Final Report: Proteome-wide Screen for RNA-dependent Proteins *non-synchronized A549 cells*

Anastasia Möller, Johannes Schadt, Sylviane Verschaeve, Tine Limberg

17.07.2023

Contents

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 analyzed contains 3680 different proteins from non-synchronized 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. Experimental Setup

To collect the mass spectrometry data, a strict protocol was followed. The non-synchronized A549 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

The data clean up consists of checking for missing values and if necessary, deleting rows of zeros. Furthermore, the columns are reordered to facilitate the separation of the dataset into two dataframes one containing Control while the other consists of the RNase group.

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 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 cannot 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 in 3074 proteins from initially 3680 proteins being 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

Since each normalization method has advantages and disadvantages, we apply three different methods to the dataframe. The **mean-value method (MVM)** is the first data normalization method 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 use 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, the smallest value per protein is added to each fraction of the corresponding protein, meaning the smallest value is now 0. Regarding the z-transformation we normalized first and then scaled to 100. Afterwards, we reduce by calculating the mean between the replicates with a higher correlation than 0.9 and scale 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.

2.4. Gaussian fit

The 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 describing the distribution are saved. To fit the parameters to our data we use the optim() function implemented in R.

2.6. Data characterization via Parameters

To identify whether a protein is RNA dependent or not, each protein is tested regarding four parameters. The goal of those four 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 A: 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 in the other sample, we define it as a significant change. An example of this is shown in figure 1.

2.6.2. Parameter B: Significant change of protein amount under local peaks

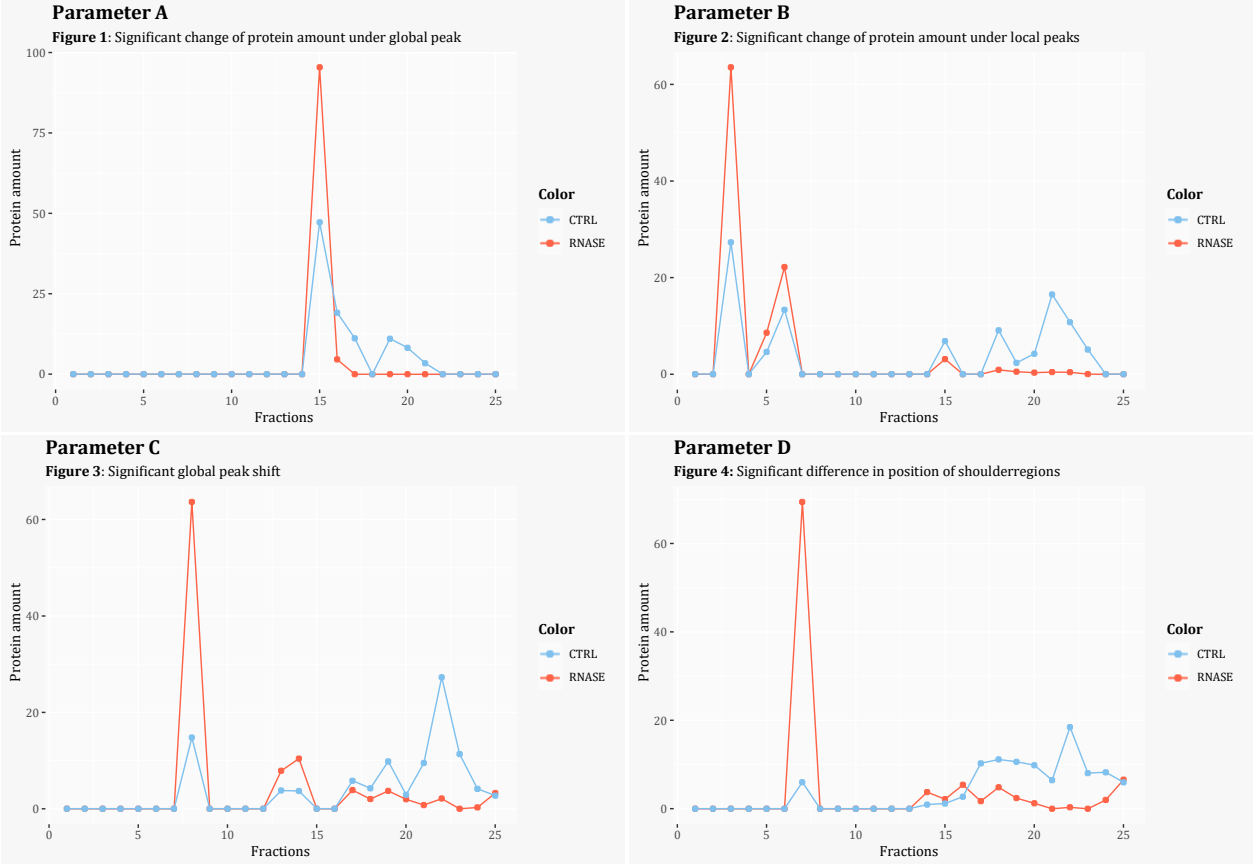
The next parameter recognizes a significant change in the total protein amount under the local peaks between the Control and RNase sample for each protein. New local peaks can occur if after RNase treatment the protein either dissociates or gains new interaction partners. To detect a significant change 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 must be higher than the y-value of its neighbor fractions. Afterwards we check if the sd of the local peak's y-value and its neighbors is higher than the sd of the y-values which each contain less than 8 % of the total protein amount (*sd.threshold*). The aim is to sort out small fluctuations between the y-values. The third criterium is that the local peak has to represent more than 3% of the total protein amount. Most of the already mentioned criteria fit to the global peaks as well and thus global maxima must be removed. For parameter B, 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. An example of this is shown in figure 2.

