DESeq2

Yoann Pradat

Supervised by P.H Cournède, D. Gautheret

November 4, 2020

MICS lab
Meta Prism

Introduction

DESeq2 Methods

- The model(s)
- Size factors estimators
- Dispersion estimators
- LFC estimators
- LFC testing

Questions

Introduction

1st paper: Anders S, Huber W: **Differential expression analysis for sequence count data**. *Genome Biol* 2010, 11:106. (∼ 11k citations, `DESeq` R package)

2nd paper: Love, M.I., Huber, W., Anders, S. **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**. *Genome Biol* 2014, 15:550. (∼ 20k citations, `DESeq2` R package)

Vignette: `https://www.bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html`

Affiliations:

- Michael I. Love -> Dana Farber Institute, Boston.
- Wolfgang Huber, Simon Anders -> EMBL, Heidelberg.

DESeq2 Methods

DESeq2 Methods

The model(s)

Counting data with Negative Binomial

Notations

- $i = 1, \ldots, n$ denote count variables (genes).
- $j = 1, \ldots, m$ denote individuals.
- $\mathrm{K_{ij}}$ count of var i in indiv j, $\mathsf{K} = \mathsf{K}_{1:n,1:m}$ count matrix.
- X_j covariates of indiv j, $\mathsf{X} = \mathsf{X}_{1:m,1:p}$ design matrix.

Counting data with Negative Binomial

Notations

- $i = 1, \dots, n$ denote count variables (genes).
- $j = 1, \dots, m$ denote individuals.
- $\mathrm{K_{ij}}$ count of var i in indiv j, $\mathsf{K} = \mathsf{K}_{1:n,1:m}$ count matrix.
- X_j covariates of indiv j, $\mathsf{X} = \mathsf{X}_{1:m,1:p}$ design matrix.

The negative binomial $\mathrm{NegBin}(r, p)$ counts the number of failures before r successes of proba p.

$$p_{\mathrm{NegBin}(r,p)}(k) = \frac{(k+r-1)!}{k!(n-1)!} p^r (1-p)^k \tag{1}$$

Other formulation with mean and dispersion

$$p_{\mathrm{NegBin}(\mu,\alpha)}(k) = \frac{\Gamma(k+\alpha^{-1})}{\Gamma(\alpha^{-1})k!} \left(\frac{1}{1+\alpha\mu}\right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1}+\mu}\right)^k \tag{2}$$

The model(s)

DESeq2 models the counts $\mathrm{K_{ij}}$ distribution conditionally to $\mathrm{X_j}$ as

$$\boxed{\mathbb{P}_{\mathrm{K_{ij}}|\mathrm{X_j}=x_j} = \mathrm{NegBin}(\mu_{ij}, \alpha_i)} \tag{3}$$

with a logarithmic link

$$\log(\mu_{ij}) = x_j^\top \beta_i$$

or rather, in order to account for indiv-specific size factors,

$$\log\left(\frac{\mu_{ij}}{s_j}\right) = x_j^\top \beta_i \tag{4}$$

The model(s)

DESeq2 models the counts $\mathrm{K_{ij}}$ distribution conditionally to $\mathrm{X_j}$ as

$$\boxed{\mathbb{P}_{\mathrm{K_{ij}}|\mathrm{X_j}=x_j} = \mathrm{NegBin}(\mu_{ij}, \alpha_i)} \quad (3)$$

with a logarithmic link

$$\log(\mu_{ij}) = x_j^\top \beta_i$$

or rather, in order to account for indiv-specific size factors,

$$\log\left(\frac{\mu_{ij}}{s_j}\right) = x_j^\top \beta_i \quad (4)$$

Model fitting: Find **estimators**

1. $\hat{s}_{1:m}$ (size factors)
2. $\hat{\alpha}_{1:n}$ (dispersions)
3. $\hat{\beta}_{1:n}$ (log fold changes)

