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

Supervisor: Dr. Maïwen Caudron-Herger

Tutor: Fabio Rauscher

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 the investigation of protein RNA-dependence, meaning that not only RNA-protein interactions, but also RNA-protein-protein interactions are detected. Untreated and RNase-treated cell lysates are separated on a sucrose gradient into fractions. The proteins in these fractions are analyzed by 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 obtained data was used to define a set of selection criteria. The optimal selection criteria for identifying RDPs were determined using principal component analysis combined with manually determined shifting behavior. For each optimal selection criterion, 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 performed to investigate if the mass spectrometry data could be used to examine *in vivo* parameters such as the specificity of RNA-protein interactions.

This analysis shows how important bioinformatical methods are in the identification of RDPs. Future work based on R-DeeP and similar methods 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 modifications, transport within 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).

Moreover, RBPs can be classified as a subcategory of RNA-interacting proteins themselves. The concept of RNA-dependence 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). In this project, we aim to identify more human RBPs and RBP-interacting proteins through R-DeeP.

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): Untreated and RNase-treated cell lysates were separated into 25 fractions on a sucrose gradient. The proteins in these fractions were identified and their amount quantified using mass spectrometry. This yielded the distribution of each protein over the 25 fractions with three replicates per fraction.

2.2 Data Cleanup

The original mass spectrometry dataset was split into two separate datasets containing the values for all RNase fractions and Control fractions, respectively. The following steps were performed on both datasets.

2.2.1 NA Values and Zero Rows

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

2.2.2 Pearson Correlation

The Pearson correlation was used to test the reproducibility of the data. In seven cases, one of the three replicates contained only zeros. These replicates were replaced with NA values as determining the Pearson correlation with these values would have been impossible. In five rows, two out of the three replicates contained only zeros. These genes were removed completely from the datasets.

The Pearson correlation r was computed as follows for each pair of replicates at a time 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. The genes for which all three correlation coefficients fell below a threshold of 0.95 were removed from both datasets. This applied to 1086 genes. If two out of three correlation coefficients were under 0.95, the replicate associated with both coefficients was replaced with NA. 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.

2.2.3 Mean and Standard Error of Replicates

After testing for reproducibility and deleting unreliable values, the arithmetic mean and standard error were 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 datasets, the Control and RNase datasets were smoothed separately. Otherwise, the determination would have been influenced by outliers and wrong peaks.

Smoothing was achieved by calculating the mean protein value across three neighboring fractions for each fraction 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 ($2 \leq x \leq 24$);* $\quad$ $f(x)$: *mass spectrometry value at fraction position x;* $\quad$ y: *smoothed value for fraction position x*

For fraction positions 1 and 25, the mean of the fractions 1 and 2 and accordingly 24 and 25 were used.

2.3.2 Normalization of Smoothed Data

In order to be able to compare different protein amounts, each protein of the smoothed datasets was normalized to a total sum of 100 across all fractions:

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

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 standard errors were scaled accordingly.

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 of each protein for the smoothed and normalized datasets were determined by identifying the fraction which contained the highest mass spectrometry value. For proteins that had two or three fractions with the highest value, the mean fraction position was used as the fraction position of the global maximum.

2.3.3.2 Local Maxima

The local maxima for the smoothed and normalized datasets were determined by comparing the protein amount of each fraction to the protein amounts in the four neighboring fractions. 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 local maxima consisting of two or three neighboring fractions with the same mass spectrometry value, the mean fraction position was calculated. To distinguish local from global maxima, all global maxima were subtracted from the local maxima dataset.

2.4 Gaussian Fitting

A Gaussian curve was fitted to the global peak and the highest local peak of each gene in the Control and RNase samples. A Gaussian curve $f(x)$ 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 peak and highest local peak, σ and h were estimated using the half width at half maximum (HWHM) and the maximum protein amount (MPA):

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

These estimated values were optimized by minimizing the sum of squares of the differences between the protein amount values of the Gaussian curve G and the smoothed and normalized protein distribution p ($min(\sum(G - p)^2)$). The optimized parameters were 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 the Identification of RNA-Dependent Proteins

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

2.5.1 Definition of Selection Criteria

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

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

the **absolute 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.

In addition, each protein was categorized as either *left*, *right*, *non*-shifting or *precipitated*. The shifting behavior was defined as following: $\Delta\mu = \mu(Control) - \mu(RNase) > +1$: left shift; $\Delta\mu = \mu(Control) - \mu(RNase) < -1$: right 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 Component Analysis

Principal component analysis (PCA) was applied to the percentage values of the selection criteria mentioned above for comparability. The variables regarding main local peaks only served for determining partially shifting proteins and were therefore not included in the PCA. The selection criteria which separated the proteins the most based on their shifting behavior in the plot of the first two principal components were chosen as the optimal selection criteria 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 RDP:

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

Furthermore, *partial* RDPs were defined as proteins that satisfy the first and second criterion and either the third or fourth criterion for RNA-dependence. More *partial* RDPs 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 RBP-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-RBPs and -non-RBPs were plotted in the biplot of the first two principal components from the PCA.