2.6.3. Parameter C: 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. An example of this is shown in figure 3.

2.6.4. Parameter D: Significant difference in position of Shoulderregions

At last, it is observed whether shoulderregions occur or disappear after the RNase treatment. A shoulderregion contains more than 2 consecutive fractions with a sd less than the *sd.threshold*. In contrast to the local peak identification the fractions with small fluctuations are kept. 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. An example of this is shown in figure 4.



2.6.5. Boundaries and Precipitated proteins

The local peakfinder cannot 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 must 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. Analysis of significant y-shift using t-test

A two-sided Welch's t-test is used to validate the results of the parameter for significant y-shifts. The t-test is performed on the global peaks of Control and RNase and its neighbors. To strengthen the power of the t-test, additional values were added in 0.25 steps between existing fractions using linear approximation. This leads to a smaller variance of each sample and an enhanced influence of the value of the global maxima which potentially differs between RNase and Control.

2.8. K-means Clustering

K-means is used to group a set of data points in a space of *d* dimensions 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 must be repeated *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 group our proteins in a two-dimensional space (fraction of control-peak and fraction of RNase) into *k=4* clusters. We perform k-means to have an alternative Method for identification of RNA-dependent Proteins, besides our Parameters.

2.9. Regression analysis

The linear regression models the mathematical relationship between a dependent variable and an independent variable. The linear equation is represented as y = mx + n. To analyze the model two values have to be taken into account. At first, if the *p-value* is above 5 %, the model must be discarded. Furthermore, the R-squared value describes the model's accuracy representing the proportion of the variance in the dependent variable that can be explained by the independent variables. Tow regression models are generated, both working with the pearson correlation between the Control and RNase protein amounts as the independent variable. The two models differ in their dependent variable. One uses the result of our four parameters, while the other uses the global shift amount in fractions. They are trained with 80 % of our dataset, so a prediction can be made for the remaining 20 %.

3. Results

3.1 Gaussian fit

On the grounds of our lists, containing the parameters for the gaussian distribution for every protein, it was possible to plot the gaussian curves. The control and RNase curves could be plotted separately or together in one plot for better comparison. Because, we used the optim() function and did not implement further parameters ourselves, the distribution did not show local peaks or shoulder regions. Therefore, the gaussian fit was only used for visualization not for further analysis. An example of this is shown in the figure 9 in the appendix.

3.2. Analysis of significant y-shift using t-test

For the z-transformed dataset the t-test identified 301 significant y-shifts (p-value ≤ 0.025) whereas the parameter A based version recognized 196 y-shifts. 127 proteins were in common. The t-test furthermore found 192 y-shifting proteins for the dataset using the mean-value method. The parameter-based version identified 253 significant y-shifts. For 106 proteins both methods returned a significant shift. Using min-max scaling, every global peak had a protein amount of 100, regardless of whether it was from the Control or RNase group, making it impossible to make meaningful comparisons.

