

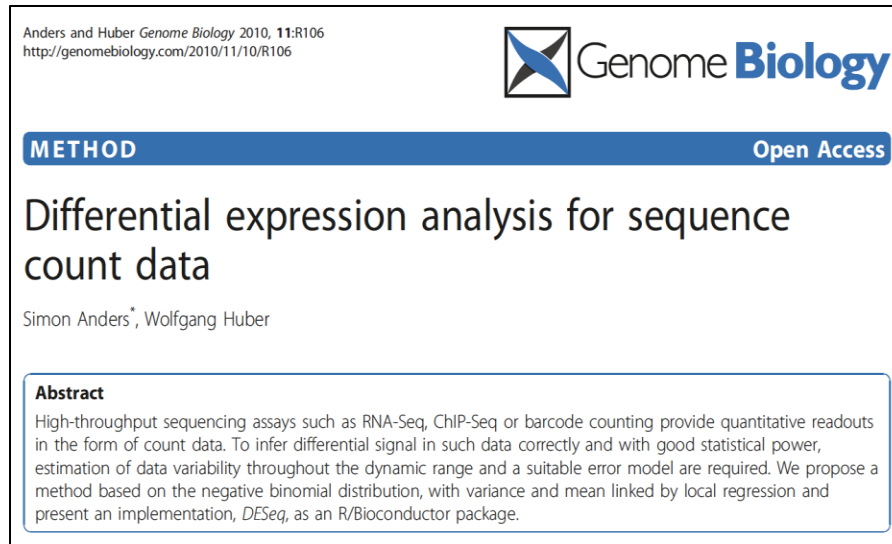


La méthodologie DESeq(2), les grands principes

Gaëlle LELANDAIS

Articles de référence (versions 1 et 2)

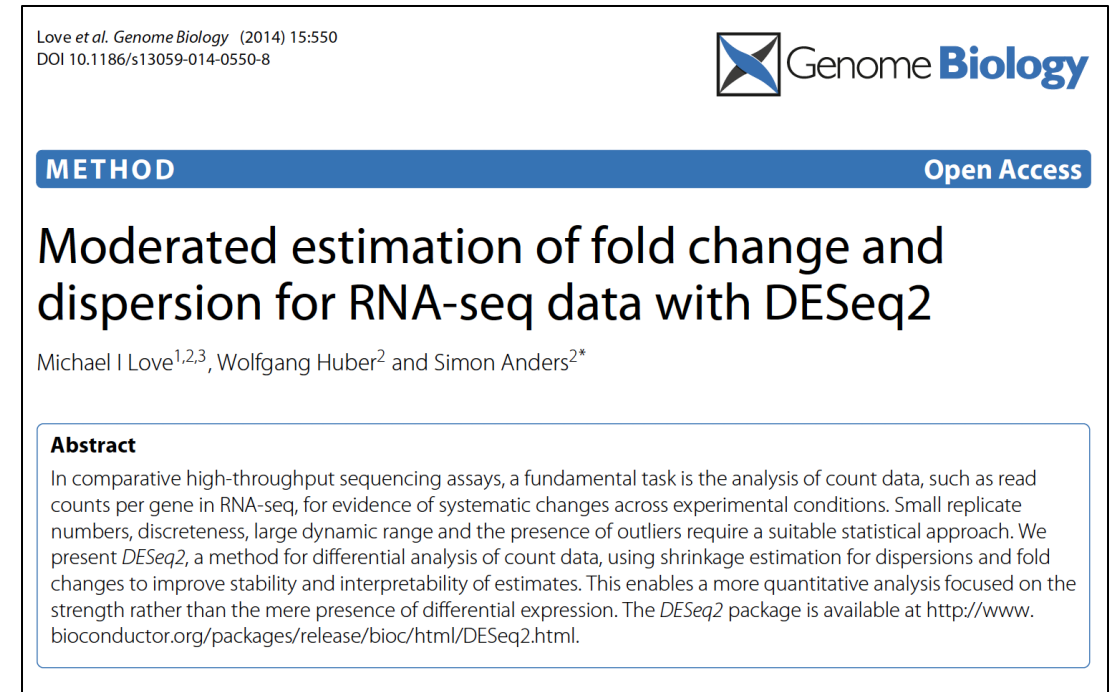
➤ Méthode DESeq (2010)



DOI: [10.1186/gb-2010-11-10-r106](https://doi.org/10.1186/gb-2010-11-10-r106)

Objectif de proposer une procédure de **test statistique** permettant de décider si (pour un gène) la différence observée est **significative**, plus grande que celle qui est attendue par des **variations aléatoires**.

