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

Contents

Page

1. Introduction **1**

2. Methods **2**

2.1 Preliminary Steps 2

2.2 Identification of RNA-dependent proteins 4

2.3 Performing Dimension Reduction Analysis 5

2.4 Developing a Linear Regression Model 6

3. Results **6**

3.1 Identification of RNA-dependent proteins 6

3.2 Performing Dimension Reduction Analysis 8

3.3 Developing a Linear Regression Model 9

4. Discussion **11**

5. References **12**

1. 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., 2016; Castello et al., 2016), 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 after 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.

2. Methods

2.1 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 (Figure 1).

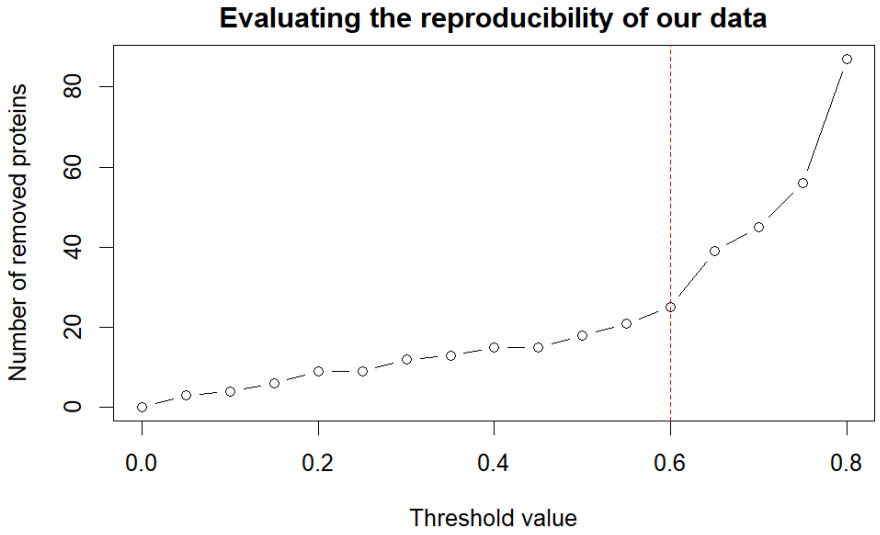

Figure 1: Number of discarded proteins dependent on the chosen threshold level. Proteins whose correlation is lower than the given threshold level are removed from the data set. One can observe a kink for a threshold level of around 0.6.

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 25 proteins from our data set. Therefore, we removed a total of 32 proteins from our data set, which amounts to around 0.5% of all proteins from our native 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 replicates for each fraction.

2.2 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 (a) the position of the global maximum of the protein amount across all 25 fractions shows an absolute shift between control and RNase of 1 fraction or greater, and (b) there is a significant difference in protein amount between the position

of the global maxima in control and the protein amount at the respective fractions 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 global maximum and significance of difference in protein amount.

For the purpose of determining the shift in position of the global maximum, 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 the averaged position of the global maximum in RNase, which were used to calculate the absolute and real magnitude of the shift.

Significance of difference in protein amount was tested using a Student's t-test. Here, for every given protein, we first determined the global maximum for each replicate in control, yielding a vector containing the magnitude and the respective fraction of each maximum. Subsequently, protein amount values at the respective fractions in RNase were determined, yielding a second vector containing protein amount values at the respective fractions. We then tested whether protein amount values from the first vector significantly differed from those of the second vector and thus tested significance of difference in protein amount for each protein in our data set, leading to a total of 7127 t-tests performed. P-values were corrected for multiple testing by controlling the false-discovery rate (FDR) according to Benjamini and Hochberg (1995), using the function `p.adjust`. Therefore, difference in protein amount was regarded as significant if the FDR-corrected p-value p was smaller than the chosen significance level of $\alpha = 0.05$.

2.3 Performing Dimension Reduction Analysis

In this step of our analysis, we looked at a data set that, for a given protein, combine all three replicates of a given condition using the mean value method, meaning that each protein can be characterized by fifty values - 25 mean protein amount values from control and 25 mean protein amount values from RNase.Thus, each protein from our data set can be depicted as a point in a fifty-dimensional space. Here, in order to identify directions in the data corresponding to biological effects, we sought to transform our data from this high-dimensional space into a lower-dimensional space with fewer relevant variables while retaining as much variance in our data as possible. For this purpose, we performed principal component analysis (PCA) on our data set using the built-in

function `prcomp`, followed by Uniform Manifold Approximation and Projection (UMAP) analysis using the function `umap`. Subsequently, we colored proteins according to previously-determined features, once with respect to their RNA-dependence status and once with respect to their shifting behavior, further differentiating between left shifting, right shifting and not shifting proteins. Left shifting proteins were defined as proteins whose maximum shifted towards smaller fractions between control and RNase condition as determined in part 2.2, i. e. proteins whose real shift in position of the global maximum was negative. Similarly, right shifting proteins showed a shift in position of the global maximum towards higher fractions, and not shifting proteins a real shift of zero.

