Proteome-wide Screen for RNA-dependent Proteins in HeLa Cells Synchronized in Interphase

Data Analysis Project 2023

Hannah Brehm, Johann Blakytny, Kira Hoffmann, Viktor Bonin

17.07.2023

Abstract

RNA-protein interactions are key mediators of integral cellular and metabolic processes. Therefore, alterations in RNA-binding proteins play a central role in the initiation and progression of many human diseases.

The method R-DeeP allows to investigate the RNA-dependence of proteins, meaning that not only RNA-protein interactions are detected but also RNA-protein-protein interactions. Untreated and RNase-treated cell lysates are separated on a sucrose gradient into fractions. The proteins in these fractions are analyzed through mass spectrometry. Based on this spectrometry data, we aim to identify RNA-dependent proteins (RDPs).

After cleaning the data, the protein distributions were smoothed using a three-fraction window. Global and local maxima of the smoothed protein distributions were then determined. Each global maximum and the highest local maximum were fitted with Gaussian curves.

The subsequent data was used to define a set of selection criteria. The optimal selection criteria for identifying RDPs were found using principal component analysis combined with manually determined shifting behavior.

For each optimal selection criterium, a threshold for RNA-dependence was set. This resulted in the identification of 539 RDPs, 274 partial RDPs and 5177 non-RDPs. These findings were compared to the RBP2GO database.

Linear regression was used to investigate if the mass spectrometry data could be used to examine cellular parameters, such as the kinetics of RNA-protein interactions.

This analysis shows how important bioinformatic methods are in the identification of RDPs. Future work based on R-DeeP and methods like it will undoubtedly have important implications in the treatment of many human diseases.

Table of Contents

1. Introduction

RNA and its interacting proteins are relevant for essential cellular activity. Especially in posttranscriptional processes, RNA-binding proteins (RBPs) mediate the regulation of posttranscriptional modification, transport through the cytoplasm, translation, and stability of the RNA template. Different RNA-binding motifs have been characterized so far allowing the linkage of proteins to ribonucleic acid (Burd and Dreyfuss, 1994). The structure of most RNA-binding domains is highly conserved; nevertheless, many RBPs bind with high specificity and affinity to distinct RNAs (Lunde *et al.*, 2007). As RBPs are crucial for certain housekeeping mechanisms, mutations or other alterations in human RBPs can have pathological effects; for instance, alterations in the expression of the *zinc-finger protein SNAI1* are associated with cancer metastasis (Gebauer *et al.*, 2021).

There are multiple classifications for interactions between proteins and RNA. RBPs can be characterized in a binary system of being "specific" or "nonspecific" where specific RBPs interact with defined RNA motifs and nonspecific RBPs bind to RNA with an absence of definite motifs (Jankowsky and Harris, 2015). Further differentiation can lead to RBP classes defined by specific RNA-binding domains such as zinc-finger domains or RNA recognition motifs. Additionally, the targets of RBPs can be considered for creating subgroups. Thus, for instance mRNA-binding or tRNA-binding RBPs are classified (Gerstberger *et al.*, 2014).

Furthermore, RBPs can be classified as a subcategory of RNA-interacting proteins themselves. The concept of RNA-dependency combines RBPs with proteins which are interacting indirectly with RNA. The latter can also be called RBP-interacting proteins. Therefore, RNA-dependent proteins (RDPs) allow for a broader comprehension of the RNA interactome (Caudron-Herger *et al.*, 2020b).

Various methods have been applied to identify RNA recognition elements so far, for example crosslinking and immunoprecipitation followed by sequencing (CLIP-seq) (Gerstberger *et al.*, 2014; 2013). This enables identification of RNAs bound to proteins. Detecting the proteins interacting with RNA requires other methods such as comprehensive identification of RBPs by mass spectrometry (CHIRP-MS) (Licatalosi *et al.*, 2020). Introducing the method R-DeeP, Caudron-Herger *et al.* promise a "proteome-wide, unbiased, and enrichment-free screen" that is solely based on disrupting the RNA-protein and RNA-protein-protein interaction (Caudron-Herger *et al.*, 2020b; 2019). Through R-DeeP, we aim to identify more human RBPs and RBP-interacting proteins in this project.

2. Material and Methods

2.1 Referenced Data

The protocol for an R-DeeP-Screen was performed on HeLa cells synchronized in interphase (Caudron-Herger *et al.*, 2020b) and the resulting protein amounts per fraction were stored in the dataframe *RDeep_HeLa_Interphase*.

2.2 Data Cleanup

The original dataset *RDeep_HeLa_Interphase* was split into two separate data frames containing the values for all RNase fractions and Control fractions respectively. These dataframes were named *RNase_RDeep* and *Ctrl_RDeep* respectively. The following steps were performed on both dataframes.

2.2.1 NA Values and Zero Rows

All rows containing only zeros, of which there were five, were deleted from the dataframes. No NA values were found.

2.2.2 Pearson Correlation

To test the reproducibility of the dataframe the pearson correlation was used. In seven cases one of the three replicates contained only zeros. These values were replaced in with NA values as determining the pearson correlation with these replicates would have been impossible. In five rows, two out of the three replicates contained only zeros. These genes were removed completely from the dataframes.

The pearson correlation r was computed via the following formula using the sample size N, individual sample points x_i and y_i, sample means $\bar{x}$ and $\bar{y}$ and standard deviations s_x and s_y:

$$r = \frac{1}{N-1} \sum_{i=1}^{N} \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y}$$

For each gene and condition this resulted in three correlation coefficients with a determined threshold of over or equal to 0.95. The genes for which all three correlation coefficients fell below that threshold were removed from both dataframes. This applied to 1335 genes. If two out of three correlation coefficients were under 0.95 the replicate associated with both coefficients was replaced with NA. This was the case for 2594 replicates. Lastly, if one of three correlation coefficients was below the determined threshold the replicate with the lower value in the other associated correlation coefficient was replaced with NA. This method was performed on 2285 replicates.

2.2.3 Mean of Replicates

After testing for reproducibility and deleting unreliable values, the arithmetic mean was calculated for all triplicates and used for further analysis.

2.3 Smoothing and Determination of Maxima

2.3.1 Smoothing the Data

To determine the global and local maxima of the dataset, the dataframes *Ctrl_Mean* and *RNase_Mean* were smoothed separately. Otherwise, the determination would have been influenced by outliers and wrong peaks.

Smoothing was achieved by calculating the mean value for each fraction through its defined neighborhood of three with the following formula (applies to all fractions between 2 and 24):

$$y = \frac{f(x-1) + f(x) + f(x-1)}{3}$$

x: fraction position ($2 \leq x \leq 24$); $f(x)$: mass spectrometry value at fraction position x; y: smoothed value for fraction position x

For fraction position 1 and 25, the mean of fraction 1 and 2 and accordingly 24 and 25 were used instead.

