Data Analysis Project 2022

Report

Institute of Pharmacy and Molecular Biotechnology

Topic 03: Proteome-wide Screen for RNA-dependent Proteins

Sub-project 2: HeLa Cells Synchronized in Mitosis

Supervisor: Dr. Maïwen Caudron-Herger

Tutor: Niklas Engel

Students: Michel Tarnow, Michèle Bennek, Lennart Müller, Sebastian Rickert

Date: 20.07.2022

Table Of Contents

Introduction

RNA-binding proteins (RBPs) are a group of proteins who share the characteristic of directly binding to RNA molecules (Gebauer et al., 2021). Thus formed complexes are termed ribonucleoprotein particles (RNPs), prominent examples of which include the spliceosom (Corley et al. 2020), the signal recognition particle (Faoro and Ataide, 2021) and the RNA-induced silencing complex (Pratt and MacRae, 2009). RBPs play key regulatory functions in the life cycle of RNA molecules, including transcription, splicing, modification, intracellular trafficking, translation and degradation (Gebauer et al., 2021).

RBPs are known to bear discrete domains that are responsible for the molecular interactions between amino acid residues and RNA nucleotides that lead to RNA binding, and which for this reason are defined as RNA-binding domains (RBDs), such as the RNA recognition motif, the K-homology domain or DEAD/DEAH helicase and zinc-finger domains (Corley et al., 2020). Although the occurrence of one or more of these domains is characteristic of RBPs, more recent research has identified RBPs that lack known RBDs but instead bind RNA through intrinsically disordered regions (Gebauer et al., 2021), indicating the heterogeneity of this group of proteins.

Since RBPs are key players in RNA metabolism, their malfunction has been associated with diverse pathological phenotypes, including cancer (Zhang et al., 2020) and neurodegenerative disorders (Maziuk et al., 2017). Therefore, identifying new RBPs may contribute to point out novel drug targets for already known conditions on the one hand and to shed new light on the diseased mechanisms of yet unresolved conditions on the other.

While the study and quantification of RBPs has received much attention in previous research, the exact number of mammalian RBPs remains a matter of debate. Techniques for proteome-wide screening of RBPs in the past heavily relied on pull-down assays of polyadenylated RBPs (Beckmann et al., 2015; Castello et al., 2012), protease digestion (Mullari et al., 2017) or UV cross-linking (Urdaneta et al., 2019), each method having its advantages and shortcomings. Thus, current estimations about the number of RBPs remain largely inconsistent across studies, leading only to a small consensus set of around 200 proteins (Caudron-Herger et al., 2019).

While not devoid of its own limitations, the here presented method provides a novel approach to the proteome-wide screening of RBPs with the goal of validating the core set of RBPs previously defined while also contributing to adding new proteins to the number of possible RBP candidates.
For this purpose, the notion of RNA dependence was introduced, defining a protein as RNA dependent when its interactome depends on RNA, while binding to RNA is sufficient but not necessary to qualify as RNA dependent. Thus, every RNA binding protein is RNA dependent, but not every RNA dependent

protein is RNA binding (Caudron-Herger et al., 2019).

The method here employed relies on the differential migration pattern of proteins in a sucrose density gradient ultracentrifugation in the presence or absence of RNA.
In density gradient centrifugation, macromolecules such as proteins are forced through a density gradient until they find a density equal to their own (Farrell, 2010). Thereby the rate of sedimentation is function of size, shape and density of the macromolecules, as well as density and viscosity of the gradient and applied centrifugal force (Raschke et al., 2009).
Here, triplicates of native (control) and RNase-treated cell lysates from HeLa cells synchronized in mitosis were loaded onto a 5% to 50% sucrose density gradient. Upon ultracentrifugation, the gradient was divided into 25 fractions and the protein amount of individual proteins per fraction determined by quantitative mass spectrometry. Because RNA-dependent proteins directly or indirectly depend on the presence of RNA, they are expected to show a different migration pattern between control and RNase-treated groups (Caudron-Herger et al., 2019).

In this report, we seek to draw a qualified conclusion about the RNA-dependence status of the 7160 proteins investigated in the provided data set using bioinformatics and statistics. For this purpose, we will define criteria and parameters the given proteins need to fulfill to qualify as RNA-dependent, and will evaluate our given data with respect to aforementioned aspects using the programming language R. Eventually, we will discuss the significance and possible limitations of our methods and findings.