➤ Méthode DESeq2 (2014)



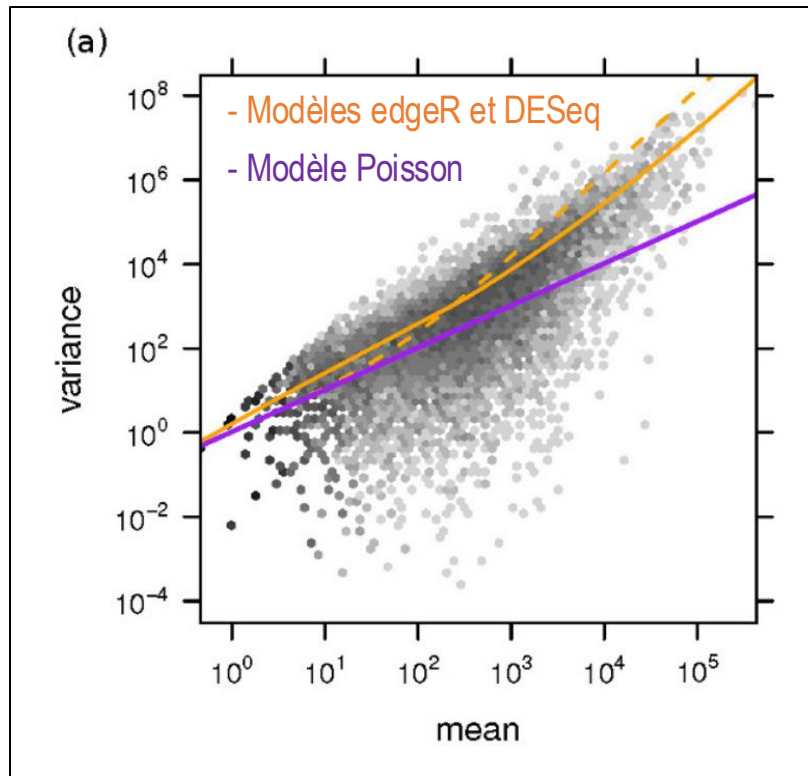
DOI: [10.1186/s13059-014-0550-8](https://doi.org/10.1186/s13059-014-0550-8)

Article rédigé à l'attention des **utilisateurs de la méthode**.
Lecture très abordable (fortement conseillée).

Modéliser la relation entre dispersion et force d'expression (principe 1)

- Une modélisation consiste à établir une relation entre variables, sous forme d'équation. Les paramètres sont estimés à partir des observations.

Anders et al. (2010) – Figure 1



- Paramètre « force d'expression »
(common scale sample mean)

$$\hat{q}_{i\rho} = \frac{1}{m_\rho} \sum_{j:\rho(j)=\rho} \frac{k_{ij}}{\hat{s}_j}, \quad (6)$$

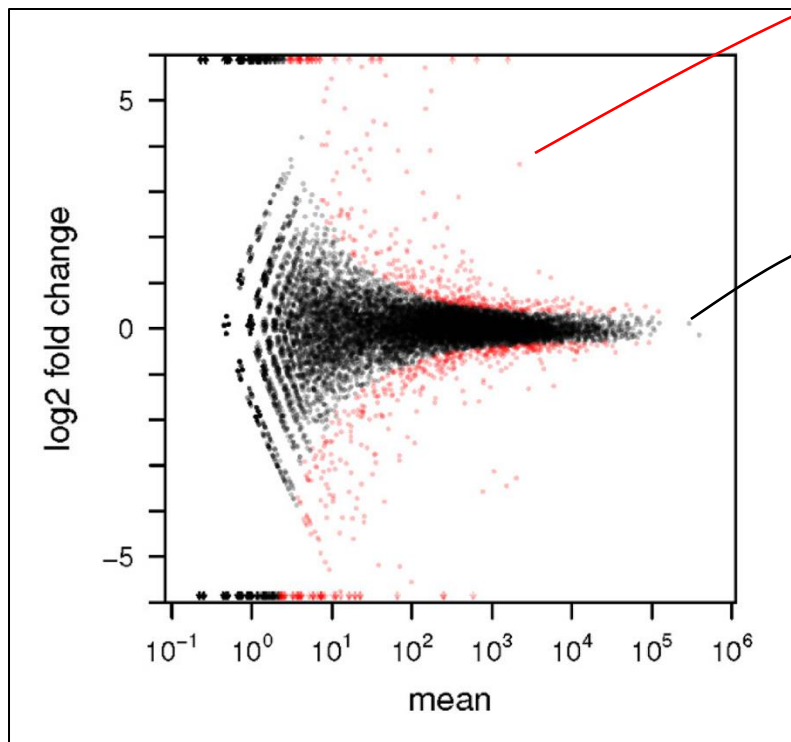
- Paramètre « dispersion »
(common-scale sample variance)

$$w_{i\rho} = \frac{1}{m_\rho - 1} \sum_{j:\rho(j)=\rho} \left(\frac{k_{ij}}{\hat{s}_j} - \hat{q}_{i\rho} \right)^2 \quad (7)$$

Anders et al. (2010)

Pénaliser la taille d'effet en fonction de la force d'expression (principe 2)

- La taille d'effet ($\log_2 \text{fold change}$) est d'autant plus fiable que la force d'expression (mean) est élevée.



