Faculté des Sciences
Aix*Marseille Université

Aix-Marseille Université M1 - Développement Logiciel et Analyses et Données

Evaluation of methods and analysis of time series-omics data to identify regulators of muscular ageing in *Drosophila melanogaster*

Rahma AIT MAHFOUD

Dr Bianca Habermann team leader
Margaux Haering supervisor
“Computational Biology” team
Institut de Biologie du Développement de Marseille

IBDM INSTITUT DE BIOLOGIE DU DÉVELOPPEMENT DE MARSEILLE

cnrs

Aix*Marseille université

ACKNOWLEDGEMENTS

I am immensely grateful for the invaluable support and assistance provided by several individuals, without whom this Master's thesis would not have been possible.

I extend my sincere appreciation to Bianca HABERMANN, the leader of the computational biology team at the IBDM Institute, for her insightful advice and exceptional supervision.

Special thanks go to Margaux HAERING, my supervisor, for her attentive listening and understanding. Her contribution was vital to the completion of this work.

I would also like to express my deep gratitude to Théo BRUNET for his invaluable assistance whenever I required it, as well as to Maria and Stephen for their constructive feedback.

My appreciation extends to Dr Frank SCHNORRER for providing me with the data I worked on.

Furthermore, I would like to express my gratitude to Thomas and Lynn for their companionship throughout these two months, which significantly enhanced this experience.

Lastly, I extend my warmest thanks to Dr TERRAPON, M1 DLAD supervisor, for granting me the opportunity to benefit from a four-month internship grant.

INTERNSHIP ORGANIZATION

This internship report presents my experience at the Institut de Biologie du Développement de Marseille (IBDM), in the Computational Biology team led by Dr. Bianca HABERMANN.

The IBDM, which stands for the Marseille Institute of Developmental Biology, is located on the Luminy campus in Marseille, France, which is a renowned university site dedicated to scientific research and education. I had the privilege of being supervised by Margaux Haering, my tutor, who guided me throughout my internship. During this period, I had the opportunity to dive into the exciting world of computational biology and to contribute to their innovative research projects.

As part of my Software Development and Data Analysis training, I wanted to deepen my knowledge in the field of developmental biology and explore the application of computational methods to complex biological problems. This is why I joined the Computational Biology team at IBDM, under the guidance of Dr Bianca HABERMANN.

The Computational Biology team is known for its expertise in the analysis of genomic, transcriptomic and proteomic data, as well as for its use of advanced computational methods to decipher the molecular mechanisms underlying developmental processes. My supervisors, Bianca HABERMANN and Margaux HAERING, played a crucial role in guiding me throughout my internship, sharing their expertise and experience in data analysis and bioinformatics tools.

CONTENTS

ABSTRACT

This report presents the work carried out during a two-month internship on the study of muscle aging in *Drosophila melanogaster*. The main objective of this internship was to evaluate methods for analyzing omics data in time series in order to identify the regulators of muscle aging. In this context, an analysis pipeline was developed to recover, normalize and filter *D. melanogaster* data, followed by Mfuzz clustering to identify different gene expression profiles associated with muscle aging. Finally, a functional enrichment analysis was carried out using the FlyEnrichR tool to identify the biological processes and molecular pathways associated with genes displaying relevant profiles linked to muscle ageing in *Drosophila.* The results obtained at this stage of the internship will be studied in depth during the remaining two months of my internship to gain a better understanding of the regulatory mechanisms of muscle ageing in *Drosophila,* and open up new research prospects in this field.

RÉSUMÉ

Le présent rapport expose le travail réalisé lors d'un stage de deux mois portant sur l'étude du vieillissement musculaire chez *Drosophila melanogaster*. L'objectif principal de ce stage était d'évaluer des méthodes d'analyse de données omiques en séries temporelles afin d'identifier les régulateurs du vieillissement musculaire. Dans ce contexte, un pipeline d'analyse a été développé pour récupérer, normaliser et filtrer les données de *D. melanogaster*, puis un profilage par le clustering Mfuzz a été réalisé pour identifier les différents profils d'expression génique associés au vieillissement musculaire. Enfin, une analyse d'enrichissement fonctionnel a été réalisée à l'aide de l'outil FlyEnrichR pour identifier les processus biologiques et les voies moléculaires associés aux gènes qui présentent des profils pertinents liés au vieillissement musculaire de la *D.melanogaste*r. Les résultats obtenus à ce stade du stage vont être étudié profondément les deux restant de mon stage afin de mieux comprendre les mécanismes régulateurs du vieillissement musculaire et ouvrent de nouvelles perspectives de recherche dans ce domaine.

ABBREVIATIONS

BRB-seq : Bulk RNA barcoding and sequencing

DESeq : Differential Expression Sequencing

FA : Females in *ad-libitum* condition

FC : Control females

GO : Gene Ontology

MA : Active males

MC : Control males

RPM : Reads per million mapped reads

TMM : Trimmed Mean of M values

1. INTRODUCTION

In the context of ageing, the loss of muscle mass and function is one of the main characteristics that affect the quality of life of individuals (López-Otín et al. 2013). However, the mechanisms involved in this muscle deterioration with advancing age are not yet well understood. In order to better understand these processes, my internship aims to use time-series expression data from ageing muscles in *Drosophila melanogaster* to identify regulators of muscular ageing.

The team led by Dr. Frank SCHNORRER is dedicated to studying muscle dynamics and understanding the molecular mechanisms governing muscle development, ageing, and function (Lemke and Schnorrer 2017). In this context, the goal of my internship is to perform an in-depth analysis of data from the SCHNORRER team, using omic time series analysis techniques. We will specifically focus on identifying genes involved in muscle ageing in *Drosophila*, which is a particularly relevant study model for ageing research.

Drosophila shares genetic and physiological similarities with humans (Figure 1) (Lemke and Schnorrer 2017), making it a powerful model organism to study the underlying mechanisms of muscle ageing. By analyzing omics data, such as RNA sequencing and proteomics data, we will be able to gain valuable insights into gene expression profiles and molecular changes associated with muscle ageing.

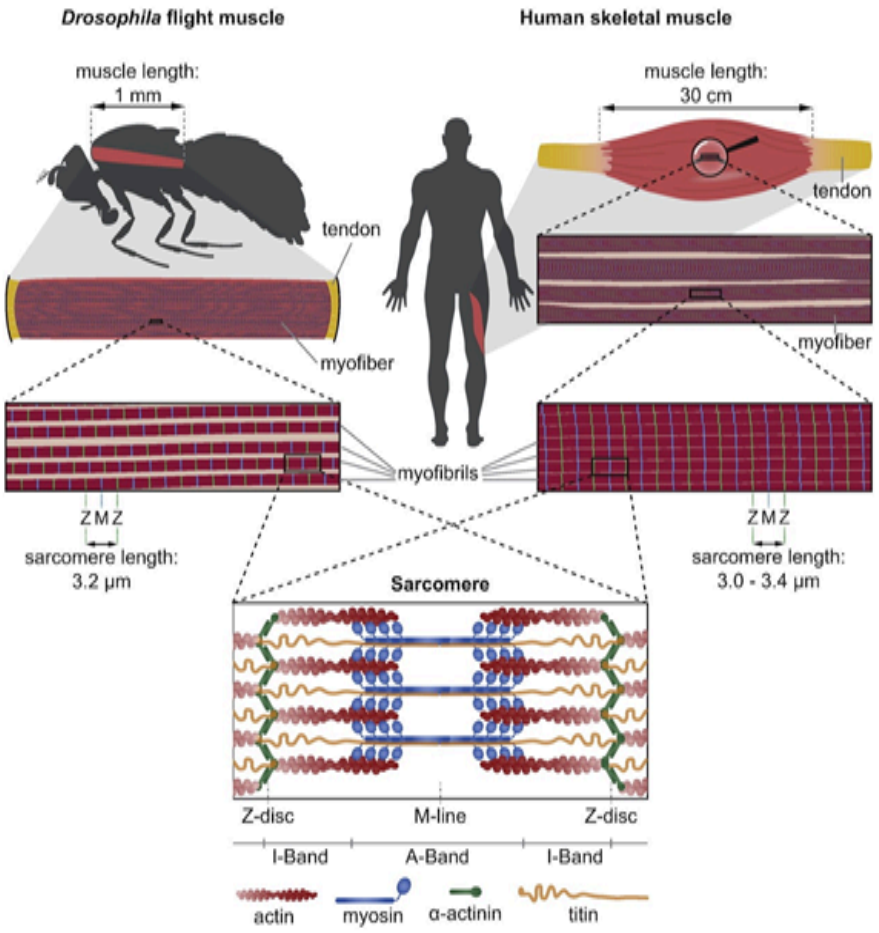


Figure.1 Similarity of muscle structures in fly versus human (Sandra B. Lemke, Frank Schnorrer, Mechanical forces during muscle development, (2017).

The short lifetime of *D. melanogaster* is a major advantage in the study of muscle ageing (Pletcher and Curtsinger 1998). While humans live on average several decades, fruit flies have a lifespan of only 2 to 3 months. This characteristic allows researchers to conduct large-scale longitudinal studies in a short period of time, accelerating the pace of discovery (Partridge and Gems 2002). They can observe muscle ageing processes and study the effects of various interventions in a relatively short period of time, facilitating experiments and rapid knowledge acquisition (Demontis and Perrimon 2010).

In addition, *D. melanogaster* is easy to manipulate genetically (Pandey and Nichols 2011). Thus, it provides an opportunity to study the effect of specific genes on muscle ageing, identify the signaling pathways involved, and understand the underlying molecular mechanisms (Demontis and Perrimon 2010). Techniques such as RNA interference (RNAi) and gene overexpression are commonly used to modulate the expression of genes related to muscle ageing (Liu et al. 2012). This allows us to determine the specific role of certain genes in muscle ageing and to explore the regulatory pathways involved (Anon n.d.-b).

Another significant advantage of *D. melanogaster* is its biological similarity to humans (Penney, Ralvenius, and Tsai 2020). Many fundamental biological processes are conserved between flies and humans despite the apparent differences between the two species (Demontis and Perrimon 2010).

Genes involved in ageing and muscle function in *D. melanogaster* often have orthologs in humans (Demontis and Perrimon 2010). Therefore, the study of *D. melanogaster* may provide valuable insights into the mechanisms underlying human muscle ageing (Demontis and Perrimon 2010).