2.6.2.2 Cluster Comparison with RBP2GO Score and t-Test

By calculating the **mean RBP2GO score** for each cluster (non-/partial/RDP) previously identified by applying the four optimal selection criteria, the RNA-dependence of a cluster 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 as $p = 0.05$.

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 RBPs in the RBP2GO database.

2.6.3 Linear Regression

In addition, the question was raised whether the parameters calculated from mass spectrometry data can be used to examine the binding dynamics of RNA-protein interactions. Exemplarily, a linear regression was utilized to calculate how well of a predictor the binding affinity of an RBP is for its binding specificity. In this case, *binding affinity* was defined as the difference in protein amount between the Control and RNase sample at the fraction position of the Control global peak and *binding specificity* was 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 consisting of 7086 proteins was reduced to 5990 proteins. Local and global maxima were determined and Gaussian curves fitted to the peaks. As can be observed in the plots of the protein *"CTCF_HUMAN"*, the smoothing and normalization steps (*Fig. 1*) as well as the peak determination and Gaussian fitting (*Fig. 2*) were successful.
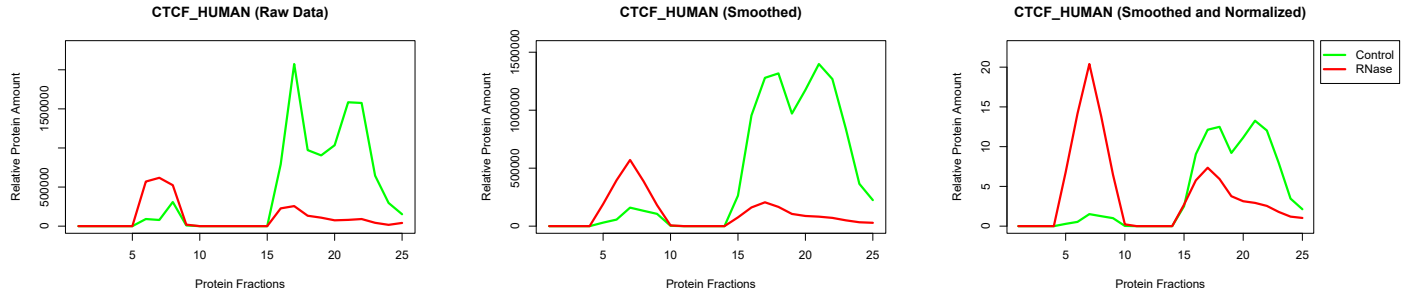
Figure 1: Data Cleanup. The protein distribution of the protein *CCCTC-binding factor (CTCF_HUMAN)* was smoothed and then normalized.
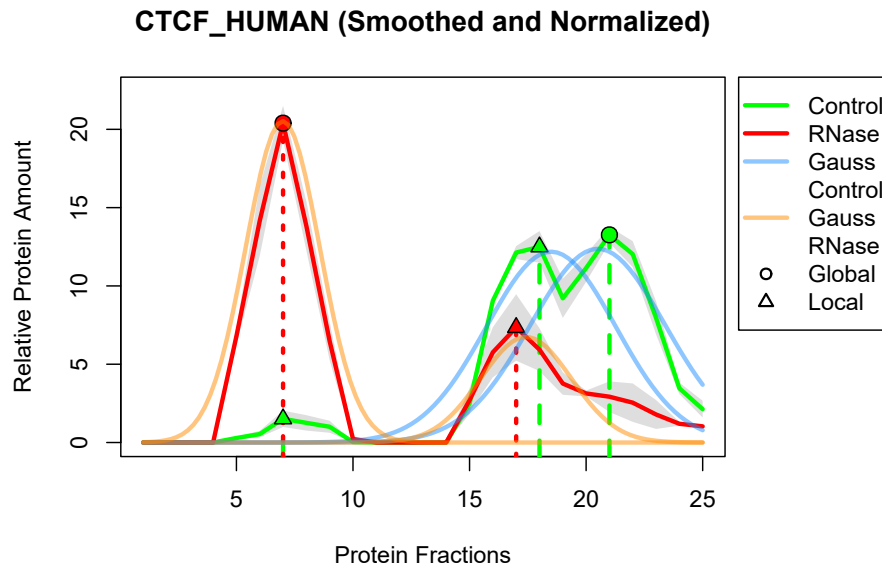

Figure 2: Local and Global Maxima and Gaussian Fitting. Local and global maxima of the protein *CCCTC-binding factor (CTCF_HUMAN)* together with fitted Gaussian curves. The gray area indicates the 95% confidence interval of the mean.

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 in 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. Most proteins are non-shifting (5153 proteins). The largest group of shifting proteins are left shifting (782 proteins). Only a few proteins are right shifting (41 proteins), and even fewer have precipitated (14 proteins).

