RiboSix –Story of an RNA-Binding Protein

Proteome-wide screen for RNA-dependent proteins particularly relevant in mitosis

Data Analysis 2025 – Project 03 Group 02 – Baumüller, Lledo Padova, Zeyrek

UNIVERSITÄT HEIDELBERG
ZUKUNFT
SEIT 1386

IPMB

The world of RBPs

In a small space called HeLa, there are many small molecules working together creating one unit. They are going through many seasons giving their all to make it work, but not necessarily everyone is working during every season. This is what makes living there for them so beautiful, after a lot of hard work many of them can gather their energy. One season is called mitosis and this is the season of our little protein RiboSix.

So, join us on his journey to discover the village of HeLa.

1 Normalization: Let's get this going

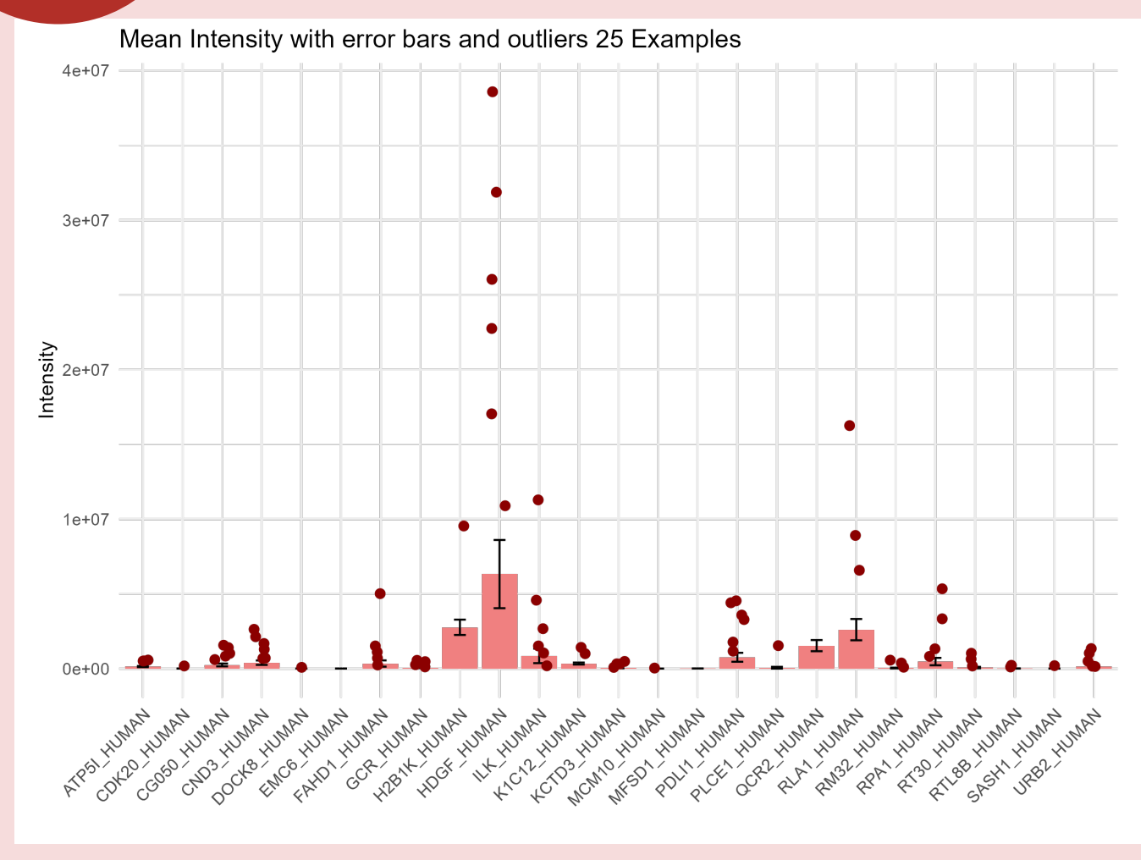


Fig. 2 Mean protein intensities across 25 fractions under control conditions: Bar plot showing the average intensity for 25 random proteins measured across 25 fractions in the Ctrl condition. Each bar represents the mean intensity per protein, with error bars indicating the standard error of the mean (SEM) across fractions. This visualization highlights the variability in abundance profiles among different proteins across the cellular gradient.

Description of MS-Dataset

- Number of rows: 7195 (Analyzed Proteins)
- Number of columns: 150 (Fractions, Reps and Treatments)
- Overall maximum intensity: ~ 15 x 10^8
- Overall minimum intensity: 0
- No NAs and only numerical values
- High variance in intensity between proteins

How we normalizes MS data for our analysis

- Average of the triplicates for every fraction and condition was computed
- To ensure comparability between proteins, each protein is scaled so that the distribution within the Ctrl and RNase conditions each sums to 100

2 Shift Analysis: Finding my friends

Descriptive Analyses – That's how I look like

Peak Analysis: For each protein profile, up to 6 peaks were identified using a slope-based function on all normalized values for control- and RNase-treatment. (Threshold : 3% of maximal signal intensity)

Shift Characteristics Protein distributions were summarized using the Center of Mass (CoM), calculated as the weighted average across all fractions.