Another reason why *D. melanogaster* has been used is the good experimental documentation of this model organism (Demontis and Perrimon 2010). Decades of research have developed an abundance of resources, protocols, and data on this species (Partridge and Gems 2002). Numerous tools and genetically engineered fly lines are also available, facilitating research and allowing scientists to focus on specific aspects of muscle ageing (Demontis and Perrimon 2010).

Experimental condition

In this study, two separate experimental conditions were set up to compare how *Drosophila* flight muscle ages. The first condition, considered the control condition, consisted of confining the fruit flies in a small box while providing them with all the necessary elements for their survival, including food. This control condition allowed us to observe how *Drosophila* age without the freedom to fly.

The second experimental condition involved the use of a larger, more complex flydome. This flydome provided the fruit flies with ample space to fly.

The objective of these two experimental conditions was to compare the effects of flight freedom on the ageing process of *Drosophila* flight muscle. Our collaborators isolated the flight muscle at regular intervals (each 5 days) and then performed BRB type Bulk RNA-sequencing (BRB-seq). These BRB-seq data were analyzed to investigate the gene expression profiles of *Drosophila* chronologically in these two conditions of no-flight and ad libitum flight conditions.

2. MATERIAL AND METHODS

The full script for the analysis pipeline used in this study is available on GitHub at :
https://github.com/Atmrahma/Internship_project/tree/main
on GitLab at:
https://gitlab.com/ibdml/internship/-/blob/origin/all_pipeline_script.rmd

2.1 Material

The R programming language version R 4.2.2 (2022-10-31 ucrt "Universal C Runtime") was mainly used during my internship. Several R packages were used, including "Package *DESeq2* version 1.38.3 "(Love, Huber, and Anders 2014) and "Package *edgeR* version 3.40.2"(Chen et al. n.d.) for normalization of sequencing data, "Package *gtools* version 3.9.4"(Bolker and al. 2022) for data array manipulation, "cluster" (Pollard and van der Laan 2005)and "Package *factoextra* version 1.0.7"(Lê, Josse, and Husson 2008) for choosing the number of clusters, "Package *Mfuzz* version 2.58.0"(Kumar and E Futschik 2007) for performing temporal clustering, "Package *org.Dm.eg.db* version 3.16.0" for converting gene names into usable identifiers, and "Package *enrichR* version 3.2"(Jawaid 2023) in combination with "*FlyenrichR*" for functional gene enrichment. For functional enrichment of clusters, the following databases were used : "Coexpression_Predicted_GO_Biological_Process_2018", "GO Cellular Component 2018", "GO Molecular Function 2018", "GO Biological Process 2018", "GO Biological Process GeneRIF", "KEGG 2019", "PPI Network Hubs from DroID 2017" and "WikiPathways 2018". The use of these packages and databases allowed for a comprehensive analysis of the data, from normalization to clustering analysis and functional enrichment.

2.1.1 Type of data used

The type of data I analyzed were BRBseq data. BRB-seq is a technique that provides high-throughput transcriptomics at a very affordable cost. This method is based on barcoding and sequencing of bulk RNA, which allows simultaneous analysis of thousands of cells in a single experiment (Alpern et al. 2019). This is achieved by only sequencing the 3' end of mRNAs, making it also accessible to low RNA amounts and high sample numbers.

It can be used to study gene expression in specific cell types or tissues. The BRB-seq method involves isolating RNA from a population of cells by the polyA tail. The RNA is then reverse transcribed into cDNA, which is fragmented and linked to a unique barcode sequence. These 3' barcoded cDNA fragments are then sequenced together(Alpern et al. 2019).

After sequencing, the reads are demultiplexed using the barcode sequences to assign each read to its cell of origin. This allows gene expression to be analyzed in thousands of individual cells simultaneously (Alpern et al. 2019).

The two main advantages of BRB-seq is first, its affordability and second, the low amount of RNA required for sequencing. Traditional RNA sequencing methods can be prohibitively expensive, but BRB-seq reduces the cost per sample by more than 100-fold, making it accessible to researchers with limited budgets (Alpern et al. 2019). Also, in our case, not 50, but 5 isolated *Drosophila* flight muscles were enough for sequencing. One limitation of this technique is the relatively shallow sequencing depth per sample, which may limit the sensitivity of the technique for low expressed genes.

2.1.2 Mfuzz clustering

Fuzzy clustering of time series expression data is a very useful technique to analyze temporal data. The Mfuzz package of R has been developed for flexible clustering of temporal gene expression data (Kumar and E Futschik 2007), based on a count matrix. The genes are grouped according to their expression patterns over time. The first step is to pre-process the gene expression data by normalizing it to make it compatible with the fuzzy algorithm. Next, it is essential to determine the number of clusters before proceeding with clustering.

From the gene expression data and the initial cluster centers, a membership matrix is calculated. This matrix assigns each gene a degree of membership to each cluster, enabling the gene's similarity within each cluster to be assessed.

As Mfuzz is a flexible clustering algorithm, it allows a gene to belong to several clusters. It is therefore recommended to repeat the clustering process several times to assess its robustness (Haering and Habermann 2021). This continues until the cluster centers converge and the membership matrix stops undergoing significant changes. Once the algorithm has converged, genes are grouped into clusters according to their similar temporal expression profiles. To facilitate interpretation and subsequent analysis, the results can be visualized graphically.

Choosing the appropriate number of clusters often requires in-depth analysis of overlaps between clusters (Haering and Habermann 2021), as well as enrichment analyses and comparisons between different numbers of clusters. This enables an informed decision to be made on the optimal number of clusters to use in the study.

2.1.3 Enrichment analysis by EnrichR-FlyEnrichR

To understand the biological processes and metabolic pathways significantly associated with a set of genes of interest, functional enrichment analysis is essential to interpret large-scale experimental results such as transcriptomic or proteomic studies.

Genes or proteins with similar biological functions tend to be similarly regulated and to interact in common metabolic pathways. Information on the biological functions, cellular processes, metabolic pathways or diseases associated with the set of genes studied is obtained

by identifying functional enrichment terms, which are previously annotated biological concepts (Chen et al. 2009).

EnrichR and FlyEnrichR (specifically designed for functional enrichment analysis in *D. melanogaster*) are versatile resources that can be used to exploit numerous functional enrichment databases (Chen et al. 2013), such as Gene Ontology, KEGG, Reactome, FlyBase and FlyMine. They take a list of genes as input, calculate over-represented terms by using an algorithm called hypergeometric distribution, and then generate graphs and summary tables to visualize the most enriched functional terms (Jawaid 2023).

Functional enrichment analysis contributes to a better understanding of the biological mechanisms underlying a set of genes, and helps generate new research hypotheses.

2.2 Methods

2.2.1 Analysis Pipeline

An R-based analysis pipeline (Figure 2) was used to examine the BRB-seq data from *D. melanogaster*, starting with data collection in the form of a table of read counts, representing gene expression levels. Various normalization methods were applied to mitigate undesirable variations between samples. The Mfuzz clustering algorithm was used to identify similar expression profiles, highlighting significant differences. Finally, the ErichR tool (FlyEnrichR) was used to identify biological terms, metabolic pathways and molecular functions enriched in gene clusters of interest.

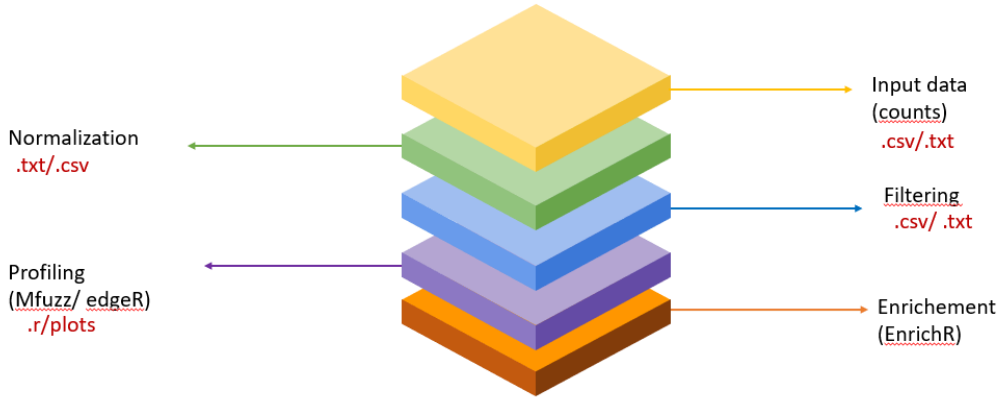


Figure.2 BRBseq analysis pipeline

The input data were CSV/TXT files that contained recorded reads from 17,874 genes, with each gene being represented in one row. The columns represented samples of *D. melanogaster* under four different conditions: active females (FA), active males (MA), control (inactive) females (FC), and control males (MC). Each sample consisted of five replicates.

The four conditions were studied chronologically, with a five-day interval from D05 to D50. In other words, we had a time series from five days to fifty days for each condition, with five replicates of each.

Firstly, the input files were normalized by TMM normalization, after having tried other normalization methods, such as Read per million (RPM), and the normalization procedure used DESeq which is a method that uses a statistical model based on a negative binomial distribution and incorporates normalization based on library size (Love et al. 2014). whereas TMM method assumes that most genes are not differentially expressed. It focuses on normalizing counts using scaling factors calculated from a stable reference gene set (Robinson and Oshlack 2010).

Next, the samples were filtered by applying a threshold criterion to eliminate rows corresponding to genes with a total number of reads below 10, 50, 100, 500 and 1000. The 1000 filter was chosen for the rest of the analysis in order to obtain more stringent results. The columns were then sorted chronologically for the samples using the 'mixedsort' function in the "gtools" package, as they were not initially ordered. Next, the condition table was divided into four separate tables (active females, active males, control females and control males), each containing five replicates per sample.

For each sample, the mean of the five replicates was calculated, resulting in a new table with only ten columns instead of fifty, each column representing a pseudo-sample (pseudo-replicate). These pseudo-samples were used for clustering using the Mfuzz package to identify groups of genes with similar expression profiles.

To perform the clustering, it was necessary to first select an appropriate number of clusters. To streamline this process and overcome its time-consuming nature, an automated approach was developed. A specific script was created to automate the selection of the optimum number of clusters, saving considerable time and effort.

Overlap between clusters was checked using Mfuzz's 'overlap.plot' function. Once the optimum number of clusters had been chosen, clustering was performed on each table containing the mean readings of the four normalized and filtered conditions.

