BINF-402 Project

Differential expression analysis of micro-RNA transcriptome between pancreas, prostate and gastrocnemius medialis tissues

Léopold Guyot

December 12, 2023

1 Introduction

This analysis investigates microRNA expression variations across three distinct tissues; prostate gland, pancreas body, and gastrocnemius medialis. For this a simple workflow will be used, consisting of quality control of the reads, followed by a filtering, then a mapping to finish we a classic differential expression analysis. The goal is to unveil tissue-specific expression patterns of miRNA. Tissue selection is strategic, anticipating closer miRNA expression patterns between prostate and pancreas, both glandular, and unique signatures in gastrocnemius medialis, a muscle tissue.

2 Methods

All the data processing was done using the language R (R Core Team, 2023) and several packages. All graphs have been realize with the ggplot2 package (Wickham, 2016) and basic data manipulations have been done with the help of the tidyr package (Wickham et al., 2023). Note that for each section, the relevant scripts are indicated. Each script name is clickable to acces code through the associated github link. The link to the github repo is https://github.com/leopoldguyot/BINF-402_Transcriptomic_Project/

2.1 Data Retrieval

Script associated : retrieve_data.R

All the data sets used in this project have been retrieve from the ENCODE database (Luo et al., 2020). Three tissues have been selectionned; pancreas body, prostate gland and the gastrocnemius medialis tissue. For each tissue, data used was coming from two distinct experiments, each comprising two replicates, thereby totaling four replicates per tissue (cf. "data/sample_table_links.csv" for file accession numbers).

The original UCSC hg38 genome was used as reference for the mapping. The NCBI accession for this genome is GCA_000001405.26.

2.2 Read Quality Control

Scripts associated : reads_mapping.R, Quality_control_stats.R, quality_control.R

A preliminary analysis done using the Rqc package (de Souza et al., 2018) revealed notable issues with the quality of the sequencing reads. Others statistics have been retrieve and used to find the problems and adapt the processing of the reads.

Red line in the summary Figure 1 depict the evolution of the quality trough the cycles for the unprocessed reads, indicate a substantial low mean quality for the initial 5 cycles (certainly due to a adapter) and a significant decrease from the 43rd cycle to the end of the reads (mean quality of the cycles of unprocessed reads is highlighted by the red line in the figure).

To address these quality issues, an initial processing step was implemented using the QuasR package (Gaidatzis et al., 2015) (for performance reasons). This step involved trim-

ming the first 5 and the last 7 cycles of the reads (additionally, reads with unidentified residues were filtered out). The result is visible with the green line in Figure 1. Even after this first step, a persistent decrease in mean quality throughout the remaining cycles was still observed.

In response to the persistent decline in mean quality, a second processing step was undertaken. In this step, reads were filtered to retain only those with a mean quality higher than 20. The outcome of this filtering is illustrated by the blue line in Figure 1. This additional processing step led to a global increase in mean quality across the entire read length and a stable read quality through cycles, the previous trend of decreasing quality has been effectively mitigated.

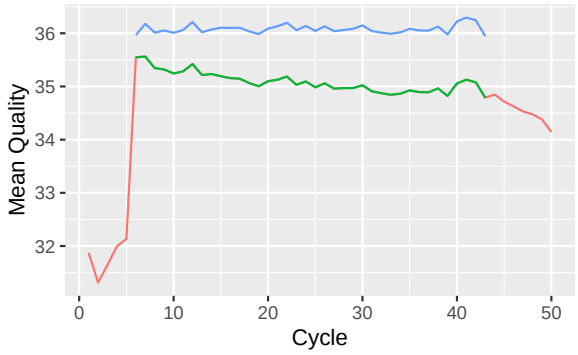

Figure 1: Graph of the mean quality (Phred score) for each cycle. Each line is composed of the mean quality values of all the reads contained on the 12 datasets (4 replicates and 3 tissues). The red line is for the unprocessed data, the green line for the data with trimmed reads and the blue line for the data with trimming and filtering.

In the following section, the mapping performance for each version of the data (unprocessed, with trimming, with trimming and filtering) will be explored. This analysis will provide insights into how the quality enhancements impact the alignment of reads.

2.3 Mapping

Script associated : reads_mapping.R

Before mapping, an index of the hg38 genome was created. The mapping was conduct on unprocessed data sets, data sets with reads trimming and data sets with trimming and filtering. Both indexing and mapping were carried out with the Rsubread package (Liao et al., 2019).

