Bioinformatics Summer School Report

Name: Marwan Mualla Aljohani
Internship Organization: Cambridge University
Supervisors/Mentors: Dr. Hugo Tavares, Dr. Thomas Krueger, Dr. Andries van Tonder
Project Title: Stage-Specific Proteomic Signatures and Expression Dynamics in ***Caenorhabditis elegans*** Development

---

1. Introduction

The nematode *Caenorhabditis elegans* is a fundamental model organism for studying metazoan development due to its well-defined genetics and invariant cell lineage. While transcriptomic studies are abundant, comprehensive proteomic analyses across its larval stages remain limited. Proteins, as the functional molecules of the cell, provide a direct view of biological processes. This project performs a bioinformatic re-analysis of a public quantitative proteomics dataset to identify protein expression dynamics across four larval stages (L1–L4), aiming to uncover stage-specific markers and long-term expression trends.

2. Objectives

1. Identify differentially expressed proteins across larval stages (L1–L4).
2. Define stage-specific protein markers and their functions.
3. Visualize global expression trends and gradual proteome changes.
4. Explore potential applications of protein signatures in human biology and disease.

3. Materials and Methodology

We re-analyzedapublisheddataset (Xia et al. 2017, PXD006676) comprising label-based quantitative mass spectrometry data from *C. elegans* larval stages L1–L4, with three biological replicates per stage. The analysis was conducted in R using a structured and reproducible workflow. Raw protein intensities were log2-transformed and normalized using median-centering (Fig. 1). Quality control was performed via PCA (Fig. 2) and distribution plots. Differential expression analysis was carried out using the limma package with stringent thresholds (FDR < 0.01, |log2FC| > 1). Functional

enrichment analysis (Gene Ontology Biological Processes) was performed using clusterProfiler. Temporal trend analysis identified proteins with monotonic expression patterns across development.

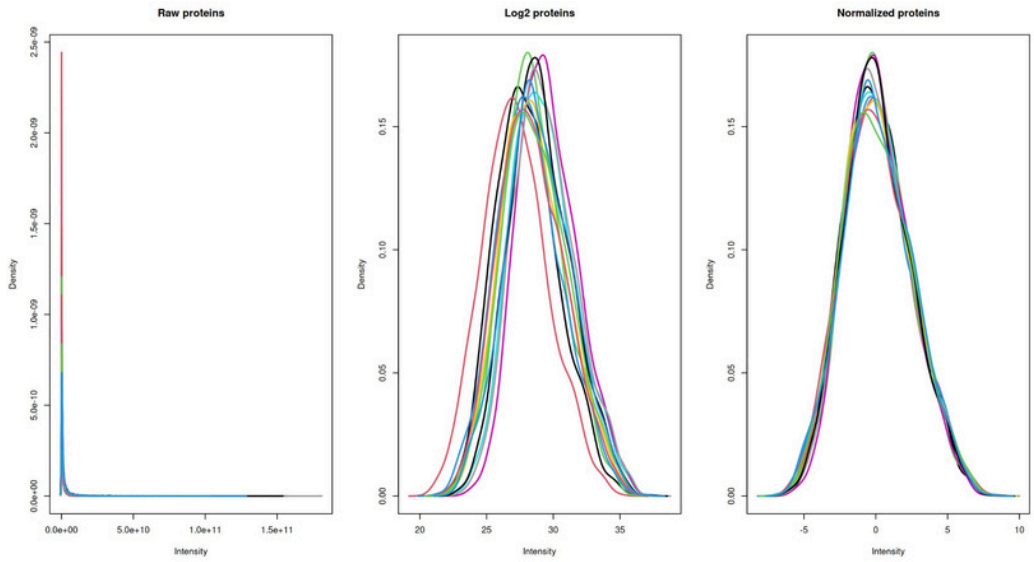


Fig. 1: Distribution of log2-transformed protein abundances across all replicates and stages before and after normalization.

