Reina Hastings

BIOL606 – Midterm 3

May 07, 2025

## *Multivariate Analysis of Neural Protein Expression in*

## *Down Syndrome Mouse Model*

## **Methods**

### Data Cleaning and Standardization

The mouse protein expression data set was retrieved from *Self-Organizing Feature Maps Identify Proteins Critical to Learning in a Mouse Model of Down Syndrome*. Categorical variables (Genotype, Treatment, Behavior) were converted to factors for statistical analysis and rows containing missing values or MouseID entries with "3429" were removed. Mouse 3429 was removed as an outlier individual.

### Feature Selection and Scaling

The 10 most relevant proteins were selected based on ANOVA p-values for Genotype, Treatment, and Behavior. Relevant proteins were visualized and assessed for normality via histograms and Q-Q plot. High-skew proteins were log-transformed.

### Principal Component Analysis (PCA)

PCA was performed to reduce the dimensionality of the data and identify key sources of variance. The top 10 selected proteins were used to construct the PCA matrix, which was mean-centered and standardized using prcomp() to ensure each variable had equal weight. Several statistical checks and analyses were performed.

* The proportion of variance explained by each principal component (PC) was calculated, and a scree plot was generated to visualize the contribution of each PC.
* The top 5 proteins contributing to PC1 and PC2 were identified based on the absolute loading magnitudes.
* Q-Q plots were used to assess the normality of PC1 and PC2 scores.
* PCA scores were combined with grouping metadata (Genotype, Treatment, Behavior) to assess experimental effects. ANOVA was performed to test the significance of each grouping variable on PC1 and PC2.
* A permutation test (1000 iterations) was used to confirm the significance of the observed variance structure.

### Redundancy Analysis (RDA)

Redundancy Analysis (RDA) was used to explore the multivariate relationships between the selected protein expression features and the experimental grouping variables (Genotype, Treatment, Behavior). RDA was chosen to provide insight into how much variance in the protein data can be explained by these known experimental factors.

The response matrix for the RDA model consisted of the top 10 proteins selected based on ANOVA results. The predictor matrix included the grouping variables Genotype, Treatment, and Behavior. The RDA model was fitted using the rda() function with the response matrix as the dependent variable and the grouping variables as independent predictors.

The overall significance of the RDA model was assessed using permutation tests. In addition to the overall model, the individual contributions of Genotype, Treatment, and Behavior were evaluated separately to determine the significance of each predictor. Axis-level significance was also tested to identify the most influential RDA components.

To visualize the results, comprehensive pairs plots were generated to explore pairwise relationships among the RDA site scores, grouped by Genotype, Treatment, and Behavior. A scree plot was used to illustrate the proportion of variance explained by each RDA axis. The top 5 proteins contributing to RDA1 and RDA2 were identified based on absolute loading magnitudes, and an RDA biplot was generated to capture the relationships between protein features, sample scores, and group centroids.

## **Results**

### Principal Component Analysis (PCA)

The first two principal components (PC1 and PC2) captured a substantial portion of the total variance, explaining ~47% and ~17% respectively, as shown in the scree plot (Figure 1) and the PCA summary. Together, these two components account for approximately 65% of the total variance, indicating that a significant portion of the data structure can be captured by just these two axes.

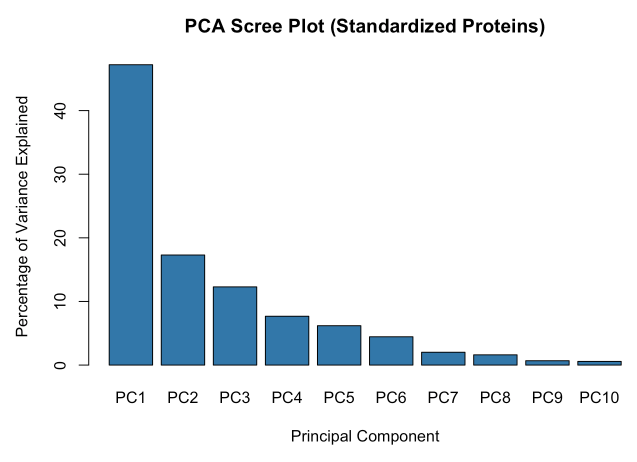


Figure 1. PCA Scree Plot

PC1 was primarily driven by *pERK\_N, DYRK1A\_N, CaNA\_N, BRAF\_N*, and *pGSK3B\_N,* which exhibited the highest absolute loadings. PC2 was influenced by *pMTOR\_N, BRAF\_N, DYRK1A\_N, P38\_N, and GFAP\_N*, suggesting these proteins contribute most to the second axis of variation.