Différence significative
(plus grande que celle attendue par
des variations aléatoires)

Différence non significative

Sélection des gènes avec une **importante** force d'expression
malgré **une** taille d'effet plus faible.

Un test statistique est réalisé pour chaque gène,
à partir de modèles estimés sur plusieurs gènes

Tables de données nécessaires à la modélisation GLM (DESeq2)

➤ Table de comptage (*counts*) ...

```
countdata <- assay( parathyroidGenesSE )
head( countdata )
```

##	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]
## ENSG000000000003	792	1064	444	953	519	855	413	365	278	1173	463	316
## ENSG000000000005	4	1	2	3	3	1	0	1	0	0	0	0
## ENSG000000000419	294	282	164	263	179	217	277	204	189	601	257	183
## ENSG000000000457	156	184	93	145	75	122	228	171	116	422	182	122
## ENSG000000000460	396	207	210	212	221	173	611	199	426	1391	286	417
## ENSG000000000938	3	8	2	5	0	4	13	22	3	38	13	10
##	[,13]	[,14]	[,15]	[,16]	[,17]	[,18]	[,19]	[,20]	[,21]	[,22]		
## ENSG000000000003	987	424	305	391	586	714	957	346	433	402		
## ENSG000000000005	0	0	0	0	0	0	1	0	0	0		
## ENSG000000000419	588	275	263	281	406	568	764	288	259	250		
## ENSG000000000457	441	211	131	115	196	266	347	133	168	148		
## ENSG000000000460	1452	238	188	102	389	294	778	162	85	339		
## ENSG000000000938	26	13	7	3	10	18	15	7	8	7		
##	[,23]	[,24]	[,25]	[,26]	[,27]							
## ENSG000000000003	277	511	366	271	492							
## ENSG000000000005	0	0	0	0	0							
## ENSG000000000419	147	271	227	197	363							
## ENSG000000000457	83	184	136	118	195							
## ENSG000000000460	75	154	314	117	233							

Beginner's guide to using the DESeq2 package. M. Love, W. Huber and S. Anders (bioRxiv, 2014)

Tables de données nécessaires à la modélisation GLM (DESeq2)

- Table de comptage (*counts*) et table des variables explicatives (plan d'expérience)

```
as.data.frame( colData(dds) )
```

##	run	experiment	patient	treatment	time	submission	study
##	SRS308866	SRR479053	SRX140504	1	Control	48h	SRA051611 SRP012167
##	SRS308868	SRR479055	SRX140506	1	DPN	48h	SRA051611 SRP012167
##	SRS308870	SRR479057	SRX140508	1	OHT	48h	SRA051611 SRP012167
##	SRS308872	SRR479059	SRX140510	2	Control	48h	SRA051611 SRP012167
##	SRS308874	SRR479062	SRX140512	2	DPN	48h	SRA051611 SRP012167
##	SRS308876	SRR479065	SRX140514	2	OHT	48h	SRA051611 SRP012167
##	SRS308878	SRR479067	SRX140516	3	Control	48h	SRA051611 SRP012167
##	SRS308880	SRR479069	SRX140518	3	DPN	48h	SRA051611 SRP012167
##	SRS308882	SRR479071	SRX140520	3	OHT	48h	SRA051611 SRP012167
##	SRS308883	SRR479072	SRX140521	4	Control	48h	SRA051611 SRP012167
##	SRS308885	SRR479074	SRX140523	4	DPN	48h	SRA051611 SRP012167
##	SRS308887	SRR479077	SRX140525	4	OHT	48h	SRA051611 SRP012167
##	sample	run.1	runsCollapsed				
##	SRS308866	SRS308866	SRR479053		SRR479053		
##	SRS308868	SRS308868	SRR479055		SRR479055		
##	SRS308870	SRS308870	SRR479057		SRR479057		
##	SRS308872	SRS308872	SRR479059		SRR479059		
##	SRS308874	SRS308874	SRR479062		SRR479062		
##	SRS308876	SRS308876	SRR479065		SRR479065		
##	SRS308878	SRS308878	SRR479067		SRR479067		

Beginner's guide to using the DESeq2 package. M. Love, W. Huber and S. Anders (bioRxiv, 2014)

- Ajustement (pour chaque gène) d'un modèle GLM (*generalized linear model*)