Mfuzz clustering was run ten times in a loop to improve the robustness of the clustering results. Member lists from the ten Mfuzz clustering runs were then extracted to convert gene names to IDs using the "org.Dm.eg.db" package. The IDs could then be used for enrichment using "FlyEnrichr".

At the end, different numbers of clusters (15, 20, 25, 30, 35 and 40) were tested, and the results compared in terms of overlap between clusters, cluster quality and information from the functional enrichment analysis. A script was developed to automate these steps, making the process less time-consuming. By using this automated approach, the time

required to perform the analysis was considerably reduced. The workflow diagram below (Figure 3) provides a clear visualization of the steps involved in the process.

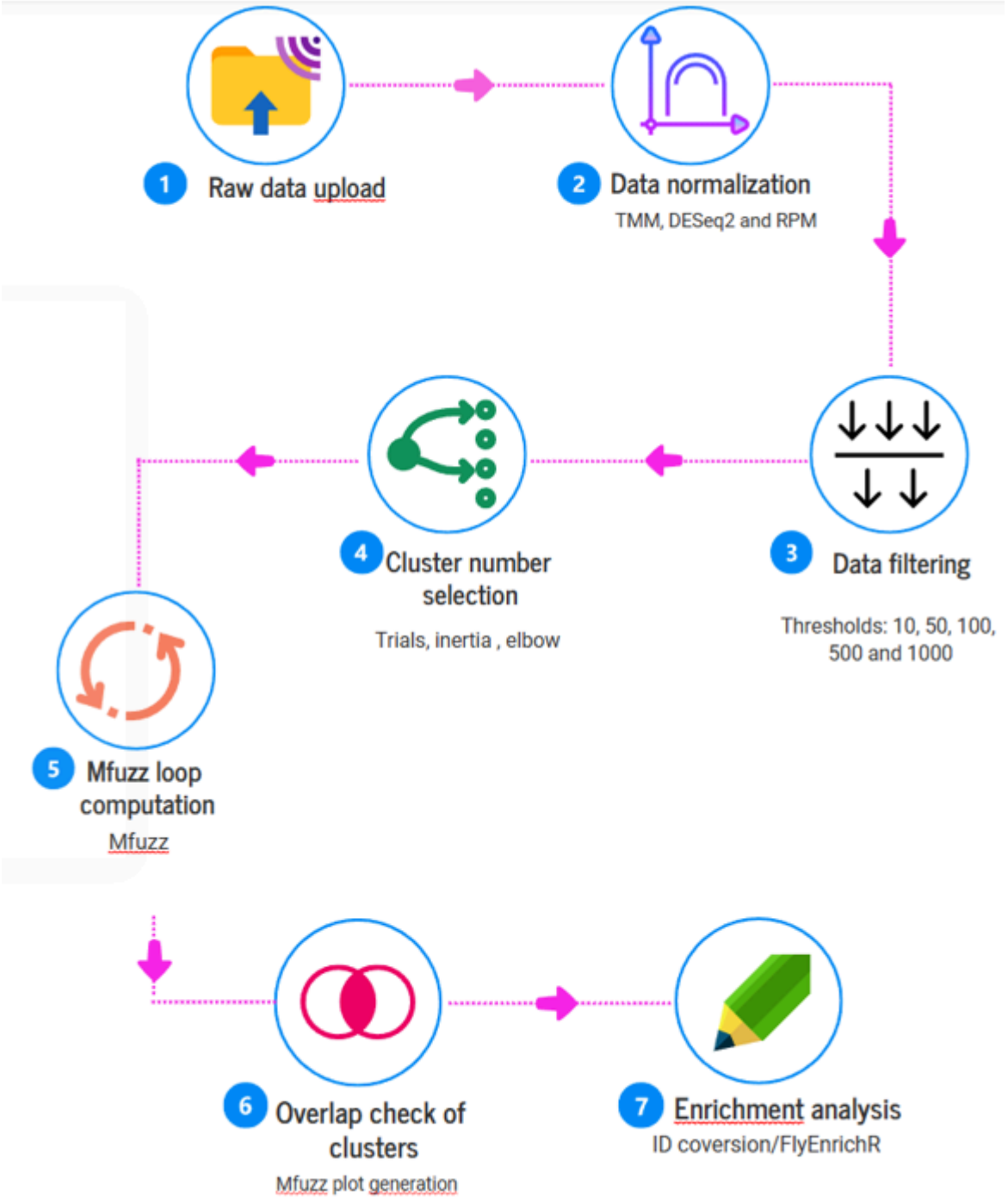


Figure.3 Automated workflow for analysis steps

3. RESULTS

3.1 Normalization

To visualize the effect of TMM normalization on the data, a boxplots for all genes was created before and after normalization.

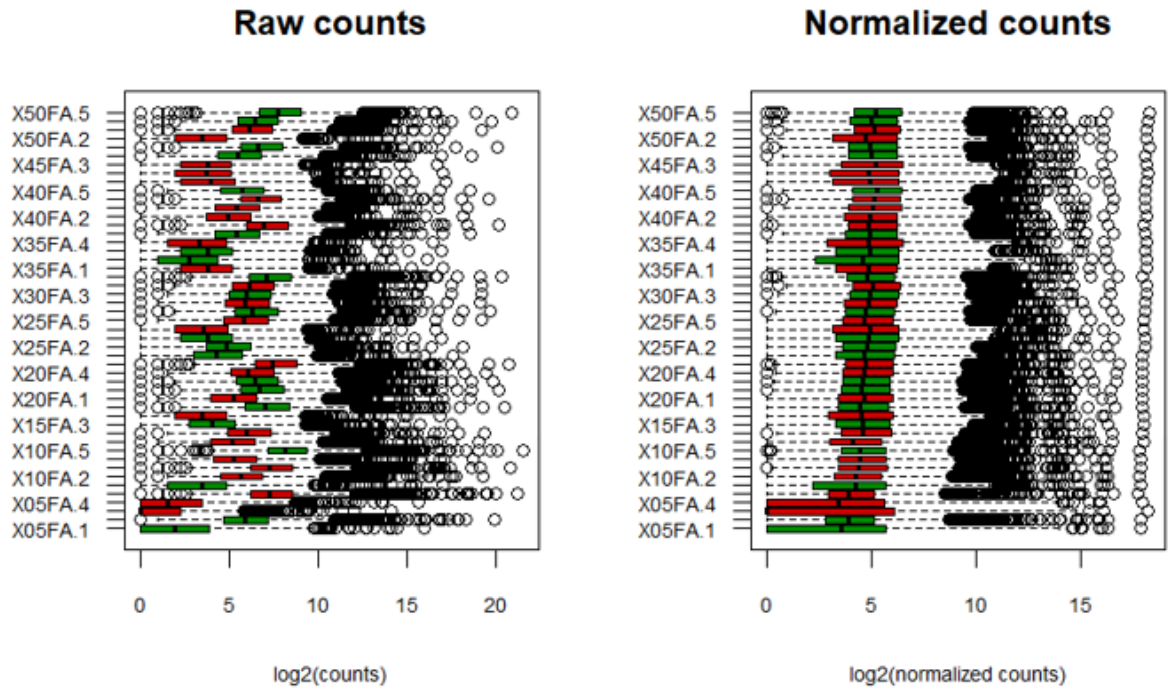


Figure.4 Comparison boxplots before and after data normalization; Distribution of samples before (left) and after (right) normalization. The x-axis represents log2(counts) and the y-axis represents samples. These boxplots were separated by color according to the experimental condition: green for the activity condition and red for the control condition.

3.2 Filtering by read counts

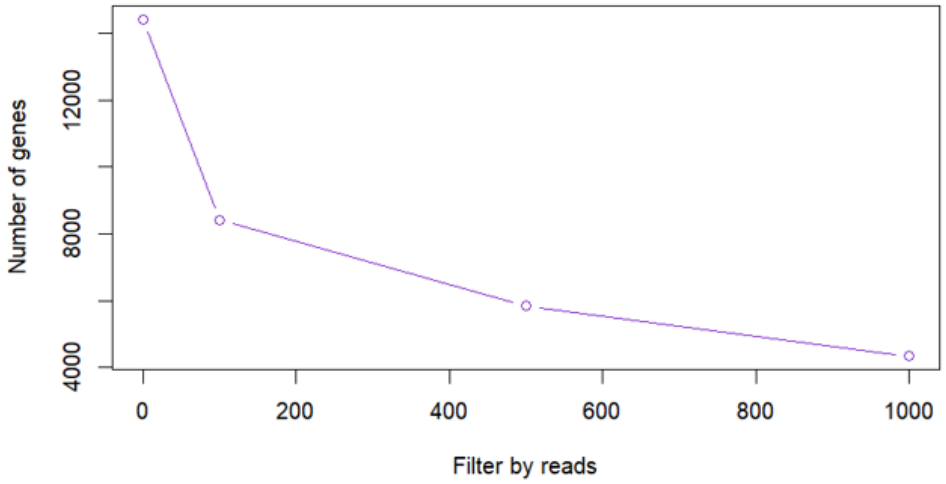


Figure.5 Variation in the number of read counts as a function of the filter parameter, increasing from 10 to 1000;

By increasing the filtering parameter from 10 to 1000, we notice that the number of reads decreases, reaching from 14.434 to 4469 for the genes which have more than 1000 reads. The number of genes expressed in muscle is generally much lower than the total number of genes present in the whole genome. According to Dr. Schnorrer, only ~2000-3000 genes are expressed in muscle. This finding indicates that, despite the reduction in the total number of genes expressed in muscle compared to the genome as a whole, we have not lost a significant amount of important information.