6
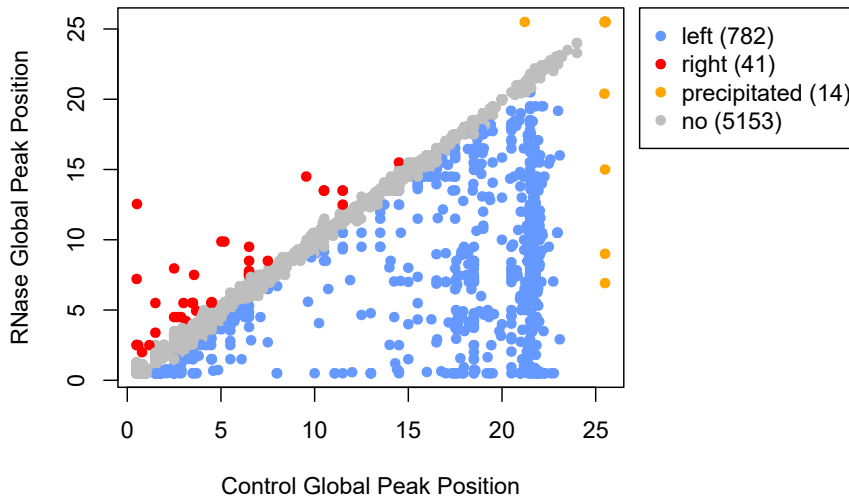
Figure 3: Shifting Behavior. All analyzed proteins were classified into left, right, non-shifting, and precipitated using the criteria mentioned in 2.5.1 and plotted using the global peak positions of the Control and RNase samples.

3.3 Results of Principal Component Analysis

In *Fig. 4 (left)*, the PCA was applied to all selection criteria and the first two principle components were plotted against each other (also included in *Appendix Fig. A.1*; for point densities see *Appendix Fig. A.2*). The different shifting behaviors of the proteins are shown by using different colors. Left shifting proteins are mostly separated from non-shifting proteins along the shift between global peaks (delta_Glb_mu_%), the overlap of the Gaussian curves (Gbl_overlap_%) and the protein amount lost or gained under global peaks after RNase-treatment (Ctrl_Gbl_prt_prop_% and RNase_Gbl_prt_prop_%). Therefore, these four criteria were defined as the optimal selection criteria. To illustrate this, *Fig. 4 (right)* shows the first two principle components plotted against each other when PCA is performed solely for the optimal selection criteria. As expected, the proteins separate into two rough clusters along these criteria.
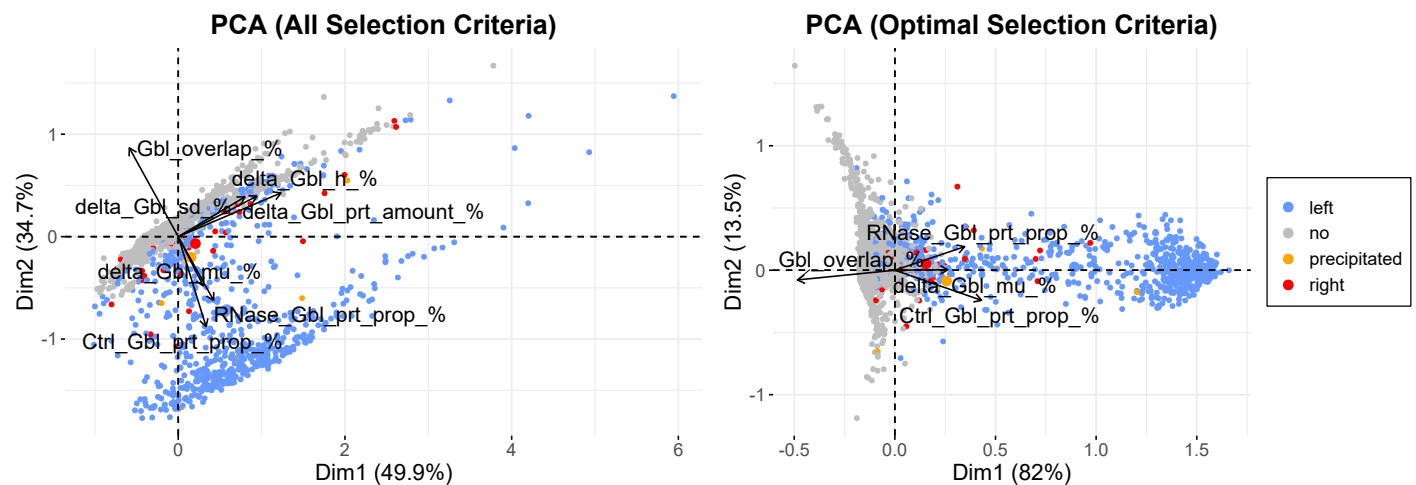

Figure 4: PCA Biplots. Principle component analysis of all selection criteria (left). Principle component analysis of optimal selection criteria (right). The first two principle components are plotted against each other, the shift classification of the proteins is shown via different colors.