2.3.2 Normalization of Smoothed data

In order to be able to compare the different protein amounts, the smoothed dataframes were normalized to 100. This means the sum of the mass spectrometry data in the different fractions for one protein added up to 100:

$$y = \frac{f(x)}{\sum_{x=1}^{25} f(x)} * 100$$

x: fraction position ($1 \leq x \leq 25$); $f(x)$: mass spectrometry value at fraction position x; y: normalized value at fraction position x The same procedure was then applied to the standard deviation.

2.3.3 Determination of Maxima

If a protein is RNA-dependent, the maximum amount of protein should be located in a different fraction in the RNase sample compared to the Control sample. To determine for which proteins this is the case, the global and local maxima of the mass spectrometry data were determined.

2.3.3.1 Global Maxima

The global maxima for the smoothed and normalized dataframes were determined by identifying the fraction of the largest mass spectrometry value in each dataframe. For proteins that had two or three fractions with the same value, the mean fraction position was calculated.

2.3.3.2 Local Maxima

The local maxima for the smoothed and normalized dataframes were determined by comparing the protein amount of each fraction position to the protein amount in the fractions in a neighborhood of $n = 2$. If the protein amount in that fraction is greater or equal to its neighborhood, it was defined as a local maximum. For the fraction positions 2 and 24, the protein amounts had to be greater than or equal to the protein amounts in the fractions 1, 3 and 4 or fractions 22, 23 and 25 respectively. The fraction positions 1 and 25 had to be greater than or equal to fractions 2 and 3 and fractions 23 and 24 respectively. For proteins that had two or three fractions with the same mass spectrometry value, the mean fraction position was calculated. To distinguish local from global peaks, all global peaks were subtracted from the local peaks dataframe.

2.4 Gaussian Fitting

A Gaussian curve was fitted to the global peak and the local peak with the highest protein amount of each gene in the Control and RNase samples. A Gaussian curve can be described using the height h, standard deviation σ and mean μ as:

$$f(x) = \frac{h}{\sqrt{2\pi} \cdot \sigma} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

These parameters were first estimated and then optimized. μ was estimated as the fraction position of the global and largest local peak, σ and h were estimated as follows:

$$\sigma = \frac{HWHM}{\sqrt{2 \cdot ln(2)}}; \qquad h = \frac{MPA}{sd \cdot \sqrt{2\pi}}$$

HWHM: half width at half maximum; MPA: maximum protein amount. These estimated values were optimized by minimizing the sum of squares of the differences between protein amount values of the Gaussian curve G and the smoothed and normalized protein distribution p ($min(\sum(G-p)^2)$). This was used to calculate the protein amount under the Gaussian curve and the overlap of the Gaussian curves between the Control and the RNase samples.

2.5 Selection Criteria for Identification of RNA-Dependent Proteins

To determine the optimal selection criteria for an RNA-dependent protein, multiple parameters were compiled and applied to a principle component analysis. The best parameters were subsequently used to differentiate between RDPs, non-RDPs and partial RDPs.

2.5.1 Defining Selection Criteria

The following parameters were compiled to determine the optimal selection criteria for defining RNA-dependent proteins:

The **absolute peak shift and the percentage of the peak shift** between the Control and RNase global peaks, between the global and main local peak of Control, and between the global and main local peak of RNase,

the **absolute change and percentage of change** in the **standard deviation** between Control and RNase, in the **global peak height** between Control and RNase, and between the the **protein amount** under the Gaussian curves of the Control and RNase peaks,

the **absolute protein amount and the percentage of protein amount** that is lost under the global peak of Control after RNase treatment, and that is gained under the global peak of RNase after RNase treatment,

the **absolute differences and the percentage of the differences in the protein amounts** of the Control and RNase samples under the main local Control peak,

the **percentage of overlap** between the Gaussian curves of the Control and RNase, between the Gaussian curves of the main local peak and global peak of Control, between the Gaussian curves of the main local peak and global peak of RNase, between the Gaussian curves of the main local peaks of Control and RNase.

A *shift* is a character variable that states whether a protein is *left*, *right*, *non* shifting or *precipitated* and is defined as following: $\Delta\mu = \mu(Control) - \mu(RNase) > +1$: right shift; $\Delta\mu = \mu(Control) - \mu(RNase) < -1$: left shift; $-1 \leq \Delta\mu = \mu(Control) - \mu(RNase) \leq +1$: no shift;$\mu(Control) > 24 \vee \mu(RNase) > 24$: precipitated

2.5.2 Principal Componant Analysis

By reducing the dimensions and applying a principle component analysis (PCA) to the percentage values of all selection criteria, the optimal parameters to select for RNA-dependent proteins were examined. Only the variables calculated in percentages were used for the PCA for better comparison and the variables regarding main local peaks only served as determining partially shifting proteins and were therefore not included in the principle component analysis. The selection criteria that separate the points the most in the plot of the first two principal components were chosen for further analysis.

2.5.3 Application of Selection Criteria

Based on the PCA, four parameters were defined as most important for separating shifting from non-shifting proteins. The following thresholds for the four parameters were consecutively set for the classification of an RNA-dependent protein (RDP):

1. Global peak shift of more than **1 fraction**
2. Protein loss under Control global peak of more than **20 %** after RNase treatment
3. Protein gain under RNase global peak of more than **20 %** after RNase treatment
4. Overlap of the Gaussian curves of less than **75 %**

Furthermore, *partially* RNA-dependent proteins were defined as proteins that satisfy the first and fourth criterion and either the second or third criterion for RNA dependence. More *partially* RNA-dependent proteins were identified using the following criteria:

1. Fraction shift between global and main local peak positions of more than **1 fraction**
2. Overlap between the Gaussian curves of global and main local peaks of less than **75 %**
3. Overlap between the Gaussian curves of the local peaks of less than **75 %**

4. Protein gain under main local peak in relation to protein distribution of other sample under the same peak of more than **20 %**
5. Protein amount at main local peak of more than **8**
6. Previous categorization as **“non-RDP”**

2.6 Further Analysis

2.6.1 Comparison with the RBP2GO Database

To evaluate the accuracy of the performed analysis, the identified RDPs were compared with their respective RDP-categorization in the RBP2GO database.

2.6.2.1 Direct Comparison

To further examine the relation between our RDP identification method and the RBP2GO classification, both RBP2GO-RDPs and -non-RDPs were plotted in the biplot of the first two principal components from PCA.

2.6.2.2 Cluster Comparison with RBP2GO Score and t-Test

By calculating the **mean RBP2GO score** for each cluster identified by applying the optimal selection criteria, the RNA-dependence of a cluster (non-/partial/RDP) was estimated. Using a two-sided Welch t-test, the significance of the differences between the mean RBP2GO score values were calculated. The H0 and H1 hypotheses were defined as following: H0: The mean values are not significantly different from each other, H1: The mean values are significantly different from each other. The false positive rate was set at $p = 0,05$.