By comparing the mapping proportions, we can clearly see the impact of quality control on the mapping performance (cf Fig.2). With no processing on the reads, the proportion is really low with a median value under 10%, some data sets have mapping proportion that do not go above 2%. After a first trimming of the start and end of the reads, the improvement is quite visible, as we go from 10% to slightly under 75% of median mapping proportion. And the minimal proportion is not going under 65%. With the extra step of filtering the reads, the median proportion slightly increase and is above 75%. And the minimal proportion is 68%.

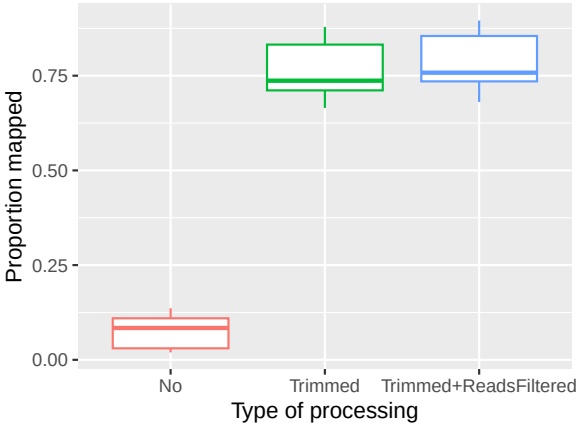

Figure 2: Graph of the proportion of reads mapped based on the read processing method used. For each category, n = 12 (4 replicate for 3 tissues). The first category "No" stands for the case with no processing.

2.4 Differential Expression Analysis

Script associated : differential_expression_analysis.R

Before running the proper differential analysis. A feature count was carried out with the Rsubread package (Liao et al., 2019). The annotation used was the one contained within the package.

The differential expression analysis was done using the DESeq2 package (Love et al., 2014) the workflow employed is inspired from the workflow presented in Chapter 4 of the "Omics Data Analysis" UCLouvain course from Laurent Gatto (Gatto and Loriot, 2023).

The big strength of DESeq2 is that it will correct the count matrix to account for potential bias between data sets. These bias are for

instance difference in sequencing depth across sample (quite present in this analysis cf. Fig 3) and the difference in library composition that can lead to bad normalisation if not taken into account. To accomplish this normalisation, DESeq2 will compute geometric mean for each feature then it will use these values and create a new matrix that consist of the counts divided by their associated geometric mean. After that it will use this new matrix to build scaling factors for each sample, and finally applying this scaling factor to the original counts.

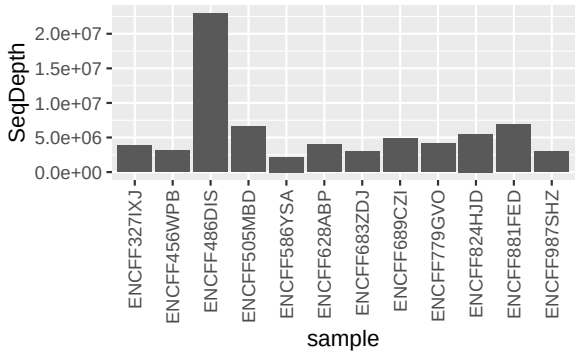

Figure 3: Graph of the sequence depth for each sample, 486DIS shows a significant higher depth. This figure show the importance of counts normalisation.

A first exploration of the normalized count matrix was done using dimension reduction method. To avoid that largely expressed features impact strongly the new dimension reduction, at the expense of features with lower expression profile, a regularized-logarithm transformation was carried out. This type of transformation allow low and high expressed features to have same weight during the dimension reduction step. A classic Principal Component Analysis was conduct. Then an unsupervised clustering algorithm (using k-means method with k = 3) was used to assess if the different tissues types can be identify based on the mi-RNA expression pattern.