DESeq2 Methods

Size factors estimators

Size factors estimators

Simple estimator Let

$$K_i^R = \left(\prod_{j=1}^{m} K_{ij} \right)^{\frac{1}{m}} \tag{5}$$

Then,

$$\hat{s}_j = \underset{K_i^R \neq 0}{\operatorname{median}} \left\{ \frac{K_{ij}}{K_i^R} \right\} \tag{6}$$

Size factors estimators

Simple estimator Let

$$K_i^R = \left(\prod_{j=1}^{m} K_{ij} \right)^{\frac{1}{m}} \tag{5}$$

Then,

$$\hat{s}_j = \underset{K_i^R \neq 0}{\text{median}} \left\{ \frac{K_{ij}}{K_i^R} \right\} \tag{6}$$

Model fitting: Find **estimators**

1. $\hat{s}_{1:m}$ (size factors) ✓
2. $\hat{\alpha}_{1:n}$ (dispersions)
3. $\hat{\beta}_{1:n}$ (log fold changes)

DESeq2 Methods

Dispersion estimators

Dispersion estimators via prior

Instead of estimating directly the dispersions α_i, they have their own distribution (prior) that is to be fitted to the data (posterior).

$$\mathbb{P}_{\alpha_i} = \mathcal{LN}\left(\alpha_{\mathrm{tr}}(\bar{\mu}_i), \sigma_d^2\right) \tag{7}$$

with $\alpha_{\mathrm{tr}}(\bar{\mu}) = a_0 + \frac{a_1}{\bar{\mu}}$, $\bar{\mu}_i = \frac{1}{m}\sum_{j=1}^{m} \frac{K_{ij}}{s_j}$.

Dispersion estimators via prior

Instead of estimating directly the dispersions α_i, they have their own distribution (prior) that is to be fitted to the data (posterior).

$$\mathbb{P}_{\alpha_i} = \mathcal{LN}\left(\alpha_{\mathrm{tr}}(\bar{\mu}_i), \sigma_d^2\right) \tag{7}$$

with $\alpha_{\mathrm{tr}}(\bar{\mu}) = a_0 + \frac{a_1}{\bar{\mu}}$, $\bar{\mu}_i = \frac{1}{m}\sum_{j=1}^{m} \frac{K_{ij}}{s_j}$.

However, the dispersions are not observed. To remedy to this, authors derive initial values of dispersion α_i^{gw} that are used to estimate α_{tr} and σ_d^2.

Dispersion fitting: Find **estimators**

1. α_i^{gw} (gene wise initial estimates)
2. $\hat{\alpha}_{\mathrm{tr}}, \hat{\sigma}_d^2$ (prior dispersions)
3. α_i^{MAP} (MAP estimators)

Remark: All genes with 0 counts are excluded from further analyses.

The model matrix X

⚠ X is the **model matrix** and it is obtained from the `DESeq2DataSet` object using the formula `colData(dds)` and `design(dds)`.

Examples with `colData(dds) =` [`condition`(0, 0, 1) `type`(A, B, C)]

1. `design` $= \sim$ condition,

$$X = \begin{bmatrix} \text{intercept} & \text{condition} \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \tag{8}$$

   For this X, the model for $\hat{\mu}_i$ **is linear** (except if weights are used).

The model matrix X

⚠ X is the **model matrix** and it is obtained from the `DESeq2DataSet` object using the formula `colData(dds)` and `design(dds)`.

Examples with `colData(dds) =` $[\texttt{condition}(0,0,1) \;\; \texttt{type}(A,B,C)]$

1. `design` $=\sim$ condition,

$$\mathrm{X} = \begin{bmatrix} \text{intercept} & \text{condition} \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \tag{8}$$

   For this X, the model for $\hat{\mu}_i$ **is linear** (except if weights are used).

2. `design` $=\sim$ condition $+$ type,

$$\mathrm{X} = \begin{bmatrix} \text{intercept} & \text{condition} & \text{type} \\ 1 & 0 & A \\ 1 & 0 & B \\ 1 & 1 & C \end{bmatrix} \tag{9}$$

   For this X, the model for $\hat{\mu}_i$ **is the NegBin GLM**.

