Data Analysis Project
Summer Term 2023

Proteome-wide Screen for RNA-dependent Proteins
HeLa synchronized cells in Mitosis

Topic 3, subtopic 2

Authors:
Verena Kantelhardt
Alexandra Kornelius
Jimena Orrego Chong

Supervisor: PD Dr. Maïwen Caudron - Herger
Tutor: Fabio Rauscher
Submission: 17th July 2023

Contents

1. Introduction

RNA- binding proteins (RBPs) are proteins able to interact with double or single stranded RNA and form a ribonucleoprotein complex (RNP) (Caudron-Herger *et al.*, 2019). They are able to change composition depending on situation, as RNA functional state, maturation or cellular context (Gebauer *et al.*, 2021).

In the last years, RBPs have gained importace, as they have been validated in cellular and molecular homeostasis (Weisse *et al.*, 2020) and in regulation of various RNA tasks, like transcription, splicing, modification, intracellular trafficking, translation and decay (Gebauer *et al.*, 2021).

RBP mutations have also been shown to cause several diseases (Kelaini *et al.*, 2021), including cancer (Qin *et al.*, 2020) and neurodegenerative conditions (De Conti *et al.*, 2017). In the case of cancer, HuR (also known as ELAVL1) is known to be an established regulator of post-transcriptional gene regulation. It is overexpressed in most human cancers (Schultz *et al.*, 2020) and is upregulated in breast cancer promoting tumorigenesis by regulating numerous proto-oncogenes, growth factors, and cytokines and is thereby supporting invasion and metastasis (Wu *et al.*, 2020). Moreover, the discovery of mislocalized hnRNP A1 in neurons in brains of multiple sclerosis patients shows that dysfunctional RBPs may play a role in the neurodegeneration of this disease (Salapa *et al.*, 2020). Identifying RBPs and later their role might contriubte to understand mechanisms and targets of different diseases and therefore help in the development of new drugs.

In this project we aim to identify RBPs through the analysis of mass spectometry data, after which we can categorize these proteins with high or low confidence as RBPs.

The analyzed data derives from lysed HeLa S3 cells in mitosis in presence or abscence of RNase after being loaded into sucrose density gradients. Different protein migration due to density differences allowed the observation of a specific protein distribution throughout the gradient (Caudron-Herger *et al.*, 2019). This distribution difference enabled the analysis of several shifts, after adding RNase, in order to identify the RNA dependence of each protein.

The data for the protein amount per fraction (per replicate) was collected using mass spectrometry and stored in a .csv-file. From this dataset, we aim to confidently identify RBPs using bioinformatic methods in R. Beyond that, we cross-referenced our results with the RBP2GO database (Caudron-Herger *et al.*, 2020) and developed a linear regression model to identify RBPs without relying on an entire data analysis protocol.

2. Material and Methods

2.1. Data description and sorting

We started off by analyzing the structure of the given dataset, then reorganized it by splitting it in two and followed by excluding experimental errors such as negative or rows with only zero values.

2.2. Check reproducibility with t-test

To ensure the integrity of our subsequent analyses, we assessed experimental reproducibility on the data we utilized. Using a student t-test, values which significantly differed from the mean within the three replicates were identified. The mechanism behind this is as soon as the values differ significantly, the variability within the sample becomes larger and in turn leads to a higher p-value. Vice versa, a low variability within the sample results in a smaller standard error and thus a lower p-value. As

representative for each row the maximum p-value was chosen, as it contains the maximal possible error per protein. Deviations greater than 10% from the mean were considered non-reproducible, leading to data exclusion. The general principle of our t-test reproducibility check is shown in R-code in Appendix Figure 1A and B.

2.3. Data scaling and normalization

After ensuring that all remaining data was highly reproducible, the data needs to be scaled. We decided to use data scaling to 100. As the first step the mean of all triplicates is calculated and subsequently used to set the total protein amount to 100 for each protein, separately for Control and RNase. This was done by dividing each mean of a triplicate by the total sum of means for each protein, lastly multiplied by 100. In essence, the normalization of our data resulted in adjustments to the absolute protein levels in each fraction while preserving the relative protein distribution across all fractions.

2.4. Shift finders

2.4.1. Max- shift

The primary approach to distinguish RBPs from non-RBPs is the max- shift method. This method involves identifying the fraction with the highest protein amount, separately for the Control and RNase treated protein samples. We subtracted the column indices of the RNase fraction with the maximal protein amount from the column indices of the Control fraction with the maximal protein amount, separately for every protein. If the result equaled 0, 1 or -1, the protein was evaluated as non-RBP, as no significant shift was detected. If the number was larger than 1, the protein showed a left shift. Analogue to this classification, if the value was smaller than -1, the protein exhibited a right shift in its maximum amount. In essence, we used this formula:

$$max\_shift = Ctrl\_fraction\_max - RNase\_fraction\_max \tag{1}$$

2.4.2. Y- shift

To include RNA-dependent proteins that did not exhibit maximum shifts, an additional method for analyzing proteins with a max- shift of 1, 0, or -1 was applied. Focusing on the relative protein amount at the maximum peak (the y-value), proteins were identified as a partial shifter if their maximal RNase protein amount differed by more than 20% from their maximal control amount.