2.6.2.3 Barplot with all Identified RDPs

Finally, a barplot was created to compare the overlap between all identified RDPs via the performed analysis and all RDPs classified as such in the RBP2GO database.

2.6.3 Linear Regression

In order to apply the obtained data on RDPs to biological questions, a linear regression was performed to evaluate how well of a predictor **binding affinity** is to **binding specificity**. Binding affinity is defined as the difference in protein amount between the Control and RNase samples at the fraction position of the Control global peak and binding specificity is defined as the reciprocal value of the full width at half maximum of the Gaussian curve fitted to the Control global peak.

3. Results

3.1 Results of Data Cleanup, Localization of Maxima and Gaussian Fitting

After performing all data cleanup steps, the original dataset *RDeep_HeLa_Interphase* was reduced from 7086 proteins to 5990 Proteins. Local and global maxima were determined and the Gaussian curve fitted to the peaks. As can be observed in the plots of “CTCF_HUMAN” (*Fig.1*), the smoothing and normalization steps as well as the Gaussian Fitting applied to the dataset (*Fig. 2*) were successful.

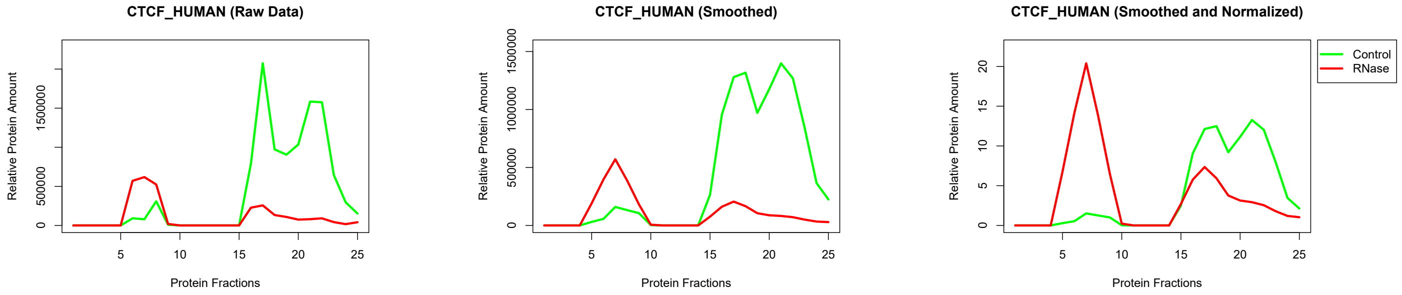

Figure 1: Data Cleanup. The protein distributions for the gene *CTCF_HUMAN* were smoothed and then normalized.

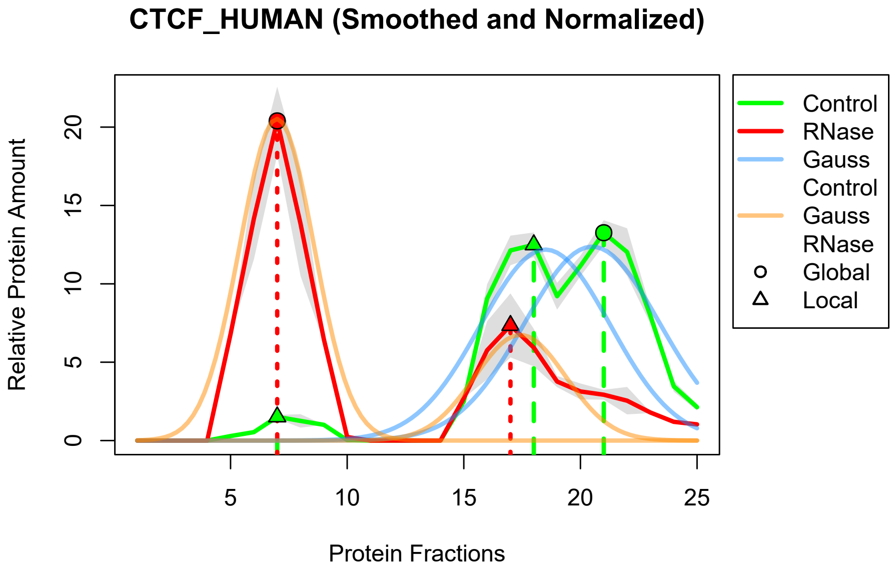

Figure 2: Local and global Maxima and Gaussian Fitting. Local and global maxima of the protein *CTCF_HUMAN* together with fitted Gaussian curves.

3.2 Shifting behavior

In *Fig. 3*, the shifting behavior is represented by plotting the global peak positions of the Control against those of the RNase samples. Non-shifting proteins lie an a straight line through the origin with a slope of 1, signifying equal positions of both global peaks. Left shifting proteins lie below this line, right shifting proteins lie above. Precipitated proteins lie close to the right as well as to the top edge which means, that either one of the peaks is in the 25th fraction. Most proteins are non-shifting (5153 proteins). The largest group of shifting proteins are left shifting (782 proteins).

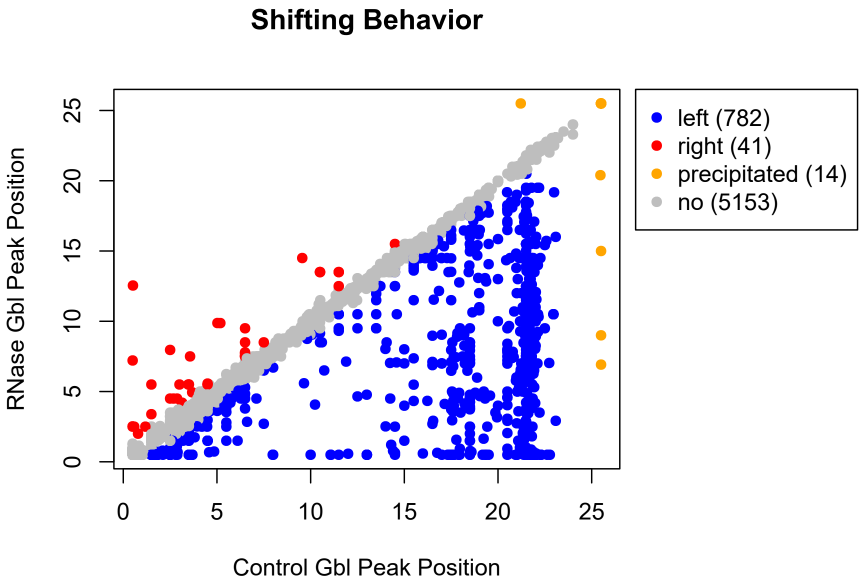

Figure 3: Shifting Behavior. All analyzed proteins were classified into left, right and no shifting and precipitated using the criteria mentioned above. They were plotted using the global peak positions of the Control and RNase samples.