3.3. RNA-dependent Proteins

Using just parameter C (global shift), we were able to characterize the shifttype, so whether the peaks show no shift, a right shift, a left shift, or are precipitated. Figure 5 visualizes our results for MVM. Using only parameter C we were able to identify 368 RNA-dependent Proteins with MVM, 349 with z-transformation and 370 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 figure 6 shows the significance of the different parameters and how they contributed to the characterization of the data. All proteins that had at least two positive parameters and all proteins that had solely a global shift were classified as RNA-dependent.

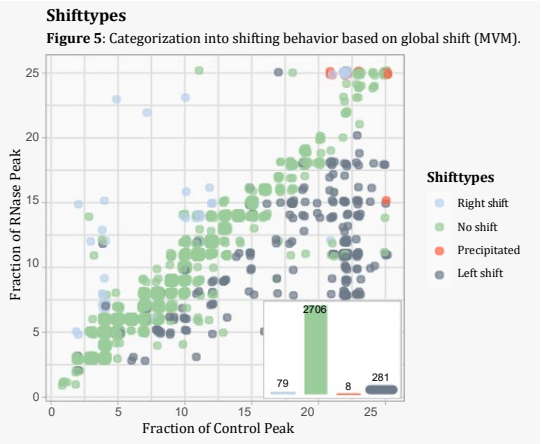


Figure 5: Categorization into shifting behavior based on global shift (MVM).

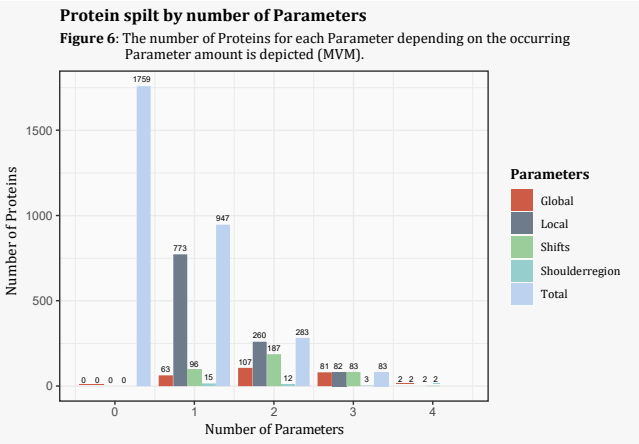


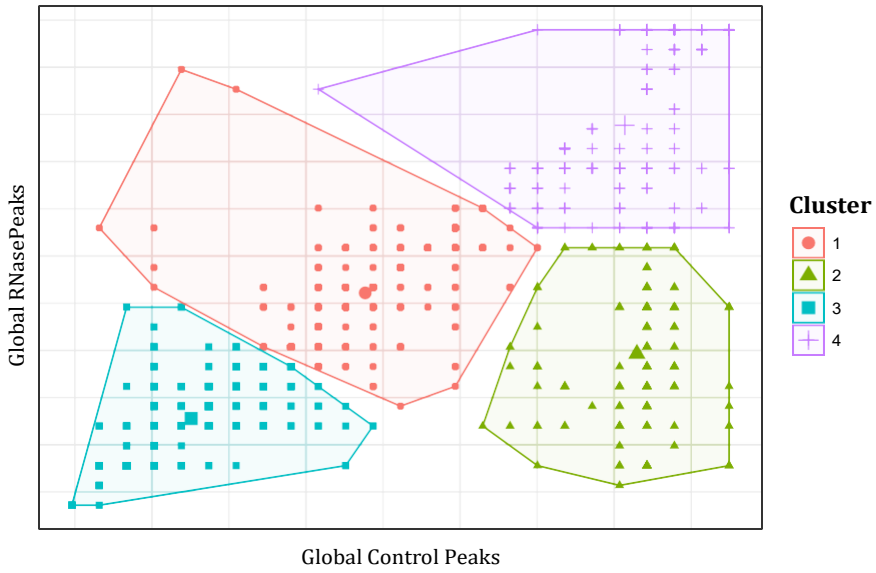
Figure 6: The number of Proteins for each Parameter depending on the occurring Parameter amount is depicted (MVM).