3.3 Clustering using Mfuzz

In this section, the results from the Mfuzz clustering and enrichment analysis are presented for the two conditions FA (active females) and FC (control females)

3.3.1 Cluster number selection

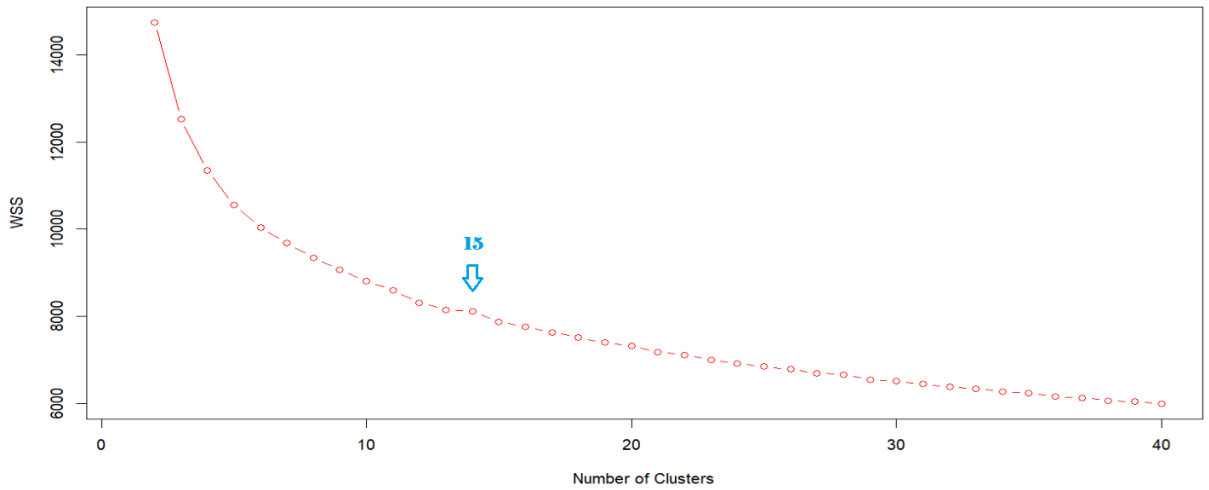


Figure.6 Elbow curve analysis; The pivotal point observed at k=15 in the elbow curve guided the selection of 15 clusters.

I first tried to estimate the cluster numbers using the Elbow method. Analysis of the kink of the curve played a key role in my decision to choose the number of 15 clusters. From this curve, I identified a key point at k=15 where the evolution of the kink becomes less significant. This indicates that adding further clusters beyond 15 would not contribute significantly to the overall structure of the data. Consequently, I took the decision to set the number of clusters at 15 for the remainder of my results.

3.3.2 Overlap test

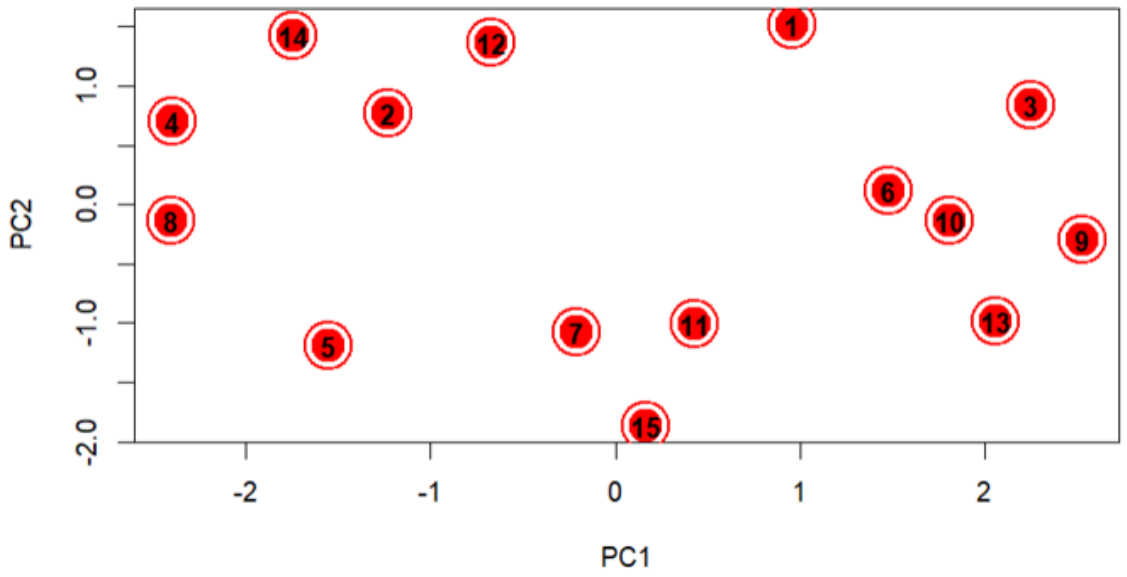


Figure.7 Overlap test results for the 15 clusters; The figure above represents the overlap of 15 clusters. Each number surrounded by a circle represents a distinct cluster.

No overlap was identified between clusters for k=15. That means that each cluster shares no profile with the others, underlining their distinctiveness.

The overlap test results for the different numbers of clusters are presented in the appendix (Figures 11, 12, 13 and 14) .

3.3.3 Clustering time-series

For the females in *ad libitum* condition (FA), Figure 8 shows the expression profiles of the 15 clusters reorganized according to expression profiles. Each graph represents a specific cluster.

Cluster results for all numbers of clusters tested are included in the appendix (Figures 15, 16 and 17).

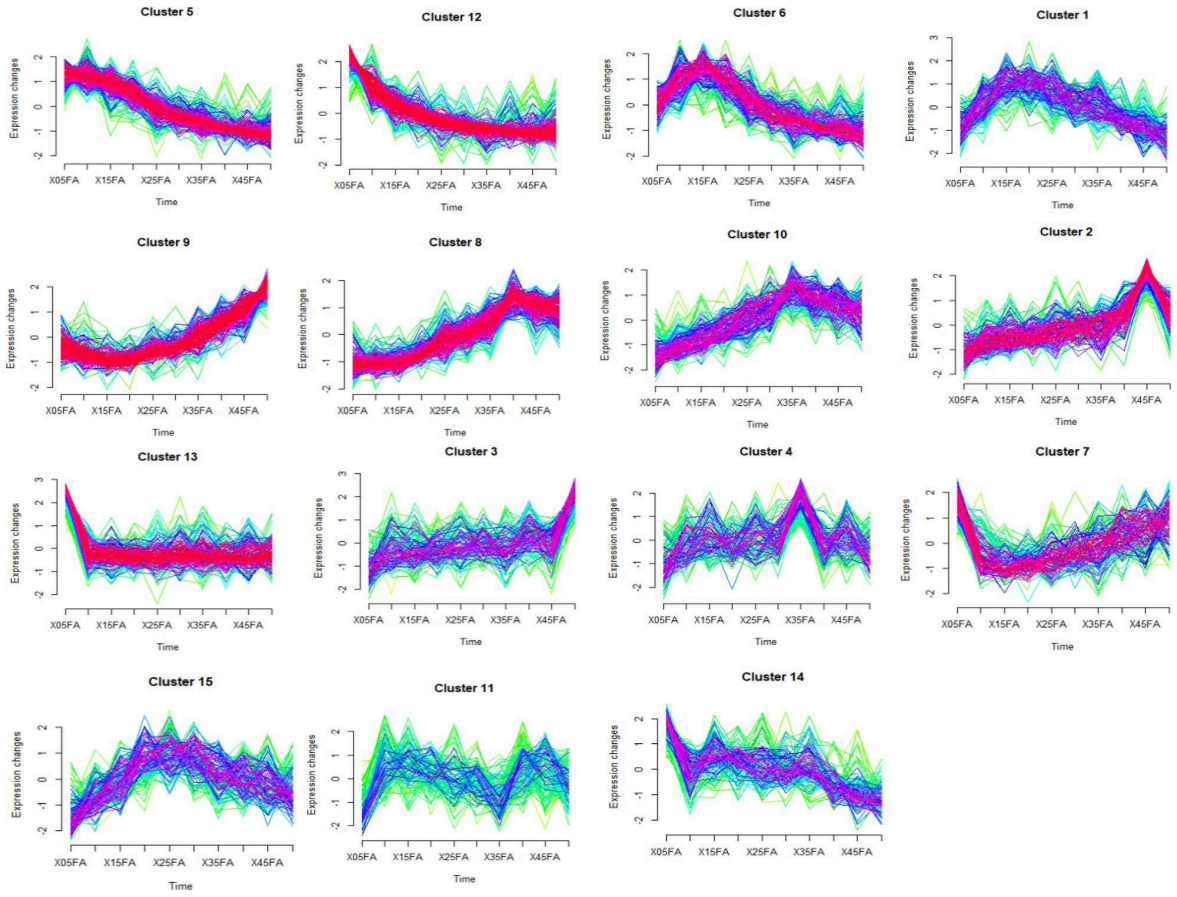


Figure.8 Exploring trends in gene expression levels across clusters in ageing females in *ad-libitum* flight conditions. The y-axis of each cluster represents the expression level, while the x-axis represents the time points (from D5 to D50). Each line of the graph represents the evolution of a gene's expression level as a function of time. The thick red line represents the core of the cluster, consisting of genes with high membership values, indicating that they closely fit the profile. Up-regulated gene clusters are reorganized on the same line of the figure, while down-regulated ones are represented on another line.

It is important to mention that there is noise visible in many of the clusters. For instance, single high peaks as seen in cluster 2 are most likely technical artifacts.

Similarly, for the flight-inhibited females, Figure 9 shows the expression profiles of the 15 clusters. Up-regulated gene clusters are grouped together on one line, while down-regulated ones are represented on another line.