2.4.3. Local shift

A second method that includes different protein shifting criteria is identifying all local peaks per protein, including the maximum peak. In a subsequent step, both the number and location of these peaks are compared. If the number of peaks varied between the control and RNase samples, the protein was classified as an RBP. Additionally, if the column indices of the local peaks varied, the protein was also classified as an RBP. This method allowed a more comprehensive analysis of the protein distribution.

$$local\_shift = \begin{cases} \text{different number of local peaks} \\ \text{different column indices of local peaks} \end{cases} \tag{2}$$

2.4.4. Curve shift

However, all methods before focused on the Control and RNase peaks of each protein. To compare the entire protein distributions, quotients were formed of every fraction of a protein. In this step every y-value of the control sample was divided by every y-value of the RNase sample. This new parameter allows the observation of all protein amount values.

$$curve\_quotients = \frac{(\text{all Control values} + 1)}{(\text{all RNase values} + 1)} \tag{3}$$

A value close to 1 indicates a close match between both samples, while larger values indicate a peak in Control (Control value is larger than the RNase value). Following the same principle, values close to 0 indicated a peak in the RNase sample (RNase value is larger than Control value), all relative to the other sample. We named this method curve shift, as it analyses changes in the entire protein amount curve. The +1 is necessary to avoid division through 0. By adding it to both sides, there is no change to identical values and only a small change to non identical values. Following the formation of quotients for every fraction, an upper and lower bound of 70% (1.7 and 0.588) was implemented, so the protein quotients needed to differ more than 70% from 1 to be considered as an RBP.

2.5. Statistical analysis

2.5.1. K- Means

As a statistical test, we perform K-Means on the results of two of our peak finders (max- shift and curve shift). In all cases the general procedure was the same. First, both data frames were tested for the optimal number of clusters using the silhouette method, which determines the amount of optimal clusters using the highest silhouette value s_i. The silhouette value is calculated using the following formula:

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)} \tag{4}$$

a_i: the average distance between the member of one cluster to all members of all other cluster
b_i: the smallest average distance between the member of one cluster to all members of all other cluster

After knowing this number, the K- Means clustering can be performed on the data, which sorts data of different dimensions and points into clusters. As a first step, a defined number of cluster-centers is determined using s_i value. Then all points are sorted into the cluster whose center is closest. Next, in each cluster, new cluster-centers are determined at the mean of all values in each cluster. Then all points are sorted to the respectively nearest cluster-center again. That is repeated until the position of the cluster and points within don't change anymore (Hartigan and Wong, 2018).
K-Means will be performed with the optimal amount of clusters and potentially another number of clusters.

2.5.2. Linear Regression

Firstly, it is assumed that our data has a linear relationship, following the formula:

$$Y = a + b \times X \tag{5}$$

a: Intercept

b: Slope

Y, X: Value on respective axis

This is visualized by fitting a linear slope into a graph where the axes are the different data series. To figure out how linear the assumed relationship actually is the R^2-value is calculated which is the squared correlation:

$$R^2 = \frac{\text{Variance(predicted values)}}{\text{Variance(measured values)}} \tag{6}$$

The closer R^2 is to 1, the better the models predictions are. Here the linear regression is performed on the data gained from the max- shift and the curve shift (Lunt, 2015).

3. Results

3.1. Data description and sorting

To start our analysis, the dimension of our dataset was recorded (7159 x 150) and we confirmed that all values are numeric. The check for experimental errors (for example proteins which escaped mass spectrometry detection and thus had an entire row equal to zero) resulted in the removal of one row from the dataset. There were no negative values in our dataset. To facilitate further analysis, the dataset was split into two, one containing all control values and the other one all RNase values. A repeated check for rows equal to zero led to the deletion of two additional proteins, finally leaving 7156 proteins in our dataset (99.96% of original dataset).

3.2. T- Test

The next step was to ensure the reproducibility of the dataset using a t-test. Using the previously described method and a significance level of 10%, 2130 rows were identified with a significant deviation in the triplicates and were excluded from our dataset of 7156 proteins, leaving us with 5026 proteins (70.3% of original dataset). More proteins in the RNase dataset had a maximum p-value exceeding 10% (1904 in comparison to 550 proteins in the Control dataset).

