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
   2. Determining Maxima and Shifts
   3. Statistical testing
   4. Comparison with the R-DeeP and CORUM database
   5. kmeans clustering
   6. regression analysis
3. Results
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 heterogenous 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 synchonized in mitosis and 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 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 (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, starting with the data from Mitosis.

If time for second dataset: After finishing the analysis of RPBs during Mitosis we applied the steps to the Interphase datatset, as our goal was to see how RBPs change according to the cellular context, which interactions are stable and which interactions occur only transiently. 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 CORUM database, to analyze further properties of the RBPs.

1. Methods
   1. Experimental Setup

The data we used was created during experiments with HeLa cells synchonized in mitosis and 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 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 (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, starting with the data from Mitosis.

If time for second dataset: After finishing the analysis of RPBs during Mitosis we applied the steps to the Interphase datatset, as our goal was to see how RBPs change according to the cellular context, which interactions are stable and which interactions occur only transiently. 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 CORUM database, to analyze further properties of the RBPs.

* 1. Description of the dataset

The 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.

As the dataset is rather large we decided to split it into 6 subsets. The subsets are divided by the sample preparation (RNAse or untreated control) as well as the replicate. 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 3 triplicate measurements.

* 1. Data Cleanup and Data Reduction

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

The 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

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

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

batch effects?

* 1. Determining Maxima and quantitative RNA dependent shifts

In order to compare the fractionwise 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. Keeping in mind the experimental setup it is essential to not only consider the global, but also the local maxima. 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 to 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 x % of the total amount of protein, this was done to ensure the elimination of non significant maxima.

We compared x % and z %

Anmerkung: im Report von Maike steht, dass sie 5% ausgewählt haben, das dies den 1 sigma Bereich repräsentieren würde, vielleicht können wir den 1 sigma und den 2 sigma bereich vergleichen?

Shifts

We plotted the amount of protein for the RNase sample, as well as the control sample and compared the maxima of the plot.

As defined by R-DeeP, a protein is defined to be an RBP, if the amount of protein will differ 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.

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

* 1. Statistical testing

Statistical testing was used to provide a more objective predication whether or not the shift in the x-direction was significant, instead of just looking at the graphical representation. For choosing a statististical test, the normality of the data had to be checked first, which was done using the Shapiro Test.

T-Test (wenn wir ihn Machen können, ansonsten non parametric tests)

Theory

* 1. Comparison with the R-DeeP and CORUM 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 were obtained by the same experimental method and we were able to compare the position and the direction of the shift.

Focusing on protein domains we then used the CORUM (comprehensive resource of mammalian protein complexes) to find common structural motifs of our RBPs.

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

Dimension reduction was used to identify similar and dissimilar groups in the dataset, for this we used k-means clustering.

K-means clustering is an iterative, unsupervised clustering approach which is based on the Euclidean Distance metric.

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

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

add kmeans plots

* 1. regression analysis
* by using complementary information gathered from the databases we wanted to see if it is possible to predict is 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
2. Discussion
3. 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.