Salmon vs STAR (DESeq2)

Tuesday, 17 February 2026 15:08

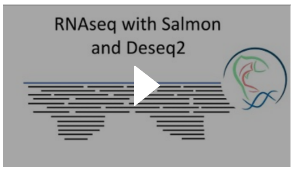
1) Introduction

The main objective of this analysis is to compare the impact of different read quantification strategies on downstream exploratory data analysis.

Two distinct workflows were employed to generate the counts for DESeq2:

- **Workflow A (Salmon)**: Fast transcript-level quantification followed by gene-level aggregation.

  RNAseq mapping with Salmon for differential expression



- **Workflow B (STAR)**: Traditional splice-aware genomic alignment followed by raw read counting

  https://ycl6.gitbook.io/guide-to-rna-seq-analysis/differential-expression-analysis/differential-gene-expression/dge-analysis-with-star-input

To ensure a controlled comparison, both workflows utilised the exact same sample metadata and final DESeq2 design formula.

2) Methods and Data Import

The technical implementation of the DESeqDataSet (dds) object differed based on the input format:

- **Salmon Output**: Since Salmon provides estimated transcript abundances (often non-integers), the tximport package was used to aggregate these to the gene level. The dds object was constructed using DESeqDataSetFromTximport().
- **STAR Output**: STAR provides an integer-based count matrix. These raw counts were imported directly into R, and the dds object was constructed using the standard countData matrix via DESeqDataSetFromMatrix().

3) Exploratory Data Analysis (PCA Comparison)

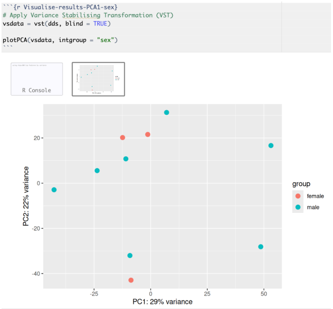
Principal Component Analysis (PCA) was performed on both datasets after applying a Variance Stabilising Transformation (vst).

PC1 (The x-axis): This is the direction of the highest variance in your data. If your samples separate into distinct groups from left to right, the factors causing that separation are the most important drivers of gene expression in your experiment.

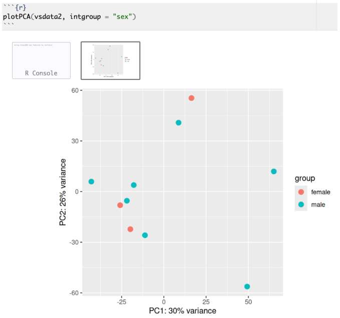
PC2 (The y-axis): This represents the second highest variance. It captures variation that is completely separate (perpendicular) to what PC1 found.

Salmon

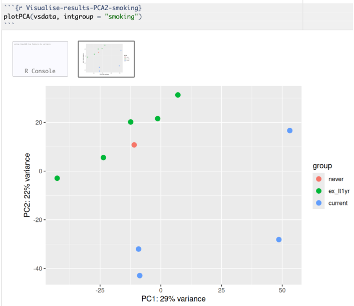
STAR



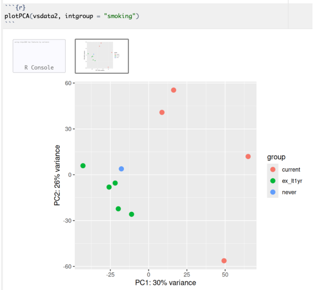
~~~
```{r Visualise-results-PCA1-sex}
# Apply Variance Stabilising Transformation (VST)
vsdata = vst(dds, blind = TRUE)

plotPCA(vsdata, intgroup = "sex")
```
~~~



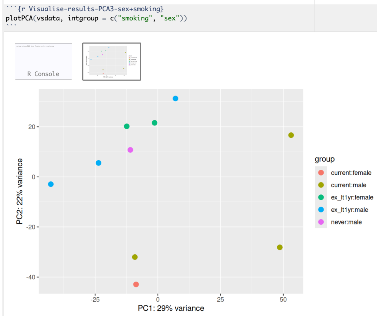
~~~
```{r}
plotPCA(vsdata2, intgroup = "sex")
```
~~~



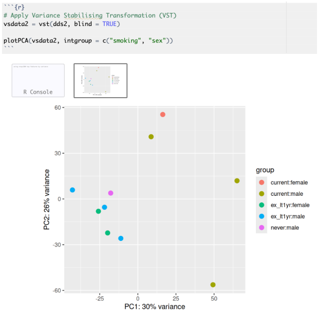
~~~
```{r Visualise-results-PCA2-smoking}
plotPCA(vsdata, intgroup = "smoking")
```
~~~



~~~
```{r}
plotPCA(vsdata2, intgroup = "smoking")
```
~~~



~~~
```{r Visualise-results-PCA3-sex+smoking}
plotPCA(vsdata, intgroup = c("smoking", "sex"))
```
~~~



~~~
```{r}
# Apply Variance Stabilising Transformation (VST)
vsdata2 = vst(dds2, blind = TRUE)

plotPCA(vsdata2, intgroup = c("smoking", "sex"))
```
~~~

What can we say from our plots:

- PC1 and PC2 explain almost the same variance in both pipelines
- Biological signals (sex, smoking) are broadly consistent.
- STAR shows slightly tighter clustering and a bit stronger separation, while Salmon looks a bit noisier with more spread/outliers. Alignment based vs quasi-mapping is not changing the conclusion.

1. **PCA coloured by sex**

Salmon

- Males and females overlap substantially
- There's no clean separation along PC1 or PC2.
- A couple of samples are quite spread out, PC1

STAR

- Still overlap, but:
  - The cloud is more compact.
  - Sex groups look slightly more structured, especially along PC2.

Interpretation:
Sex contributes some variation, but it's secondary. STAR captures this a bit more cleanly; Salmon introduces slightly more dispersion.

2. **PCA coloured by smoking status**

Salmon

- Smoking groups (never / ex / current) show partial separation.
- Current smokers look more spread out, especially along PC1.
- There are a couple of strong outliers.

STAR

- Separation by smoking is clearer, particularly:
  - Current smokers shift along PC1.
  - Never and ex-smokers cluster more tightly.
- Less scatter overall.

Interpretation:
Smoking is a stronger signal than sex, and STAR makes this more obvious. Salmon still captures it, but with more noise.

3. **PCA coloured by sex + smoking combined**

Salmon

- Groups are distinguishable, but:
  - Considerable overlap between combinations.
  - Larger spread within groups.
- One or two samples dominate the extremes of PC1.

STAR

- Same biological pattern as Salmon, but:
  - Tighter clusters
  - Cleaner gradients across PC1/PC2

Interpretation:
Both pipelines agree on the relative relationships between groups, but STAR gives a cleaner structure.

Conclusion:
In both pipelines, smoking status explained more variance than sex, while sex alone did not strongly separate samples. STAR-based quantification resulted in tighter clustering and slightly clearer group separation, whereas Salmon exhibited increased dispersion, likely reflecting transcript-level variability