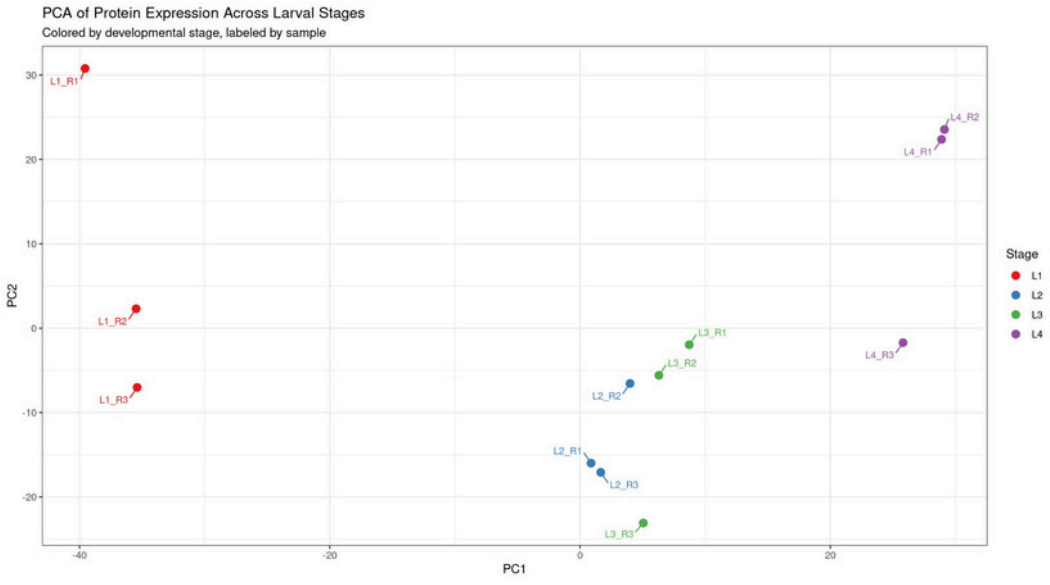


Fig. 2: Principal Component Analysis (PCA) plot of normalized protein data.

4. Results and Discussion

We identified 1,960 high-confidence proteins quantified across all stages. Our analysis revealed: The largest proteomic changes occur between the most distant stages (L4 vs L1), while the smallest changes occur between adjacent stages (L3 vs L2), indicating that changes accumulate over developmental time (Fig. 3). The L1 and L4 stages exhibit the highest number of stage-specific proteins, while L2 and L3 show very few, acting as transitional periods (Fig. 6). Functional enrichment showed L1 is highly enriched for metabolic and biosynthetic processes (Fig. 7), while L4 is enriched for reproductive and structural maturation functions (Fig. 8). We identified proteins with gradual increasing or decreasing expression trends across all stages, representing long-term biological programs (Fig. 4, Fig. 5).

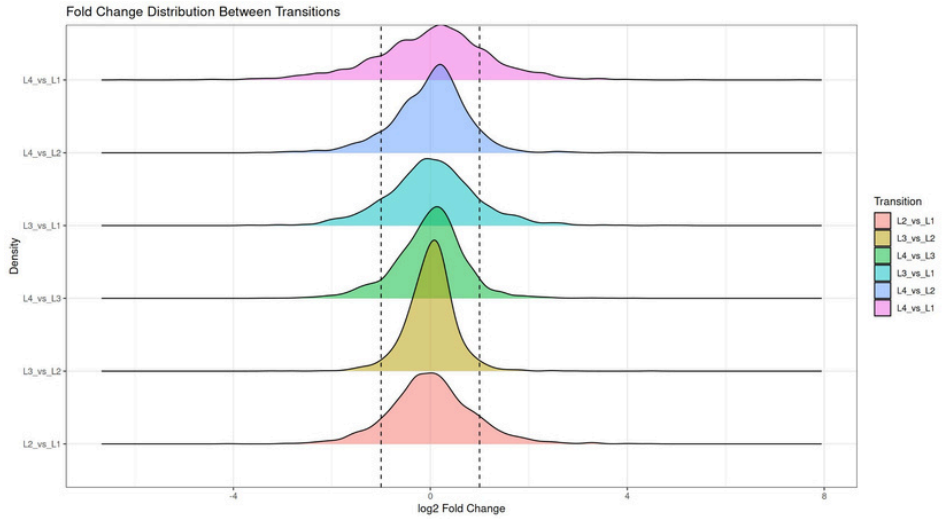


Fig. 3: Density plots of log2 fold changes (log2FC) for all stage comparisons.

