Variance All the Way Down: Quantifying the Uncertainty Introduced at each Stage of an End-to-End RNA-Seq Analysis

Art Tay

Abstract

In the realm of RNA-Seq research, rigorous data preprocessing is a critical foundation for meaningful analysis. Despite its importance, this preprocessing involves numerous stages, each introducing potential sources of variance. While previous studies have examined the overall variance across entire RNA-Seq pipelines, (Arora et al. 2020) (Tong et al. 2020), (Vieth et al. 2019), the impact of individual stages remains less understood. We propose a comprehensive investigation into the variance introduced at each stage of RNA-Seq preprocessing. Our goal is to quantify these variances, study their distributions, and understand their statistical implications on downstream modeling. This will include exploring the multitude of decisions researchers face — from quality control to normalization and feature selection — and evaluating how these choices propagate uncertainty through the analysis. Of particular interest is whether variance amplifies due to interactions between decisions made at different stages. By modeling these interactions, we aim to identify cases where suboptimal combinations of preprocessing choices exacerbate variability, potentially distorting biological interpretations. Finally, we will assess various bias correction methods and uncertainty quantification strategies to incorporate into final models. This work aims to provide researchers with actionable insights and robust statistical tools to mitigate preprocessing-induced variance, ultimately enhancing the reliability and reproducibility of RNA-Seq studies.

Preliminary Results

Preliminary Methodology Section

Table 1: Basic RNA-Seq Differential Analysis End-to-End Pipeline

Pipeline Steps	Software	Options	Choices
1. Pull SRA data from the NIH.	prefetch	NA	NA
2. Compute quality scores.	fasterq-dump	–skip-technical –threads X	Boolean Integer
3. Filter low quality reads.	fastp	–qualified_quality_phred X –length_required X	Integer Integer
4. Trim excess bases.	fastp	–trim_poly_g –trim_ploy_x	Boolean Boolean
5. Align reads to a genome.	Various	Default	STAR, HISAT2 Salmon, Kallisto
6. Count genes.	featureCounts	-Q (minimum map quality.)	Integer

		-C (require pair agreement.) -M (allow multi-map.)	Boolean Boolean
7. Count normalization.	edgeR	calcNormFactors(method='X')	TMM, RLE, upperquartile
8. Differential expression analysis.	edgeR	Default	NA

Quality Score Variance Due to Fasterq-dump Options

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("ShortRead")
BiocManager::install("Rsubread")
```

```
library(ShortRead)

sample_1_fq_1 <- readFastq("./data/dump_1/SRR31476642.fastq")
sample_1_fq_2 <- readFastq("./data/dump_2/SRR31476642.fastq")
sample_1_fq_3 <- readFastq("./data/dump_3/SRR31476642.fastq")
```

```
sample_1_fq_1_qual <- as(quality(sample_1_fq_1), "matrix")
sample_1_fq_2_qual <- as(quality(sample_1_fq_2), "matrix")
sample_1_fq_3_qual <- as(quality(sample_1_fq_3), "matrix")

sample_1_fq_13_qual_diff <- sample_1_fq_1_qual - sample_1_fq_3_qual
sample_1_fq_12_qual_diff <- sample_1_fq_1_qual - sample_1_fq_2_qual
```

```
mean(sample_1_fq_13_qual_diff)
mean(sample_1_fq_12_qual_diff)
```

Comparing Alignment Accuracy

```
fastq_files <- list.files(
    path = "./data/dump_1", pattern = "\\.fastq$", full.names = TRUE
)
```

```
library(Rsubread)

buildindex(basename="hg19_g1k",
           reference="./data/human_g1k_v37.fasta",
           memory=3600
)

align_reads <- function(file, index_base, output_dir) {
  align(
    index = index_base,
    readfile1 = file,
```

```
    output_file = file.path(output_dir, paste0(basename(file), ".bam")),
    nthreads = 4
  )
}
```

```
trim_reads <- function(file, quality_threshold = 20, min_length = 30) {
    fq <- readFastq(file)
    fq_filtered <- fq[
        alphabetScore(quality(fq)) >= quality_threshold & width(fq) >= min_length
    ]
    output_file <- sub(".fastq", "_trimmed.fastq", file)
    writeFastq(fq_filtered, output_file, compress = FALSE)
}
```


Arora, S., Pattwell, S. S., Holland, E. C., and Bolouri, H. (2020), "Variability in estimated gene expression among commonly used RNA-seq pipelines," *Scientific reports*, Nature Publishing Group UK London, 10, 2734.

Tong, L., Wu, P.-Y., Phan, J. H., Hassazadeh, H. R., Tong, W., and Wang, M. D. (2020), "Impact of RNA-seq data analysis algorithms on gene expression estimation and downstream prediction," *Scientific reports*, Nature Publishing Group UK London, 10, 17925.

Vieth, B., Parekh, S., Ziegenhain, C., Enard, W., and Hellmann, I. (2019), "A systematic evaluation of single cell RNA-seq analysis pipelines," *Nature communications*, Nature Publishing Group UK London, 10, 4667.