Final Report

Proteome-wide screen for RNA-dependent Proteins

Supervisor Dr Maiwen Caudron Herger

Team 05 – HeLa Mitosis

University of Heidelberg - IPMB

Molecular Biotechnology

Summer Term 2022

19.07.2022

Laura Herrfurth

Katharina Lotter

Kiren Nadeem

Marie Lulu Salein

Table of Contents

1. Introduction
   1. RNA binding proteins
   2. Experimental setup
2. Methods
   1. Data Cleanup and Data Reduction
      1. Data Reduction
      2. Removal of batch effects
      3. Reproducibility of the data
   2. Determining Maxima and Shifts
   3. T-test for validation of shifts
   4. kmeans clustering
   5. regression analysis
3. Results
   1. Comparison with the R-DeeP and CORUM database
4. Discussion
5. Literature
6. Introduction
   1. RNA binding proteins

RNA plays a fundamental role in all cellular processes, ranging from the cellular localization of proteins, to translating mRNA and post transcriptional alterations. The interaction of RNA and proteins enables a fine tuned regulation of RNAs via degradation and stabilization (Alberts 2015), (Nelson 2021). In recent years the field of proteomics, which describes the analysis of the proteins in a cell or tissue at a given time under specific conditions(Ginsburg and Willard 2009), has begun to focus on the discovery and understanding of molecular mechanisms having to do with RNA dependent proteins. However, there are still many processes which have yet to be fully understood.

Instead of just focusing on proteins which bind RNA directly, the analysis also includes RNA dependent proteins. A protein is considered RNA dependent, when its interactome depends on RNA. This does not necessarily mean the protein binds directly to RNA, instead of having classical RNA binding domains (Lunde, Moore et al. 2007), these proteins can bind RNA via intrinsically disordered domains (Caudron-Herger, Rusin et al. 2019)

For simplicity, if not stated otherwise the term RBP will be used for both RNA binding and RNA dependent proteins in this report.

The interaction between RNA and proteins is highly diverse and dynamic, the binding can either be cooperative or antagonistic and responds to cellular as well as environmental stimuli, which enables a fine tuned regulation of RNA. Another aspect important for this interaction is the flexibility of the RNAs tertiary structure. There are protein binding domains in RNA which are found repeatedly, such as the RNA recognition motif or the K homology domain, but the RNA interactome is also dependent on less specific interactions such as hydrogen bonds formed with the 2’OH group or stacking interactions between pi-bonds. RBPs are enriched in intrinsically disordered regions, that loosen when the site of interaction becomes more rigid. The interaction between RNA and the protein is best described as a dynamic rearrangement process, which can contain backbone shifts or the flipping of bases. RBP interaction can be cooperative or antagonistic (Sternburg and Karginov 2020). Moreover, in some cases the combination of multiple RNA binding regions enables RBPs to exhibit a highly specific interaction pattern with certain RNA sequences, while at other times RBPs do not exhibit specific or high-affinity binding in order to enable transient interactions with RNA (Corley, Burns et al. 2020)

At the moment there is a large study gap in RBP research, the common pull of studies only contains 215 proteins, which only amounts to approximately 10% of potential RBP candidates, so there is a need for more identification methods and more studies in this field to create a better overlap. The identification still remains a major challenge in RNA biology (Caudron-Herger, Rusin et al. 2019).

Malfunctions in RBPs have severe consequences and are associated with a wide variety of diseases. The mutated splicing factor 3b subunit 1 causes cancer progression in CLL and the heterogeneous nuclear ribonucleoprotein A1 is an important protein in the progression and development of neurodegenerative disorder ALS (Kelaini, Chan et al. 2021). Therefore RBPs have present a high potential in the field of translational research, as targeting RBPs introduces new treatment approaches and perhaps the possibility of novel cancer therapeutics using siRNA, RNA knockdown or RNA aptamers (Kelaini, Chan et al. 2021). For instance, the depletion of the YTHDF2 gene, which encodes for an RBP, can induce apoptosis in human triple negative breast cancer cell lines. The RBP contributes to the epithelial mesenchymal transition. Mice with a YTDF2 knockout were found to have a decreased tumor volume, in comparison with mice that expressed YTDF2 (Einstein, Perelis et al. 2021).

* 1. Experimental Setup

The data we used was created during experiments with HeLa cells synchronized in interphase. During mitosis the chromosomes, which have been replicated previously are divided into two new nuclei while interphase is mainly characterized by cell growth, DNA replication and preparation for mitosis (Alberts).