7

3.4 Application of Selection Criteria and Determination of RDPs

Application of the optimal selection criteria identified 539 proteins as RDPs, 274 as partial RDPs, and 5177 as non-RDPs (individual proteins are illustrated in *Appendix Fig. A.4*). All proteins are plotted in a biplot of the first two principle components and colored according to their classification (*Fig. 5, left* and also *Appendix Fig. A.3, left*). Two main clusters are formed, one containing RDPs (blue) and one containing non-RDPs and partial RDPs (gray and orange).

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 RBPs 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 RBPs distribute homogeneously across the two clusters with no clear observable trend (see also *Appendix Fig. A.3, right*).
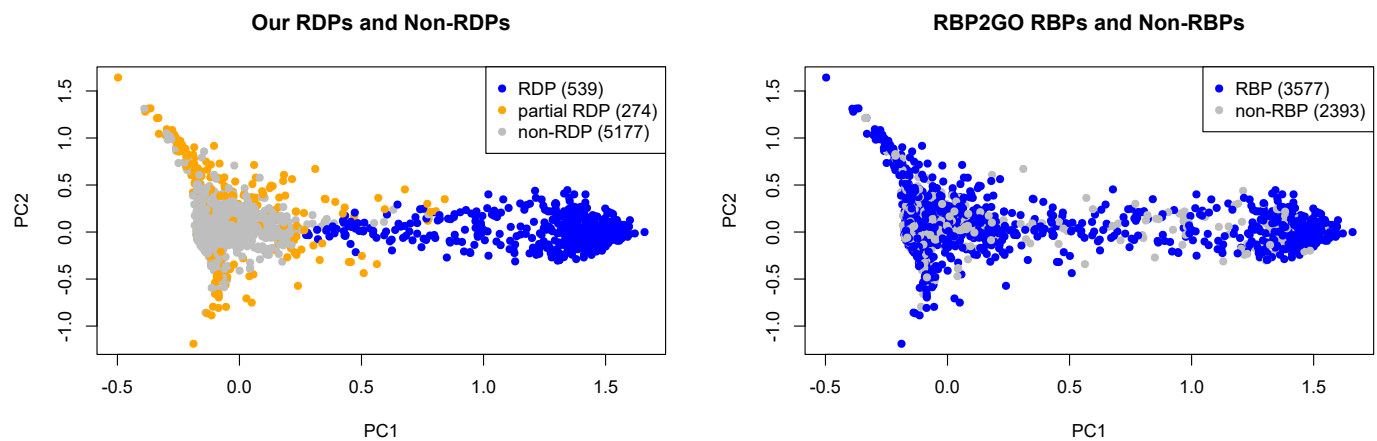
Figure 5: Identification of RDPs and Comparison to RBP2GO. The first two principle components of the optimal selection criteria PCA are plotted against each other. The classification of the proteins (RDP, partial RDP, non-RDP) is shown by using different colors (left). The classification of the proteins (RBP, non-RBP) from the RBP2GO database is shown by using different colors (right).

3.6 Cluster Comparison with RBP2GO Score and t-Test

The average RBP2GO score of each cluster was calculated. The average RBP2GO score of the RDPs is 21.5 with a standard error of 0.9, the average score of the partial RDPs is 13.6 with a standard error of 0.9 and the average score of the non-RDPs is 6.6 with a standard error of 0.1.
A two-sided Welch t-test was used to evaluate the significance of the differences between the mean RBP2GO scores: 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*).
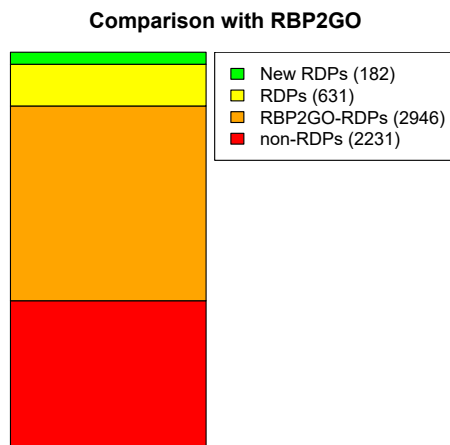
Figure 6: Bar Plot. The new proteins classified as RDPs via the performed analysis and within the RBP2GO database are illustrated in this bar plot.

