

DMD vs WT Single-End RNA-seq Analysis Pipeline Using Nextflow DSL2

Molecular Biology and Biotechnology Master's Programme
Assignment Report

Student: Berfin Nida Öztürk

Date: 25 May 2026

Contents

- 1 1. Introduction
- 2 2. Methods
- 3 3. Results
- 4 4. Supplementary Items (SI)

1. Introduction

Duchenne muscular dystrophy (DMD) is a progressive neuromuscular disorder caused by mutations in the dystrophin gene. RNA-seq provides a robust approach to investigate transcriptomic differences between DMD and wild-type (WT) muscle samples. In this study, a modular and reproducible workflow was implemented with Nextflow DSL2 for single-end WT vs DMD Δ Ex51 mouse tibialis anterior muscle RNA-seq data.

The selected public dataset was GSE156496 / SRP278118 (mouse DMD Δ Ex51 model RNA-seq dataset), including two WT and two DMD Δ Ex51 samples. The aim was to generate a reproducible end-to-end pipeline from raw FASTQ files to a transcript-level expression matrix.

2. Methods

2.1 Dataset and sample selection

- WT1 = SRR12478073 = GSM4732625 = TA_WT_1 RNA-seq
- WT2 = SRR12478074 = GSM4732626 = TA_WT_2 RNA-seq
- DMD1 = SRR12478076 = GSM4732628 = TA_D51_4 RNA-seq (Delta 51)
- DMD2 = SRR12478077 = GSM4732629 = TA_D51_5 RNA-seq (Delta 51)

2.2 Reference preparation

Mus musculus GRCm39 Ensembl cDNA transcriptome FASTA was used as reference. Salmon index was built with k=31.

2.3 Pipeline steps

- 1 FastQC raw read quality control
- 2 fastp single-end read trimming and filtering
- 3 Salmon transcript quantification
- 4 MultiQC report generation
- 5 Custom Python/pandas expression matrix generation

2.4 Outputs

- results/fastqc/
- results/fastp/
- results/salmon/
- results/multiqc/multiqc_report.html
- results/matrix/expression_matrix.tsv

3. Results

The pipeline completed successfully with exit code 0. All workflow processes, including MultiQC and EXPR_MATRIX, completed without unresolved errors.

The expression matrix was generated successfully at results/matrix/expression_matrix.tsv.
The first columns are: Name, DMD1, DMD2, WT1, WT2.

A descriptive TPM-based comparison was added to compute DMD mean TPM, WT mean TPM, and log2FC using pseudocount 1. This is a descriptive summary and not a formal statistical differential expression test.

4. Supplementary Items (SI)

- GitHub repository: <https://github.com/berfinida/pipeline>
- GEO: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE156496>
- SRA project: <https://www.ncbi.nlm.nih.gov/sra?term=SRP278118>
- Runs: <https://www.ncbi.nlm.nih.gov/sra/SRR12478073>,
<https://www.ncbi.nlm.nih.gov/sra/SRR12478074>,
<https://www.ncbi.nlm.nih.gov/sra/SRR12478076>,
<https://www.ncbi.nlm.nih.gov/sra/SRR12478077>
- Run command: `nextflow run main.nf -profile conda --samplesheet samplesheet.csv --outdir results --salmon_index "$PWD/refs/salmon_index" -resume`
- Software: Nextflow DSL2, FastQC, fastp, Salmon, MultiQC, Python (pandas)
- Raw FASTQ files and Salmon index are excluded from GitHub due to file size