The cells are centrifuged and lysed. One sample is treated with RNase, the other one is left untreated. The sample of interest is placed on top of the gradient. It then travels through the gradient until it reaches the point at which their density matches the one of the surrounding sucrose. The sucrose gradient is divided into 25 fractions. If the protein of interest is a RBP it migrates to a different fraction in the sucrose gradient, depending on the presence or absence of RNA. After ultracentrifugation the control and RNase samples were further processed for Coomassie staining, western blot of mass spectrometry. This method offers a quantitative approach which is able to detect both RNA binding and RNA dependent proteins. It further divides RBPs into right shift and left shift proteins (was genau das bedeutet wird in den Results oder der discussion erklärt) (Caudron-Herger, Rusin et al. 2019).

During this project our aim was firstly to find out which proteins are RBPs by analyzing the shifts between the RNase and the control samples. We then wanted to see, if we could find RBPs associated with Actin, as it is an important protein for the cytoskeleton, which has to be regulated during the cell cycle (Alberts 2015).

Lastly we compared the RBPs we found to the R-DeeP database and the UniProt database, to analyze further properties of the RBPs, such as domains which were found frequently and their function in the cell.

1. Methods
   1. Experimental Setup

The Interphase HeLa cells are centrifuged and lysed one sample is treated with RNase, the other one is left untreated. The sample of interest is placed on top of the gradient. It then travels through the gradient until it reaches the point at which their density matches the one of the surrounding sucrose. The sucrose gradient is divided into 25 fractions. If the protein of interest is an RBP it migrates to a different fraction in the sucrose gradient, depending on the presence or absence of RNA. After ultracentrifugation the control and RNase samples were further processed for Coomassie staining, western blot of mass spectrometry. This method offers a quantitative approach which is able to detect both RNA binding and RNA dependent proteins. It further divides RBPs into right shift and left shift proteins, as well as fully shifting and partially shifting, where the global maximum remains static but the smaller peaks shift. (Caudron-Herger, Rusin et al. 2019).

* 1. Description of the dataset

The HeLa Interphase dataset consists of 7086 rows that represent the proteins and 150 columns which represent the 25 fractions of both the RNase sample and the untreated control sample The measurements were made in triplicates.

The class of the values was transformed to numerics for further calculations. Doubles?

As the dataset is rather large we decided to split it into 6 subsets. For example, colRep1rnase contains all columns with measurements from the first replicate in the samples treated with RNase. In the next step the 3 subsets, describing either the control or the RNase samples were joined per column, in order to get one dataset or each condition, that contained all three replicates measurements.

* 1. Data Cleanup and Data Reduction

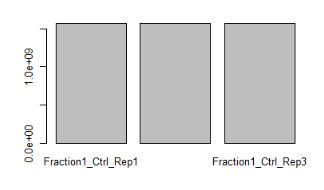
Our first step was to remove rows (proteins) which only contained zeroes, as well as proteins where either values of the RNase and the control sample were comprised of only zeroes, since these proteins cannot be compared. This led to a reduction to 7081 proteins. All remaining zeroes were transformed to be NA. (haben wir die NAs dringelassen?)

Discussion: We decided against implementing a threshold for zero values, as this might have led to the loss of proteins which show only very narrow peaks, and therefore the loss of important RBPs

* + 1. Normalization

By the nature of the experimental setup by which our data was obtained heterogeneous amounts of proteins were used between replicates, but for further steps it is crucial the replicates are compatible, which is why fraction-wise normalization was performed.

First, we adapted the amount of protein per fraction for each of the three replicates to be the mean of the two means lying closest to each other. We compared the fractionwise sum of Ctrl and RNase before normalization and visualized them with a barplot. Then the vector mean\_frac\_rep which contained the mean for each fraction in each condition and replicate was created. We used this vector to create another vector that contained the mean value of the two mean values for the protein amount of the replicates closest to each other in each fraction for RNase and control. Therefor we generated a for loop that compared the three values by using an if and an else if condition to compare two variables at a time and check which ones are closest to each other. The normalized data was stored in mean\_rep and again visualized with a box plot that now showed the same height for all replicates, indicating the same amount of protein. To increase interpretability we then normalized the protein amount of each replicate for each protein to 100 using the apply function. The amount of protein is now interpretable as a percentage, which we used when checking for local maxima and for the t-test. The values were stored in six different dataframes, one for each replicate and condition.