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 increases with an increasing difference between the protein amount of Control and RNase samples at the global Control peak position.
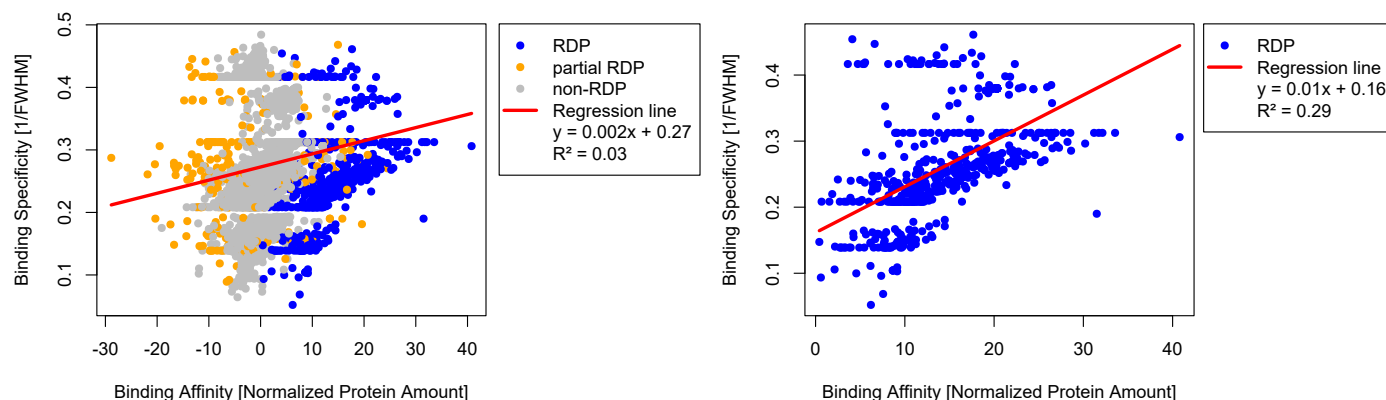
Figure 7: Linear Regression. A linear regression was applied to the binding affinity and binding specificity for all proteins (left) and only for the RDPs identified in this analysis (right).

4. Discussion

In conclusion, different parameters were calculated from the protein distributions provided by Caudron-Herger *et al.* for HeLa cells synchronized in interphase. Based on these parameters, the RNA-dependence of these proteins was determined.

One step in this analysis was smoothing (*Fig. 1*). Smoothing was performed to help with global and local peak determination because the averaged protein distributions were sometimes too jagged for an unambiguous peak determination (*Fig. 2*). However, this approach has three problems: Firstly, single

fraction peaks result, upon smoothing, in multiple peaks in neighboring fractions. This problem was fixed by calculating the mean fraction position for multiple adjacent peaks. Secondly, smoothing sometimes caused the fraction position of peaks to change. However, this change was very little (no more than one fraction position) and had no influence on analysis. Thirdly, 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 the classification of partial RDPs and applying them to the highest 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*).

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 database (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 database is the RBP2GO score. The mean RBP2GO score of all identified RDPs was 21.5, compared to 13.6 for partial 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 t-test. It is noticeable that the p-value of the t-test for the difference between the mean RBP2GO score values of RDPs and non-RDPs is by far the smallest (p(RDP, non-RDP) = 8.5e-54). However, these findings are of limited use because the RBP2GO score is calculated very differently from the analysis presented here (Caudron-Herger *et al.*, 2020a).
The identified RDPs were also compared directly to the RBP2GO database (*Fig. 6*): 2231 genes (37 % of all analyzed genes) which are categorized as non-RDPs are also listed as non-RBPs in the RBP2GO database. 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 database. These "new" RDPs include proteins which are known to be associated with RNA (e.g. *CDK13_HUMAN*, *Appendix Fig. A.5*) (Berro *et al.*, 2008).

The differences between the presented results and the RBP2GO database are possibly the result of differences in the human cell lineages and experimental methods used and different stages of the cell cycle analyzed. Notably, only the proteome present in interphase was analyzed here, so suboptimal comparability with the RBP2GO database, which draws its data from many orthogonal experiments, should be expected.

Lastly, a linear regression was performed to test whether molecular interaction parameters could be determined based on the mass spectrometry data. One could assume that an RDP which binds tightly to its RNA target also forms very specific structures and, therefore, has a high binding specificity. This relation was somewhat observed when applying linear regression modelling only to the identified RDPs (*Fig. 7, right*). A linear regression of all proteins, partial RDPs or non-RDPs result in a lower coefficient of correlation (*Fig. 7, left* and *Appendix Fig. A.6*). However, the experiment was not set up to investigate the kinetics of RNA-protein interactions. Other methods that have been developed especially to investigate the kinetics of RNA-protein interactions would be more suitable (see for instance Sharma *et al.*, 2021).

The results of this analysis show that, besides the experimental methods used, bioinformatical analysis also plays a major role in determining the RNA-dependence of proteins. Further improvements in analysis of 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. Mol. Cell *75*, 184-199.

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 Res. *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. Nat. Protoc. *15*, 1338-1370.

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

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

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

Jankowsky, E., and Harris, M.E. (2015). Specificity and nonspecificity in RNA–protein interactions. Nat. Rev. Mol. Cell Biol. *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. Nat. Rev. Mol. Cell Biol. *8*, 479-490.