3.3. Normalization

To facilitate a comparison all protein values were scaled to a relative percentage scale. Even after converting the absolute protein amount to the relative protein amount, the overall trend in the distribution of the protein amount stays comparable. This is visualized at a randomly chosen protein, CASP7_HUMAN (see Figure 1 A and B). Another option tested was MinMax scaling. In the results a lost of height difference between the y-values was recorded ocasionally. This effect is also visible in CASP7_HUMAN (see Appendix Figure 1D and E). To conserve most of the original data distribution, the following steps were executed with data which was scaled to 100.

3.4. Shift finders

3.4.1. Max shift

Using a margin of more than one shifted fraction between the Control and RNase sample, 581 proteins were identified as RBPs (11.6% of all analysed proteins). A protein which visualizes the max- shift method is ABCF1_HUMAN, as seen in Figure 1C.

3.4.2. Y- shift

Through the additional y- shift, which was applied to all non-RBPs from the max- shift, an additional 503 proteins were identified as partially RNA dependent (partial shifts). These were merged with the max shift results, so that all could be summarized in one graph. An example protein which was labeled as non-RBP by the max shift but shows shifting characteristics (significant changes between the control and RNase sample distribution), is ABCB7_HUMAN, shown in Figure 1D. An application of the y- shift to all proteins, not only to the non-max- shifters, can easily lead to wrong classifications, as visualized in HNRPC_HUMAN. This protein shows sure shifting characteristics, but since both maximum peaks are almost the same height, the y- shift would assign it to non-RBPs (see Appendix Figure 1C). Thus, the y- shift was only used on proteins with no max- shift.

3.4.3. Local shift

By applying the condition that RBPs have either different number or localizations of peaks, a total of 1570 proteins were identified as RBPs. Several proteins which were not detected by the combined max- and y-shift but exhibited sure shifting characteristics were identified in this method. One example for this would be SQOR_HUMAN (see Figure 1E). None of the previous methods (max- shift or y-shift) would have identified this protein as a shifter, even though it has been shown to exhibit an RNA dependent shift and is classified as RBP (Caudron-Herger *et al.*, 2019).

3.4.4. Curve shift

Using the curve shift method to form a quotient and setting a threshold of 70% deviation from 1, 1241 proteins were identified as RBPs. Out of these, 194 proteins were not identified by the three methods (max- shift, y- shift and local shift) before. One example of such a protein is shown in Figure 1E, the protein PELO_HUMAN. Again, it is a confirmed shifter (Caudron-Herger *et al.*, 2019), however not identified in any previous method. The curve shift identified a large shift in the fraction on the right of the peak and thus classified the protein as a shifter. The difference in the height of the maximum peaks was not significant enough to be identified by the y- shift. Another example is the protein PHLB2_HUMAN shown in Appendix Figure 1F, where the only significant difference between Control and RNase treated sample was a y- shift in a local peak, which the curve shift identified. For further analysis, we took the quotient with the highest deviation from 1 for each protein and saved it to a new dataframe.

3.4.5. Identified RBPs and non RBPs

To summarize all findings from our shift analysis in one graphic, we used a Venn Diagram to illustrate all overlaps. 492 proteins are identified as RBPs by all methods combined, and another 656 proteins by at least two methods (see Figure 1G). To visualize the partial shifters, the y- shift was separately plotted as an addition in a new Venn Diagram in Figure 1H. Here it is clearly visible how the y- shift is an addition to the max- shift, as it has no overlaps at all. All overlaps between the y- shift and the other shift analysis methods show that both curve shifters and local shifters identify some partial shifters, but do not directly differentiate between absolute and partial shifter Important to note here

is that the local shift method shows the least overlap with any other method, it solely identifies 678 proteins as RBP, which were all classified as non-RBPs by the other methods.

3.4.6. Comparison of methods with known RBPs

Following the comparison of our shift-finders with each other is the comparison of our identified RBPs to previously identified RBPs. Therefore, our results were cross-referenced with the RBP2GO database (Caudron-Herger *et al.*, 2020). All known RBPs were loaded into a data frame. The goal is to check how many of our identified RBPs are already found in a publicly known RBP dataset, by checking the overlap of the the row names in the datasets. The results for each shift-finding approach separately can be found in Appendix Figure 1 G-I, a summary can be seen in Figure 1 I. Important to mention is that our max- shift (including the additional y- shift) had the highest percentage of newly identified RBPs (27.12%). On the other hand, the curve shift showed the highest conformity with the RBP2GO database, newly identifying 22.32% as RBPs (from all proteins identified as RBP via curve shift). In between lies the local shift with 23.25% newly identified RBPs. In total, 82.5% of proteins identified by all three shift finders and 77.7% of proteins identified by two or more shift finders are previously identified RBPs.