3.3 Results of PCA

In *Fig. 4 (left)* the PCA was applied to all selection criteria and the first and second principle component were plotted against each other. The different shifting behaviors of the proteins are shown by using different colors. Left shifting proteins (blue), the largest shifting group, spread along the following four selection criteria: delta_Glb_mu_%, Ctrl_Gbl_prt_prop_%, RNase_Gbl_prt_prop_% and negative Gbl_overlap_%. Therefore, they were defined as the optimal selection criteria. To illustrate this, *Fig. 4 (right)* shows the first and second principle component plotted against each other when PCA is performed solely of the optimal selection criteria. This shows that shifting proteins spread in the direction of delta_Glb_mu_%, Ctrl_Gbl_prt_prop_% and RNase_Gbl_prt_prop_%, whereas non-shifting proteins spread in the direction of the delta_Glb_mu_%. Since delta_Gbl_mu_% is the longest vector, it has the largest influence on the shifting behavior.

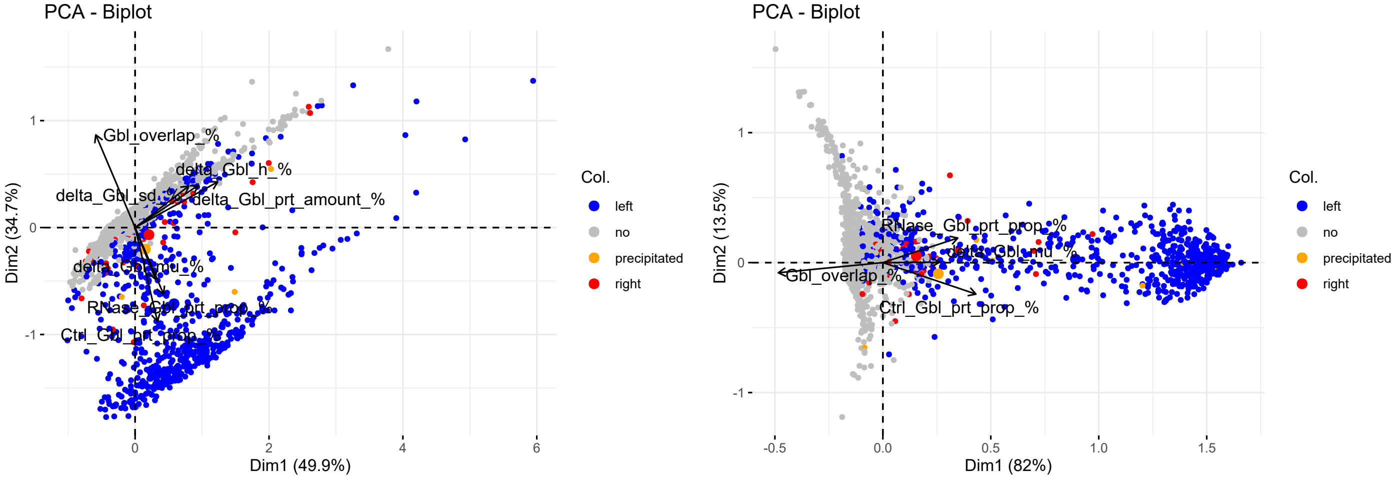

Figure 4: PCA Biplots. Principle Component Analysis of all selection criteria (left). Principle Component Analysis of optimal selection criteria (right). The first two principle components are plotted against each other, the shift classification of the proteins is shown via different colors.

3.4 Application of Selection Criteria and Determination of RDPs

All proteins are plotted in a biplot of the first and second principle components and colored depending on the classification based on the optimal selection criteria *(Fig. 5, left)*. Two main clusters are formed, one containing RDPs (blue) and one containing non-RDPs and partial RDPs (gray and orange). As a result of the optimal selection criteria, 539 proteins are classified as RDPs (blue), 274 are classified as partial RDPs (orange) and 5177 are classified as non-RDPs (gray).

3.5 Comparison with the RBP2GO Database

To determine the effectiveness of our method, the proteins classified as RDPs in our analysis were compared to the proteins classified as RDPs in the RBP2GO database.

Firstly, both RBP2GO-RDPs and -non-RDPs were plotted in the biplot of the first two principal components used in our analysis. As depicted in *Fig. 5 (right)*, the RBP2GO-classified RDPs distribute homogeneously across the two clusters with no clear observable trend.

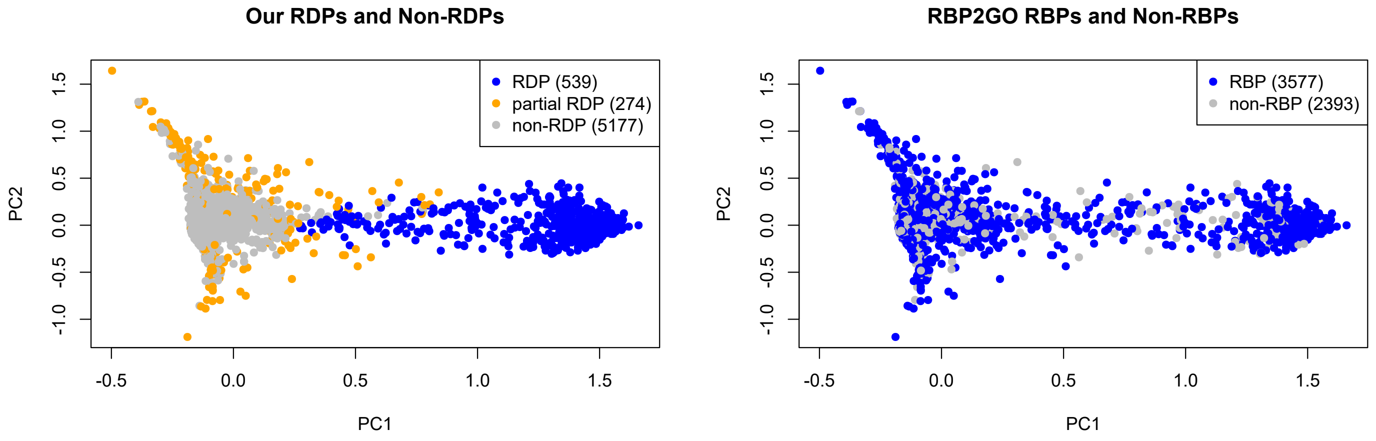

Figure 5: Identification of RDPs and Comparison to RBP2GO. The first two principle components of the optimal selection criteria PCA are plotted against each other. The classification of the proteins (RDP, partial RDP, non-RDP) achieved in the performed analysis is shown by using different colors (left). The classification of the proteins (RDP, non-RDP) from the RBP2GO database is shown by using different colors (right).