$$Shift\ distance\ = CoM\ Ctrl\ - CoM\ RNase$$

Left shift: distance > 0 ; Right shift: distance < 0 ;
No shift: distance ~ 0

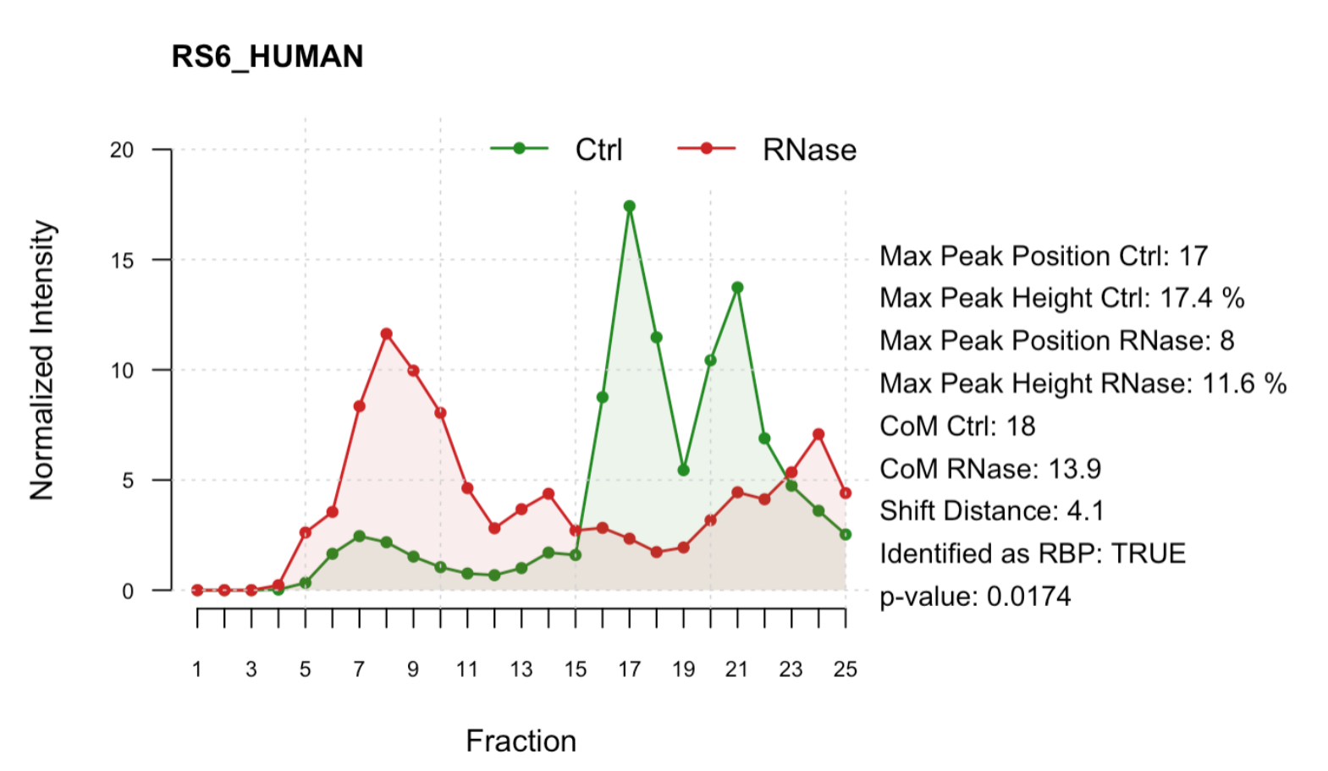


Fig. 3 Intensity profile of RS6 : *Plot shows normalized signal distributions as well as extracted descriptive parameters such as peak positions, peak heights, and shift distance, t-test results.*

$$CoM = \frac{\sum_{i=1}^{25} fraction_i * \ intensity_i}{\sum_{i=1}^{25} intensity_i}$$

Statistical Analysis – Do I have a connection with RNA?

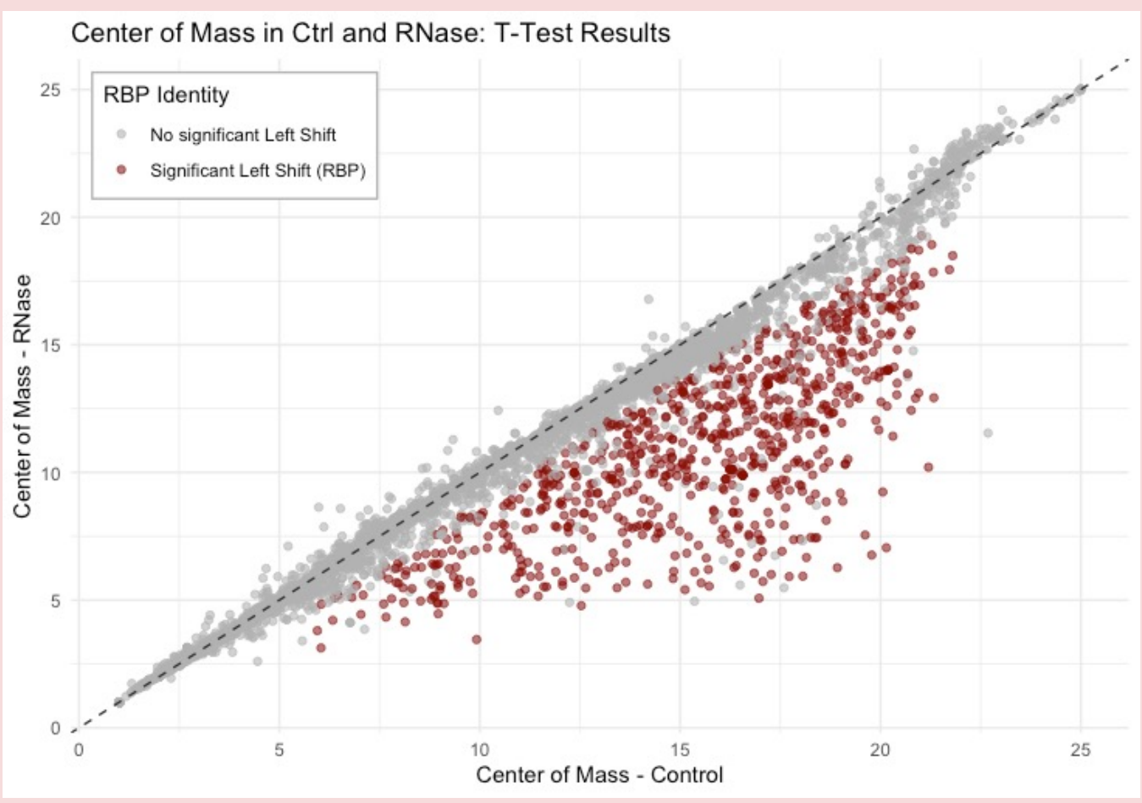


Fig. 4 Visual Presentation of T-Test Results: *Using a scatterplot of CoM-Ctrl and CoM-RNase to visualize shift distance and direction.*

T-Test: To statistically assess RNA dependence, shift distances from CoM values across all replicates were computed. A Shapiro-Wilk test was performed to confirm normality. If normally distributed, a one-sided t-test was used to assess whether the mean shift exceeded 1.

