

# Proteome wide screen for RNA-dependent Proteins in interphasic HeLa cells

Data Analysis project Summer Term 2022 supervised by Dr. Caudron-Herger

Bolz, C., Bonsen, M., Pott, M., Simon, M.

18.07.2022

## Introduction

RNA and proteins represent a symbiotic system. Proteins need RNA as a template for their biosynthesis as stated by Francis Crick in his “Central Dogma of Molecular Biology” (Crick, 1970). The opposite is also true, studies have shown that proteins need RNA for their catalytic activity, for instance in the RNA-induced silencing complex (Pratt and Macrae, 2009; Wilson and Doudna, 2013). Additionally, some RNA sequences depend on proteins for their synthesis and stability (Kishore *et al.*, 2010). A family of proteins which illustrates this symbiotic relationship are the RNA-binding proteins (RBPs). RBPs present a class of proteins whose interactome depends on RNA. They were shown to play a crucial role in RNA metabolism (Kishore *et al.*, 2010), cancer development (Wei *et al.*, 2022), and genetic disease (Gebauer *et al.*, 2021). Therefore, a deeper understanding of RNA binding and RNA-protein interaction strengthens our ability to adjust and manipulate the cellular mechanisms affected.

RBPs can be categorized into “true” RBPs (e.g. DICER, NPM3), those RNA-binding proteins which directly bind to RNA, and “RBP interacting proteins” which merely interact with “true” RBPs (eg. RBBP7). Furthermore, “true” RBPs can be subcategorized into “RNA-dependent”, meaning relying on RNA for their whole and correct function (eg. DICER) and “partially RNA dependent”, those that only require RNA for certain functions or transport (eg. NPM3) (see Fig. 1) (Caudron-Herger *et al.*, 2019; Corley *et al.*, 2020).

Several approaches to study RNA-protein interaction and to identify new RBPs were established such as RaPID and CLIP-Seq. These methods either analyze the interaction between the RNA of interest and additional proteins (RaPID) or the interaction between a protein of interest and the different RNA with which it is interacting (CLIP-Seq) (Qin *et al.*, 2021). Therefore, approaching the global study of RNA-binding protein, interaction networks (Sternburg and Karginov, 2020) has become a matter of interest. A method published by Caudron-Herger *et al.* enables analysis and quantification of whole cell interactomes. Furthermore, this method allows for identification of new RBPs through RNase treatment and density gradient ultracentrifugation. Subsequently, the resulting fraction shifts of proteins identified via mass spectrometry were analyzed using bioinformatic techniques (Caudron-Herger *et al.*, 2019, Caudron-Herger *et al.*, 2020).

In this project, we identified RBPs and possible RBP candidates using bioinformatics in R. Beyond that, we further identified contributing variables in our data differentiating RBPs without relying on our entire analysis protocol. We cross-referenced our results with known databases such as R-DeeP (<https://r-deep.dkfz.de/>; Caudron Herger *et al.*, 2019), UniProt (<https://www.uniprot.org/>) and RBP2GO (<https://rbp2go.dkfz.de/>; Caudron-Herger *et al.*, 2021). Our dataset focusing on interphase synchronized HeLa S3 cells was obtained by the method published by Caudron-Herger *et al.* cited above.

# Methods

## Generation of the dataset

Interphasic HeLa S3 were processed as published before (see: Caudron-Herger *et al.*, 2019). The data for the amount of protein per fraction, condition and replicate was collected in arbitrary units and stored in a .csv-file.

## Clean-Up and sorting of the dataset

Using the R package tidyverse, two separate data frames were generated containing either all untreated (“\_ctrl”) or RNase treated (“\_RNA”) replicates. Both data frames, as well as the raw data set, were screened for rows containing zeros only. Those zero rows were removed from the data frame and stored in a new data frame.

## Normalization

To rule out batch to batch effects and technical error, the protein amount per replicate (“Rep”) and fraction (“Frac”) were additionally normalized with respect to each sample. The normalized results were then visualized by plotting the protein distribution in both samples.