3.6 Cluster Comparison with RBP2GO Score and T-test

To determine the significance of our clusters, the average RBP2GO score of these clusters was compared with a t-test. The average RBP2GO score of our RDPs is 21.5 with a standard error of 0.9, the average score of our partial RDPs is 13.6 with a standard error of 0.9 and the average score of our non-RDPs is 6.6 with a standard error of 0.1.
For the comparison of the different clusters following p-values were calculated: p(RDP, non-RDP) = 8.5e-54, p(RDP, partial RDP) = 8.2e-10, p(partial RDP, non-RDP) = 2.0e-12. All clusters have significant p-values since each p-value is smaller than the False Positive Rate of $p = 0.05$.

3.7 Bar Plot

Comparing the RDPs identified via our presented analysis with the RDPs classified as such in the RBP2GO database, 631 proteins were identified as RDPs in both analyses, 2946 proteins were listed as RDPs in the database but were not identified as RDPs in our performed analysis, and 182 proteins which we determined as RDPs are not classified as such on the RBP2GO database (*Fig. 6*).

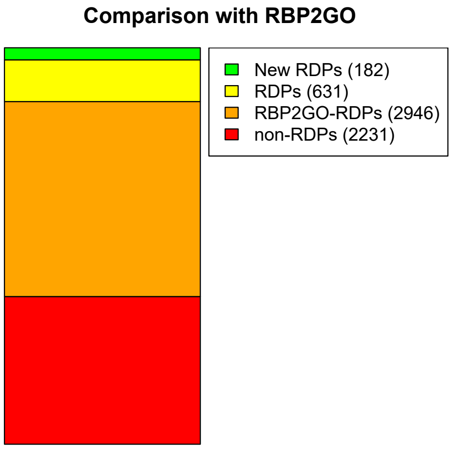

Figure 6: Bar Plot. The new proteins classified as RDPs via the performed analysis and the RBP2GO database are illustrated in this bar plot.

3.8 Linear Regression

Performing a linear regression to test whether binding affinity is a good predictor of binding specificity returns a very low coefficient of determination for all analyzed proteins ($R^2 = 0.03$) (*Fig. 7, left*) and a rather low coefficient of determination for only the RDPs ($R^2 = 0.29$) (*Fig. 7, right*). It is noticeable that the calculated correlation between binding affinity and specificity is positive in both cases, meaning that on average the reciprocal value of the full width at half maximum decreases with an increasing difference between the protein amount of Control and RNase samples at the global Control peak position.

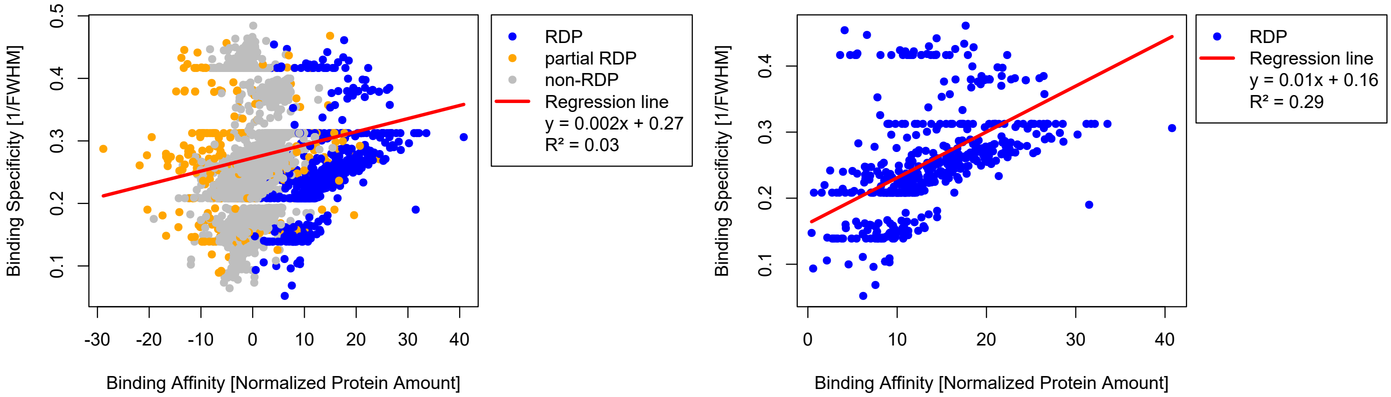

Figure 7: Linear Regression. The linear regression is applied to the binding affinity and the binding specificity for all proteins (left) and for only RDPs (right) classified in this analysis.

4. Discussion

In conclusion, the protein distributions provided by Caudron-Herger *et al.* for HeLa cells synchronized in interphase were cleaned and normalized. Next, the arithmetic mean for each fraction was calculated. The averaged protein distributions were smoothed using a three fraction sliding window and normalized (*Fig. 1*).

Smoothing was performed to help with global and local peak determination, because the averaged protein distributions were sometimes too jagged for unambiguous peak determination (*Fig. 2*).

However, this approach has three problems: First, single fraction peaks result, upon smoothing, in multiple peaks in neighboring fractions. This problem was fixed by calculating the mean fraction position for multiple peaks.
Second, smoothing sometimes caused the fraction position of peaks to change. But this did not have an influence on analysis, because the change in fraction position between raw and smoothed data was very slight (no more than one fraction position).
Third, smoothing changed the protein amount which sometimes led to global peaks in the raw data becoming local peaks in the smoothed data and *vice versa*. This problem was partly fixed by defining additional selection criteria for partial RDPs using the main local peaks.

Next, the shifting behavior of the proteins was analyzed. There were 782 left shifting, 41 right shifting, 5153 non-shifting and 14 precipitated proteins (*Fig. 3*).

PCA helped with determining the optimal selection criteria for an RDP, namely the fraction shift between global peaks, the overlap of the Gaussian curves fitted to the global peaks and the protein amount gained or lost under the global peaks after treatment with RNase (*Fig. 4*). Two rough clusters emerged from plotting the first two principal components. In the right plot, most of the non-shifting proteins were in the left cluster and most of the left shifting proteins were in the right cluster.

Application of the selection criteria resulted in 539 RDPs, 274 partial RDPs and 5177 non-RDPs (*Fig. 5, left*). These results compare well with the genes' identified shifting behavior of which 823 are categorized as either right or left shifting compared to 813 identified RDPs and partial RDPs. Note that this does not mean that the individual genes are classified in the same way (compare *Fig. 4, right* with *Fig. 5, left*).

The identified RDPs are only partly comparable to the RBP2GO databank (813 RDPs and partial RDPs and 5177 non-RDPs compared to 3577 RDPs and 2393 non-RDPs) (*Fig. 5*).
Another way to compare the identified RDPs with the RBP2GO databank is the RBP2GO score. The means of the RBP2GO scores for RDPs, partial RDPs and non-RDPs was calculated. The mean of the RBP2GO scores for the identified RDPs was 21.5, compared to 13.6 for partial RDPs and 6.6 for non-RDPs. The differences in the mean values between RDPs, partial RDPs and non-RDPs are all highly significant based on p-values calculated using a two-sided Welch two-sample t-test. It is noticeable that the p-value of the t-test for the difference between the mean values of the RBP2GO score of RDPs and non-RDPs is by most the smallest (p(RDP, non-RDP) = 8.5e-54 compared to p(RDP, partial RDP) = 8.2e-10 and p(partial RDP, non-RDP) = 2.0e-12). But these findings are of limited use because the RBP2GO score was calculated very differently from the analysis presented here (Caudron-Herger *et al.*, 2020a).

