Trabajo Fin de Máster

Papel de los microRNAs en la regulación del metabolismo del tejido adiposo.

Role of microRNAs in adipose tissue matabolism

Autor

José Andrés Castillo Rivas

Directores

Silvia Lorente Cebrián
José Miguel Arbonés Mainar

Master in Biophysics and Quantitative Biotechnology

FACULTAD DE CIENCIAS DE LA SALUD Y EL DEPORTE
2024–2025

Dedication

Acknowledgments

Abstract

Table of contents

List of Tables

List of Figures

Chapter 1

Introduction

Chapter 2

Justification of the Topic

Chapter 3

Objectives

lalallallalalalallalala

Chapter 4

Problem Statement

lalallallalalalallalala

Chapter 5

Hypothesis

Chapter 6

State of the Art

lalallallalalalallalala

Chapter 7

Theoretical Framework

lalallallalalalallalala

Chapter 8

Methodological Framework

8.1 Input Data

The small RNA sequencing (sRNA-seq) data used in this study were generated on the Illumina sequencing platform and provided in FASTQ format. The FASTQ files correspond to 78 samples of subcutaneous adipose tissue from obese patients belonging to the FATe cohort. The patients included both men and women aged between 22 and 61 years, classified into three degrees of hepatic steatosis: normal liver (grade 0, 0 points), mild hepatic steatosis (grade 1, 1 to 3 points), moderate hepatic steatosis (grade 2, 4 to 6 points), and severe hepatic steatosis (grade 3, 7 to 9 points). The data is summarized in.

8.2 Analysis of sRNA-seq Data with *nf-core/smrnaseq*

For the analysis of small RNA sequencing (sRNA-seq) data, version 2.4.0 of the nf-core/smrnaseq pipeline (Peltzer et al., 2024) was used, which is specifically designed for the automated processing of microRNA data. This pipeline facilitates quality control, filtering, and quantification of microRNAs and their variants, and it was executed using the Docker profile to ensure reproducibility and compatibility across operating systems.

8.2.1 Execution of the *nf-core/smrnaseq* Pipeline

The installation of *nf-core/smrnaseq‘* was carried out following the instructions provided by the authors, available at https://nf-co.re/smrnaseq/2.4.0.

To ensure the proper installation and execution of the pipeline, the following key components were installed beforehand:

1. *Nextflow*: Version 24.04.4 of Nextflow was used, following the detailed instructions at https://nf-co.re/usage/installation.

2. *Java Runtime Environment (JRE)*: Version 11.0.25 of the Java Runtime Environment was installed, as it is required for compatibility with Nextflow and the *nf-core/smrnaseq* pipeline.

To ensure the reproducibility of results and facilitate pipeline execution, the authors recommend installing one of the available Docker containers. These containers include the necessary instructions and configurations for running the pipeline. This configuration is specified at runtime using the `profile` argument. In this analysis, the Docker image of *nf-core/smrnaseq*, available at https://hub.docker.com/r/nfcore/smrnaseq, was used.

The pipeline was executed on a server with 8 CPUs, 16 GB of RAM, and a Linux operating system. The following command was used in the terminal, which configures the main options, including the reference genome, input data, and output file location:

```
nextflow run nf-core/smrnaseq -r 2.4.0
-profile docker,ci
--genome GRCh38
--input '/home/joshoacr13/Documentos/TFM/nfcore-smrnaseq/input/samples.csv'
--fasta 'https://github.com/nf-core/test-datasets/raw/smrnaseq/reference/genome.fa'
--mirtrace_species 'hsa'
--outdir /home/joshoacr13/Documentos/TFM/nfcore-smrnaseq/workdir
-resume -c /home/joshoacr13/Documentos/TFM/nfcore-smrnaseq/nextflow_memory.config
--save_intermediates FALSE
```

8.2.2 Description of the Parameters Used

- `-profile docker,ci`: Runs the pipeline inside a Docker container to ensure reproducibility and sets up a continuous integration (CI) profile.
- `--genome GRCh38`: Specifies the human genome (version GRCh38) as the reference for sequence mapping.
- `--input`: Provides the path to the CSV file containing metadata and the paths to the FASTQ files.
- `--fasta`: URL to the FASTA file of the reference genome.
- `--mirtrace_species hsa`: Defines the species as Homo sapiens (hsa) for microRNA analysis with miRTrace.
- `--outdir`: Sets the working directory for the processed results.
- `-resume`: Allows continuation of a previous analysis without restarting from the beginning.
- `-c`: Specifies a custom configuration file (nextflow_memory.config) to adjust resource usage.
- `--save_intermediates FALSE`: Prevents the storage of intermediate files to save disk space.

8.2.3 Analysis Workflow and Tools Used

The *nf-core/smrnaseq* pipeline performs the following steps:

1. **Quality Control** An initial quality assessment of the raw reads was conducted using *FastQC* (version 0.12.1). Additionally, 3' adapter trimming was performed using *fastp* (version 0.23.4), followed by quality and length filtering. A second quality assessment of the trimmed reads was conducted with *FastQC*.

