives	708	703	703
False Negatives	1988	2012	1991

	mvm	zt	mms
FNR	0.8543	0.8646	0.8556
FPR	0.0354	0.0422	0.0422
Precision	0.9288	0.9104	0.9155

mvm	shift?	zt	shift?	mms	shift?
PBIR3_HUMAN	0	PBIR3_HUMAN	0	PBIR3_HUMAN	0
MPP6_HUMAN	0	MPP6_HUMAN	0	MPP6_HUMAN	0
F207A_HUMAN	1	F207A_HUMAN	1	F207A_HUMAN	1
CCD58_HUMAN	0	CCD58_HUMAN	0	CCD58_HUMAN	0
BZW2_HUMAN	0	BZW2_HUMAN	0	BZW2_HUMAN	0
UBIM_HUMAN	1	UBIM_HUMAN	1	UBIM_HUMAN	1
ACOC_HUMAN	0	ACOC_HUMAN	0	ACOC_HUMAN	0
PHB_HUMAN	0	PHB_HUMAN	0	PHB_HUMAN	0
UTRO_HUMAN	0	UTRO_HUMAN	0	UTRO_HUMAN	0
RT36_HUMAN	1	RT36_HUMAN	1	RT36_HUMAN	1
DIEXF_HUMAN	0	DIEXF_HUMAN	0	DIEXF_HUMAN	0
BZW1_HUMAN	0	BZW1_HUMAN	1	BZW1_HUMAN	0
WDR92_HUMAN	0	WDR92_HUMAN	0	WDR92_HUMAN	0
13	3	13	4	13	3

mvm	shift?	zt	shift?	mms	shift?
PBIR3_HUMAN	0	PBIR3_HUMAN	0	PBIR3_HUMAN	0
MPP6_HUMAN	0	MPP6_HUMAN	0	MPP6_HUMAN	0
F207A_HUMAN	0	F207A_HUMAN	0	F207A_HUMAN	0
CCD58_HUMAN	0	CCD58_HUMAN	0	CCD58_HUMAN	0
BZW2_HUMAN	0	BZW2_HUMAN	0	BZW2_HUMAN	0
UBIM_HUMAN	0	UBIM_HUMAN	0	UBIM_HUMAN	0
ACOC_HUMAN	0	ACOC_HUMAN	0	ACOC_HUMAN	0
PHB_HUMAN	0	PHB_HUMAN	0	PHB_HUMAN	0
UTRO_HUMAN	0	UTRO_HUMAN	0	UTRO_HUMAN	0
RT36_HUMAN	1	RT36_HUMAN	1	RT36_HUMAN	1
DIEXF_HUMAN	0	DIEXF_HUMAN	0	DIEXF_HUMAN	0
BZW1_HUMAN	0	BZW1_HUMAN	0	BZW1_HUMAN	0
WDR92_HUMAN	0	WDR92_HUMAN	0	WDR92_HUMAN	0
13	1	13	1	13	1

But it has to be taken into account that the method used in the experiment is not able to detect all RNA-dependent proteins. The RDeep Database uses the results of many different papers that used all sorts of different experiments to identify RNA-dependent proteins. So it is impossible for our results to have a low false-negative rate. (**unterschiedliche Zellstadien?**) More important for us would be to look at the false-positive rate and precision of our results: Through all methods, normalization via mean-values method was the best. It led to the lowest false-negative rates, the lowest false-positive rates and the highest precision. z-transformation seems to have been the worst of the three methods. So, mean-value method is the normalization method that should be used when analyzing this sort of data. (**Why?**) The following applies to our MVM-results, but is similar to our z-transformation results and MMS results. If we look at the results of kmeans, our parameters and global shift alone, we observe that kmeans is the most precise, but also leaves out a lot more RNA-dependent proteins that were detected by other methods. This is because, we only chose one cluster, that probably contained right shift proteins, leaving out all precipitated and left shifting proteins. To increase the number of proteins found this way, we would have to create more clusters and choose them accordingly. Looking at the false negative rate of the results of only the global shift, it is clear that the global shift already detects more RNA-dependent proteins than kmeans, but also has a higher false positive rate.

Still, the low false-positive rate of our global shift results is quite low, and validates our decision to qualify proteins that solely have a global shift, but no other positive parameter, as RNA-dependent. If we look at the results of our four parameters, we see the same pattern. We found way more RNA-dependent proteins, but the false positive rate increased as well. In

summary: If we want to be sure that the proteins we identify as RNA dependent, really are RNA dependent, we should choose kmeans. But we have to embrace the possibility that we miss a lot of proteins. If we want to detect many proteins, we should choose our parameters, but take into account that our precision will go down by about 4-5% depending on the normalization method. Looking at the global shift as sole parameter, would be a compromise between precision and quantity. But even with kmeans clustering, our false positive rate is at about 1%, so the results should be verified by another method. Using our methods, we were able to detect 13 new proteins that are not part of the RDeep Database, of which 1-4 are RNA-dependent, depending on the method used. These proteins could be analysed further with additional experiments, and be added to RDeep.

**Regressionsanalyse evtl. schauen welche Proteine RNA-binding sind und was sie machen? shift types?**

## **5.Outlook**

nochmal ein paar sätze zum biological background.

## **6. Literature**

Gebauer et al., RNA-binding proteins in human genetic disease, 2020, Nature Reviews Genetics

Caudron-Herger et al., R-DeeP: Proteome-wide and Quantitative Identification of RNA-Dependent Proteins by Density Gradient Ultracentrifugation, 2019, Molecular Cell

Corley et al., How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms, 2020, Molecular Cell