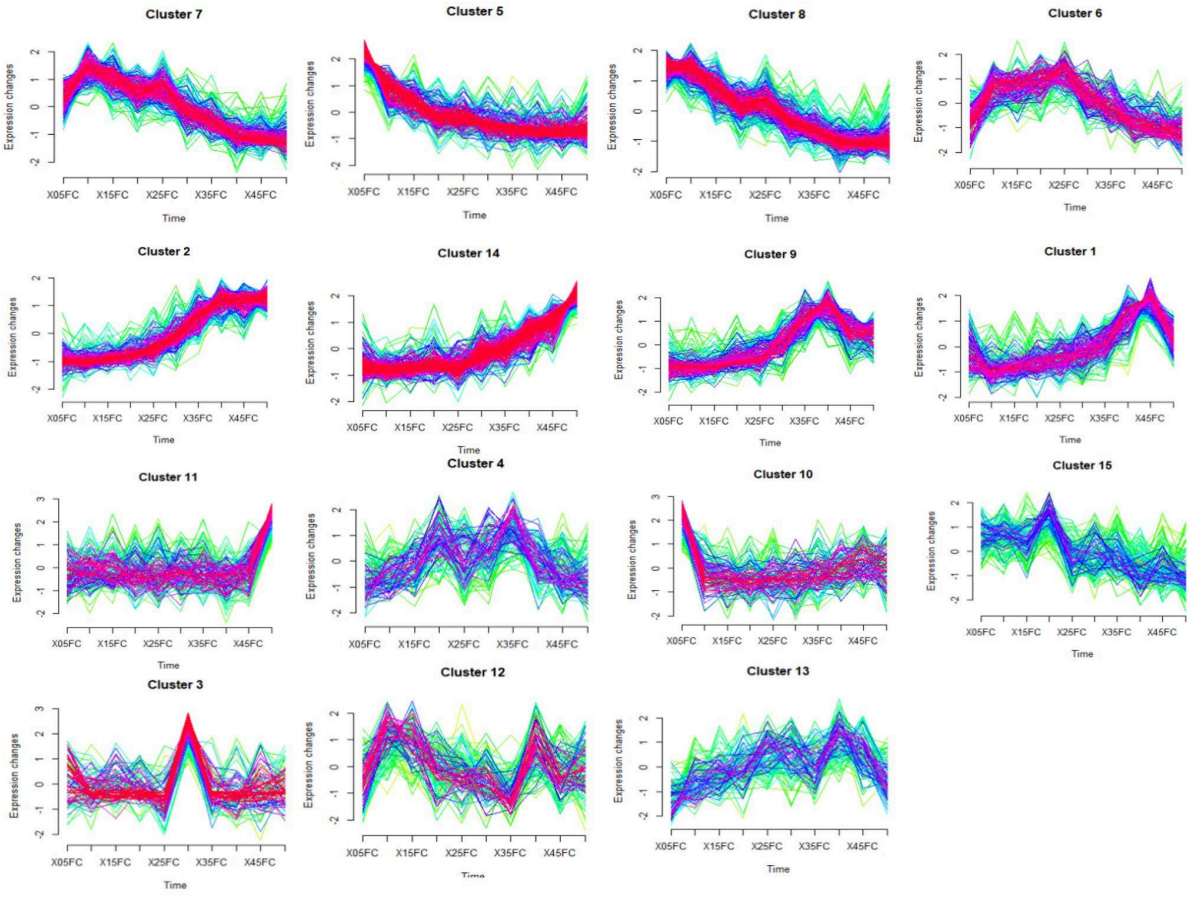


Figure.9 Exploring trends in gene expression levels across clusters (for the legend, please refer to Figure 8).

3.4 ID conversion/Enrichment

Before proceeding with enrichment analysis, it was essential to convert gene names into identifiers or symbols supported by the enrichment analysis package. For this, I used the converter "org.Dm.eg" package. Below is a table illustrating some examples of gene names before and after conversion:

Table 1. Examples of converting gene names to identifiers/symbols

Gene name (Before conversion)	**Gene ID/Symbol** (After conversion)
FBgn0000228	Bsg25D
FBgn0003134	Pp1alpha-96A
FBgn0003495	spz
FBgn0004240	DptA

FBgn0004629	Cys
FBgn0010225	Gel

This conversion was crucial to ensure a correct match between the genes and the databases used for enrichment.

Functional enrichment grouped by p-value was performed to identify the biological processes and molecular functions associated with the clusters with the most relevant profiles. The results of functional enrichment revealed important information about the specific biological pathways and functions regulated in these clusters.

In this section we will only interpret the first four down-regulated clusters for *ad libitum* condition females.

Taking the down-regulated cluster 5 as an example, the figure below shows the enrichment result obtained for this cluster,

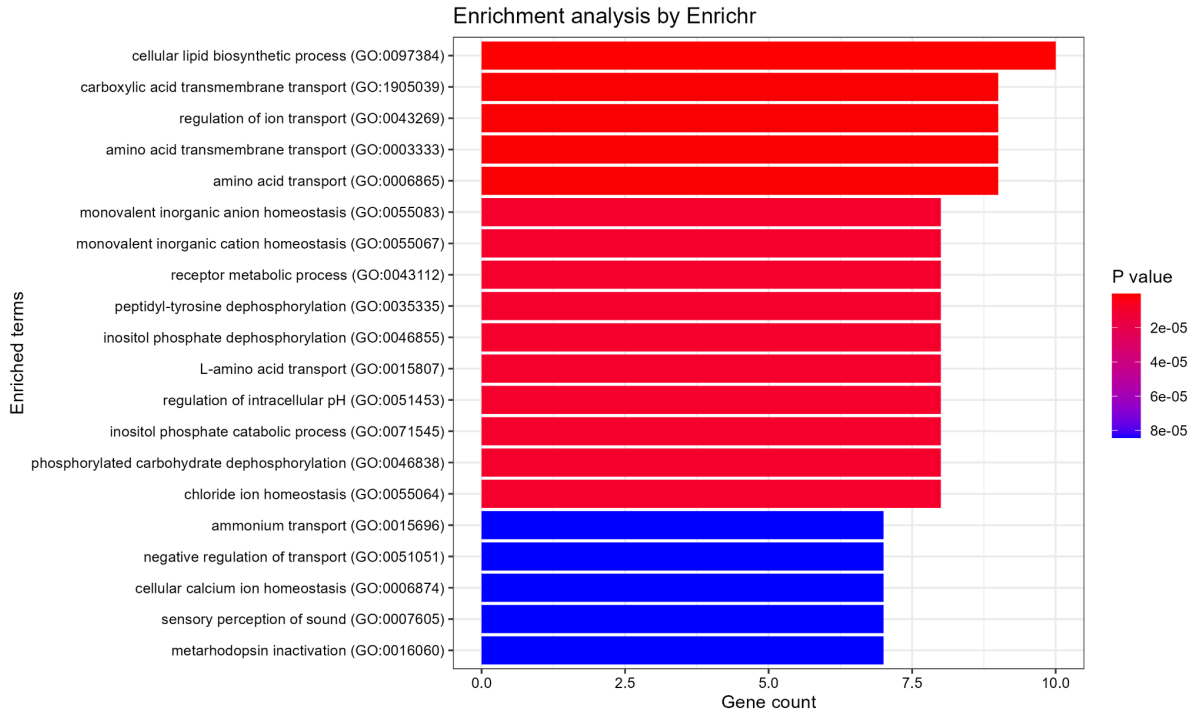


Figure.10 Functional enrichment of the down-regulated cluster number 5 in the FA condition. Terms are ranked according to their p-value and are represented by two colors (red for potentially enriched terms and blue for non-significant terms). This cluster is characterized by down-regulation of processes linked to cellular lipid biosynthesis, transmembrane transport of carboxylic acids, regulation of ionic transport, amino acid transport and receptor metabolism. It also regulates intracellular pH balance.

These results suggest that muscle aging may be associated with altered lipid biosynthesis, amino acid and ion transport, as well as disturbances in intracellular pH balance.

Cluster 1 (Appendix-Figure 19) shows down-regulation of processes related to larval or pupal morphogenesis, post-embryonic animal organ morphogenesis, protein localization in synapses, regulation of cell morphogenesis involved in differentiation, regulation of neuronal death, actin filament bundle assembly, actin filament polymerization and growth regulation. These results fit with the observation that there is a general trend of down-regulation of genes at an early stage in the time-series, where organ morphogenesis is clearly finished and the genes invovled in the enriched processes are no longer needed.

The cluster 12 (Appendix-Figure 20) shows down-regulation of several metabolic processes, including phospholipid and purine nucleobase catabolism, as well as xanthine metabolism. There is also down-regulation of processes linked to the regulation of cell death and actin filament bundle assembly. In addition, there is a reduction in intracellular pH and sodium-independent transport of organic anions. We suggest that in muscle aging, there may be altered lipid and nucleobase metabolism, dysfunction in cell death and actin remodeling processes, as well as disruption of intracellular pH balance and anion transport.

The cluster 6 (Appendix-Figure 21) highlights down-regulation of processes related to proteasome assembly, the ERAD (endoplasmic reticulum-associated degradation) pathway, regulation of ATP oxidation, targeting of post-translational proteins to the Golgi apparatus membrane, as well as regulation of assembly of the RNA polymerase II transcription preinitiation complex. These results suggest an alteration in protein degradation, protein complex assembly and transcription regulation.

DISCUSSION

In my internship, I analysed time-series data of Drosophila flight muscle ageing using a fuzzy clustering algorithm and tried to interpret the resulting data biologically.

I was confronted with the analysis of BRBseq type data. This type of data is known for its complexity and its difficulty to be processed, analyzed and interpreted due to the noise present in the results (Alpern et al. 2019). In this part, we will discuss the challenges encountered when analyzing *D. melanogaster* data to study muscle ageing, as well as the different methods used to overcome these problems and improve the robustness of the results to be able to obtain relevant biological information.

BRBseq data is often subject to noise (Figure 22), making it difficult to distinguish between real signals and artifacts. These noises can be due to various factors, such as technical variations, experimental errors or biological variations.

Given this specificity, several tests were performed to fit the data and minimize unwanted noise. This step was crucial to ensure the quality of the results obtained and to guarantee their reliability. For this, I tested different normalization and filtering methods, such as TMM, RPM normalization and the normalization by the DESeq2 package, as well as thresholding filtering, starting with 10 going up to 1000. In order to reduce noise while preserving biologically relevant signals.

By choosing the filter of 1000, it was noted that the number of genes decreased, reaching from 14,434 to 4469 genes 1000. Despite the significant drop, the number of genes expressed in muscle is generally much lower than the total number of genes present in the whole genome. The reduction in the total number of genes expressed in the muscle compared to the genome as a whole does not affect the quality of biological information.

Another particularity of our approach was the use of Mfuzz clustering. This clustering method was chosen for its ability to process complex biological data. However, to ensure the robustness of our results, we ran the clustering 9 times randomly. This strategy allowed us to better understand the inherent variability in the data and to produce more reliable results (Haering and Habermann 2021).

Taking into account the commonalities obtained from functional enrichment analysis of the four down-regulated profiles 1,5, 6, and 12 for females in *ad libitum* condition, we find alterations in several metabolic processes and dysfunction in some processes.

The results suggest that muscle aging is associated with metabolic alterations, dysfunction in cell death and actin remodeling processes, as well as disturbances in ion balance, amino acid transport, protein degradation and transcription regulation.

These alterations could contribute to reduced muscle function, loss of muscle mass and other manifestations of muscle ageing.

CONCLUSION

Promising results have been obtained by an analysis pipeline in R developed during the first two months of my internship. In-depth interpretation of the results and deduction of an answer to the initial biological question will be the next stage of my internship.

The following two months will be devoted to implementing a functional enrichment extension based on this pipeline in the RNAFuzzy application, using the FlyEnrichR tool. This extension will provide the user with advanced functional analysis of *D. melanogaster* data and a better understanding of the underlying biological mechanisms.