Initial gene-wise dispersion estimators

Initial estimation of gene-wise dispersions as a minimum

$$\alpha_i^{\text{init}} = \min(\alpha_i^{\text{rough}}, \alpha_i^{\text{moment}}) \tag{10}$$

with

$$\alpha_i^{\text{rough}} = \frac{1}{m-p} \sum_{j=1}^{m} \frac{(\tilde{K}_{i,j} - \tilde{\mu}_{i,j})^2 - \tilde{\mu}_{i,j}^2}{\tilde{\mu}_{i,j}^2}, \quad \begin{cases} \tilde{\mu}_{i,1:m} = \mathsf{X}\hat{\tilde{\beta}}_i \text{ (linear for all } \mathsf{X}) \\ \hat{\tilde{\beta}}_i = \operatorname{argmin}_\beta \|\widetilde{\mathsf{K}}_{\mathsf{i,1:m}} - \mathsf{X}\beta\|_2^2 \end{cases}$$

$$\alpha_i^{\text{moment}} = \frac{\sigma^2(\widetilde{\mathsf{K}}_{\mathsf{i,1:m}}) - \mu(\widetilde{\mathsf{K}}_{\mathsf{i,1:m}})\mu(\hat{\mathsf{S}}_{1:m}^{-1})}{\mu(\widetilde{\mathsf{K}}_{\mathsf{i,1:m}})^2}$$

Initial gene-wise dispersion estimators

Initial estimation of gene-wise dispersions as a minimum

$$\alpha_i^{\mathrm{init}} = \min(\alpha_i^{\mathrm{rough}}, \alpha_i^{\mathrm{moment}}) \tag{10}$$

with

$$\alpha_i^{\mathrm{rough}} = \frac{1}{m-p} \sum_{j=1}^{m} \frac{(\tilde{K}_{i,j} - \tilde{\mu}_{i,j})^2 - \tilde{\mu}_{i,j}^2}{\tilde{\mu}_{i,j}^2}, \quad \begin{cases} \tilde{\mu}_{i,1:m} = \mathsf{X}\hat{\hat{\beta}}_i \text{ (linear for all } \mathsf{X}) \\ \hat{\hat{\beta}}_i = \mathrm{argmin}_\beta \, \|\widetilde{\mathsf{K}}_{\mathsf{i,1:m}} - \mathsf{X}\beta\|_2^2 \end{cases}$$

$$\alpha_i^{\mathrm{moment}} = \frac{\sigma^2(\widetilde{\mathsf{K}}_{\mathsf{i,1:m}}) - \mu(\widetilde{\mathsf{K}}_{\mathsf{i,1:m}})\mu(\hat{\mathsf{S}}_{1:m}^{-1})}{\mu(\widetilde{\mathsf{K}}_{\mathsf{i,1:m}})^2}$$

`DESeq2` restricts by default the dispersion estimates as follows

$$\alpha_i^{\mathrm{init}} = \min(\max(10^{-8}, \alpha_i^{\mathrm{init}}), \max(10, m)) \tag{11}$$

Iterative MLE gene-wise dispersion estimators

```
Result: $\alpha_i^{\mathrm{gw}} = \alpha_i^{(T)}$
initialization $\alpha_i^{(0)} = \alpha_i^{\mathrm{init}}$;
for $t = 1, \dots, T$ do
    $\hat{\mu}_{i,1:m}^{(t)} = \begin{cases} \hat{\mathsf{S}}_{1:\mathrm{m}} \odot \tilde{\mu}_{i,1:m} & \text{if linear model} \\ \underset{\mu_{1:m}}{\mathrm{argmax}} \prod_{j=1}^{m} p_{\mathrm{NegBin}(\mu_\mathrm{j}, \widehat{\alpha}_\mathrm{i}^{(\mathrm{t}-1)})}(K_{i,j}) & \text{otherwise} \end{cases}$ ;
    $\hat{\alpha}_i^{(t)} = \begin{cases} \underset{\alpha}{\mathrm{argmax}} \frac{1}{\sqrt{\det(\mathsf{X}^\top \mathsf{W} \mathsf{X})}} \prod_{j=1}^{m} p_{\mathrm{NegBin}(\widehat{\mu}_{\mathrm{i,j}}^{(\mathrm{t})}, \alpha)}(K_{i,j}) & \text{if DESeq2 type} \\ \texttt{overdispersion}(y = \mathsf{K}_{\mathrm{i,1:m}}, \mu = \hat{\mu}_{i,1:m}^{(t)}, \mathsf{X} = \mathsf{X}) & \text{if glmGamPoi type} \end{cases}$ ;
end
if estimator $\alpha_i^{(T)}$ did not converge and $\alpha_i^{(T)} > 10^{-7}$, then

    $\alpha_i^{\mathrm{gw}} = \underset{\alpha}{\mathrm{argmax}} \frac{1}{\sqrt{\det(\mathsf{X}^\top \mathsf{W} \mathsf{X})}} \prod_{j=1}^{m} p_{\mathrm{NegBin}(\widehat{\mu}_{\mathrm{j}}^{(\mathrm{T})}, \alpha)}(K_{i,j})$   on a grid (fitDispGrid)
```