2.4 Developing a Linear Regression Model

Last, for all proteins included in our data set, we sought to develop a linear regression model that features variables from our previous analysis steps. Here, we tried to predict the absolute and the real shift in position of the global maximum using the correlation between the mean protein amount values between control and RNase for each given protein. This approach is based on the rationale that proteins with low correlation between control and RNase are expected to show distinct alterations in the distribution of protein amount values across all 25 fractions between both conditions, which in turn may increase the chance for a protein to show a shifting behavior by our definitions as stated above. For that purpose, we randomly split our data set so that 80% of all proteins included served to train our model while the remaining 20% were used to test it. We used the built-in function `lm` to do the training and `lm.predict` for the purpose of testing.

3. Results

3.1 Identification of RNA-dependent proteins

According to aforementioned methods and criteria, we computed the RNA-dependence status for all 7127 still included in our data set. Here, we found 1320 proteins that only fulfilled the first criterion (shift) for proteins to qualify as RNA-dependent and 1547 proteins that exclusively fulfilled the second criterion (significant t-test) but not the first one. The intersection between both groups amounted to 657 proteins, i.e. we identified a total of 657 as RNA-dependent, which makes up around 9.2% of all proteins tested. Using the two additional data sets we've

been granted, we can check whether the proteins we identified to be RNA-dependent are in fact RNA-binding or not (Table 1).

Table 1: Comparison between the number of RNA-dependent (R-Deep) an non-RNA-dependent proteins (Non-R-Deep) as identified by our analysis and the number of RNA-binding (RBP) or non-RNA-binding proteins (Non-RBP) set to occur in our data set.

	R-Deep	Non-R-Deep	Sum
RBP	562	3961	4523
Non-RBP	91	2460	2551
Sum	653	6421	7074

Because the notion of RNA-dependence embeds that of RNA-binding, we can approximate a number of key performance metrics to assess the quality of our analysis (Table 2).

Table 2: Selection of different performance metrics to evaluate the quality of our analysis. FNR, False Negative Rate; FPR, False Positive Rate; FDR, False Discovery Rate.

	Number of Proteins		Percent
False Positives	91	FNR	87.6
True Positives	562	FPR	3.67
False Negatives	3961	FDR	13.9
True Negatives	2460	Precision	86.1
		Recall	12.4

Using these metrics, we can conclude that (a) our analysis shows strengths in its relatively high precision, i. e. a relatively high rate by which RNA-dependent proteins as identified by our analysis are in fact RNA-dependent, and that (b) our analysis is limited in that it only shows a relatively low recall, i. e. it does only a poor job in retrieving truly RNA-dependent proteins from the set of all RNA-dependent proteins in our data set.

3.2 Performing Dimension Reduction Analysis

Dimension reduction analysis was performed by aforementioned methods and proteins colored according to their shifting behavior (Figure 2).