Model and normalization

The read count K_{ij} for gene i in sample j is described with a GLM of the negative binomial family with a logarithmic link:

$$K_{ij} \sim \text{NB}(\text{mean} = \mu_{ij}, \text{dispersion} = \alpha_i) \quad (1)$$

$$\mu_{ij} = s_{ij}q_{ij}$$

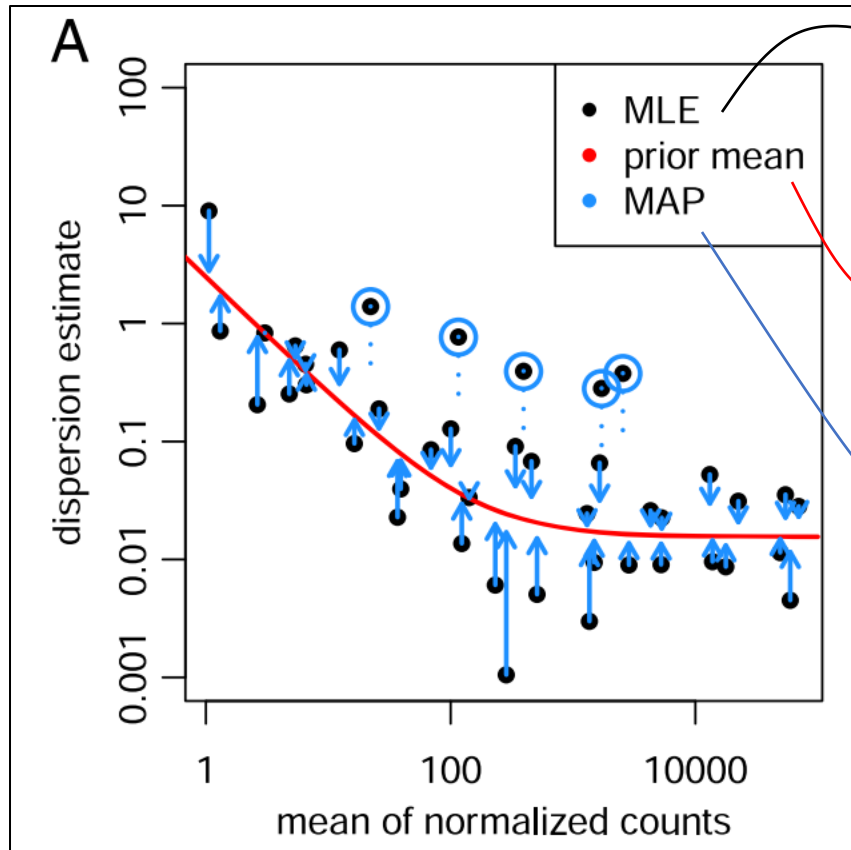
$$\log q_{ij} = \sum_r x_{jr} \beta_{ir}. \quad (2)$$

Love et al. (2014)

Objectif d'exprimer les valeurs de comptage comme une combinaison linéaire de facteurs explicatifs, directement reliés au plan d'expérience.

Procédure de « shrinkage » (quelques notions)

- Littéralement « contraction » ou « rétrécissement », cette procédure améliore les estimations de paramètres (sans augmenter le nombre de réplicats).