REFERENCES

1. Alpern, Daniel, Vincent Gardeux, Julie Russeil, Bastien Mangeat, Antonio C. A. Meireles-Filho, Romane Breysse, David Hacker, and Bart Deplancke. 2019. 'BRB-Seq: Ultra-Affordable High-Throughput Transcriptomics Enabled by Bulk RNA Barcoding and Sequencing'. *Genome Biology* 20(1):71. doi: 10.1186/s13059-019-1671-x.
2. Bolker [aut, Ben, cre, Gregory R. Warnes, and Thomas Lumley. 2022. 'Gtools: Various R Programming Tools'.
3. Chen, Edward Y., Christopher M. Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R. Clark, and Avi Ma'ayan. 2013. 'Enrichr: Interactive and Collaborative HTML5 Gene List Enrichment Analysis Tool'. *BMC Bioinformatics* 14(1):128. doi: 10.1186/1471-2105-14-128.
4. Chen, Jing, Eric E. Bardes, Bruce J. Aronow, and Anil G. Jegga. 2009. 'ToppGene Suite for Gene List Enrichment Analysis and Candidate Gene Prioritization'. *Nucleic Acids Research* 37(Web Server issue):W305-311. doi: 10.1093/nar/gkp427.
5. Chen, Yunshun, Davis McCarthy, Pedro Baldoni, Mark Robinson, and Gordon Smyth. n.d. 'EdgeR: Differential Analysis of Sequence Read Count Data User's Guide'.
6. Demontis, Fabio, and Norbert Perrimon. 2010. 'FOXO/4E-BP Signaling in Drosophila Muscles Regulates Organism-Wide Proteostasis during Aging'. *Cell* 143(5):813–25. doi: 10.1016/j.cell.2010.10.007.
7. Haering, Margaux, and Bianca H. Habermann. 2021. 'RNfuzzyApp: An R Shiny RNA-Seq Data Analysis App for Visualisation, Differential Expression Analysis, Time-Series Clustering and Enrichment Analysis'.
8. Jawaid, Wajid. 2023. 'EnrichR: Provides an R Interface to "Enrichr"'.
9. Kumar, Lokesh, and Matthias E Futschik. 2007. 'Mfuzz: A Software Package for Soft Clustering of Microarray Data'. *Bioinformation* 2(1):5–7. doi: 10.6026/97320630002005.
10. Lê, Sébastien, Julie Josse, and François Husson. 2008. '**FactoMineR**: An *R* Package for Multivariate Analysis'. *Journal of Statistical Software* 25(1). doi: 10.18637/jss.v025.i01.
11. Lemke, Sandra B., and Frank Schnorrer. 2017. 'Mechanical Forces during Muscle Development'. *Mechanisms of Development* 144:92–101. doi: 10.1016/j.mod.2016.11.003.
12. Liu, Nan, Michael Landreh, Kajia Cao, Masashi Abe, Gert-Jan Hendriks, Jason R.

Kennerdell, Yongqing Zhu, Li-San Wang, and Nancy M. Bonini. 2012. ‘The MicroRNA MiR-34 Modulates Ageing and Neurodegeneration in Drosophila’. *Nature* 482(7386):519–23. doi: 10.1038/nature10810.

13. López-Otín, Carlos, Maria A. Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. 2013. ‘The Hallmarks of Aging’. *Cell* 153(6):1194–1217. doi: 10.1016/j.cell.2013.05.039.
14. Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. ‘Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2’. *Genome Biology* 15(12):550. doi: 10.1186/s13059-014-0550-8.
15. Pandey, Udai Bhan, and Charles D. Nichols. 2011. ‘Human Disease Models in Drosophila Melanogaster and the Role of the Fly in Therapeutic Drug Discovery’. *Pharmacological Reviews* 63(2):411–36. doi: 10.1124/pr.110.003293.
16. Partridge, Linda, and David Gems. 2002. ‘Mechanisms of Aging: Public or Private?’ *Nature Reviews Genetics* 3(3):165–75. doi: 10.1038/nrg753.
17. Penney, Jay, William T. Ralvenius, and Li-Huei Tsai. 2020. ‘Modeling Alzheimer’s Disease with IPSC-Derived Brain Cells’. *Molecular Psychiatry* 25(1):148–67. doi: 10.1038/s41380-019-0468-3.
18. Pletcher, Scott D., and James W. Curtsinger. 1998. ‘Mortality Plateaus and the Evolution of Senescence: Why Are Old-Age Mortality Rates so Low?’ *Evolution* 52(2):454–64. doi: 10.2307/2411081.
19. Pollard, K. S., and M. J. van der Laan. 2005. ‘Cluster Analysis of Genomic Data’. Pp. 209–28 in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, *for Biology and Health*, edited by R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, and S. Dudoit. New York, NY: Springer.
20. Robinson, Mark D., and Alicia Oshlack. 2010. ‘A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data’. *Genome Biology* 11(3):R25. doi: 10.1186/gb-2010-11-3-r25.