Sharma, D., Zagore, L.L., Brister, M.M., Ye, X., Crespo-Hernández, C.E., Licatalosi, D.D., and Jankowsky, E. (2021). The kinetic landscape of an RNA-binding protein in cells. Nature *591*, 152-156.

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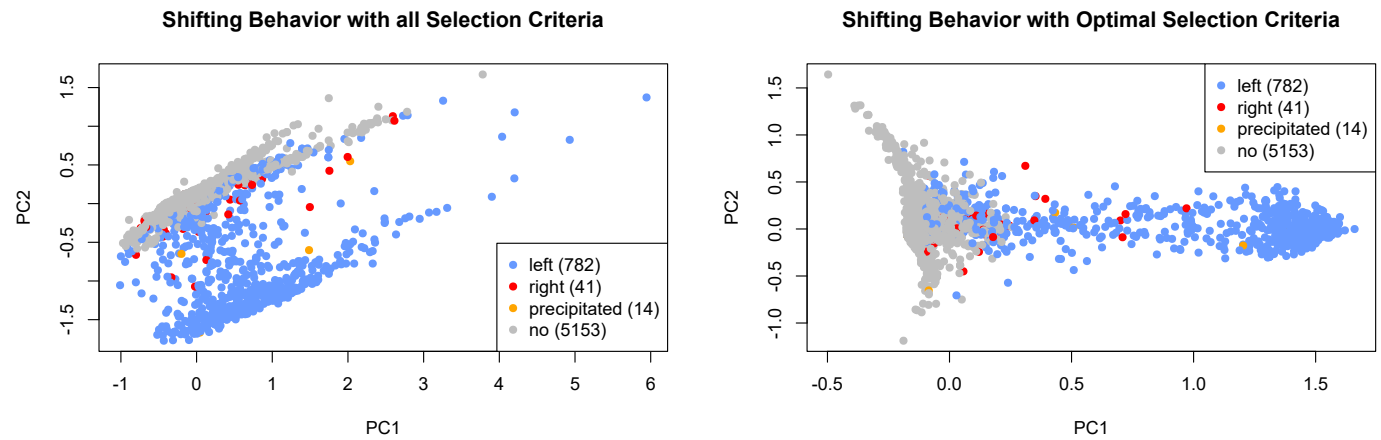
Figure A.1: PCA with Shifting Behavior. The first two principle components of the PCA applied to all selection criteria (left) and of the PCA applied to the optimal selection criteria (right) are plotted against each other. The shifting classification of the proteins is shown by different colors.
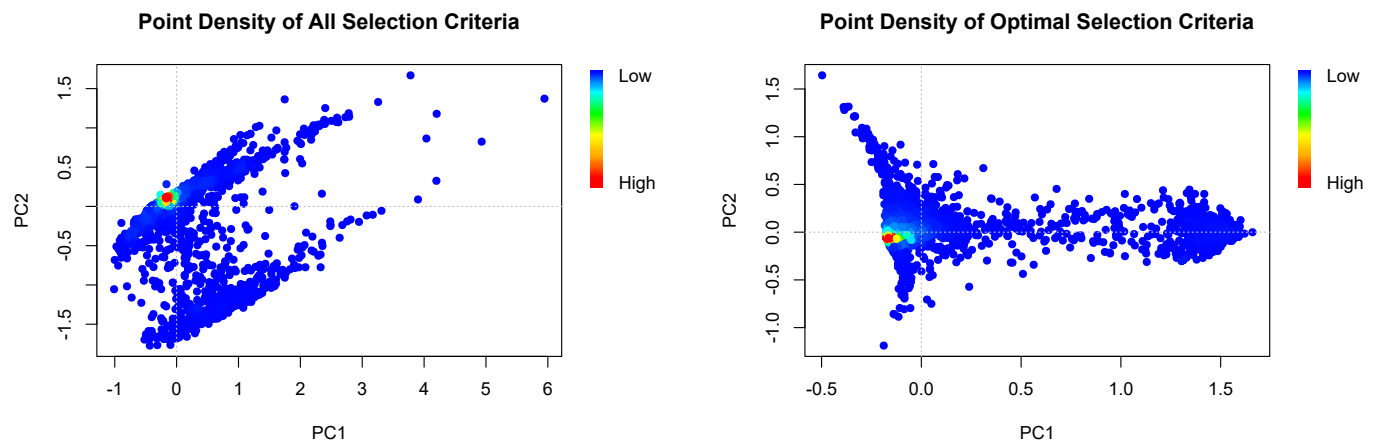