Scatter plots were generated for PC1 and PC2 colored by Behavior (Figure 2). Scatter plots colored by Treatment, and Genotype are in the supplemental figure section. Statistical testing using ANOVA indicated that Genotype and Behavior significantly influenced PC1 (p < 0.001), while Treatment had significant effects on PC2 (p < 0.001) (Table 1).

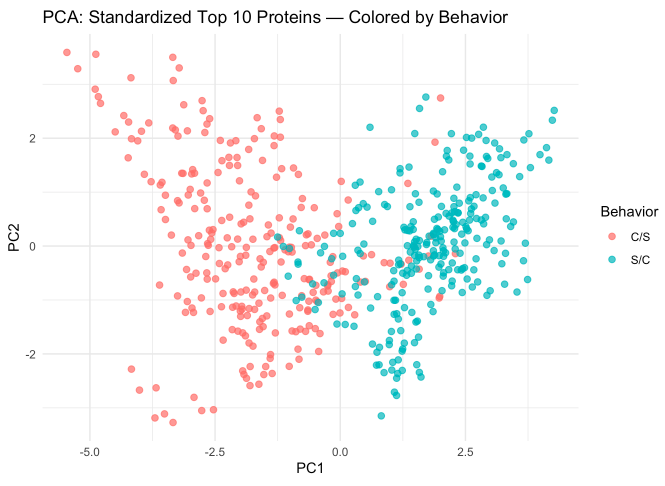


Figure 2. PCA Scatter Plot Colored by Behavior

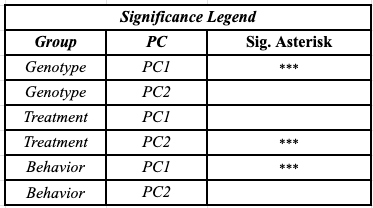


Table 1. PC Scatter Plot Significance Legend

Q-Q plots for PC1 and PC2 (Figure 3) confirmed approximately normal distributions. The slight deviation in the tails could be evidence of subpopulation effects.

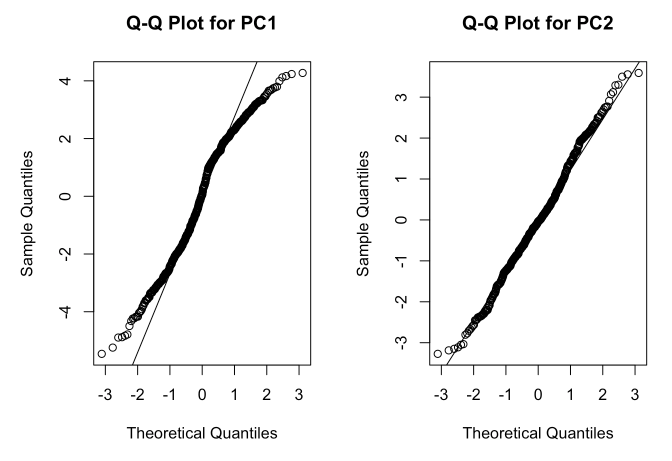


Figure 3. PC Q-Q Plots

A permutation test with 1000 iterations confirmed that the observed variance structure in PC1 was unlikely to have occurred by chance (p = 0.002). The rest of the PCs likely did not contribute significantly to the observed variance in the data.

### Redundancy Analysis (RDA)

The RDA model explained about half of the total variance in the protein expression data, with the constrained component capturing 53% of this variance and the remaining 47% attributed to residual variation according to the RDA model summary output. The first RDA axis (RDA1) accounted for 48.5% of the constrained variance, while the second axis (RDA2) explained 3.7%, as illustrated in the scree plot (Figure 4).