2. **miRNA Quality Control** A more specific quality control for miRNAs was performed using *mirtrace* (version 1.0.1). This tool allowed us to:

- **Verify Read Length Distribution**: The majority of reads fell within the expected range of 18–24 nucleotides, indicative of high-quality small RNA data.

- **Identify Contaminants**: Potential contaminants such as tRNA, rRNA, and other non-target molecules were flagged.
- **Taxonomic Classification**: Reads were classified taxonomically to ensure that most sequences originated from the organism of interest (*Homo sapiens*).

Samples that failed to meet the minimum quality thresholds, as determined by *mirtrace*, were excluded from further analysis to maintain data integrity and reliability.

3. **Contaminant Filtering**: Contaminant reads identified in the quality control step were removed using *Bowtie2* (version 1.3.0) . Reads were aligned against reference databases specific to: rRNA (ribosomal RNA), tRNA (transfer RNA), piRNA (PIWI-interacting RNA) and ncRNA (non-coding RNA)

This filtering step ensured enrichment for miRNA-specific reads, reducing noise and improving the precision of downstream analyses.

4. **MicroRNA Quantification**:

- **Alignment**: The filtered reads were aligned against mature microRNA sequences in the miRBase database using *Bowtie1*. Unmapped reads were aligned against "hairpin" sequences to identify microRNA precursors.
- **Post-Alignment Processing**: *SAMtools* was used to process the mapping results.
- **Quantification and Normalization**: Initial quantification was performed with *edgeR*, generating normalized count tables (TMM) for detected microRNAs. Exploratory graphs were generated, including a multidimensional scaling (MDS) analysis to cluster samples and a heatmap to evaluate similarities among them.

5. **IsomiR Annotation**: The collapsed reads were processed with *mirtop* (version 0.4.28) to identify microRNA variants (isomiRs). This analysis allows for the mapping and annotation of variants related to length and sequence modifications of mature microRNAs.

The mirtop tool employs the Blending Analysis technique to process and integrate miRNA data, ultimately generating a count matrix that accurately represents the expression levels of these molecules in the analyzed samples. This method includes the following essential steps:

- **Read Grouping**: The miRNA reads are grouped from the processed data, ensuring that different variants and reference sequences are integrated coherently.
- **Adjustment for Variants**: Both miRNA variants (isomiRs) and standard reference sequences extracted from databases such as miRBase are considered. This adjustment is fundamental to obtaining an accurate representation of miRNA expression in the analyzed samples.

The application of *Blending Analysis* allows for the generation of a more robust and comprehensive count matrix, thus facilitating subsequent differential expression analysis.

6. **Analysis and Visualization of Results**: The overall pipeline metrics, encompassing quality assessments, mapping statistics, and expression analysis results, were consolidated and summarized using *MultiQC* (version 1.25.1). This versatile tool that aggregates output from various bioinformatics analyses into a unified, interactive report, enabling an efficient overview of the data processing workflow.

8.3 Differential expression analysis

8.3.1 Working Environment and Computational Resources

For the differential expression analysis, the R statistical software (R Core Team, 2024), version 4.4.1 (2024-06-14) (https://cran.r-project.org/), was used. This analysis was performed using the RStudio integrated development environment (IDE) (Posit team, 2023), version 2023.12.0+369, designed for Ubuntu Jammy (https://www.rstudio.com/).

The script used to perform the differential expression analysis is available in the file "`miRNA_steatosis.qmd`" which can be accessed at the following link: https://github.com/joshoandres13/miRNAs.

To execute this script, several libraries must be installed. Some of these are standard libraries available on CRAN, while others are specific to sequencing data analysis and are part of the Bioconductor project (Morgan, 2024) (https://www.bioconductor.org/, version 3.19.1).The libraries utilized in this analysis include *isomiRs*(Pantano & Escaramis, 2024) and *DESeq2* (Love et al., 2014), both of which are part of Bioconductor.

8.3.2 Data Import and Preparation

The starting data consisted of `.tsv` format files generated by the *mirtop* tool, which is integrated into the *nf-core/smrnaseq* pipeline. These files contained raw isomiR counts for each sample. They were then imported into R and combined into a single count matrix, where:

- Rows represent the identified isomiRs.
- Columns correspond to the experimental samples.

Additionally, a metadata matrix was created to describe the experimental conditions of each sample, including variables such as `sex` and `steatosis`.

Using the *mirtop* count matrix and the metadata matrix, an object of class `IsomirDataSeq` was created. This object is fundamental within the *isomiRs* package, as it allows for efficient management of the information derived from small RNA sequencing studies, facilitating differential expression analysis and the interpretation of biological results.

8.3.3 Filtering and Processing of isomiRs

8.3.3.1 Filtering of isomiRs with Low Read Counts

The filtering process allows for the grouping of isomiRs into different categories, assigning them to a single variant associated with a miRNA. This grouping is crucial to ensure consistency and accuracy in differential expression analyses.