Methods

Preliminary Steps

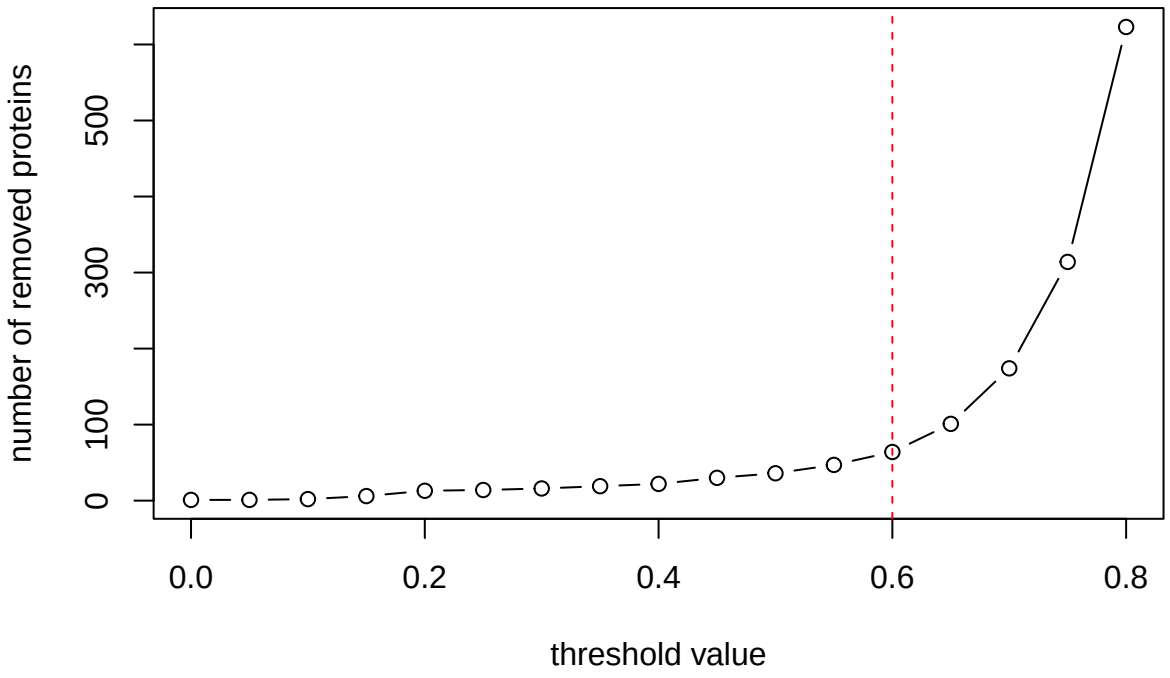
Before interpreting our data with respect to RNA dependent proteins, we performed a set of preliminary steps to ensure the validity and conclusiveness of our approach.

For the purpose of cleaning our data, we first tested whether every column's data type was numeric, which was the case for all 150 columns. Subsequently, we removed all proteins that contained negative protein amounts or only zeros in at least one of the three replicates for each condition (control and RNase), which was the case for a total of 7 proteins. Although replacement strategies do exist, we decided against keeping these proteins in our data set because these anomalies were not compatible with our further analysis steps and indicated an error in measurement.

In theory - all other things being equal - replicating the same experiment three times should produce three times the same result. This is equivalent to saying that, for a given fraction, the sum of protein amount values across all included proteins of the data set should be equal for all three replicates. However, this is not the case because in reality, experimental conditions aren't ideal and are subject to various sources of distortion, including naturally occurring statistical fluctuations and limited measurement accuracy. Thus, in order to ensure comparability of our data, it was necessary to perform a column-wise normalization step. For this purpose, we computed the sums of protein amount values for each replicate of a given fraction across all 7152 proteins still included in our data set. Subsequently, we computed the mean of the two closest sums, i. e. the two sums with the smallest absolute difference to each other. Using this mean value, we defined a normalization factor for each replicate of each fraction by dividing the sum of a given replicate by aforementioned mean value. This produced a vector of three normalization factors for each fraction. Last, protein amount values of a given replicate were multiplied with their corresponding normalization factor, resulting in a column-wise normalization in that for a given fraction, the sum of protein amount values across all included proteins of the data set was now equal for all three replicates.

Furthermore, since all proteins investigated in our data set naturally show a heterogeneous expression level in living cells, it was necessary to perform an additional row-wise normalization step to ensure comparability between different proteins. For this purpose, we set the protein amount to be 100 for each replicate in each condition. After this step, the sum of the protein amount across all 25 fractions was equal to 100 for each replicate. In other words, normalizing our data changed the absolute protein levels for each fraction while the protein amount in each fraction relative to the total protein amount across all fractions stayed the same.