[Wecken Sie das Interesse Ihrer Leser mit einem passenden Zitat aus dem Dokument, oder verwenden Sie diesen Platz, um eine Kernaussage zu betonen. Um das Textfeld an einer beliebigen Stelle auf der Seite zu platzieren, ziehen Sie es einfach.]

Figure 1: amount of protein in Rep1, Rep2 and Rep 3 before (left) and after normalization (right)

Again we used the two dataframes TripCtrldat,TripRnasedat (haben die einen Namen?) that combined the three replicates for each condition. The mean for each (im Aktiv oder passiv schreiben?)

Theory:

* if the values from the replicates vary, continuing with all 3 values will lead to inaccuracies during the next analysis steps, therefore the data had to be normalized
  1. Batch effects­­

Batch effects describe variations in the data or outliers caused by technical means without biological significance and which therefore might lead to over-interpretation of certain results or false conclusions. We found such batch effects in 23 proteins. To remove them we compared the values of all three replicates for each fraction in each condition. If two values are zero, the third value (x) was set to zero as well. If two values (x,y) had numeric values and only the third measurement was 0 it was set to be the mean of x and y. To perform this, we used a for loop with 2 different if conditions.

* 1. Determining Maxima and quantitative RNA dependent shifts

Methode+ Wie viele Shifts, wie viele sind right und left shifting!

In order to compare the fraction wise change in protein amount after treating the sample with RNase, we focused on the maxima of the plots and analyzed, whether the maxima of the graphs between RNase and control sample would align of differ.

The global maxima of the plot represent the fraction of the sucrose gradient, which contains the highest amount of protein. We found this fraction searching for the maximum using the which.max() function. The fraction and the global values for each protein and replicates are stored in the dataframe fraction\_abs\_max, the first column it the first replicate of the control. ?? hab ich die richtig verstanden? (Laura: In the matrix fraction\_abs\_max is for every replicate the fraction of the main maxima saved to be able to compare them.)

We verified the global maxima, by checking whether there is a distance of more than one fraction between the replicates by using the function is.TRUE, which gave us the output TRUE if the difference is only one fraction, therefore indicating the reproducibility of this protein, proteins that were not reproducible were removed (wie viele?) 411

Keeping in mind the experimental setup, it is essential to not only consider the global, but also the local maxima. As not only fully but also partially shifting RBPs exist it is crucial to also include local maxima, since their shifts are of biological significance as well. If after RNase treatment RNA and the protein either dissociate or gain new interaction partners, the molecules will migrate to different positions in the gradient, depending on their mass. This might for instance cause the appearance of new maxima.

A characteristic global and local maxima have in common, is that the value for y (representing the amount of protein in our plots) is higher than its surroundings. This was chosen to be the first of our selection criteria. We defined a value as the (local) maximum, if both the value to the left and to the right are smaller. As this resulted in too many positions, which contained unspecific, relatively small maxima, which also could have been due to outliers we further implemented a y-threshold. The position was only considered a maxima, if the fraction contained at least 5% of the total amount of a protein. This was done to ensure the elimination of non-significant maxima. We chose to set this value to 5%, as the 2δ area approximates to 95%, therefore non-significant local maxima are eliminated in this step. Additionally we compared the number of proteins we would throw out with a threshold of 3% and 7% and agreed to the threshold in the middle of 5%.

We compared x % and z %

Shifts

For a better usability of the data, we merged the curves of the three replicates for each condition. In order to receive this average curve for every protein we generated a matrix for every protein (corresponding to one row in FullDat), whose columns are the 25 fractions, and the rows are the values of the protein amount for each replicate. ??? We generated a matrix with the average value of the protein amount of the three replicates for each protein per fraction. Hab ich die Matrix so richtig verstanden (Codezeile 840- 854).

We now have plotted two curves, one of the control and one for the protein after the RNase treatment, which are now merged into one graph for better comparison.

Bild einfügen!

As defined by R-DeeP, a protein is defined to be an RBP, if the amount of protein differs between the control sample and the RNase sample. To characterize the shifts, we considered the position of the maxima, as well as the distance and direction. ??

abs\_max\_compare???

Right shifts indicate a positive change in the density of the protein. This might be that the absence of RNA causes new binding sites in the proteins to become available and the protein gains interaction partners after the loss of RNA or could form complexes. These proteins might be RNA dependent rather than RNA binding.