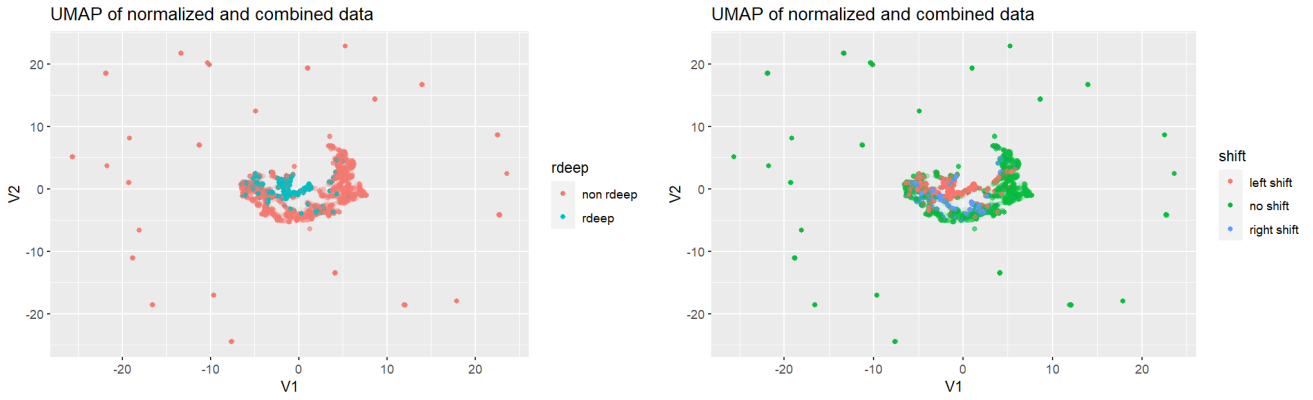

Figure 2: PCA and subsequent UMAP analysis applied to our normalized and mean-combined data set. Left: proteins colored to their RNA-dependence status; Right: proteins colored to their shifting behaviors. Overplotting was reduced by increasing transparency for each point in the scatter plot.

Here, it can be seen that proteins with the same shifting behavior generally tend to organize themselves into regionally distinct groups in this low-dimensional representation. Expectedly, these findings suggest that there may exist an underlying group-like structure in our data set that seems to be valid even outside of our arbitrarily-defined groups of shifting behaviors. In other words, depicting proteins as done here is visual evidence that clustering proteins by shifting behavior may be a good choice for identifying sub-structures in our data set because it is based upon immanent groups of proteins in a low-dimensional space. As a follow-up analysis, we took a closer look at the distribution of shifting behaviors in our data set in order to identify directions in the data that may point towards corresponding biological effects. For that purpose, we plotted the position of the global maximum in control against the global maximum in RNase using our results from analysis part 2.2 (Figure 3).

Notably, this representation visualizes that there exist almost no proteins that show a drastic right shift - i. e. a shift towards higher fractions between control an RNase - while the latter can be observed for left-shifting proteins. Referring back to the introduction, we know that in density gradient centrifugation, macromolecules are forced through a density gradient until they find a density equal to their own (Farrell, 2010), and that thereby the rate of sedimentation depends, among others, on size, shape and density of the macromolecules (Raschke et al., 2009).

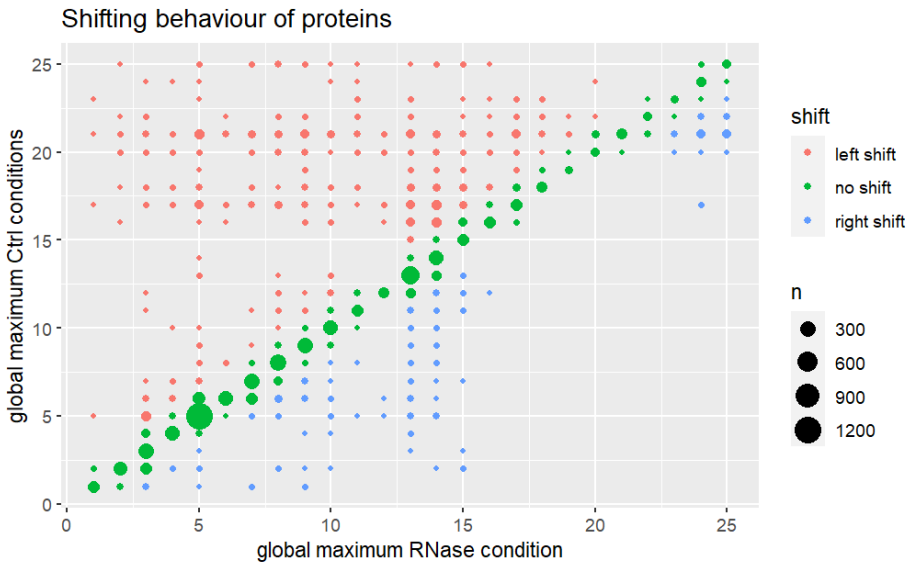

Figure 3: Position of global maximum in control against position of global maximum in RNase. For each observable shift, the number of proteins n that display this specific shift is represented by the size of the points.

Therefore, these findings suggest that, generally speaking, there may exist a tendency for proteins to experience a loss in density and/or conformational changes that are unfavorable for migration in a density gradient, i. e. slow down the rate of sedimentation, as a result of RNase treatment.

3.3 Developing a Linear Regression Model

Linear regression was performed using aforementioned methods and functions. First, we sought to predict the real values in shift of the maxima between control and RNase using the correlation values computed earlier (Figure 4).