794 proteins exhibited a significant left shift and were classified as RBPs including RiboSix !

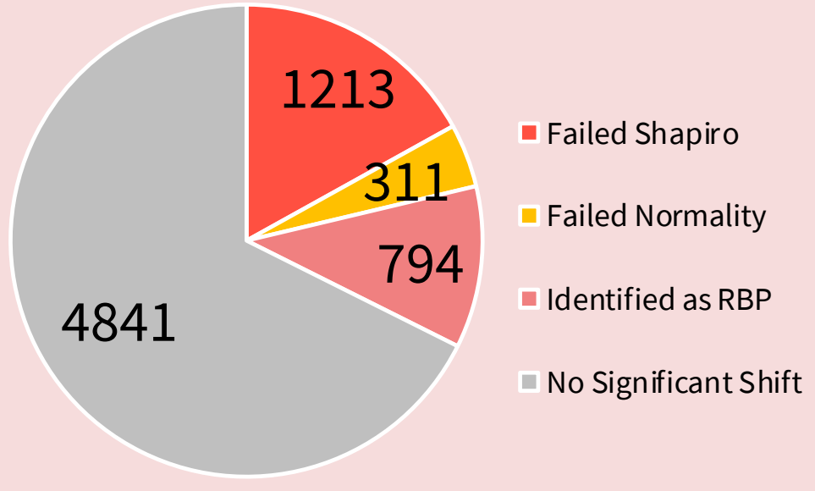
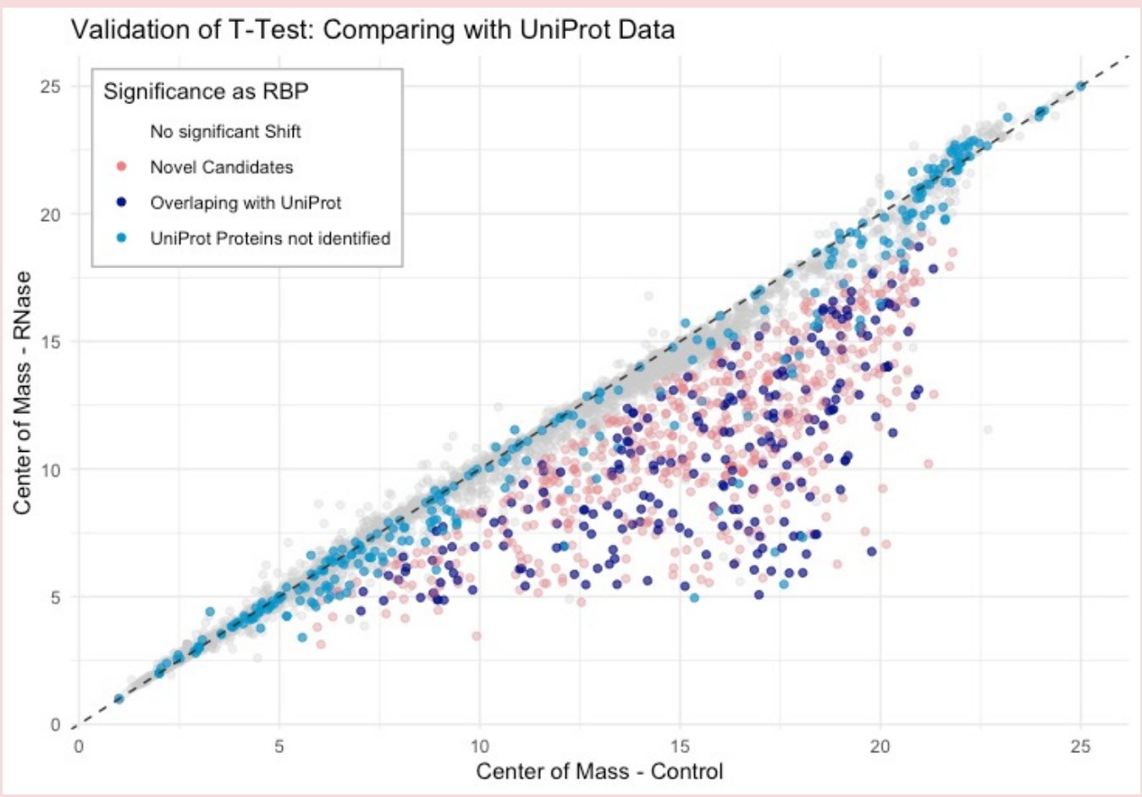


Fig. 5 Results and Limitations of Shift Significance Testing: *Outcome of t-test and pipeline evaluation for all proteins, with representation of all excluded proteins.*

Validation of Test Results – Comparing with UniProt



UniProt was used as a reference to identify proteins previously annotated as RNA-binding or RNA-interacting.

- 3,114 human proteins with RNA-binding function are listed in UniProt (based on experimental data and literature)
- 543 were present in our dataset
- **230 were correctly identified as RNA-dependent (hit rate: 42.4%)**
- 564 identified remaining proteins are novel candidates

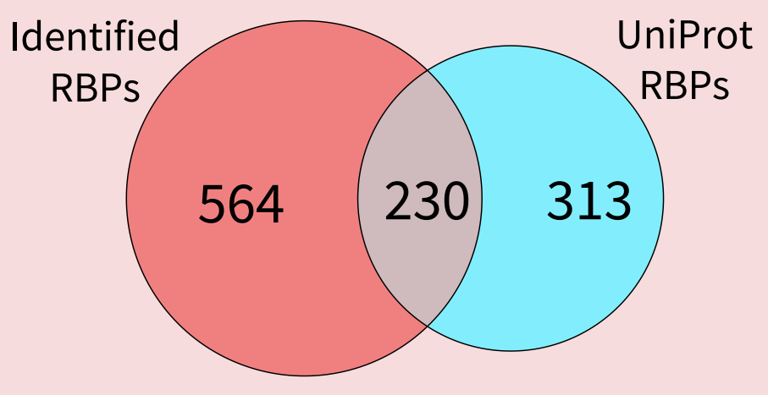


Fig. 6 Novel Candidates and known RBPs: *Intersecting between previously noted UniProt RBPs analysed in our sample and RBPs identified by this pipeline*

Linear Regression: Maybe it's better not to step on the scale

Hypothesis: In theory, heavier proteins migrate to deeper fractions in a sucrose gradient, so we therefore hypothesized, that a protein's peak position after RNase treatment might reflect its molecular weight. Monomeric molecular weights were retrieved from UniProt. To illustrate the expected relationship, we included five standard reference proteins from *Caudron- Herger et al. (2019)* with known mass and elution positions.