### Fraction-wise normalization

Normalization for each fraction was performed applying the following equation:

$$Protein(norm) = \frac{maxcolsums}{ColSumme} * Protein(before)$$

*Protein(norm)* describes the normalized amount of a single protein per fraction and replicate. *ColSumme* represents the total amount of protein per fraction and replicate. For each fraction, one maximum was chosen. *maxcolsums* is the selected maximum total protein amount per fraction and replicate. The parameter *Protein(before)* describes the arbitrary amount of a single protein per fraction and replicate prior to normalization.

### Scaling the protein distribution per replicate and condition

For better comparison, the protein amount per fraction was converted to a relative percentage scale applying the following equation:

$$Protein(relative) = \frac{Protein(norm)}{RowSum} * 100$$

*Protein(relative)* describes the relative protein amount per fraction in relation to the total amount of the protein per replicate. *Protein(norm)* represents the normalized amount of a single protein per fraction and replicate. The total amount of a single protein per replicate is given by *RowSum*.

Next, data frames for each individual fraction and condition were created, containing the triplicate values for each protein.

## Calculation of means with standard deviation

The mean and standard deviation of the triplicates for each protein in each fraction were calculated using the built-in functions *mean()*-( $\bar{x}$ ) and *sd()*-function( $\sigma$ ) of R. Outliers were detected using  $\bar{x} \pm 1\sigma$  as a cut-off. All values below and above the three sigma cut-off were replaced with NA and not considered in the following calculation of the mean.

## Shapiro-Wilk Test

To test the normal distribution of the triplicate values, the Shapiro-Wilk test was chosen due to its high power in small populations compared to other tests. The test was performed using the built-in R function *shapiro\_test()*.

## Determination of Maxima

We determined the global maximum for the RNase-treated and untreated sample for all proteins. The protein content (y-value) was compared to the two neighbors right and left of the analyzed fraction (x-position). For fractions 1 and 25 only the neighbors right or left of the fraction could be compared due to border limitations. For fractions 2 and 24 only one neighbor could be compared left or right. Obtained values were stored in a separate data frame.

## Detection of Protein Shifts

In our analysis, we considered shifts in the global maximum comparing RNA-treated and untreated samples as a proxy for the presence of RNA. We required “shifting proteins” show both a significant x-shift and y-shift to improve precision.

### X-shifts

First, we compared the x-position of the global maximum in both samples. To quantify the shift, we determined the difference in the fraction number of the control group and the RNase treated sample.

$$x - shift = |fracs\_max\_ctrl| - |fracs\_max\_RNase|$$

The variables *fracs\_max\_ctrl* and *fracs\_max\_RNase* describe the x-position of the global maxima for control and RNase treatment respectively. *x-shift* is the resulting value and used for Shift-direction determination.

The following convention regarding x-shifts was used:

1. Left-shift: Value < 1
2. Right-shift: Value > 1
3. No-Shift: Value = 0

### Y-shifts

The total value of the y-shift was calculated as the difference between the y-values of the global maxima for both conditions.

$$y - shift\_total = |absolute\_max\_ctrl| - |absolute\_max\_RNase|$$

The variables *absolute\_max\_ctrl* and *absolute\_max\_RNase* describe the y-value of the global maxima for control and RNase treatment respectively. *y-shift\_total* is the resulting value.

## Statistical analysis

To identify significant y-shifts, we used statistical analysis to determine the difference in the relative protein amount (y-value) for each protein in the global maximum fraction of RNase-treated and untreated samples. All proteins with x-shift values  $\geq |1|$  were considered.

## F-Test

To determine whether our sample variances were comparable and therefore suitable for the two-tailed, unpaired t-test, a two-sided F-Test was performed on the triplicates of each protein and fraction using the built-in R function *var.test()*. The significance level (p-value) was set to  $\alpha = 0.01$ . The test was deemed positive if  $p - value > 0.01$ .

If *var.test()* was performed on all zero samples, NA/NaN was returned. Therefore, the F-test filtered out those samples that were not relevant for maxima analysis anyways. Since only the comparison of y-values at the x-positions of the global maxima mattered for our further analysis, proteins that failed the F-test (p-value < 0.01) at those x-positions were excluded. To quantify how many proteins failed F-test RNase- and ctrl-maximum spots were analyzed. To refer the F-test results back to the actual proteins, a more visual matrix was created labeled *p\_value\_matrix*.