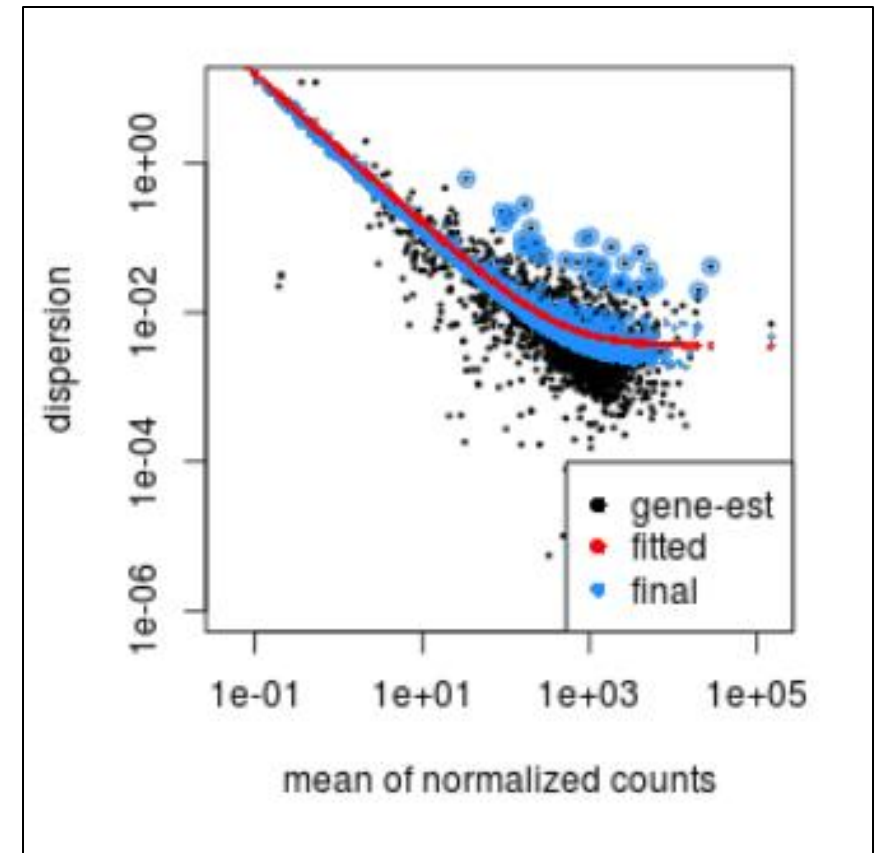
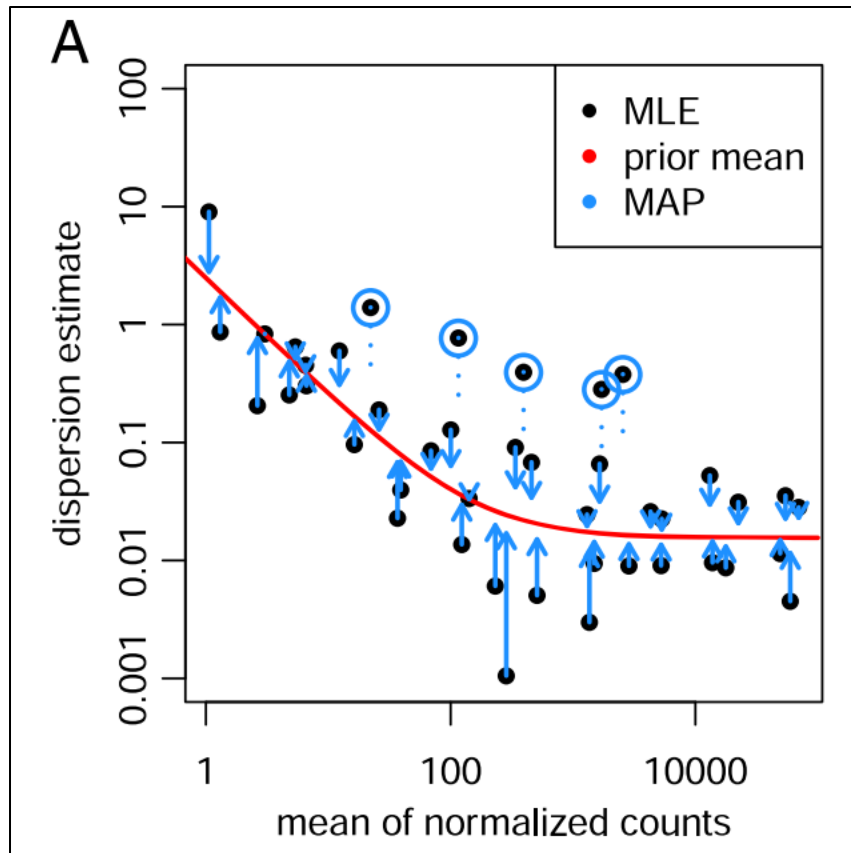
Etape 1 : Estimation des paramètres de *dispersion* et de *mean of normalized counts*, pour chaque gène (pour un gène donné, seules les données de ce gène sont utilisées).

Etape 2 : Procédure d'ajustement (courbe rouge), afin de révéler la relation entre les paramètres de dispersion et de force d'expression (les données de tous les gènes sont utilisées).

Etape 3 : Nouvelle estimation du paramètre de dispersion des gènes (utilisation des valeurs de *prior mean* de l'étape 2).