The identification of RDPs was also compared directly to the RBP2GO databank (*Fig. 6*): 2231 genes (37 % of all analyzed genes) which are categorized as non-RDPs are also listed as non-RBPs in the RBP2GO databank. In addition, 631 (11 %) RDPs are included as RBPs in RBP2GO. Yet, nearly half of the analyzed genes (2946 (49 %)) are categorized as non-RDPs, even though RBP2GO lists them as RBPs. Moreover, 182 (3 %) RDPs are listed as non-RBPs in the RBP2GO databank. These "new" RDPs include proteins from which it is known that they interact with RNA either directly or indirectly (e.g. *CDK13_HUMAN*, *Fig. A.6*) (Berro *et al.*, 2008).

The differences between these results and the RBP2GO databank are possibly the result of different human cell lineages used, different stages of the cell cycle analyzed and different experimental methods. Notably, the analysis presented here is only for HeLa cells synchronized in interphase, so poor comparability with the RBP2GO databank, which draws its data from many orthogonal experiments, should be expected.

Linear regression was used to examine if biological questions could be answered using the protein distributions. It would be expected that an RDP which binds tightly to its RNA target also forms very specific structures and, therefore, has a high binding specificity. This was partially observed when applying linear regression modeling only to the identified RDPs (*Fig. 7, right*). Analysis of all proteins or partial RDPs or non-RDPs show a lower coefficient of correlation (*Fig. 7, left* and *Fig. A.7*). However, due to the highly diverse nature of RNA-protein interactions further research would be needed.

The results of this analysis show that, besides the experimental methods used, bioinformatical analysis also plays a major role in determining the RNA-dependency of proteins. Further improvements of analysis in RNA-protein interactions will undoubtedly lead to better understanding and new therapies of human diseases associated with RBPs and RDPs.

5. References

Berro, R., Pedati, C., Kehn-Hall, K., Wu, W., Klase, Z., Even, Y., Genevière, A.M., Ammosova, T., Nekhai, S., and Kashanchi, F. (2008). CDK13, a new potential human immunodeficiency virus type 1 inhibitory factor regulating viral mRNA splicing. J Virol *82*, 7155-716

Burd, C.G., and Dreyfuss, G. (1994). Conserved Structures and Diversity of Functions of RNA-Binding Proteins. Science *265*, 615-621.

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.

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

Caudron-Herger, M., Wassmer, E., Nasa, I., Schultz, A.-S., Seiler, J., Kettenbach, A.N., and Diederichs, S. (2020b). Identification, quantification and bioinformatic analysis of RNA-dependent proteins by RNase treatment and density gradient ultracentrifugation using R-DeeP. Nature Protocols *15*, 1338-1370.

Gebauer, F., Schwarzl, T., Valcárcel, J., and Hentze, M.W. (2021). RNA-binding proteins in human genetic disease. Nature Reviews Genetics *22*, 185-198.

Gerstberger, S., Hafner, M., and Tuschl, T. (2013). Learning the language of post-transcriptional gene regulation. Genome Biology *14*, 130.

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

Jankowsky, E., and Harris, M.E. (2015). Specificity and nonspecificity in RNA–protein interactions. Nature Reviews Molecular Cell Biology *16*, 533-544.

Licatalosi, D.D., Ye, X., and Jankowsky, E. (2020). Approaches for measuring the dynamics of RNA–protein interactions. WIREs RNA *11*, e1565.

Lunde, B.M., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for efficient function. Nature Reviews Molecular Cell Biology *8*, 479-490.

6. Appendix

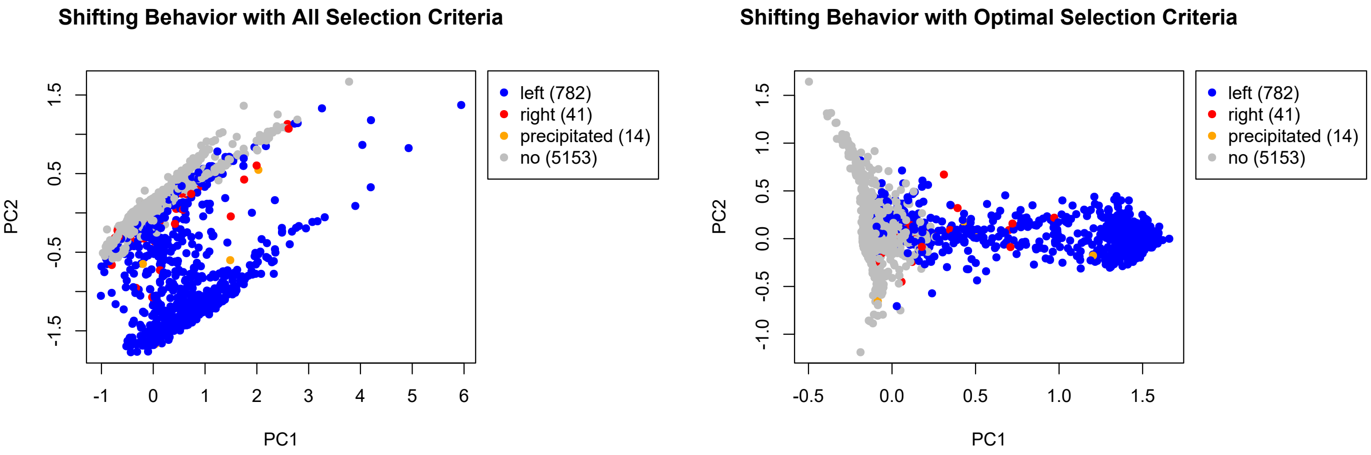

Figure A.1: PCA with Shifting Behavior. The first two principle components of the PCA applied to all selection criteria (left) and of the PCA applied to the optimal selection criteria (right) are plotted against each other. The shifting classification of the proteins is shown by different colors.

