Dana-Farber Cancer Institute | Department of Data Science

Deeper differential expression analysis with shrinkage correction

Jared Brown
Postdoctoral research fellow
Lab of Rafael Irizarry · DFCI Data Science

21 May 2024

Follow-along preliminaries

Dana-Farber Cancer Institute | Department of Data Science

Software:

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

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

`DESeq2, pasilla, ggplot2, matrixStats, apeglm, ashr, TENxPBMCData, Matrix, irlba, scran, scater`

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/

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

DE definitions: data structure

Dana-Farber Cancer Institute | Department of Data Science

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

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

DESeq2 quick-start analysis

Switch over to `bulkRNA_AnalysisScript.R`

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

DE definitions: what are we testing

Dana-Farber Cancer Institute | Department of Data Science

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

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

DE definitions: typical model spec.

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

$$y_{gj} \sim NB\left(\mu_{gj}, \phi_g\right)$$
$$\log\left(\mu_{gj}\right) := \vec{\beta}_g^T [X]_j + \log(s_j)$$

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

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

NB GLM: Comp. to linear regression

Dana-Farber Cancer Institute | Department of Data Science

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

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

is equivalent to:

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

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

NB GLM: Link function

Dana-Farber Cancer Institute | Department of Data Science

Identity link (Normal):

$$\mu_{gj} = \vec{\beta}_g^T [X]_j = \beta_{g0} x_j + \beta_{g1} x_j \cdots + \beta_{gK} x_j$$

Log link (Negative Binomial):

$$\log(\mu_{gj}) = \vec{\beta}_g^T [X]_j + \log(s_j)$$
$$\Rightarrow \mu_{gj} = (e^{\beta_{g0} x_j} e^{\beta_{g1} x_j} \dots e^{\beta_{gK} x_j}) s_j$$

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

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

(Pre) Calculation of normalization offset

Dana-Farber Cancer Institute | Department of Data Science

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)

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

DESeq2 normalization default

Dana-Farber Cancer Institute | Department of Data Science

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

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

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

Dana-Farber Cancer Institute | Department of Data Science

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 ϕ
- Reduced power to identify significant β_{gk}

So pool information across genes with similar expression levels…

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

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

DESeq2 dispersion visualization

Dana-Farber Cancer Institute | Department of Data Science

Switch over to `bulkRNA_AnalysisScript.R`

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

Multiple covariates

Standard multiple-regression is encoded in the design matrix X as:

$$\log(\mu_{gj}) = \vec{\beta}_g^T [X]_j + \log(s_j)$$

- Factors encoded in usual "dummy variable" format
- Default model specification chooses first factor as intercept
- For three group model, encoded as:

$$\log(\mu_{gj}) = \beta_{g1}x_j + \beta_{g2}x_j + \beta_{g3}x_j = \beta_{g1} \mid j \in \mathcal{A}$$
$$\log(\mu_{gj}) = \beta_{g1}x_j + \beta_{g2}x_j + \beta_{g3}x_j = \beta_{g1} + \beta_{g2} \mid j \in \mathcal{B}$$
$$\log(\mu_{gj}) = \beta_{g1}x_j + \beta_{g2}x_j + \beta_{g3}x_j = \beta_{g1} + \beta_{g3} \mid j \in \mathcal{C}$$

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

Contrasts and complex comparisons

We can use "contrasts" or *linear combinations of coefficients* to test non-default or complex hypotheses

- Typical DESeq2 reduces general *contrasts* to the difference between the sums of two groups of coefficients (general linear combination possible):

$$c = \sum_{k \in \mathcal{C}_1} \beta_k - \sum_{k \in \mathcal{C}_2} \beta_k$$

$$H_0 : c = 0 \quad H_1 : c \neq 0$$

- Equivalent to testing whether fold-change (ratio) between groups is different from 1; using groups from before:

$$\frac{\mu_{gB}}{\mu_{gC}} = \frac{e^{\beta_1+\beta_2}}{e^{\beta_1+\beta_3}} = e^{(\beta_1+\beta_2)-(\beta_1+\beta_3)} = e^{\beta_2-\beta_3} \Rightarrow c = \beta_2 - \beta_3$$

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

DESeq2 design matrix and contrasts

Switch over to `bulkRNA_AnalysisScript.R`

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

Shrinkage and calling sig. features

Goals:

- Maximize power
- Maintain control on FDR (or something like it…)
- Improve interpretability/reliability