Using these findings, we could assess whether or not our linear model serves well in modelling the relationship between real values in shift and correlation. Most importantly, one can observe a p-value of the F-test of $p < 2.2 * 10^{-16}$, meaning that for a given α of $\alpha < 0.05$ our regression model is significantly better than the null-model, i. e. a regression model where for a given correlation the real values in shift are simply predicted as the mean of all real shifting values used to train the model. Likewise, since both p-values are significant ($p < 2 * 10^{-16}$), we can exclude that both intercept and slope of our model are actually equal to zero. However, our model could only reach a coefficient of determination (R^2) of around 0.6, meaning that only around 60% of the total variance in the dependent variable is explained by our regression model. Furthermore, judging by the way the regression line is located in the data cloud (Figure 4, right), it can be said that our model

```
Call:
lm(formula = shift ~ correlation, data = df_lm_train)

Residuals:
     Min       1Q   Median       3Q      Max
-21.2604  -0.3627  -0.2122   0.3780  16.1125

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.0999     0.1413  -92.73   <2e-16 ***
correlation  13.4626     0.1525   88.28   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.266 on 5719 degrees of freedom
Multiple R-squared:  0.5768,Adjusted R-squared:  0.5767
F-statistic:  7794 on 1 and 5719 DF,  p-value: < 2.2e-16
```

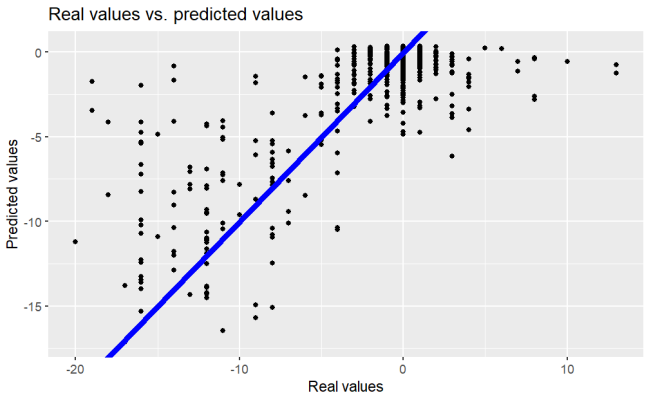

Figure 4: Left: summary of key parameters of the first regression model. Right: real values in shift as predicted by the first regression model in comparison to the values computed by our analysis. Note that our model does only a poor job in predicting real values in shift for right shifting proteins.

serves well in describing left shifts but does only a poor job in describing right ones. As a follow-up quality assessment, we tested whether or not the assumption that the relationship between the real values in shift and correlation is linear was actually valid to make. For that purpose, we looked at different properties of the residuals (Figure 5).

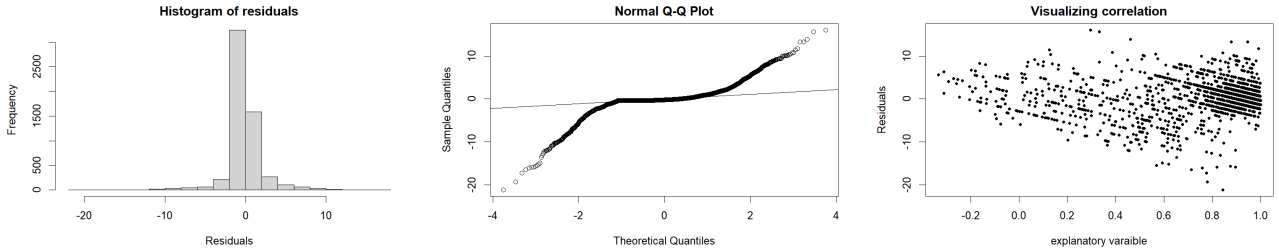

Figure 5: Graphical representation of different properties of the resiudals from our linear model. Left: histogram of residuals; note that the mean value is zero. Middle: Q-Q plot plotting quantiles of the distribution of our residuals against those of a normal distribution to assess distribution of residuals. Right: scatter plot displaying residuals against our explanatory variable (correlation).