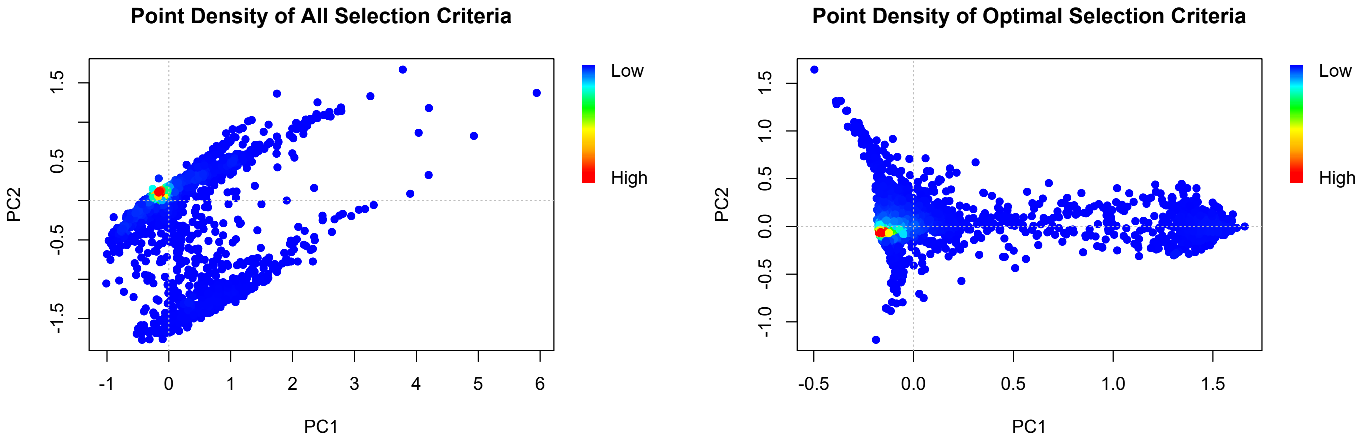

Figure A.2: PCA Densitiy Plots. The first two principle components of the PCA applied to all selection criteria (left) and of the PCA applied to the optimal selection criteria (right) are plotted against each other. The point density of the proteins is shown by the usage of different colors.

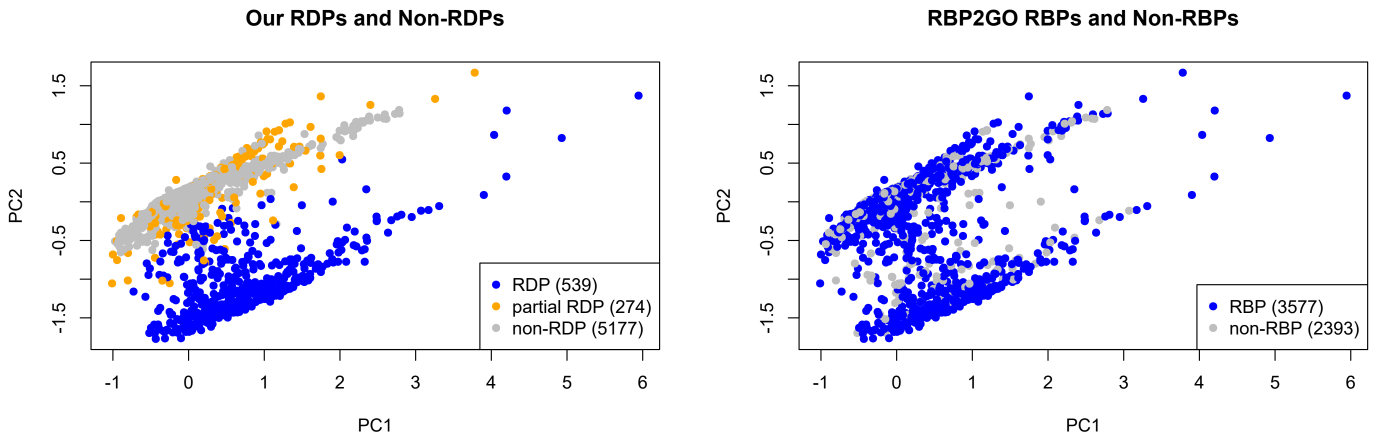

Figure A.3: Identification of RDPs and Comparison to RBP2GO. The first two principle components of the PCA for all selection criteria are plotted. The classification of the proteins (RDP, partial RDP, non-RDP) achieved in the performed analysis is shown by using different colors (left). The classification of the proteins (RDP, non-RDP) from the RBP2GO database is shown by using different colors (right).

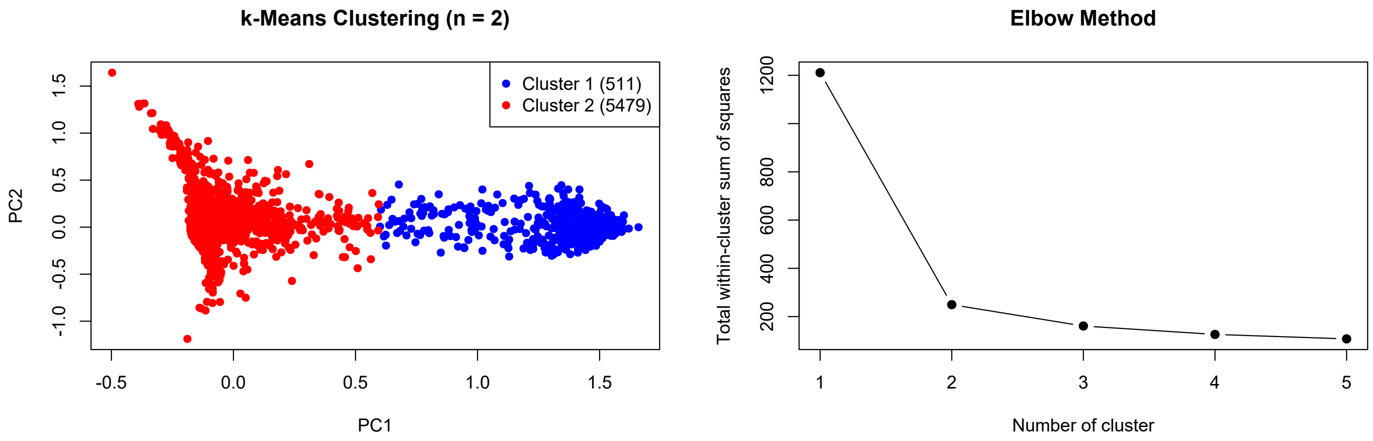

Figure A.4: *k*-Means Clustering of PCA Biplot of Optimal Selection Criteria. The elbow plot for determination of optimal number of clusters (right) and applied clusters to the biplot of the optimal selection criteria from PCA (left).