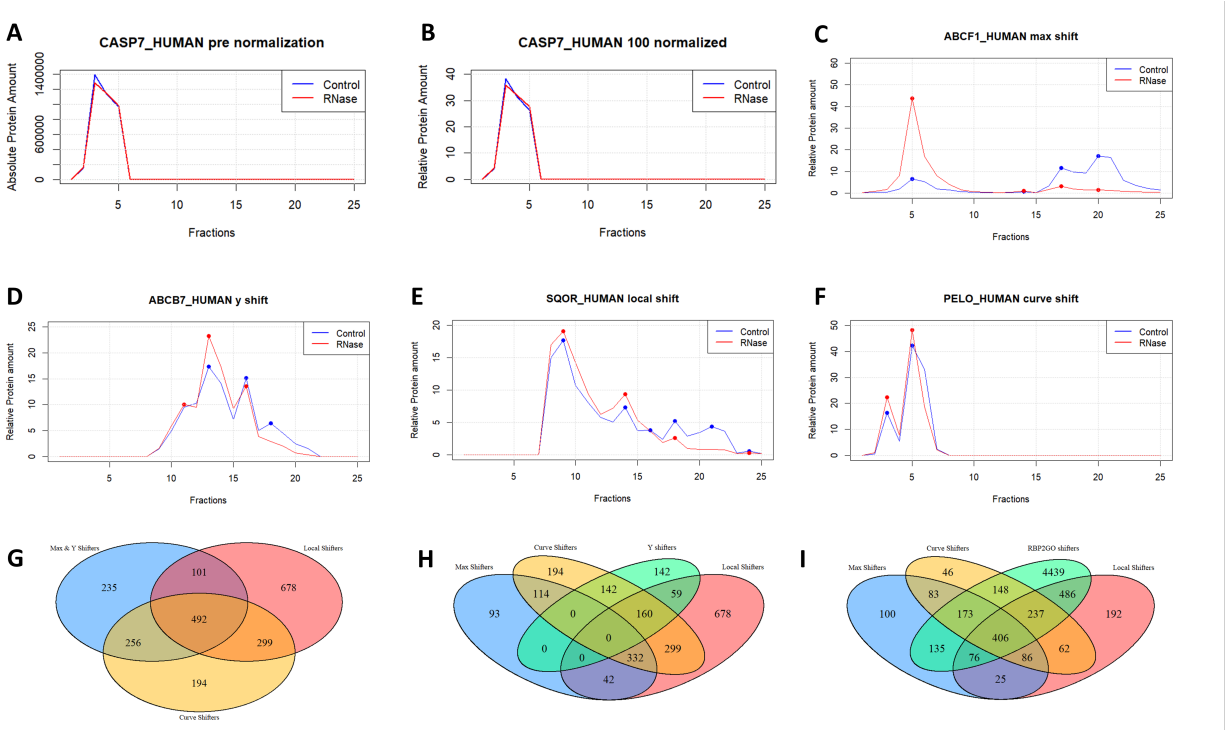

Figure 1: Results of used methods: Protein pre and post normalization (A,B). Visualization of the max-shift (C), y-shift (D), local shift (E), curve shift (F). Overlaps of all methods using a Venn Diagram (G,H). Overlaps of all methods with RBP2GO database using a Venn Diagram (I).

3.5. Statistical analysis

3.5.1. K- Means

In order to cluster RBPs identified with the max- shift and curve shift, the silhouette method determined in both cases that 2 clusters are the optimal number of clusters as seen in Appendix Figure 2A and B. However, in the following we performed K-Means with two clusters as well as with 3 clusters.

Seen in Figure 2 are the results of the K-Means clustering. For the clustering of max- shifters, the K-Means clustered proteins found in lower and in upper fractions (See Figure 2A, Figure 2C). And as shown in Figure 2B clustered together proteins that had a low max- quotient and those with a high max- quotient.