Using these findings, we conclude that the linearity assumption is valid since our residuals (a) show a mean value of zero, (b) are approximately normally distributed and (c) do not correlate with our explanatory variable ($corr = 4.958122 * 10^{-16}$). Last, having trained our model and tested its validity, we wanted to assess its performance in predicting real values in shift using correlation values for a given protein using the remaining 20% of our data set. For that purpose, we used the built-in function `lm.predict` to test in what fraction of cases the real value in shift happens to be within the 95 percent confidence interval as predicted by our model. In that way we computed that our model was accurate only in around 6% of cases. Because of that - and because this first model showed a tendency to preferably explain left shifts but not right ones - we developed

another regression model that sought to predict not the real but the absolute value in shift from correlation, all else being equal (Figure 6).

```
Call:
lm(formula = shift ~ correlation, data = df_lm_train_ab)

Residuals:
     Min       1Q   Median       3Q      Max
-12.0198  -0.4977  -0.1277  -0.0555  20.8095

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.7272     0.1235   111.2   <2e-16 ***
correlation -13.6717     0.1333  -102.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.981 on 5719 degrees of freedom
Multiple R-squared:  0.6478,Adjusted R-squared:  0.6477
F-statistic: 1.052e+04 on 1 and 5719 DF,  p-value: < 2.2e-16
```

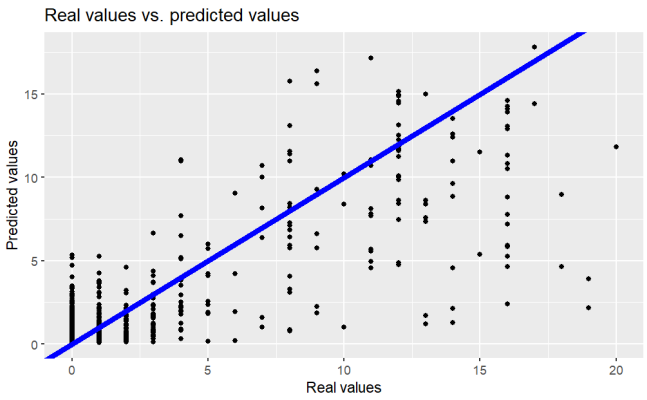

Figure 6: Left: summary of key parameters of the second regression model. Right: real values in shift as predicted by the second regression model in comparison to the values computed by our analysis.

Again, we can find that this model serves significantly better in predicting absolute values in shift than the null-model (p-value of F-test $< 2.2 * 10^{-16}$). Assuming linearity was found to be valid by aforementioned methods. However, while having a very similar coefficient of determination ($R^2 = 0.6466$), we found that this model using absolute values in shift rather than real ones was more accurate in predicting values in shift for new proteins in that the predicted values were found to lie within the 95 percent confidence interval in around 20% of cases.

4. 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??

5. References

Beckmann, B.M., Castello, A., and Medenbach, J. (2016). The expanding universe of ribonucleoproteins: Of novel RNA-binding proteins and unconventional interactions. Pflügers Archiv-European Journal of Physiology *468*, 1029–1040.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) *57*, 289–300.

Castello, A., Fischer, B., Frese, C.K., Horos, R., Alleaume, A.-M., Foehr, S., Curk, T., Krijgsveld, J., and Hentze, M.W. (2016). Comprehensive identification of RNA-binding domains in human cells. Molecular Cell *63*, 696–710.

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.

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.

Faoro, C., and Ataide, S.F. (2021). Noncanonical functions and cellular dynamics of the mammalian signal recognition particle components. Frontiers in Molecular Biosciences 420.

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

Maziuk, B., Ballance, H.I., and Wolozin, B. (2017). Dysregulation of RNA binding protein aggregation in neurodegenerative disorders. Frontiers in Molecular Neuroscience *10*, 89.

Mullari, M., Lyon, D., Jensen, L.J., and Nielsen, M.L. (2017). Specifying RNA-binding regions in proteins by peptide cross-linking and affinity purification. Journal of Proteome Research *16*, 2762–2772.

Pratt, A.J., and MacRae, I.J. (2009). The RNA-induced silencing complex: A versatile gene-silencing machine. Journal of Biological Chemistry *284*, 17897–17901.

Urdaneta, E.C., Vieira-Vieira, C.H., Hick, T., Wessels, H.-H., Figini, D., Moschall, R., Medenbach, J., Ohler, U., Granneman, S., Selbach, M., et al. (2019). Purification of cross-linked RNA-protein complexes by phenol-toluol extraction. Nature Communications *10*, 1–17.

Zhang, B., Babu, K.R., Lim, C.Y., Kwok, Z.H., Li, J., Zhou, S., Yang, H., and Tay, Y. (2020). A comprehensive expression landscape of RNA-binding proteins (RBPs) across 16 human cancer types. RNA Biology *17*, 211–226.