3.4. K-means Clustering

We wanted to cluster the Proteins depending on their global control peaks and their global RNase peaks. To find out how many clusters were optimal in theory we used the elbow method that showed us, that two clusters are optimal. However, looking at the biological background, two clusters didn't help us to identify RNA-dependent proteins, but rather grouped them into heavy and light proteins. To gain useful results from k-means, we forced it to create four clusters. K-means clustered the proteins as shown in figure 7. We chose the cluster at the bottom right for every normalization method, to be RNA-dependent. For MVM it is cluster two as depicted in Figure 7. The other clusters could not be characterized. There were no visible differences between the RNA-dependent proteins clusters of the different normalization methods. Using k-means we were able to identify 160 RNA-associated Proteins for MVM, 155 for z-transformation and 159 for MMS.

Clustering of Global Peaks

Figure 7: K-means clustering using Global Control and Global RNase Peaks (MVM).



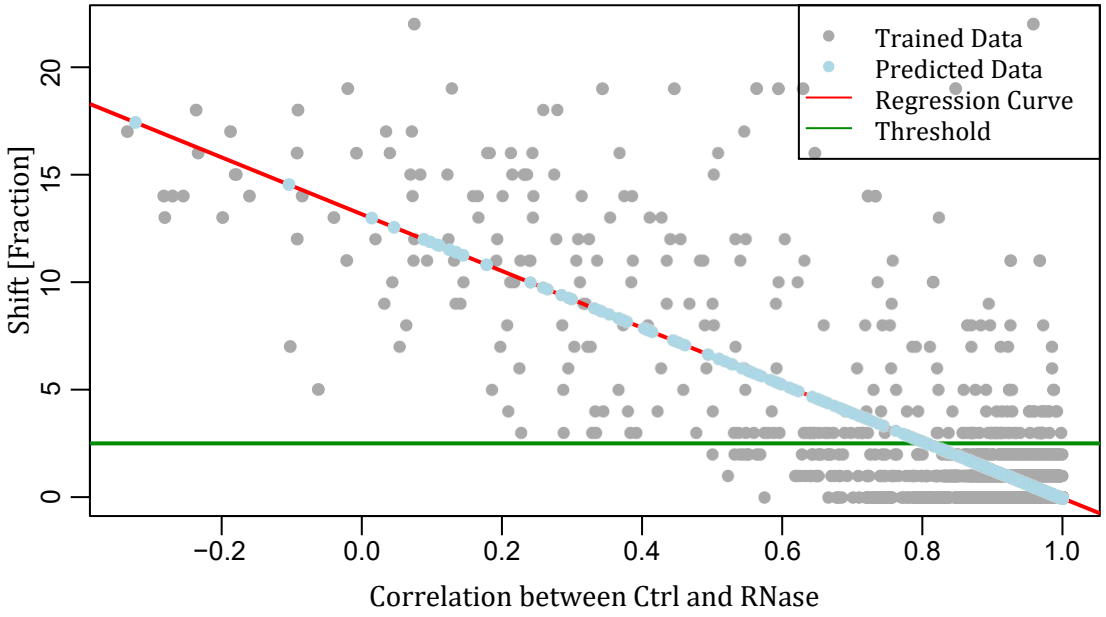
3.5. Regression analysis

The two already described regression models showed both a p-value below 0.05 and therefore they could be used for further analysis. The highest R-squared values were detected for the z-transformed data: the model based on the parameters showed a R-squared value of 0.43, while in the model using the global shift amount the R-squared value was 0.66.

The model with the higher R-squared value is visualized in Figure 8. Each protein is depicted as a dot. As expected, the majority of the trained data, marked in grey, showed a high correlation and a shift under 2 fractions. On the other hand, those with a small correlation had a shift upwards of 2 fractions. As already decided for the parameter **Significant fraction-shift of global peak**: a protein was classified as RNA-dependent protein if it had a shift higher than two fractions. The threshold is marked as a green line in the plot.

Regression Analysis

Figure 8: Regression curve with predicted data based on the global shift amount.



3.6. Comparison with Database

We compared our results with the R-Deep 2.0 database. Table 1 gives an overview of the identified Rdeeps compared to the database. It shows how many true positives, false positives, true negatives and false negatives per identification and normalization method were found.