To access the differential analysis, DESeq2 employs a step that will estimate the dispersion parameter of the negative-binomial distribution. This estimation rely to the assumption that features of similar expression levels have similar dispersions and will thus use information coming from similar expressed features to estimate the dispersion values. Then it proceeds with extra step that will lead to reduce false positives in the differential expression analysis. These steps include, fitting the dispersion values to then shrink the dispersion values toward the values predicted by the curve. The results of the dispersion estimation is visible on the Figure 4

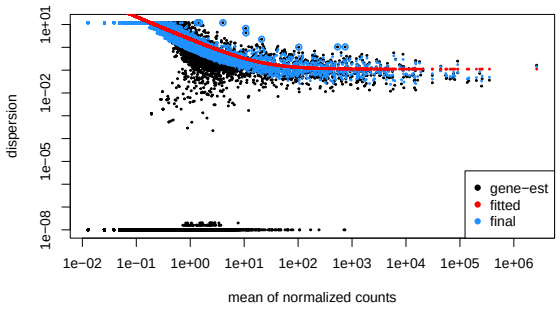

Figure 4: Graph of the dispersion estimation carried out by DESeq2. Black dot show the estimated dispersion for each feature in function of the mean expression level. The red line is the fitted curve to feature-wise dispersion estimates. Blue points are the new values obtained with the shrinkage of the initial toward the predicted values. Black points circled with blue show features with too high dispersion that are not shrunk.

DESeq2 will then fits a generalized linear model. This model will give us the log2 Fold Change between two sampled type and its associated p-value (the p-value is obtained with Wald test). Due to the high number of statistical tests that need to be carried out, DESeq2 uses Benjamini-Hochberg method to adjust pvalues. To decrease the loss of statistical power associated with the multiple testing correction, DESeq2 filter out the tests that have almost no chance to show a significant fold change.

The results of this analysis will be explored on the Results section.

3 Results an Discussion

3.1 Graphic exploration

The dimension reduction (cf. Fig 5) show that the 3 sample groups are well separated within the new dimension space (based on the features dimensions). Although visually distinct groupings are apparent, their separation lacks numerical interpretability. Therefore, an unsupervised clustering was used to obtain more robust

group assignments. This clustering showed a correct classification of the sample, and the ratio between "between-cluster sum of squares" and "total sum of squares" was 99%. This means that 99% of the distance between points is explained by the groups. As a result, this prove that these 3 tissue have quite different mi-RNA expression profile.

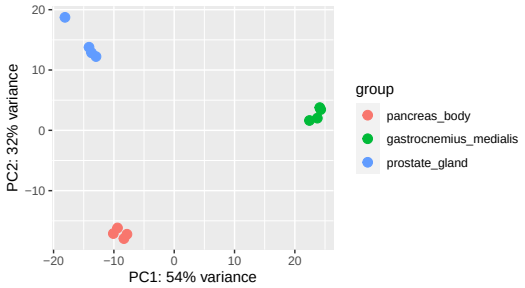

Figure 5: PCA plot, showing the two first components axis. Variance associated with component is indicate on legend axis. The PCA was compute on the top 500 features by variances.

3.2 Differential expression

A first exploration of the differential expression analysis result can be done using volcano plot, that represents the adjusted p-value by the log2 of fold change. Since the analysis is based on comparison between two sample group, the results were produce to compare Pancreas versus Prostate and Pancreas versus Gastrocnemius medialis. The choice behind these pairs of comparison is to answer the initial hypothesis("anticipating closer miRNA expression patterns between prostate and pancreas, both glandular, and unique signatures in Gastrocnemius medialis, a muscle tissue" cf. Introduction). Visually, the two volcano plots (cf Fig. 6) seems to be quite similar, both show an important number of mi-RNA with an considerable log2 Fold change (greater than 1) and a low adjusted p-value (less than 0.05). In fact if we count the number of these mi-RNAs, we have 558 for the comparison with Gastrocnemius medialis and 597 for the comparison with prostate. This difference is not significant (chi test pval = 0.2705) therefore we can not say that there is a pair of tissue type that have much similar mi-RNA expression pattern. Thus we reject the initial hypothesis.

4 Conclusion

To conclude this report, I want to highlight things I could add to my analysis workflow. For the quality control, it would maybe be better to trim the read ends selectively, maybe cut off cycle that have low quality values. A newer version of the genome would maybe increase the proportion of read mapped. As for the sample chosen, maybe it would be better to choose two different tissues type and have more replicate. This would simplify and improve the differential expression analysis. Other results could be explored, for instance comparing the normalized counts of the highest expressed features across the tissue types, run a Gene Ontology on the highest expressed mi-RNA would also be a great idea.