## Students T-Test

The two-tailed, unpaired t-test was used to identify proteins with significant y-value changes in global maxima x-positions after treatment. P-Values were calculated for each triplicate per protein per fraction using the built-in R function *t.test()*. Fractions with a global maximum were compared to the significance level  $\alpha = 0.05$ . The test was deemed positive if  $p - value < 0.05$ . Therefore, the change was considered significant.

Since the t-test was performed on y-values of x-positions of global maxima for both conditions separately, the returned results were either TRUE/TRUE, TRUE/FALSE, FALSE/TRUE or FALSE/FALSE for the ctrl and RNase maxima. Only proteins with positive t-tests (p-value < 0.01) for the global maximum in both conditions (TRUE/TRUE) were considered for further analysis.

## Identification of potential RBPs by analysis of x- and y-shift

Potential RBPs were selected by filtering out proteins with significant y-shift following the t-test but no significant x-shift in the global maxima. Cut-Off conditions was defined as: significant y-shift and x-shift < 1.

Proteins filtered out by this method were removed from the data frame and stored separately for further analysis.

## Checking for false positive and false negative results

False-positive and false-negative proteins were determined by cross-referencing with two data sets which were provided by Maiwen Caudron-Herger at Prof. Dr. Sven Diederichs Lab (DFKZ) (unpublished Data). The data contains information on RBPs and non-RBPs previously identified by different researchers.

Precision is defined as:

$$precision = \frac{TP}{TP+FP}$$

Accuracy is defined as:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

TP = True positives, FP = False positives, TN = True negatives, FN = False negatives

## Identification of potential RBD Proteins

Potential RNA-binding dependent (RBD) protein candidates were identified by cross-referencing false-positive proteins with the data set provided by Maiwen Caudron-Herger at Prof. Dr. Sven Diederichs Lab (DFKZ) (unpublished Data).

## k-means clustering

Using the libraries *corrplot*, *cluster* and *factoextra* k-means clustering was conducted using k-means algorithm. First we defined a function that applies the following equation to all proteins

$$Quotient = \frac{mean\_RNase(Protein)}{mean\_ctrl(Protein)}$$

The resulting values were stored in a matrix labeled *q\_Hela\_mat*. *mean\_RNase(Protein)* describes the mean value for each protein and fraction in RNase condition. *mena\_ctrl(Protein)* describes the mean value for each protein and fraction in ctrl condition. Then we selected the maximum value (*maxi\_RNase*) for each protein and also calculated the overall quotient sum (*q\_mat\_sums*) for each protein and condition. These values were stored in the data frame *abs\_mat\_both\_q*.

Using the *factoextra* function *fviz\_nbclust()* we created a silhouette plot to determine the optimal cluster number. With the built-in R function *k-means()* we clustered our proteins using the previous determined variables *q\_mat\_sums* and *maxi\_RNase* into three clusters.

**Cluster A:** Non-shifter

**Cluster B:** Potential shifter

**Cluster C** Sure-shifter

To check the quality of our clustering we created different silhouette plot using *factoextra*. To further analyse the results of our clustering we applied the functions created for cross-referencing our x-shift results. ##  
Linear regression analysis

To train our linear regression model we selected 5000 random proteins from our raw data set. For these proteins we selected the values *maxi\_RNase* and *q\_mat\_sums* from the data frame *abs\_mat\_both\_q*. We created a model to predict the x-shift based on those values. To test our regression model the remaining 2081 proteins were applied to the model. We crossreferenced the results with the data provided by Maiwen Caudron-Herger.