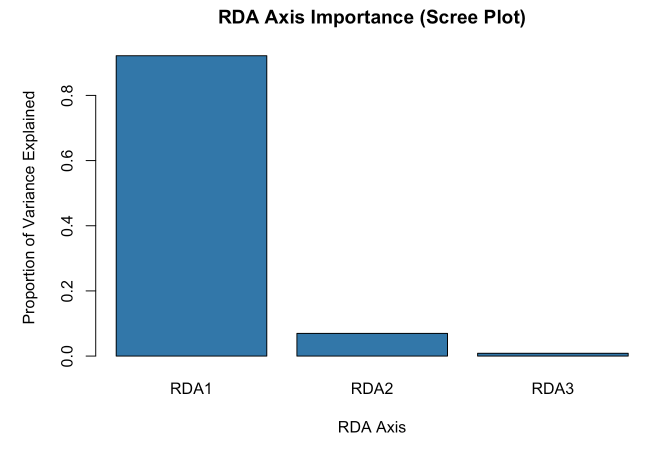


Figure 4. RDA Scree Plot

Permutation testing confirmed that the overall RDA model was highly significant (F₃,₅₃₃ = 197.62, p < 0.001), indicating that the grouping variables (Genotype, Treatment, Behavior) collectively explain a substantial portion of the observed variance. Individual term significance tests revealed that each grouping variable independently contributed significantly to the model (Genotype: F₁,₅₃₃ = 54.07, p < 0.001, Treatment: F₁,₅₃₃ = 39.86, p < 0.001, Behavior: F₁,₅₃₃ = 498.94, p < 0.001, further supporting the influence of these factors.

When examining the significance of individual axes, RDA1 (F₁,₅₃₃ = 546.51, p < 0.001), RDA2 (F₁,₅₃₃ = 41.27, p < 0.001), and RDA3 (p = 0.003) were all significant, confirming that the first three RDA axes capture meaningful structure in the data.

The top 5 proteins contributing to RDA1 and RDA2 included *CaNA\_N, pERK\_N, DYRK1A\_N, BRAF\_N,* and *pMTOR\_N*, which were identified based on the largest absolute loading magnitudes. These proteins likely play key roles in differentiating the experimental groups.

Variance inflation factor (VIF) analysis indicated minimal multicollinearity among the predictor variables, with all VIF values below 1.01, confirming that the model was not overly influenced by redundancy among the predictors.

BehaviorC/S and BehaviorS/C mice are separated along RDA1 suggesting overall protein expression differences between behavior groups. TreatmentMemantine and TreatmentSaline mice are separated along RDA2 suggesting more overall protein expression differences between treatment groups. *CaNA\_N* has the strongest influence along RDA1, pointing to the far left, indicating it is a key driver of the variation, likely distinguishing BehaviorC/S mice. *pERK\_N, DYRK1A\_N,* and *BRAF\_N* are clustered, suggesting they share a similar influence. Behavior groups are clearly separated along RDA1, primarily driven by *CaNA\_N, pERK\_N, DYRK1A\_N,* and *BRAF\_N* (negative loading). Overall, the biplot suggests that genotype, treatment, and behavior each contribute significantly to the observed variation in protein expression, but along different dimensions.