To minimize technical noise in the data and focus on biologically relevant signals, a strict filtering criterion was applied. Only isomiRs with at least 20 reads in at least 40 samples were retained. This step is essential to eliminate sequences with low representation that could affect the robustness of subsequent analyses.

8.3.4 Gene Expression Data Analysis

Gene expression analysis was conducted using the *DESeq2* package in R, which models count data and performs statistically robust tests to identify significant differences in gene expression. The analysis steps are outlined below:

8.3.4.1 Data Preparation

The analysis began with the creation of a `DESeqDataSet` object from a count matrix and a metadata table describing the experimental conditions. In this case, the count matrix contained expression data for isomiRs (variants of a single RNA), and the experimental design included two variables: **sex** and **steatosis**.

8.3.4.2 Model Fitting

To evaluate the effect of steatosis on isomiR expression, a full model including both variables was fitted. A reduced model excluding the steatosis variable was then fitted, allowing a comparison between the two models using the Likelihood Ratio Test (LRT).

8.3.4.3 Obtaining Results

The results of the analysis were obtained using the `results()` function, which provides a data frame containing information about log2 fold changes and adjusted p-values.

8.3.4.4 Filtering Significant Results and Visualization

The criteria established to identify significant isomiRs:

- *Strict criterion*: Selected isomiRs with an adjusted p-value (padj) less than 0.05 and an absolute log2 fold change greater than 1 for the higly expresed and less than 1 for the less expressed.

To facilitate result interpretation, a scatter plot was generated showing log2FoldChange on the x-axis and -log10 p-value on the y-axis. Points were colored red to indicate significant isomiRs and black for non-significant ones. The significants will be used in the following steps.

8.4 Target mRNA Selection and Validation Using *multiMiR*

The identification of mRNA targets was performed using the *multiMiR* bioinformatics package (Ru et al., n.d., 2014), version 2.4.0 in R. *multiMiR* facilitates systematic search and annotation of microRNA targets, providing functional analysis to elucidate biological mechanisms. For this analysis, only validated interaction data were used.

8.4.1 Filtering Parameters

The validated target table provided by *multiMiR* was used during the selection process. Key columns included:

1. **database**: Source database of validated interactions, such as *miRTarBase*, *TarBase*, or *miRecords*.
2. **mature_mirna_id**: Standard format identifier for the miRNA.
3. **target_symbol**: Target gene symbol.
4. **experiment**: Experimental methods used for validation, including luciferase assays, Western blot, or qRT-PCR.

5. **support_type**: Level of experimental support, such as “Functional MTI” (miRNA-mRNA functional interaction).
6. **pubmed_id**: References to PubMed articles reporting the interaction.
7. **type**: Specifies whether the interaction is “validated” or “predicted.”

8.4.2 Selection Criteria

To ensure reliable results, databases were filtered according to update criteria and the following selection parameters: - Databases up-to-date at the time of analysis were prioritized (*miRTarBase* and *TarBase*).

- Only interactions classified as “validated” were included.
- Interactions backed by robust experimental methods, such as luciferase assays or Western blot, were prioritized.
- Interactions with functional support (“Functional MTI”) and verifiable references in PubMed were selected.

This approach ensured the identification of mRNA targets with high reliability and experimental backing, facilitating the analysis of potential regulatory functions of the studied miRNAs.

8.4.3 Functional Analysis

To explore biological functions associated with validated target genes, Gene Ontology (GO) enrichment analysis was conducted using the *clusterProfiler* package (Xu et al., 2024) in R. This analysis identified biological processes, molecular functions, and cellular components involving miRNA-regulated genes.

1. **Data Preparation**:
   - Symbols for validated genes (*target_symbol*) associated with selected miRNAs were extracted using *multiMiR*, with duplicates removed.
2. **GO Enrichment Analysis**:
   - The `enrichGO()` function from *clusterProfiler* was used with the following parameters:
     - `OrgDb`: Human gene database from *org.Hs.eg.db* (Carlson, 2024).
     - `keyType`: Key type defined as “SYMBOL”.
     - `ont`: Ontology type analyzed, including “ALL” (biological processes, molecular functions, and cellular components).
     - `pAdjustMethod`: False discovery rate (FDR) adjustment method using Benjamini-Hochberg.
     - `qvalueCutoff` and `pvalueCutoff`: Cutoff values set to 0.05 to select significant results.
3. **Results and Visualization**:
   - A bar plot of the top 10 enriched categories in biological processes (GO:BP) was generated, showing statistical significance and the number of genes associated with each category.
   - The plot highlighted key biological processes related to the activity of miRNA target genes.

