# Deeper differential expression analysis with shrinkage correction

Jared Brown

Postdoctoral research fellow

Lab of Rafael Irizarry · DFCI Data Science

20 May 2024

# Follow-along preliminaries

## Software:

- R: https://cran.r-project.org/
- Rstudio: https://posit.co/download/rstudio-desktop/
- R packages:

```
install.packages("BiocManager")
BiocManager::install(c("DESeq2”, “pasilla”, “ggplot2”))
```

## Analysis scripts:

- https://github.com/JBrownBiostat/DifferentialExpressionTraining_May2024

Jared Brown · Postdoctoral research fellow · DFCI Lab of Rafael Irizarry
jbrown@ds.dfci.harvard.edu · github.com/JBrownBiostat

# Other resources

From the original developers, Michael Love, Simon Anders, Wolfgang Huber:
https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

General (single-cell RNA) pipelines from some greats in the field:

https://bioconductor.org/books/3.17/OSCA/

Carpentry workshop from Bioconductor:

https://carpentries-incubator.github.io/bioc-rnaseq/

# DE definitions: data structure

A $G \times J$ matrix of abundance measures across:

- $G$ many features: ***genes***, transcripts, exons, protein binding peaks, methylation sites, etc.
- $J$ many samples: ***experimental libraries***, single cells, spatial spots, binned pixels, etc.
- First few lines from the *pasilla* dataset of pasilla gene knock-down in drosophila melanogaster

| | Samp1 | Samp2 | Samp3 | Samp4 | Samp5 | Samp6 | Samp7 |
|---|---|---|---|---|---|---|---|
| FBgn0000003 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| FBgn0000008 | 92 | 161 | 76 | 70 | 140 | 88 | 70 |
| FBgn0000014 | 5 | 1 | 0 | 0 | 4 | 0 | 0 |
| FBgn0000015 | 0 | 2 | 1 | 2 | 1 | 0 | 0 |
| FBgn0000017 | 4664 | 8714 | 3564 | 3150 | 6205 | 3072 | 3334 |
| FBgn0000018 | 583 | 761 | 245 | 310 | 722 | 299 | 308 |
| FBgn0000022 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| FBgn0000024 | 10 | 11 | 3 | 3 | 10 | 7 | 5 |
| FBgn0000028 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |
| FBgn0000032 | 1446 | 1713 | 615 | 672 | 1698 | 696 | 767 |

# DESeq2 quick-start analysis

Switch over to `bulkRNA_AnalysisScript.R`

# DE definitions: what are we testing

## Most generally:

- Is the effect size of a given combination of covariates significantly non-zero

## In practice:

- Is the abundance of gene $g$ in condition $A$ significantly different than in condition $B$ controlling for appropriate nuisance variation

## Implied comparison of interest:

- Ratios measured as log (2) fold-changes

# DE definitions: typical model spec.

DESeq2 (and other models) assume observed expression is effectively modeled as a *Negative Binomial GLM*, i.e.:

$$y_{gj} \sim NB\left(\mu_{gj}, \phi_g\right)$$
$$\log\left(\mu_{gj}\right) \coloneqq \left[X\vec{\beta}\right]_{gj} + \log(s_j)$$

For observed counts $y_{gj}$, design matrix $X$, coefficient vector $\vec{\beta}$, normalization factor/offset $s_j$, and dispersion parameter $\phi_g$.

# NB GLM: Comp. to linear regression

Traditional "simple linear regression" can be written as a Normal GLM:

$$y_{gj} = \left[X\vec{\beta}\right]_{gj} + \epsilon_{gj} \;\; s.t. \;\; \epsilon_{gj} \sim N\left(0, \sigma_g^2\right)$$

is equivalent to:

$$y_{gj} \sim N\left(\mu_{gj}, \sigma_g^2\right)$$
$$\mu_{gj} \coloneqq \left[X\vec{\beta}\right]_{gj}$$

# NB GLM: Link function

Identity link (Normal):

$$\mu_{gj} = \left[X\vec{\beta}\right]_{gj} = x_{g0}\beta_0 + x_{g1}\beta_1 \cdots + x_{gK}\beta_K$$

Log link (Negative Binomial):

$$\log\left(\mu_{gj}\right) = \left[X\vec{\beta}\right]_{gj} + \log(s_j)$$
$$\Rightarrow \mu_{gj} = (e^{x_{g0}\beta_0} e^{x_{g1}\beta_1} \ldots e^{x_{gK}\beta_K})s_j$$

Action of log link indicates data $Y$ should be raw expression and not normalized expression

# (Pre) Calculation of normalization offset

In DESeq2 (and other methods) normalization factors $s_j$ are offsets:

- $s_j$ are calculated prior to model estimation (fixed)
- $s_j$ have an implicit fixed coefficient $\beta_s = 1$

Popular norm. factor calculation methods include:

1. Reads-per-kilobase-million (RPKM)
2. Transcripts/Counts-per-million (TPM/CPM)
3. Library-size normalization
4. Trimmed-mean-of-m-values (TMM, default in edgeR)

# DESeq2 normalization default

Method of *Median of Ratios* and *Scran*:

- Median of Ratios (MR) from Anders and Huber 2010
- Scran based on Lun, Bach, and Marioni 2016

Both methods based on the median of ratios relationship:

$$\hat{s}_j := median_g\left\{\frac{\widehat{y_{gj}}}{\left(\prod_g^G \widehat{y_{gj}}\right)^{\frac{1}{G}}}\right\} = median_g\left\{\frac{\widehat{y_{gj}}}{\exp\left[\frac{1}{G}\sum_g^G \log\left(\widehat{y_{gj}}\right)\right]}\right\}$$

For MR, $\widehat{y_{gj}} = y_{gj}$ and $s_j = \hat{s}_j$

For Scran, $\widehat{y_{gj}}$ is based on a pooling of cells and $s_j$ is derived from a deconvolution of (multiple) estimates $\hat{s}_j$ that are based on sample $j$

# DESeq2 bulk custom size factors

Switch over to `bulkRNA_AnalysisScript.R`

Jared Brown · Postdoctoral research fellow · DFCI Lab of Rafael Irizarry
jbrown@ds.dfci.harvard.edu · github.com/JBrownBiostat

# NB GLM: Dispersion estimates

Dispersion is analogous to variance, specifically:

$$y \sim NB(\mu, \phi) \Rightarrow \mathbb{E}[y] = \mu; \ \mathbb{V}[y] = \mu + \phi\mu^2$$

In typical bulk experiments, replicate counts are low:

- High standard error on fitted $\phi$
- Reduced power to identify significant $\beta_{gk}$

So pool information across genes with similar expression levels...

*Note:* MLE of ϕ only exists when sample variance is greater than sample mean...

# DESeq2 dispersion visualization

Switch over to `bulkRNA_AnalysisScript.R`