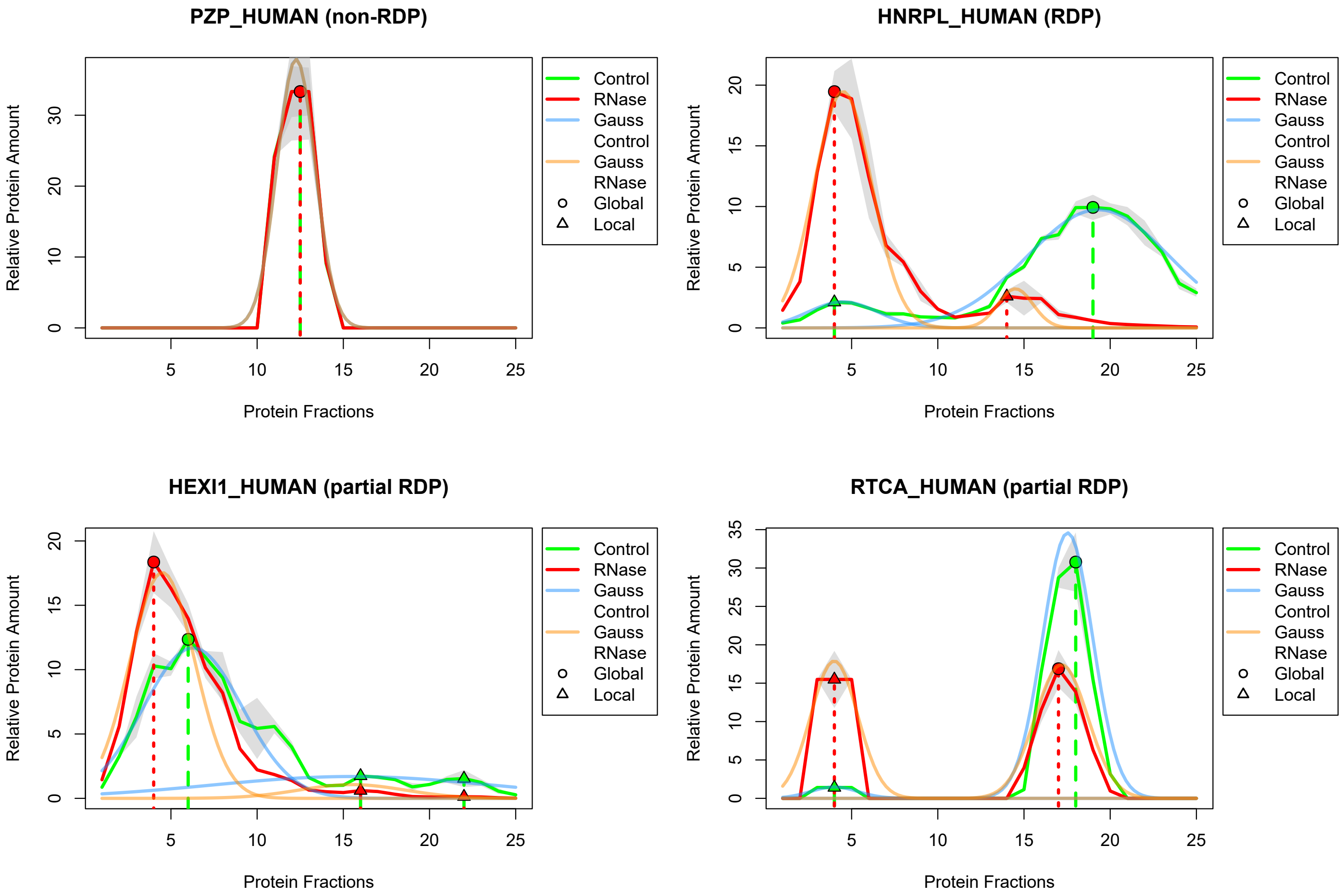

Figure A.5: Application on Selection Criteria. Protein and Gaussian distributions for the Control and RNAse sample for following for different proteins are plotted for comparison: *PZP_HUMAN* (non-RDP), *HNRPL_HUMAN* (RDP), *HEXI1_HUMAN* (partial RDP), *RTCA_HUMAN* (partial RDP).

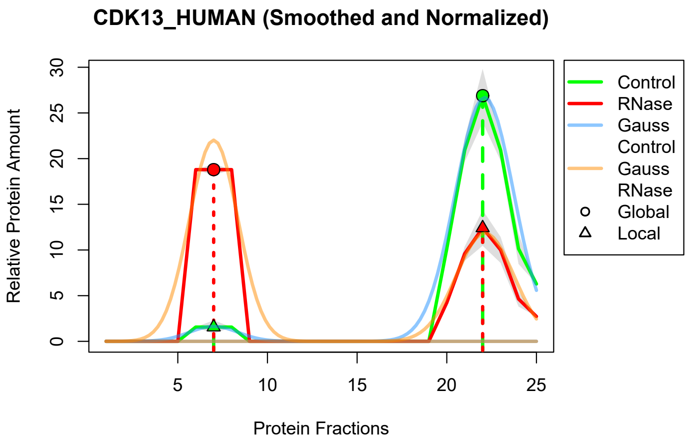

Figure A.6: New RDP with RNA associating. *CDK13_HUMAN*: Cyclin-dependent kinase 13. “Cyclin-dependent kinase which displays CTD kinase activity and is required for RNA splicing.

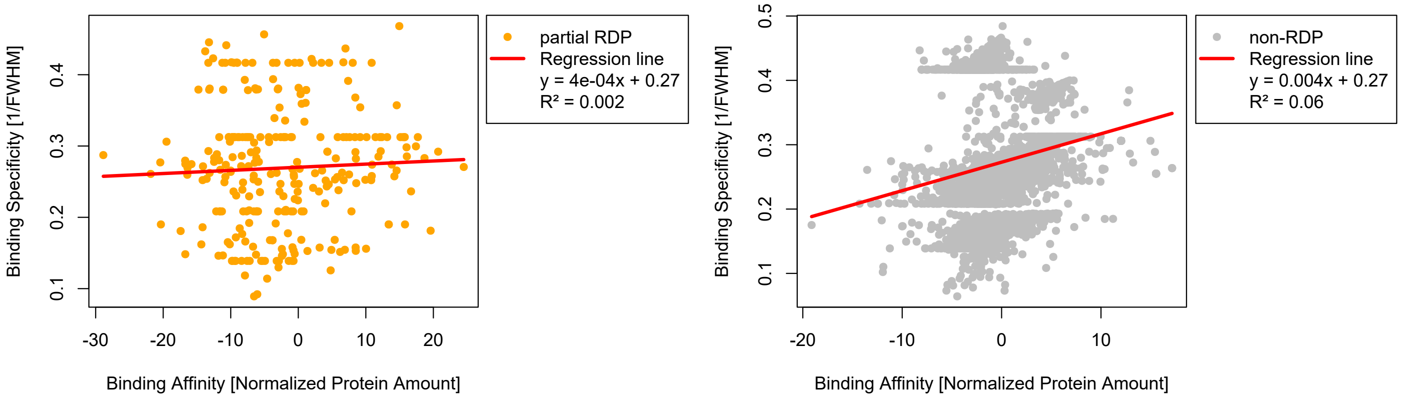

Figure A.7: Linear Regression for Partial RDPs and Non-RDPs. The linear regression is applied to the binding affinity and the binding specificity for only partial RDPs (left) and for only non-RDPs (right). Both linear regressions have considerably smaller Rˆ2 values compared to the linear regression of only RDPs.