Note 1: traditional methods for controlling FDR tend to be under-powered

Note 2: leveraging extra information can boost power

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

q-value and local false discovery rate

q-value (Corollary 2, J. Storey 2003) for statistic t and rejection region Γ_α:

$$q(t) \coloneqq \inf_{\{\Gamma_\alpha : t \in \Gamma_\alpha\}} \mathbb{P}(H = 0 | T \in \Gamma_\alpha)$$

lfdr for a typical null hypothesis ($\beta_j = 0$):

$$lfdr_j \coloneqq \mathbb{P}(\beta_j = 0 | \hat{\beta}_j, \hat{s}_j, \hat{\pi})$$

For a particular set of observed effect sizes, ordered by lfdr,

$$q(\hat{\beta}_{(j)}) = \frac{1}{j} \sum_{i=1}^{j} lfdr_{(i)}$$

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

Local false sign rate and s-value

lfsr (Eqn. 2.7 from M. Stephens 2017):

$$lfsr_j := \min\left[\mathbb{P}\left(\beta_j \geq 0 \middle| \hat{\pi}, \hat{\beta}, s\right), \mathbb{P}\left(\beta_j \leq 0 \middle| \hat{\pi}, \hat{\beta}, s\right)\right]$$

Tukey (1991): "All we know about the world teaches us that the effects of A and B are always different – in some decimal place – for any A and B"

For a particular set of observed effect sizes, ordered by lfsr,

$$s\left(\hat{\beta}_{(j)}\right) = \frac{1}{j}\sum_{i=1}^{j} lfsr_{(i)}$$

In practice lfsr can be more powerful than lfdr in the sense that calculated lfsr is closer to the true lfsr while still being conservative

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

Adaptive SHrinkage (ASH) model (1)

Model from *M. Stephens 2017*

Consider *true* effects (model coefficients) $\beta = (\beta_1, \dots, \beta_J)$ and observed effect sizes $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_J)$ with corresponding estimated standard errors $\hat{s} = (\hat{s}_1, \dots, \hat{s}_J)$

We can conduct hypothesis testing based on the model of the true effects:

$$\mathbb{P}(\beta | \hat{\beta}, \hat{s}) \propto \mathbb{P}(\beta | \hat{s}) \mathbb{P}(\hat{\beta} | \beta, \hat{s})$$

Model the conditional distribution on true effects and observed effects as:

$$\mathbb{P}(\beta | \hat{s}, \pi) = \prod_j g(\beta_j | \pi)$$

$$g(\cdot | \pi) = \pi_0 \delta_0(\cdot) + \sum_{k=1}^{K} \pi_k N(\cdot | 0, \sigma_k^2)$$

$$\mathbb{P}(\hat{\beta} | \beta, \hat{s}) = \prod_j N(\hat{\beta}_j | \beta_j, \hat{s}_j^2)$$

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

Adapative SHrinkage (ASH) model (2)

Integrating out β reveals a convolution of normal such that:

$$\mathbb{P}(\hat{\beta}|\hat{s},\pi) = \prod_j \left[\pi_0 N(\hat{\beta}_j|0,\hat{s}_j^2) + \sum_{k=1}^{K} \pi_k N(\hat{\beta}_j|0,\sigma_k^2+\hat{s}_j^2)\right]$$

Note 1: π_0 is a direct estimate of the proportion of null effects

Note 2: in practice optimization penalizes to prefer large π_0

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

Adapative SHrinkage (ASH) model (3)

Estimates of π give a closed form distribution for the posterior on β:

$$\mathbb{P}(\beta|\hat{\beta},\hat{s},\pi) \propto \prod_j \left[\pi_0\delta_0(\beta_j) + \sum_{k=1}^{K} \pi_k N(\beta_j|0,\sigma_k^2)\right] N(\hat{\beta}_j|\beta_j,\hat{s}_j^2)$$

- Posterior *mean* yields shrunken estimate of effect size
- Posterior tails and point mass at 0 give lfsr/s-values

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

Shrinkage/FDR control examples

Switch over to `bulkRNA_AnalysisScript.R`

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

Some considerations for single cell

Dana-Farber Cancer Institute | Department of Data Science

Data sparsity causes normalization problems

- Use scran

Data sparsity also affects lower bound on fitted dispersion parameters

- Set `minmu = 1e-6`

High sample (cell) counts trigger outlier correction

- Set `minReplicatesForReplace = Inf`

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