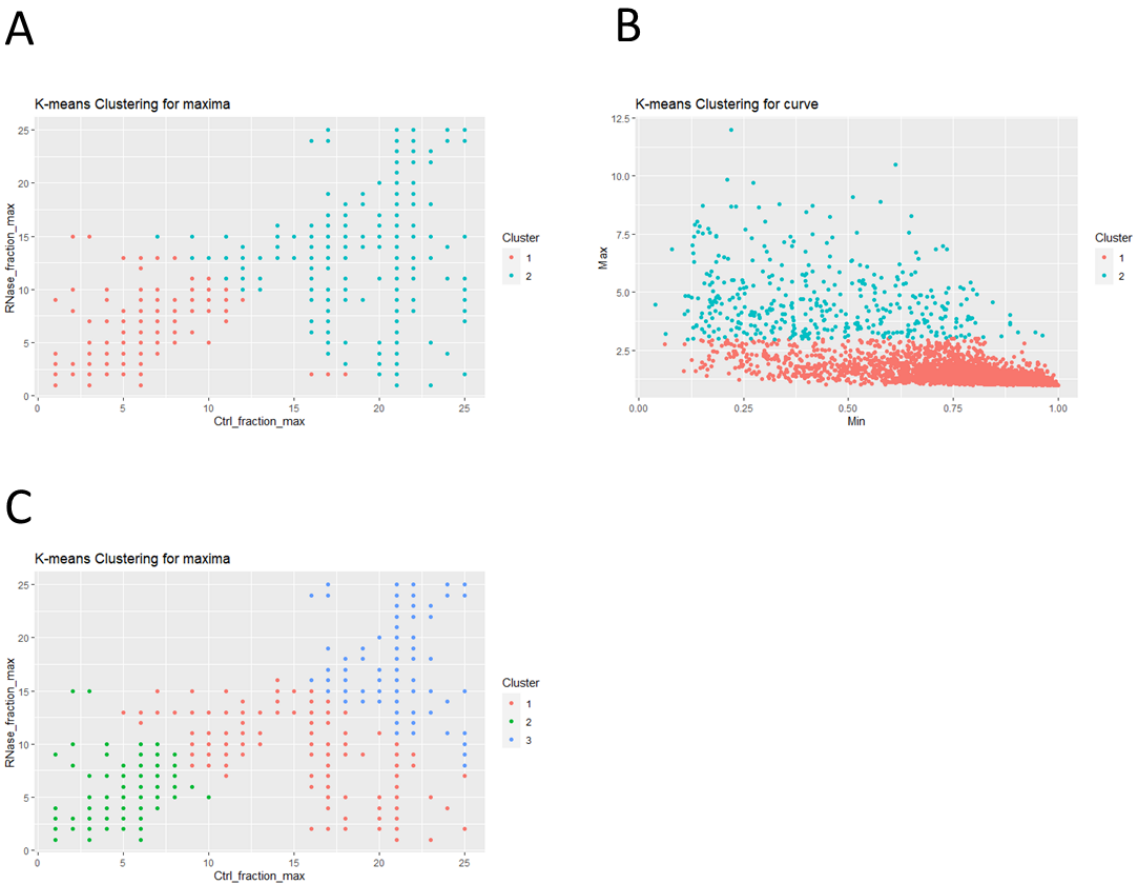

Figure 2: K-Means clustering: Performed on different data with different number of cluster; K-Means clustering was performed on the dataframe gained from shifters in maxima (A and C) and on the dataframe created of curve values (B). The clustering was performed with two (A and B) and three clusters (C) with the maxima shifters.

3.5.2. Linear Regression

It was figured that the linear relationship between max- shift and curve shift is better than to local shift The result of the linear regression show that most values do not lay on the linear slope, as shown in figure 3A . That leads to an R^2 value of 0.2658505. The p-value is 2.2×25^{-16} and the root mean squared error is 1.858949. The diagnostic QQ-plot shown in figure 3B visualizes that the residuals do not lay on the theoretical linear slope shown in blue.

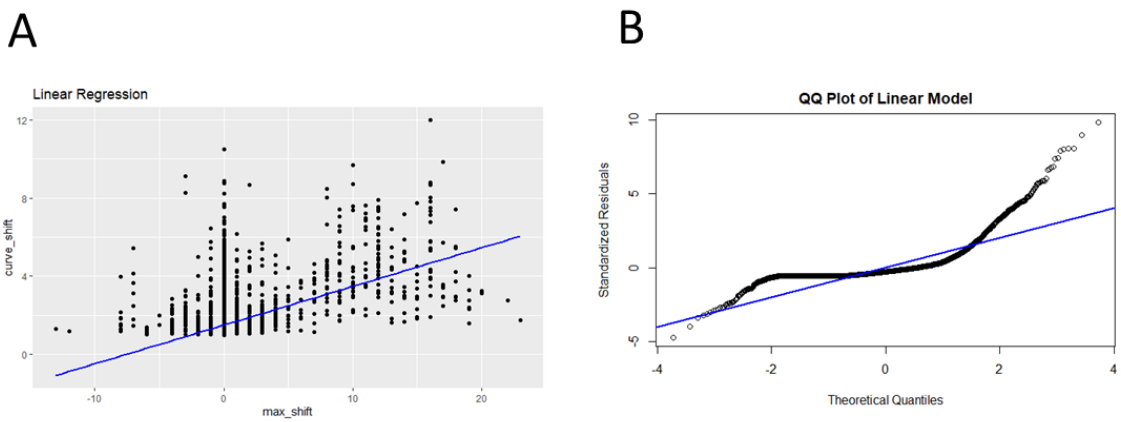

Figure 3: Results of the linear Regression: The fitting of the linear slope on the curve shift and the max- shift data (A). The diagnostic QQ-plot showing the residuals in comparison to a linear slope (B).