APPENDIX

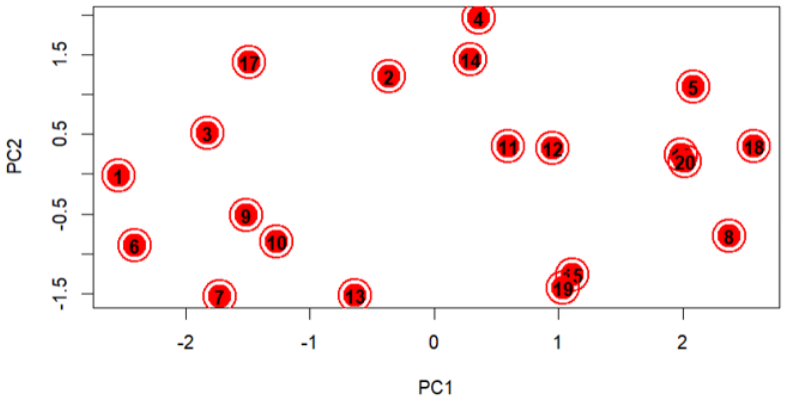


Figure. 11 Overlap test results for the 20 clusters

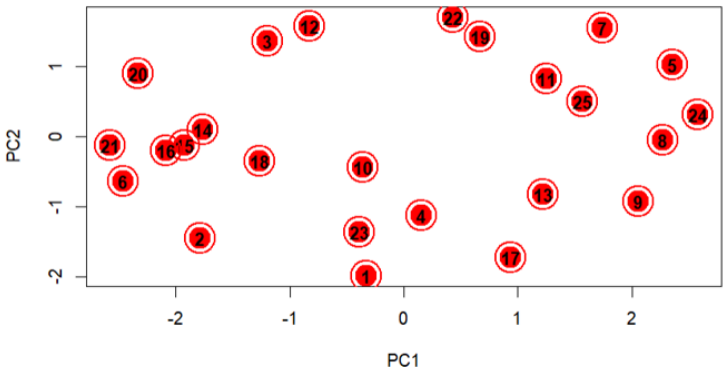


Figure.12 Overlap test results for the 25 clusters

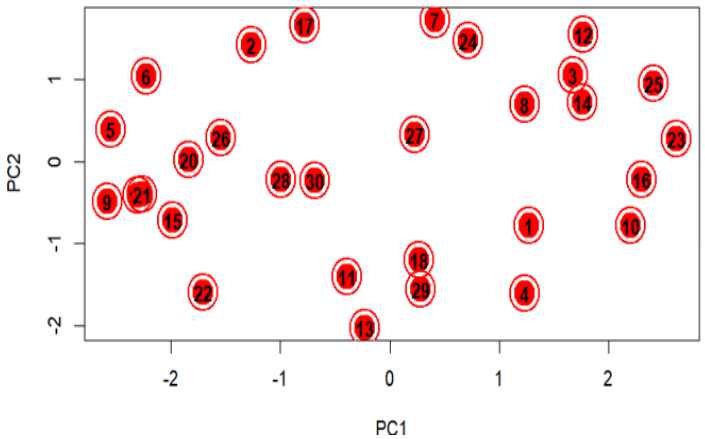


Figure.13 Overlap test results for the 3O clusters

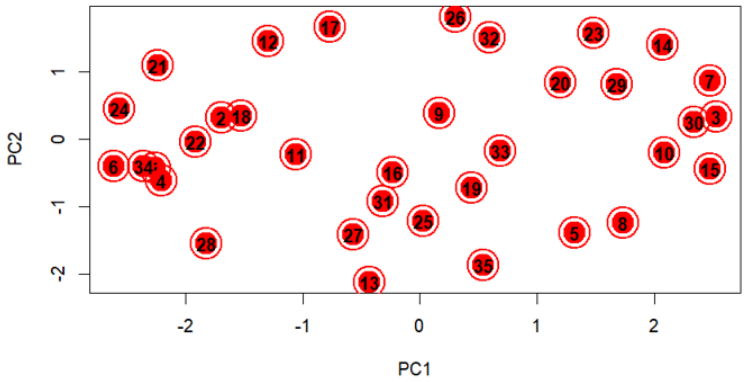


Figure.14 Overlap test results for the 35 clusters

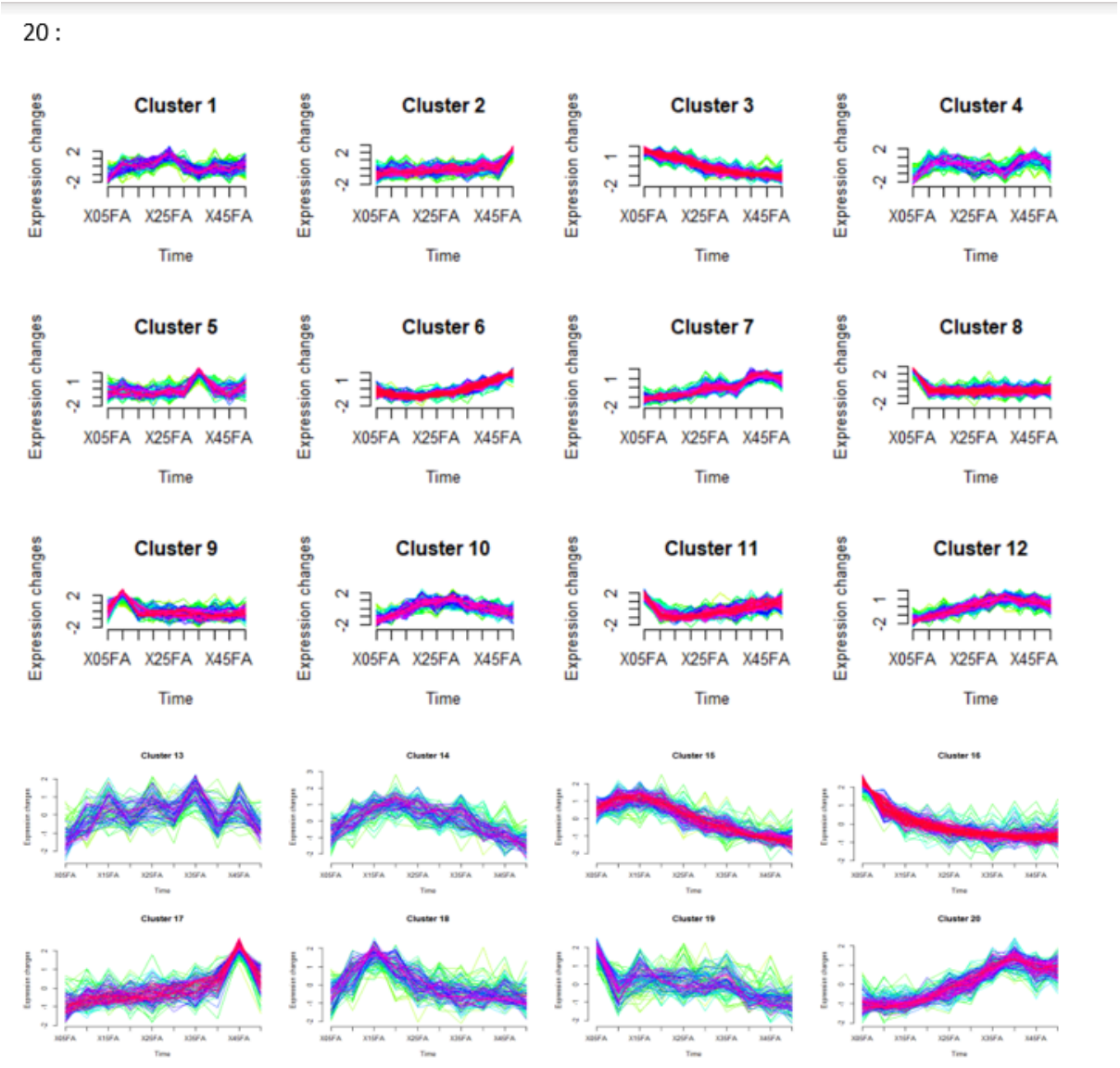
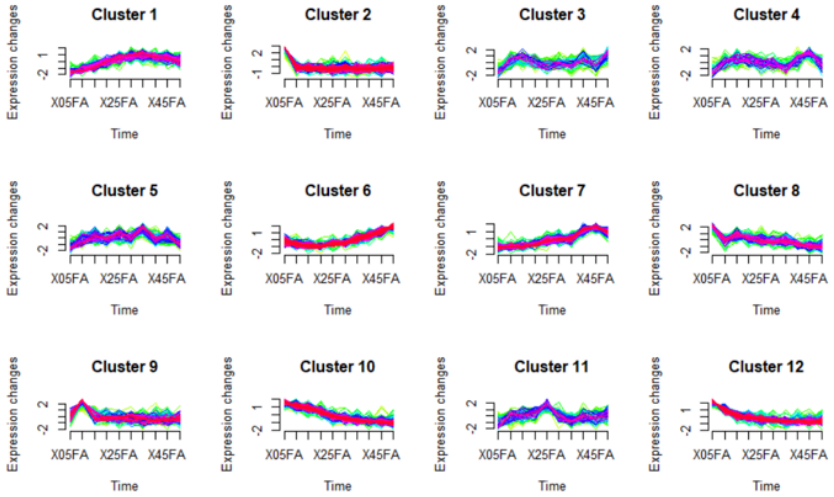
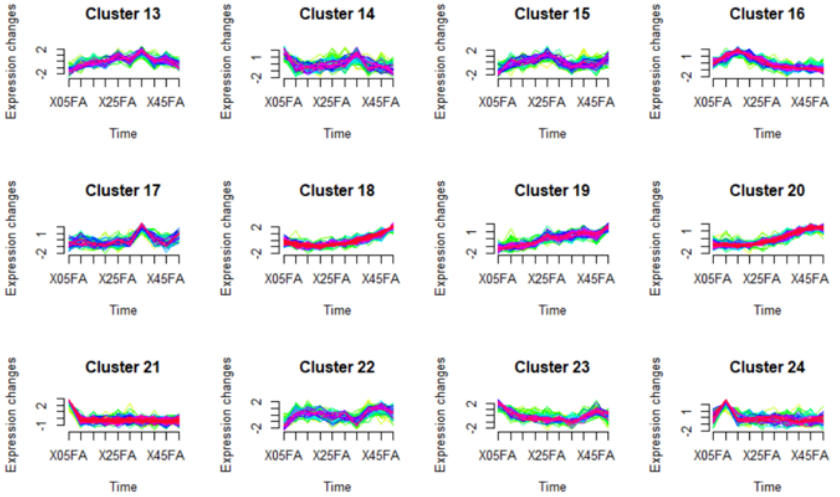


Figure.15 Exploring trends in gene expression levels across 20 clusters (Cf- Figure 8).

a



b



c

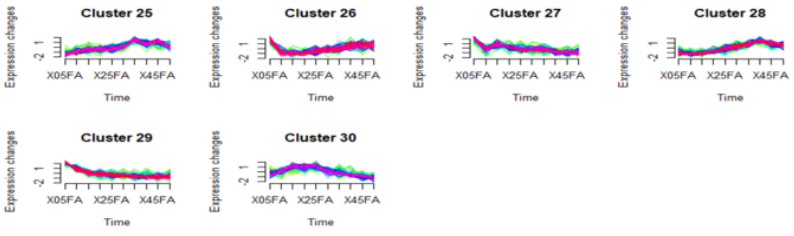
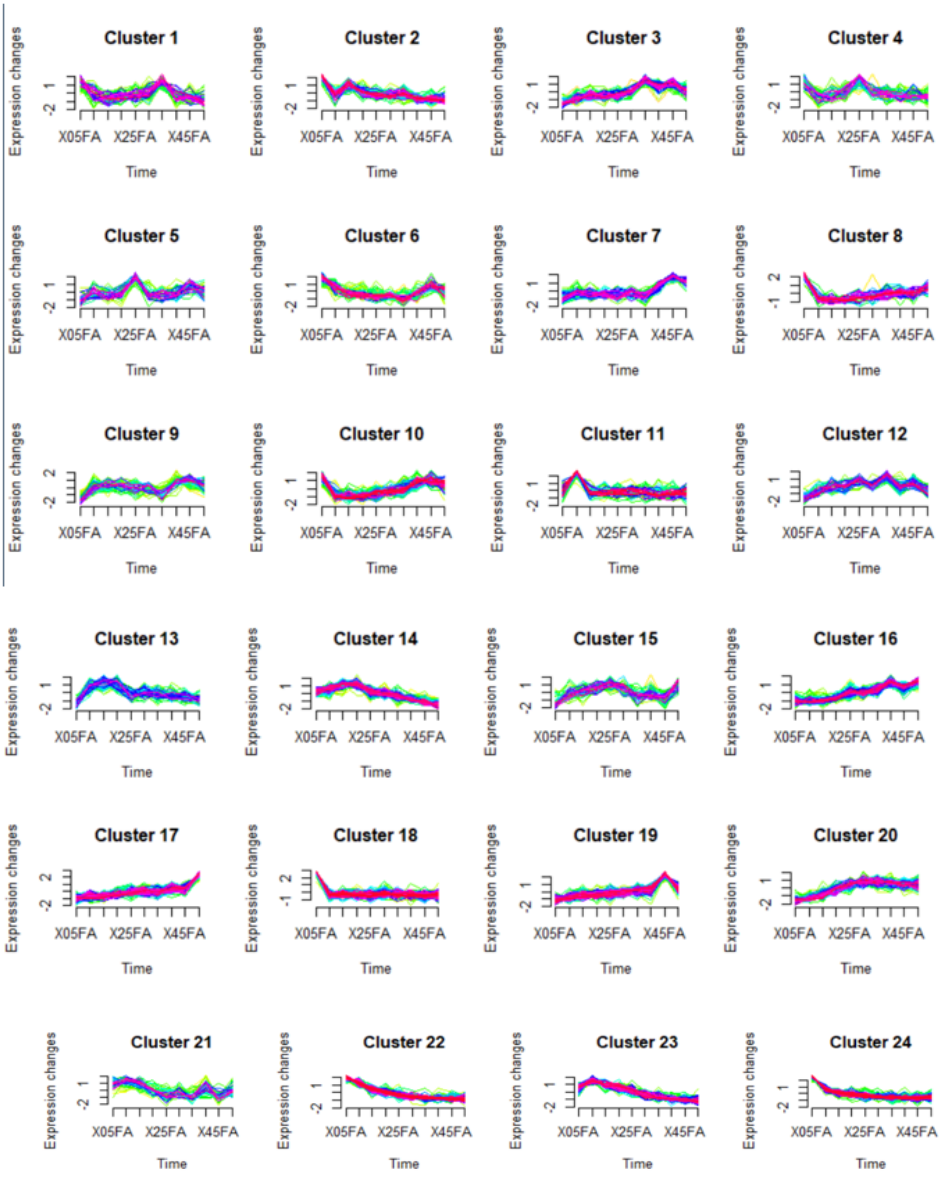


Figure.16 a,b,c Exploring trends in gene expression levels across 20 clusters (Cf- Figure 8).


Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Cluster 10
Cluster 11
Cluster 12
Cluster 13
Cluster 14
Cluster 15
Cluster 16
Cluster 17
Cluster 18
Cluster 19
Cluster 20
Cluster 21
Cluster 22
Cluster 23
Cluster 24
Expression changes
Time
X05FA
X25FA
X45FA

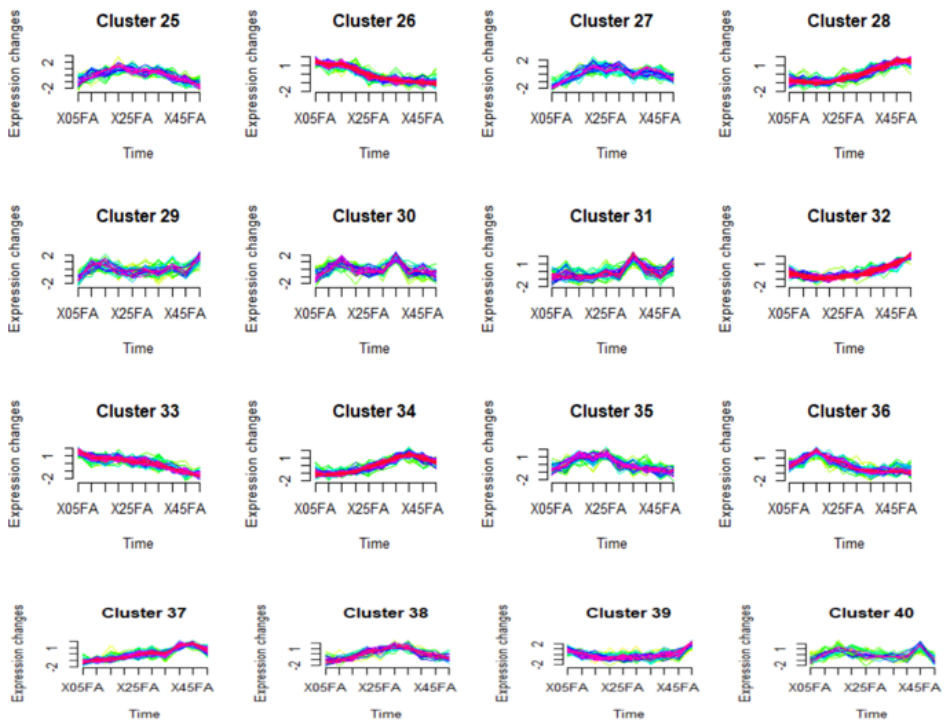


Figure.17 a,b,c Exploring trends in gene expression levels across 40 clusters (Cf- Figure 8).

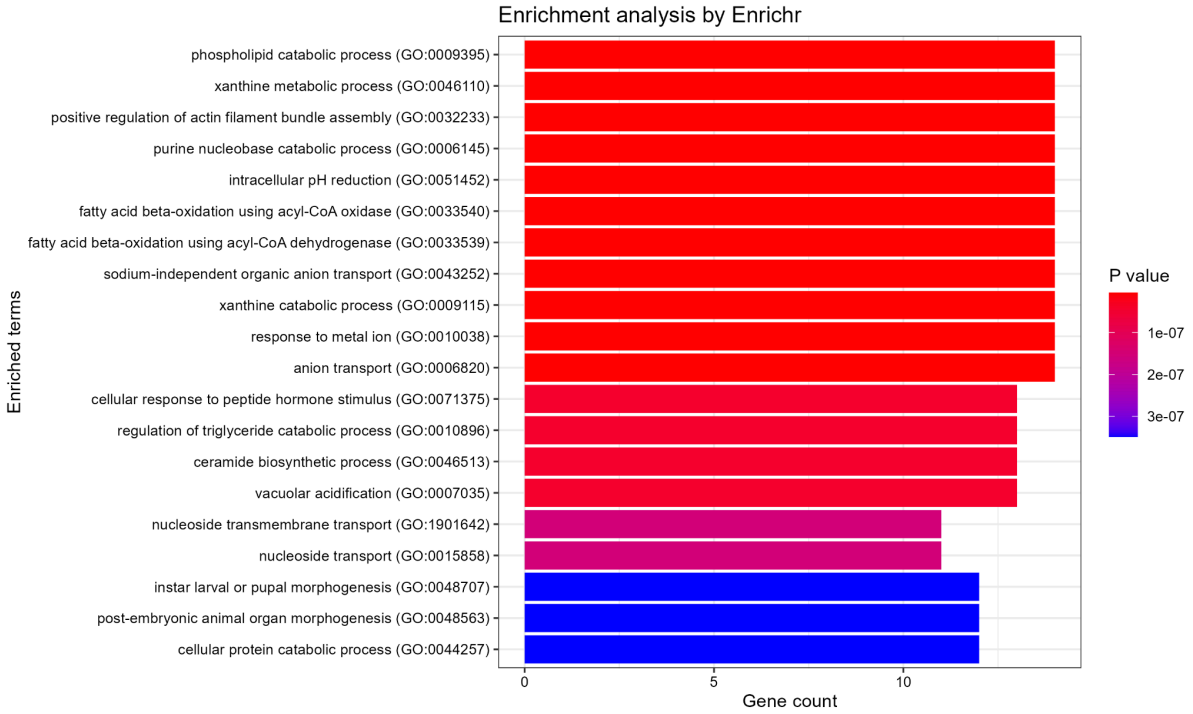


Figure.19 Functional enrichment of the down-regulated cluster number 1 in the FA condition

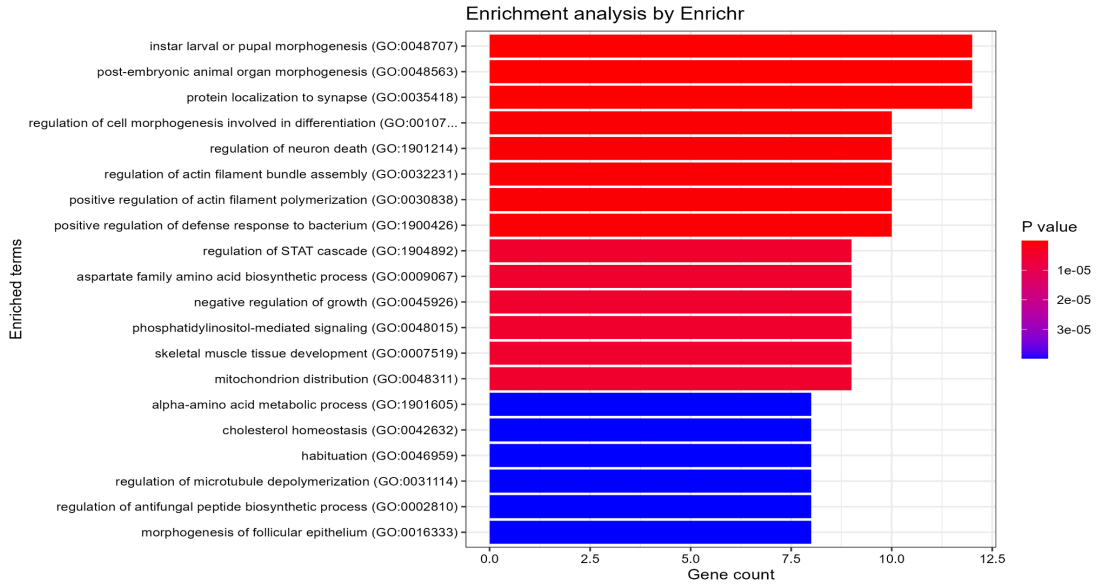


Figure.20 Functional enrichment of the down-regulated cluster number 12 in the FA condition

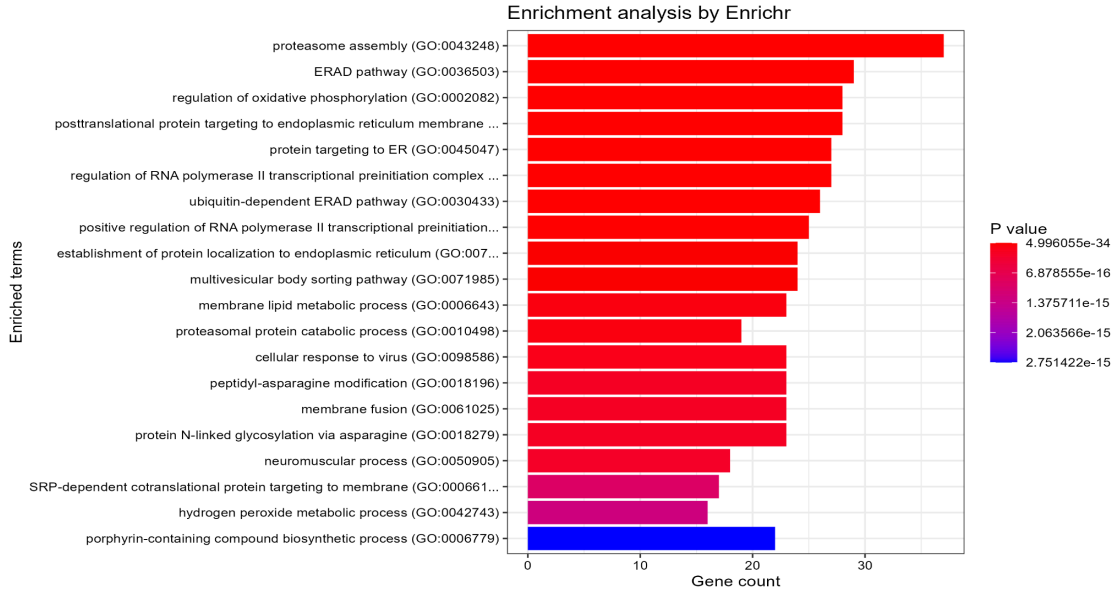


Figure.21 Functional enrichment of the down-regulated cluster number 6 in the FA condition

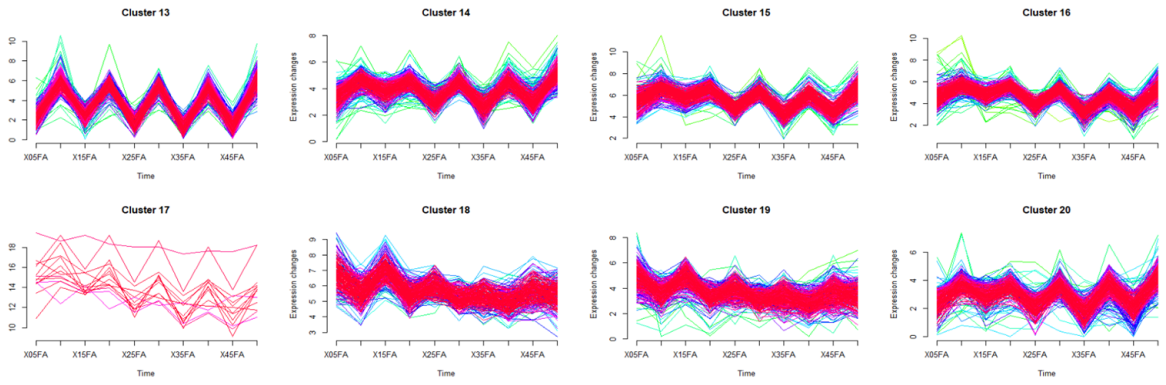


Figure.22 Noisy clusters before data normalization