Next, we wanted to evaluate the reproducibility of our data. For this purpose, we computed the correlation (Pearson) between replicates within each condition in all possible combinations (for example, correlation between control replicate 1 and 2, 1 and 3 and 2 and 3). This allowed us to sort out single proteins whose data did not seem to be reproducible. We defined a threshold correlation value below which the data for given proteins was regarded as non-reproducible. For this purpose, we looked at the number of proteins we would have to discard for each threshold level between 0 and 0.8 in 0.05 steps:



Considering this, we identified a threshold level of 0.6 as a good compromise between choosing a preferably high threshold level on the one hand and avoiding to remove too many proteins from our data set on the other. This led to the removal of another 64 proteins from our data set.

Last, for each protein and condition, all three replicates were combined into a single vector of protein amount values using the mean value between fractions. Although more robust against outliers, we decided against using the median for the purpose of combining replicates since many proteins contained fractions with a protein level of zero, which would have biased our further analysis.

Identification of RNA-dependent proteins

As mentioned earlier, RNA-dependent proteins are expected to show an altered migration pattern in a sucrose density gradient after ultracentrifugation in the presence or absence of RNA. After performing

aforementioned preliminary steps, we set to identify possible RNA-dependent proteins by interpreting this altered migration pattern, that is the distribution of normalized protein amount values across all 25 fractions for both conditions, control and RNase. For this it was necessary to define selection criteria proteins have to fulfill in order to qualify as RNA-dependent. Here, proteins are considered RNA-dependent when (i) the position of the global maximum of the protein amount across all 25 fractions shows a shift between control and RNase of 1 fraction or greater, and (ii) there is a significant difference in protein amount between the position of the global maximum in control and the respective fraction in RNase. This meant that first, we had to identify the global maximum for each protein in control and RNase condition, so that subsequently we could test for shift in position of the maximum and significance of difference in protein amount.

For the purpose of determining global maxima, we used aforementioned vectors that combine all three replicates of a given condition using the mean value method. Global maxima for all proteins were then identified by determining the greatest value of protein amount across all 25 fractions for a given protein. This resulted in two values, the averaged position of the global maximum in control and averaged position of the global maximum in RNase, wich were used to calculate the absolute magnitude of the shift.

Significance of difference in protein amount was tested using a Student's t-test. Here, we tested whether the protein amount in each RNase replicate significantly differed from the protein amount at the averaged position of the global maximum in control. Thus, we tested significance of difference in protein amount for each protein in our data set, leading to a total of 7127 t-tests performed on our data set, each test being significant if the computed p-value p was smaller than or equal to the chosen significance level α ($p \leq \alpha$). At the same time, the probability of an individual statistical test to be false-positive (type I error, i. e. the mistaken rejection of an actually true null hypothesis) is also equal to the significance level α, meaning that for a given α of 0.05, the probability of the individual t-test to produce a false-positive result is equal to 5%. However, performing multiple simultaneous statistical tests on the same data set, each having the probability α to produce a false-positive result, leads to a greater probability of making a type I error with respect to the whole family of performed t-tests than when considering each test individually, a phenomenon referred to as multiple testing problem. Thereby the probability of making one or more false discoveries when performing multiple hypotheses tests is given by the family-wise error rate (FWER; Fahrmeier Statistik): $FWER = 1 - (1 - \alpha)^k$, with k being the number of independent t-tests to be performed on the given data set and α being the significance level of the individual t-test. Thus, choosing the generally used α of 0.05 would lead to a probability α^* of making one or more false discoveries with respect to all 7127 t-tests performed of $\alpha^* = 1 - (1 - 0.05)^{7127} \approx 1$, meaning around 100%. To counteract this, we used the Bonferroni correction, defining our new significance level as $\alpha_{new} = \frac{\alpha_{old}}{k}$. Therefore,

t-tests werde regarded as significant if the computed p-value p was smaller than or equal to the the new significance level of $\alpha_{new} = \frac{0.05}{7127} \approx 7.02 * 10^{-6}$, leading to a new α^* of $\alpha^* = 1 - (1 - \frac{0.05}{7127})^{7127} \approx 0.05$.

Results

Discussion

- T-Test mit nur 2 Replikaten möglich, aber Mehraufwand für 7 Proteine nicht zumutbar, deshalb einfach entfernt + Mittelwert als Replacement
- <0.6 in mind. einem Vergleich o allen 3en??
- Abbildung, um column-wise normalization zu erklären??