However, most proteins did not follow the expected trend!

Results: Predicting molecular weight by maximal peak position

- Spearman Correlation: $\rho = 0.014, p = 0.25$
- Linear Regression: $R^2 = 0.00017, p = 0.26$ *(F-Test)*

Further Analysis:

- All tests we repeated using the Center of Mass (CoM) instead of peak position.
- We also tested the hypothesis that many proteins may remain in RNA-independent complexes after RNase treatment, which could distort elution profiles. To test this, we removed all 2200 proteins listed in the CORUM* database.

→ No improvement in correlation or regression

Discussion:

Elution does not only depend on weight, but also on shape, size and density, so peak based features might be too simplistic. CORUM does not reflect all protein interactions, that might influence elution.

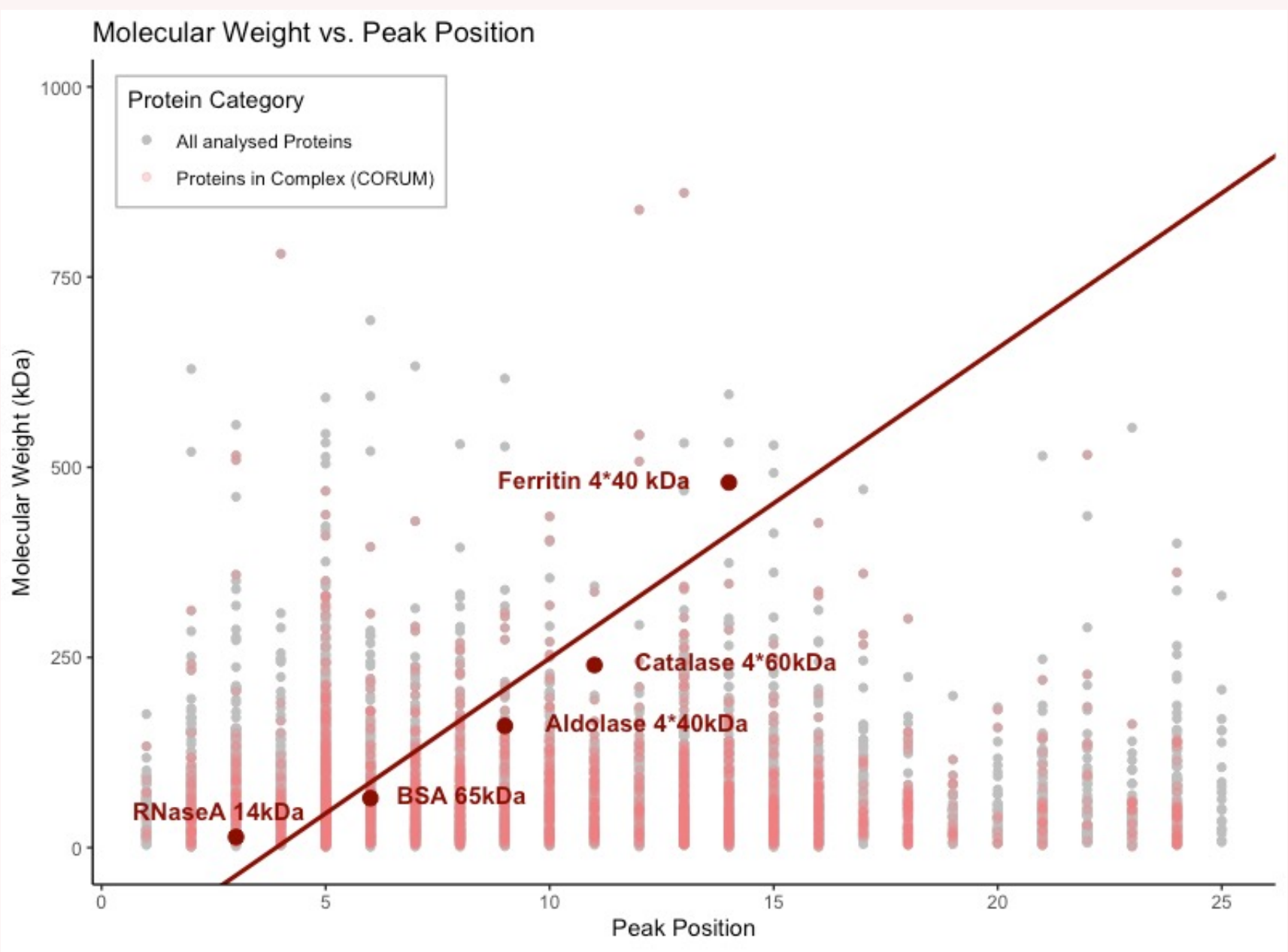


Fig. 10 Relation of monomeric molecular weight and maximal peak position : *Scatterplot of all analyzed proteins, showing their monomeric molecular weight (UniProt) versus maximal peak position in the RNase condition. Red line shows expected linear trend based on five reference proteins (Caudron-Herger et al., 2019).*