Left shifts imply less density after the digestion of RNA by RNase. These proteins seem to be RNA binding proteins, that become lighter without the RNA.

Reproducibility:

We checked whether our data was reproducible with the alignment of the maxima. The global maxima were the same for each replicate, for the local maxima we implemented the condition that the difference could only be one fraction. Due to the verification of the reproducibility between replicates we assumed normality of the data which was used for statistical analysis.

* 1. Statistical testing

Statistical testing was used to provide a more objective predication whether or not the shift was significant instead of solely relying on optical perception and graphical representation. We chose to do this by using a t-test. As we assumed normality of the data due to its reproducibility the data fulfills the necessary condition for performing a t-test. Since we already implemented conditions for shifts in the x-direction we now focused on the y-shift, to tell, whether it is significant or not. While the-x shift shows the direction of the shift and can be used as an indicator whether the protein gains or loses interaction partners after RNase treatment, the y-shift between maxima represents the amount of protein, which changes the fraction after RNase treatment.

Add plots for x and y-shifts

Figure x

x-shift of a protein (left) and y-shift of a protein (right). For the x-shifts we predefined the condition of a distance of at least two fractions and to determine, whether the y-shift was significant we performed a t-test on the protein.

In order to avoid coding mistakes and to focus on the interpretation of our results we used the predefined t.test function in R. We set our significance level to be 5% and used a two-sided t-test, therefore the shift of each protein which had a p-value smaller than 0.025 was determined significant. For the t-test we looked at the global maxima and the local maxima separately. Two vectors with the maxima of the RNase and the control were created and compared. For the global maxima 869 out of the previously determined 1118 proteins showed a significant shift. We then analyzed the local maxima, using for-loops

As we performed many t-tests in a row, we wanted to avoid the cumulation of alpha-errors, which is why we adjusted the p-value using an FDR correction. However, all shifts of the global maxima were still significant after FDR correction. For the local maxima.

Anzahlen hinzufügen

* 1. Dimension reduction analysis
     1. k-means clustering

K-means clustering is an iterative, unsupervised clustering approach which is based on the Euclidean Distance metric. We used k-means to see, whether for certain properties of the proteins, such as the number of shifts, groups could be identified.

The amount k of centers is pre-defined, each point is the assigned to the point closest to the centers. For each cluster the center of gravity is determined and the points are assigned to the closest center again. This process is repeated up to when a certain number of maximum iterations is reached, or when no point should be repeated. Fluctuations indicate the lack of a clear cluster structure, however this method will identify clusters, regardless of their biological significance

Therefore, the amount of centers has a great influence on the results. The two most common methods to determine the ideal number of centers are the Ellbow and the Silhouette plot. When using the Ellbow plot all pairwise distances squared between members of the same cluster are added, which gives the within square distance value (WSS). WSS is plotted against the number of cluster k, in general it decreases with increasing k, however upon reaching the optimal number of clusters adding more clusters does not improve WSS, which can be seen as a kink in the curve. As this method will not always produce a clear kink the silhouette plot can be used as well to determine k.

Add our plots

As seen in the plots the optimal number of clusters was determined to be x.

This method calculates the mean distance to all members of its own cluster, as well as the smallest average distance bi to members of all other clusters for each element.

Add Equation

The si value ranges from -1 to +1, the higher it is, the better the object is clustered, a si value close to zero can indicate an ambiguous object, which cannot be assigned clearly to a cluster. We chose x, as this had the highest silhouette value.

For k means x clusters were chosen

We clustered

….

Again we made use of the r-function kmeans() to avoid inaccuracies due to programming errors. This function performed k-means clustering as described above.

add kmeans plots

* 1. Regression analysis
* by using complementary information gathered from the databases we wanted to see if it is possible to predict if a protein is an RBP
* Model building: predicting one variable using one or several other variables
* parameters of the regression line are estimated using the least square method

1. Results

Erklären wie viele Proteine, wir für was gefunden haben

After performing these steps, we wanted to see, whether the proteins we determined to be RBPs are accurate in comparison with other databases. For reference we used R-DeeP

The results of this comparison are as following

Prozentzahl einfügen, wie gut unsere Trefferquote ist.

In conclusion our analysis was successful, but has room for improvement.