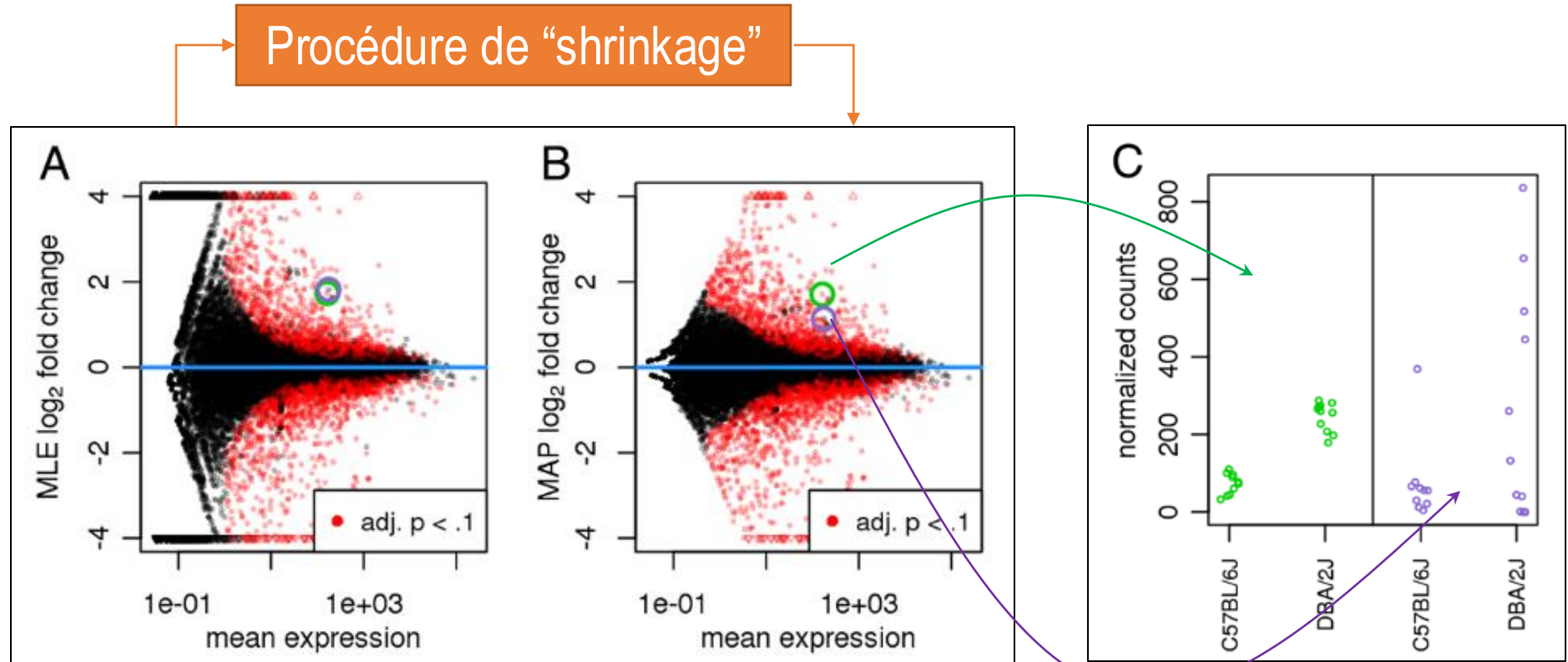
- Procédure de détection de gènes *outliers* pour le paramètre de dispersion (étape 3 non appliquée).

Procédure de « shrinkage » (résultats sur tous les gènes)



Taille d'effet, dispersion et force d'expression

(bilan DESeq2 sur l'ensemble des gènes)



Love et al. (2014) – Figure 2



Table finale des résultats

- L'application de la méthode DESeq2 se conclut par la compilation des différents paramètres calculés dans un tableau.

```
res <- results( dds )
res
## log2 fold change (MAP): treatment OHT vs Control
## Wald test p-value: treatment OHT vs Control
## DataFrame with 4000 rows and 6 columns
##           baseMean log2FoldChange      lfcSE      stat      pvalue
##           <numeric>      <numeric> <numeric> <numeric> <numeric>
## ENSG000000000003    616.390      -0.04917    0.0868    -0.5667    0.57094
## ENSG000000000005         0.553      -0.08167    0.1658    -0.4927    0.62225
## ENSG000000000419    305.400      0.10779    0.0945     1.1405    0.25410
## ENSG000000000457    184.341      0.00189    0.1196     0.0158    0.98742
## ENSG000000000460    208.305      0.44504    0.1371     3.2456    0.00117
## ...
## ENSG00000112851    1071.1      -0.04915    0.0628    -0.7827     0.434
## ENSG00000112852      11.8        0.24846    0.3184     0.7803     0.435
## ENSG00000112855     257.8        0.03485    0.1019     0.3421     0.732
## ENSG00000112874     311.4        0.00817    0.1058     0.0772     0.938
## ENSG00000112877      44.4       -0.01252    0.2085    -0.0600     0.952
##           padj
##           <numeric>
## ENSG000000000003    0.9984
## ENSG000000000005         NA
## ENSG000000000419    0.9894
## ENSG000000000457    0.9984
## ENSG000000000460    0.0995
## ...
## ENSG00000112851     0.998
## ENSG00000112852         NA
## ENSG00000112855     0.998
## ENSG00000112874     0.998
## ENSG00000112877         NA
```

Paramètre	Signification
baseMean	Force d'expression, valeur moyenne des comptages (counts) normalisés.
log2FoldChange	Taille d'effet, différentiel d'expression entre les conditions comparées.
lfcSE	Erreur standard associée à la taille d'effet.
stat	Statistique du test d'hypothèses.
pvalue	Valeur P calculée à partir de la statistique de test, en utilisant la loi H0.
padj	Valeur P corrigée, prise en compte de la problématique des tests multiples.