** CORUM is a curated database of experimentally validated protein complexes in mammals.*

Our Goal: Hunting RNA-Binding Proteins in the DeeP

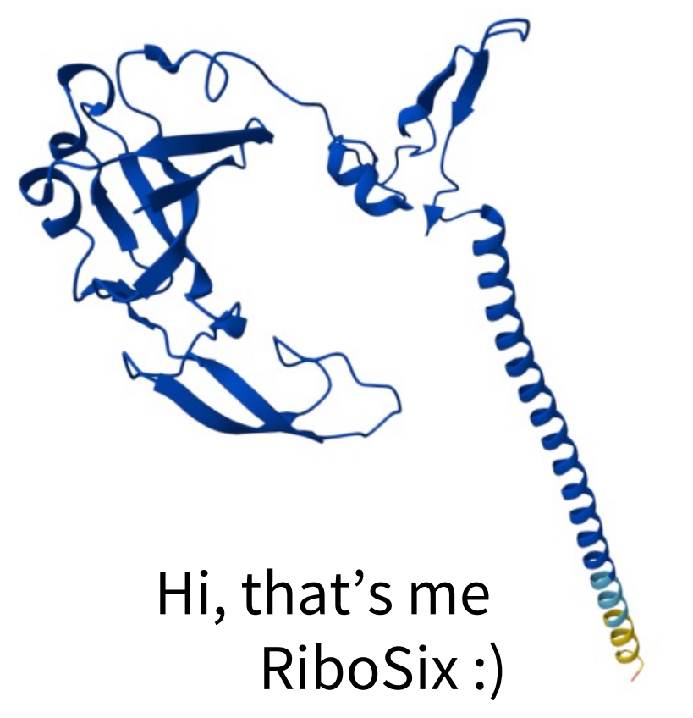
The main goal of our project was to identify RNA-binding proteins (RBPs) in mitotic HeLa cells. To achieve this, proteins were fractionated with and without RNase treatment. Each sample was separated into 25 fractions, and protein intensities were measured using mass spectrometry in triplicates.

To uncover potential RBPs, we performed the following key steps:

- Reproducibility analysis
- Normalization of the data
- Peak characterization
- Shift analysis, where a left shift in the RNase condition indicates RBP behavior

To gain deeper insights, we extended the analysis by:

- Identifying RBPs specifically active during mitosis
- Clustering peak characteristics to reveal potential complexes
- Performing linear regression to predict molecular weight from peak data



Hi, that's me
RiboSix :)

Reproducibility Analysis: Am I real?

Theory: High reproducibility is indicated by strong correlations between replicates of the same fraction

Method:

- Reproducibility was assessed via Spearman correlation between all replicate & fraction combinations, performed separately for RNase and control conditions
- Resulting correlation coefficients (r-values) were visualized in heatmap

Results: High reproducibility visible on the heatmap as a distinct diagonal pattern forming 3×3 blocks, consistently observed across both treatment conditions **→ R-DeeP results are reproducible**

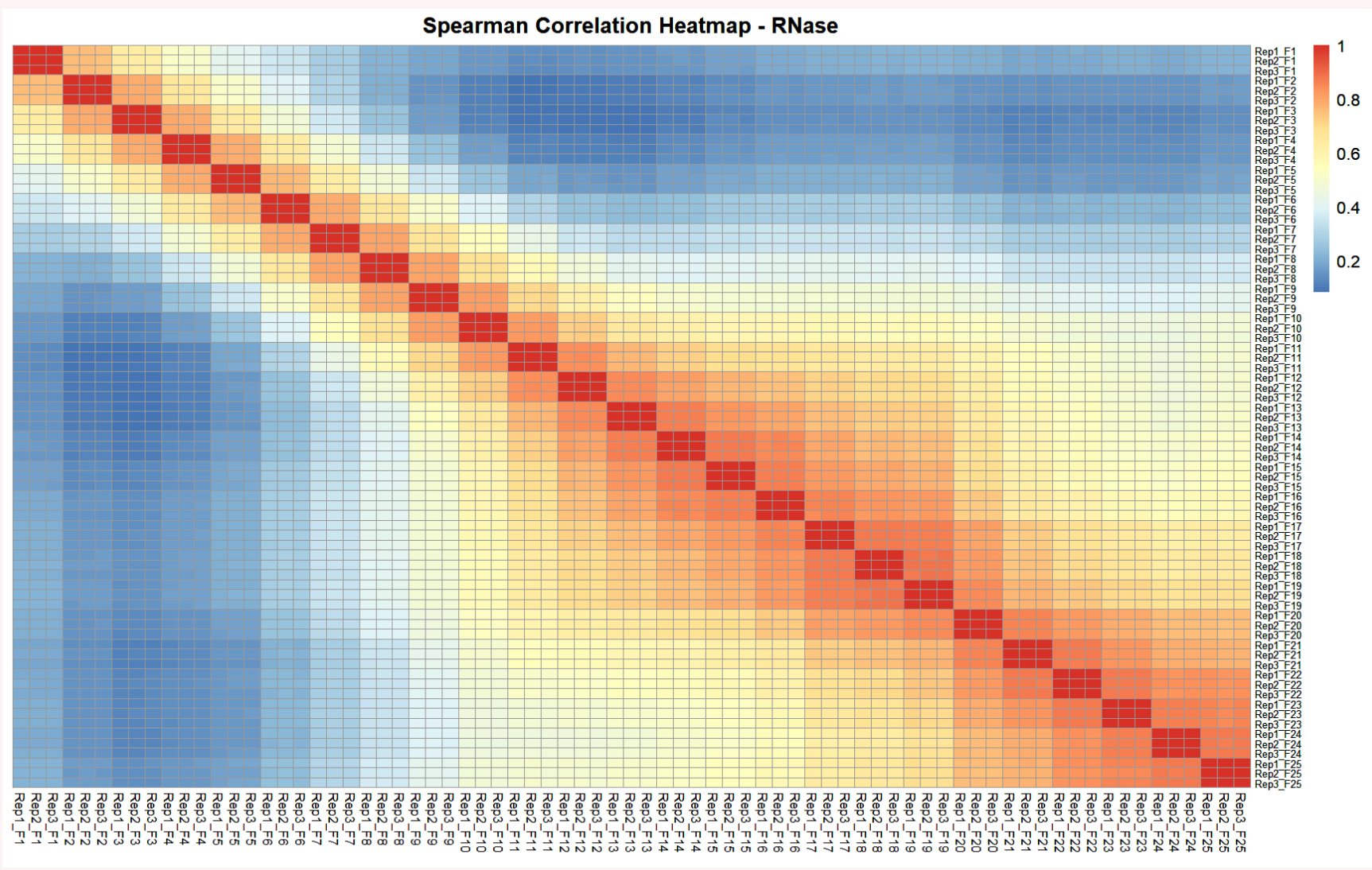


Fig. 1 Reproducibility heatmap (RNase, Spearman correlation): Heatmap displays pairwise Spearman correlation coefficients between all replicate–fraction combinations under RNase treatment. High correlations within 3×3 diagonal blocks indicate strong reproducibility across corresponding fractions.