Fit the dispersion prior mean

Dispersion fitting: Find **estimators**

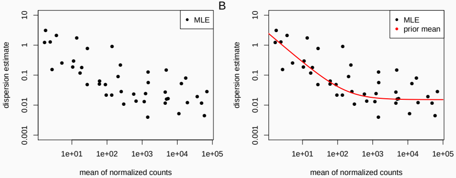
1. α_i^{gw} (gene wise initial estimates) ✓
2. $\hat{\alpha}_{\mathrm{tr}}, \hat{\sigma}_d^2$ (prior dispersions)
3. α_i^{MAP} (MAP estimators)

1. **Trend**

The prior model is

$$\mathbb{P}_\alpha = \mathcal{LN}\left(a_0 + \frac{a_1}{\bar{\mu}}, \sigma_d^2\right)$$

DESeq2 uses



$$\begin{cases} \text{10 iterations of} \quad \mathbb{P}_{\alpha|\bar{\mu}=\bar{\mu}} = \Gamma\left(a_0 + \frac{a_1}{\bar{\mu}}, \phi\right) & \text{if type=}\texttt{parametric} \quad \text{with } a_0^{(0)} = 0.1, a_1^{(0)} = 1 \\ \texttt{locfit} & \text{if type=}\texttt{locfit} \text{ or failed } \texttt{parametric} \\ \texttt{loc\_median\_fit} & \text{if type=}\texttt{glmGamPois} \end{cases}$$

⚠ Γ regression ignores points with log residual outside $[10^{-4}, 15]$.

Fit the dispersion prior mean

Dispersion fitting: Find **estimators**

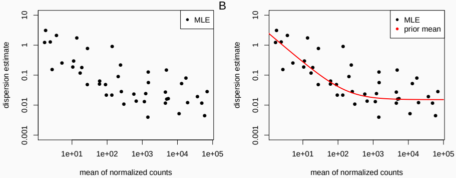
1. α_i^{gw} (gene wise initial estimates) ✓
2. $\hat{\alpha}_{\mathrm{tr}}, \hat{\sigma}_d^2$ (prior dispersions)
3. α_i^{MAP} (MAP estimators)

1. **Trend**

The prior model is

$$\mathbb{P}_\alpha = \mathcal{LN}\left(a_0 + \frac{a_1}{\bar{\mu}}, \sigma_d^2\right)$$

DESeq2 uses



$$\begin{cases} \text{10 iterations of} \quad \mathbb{P}_{\alpha|\bar{\mu}=\bar{\mu}} = \Gamma\left(a_0 + \frac{a_1}{\bar{\mu}}, \phi\right) & \text{if type=}\texttt{parametric} \quad \text{with } a_0^{(0)} = 0.1, a_1^{(0)} = 1 \\ \texttt{locfit} & \text{if type=}\texttt{locfit} \text{ or failed } \texttt{parametric} \\ \texttt{loc\_median\_fit} & \text{if type=}\texttt{glmGamPois} \end{cases}$$

⚠ Γ regression ignores points with log residual outside $[10^{-4}, 15]$.