Table 1: Overview of identified Rdeeps compared to database - numbers.

	Global shift			Parameters			K-means		
	MVM	zt	MMS	MVM	zt	MMS	MVM	zt	MMS
True Positives	301	295	302	356	380	316	158	153	157
False Positives	64	52	65	105	85	77	2	2	2
True Negatives	2148	2160	2147	2107	2127	2135	2210	2210	2210
False Negatives	540	546	539	485	461	525	683	688	684

Table 2 provides an overview of how well each method worked by depicting the false negative rate (FNR), the false positive rate (FPR) and the precision.

Table 2: Overview of identified Rdeeps compared to database - rates.

	Global shift			Parameters			K-means		
	MVM	zt	MMS	MVM	zt	MMS	MVM	zt	MMS
FNR	0.6421	0.6492	0.6409	0.5767	0.5482	0.6243	0.8121	0.8181	0.8133
FPR	0.0289	0.0235	0.0294	0.0475	0.0384	0.0348	0.0009	0.0009	0.0009
Precision	0.8247	0.8501	0.8229	0.7722	0.8172	0.8041	0.9875	0.9871	0.9874

There were 22 proteins, that were not present in the R-Deep 2.0 database. Whether our analysis characterized them as RNA-associated or not is shown in table 3 in the appendix.

4. Discussion

The goal was to identify RNA-dependent proteins using the data provided by mass-spectrometry. Our discussion will first focus on the t-test followed by the regression analysis and will end with the examination of our identified number of RDeeps in comparison with the database RDeep 2.0. The low overlap between the results of the t-test and the Parameter A: "Significant change of protein amount under global peak" was an indicator for a not high enough precision of the t-test to use it for further analysis. Regarding the regression analysis, the R-squared values differed notably. This can be explained with the following information: For the model based on the results of the parameters, we used zeros for non-RDeep proteins and ones for proteins we classified as RDeep, resulting in a low range and therefore a worse linear relationship. The second model had a higher range because we used the shift amount in fractions as the dependent variable, thus had a better linear relationship, shown by the higher R-squared value. To find the RDeep proteins, 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 54,8%. This could be, because our criteria were too strict, or because we did not include enough parameters. 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 2.0 database used the results from another analysis, that naturally, had different results. It is more important to look at the false-positive rate and precision of our results: Considering the global shift and all four parameters, normalization via z-transformation worked best. It led to the lowest false-negative rates, the lowest false-positive rates and the highest precision. Looking at our k-means results, MVM offered the highest precision. So, z-transformation or mean-value method are the normalization methods that should be used when analyzing this sort of data, depending on the method used for analysis. The following applies to our z-transformation-results but is similar to our MVM results and MMS results. If we look at the results of k-means, our parameters and global shift alone, we observed that k-means was the most precise, but also leaves out a lot more RNA-dependent proteins that were detected by other methods. This was because, we only chose one cluster, that probably contained left shift proteins, leaving out all precipitated and right 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 was clear that the global shift already detected more RNA-dependent proteins than k-means, but also had a higher false positive rate. Still, the false-positive rate of our global shift results was quite low and validates our decision to qualify proteins that solely had a global shift, but no other positive parameter, as RNA-dependent. If we look at the results of our four parameters, we saw 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 we identify RNA dependent proteins correctly, we should choose k-means. 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 would go down by about 2-4% depending on the normalization method. Looking at the global shift as the only parameter, would have been a comprise between precision and quantity. But even with k-means 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 22 new proteins that are not part of the R-Deep 2.0 database, of which, according to our parameters 3-4 are RDeep, depending on the normalization method used. These proteins could be analyzed further with additional experiments and be added to R-Deep 2.0.

5. Outlook

With our analysis we verified numerous RDeeps and identified 3 to 4 new ones. Nevertheless, for medical applications their cellular functions, and potential mutations causing the proteins to malfunction which leads to disease, have to be investigated.

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.

7. Appendix

As described in section 3.1. the following graph shows the Gaussian fit for the protein MYOME_HUMAN.

