Exploring the Impact of RNA-Seq Pipeline Choices on Differential Expression Variance

Hunter Schuler, Art Tay

March 31st, 2025

Introduction

RNA sequencing (RNA-Seq) is a powerful technique for studying gene expression. However, the results of RNA-Seq analyses are highly sensitive to the choices made throughout the processing pipeline, including alignment methods, read quality thresholds, data normalization techniques, and even the options used during the raw data download step. These choices can significantly influence downstream analyses, particularly in differential expression (DE) analysis, which is a key step in understanding biological differences between conditions.

The aim of this project is to explore how various discretionary decisions made during RNA-Seq data processing impact the variance of differential expression results. In particular, we are interested in how factors such as read extraction options, the choice of aligner, the filtering criteria applied to the data, and normalization methods influence the consistency of DE results across multiple quality thresholds.

Question of Interest

> How do discretionary choices made during RNA-Seq pipeline processing, such as ‘fasterq-dump‘ options, quality filtering threshold, the choice of aligner, and normalization method impact the variance of differential expression results?

We hypothesize that differences in these choices will lead to significant variance in DE results, particularly in terms of how consistently differentially expressed genes are identified across pipeline variations. This variance could introduce substantial uncertainty into the interpretation of gene expression data, influencing biological conclusions.

Ideas for Exploration

To address this question, we will explore the following factors throughout the RNA-Seq processing pipeline:

1. **fasterq-dump Options:** The extraction of FASTQ data from raw sequencing data using different ‘fasterq-dump‘ options can introduce variability into the input data quality. We will test multiple ‘fasterq-dump‘ settings and investigate their effect on read quality and subsequent alignment performance.

2. **Quality Thresholds:** The filtering of raw reads based on quality thresholds can significantly impact the amount of data available for downstream analysis. We will explore how adjusting these thresholds affects alignment results and how this, in turn, affects the variance in differential expression analyses.

3. **Choice of Aligner:** Different RNA-Seq aligners (e.g., STAR, Rsubread, HISAT2, etc.) use different algorithms, which may lead to varying alignment results, affecting downstream analyses such as differential expression. We will compare the alignment accuracy of these tools and assess how their output impacts the variance of DE results.

4. **Normalization Methods:** Various methods for normalizing RNA-Seq data (e.g., TPM, FPKM, RPKM, TMM) can lead to different interpretations of gene expression levels. We will examine how each normalization method affects the variance in DE results and whether any method consistently reduces the variance across pipeline variations.

Unique Contribution: This project’s unique contribution lies in its comprehensive exploration of how RNA-Seq processing choices impact the variance in differential expression results. By systematically varying multiple stages of the RNA-Seq pipeline, we will identify (or rule out) key factors that contribute to inconsistency in downstream analyses, providing insight into how to optimize RNA-Seq pipelines for robust and reliable results.

Preliminary Results

As part of our preliminary analysis, we used the Ovarian Cancer RNA-Seq dataset (SRR31476642.fastq) to test two different aligners, STAR and Rsubread, at various quality thresholds. These were tested at minimum average (per read) Phred score levels of 25, 30, and 39. For both tools, we compared the number of aligned reads to the number of filtered reads, observing that Rsubread aligned nearly 100% of filtered reads across all thresholds, while STAR aligned only about 80%. This result suggests potential differences in alignment sensitivity between the two tools, which could influence downstream differential expression analysis.

We also applied different quality thresholds and normalization methods, and visualized the effect of these factors on the variance of the differential expression results. These preliminary results indicate that variations in the RNA-Seq pipeline choices lead to noticeable differences in the DE results, with some choices (e.g., quality filtering) contributing more to the variance than others.

We are leveraging the computational power of SMU’s High-Performance Computing system, ManeFrame III, to efficiently run parallelized RNA-Seq pipeline analyses. Our preliminary results utilize a SLURM script to create pairwise combinations of task parameter values. These combinations are then processed concurrently across multiple compute nodes, with each task invoking an R script that performs the necessary alignment and downstream analysis. By distributing the computational load across multiple processors, we significantly reduce the time required to process large RNA-Seq datasets, allowing us to explore a wide range of pipeline configurations and their effects on differential expression variance.

The following figure shows the preliminary results of alignment accuracy between Rsubread and STAR at different quality thresholds:

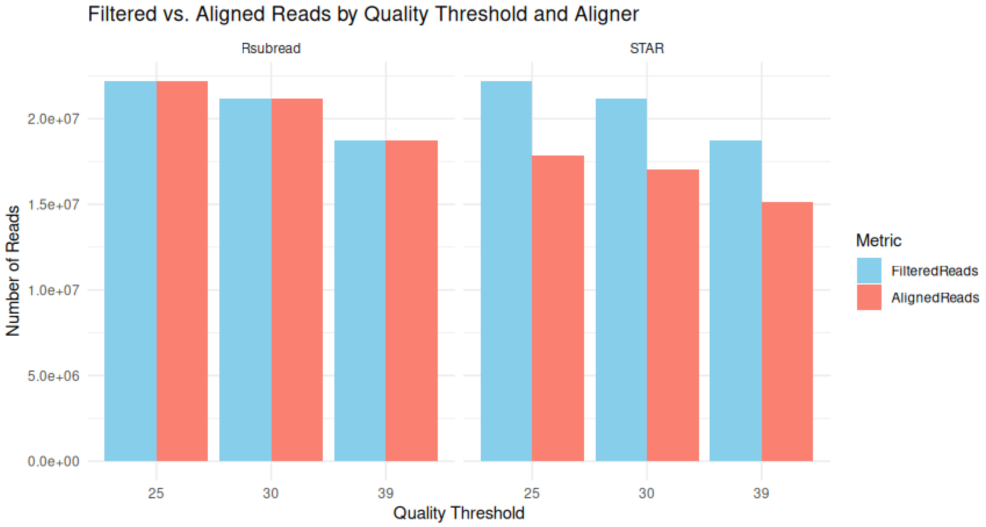

Figure 1: Alignment Accuracy of Rsubread and STAR at Different Quality Thresholds.

While these preliminary results suggest alignment sensitivity differences between Rsubread and STAR, further analysis will be conducted using additional aligners such as Salmon, HISAT2, and Kallisto, as well as varying read extraction options and normalization methods to assess the overall impact on differential expression variance.

Planned Next Steps

The next steps in the project include:

1. **Impact of ‘fasterq-dump‘ Options:** We will further investigate how variations in ‘fasterq-dump‘ settings influence the data quality and impact downstream analyses.

2. **Testing Additional Aligners:** We plan to expand our analysis to include other RNA-Seq aligners (e.g., HISAT2, Kallisto, and Salmon) to assess their performance across the various factors we are investigating.

3. **Exploring More Quality Thresholds:** Additional quality thresholds will be tested to better understand how stringent filtering criteria affect differential expression results. This will give us a more granular understand of the effect that threshold selection has.

4. **Comparing Different Normalization Methods:** We will explore different normalization methods (e.g., TPM, TMM) to evaluate how they impact the consistency of DE results across different pipeline choices.

Conclusion

The impact of RNA-Seq pipeline choices on differential expression analysis is an important, but often overlooked, aspect of RNA-Seq data analysis. This project aims to systematically explore how variations in aligners, quality thresholds, normalization methods, and raw data extraction options influence the variance of DE results. By identifying key sources of variability in RNA-Seq workflows, we aim to optimize RNA-Seq pipelines for more reproducible and robust differential expression analysis. The results of this study will provide important guidance for researchers in the RNA-Seq field, helping them make more informed decisions about the tools and settings used in their analyses.