4. Discussion

RNA-binding proteins (RBPs) directly interact with RNA molecules and play a crucial role in maintaining cellular functions such as RNA processing and translation. Dysfunctions in RBPs are associated with diseases such as Multiple Sclerosis or cancer, for which the identification of RBPs and their complex interactions may help to optimize treatment. In this study, we aimed at confidently identifying RBPs in synchronized HeLa cells in mitosis.

First of all, given triplicates of every protein provided a stable base and good opportunity to analyze the reproducibility using a t-test. Due to the small sample size of three replicates, we decided on a 10% significance level to allow for a more flexible threshold. This increases the statistical power by reducing false negatives (Type II error). However, it also increases the possibility of false positives (Type I errors). The issue could be solved with a higher sample size; however, this always comes with a considerable financial cost. As only the maximum p-value per row was regarded, one repetition with high deviation sufficed to exclude the entire protein, which gave us very viable data to continue with the data scaling and the shift identifications.

As expected, more fluctuation was found in RNase than Control triplicates. This underscores that the effect of RNase treatment on proteins can be diverse, potentially leading to varying results. Using a t-test to confirm a similar observed effect in all triplicates proves crucial in our analysis to avoid inaccurate conclusions. However, in total much data was lost or left uninterpreted. To avoid this, a higher significance level could be set, however this is always at the risk of including more erroneous results.

Furthermore, we did not use any smoothing techniques such as a Gaussian fit, and solely relied on the 25 discrete fractions. This simplification leads to the loss of a more accurate position of the peaks (if they lie between fractions), possibly causing RPBs to avoid peak detection. Using a gaussian fit would allow a more realistic interpretation, however it also leads to some generalization and would have hugely complicated later shift analysis.

One major source of uncertainty arises from the different criteria chosen to define RNA-dependent proteins. Here, we used four methods in total: max- shift, y- shift, local shift and curve shift. We used the y- shift to improve the max- shift, as both methods separately cannot consider some important shifting characteristics but complement each other nicely. The y-shift margin was set at 20%, as this threshold allowed us to confidently avoid false positive detections while capturing significant y-shifters. We decided that all proteins identified with this y- shift are considered partial shifters, as the column of the maximum peak was unchanged. This means that the highest protein amount still accumulated in the same fraction, thus RNA probably plays a rather minor role in this protein.

A potential flaw in our local shift is that it is most likely to be influenced by small fluctuations. If any of the peaks vary in just one fraction, it would be identified as shifter. The max- shift avoids this by having a margin of more than one shift. However, the local shift allows a more comprehensive analysis in the total protein distribution. As seen in the high conformity of our local shift with known RBPs, the local shift constitutes an important analysis approach.

For the curve shift, the margin was set to 70%. This is a quite high but reliable margin, as it allows to confidently avoid identifying false RBPs by chance. Using a 25% margin would e.g. result in a significantly higher amount of identified RBPs, however the possibility to falsely include statistical fluctuations rises considerably. By the addition of +1 to all Control and RNase values to avoid dividing through zeros, we produce an error as we change the proportions. This error would be reduced if instead of adding +1, we would only add +0.1 or even smaller values to both sides. However, we didn't realize this until later on, but it would be a possible optimization for future analysis. Overall, the curve shift has a unique approach to shifting characteristics, as it focuses on the entire protein distribution, not only the peaks.

In summary, 1148 proteins were identified by two or more shift finders, and we consider all those as high confidence RBPs or high confidence partial shifters. However, not all have been previously identified, 22.3% are newly identified. A possible reason for this difference between our results and public data is that we used data from cells synchronized in mitosis, and RNA interaction may depend on different cell cycle stages. Therefore, using this information, our new results possibly aid to a more comprehensive analysis of protein behavior during mitosis.

In statistical analysis, after the usage of K- Means on max- shift and curve shift data, hope was held that it would be possible to automatically differentiate between left-shifters, right-shifters and non-shifters. Two clusters are shown as that is the optimal number of cluster and three in the case of the max- shift in hopes to enable a differentiation. After revising the results gained from the different K-Means, it became clear that differentiation is not possible using K-Means on these dataframes. Thus, it is not possible to replace our workflow following the peak finders to identify RBPs using K-Means clustering. Changing the K-Means algorithm did not influence the results significantly.

K-Means clustering of local shift and y- shift is not shown, as the results were both unable to differentiate RBPs and non-RBPs.