Figure A.2: PCA Densitiy Plots. The first two principle components of the PCA applied to all selection criteria (left) and of the PCA applied to the optimal selection criteria (right) are plotted against each other. The point density of the proteins is shown by the usage of different colors.
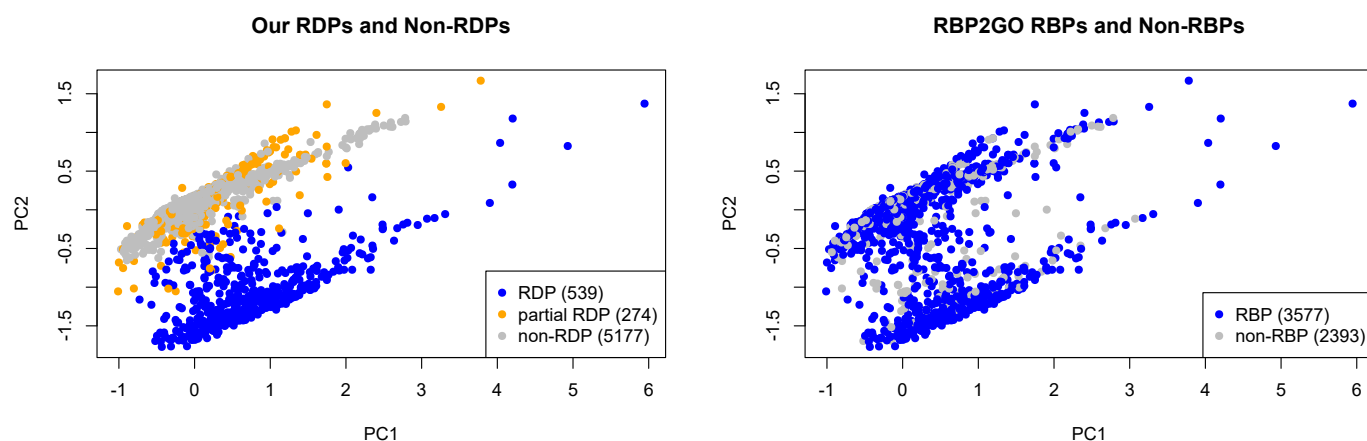
Figure A.3: Identification of RDPs and Comparison to RBP2GO. The first two principle components of the PCA for all selection criteria are plotted. The classification of the proteins (RDP, partial RDP, non-RDP) achieved in the performed analysis is shown by using different colors (left). The classification of the proteins (RDP, non-RDP) from the RBP2GO database is shown by using different colors (right).
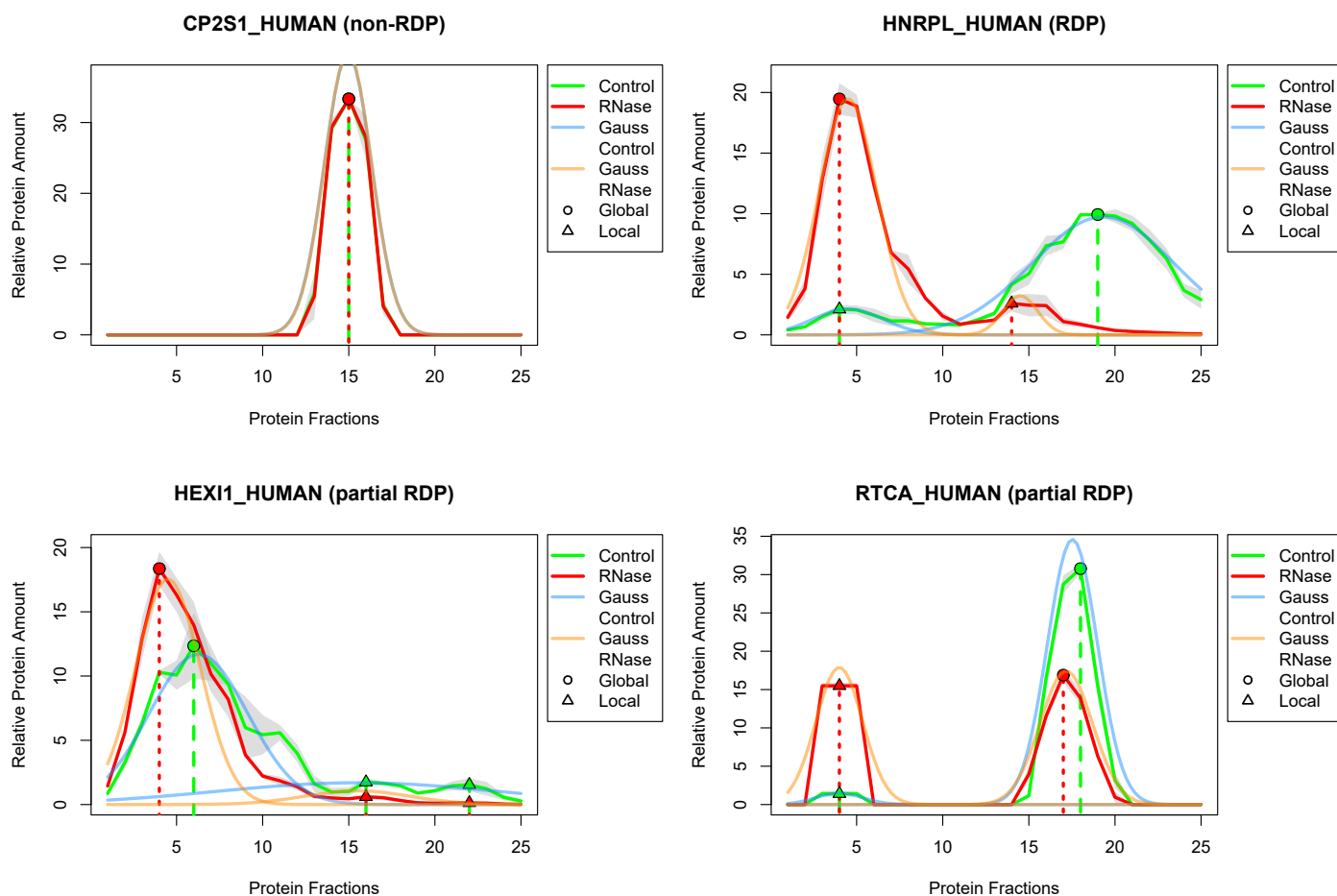
Figure A.4: Application of Selection Criteria. Protein and Gaussian curves for the Control and RNase sample for the following four different proteins are plotted for comparison: *Cytochrome P450 2S1 (CP2S1_HUMAN)* (non-RDP), *Heterogeneous nuclear ribonucleoprotein L (HNRPL_HUMAN)* (RDP), *Hexamethylene bisacetamide-inducible protein 1 (HEXI1_HUMAN)* (partial RDP), *RNA 3'-terminal phosphate cyclase (RTCA_HUMAN)* (partial RDP).

