BINF-402 Project

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

Léopold Guyot

December 9, 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. 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 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 trimming 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 through-

out 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.

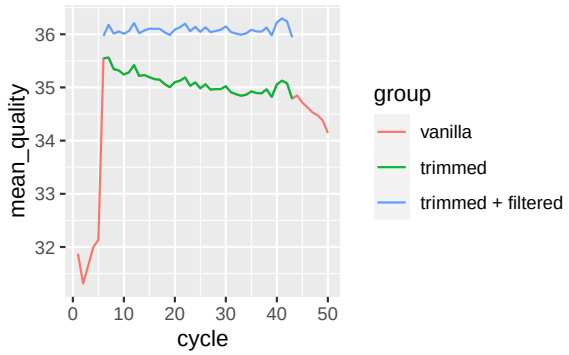

Figure 1: Graph of the mean quality for each cycle.

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

present mapping stats

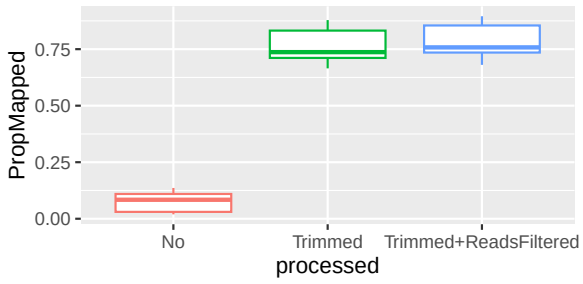

Figure 2: Caption of the figure.

2.4 Differential Expression Analysis

Script associated : differential_expression_analysis.R

3 Results

Present the overall results => Differential Expression Analysis

4 Discussion

???? maybe not include it => things I could Improve

5 Conclusion

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.

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/`.