Before performing linear regression, data with a possibly good linear relationship had to be chosen.

Hereby, the decision fell on the max- shift and the curve shift, as a max- shift would likely mean that the maximal curve shift values are far from 1. The stronger the difference of protein amounts at the maxima of RNase in comparison to Control, the further the quotient value of the curve shift deviates from 1. To verify that further it was also checked whether the local shift has a better linear relationship to either, which it does not.
With an R^2 value of 0.2658505 the linear regression model only covers 26,5% variance. The p-value is 2.2×10^{-16} which is proof of the high significance of the model despite the low R^2 value. The root mean square error is 1.858949 and thereby explains that our model is not fit for identifying RBPs based on its training. That is caused by the scatter of the data, which could be caused by the fact that the max-shift quantified the amount of difference of RNase and Control values on the x-axis, whereas the curve shift is based of the y-values of the curves. The scatter of data is additionally visualized in the QQ-plot in Figure 3 B showing the residuals.

To sum everything up, interpreting data always involves some assumptions, simplifications, and biases. By using different approaches, we try to cover as much uncertainty as possible, but it is impossible to consider all exceptions. No method can be labeled as the perfectly right approach. Further optimization, for example by using more statistically based and not our slightly arbitrarily thresholds for our methods or including a Gaussian fit would potentially improve the reliability and accuracy of our data. But all in all, we are confident that our shift-finders pose valuable methods for the identification of RPBs.

References

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

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.e10.

De Conti, L., Baralle, M., and Buratti, E. (2017). Neurodegeneration and RNA-binding proteins. Wiley Interdiscip Rev RNA *8*.

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

Hartigan, J. A., and Wong, M. A. (2018). A K-Means Clustering Algorithm. Journal of the Royal Statistical Society Series C: Applied Statistics *28*, 100–108.

Kelaini, S., Chan, C., Cornelius, V. A., and Margariti, A. (2021). RNA-Binding Proteins Hold Key Roles in Function, Dysfunction, and Disease. Biology (Basel) *10*.

Lunt, M. (2015). Introduction to statistical modelling: linear regression. Rheumatology *54*, 1137–1140.

Qin, H., Ni, H., Liu, Y., Yuan, Y., Xi, T., Li, X., and Zheng, L. (2020). RNA-binding proteins in tumor progression. Journal of Hematology Oncology *13*, 90.

Salapa, H. E., Hutchinson, C., Popescu, B. F., and Levin, M. C. (2020). Neuronal RNA-binding protein dysfunction in multiple sclerosis cortex. Ann Clin Transl Neurol *7*, 1214–1224.

Schultz, C. W., Preet, R., Dhir, T., Dixon, D. A., and Brody, J. R. (2020). Understanding and targeting the disease-related RNA binding protein human antigen R (HuR). WIREs RNA *11*, e1581.

Weisse, J., Rosemann, J., Krauspe, V., Kappler, M., Eckert, A. W., Haemmerle, M., and Gutschner, T. (2020). RNA-Binding Proteins as Regulators of Migration, Invasion and Metastasis in Oral Squamous Cell Carcinoma. International Journal of Molecular Sciences *21*, 6835.

Wu, X., Gardashova, G., Lan, L., Han, S., Zhong, C., Marquez, R. T., Wei, L., Wood, S., Roy, S., Gowthaman, R., Karanicolas, J., Gao, F. P., Dixon, D. A., Welch, D. R., Li, L., Ji, M., Aubé, J., and Xu, L. (2020). Targeting the interaction between RNA-binding protein HuR and FOXQ1 suppresses breast cancer invasion and metastasis. Communications Biology *3*, 193.