Finding Mitosis-Specific RBPs: Clocking in for the season 3

Comparative Shift Analysis

Shift analysis pipeline was applied to non-synchronized HeLa cells. To visualize similarities and differences a scatterplot compares shift distances between mitotic and non-synchronized cells.

- 376 Proteins highlighted in pink were identified as RBPs in both samples
- 298 Proteins highlighted in dark gray where only identified in non-sychronised cells and might not be active in mitosis
- **237 Proteins highlighted in red show significant RNA dependency exclusively during mitosis.** One of which is RiboSix, suggesting that mitosis is his active season in the village of HeLa.

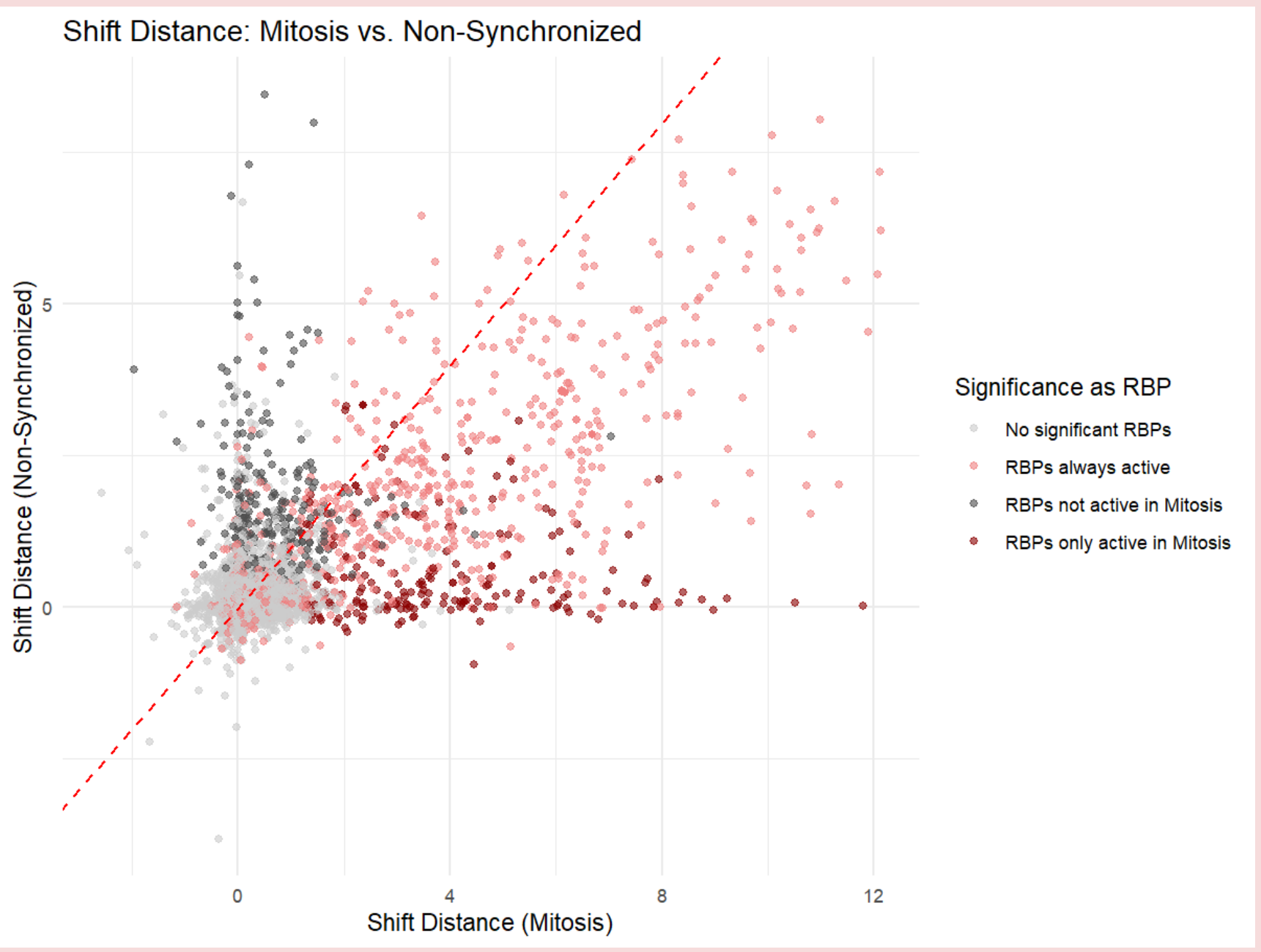


Fig. 7 ***Comparative Shift Scatterplot (Mitosis vs. Non-Synchronized):*** Scatterplot displays shift distances derived from center of mass (CoM) values for all proteins under both conditions. Each point represents one protein, color-coded by statistical significance. The red dashed identity line marks equal shift behavior; proteins below the line show mitosis-specific leftward shifts, suggesting RNA dependency unique to mitosis.

Complex Analysis: Finding my Family 4

Hypothesis: We based our clustering analysis on the idea that proteins forming a complex should co-migrate within the same fraction—at least under control conditions. If they are physically associated, they are expected to shift together and show peak abundance in the same MS fraction.

DBSCAN: is a clustering algorithm that considers point density and distance.

- **ε:** The maximum distance between two points to be considered neighbors.
- **MinPts:** The minimum number of neighbors (within ε distance) to form a core point, border points are those within ε of a core point.