## Results

### Clean-Up and normalization

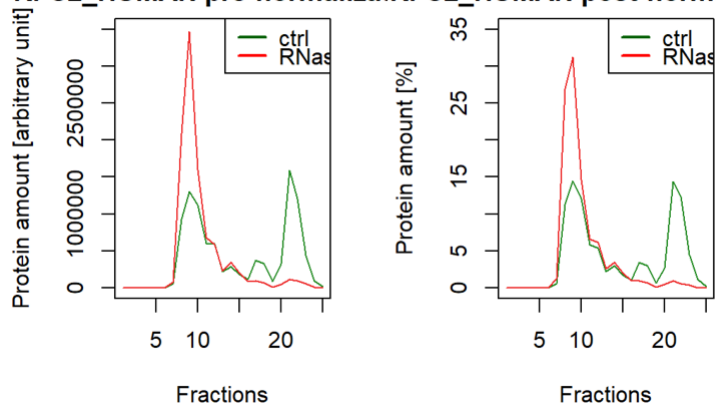
The initial data set consisted of 7086 proteins. After clean-up 7081 proteins remained for further analysis. Normalization was visualized for several reported sure-shifters and non-shifters (see Caudron-Herger *et al.*, 2019). The selected proteins are:

Sure-shifters: Sin3A\_HUMAN, HDAC1\_HUMAN, HNRPU\_HUMAN, RFC2\_HUMAN

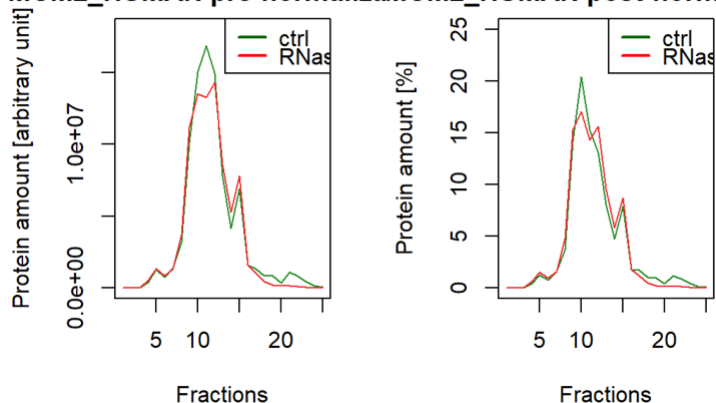
Non-shifters: ASNS\_HUMAN, MCM2\_HUMAN, MCM3\_HUMAN

Through normalization the arbitrary values were converted into relative percentage scale. Yet, the overall distribution of the protein amount per fraction remained comparable to the distribution prior to normalization. Therefore, the normalization worked for sure-shifters as well as non-shifters. For instance, figure 2 displays the normalization of a sure-shifter (RFC2\_HUMAN) and a non-shifter (MCM2\_HUMAN).

### A RFC2\_HUMAN pre-normalizatRFC2\_HUMAN post-normaliza



### B MCM2\_HUMAN pre-normalizaMCM2\_HUMAN post-normaliza



**Figure 1: Comparing pre and post normalization** (A): Comparison of the sure-shifter RFC2\_HUMAN prior to and post normalization. (B): Comparison of the non-shifter MCM3\_HUMAN prior to and post normalization. For both selected proteins, the overall trend of distribution remains comparable.

## Calculation of mean with sd and Shapiro-Wilk Test

With the proposed  $1\sigma$  rule we were able to exclude one value for 122175 triplicates. Shapiro-Wilk Test concluded that all values obtained for each protein were normally distributed for both conditions. Therefore they were suitable for statistical analysis with F-Test and two-tailed unpaired T-test.

## Statistical analysis and Identification of potential RBPs by analysis of x- and y-shift

The obtained x- and y-shift as well as the fraction of the global maxima, shift direction and y-shift values were stored in a data frame labeled *abs\_max\_both*.

7081 proteins were tested with F-Test. Only the obtained p-values for fractions where a global maximum was detected were deemed relevant for identification of potential RBP candidates. Of 7081 proteins 1196 failed the test and were excluded from further analysis. 5885 remained. Quantification of FALSE Results in F-Test resulted that 10 % of all analyzed proteins fail F-Test. This means that either one or both obtained maxima for triplicate measurements vary in their variance in one treatment compared to the other. Because it seems like some FALSE F-test results originate from outlier values (see Discussion) and otherwise we would experience a Precisiongreat exclusion of data, significance level was set to  $\alpha = 0.01$  instead of the previously usual  $\alpha = 0.05$ . With  $\alpha = 0.01$  closer to 5 % of data will be excluded based on a FALSE F-test result. In this compromise some more likely faulty data will still be excluded while retaining as much data for further

analysis as logically arguable.