Chapter 9

Results

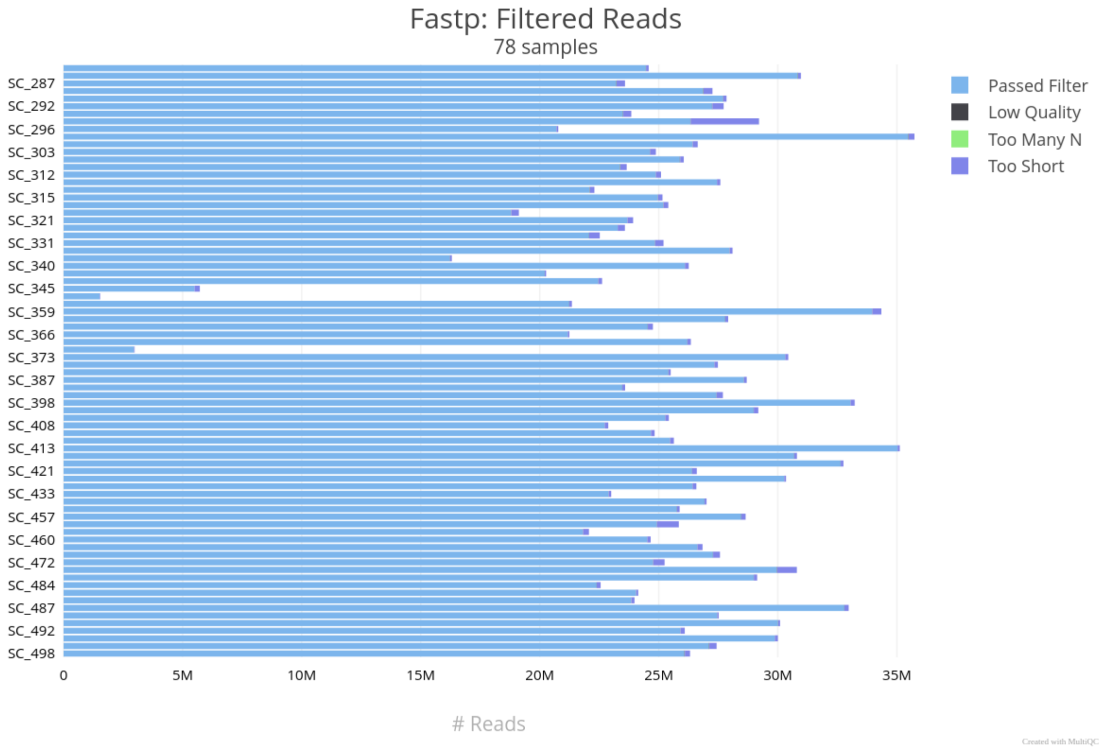


Figure 9.1: Fastp: Filtered Reads

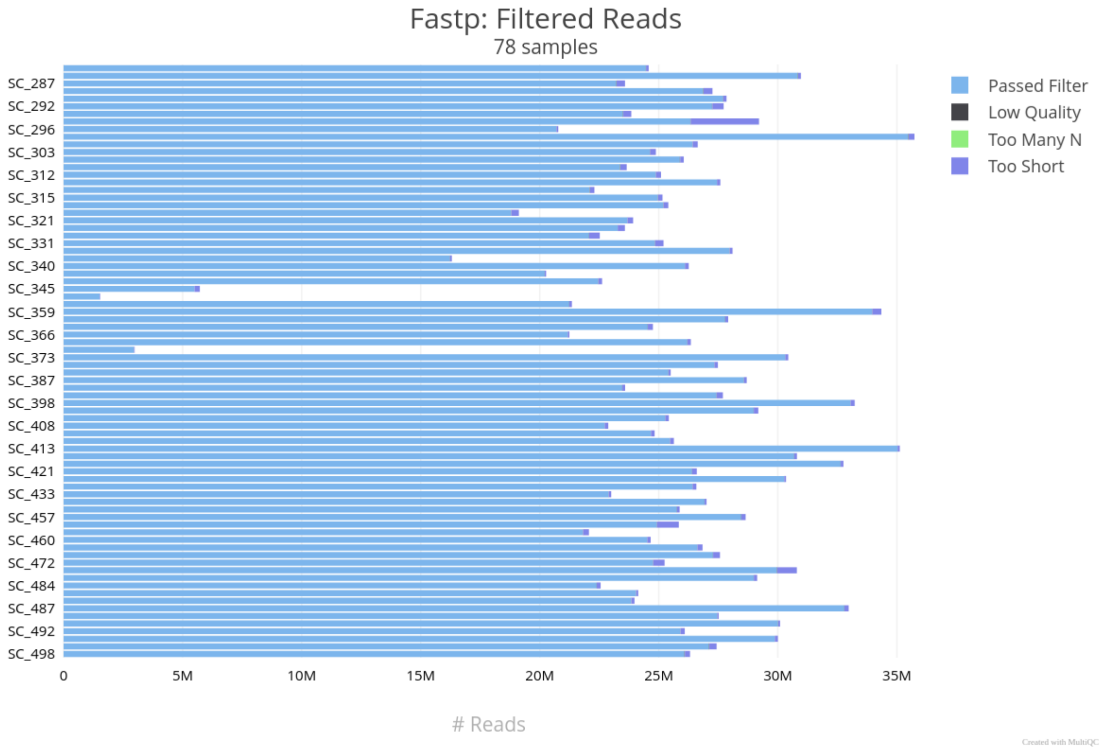


