Bioinformatics summer school report

Name: Balqees Abbas Yamani
Internship organization: Cambridge University
Mentor: Thomas Krueger
Project title: Differentially enriched neural biological processes between L1 and L4 in Caenorhabditis elegans
Codes: https://github.com/Balqeesya/My-Rcodes.git

Differentially enriched neural biological processes between L1 and L4 in Caenorhabditis elegans

introduction

Caenorhabditis elegans is a transparent nematode, ~1 mm in length, having a short 3-day life cycle and 2–3week lifespan at 20 °C. It contains four larval stages (L1–L4) prior to adulthood and can be easily cultured in the lab, making it a very good model organism.
Its nervous system is highly complex, consisting of 302 neurons and 56 glial cells (~37% of somatic cells) in 118 neuron types. Neurons are generated in three major waves: embryogenesis, late L1 stage, and L2 stage.
Because the entire neuronal lineage, morphology, and fluorescent markers are characterized, and because powerful genetic tools are available, C. elegans presents a unique model to examine the gene regulatory programs that build a nervous system.

gap

The C. elegans nervous system has been well studied at the cellular level, but protein expression and functional level research of neural proteins is scarce.

Project objectives

Understanding the stages of nervous system development from the perspective of the biological processes of neuronal proteins.

Methodology and approach

Data source

Protein level data and sample information were provided by the University of Cambridge from the Proteomic Characterization of Caenorhabditis elegans Larval Development research dataset.

Data import and preprocessing

To import, store, and manipulate my data, I used the Bioconductor package called Quantitative features for mass spectrometry(QFeatures), which the QFeatures object contains multiple SummarizedExperiment objects that work as data containers made up of three different parts.

Missing data were zeros. I've turned them to NA by (zeroIsNA), then proteins with excessive NA were filtered, no other filtering was required because the data was already at the protein level.

Data transformation and normalization
First, I plot the data distribution before applying Log2. The data was skewed toward zero, and that is normal because proteins have low abundance within the cell, but leaving it skewed with no Gaussian distribution will not allow us to be able to perform statistical tests later, so Log2 was applied to convert the data into a symmetrical, Gaussian distribution.

To reduce any technical variation between samples, we should normalize the data I chose median centering normalization as the original research did.

Exploratory Data Analysis
Exploratory Data Analysis was performed using principal component analysis (PCA) on normalized data and visualization of PCA plot and scree plot.

Differential expression analysis
To test proteins abundance between stages, I had the find the best model, which I used a linear modeling framework called limma to test pairwise contrast between the four larval stages and with no intercept, stabilize variance estimates using empirical Bayes, for significance testing I used Benjamini–Hochberg adjusted p-values (FDR < 0.01) like the original research method than visualized the significance by volcano plot .

Clustering and Biological Interpretation
A heatmap was performed to show the abundance,
To choose the right number of clusters, I performed hierarchical clustering using Ward.D2 on z score scaled abundance. Then, based on it I chose to cluster the proteins into four groups each representing a different expression pattern