2. **Variance**

$$\widehat{\sigma_d^2} = \max\{s_{\mathrm{lr}}^2 - \psi_1(\frac{m-p}{2}), 0.25\}$$

with s_{lr}^2 a robust estimator

$$s_{\mathrm{lr}}^2 = \operatorname*{mad}_i\{\log(\alpha_i^{\mathrm{gw}}) - \log \alpha_{\mathrm{tr}}(\bar{\mu}_i)\}$$

MAP dispersion estimators

Dispersion fitting: Find **estimators**

1. α_i^{gw} (gene wise initial estimates) ✓
2. $\hat{\alpha}_{\mathrm{tr}}, \hat{\sigma}_d^2$ (prior dispersions) ✓
3. α_i^{MAP} (MAP estimators)

MAP dispersion estimators

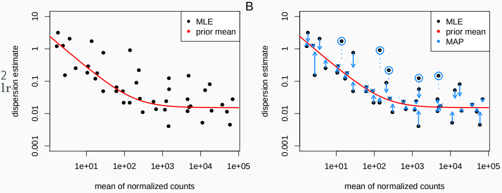
Dispersion fitting: Find **estimators**

1. α_i^{gw} (gene wise initial estimates) ✓
2. $\hat{\alpha}_{\mathrm{tr}}, \hat{\sigma}_d^2$ (prior dispersions) ✓
3. α_i^{MAP} (MAP estimators)

1. α_i^{gw} is an **outlier** if

   $\log(\alpha_i^{\mathrm{gw}}) - \log \hat{\alpha}_{\mathrm{tr}}(\bar{\mu}_i) > 2s_{\mathrm{lr}}^2$

   Then, $\alpha_i^{\mathrm{final}} = \alpha_i^{\mathrm{gw}}$.
2. Otherwise,



$$\alpha_i^{\mathrm{final}} = \underset{\alpha}{\operatorname{argmax}}\, p_{\alpha_i|\mathrm{K_{i,1:m}}=\mathsf{k}_{\mathsf{i,1:m}}}(\alpha) \propto \prod_{j=1}^{m} p_{\mathrm{K_{i,j}}|\alpha_i=\alpha}(k_{ij})\, p_{\alpha_i}(\alpha) \tag{12}$$

Model fitting: Find **estimators**

1. $\hat{s}_{1:m}$ (size factors) ✓
2. $\hat{\alpha}_{1:n}$ (dispersions) ✓
3. $\hat{\beta}_{1:n}$ (log fold changes)

DESeq2 Methods

LFC estimators

LFC model and estimation

Reminder:

$$\boxed{\mathbb{P}_{\mathrm{K_{ij}|X_j}=x_j} = \mathrm{NegBin}(s_j e^{x_j^\top \beta_i}, \alpha_i)} \tag{13}$$

As for dispersions, author set a prior on each $\beta_{i,r}$,

$$\mathbb{P}_{\beta_{i,r}} = \mathcal{N}\left(0, \sigma_r^2\right) \tag{14}$$

LFC fitting: Find **estimators**

1. β_i^{MLE} (initial estimates)
2. $\hat{\sigma}_r^2$ (prior fitting)
3. β_i^{MAP} (MAP estimators)

Prior fitting

1. Initial estimation of gene-wise LFC as a minimum

$$\beta_i^{\mathrm{MLE}} = \underset{\beta}{\operatorname{argmax}} \prod_{j=1}^{m} p_{\mathrm{NegBin}(\hat{\mathrm{s}}_{\mathrm{j}} \mathrm{e}^{\beta^{\mathrm{T}} \mathrm{x}_{\mathrm{j}}}, \hat{\alpha}_{\mathrm{i}}^{\mathrm{final}})}(K_{i,j}) \tag{15}$$