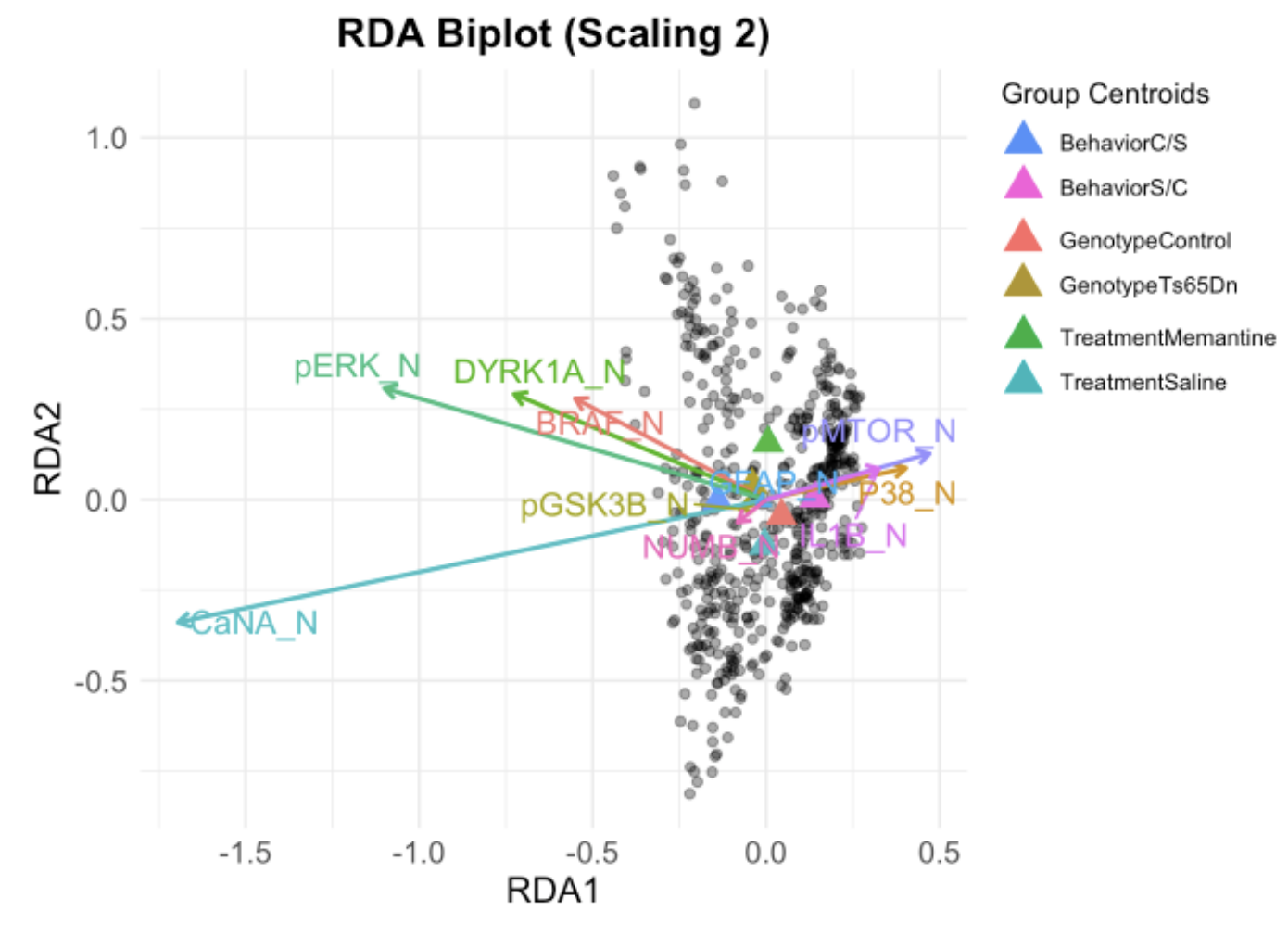


Figure 5. RDA Biplot

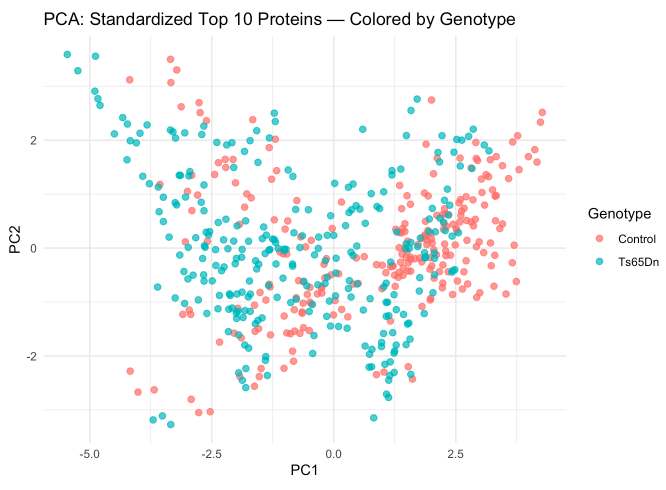
### **Literature Cited**

Higuera, Clara et al. “Self-Organizing Feature Maps Identify Proteins Critical to Learning in a Mouse Model of Down Syndrome.” PLoS ONE 10 (2015): n. pag.

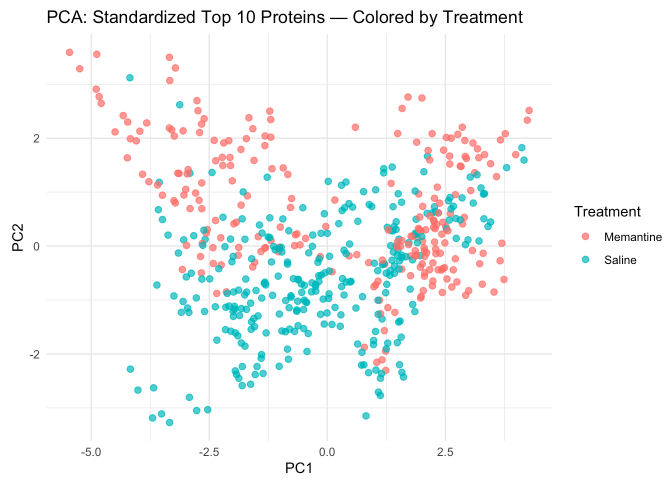
R Development Core Team. 2018. R: A Language and Environment for Statistical Computing. R

Foundation for Statistical Computing, Vienna, Austria.

### **Supplemental Figures**



Supplemental Figure 1. PCA Scatter Plot Colored by Genotype



Supplemental Figure 2. PCA Scatter Plot Colored by Treatment