Figure 9.2: Fastp: Filtered Reads

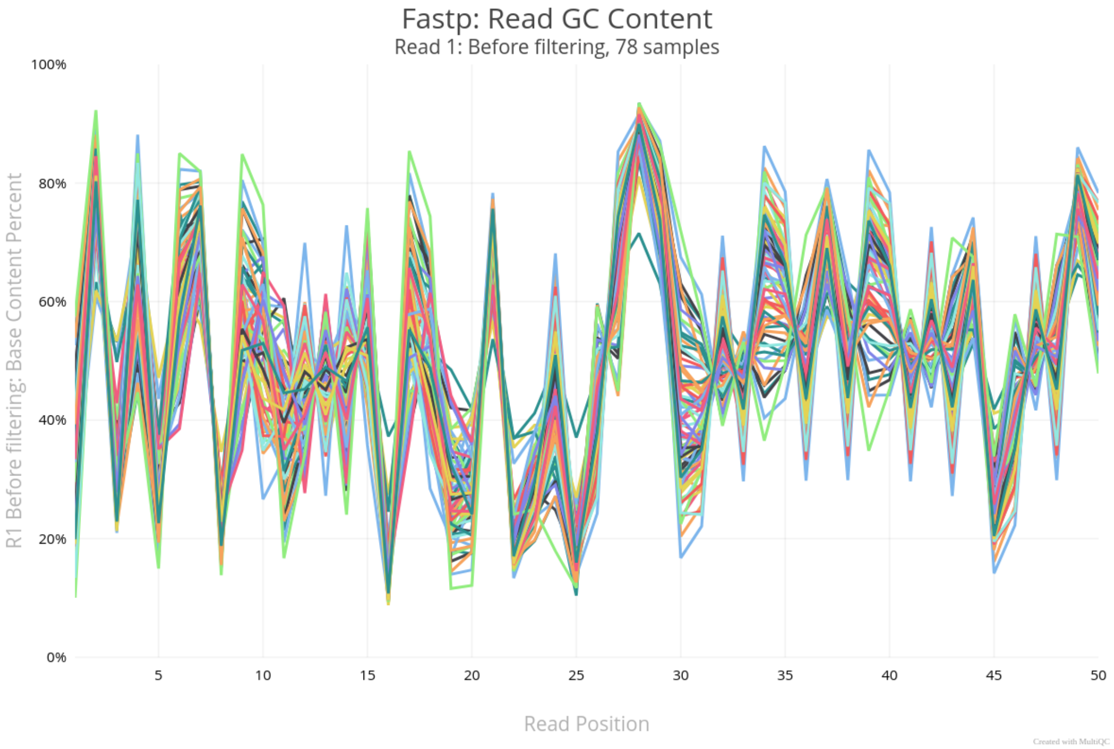


Figure 9.3: Fastp: Read GC Content