Two-tailed unpaired t-test was applied to the remaining 5885 proteins per fraction ctrl vs RNase. P-Values were checked for significance at previous determined maxima positions. We identified 2929 potential shifters. 5838 proteins were annotated in the cross-referencing data set. These 5838 proteins were analysed for correct prediction. The cut-off  $>1$  for the x-shift allowed us to identify 513 true RBP (TP) with 121 false-positives (FP) out of 634 total. Of the 121 false-positives six were identified with cross-referencing as RBD proteins. 5214 proteins were predicted as non-RBP. Cross-referencing revealed that 3600 proteins were false-negative (FN) and 1614 were true-negative (TN). This results in a precision of 80.914 % and an accuracy of 36.43 % (see **Fig.2C**)

With T-Test based analysis, we detected one out of four previously selected sure-shifters and all non-shifters.

## K-Means Clustering

The clustering analysis concluded that 73.88 % of our data was represented. Values were assigned with 81 % accuracy to our cluster.

The clustering analysis was conducted with three clusters (see **Fig. 2B**). For Cluster 1 (C) 53 true RBP (96.4%) and 2 false RBP (3.6%) were detected. Cluster 2 (B) presented a similar accuracy of 318 true RBP (96.7%) and 10 false RBP (3%). Therefore, we decided to summarize Cluster B and C, as both accurately predicted true RBPs. For Cluster 3 (A) 2557 true non-RBP (38.2%) and 4089 false RBP (61.1%) were detected. 1 RBD was detected out of the false-RBP proteins. The results are displayed in **Fig. 2C - Clustering**.

With the clustering, we detected none of our previously selected sure-shifters and all non-shifters.

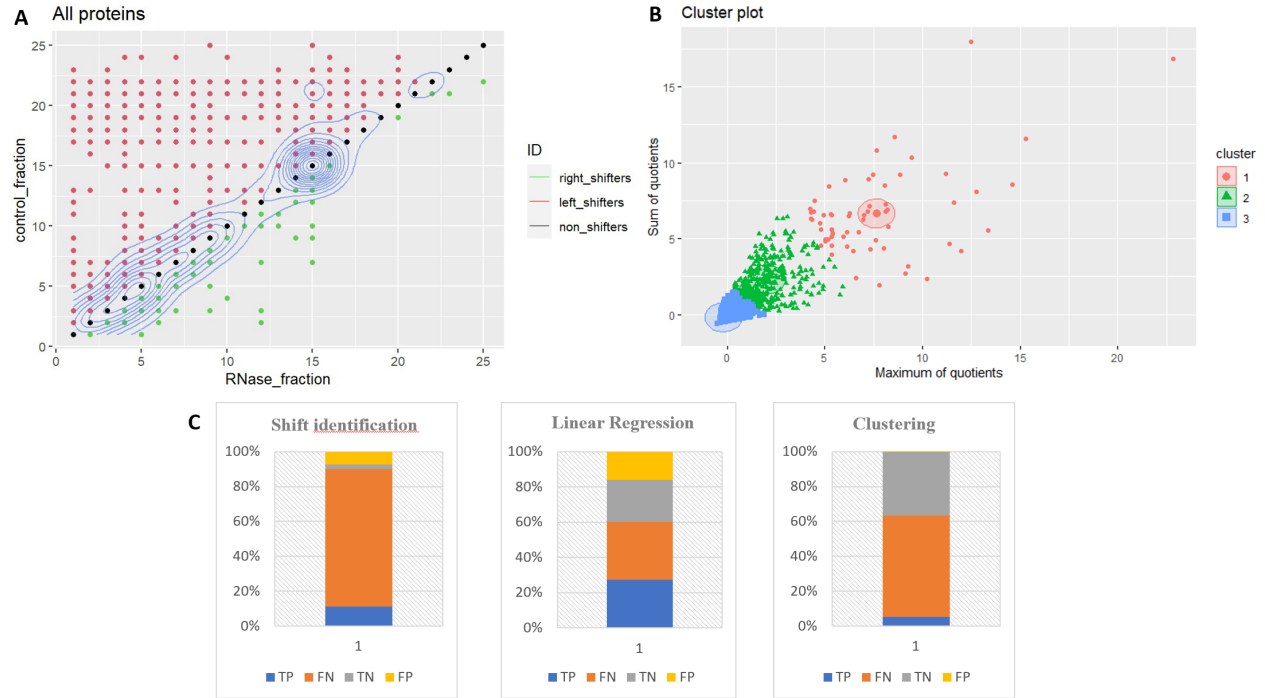
## K-means Clustering vs. Statistical Analysis