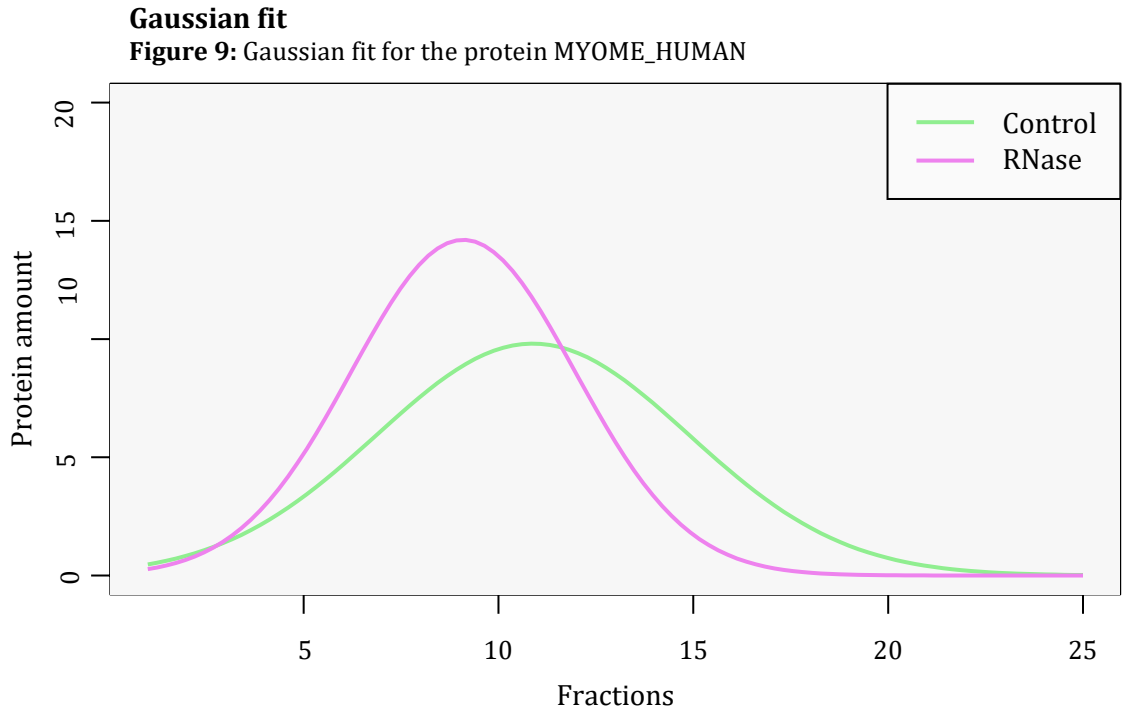


Figure 9: Gaussian fit for the protein MYOME_HUMAN

The following table shows the names of the 22 proteins not included in the R-Deep 2.0 database 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 k-means.

Table 3: Proteins not included in R-Deep 2.0 database.

	Name	RNA dependent according to parameters			RNA dependent according to k-means		
		MVM	zt	MMS	MVM	zt	MMS
1	OREX_HUMAN	0	0	0	0	0	0
2	DFFB_HUMAN	0	0	0	0	0	0
3	TISB_HUMAN	0	0	0	0	0	0
4	GDC_HUMAN	0	0	0	0	0	0
5	NAPSA_HUMAN	0	0	0	0	0	0
6	GGYF1_HUMAN	0	0	0	0	0	0
7	RNH1_HUMAN	1	1	1	0	0	0
8	MIS_HUMAN	0	0	0	0	0	0
9	CTDS2_HUMAN	0	0	0	0	0	0
10	MT2_HUMAN	0	0	0	0	0	0
11	SC6A4_HUMAN	0	0	0	0	0	0
12	DLGP3_HUMAN	1	1	1	0	0	0
13	LRP4_HUMAN	0	0	0	0	0	0
14	PEX2_HUMAN	0	1	0	0	0	0
15	APLP2_HUMAN	0	0	0	0	0	0
16	TAF12_HUMAN	0	0	0	0	0	0
17	VMAT1_HUMAN	1	0	1	0	0	0
18	PHIP_HUMAN	0	0	0	0	0	0
19	KLC3_HUMAN	0	0	0	0	0	0
20	MTU1_HUMAN	0	0	0	0	0	0
21	NXT1_HUMAN	1	1	1	0	0	0
22	MBD4_HUMAN	0	0	0	0	0	0
total	22	4	4	4	0	0	0