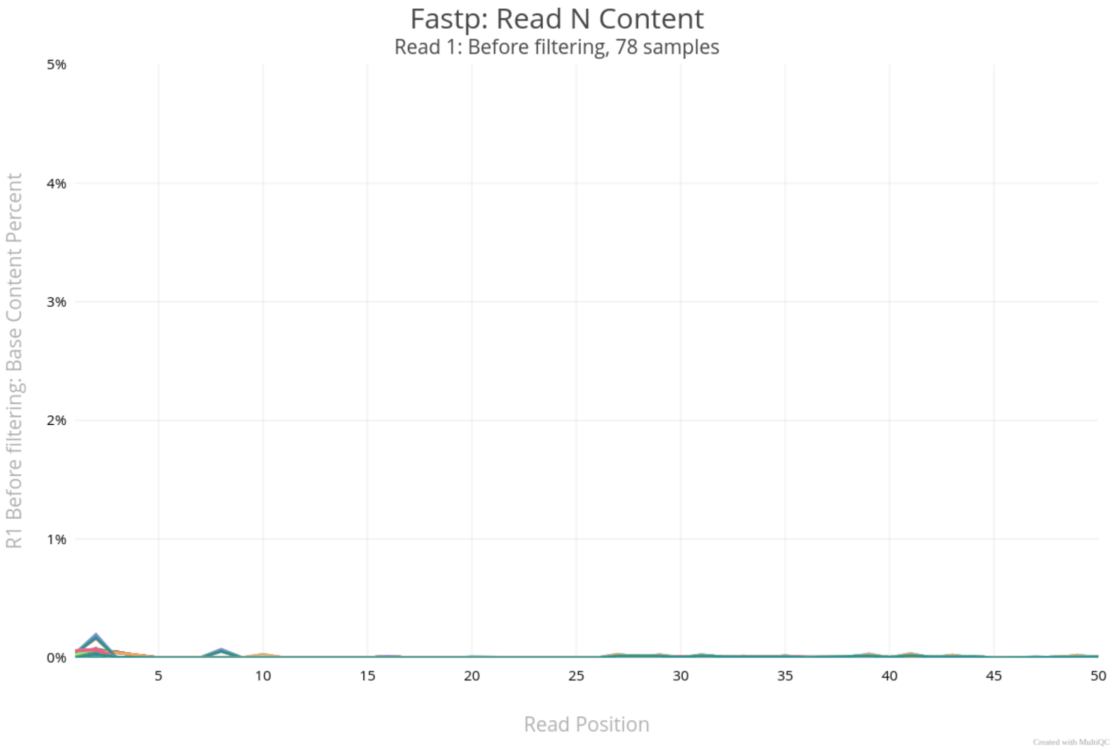


Figure 9.4: Fastp: Read N Content

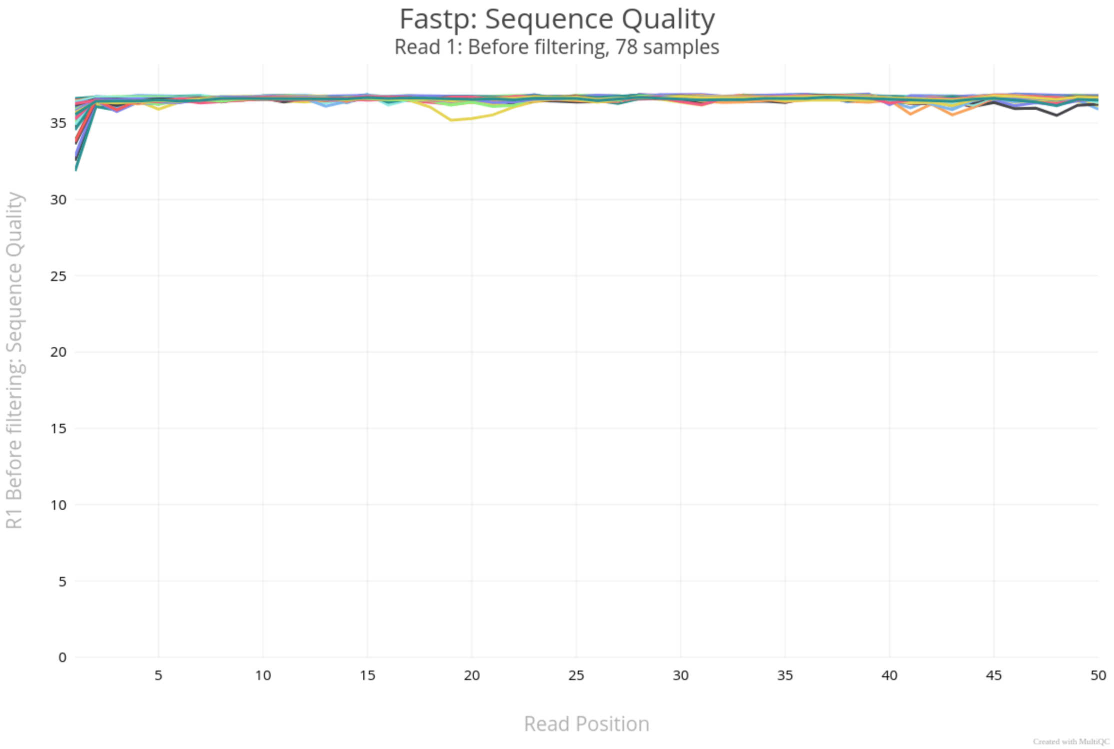
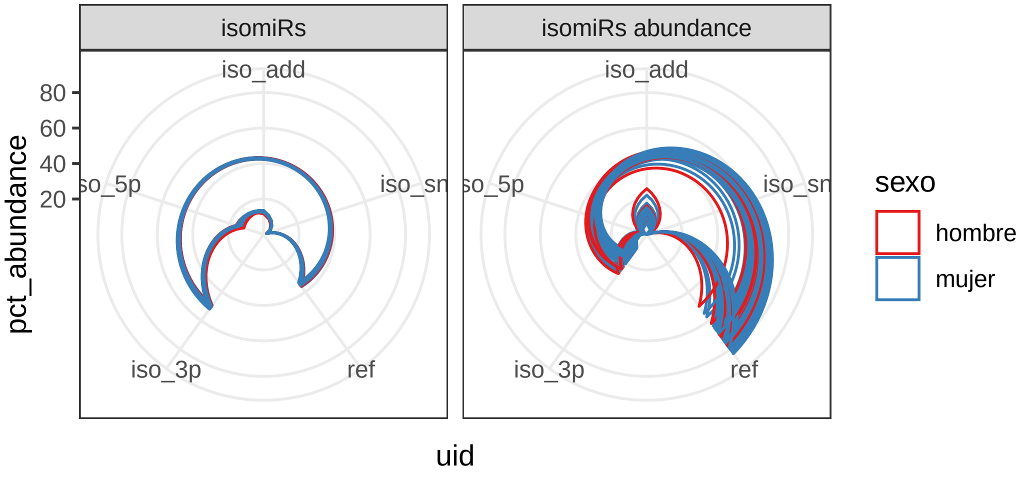