Beginner's guide to using the DESeq2 package. M. Love, W. Huber and S. Anders (bioRxiv, 2014)

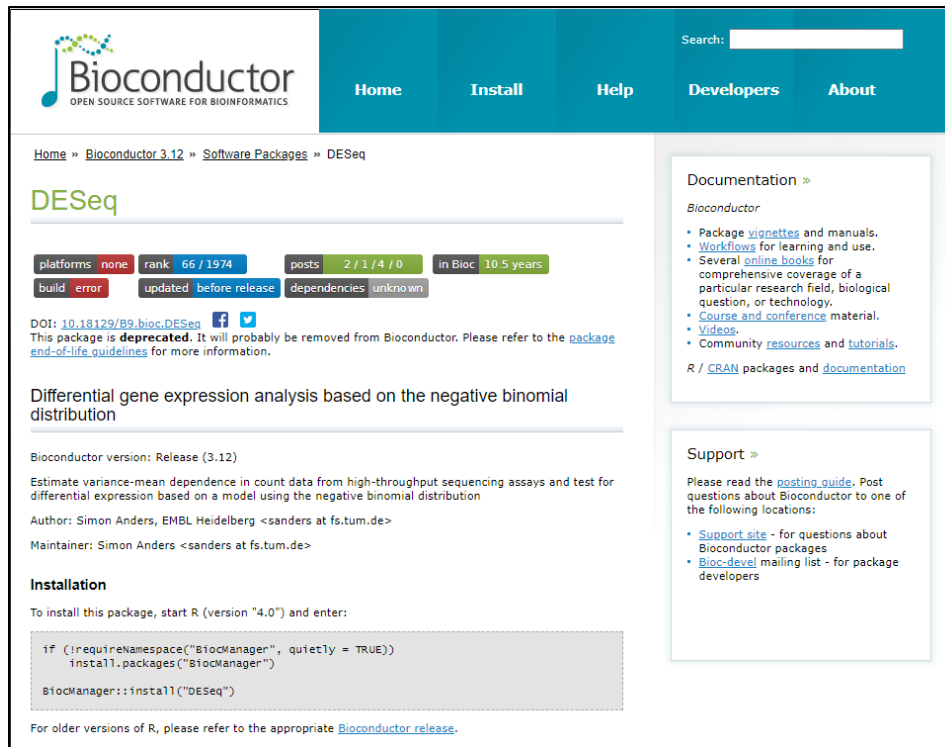
Bilan

- La méthodologie DESeq(2) est une des méthodologies d'analyse différentielle les plus utilisée. Elle combine les trois notions critiques de l'analyse différentielle.

Taille d'effet

Dispersion

Force expression



Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home Install Help Developers About

Search:

Home » Bioconductor 3.12 » Software Packages » DESeq

DESeq

platforms: none rank: 66 / 1974 posts: 2 / 1 / 4 / 0 in Bioc: 10.5 years
build: error updated: before release dependencies: unknown

DOI: 10.18129/B9.bioc.DESeq
This package is **deprecated**. It will probably be removed from Bioconductor. Please refer to the [package end-of-life guidelines](#) for more information.

Differential gene expression analysis based on the negative binomial distribution

Bioconductor version: Release (3.12)

Estimate variance-mean dependence in count data from high-throughput sequencing assays and test for differential expression based on a model using the negative binomial distribution

Author: Simon Anders, EMBL Heidelberg <sanders at fs.tum.de>
Maintainer: Simon Anders <sanders at fs.tum.de>

Installation

To install this package, start R (version "4.0") and enter:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