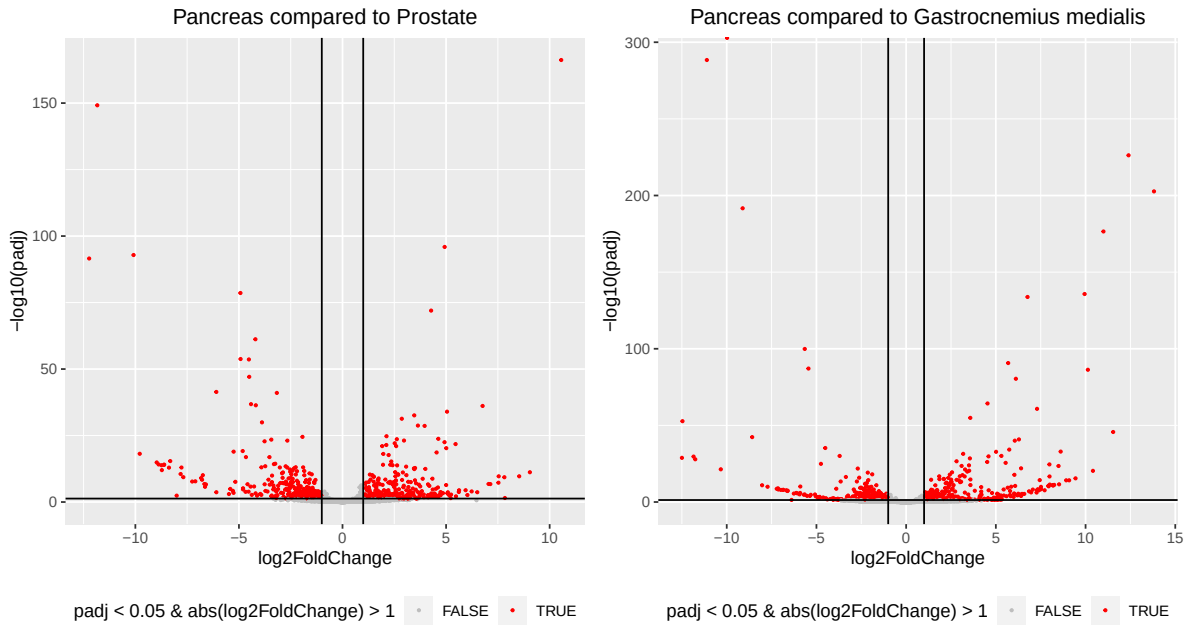

Figure 6: Volcano plot of the two differential expression analysis. Red points represents mi-RNA with an adjusted p-value lower than 0.05 and an absolute log2 Fold change higher than 1 (black lines are threshold values). The volcano plot on the left show the results for the comparison between Pancreas body and Prostate gland, as the right one show the comparison between Pancreas body and Gastrocnemius medialis.)

References

W. de Souza, B. S. Carvalho, and I. Lopes-Cendes. Rqc: A Bioconductor package for quality control of high-throughput sequencing data. *Journal of Statistical Software, Code Snippets*, 87(2):1–14, 2018. doi: 10.18637/jss.v087.c02.

D. Gaidatzis, A. Lerch, F. Hahne, and M. B. Stadler. Quasr: Quantification and annotation of short reads in r. *Bioinformatics*, 31(7):1130–1132, 2015. doi: 10.1093/bioinformatics/btu781. PMID:25417205.

L. Gatto and A. Loriot. Wsbim2122: Omics data analysis, 2023. URL `https://github.com/UCLouvain-CBIO/WSBIM2122`.

Y. Liao, G. K. Smyth, and W. Shi. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research*, 47:e47, 2019. doi: 10.1093/nar/gkz114.

M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15:550, 2014. doi: 10.1186/s13059-014-0550-8.

Y. Luo, B. C. Hitz, I. Gabdank, J. A. Hilton, M. S. Kagda, B. Lam, Z. Myers, P. Sud, J. Jou, K. Lin, et al. New developments on the encyclopedia of dna elements (encode) data portal. *Nucleic acids research*, 48(D1): D882–D889, 2020.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL `https://www.R-project.org/`.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org`.

H. Wickham, D. Vaughan, and M. Girlich. *tidyr: Tidy Messy Data*, 2023. URL `https://CRAN.R-project.org/package=tidyr`. R package version 1.3.0.