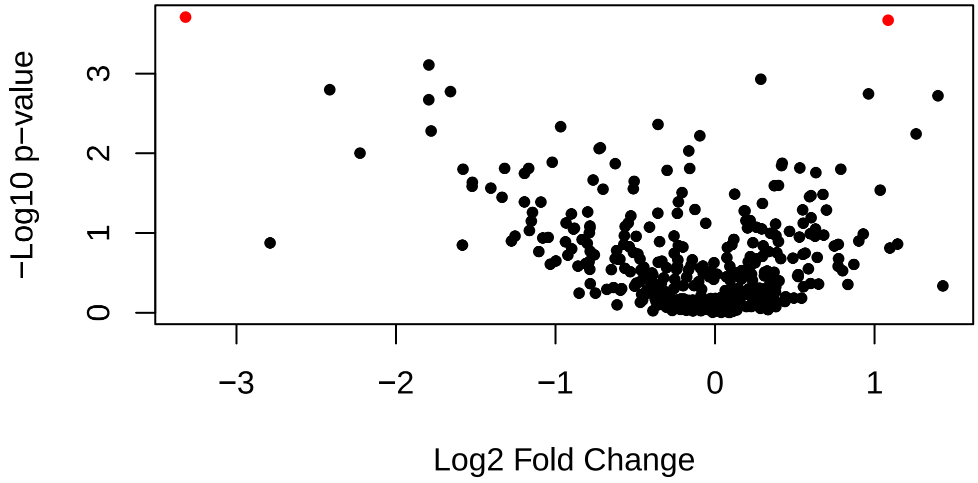
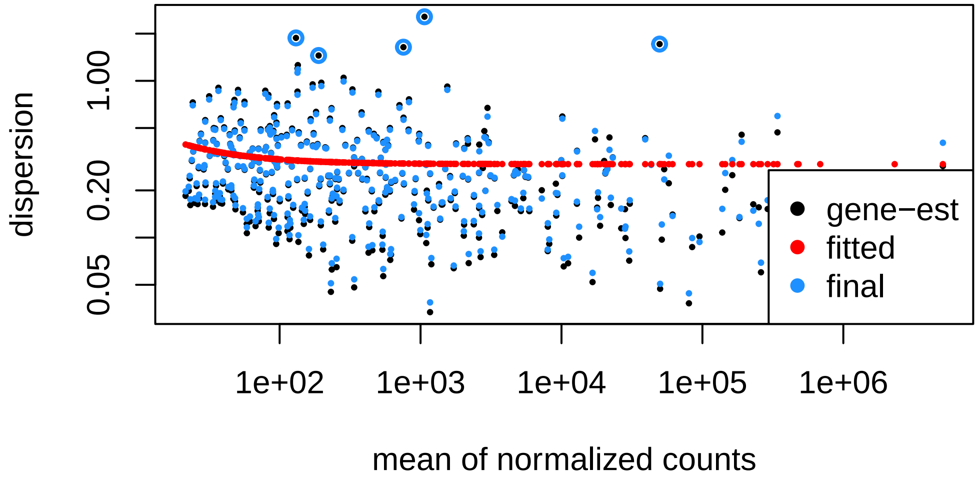
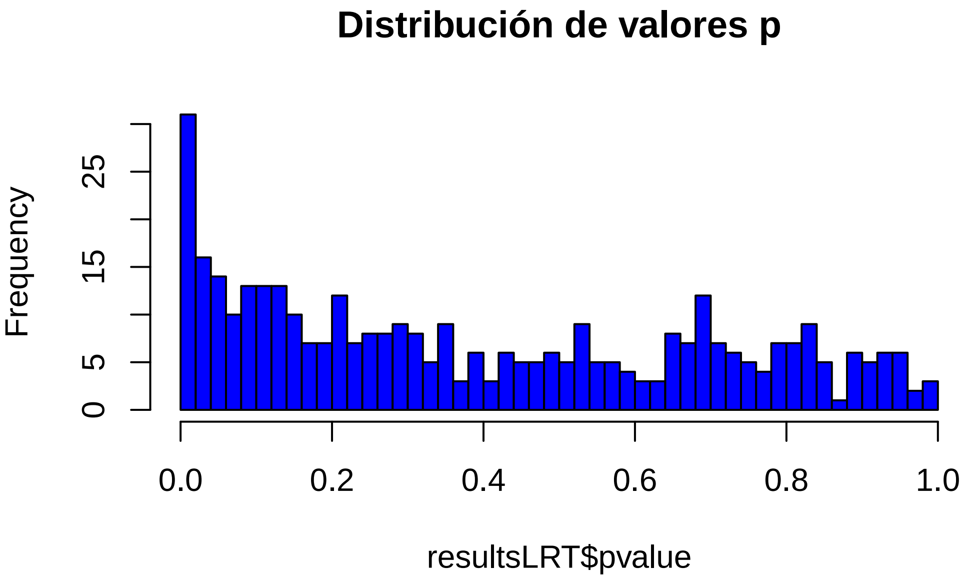