Comparison of the identified true RBP via clustering and statistical analysis shows, that out of 371 RBP from the clustering analysis, 189 (50,9%) were not identified by the statistical tests with more identified RBP overall. 182 (49,1%) true RBP identified via clustering were also identified by the statistical tests.

## Linear Regression model

Since the regression takes different random proteins for the training every time, the shown results were generated once. Out of 2081 tested proteins, 883 RBP were predicted of which 525 (59,4%) turned out as true RBPs and 347 (39,3%) were false RBPs. 1198 non-RBPs were predicted. 470 (39,2%) were true non-RBPs while 709 (59,2%) were false non-RBPs. 8 RBDs were detected out of the false-RBP proteins. The results are displayed in **Fig. 2C - Linear Regression**.

With the regression, we detected none of our previously selected sure-shifters and one out of three non-shifters.



**Figure 3: Overview of obtained results** (A) Plot of all 7081 proteins, calculated global maxima were plotted against each other, proteins can be classified as either right-shifter (green), left-shifter (red) or non-shifter (black). Density (blue) indicates that most proteins are non-shifters. (B) k-means clustering of our proteins as described in methods, three distinct clusters can be visualized, Cluster 1 = sure-shifters (red), Cluster 2 = possible shifters (green), Cluster 3 = non-shifters (blue). (C) Overview of identification results for Shift based identification = Shift identification, Linear Regression model and k-means clustering.

## Discussion



## References

- Caudron-Herger, M., Jansen, R.E., Wassmer, E., and Diederichs, S. (2021). RBP2GO: a comprehensive pan-species database on RNA-binding proteins, their interactions and functions. *Nucleic Acids Research* 49.
- 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.
- Caudron-Herger, M., Wassmer, E., Nasa, I., Schultz, A.-S., Seiler, J., Kettenbach, A.N., and Diederichs, S. (2020). Identification, quantification and bioinformatic analysis of RNA-dependent proteins by RNase treatment and density gradient ultracentrifugation using R-DeeP. *Nature Protocols* 15, 1338-1370. 10.1038/s41596-019-0261-4.
- Corley, M., Burns, M.C., and Yeo, G.W. (2020). How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. *Molecular Cell* 78, 9-29.
- Crick, F. (1970). Central Dogma of Molecular Biology. *Nature* 227, 561-563.
- Gebauer, F., Schwarzl, T., Valcárcel, J., and Hentze, M.W. (2021). RNA-binding proteins in human genetic disease. *Nature Reviews Genetics* 22, 185-198.
- Kishore, S., Lubner, S., and Zavolan, M. (2010). Deciphering the role of RNA-binding proteins in the post-transcriptional control of gene expression. *Briefings in Functional Genomics* 9, 391-404.
- Li, W., Deng, X., and Chen, J. (2022). RNA-binding proteins in regulating mRNA stability and translation: roles and mechanisms in cancer. *Seminars in Cancer Biology*.
- Pratt, A.J., and Macrae, I.J. (2009). The RNA-induced Silencing Complex: A Versatile Gene-silencing Machine. *Journal of Biological Chemistry* 284.
- Qin, W., Cho, K.F., Cavanagh, P.E., and Ting, A.Y. (2021). Deciphering molecular interactions by proximity labeling. *Nature Methods* 18, 133-143.
- Sternburg, E.L., and Karginov, F.V. (2020). Global Approaches in Studying RNA-Binding Protein Interaction Networks. *Trends in Biochemical Sciences* 45, 593-603.
- Wilson, R.C., and Doudna, J.A. (2013). Molecular Mechanisms of RNA Interference. *Annual Review of Biophysics* 42, 217-239.