2. Variance estimator robust against LFC outliers

$$\hat{\sigma}_r = \frac{Q_{|\beta_r^{\mathrm{MLE}}|}(1-p)}{Q_N(1-p/2)}$$

p is set to 0.05 by default

LFC fitting: Find **estimators**

1. β_i^{MLE} (initial estimates) ✓
2. $\hat{\sigma}_r^2$ (prior fitting) ✓
3. β_i^{MAP} (MAP estimators)

Posterior fitting

1. **LFC final estimator**

$$\beta^{\text{final}}_{i,1:p} = \underset{\beta}{\text{argmax}} \frac{1}{\sqrt{\det(\mathsf{X}^\top \mathsf{W}\mathsf{X})}} p_{\beta_i|\mathrm{K_{i,1:m}}=\mathsf{k_{i,1:m}}}(\beta)$$

$$\propto \frac{1}{\sqrt{\det(\mathsf{X}^\top \mathsf{W}\mathsf{X})}} \prod_{j=1}^{m} p_{\mathrm{K_{i,j}}|\beta_i=\beta}(k_{ij}) p_{\beta_i}(\beta)$$

i.e

$$\beta^{\text{final}}_{i,1:p} = \underset{\beta}{\text{argmax}} \sum_{j=1}^{m} \log p_{\mathrm{NegBin}(\mu_j(\beta),\hat{\alpha}_i)}(K_{i,j}) - \frac{1}{2}\log\det(\mathsf{X}^\top \mathsf{W}\mathsf{X}) - \sum_{r=1}^{p} \frac{\beta_r^2}{2\sigma_r^2}$$

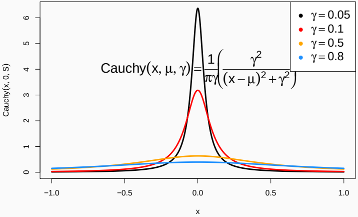
2. **Estimator of covariance LFC estimator** Also estimate the covariance matrix $\Sigma_i = \widehat{\mathrm{Cov}}\left(\beta^{\text{final}}_i\right)$ for the tests.

Other LFCs priors

Authors observed "*normal prior can sometimes produce too strong of a shrinkage*". From v1.18, additional priors may be used

1. `apeglm` adaptive t prior from the `apeglm` package (Zhu, Ibrahim and Love, Bioinformatics 2018). The prior is

$$\mathbb{P}_{\beta_{ir}} = \mathrm{Cauchy}(0, S_r) \qquad (16)$$



2. `ashr` from `ashr` package (Stephens, Biostatistics 2016). New approach to bridge the gap between **FDR** and **estimation** using "local false sign rate". Assuming there are effect and SE estimates, $\hat{\beta}_{i,1:p}$ and $\hat{S}_{i,1:p}$, ashr computes

$$p_{\beta_i|\hat{\beta}_i,\hat{S}_i}(\beta) \propto p_{\hat{\beta}_i|\beta_i,\hat{S}_i}(\hat{\beta}_i) p_{\beta_i|\hat{S}_i}(\beta)$$

with

$$p_{\hat{\beta}_i|\beta_i,\hat{S}_i}(\beta) = \prod_{r=1}^{p} \mathcal{N}(\beta_r|\beta_{i,r}, \hat{S}^2_{i,r}) \quad , \mathbb{P}_{\beta_i|\hat{S}_i} = \pi_{0,i}\delta_0 + \sum_{k=1}^{K} \pi_{k,i}\mathcal{N}(0, \sigma^2_{i,k}).$$

DESeq2 Methods

LFC testing

Wald testing

1. Using the LFC estimator and the estimation of the covariance of this LFC estimator, one may form Wald statistics

$$\frac{\beta_{i,r}^{\text{final}}}{\sqrt{\Sigma_{i,rr}}} \tag{17}$$

2. **Only** the p-values for the genes that **individually pass the independent filtering step** are adjusted using BH procedure.
   Independent filtering: threshold on

$$\bar{K}_i = \frac{1}{m}\sum_{j=1}^{m}\tilde{K}_{ij}$$

Ref: Wolfgang Huber: Independent filtering increases detection power for high-throughput experiments. PNAS (2010), `http://dx.doi.org/10.1073/pnas.0914005107`

Questions