13
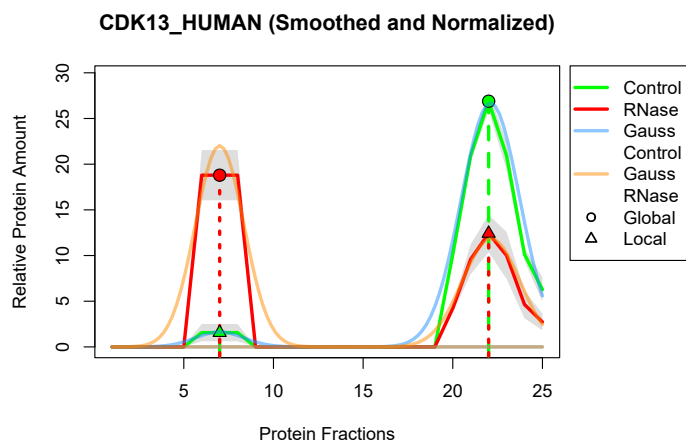
Figure A.5: New RDP with an RNA Association. *Cyclin-dependent kinase 13 (CDK13_HUMAN)* is involved in many key cellular RNA processes such as splicing and transcription (Berro *et al.*, 2008).
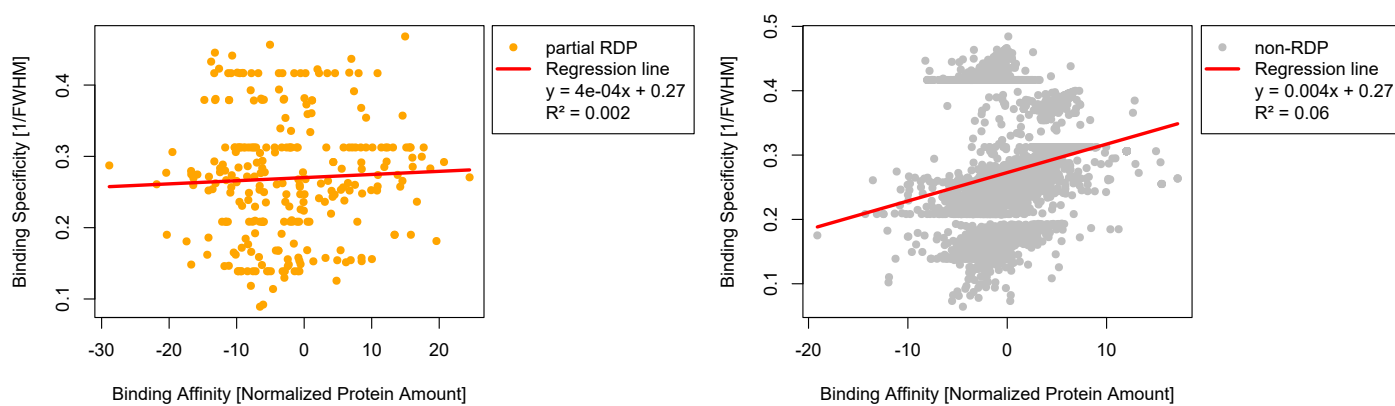
Figure A.6: Linear Regression for Partial RDPs and Non-RDPs. Linear regression was applied to the binding affinity and the binding specificity for only partial RDPs (left) and for only non-RDPs (right). Both linear regressions have considerably smaller R^2 values compared to the linear regression of only RDPs.