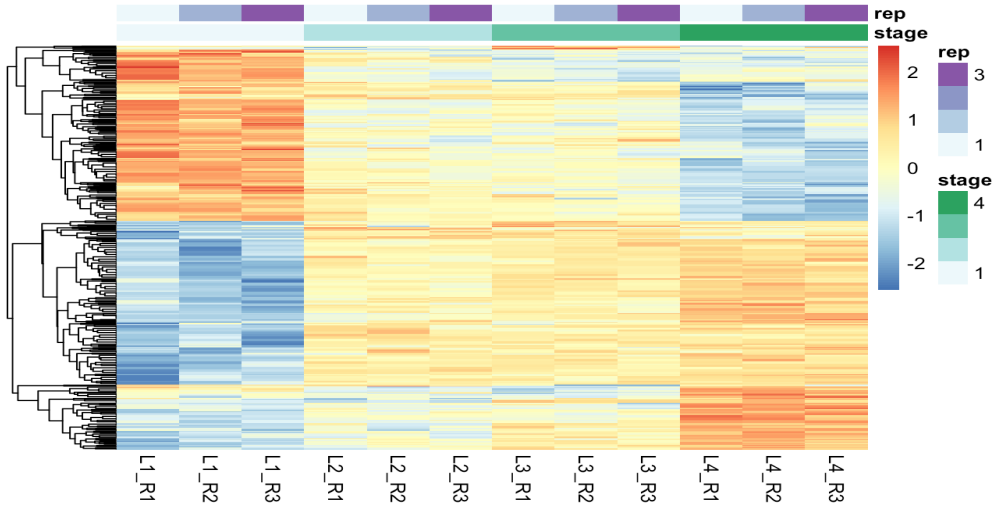
Proteins of higher significance (per contrast and per cluster) were subjected to a Gene Ontology (GO) Biological Process enrichment analysis using the clusterProfiler package and the org.Ce.eg.db annotation. EnrichGO and compareCluster functions were used to determine enriched biological processes per protein cluster. The visualization included the following: Dot plots for enriched terms ordered by the adjusted p-values; UpSet plots for the overlapping of enriched terms and faceted enrichment plots for the cluster specific processes.

Enrichment showed a specific biological process in each stage, and to understand the neuronal protein biological process in each stage, I extracted GO terms specific to the neural system. In order to monitor developmental regulation, proteins that were classified to these categories were displayed in scatter plots (z-scores L1 vs. L4).

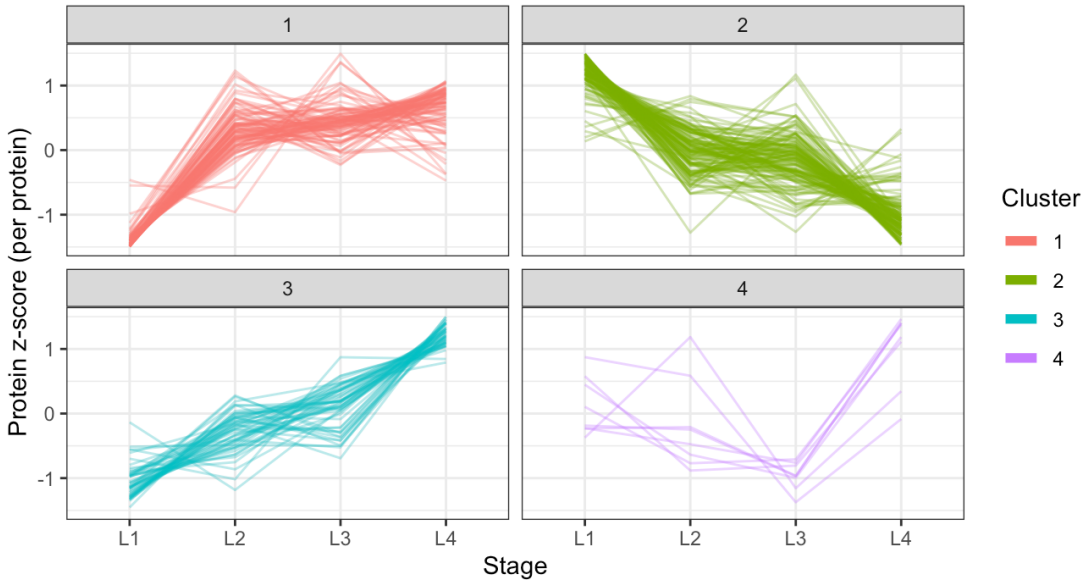
Then compare the stage specific neural biological process to what biological and cellular research says.