Choosing parameters based on control proteins

To validate the efficiency in clustering a heatmap with a specific scoring logic was created. Parameters were adjusted accordingly **to ε = 0.7 and MintPts = 4.**

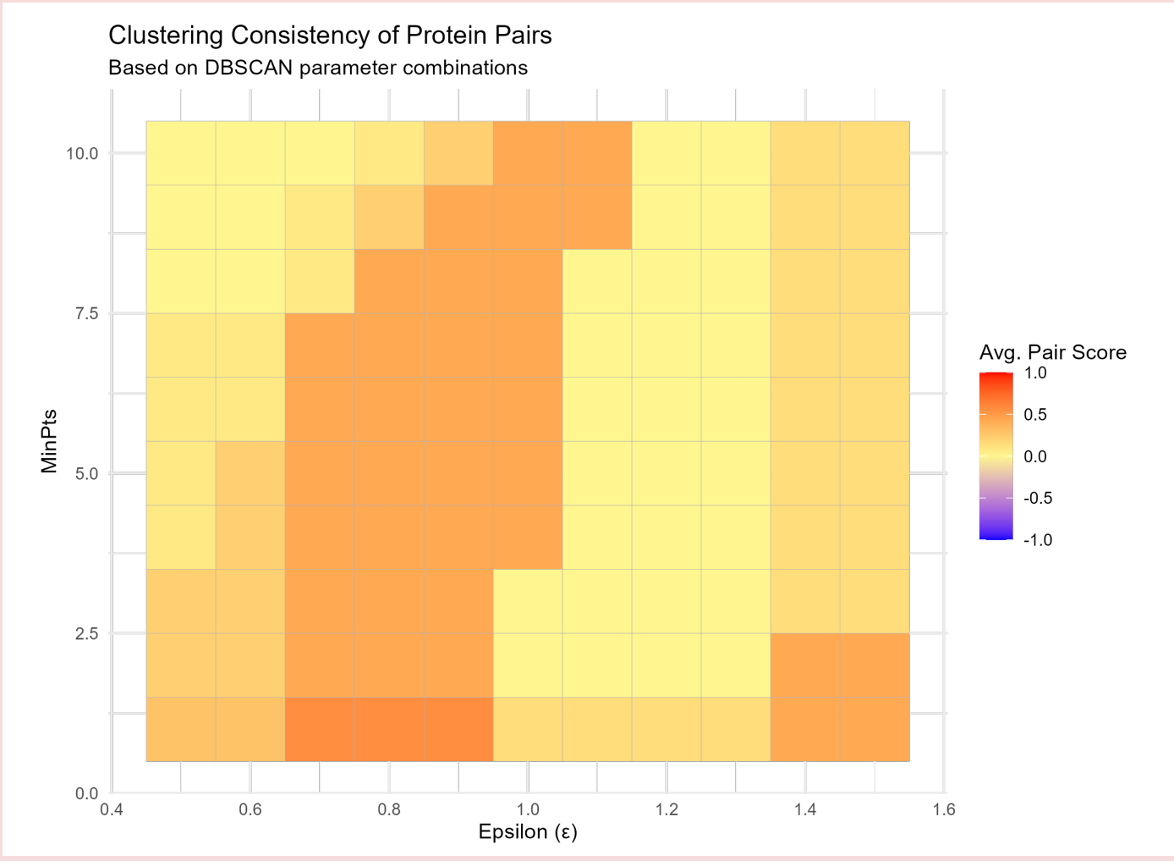


Fig. 9 ***Heatmap of accuracy for combinations of Parameters for DBSCAN:*** *Accuracy calculated on pos. and neg. controls, from* ε (0.5-1.5) and MintPts (1-10). Lower and higher ε lower the accuracy.

Positive control: 4 proteins of our mitosis specific RBPs known to be in the 40S Ribosomal Complex