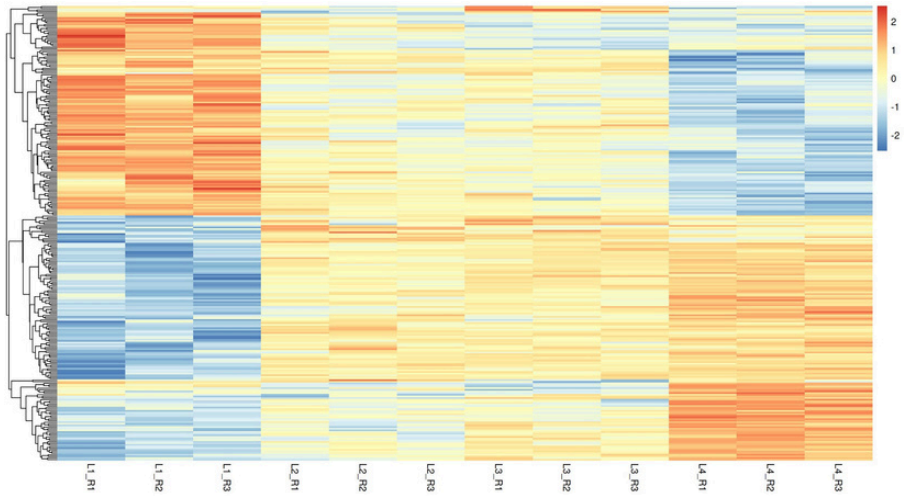


Fig. 4: Heatmap of proteins with strictly increasing or decreasing abundance from L1 to L4.

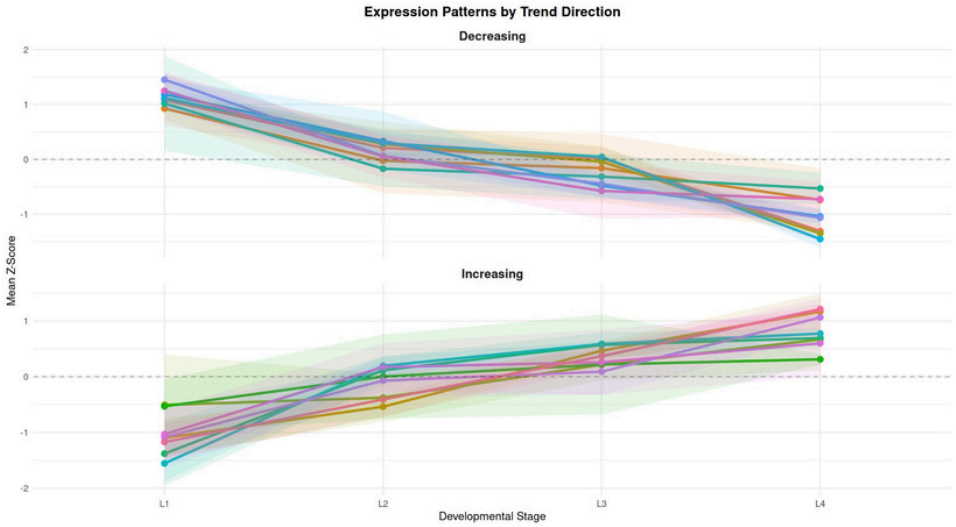


Fig. 5: Line plots of Z-scored expression for representative proteins with sustained trends.

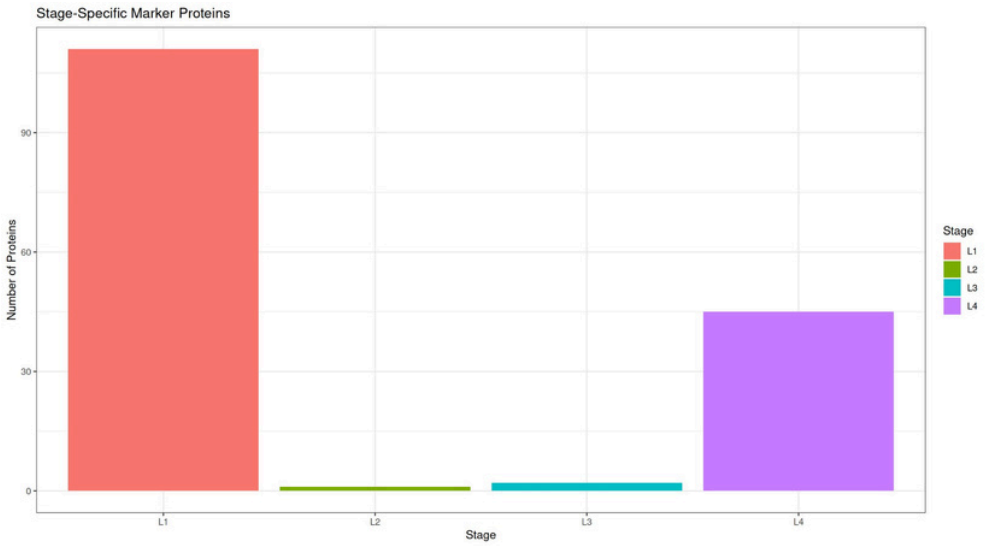


Fig. 6: Bar chart showing the count of stage-specific marker proteins.