BiocManager::install("DESeq")
```

For older versions of R, please refer to the appropriate [Bioconductor release](#).

Documentation »

Bioconductor

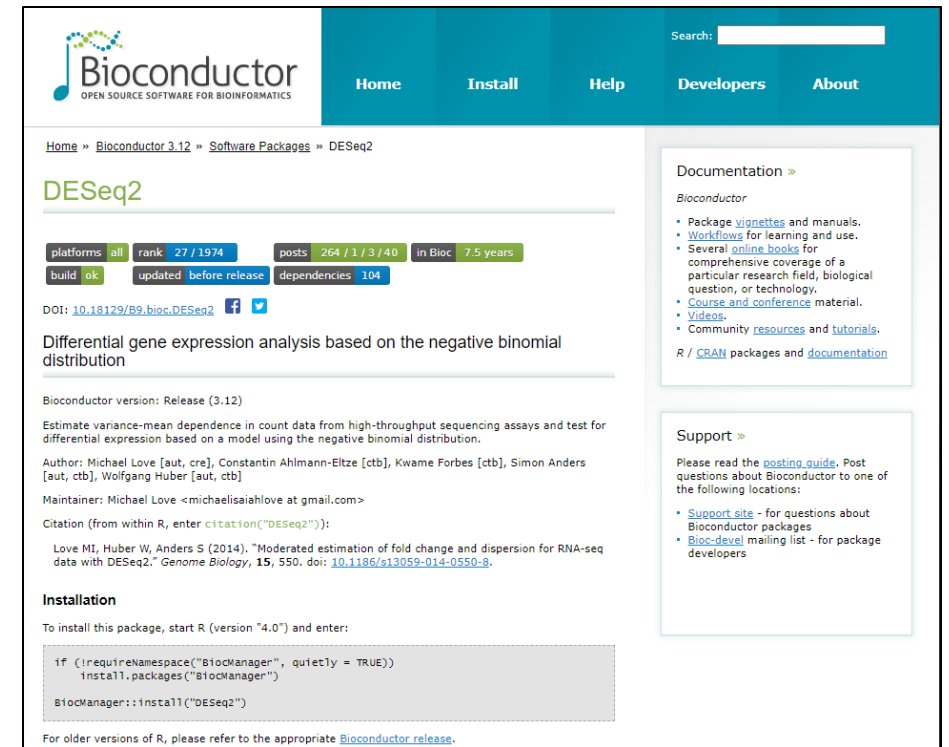
- Package [vignettes](#) and manuals.
- [Workflows](#) for learning and use.
- Several [online books](#) for comprehensive coverage of a particular research field, biological question, or technology.
- [Course and conference](#) material.
- [Videos](#).
- Community [resources](#) and [tutorials](#).

R / [CRAN](#) packages and [documentation](#)

Support »

Please read the [posting guide](#). Post questions about Bioconductor to one of the following locations:

- [Support site](#) - for questions about Bioconductor packages
- [Bio-devel](#) mailing list - for package developers



Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home Install Help Developers About

Search:

Home » Bioconductor 3.12 » Software Packages » DESeq2

DESeq2

platforms: all rank: 27 / 1974 posts: 264 / 1 / 3 / 40 in Bioc: 7.5 years
build: ok updated: before release dependencies: 104

DOI: 10.18129/B9.bioc.DESeq2

Differential gene expression analysis based on the negative binomial distribution

Bioconductor version: Release (3.12)

Estimate variance-mean dependence in count data from high-throughput sequencing assays and test for differential expression based on a model using the negative binomial distribution.

Author: Michael Love [aut, cre], Constantin Ahlmann-Eltze [ctb], Kwame Forbes [ctb], Simon Anders [aut, ctb], Wolfgang Huber [aut, ctb]
Maintainer: Michael Love <michaelisaiahlove at gmail.com>

Citation (from within R, enter `citation("DESeq2")`):
Love MI, Huber W, Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biology*, 15, 550. doi: [10.1186/s13059-014-0550-8](#).

Installation

To install this package, start R (version "4.0") and enter:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

BiocManager::install("DESeq2")
```

For older versions of R, please refer to the appropriate [Bioconductor release](#).

Documentation »

Bioconductor

- Package [vignettes](#) and manuals.
- [Workflows](#) for learning and use.
- Several [online books](#) for comprehensive coverage of a particular research field, biological question, or technology.
- [Course and conference](#) material.
- [Videos](#).
- Community [resources](#) and [tutorials](#).

R / [CRAN](#) packages and [documentation](#)

Support »

Please read the [posting guide](#). Post questions about Bioconductor to one of the following locations:

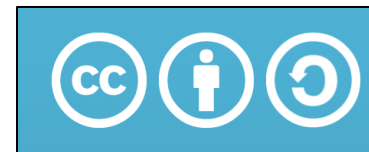
- [Support site](#) - for questions about Bioconductor packages
- [Bio-devel](#) mailing list - for package developers

<https://bioconductor.org/packages/release/bioc/html/DESeq.html>

<https://bioconductor.org/packages/release/bioc/html/DESeq2.html>

Ressource(s) complémentaire(s)

- Commentaires enregistrés du cours (13 minutes) :
 - Lien Youtube : <https://youtu.be/XzbzaDNqXhQ>
- Exemple de mise en application avec le logiciel R (32 minutes) :
 - Lien Youtube : https://youtu.be/BNz7QkuHY_U
- Interview de Michaël Love (1 heure 31 minutes) :
 - Lien podcast : <https://bioinformatics.chat/deseq2>



Sauf mention contraire, ce contenu est mis à disposition selon les termes de la licence Creative Commons Attribution - Partage dans les mêmes conditions 4.0 International (CC BY-SA 4.0)

Gaëlle LELANDAIS

Version du document : 11/02/2025