Figure 9.5: Fastp: Sequence Quality

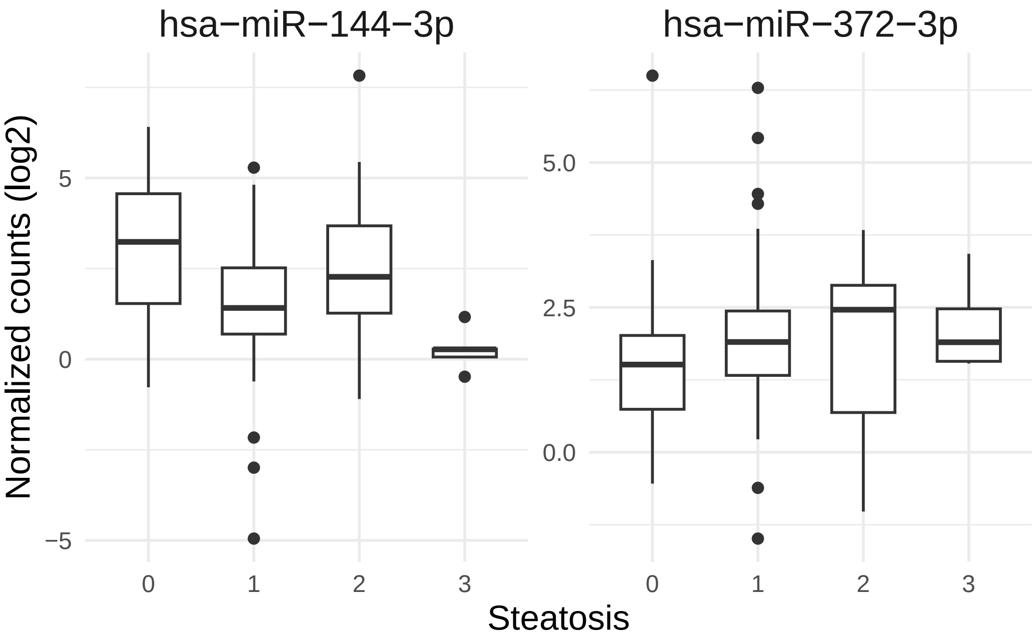

isomiRs
isomiRs abundance
pct_abundance
80
60
40
20
iso_add
iso_sn
so_5p
iso_3p
ref
uid
sexo
hombre
mujer


−Log10 p−value
0
1
2
3
−3
−2
−1
0
1
Log2 Fold Change


dispersion
0.05
0.20
1.00
1e+02
1e+03
1e+04
1e+05
1e+06
mean of normalized counts
● gene−est
● fitted
● final



x
hsa-miR-144-3p
hsa-miR-372-3p



Chapter 10

Discussion

Chapter 11

Conclusions

Chapter 12

Recommendations

lalallallalalalallalala

Appendix A

Appendices

References

Carlson, M. (2024). *Org.hs.eg.db: Genome wide annotation for human*.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*, 550. https://doi.org/10.1186/s13059-014-0550-8

Morgan, M. (2024). *BiocVersion: Set the appropriate version of bioconductor packages*. https://doi.org/10.18129/B9.bioc.BiocVersion

Pantano, L., & Escaramis, G. (2024). *isomiRs: Analyze isomiRs and miRNAs from small RNA-seq*. https://doi.org/10.18129/B9.bioc.isomiRs

Peltzer, A., Trigila, A., Pantano, L., Ewels, P., Wang, C., Espinosa-Carrasco, J., Schcolnicov, N., Mohr, C., bot, nf-core, Menden, K., Patel, H., Sturm, G., CKComputomics, Cabus, L., Keys, K. L., Guizard, S., Garcia, M. U., Syme, R., Talbot, A., … Tommaso, P. D. (2024). *Nf-core/smrnaseq: v2.4.0 - 2024-10-14 - gray zinc dalmation patch*. Zenodo. https://doi.org/10.5281/ZENODO.3456879

Posit team. (2023). *RStudio: Integrated development environment for r*. Posit Software, PBC. http://www.posit.co/

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ru, Y., Kechris, K. J., Tabakoff, B., Hoffman, P., Radcliffe, R. A., Bowler, R., Mahaffey, S., Rossi, S., Calin, G. A., Bemis, L., & Theodorescu, D. (2014). The multiMiR r package and database: Integration of microRNA target interactions along with their disease and drug associations. *Nucleic Acids Research*, *42*(17), e133. https://doi.org/10.1093/nar/gku631

Ru, Y., Mulvahill, M., Mahaffey, S., & Kechris, K. (n.d.). *multiMiR: Integration of multiple microRNA-target databases with their disease and drug associations*. https://github.com/KechrisLab/multiMiR

Xu, S., Hu, E., Cai, Y., Xie, Z., Luo, X., Zhan, L., Tang, W., Wang, Q., Liu, B., Wang, R., Xie, W., Wu, T., Xie, L., & Yu, G. (2024). Using clusterProfiler to characterize multiomics data. *Nature Protocols*. https://doi.org/10.1038/s41596-024-01020-z