Results

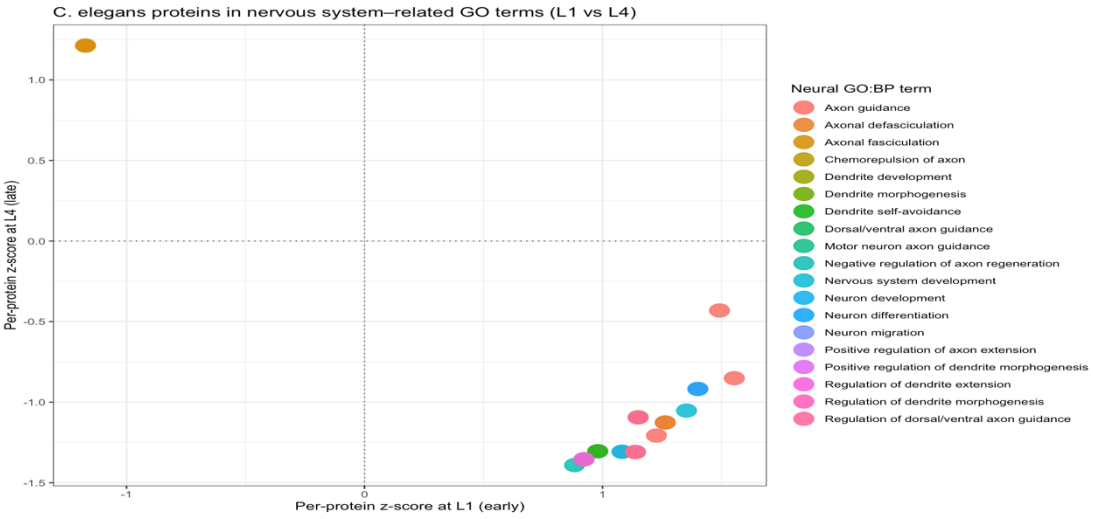
Our first figure and result show the protein abundance in the 4 larval stages and the largest difference between L1 and L4. This Heatmap of 1961 protein expressions across four *C. elegans* larval stages (three replicates each). Orange red = high expression; blue = low expression.



To get a deeper understanding of the pattern, I did the clustering, each line shows a protein's per-protein z score across stages L1–L4, and panels group proteins with similar temporal trajectories. Together, these patterns indicate early induced, early repressed, late increasing, and transiently suppressed programs, suggesting stage specific functions.



Scatter plot comparing GO biological processes linked to the nervous system in *C. elegans* at L1 and L4. Each point = a GO term from protein expression patterns. Nervous system processes dominate at L1.



The results of the gene ontology term related to the functions of neuronal proteins are consistent with research on the developmental stages of the nervous system in C. elegans(Sulston and Horvitz, 1977; Sulston et al., 1983), where most developmental stages occur in embryogenesis and post embryonic in the first and second stages of larval development.

Challenges and obstacles encountered

Lack of time as only 3 days were scheduled to conduct the analysis and the data was in the protein level and not the peptide level.

Next steps and future work
Kegg pathways analysis

References

- Xia T, Horton ER, Salcini AE, Pocock R, Cox TR, Erler JT. Proteomic Characterization of Caenorhabditis elegans Larval Development. Proteomics. 2018 Jan;18(2). doi: 10.1002/pmic.201700238. Epub 2017 Dec 27. PMID: 29178193.
- (2013) *Hermaphrodite - Nervous System - General overview*. Available at: https://www.wormatlas.org/hermaphrodite/nervous/Neuroframeset.html (Accessed: 21 August 2025).
- Sulston, J.E. and Horvitz, H.R. 1977. Post-embryonic cell lineages of the nematode *Caenorhabditis elegans*.
- Sulston, J.E., Schierenberg, E., White, J.G. and Thomson, J.N. 1983. The embryonic cell lineage of the nematode *Caenorhabditis elegans*.
- Hobert O. Neurogenesis in the nematode Caenorhabditis elegans. In: WormBook: The Online Review of C. elegans Biology [Internet]. Pasadena (CA): WormBook; 2005-2018. Available from: https://www.ncbi.nlm.nih.gov/books/NBK116086/