* 1. Comparison with the R-DeeP and UniProt database

The aim was to see if we could find properties, such as common domains or a similar isoelectric points, which were shared by the majority of RBPs. For this we first compared the RBPs we found with the R-DeeP database, which provided a visualization of protein shifts. Moreover, the data was obtained by the same experimental method and we were able to compare the existence and direction of the shift. When comparing our RBPs with R-Deep we found matches for several of them, meanwhile for some proteins there was no data yet, or the protein was not identified as an RBP by R-deep, the latte probably being due to differences in the statistical testing and predefined conditions for an RBP. For further analysis, we focused on the proteins which had also been identified as an RBP by R-DeeP. We found that right-shifts were rare events and were often also associated with precipitated proteins, the majority of our proteins had a left-shift, indicating a loss of interaction partners after RNase treatment.

Using UniProt we were unable to see many proteins, that were associated in any way with actin, what we found out was that many proteins had important functions in chromatin reorganization, transcription and translation regulation or the transition of phases in the cell cycle. We found several RBPs that were important for the function and the organization of the nucleosome but also for the organization of cytoskeletal keratin, brain and neural development, as well as the function of the ribosome. Looking at the pathology of these proteins we saw, that a dysregulation often leads to cancer.

* Zinc-finger (these proteins are also DNA binding)
* RRM Definition? which we expected was found in several cases
* amino acid bias: containing basic and acidic amino acid residues, less frequently polar residues
* almost every protein contains disordered regions
* Q-motif, dead box (these proteins all belonged to the same family)
* often coiled coil

Diseases

H2A1A\_HUMAN: monoubiquitinylation may be involved in breast cancer

other proteins: Ehlers-Danlos Syndrome, neurodevelopmental disorders and developmental delay, corneal dystrophy

To-do: für die Interpretation dieses Teils die anderen beiden vorgeschlagenen Paper nochmal durchgehen

1. Discussion

During this project by using R-Studio we were able to determine which proteins in a dataset were RNA dependent proteins, by analyzing and comparing the shift in the amount of protein before and after the treatment with RNase.

Due to time constraints we were unable to work with the second dataset as well, comparing the two datasets would have given insight on more transient RBP interactions and how they can change during the cell cycle. This might have resulted in finding more proteins, which were associated with actin.

RNA-dependent proteins were shown to be involved in almost all cellular processes, mainly in functions performed by RNA, such as transcription, translation and the function of the ribosome, but also in brain development, the regulation of the cell cycle and many others, showing their diverse range and therefore the importance of research on RNA dependent and RNA associated proteins, which continues to be a promising field of research.

1. Literature

Alberts, B. (2015). Molecular Biology of the cell. New York, NY, Garland Science.

Caudron-Herger, M., S. F. Rusin, M. E. Adamo, J. Seiler, V. K. Schmid, E. Barreau, A. N. Kettenbach and S. Diederichs (2019). "R-DeeP: proteome-wide and quantitative identification of RNA-dependent proteins by density gradient ultracentrifugation." Molecular cell **75**(1): 184-199. e110.

Corley, M., M. C. Burns and G. W. Yeo (2020). "How RNA-binding proteins interact with RNA: molecules and mechanisms." Molecular cell **78**(1): 9-29.

Einstein, J. M., M. Perelis, I. A. Chaim, J. K. Meena, J. K. Nussbacher, A. T. Tankka, B. A. Yee, H. Li, A. A. Madrigal and N. J. Neill (2021). "Inhibition of YTHDF2 triggers proteotoxic cell death in MYC-driven breast cancer." Molecular Cell **81**(15): 3048-3064. e3049.

Ginsburg, G. S. and H. F. Willard (2009). Essentials of genomic and personalized medicine, Academic Press.

Kelaini, S., C. Chan, V. Cornelius and A. Margariti (2021). RNA-Binding Proteins Hold Key Roles in Function, Dysfunction, and Disease. Biology 2021, 10, 366, s Note: MDPI stays neutral with regard to jurisdictional claims in published ….

Kelaini, S., C. Chan, V. A. Cornelius and A. Margariti (2021). "RNA-binding proteins hold key roles in function, dysfunction, and disease." Biology **10**(5): 366.

Lunde, B. M., C. Moore and G. Varani (2007). "RNA-binding proteins: modular design for efficient function." Nature reviews Molecular cell biology **8**(6): 479-490.

Nelson, D. L., Cox, Michael M. , Hoskins Aron A. (2021). Lehninger principles of biochemistry. Macmillan International, Higher Education.

Sternburg, E. L. and F. V. Karginov (2020). "Global approaches in studying RNA-binding protein interaction networks." Trends in Biochemical Sciences **45**(7): 593-603.