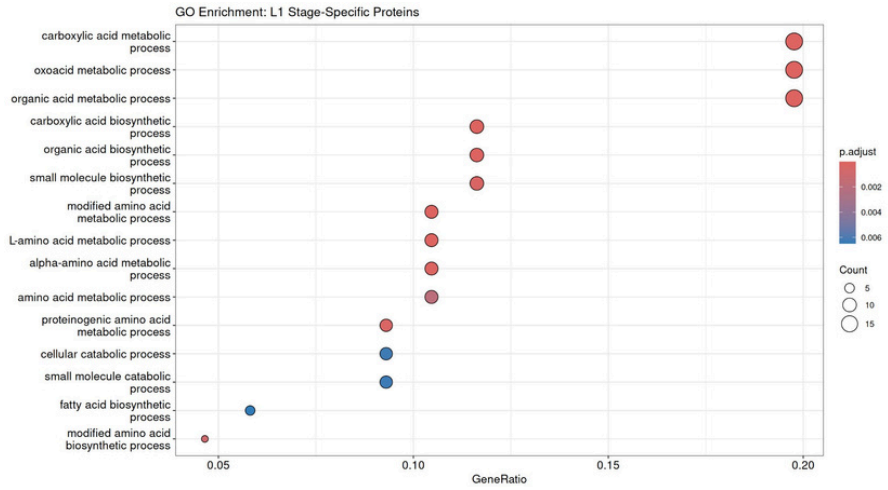


Fig. 7: Gene Ontology (GO) enrichment analysis for L1 stage-specificproteins.

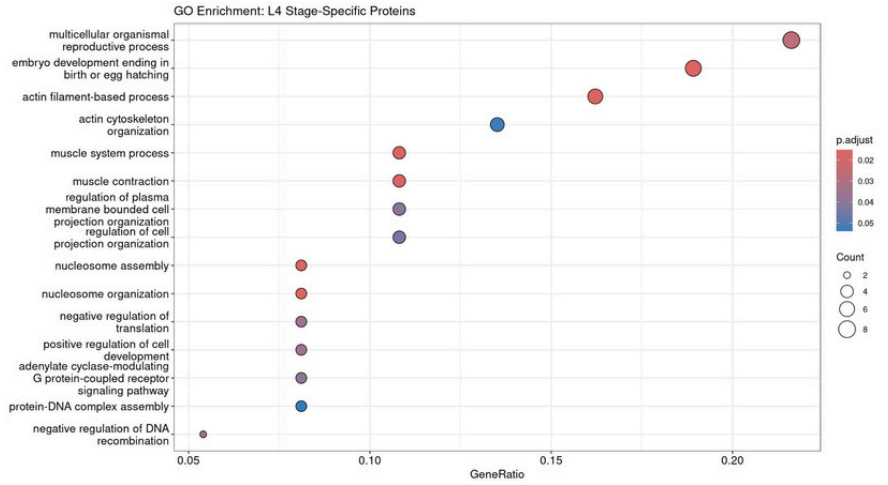


Fig. 8: GO enrichment analysis for L4 stage-specific proteins.

5. Challenges and Solutions

- Challenge: Managing a large dataset with missing values and ensuring only high-confidence proteins were analyzed.
- Solution: Implemented stringent filtering to include only proteins quantified in all replicates across all stages.
- Challenge: Choosing appropriate statistical thresholds for differential expression.
- Solution: Used a combination of fold change (|log2FC| > 1) and false discovery rate (FDR < 0.01) to ensure biologically and statistically significant results.
- Challenge: Biologically interpreting lists of significant proteins.

- Solution: Used Gene Ontology enrichment analysis to identify overrepresented biological functions in stage-specific protein sets.

6. Next Steps and Future Work

- Integrate with existing transcriptomic data for multi-omics perspective
- Publish reproducible code and results on GitHub for community use
- Extend analysis to stress-response or mutant backgrounds

7. Conclusion

This project provides a comprehensive, protein-level view of C. elegans larval development. We characterized distinct proteomic signatures for the L1 and L4 stages and revealed gradual expression trends that span the entire developmental timeline. The analysis pipeline developed is reproducible and can be applied to other temporal proteomics studies. These findings deepen our understanding of the molecular mechanisms that drive development and provide a resource for further functional studies.

8. References

Xia, T. et al. Proteomic characterization of C. elegans larval development.