# **Variance All the Way Down:** Exploring the Impact of RNA-Seq Pipeline Choices on Differential Expression Variance

Hunter Schuler and Art Tay

## Analysis

### NIH Baseline

```
# Meta-Data from:
# https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA1189593&o=acc_s%3Aa

sample_names <- c(
  "gene",
  "SRR31476642",
  "SRR31476643",
  "SRR31476644",
  "SRR31476645",
  "SRR31476646",
  "SRR31476647",
  "SRR31476648",
  "SRR31476649",
  "SRR31476650"
)

treatments <- c(
    "DMSO",
    "DMSO",
    "DMSO",
    "DMSO",
    "EPZ015666",
    "EPZ015666",
    "EPZ015666",
    "DMSO",
    "DMSO"
)
```

```
nih_count_matrix <- read.csv("./data/GSE282674_ovcar4_count_table.csv")
colnames(nih_count_matrix) <- sample_names
nih_count_matrix <- nih_count_matrix |>
    mutate(gene = str_remove(gene, "\\..*$"))

nih_count_vector <- nih_count_matrix |>
    pivot_longer(cols = -gene, names_to = "sample", values_to = "count")

nih_dgelist <- DGEList(
    counts = nih_count_matrix[, -1],
```

```
    genes = nih_count_matrix$gene,
    group = as.factor(treatments)
)
nih_dgelist <- nih_dgelist[filterByExpr(nih_dgelist), , keep.lib.sizes = FALSE]
nih_dgelist <- estimateDisp(nih_dgelist)

design <- model.matrix(~as.factor(treatments))
nih_fit <- glmFit(nih_dgelist, design)
nih_LRT <- glmLRT(nih_fit, coef = 2)

nih_p_values <- topTags(nih_LRT, n = Inf)$table
nih_p_values <- data.frame(
    gene = nih_p_values$genes,
    p_value = nih_p_values$PValue
)
```

### Salmon Test Case

```
salmon_test <- read.csv(
    "./data/salmon/salmon_count_matrices/gene_count_matrix_Q21_L32_G1_X1.csv"
)
colnames(salmon_test) <- sample_names

salmon_test_long <- salmon_test |>
    pivot_longer(cols = -gene, names_to = "sample", values_to = "count")

# Compute count variance.
joined_counts <- full_join(
    salmon_test_long, nih_count_vector,
    by = join_by(gene, sample)
) |> mutate(across(where(is.numeric), ~replace_na(., 0)))

count_var <- sum((joined_counts$count.x - joined_counts$count.y)^2)
count_sd <- sqrt(count_var / nrow(joined_counts))
```

```
# Compute p_value standard deviation.
salmon_dgelist <- DGEList(
    counts = salmon_test[, -1],
    genes = salmon_test$gene,
    group = as.factor(treatments)
)

salmon_dgelist <- salmon_dgelist[
    filterByExpr(salmon_dgelist), , keep.lib.sizes = FALSE
]
salmon_dgelist <- estimateDisp(salmon_dgelist)

design <- model.matrix(~as.factor(treatments))
salmon_fit <- glmFit(salmon_dgelist, design)
salmon_LRT <- glmLRT(salmon_fit, coef = 2)

salmon_p_values <- topTags(salmon_LRT, n = Inf)$table
salmon_p_values <- data.frame(
```

```
    gene = salmon_p_values$genes,
    p_value = salmon_p_values$PValue
)

joined_p_values <- full_join(
    salmon_p_values, nih_p_values, by = join_by(gene)
) |> mutate(across(where(is.numeric), ~replace_na(., 1)))

p_value_var <- sum((joined_p_values$p_value.x - joined_p_values$p_value.y)^2)
p_value_sd <- sqrt(p_value_var / nrow(joined_p_values))
```