Negative control: UIMC1_HUMAN and LPPRC_HUMAN

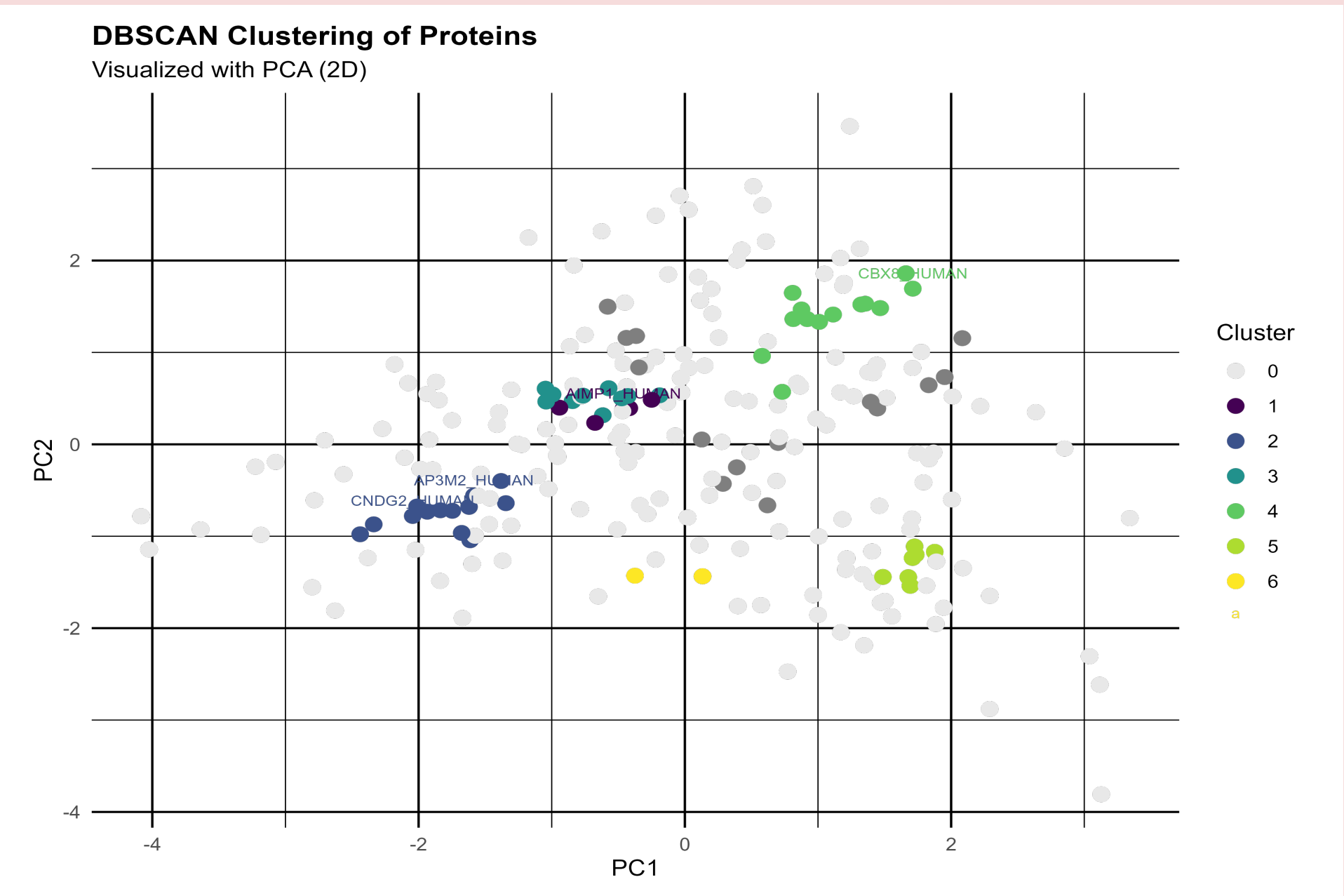
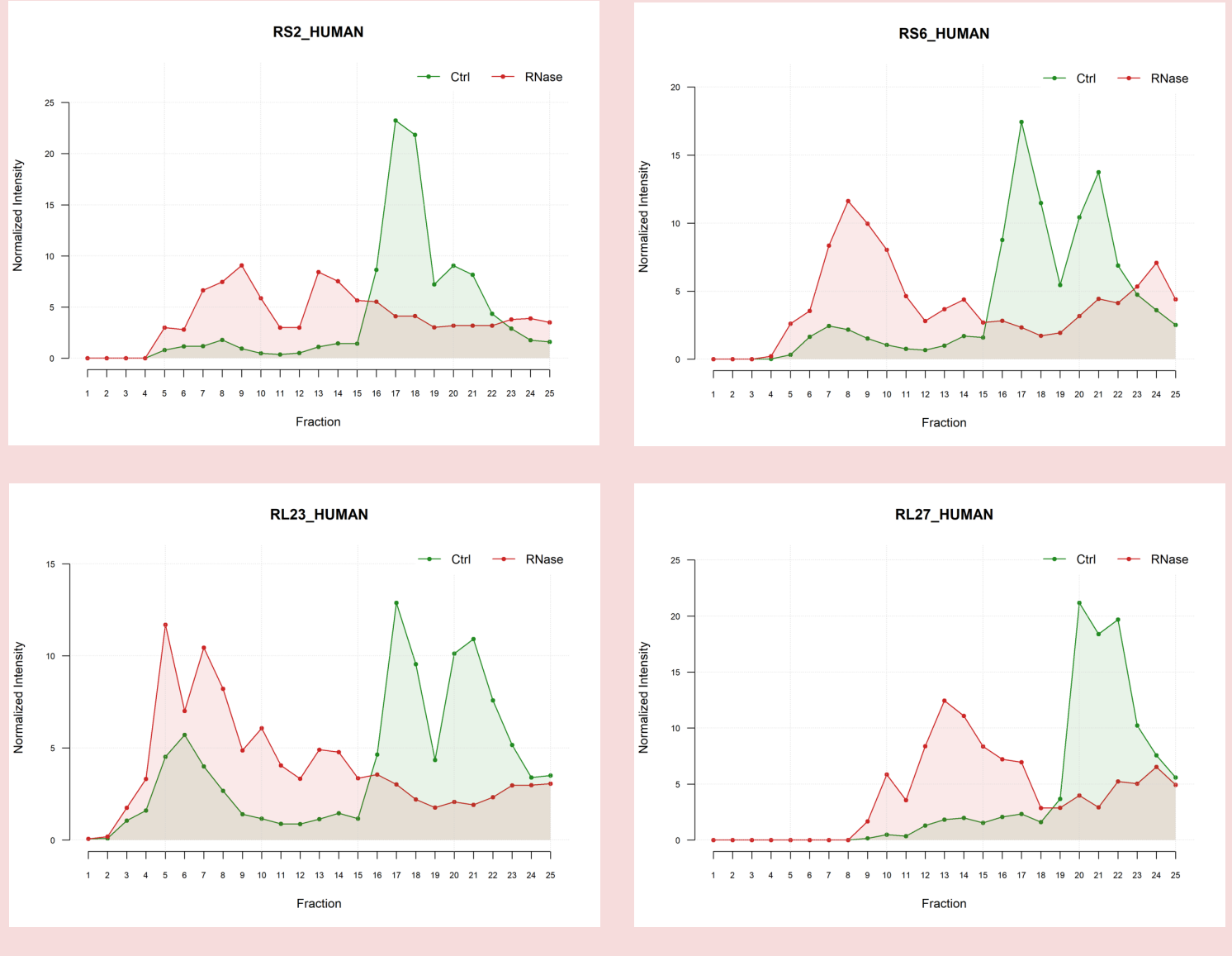


Fig. 8 Proteins in 2D PCA showing results from clustering method DBSCAN: Only Proteins from RBPs in Mitosis where clustered. Dimension reduction on Data from Ctrl : COM and Peak height. **ε = 0.7 and MintPts = 4.**



Hey there are other that look just like me :)

→ I found my family in cluster 4. See there are some friends out of the Nop56p-associated pre-rRNA complex

Representative Results for Cluster 4 (13 Proteins) :

- From 40S Ribosomal Complex: 3 out of 4 proteins were clustered
- From Nop56p-associated pre-rRNA complex: 4 out of 9 were clustered

Main Findings

→ Out of 7.159 analyzed proteins **749 RNA-binding proteins (RBPs)** were identified based on significant distribution

→ Of these, 230 RBPs are known (UniProt-annotated), while **564 are likely novel RBP candidates**.

→ **237 RBPs are specifically active during mitosis**, confirmed by direct comparison with non-synchronized HeLa cells.

→ R-DeeP successfully enables **complex analysis:** known and potentially novel **protein complexes** can be detected based on clustering of distribution profiles using DBSCAN.

→ No valid linear regression model could predict molecular weight from gradient shifts. This approach might be to simplistic sice elution alsp depends on shape, **protein density and protein interactions.**

On this journey RiboSix learned three things: he's an RNA-dependent protein, he might be active exclusively in mitosis, and — perhaps most meaningfully — he's not alone. He shares his fraction with others. He belongs to a complex. He found his family!