A. Appendix

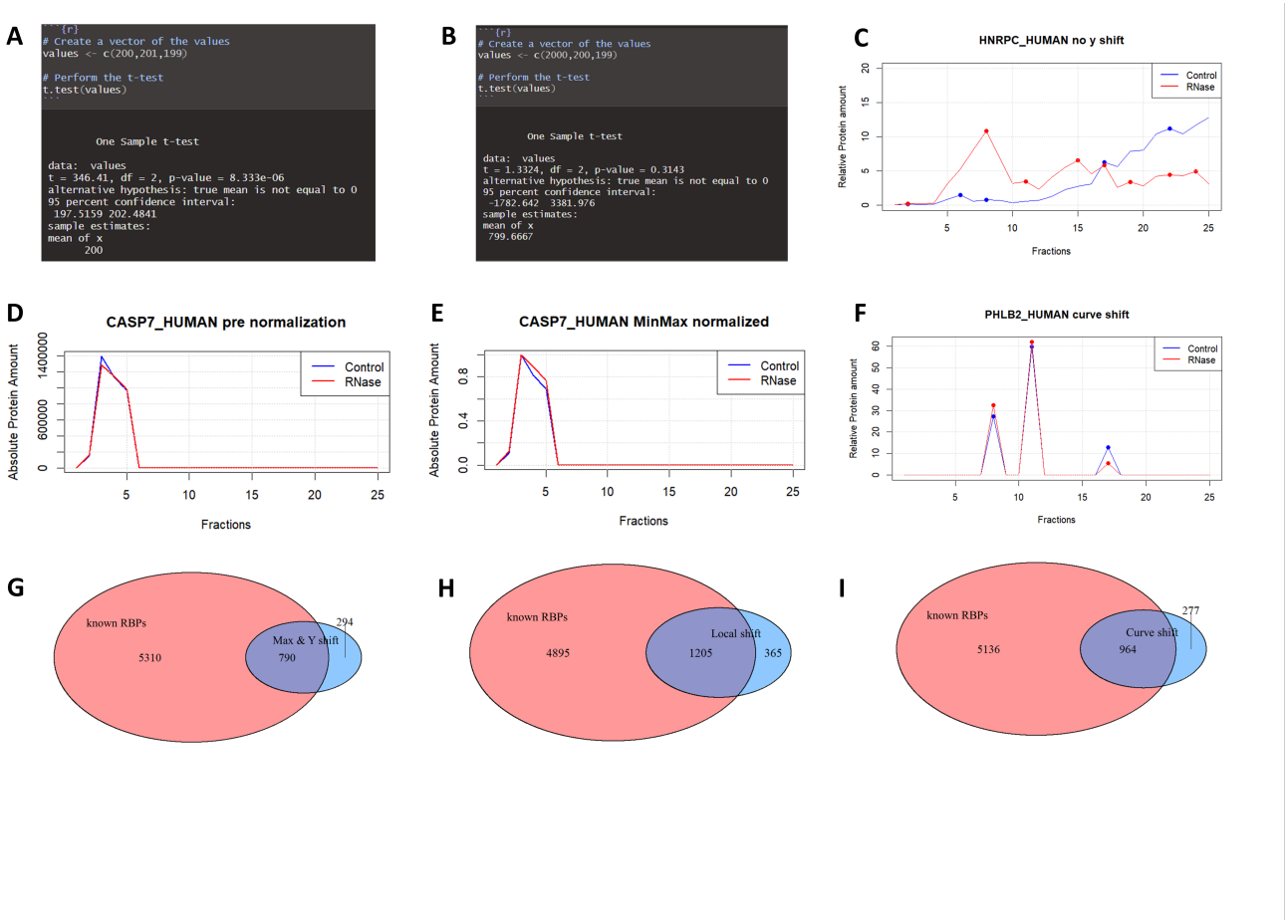

Figure 1: Principle of our t-test reproducibility check in R for small (A) and (B) large deviations. Visualization of y-shift omissions (C) . Protein pre and post MinMax normalization (D,E). Visualization of additional curve shift (F). Overlaps of each method with the RBP2GO database using a Venn Diagram, for max- y-shift (G), local shift (H) and curve shift (I).

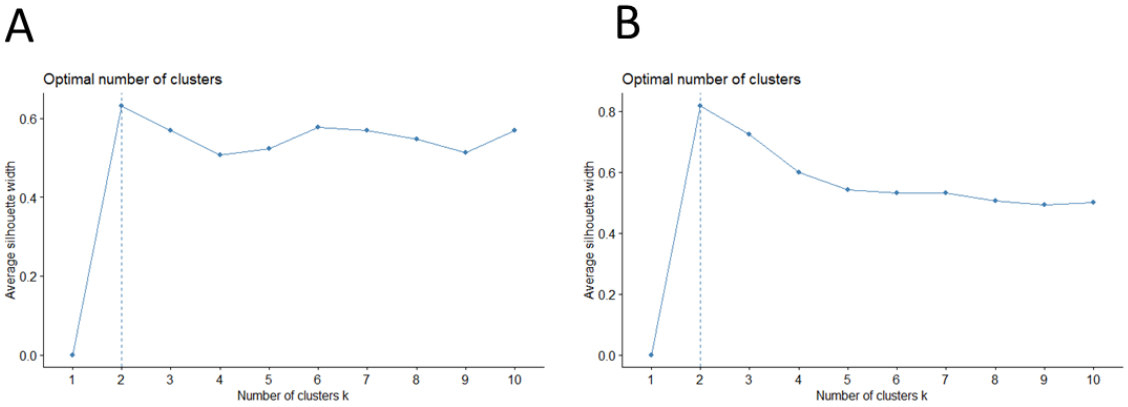

Figure 2: Results of a silhouette method for maxima (A) and curve (B) dataframe. In both cases the optimal number of clusters is two cluster.