WELCOME TO THE HANDS-ON SESSION OF THE WORKSHOP

"Incorporating biological sex as a variable in the design and analysis of biomedical research experiments"

24 June 2024 - Biel/Bienne Congress Centre

Introduction

This hands-on tutorial is designed to foster a critical perspective on integrating the sex dimension in bioinformatics. Through a practical exercise focusing on **Differential Expression (DE) analysis**, we will together examine the implications of the conclusions from a published study in light of the results of our re-analysis. You can access the full text of the study here: Chucair-Elliott et al., 2019.

In brief, the study by Chucair-Elliott et al. investigated the effects of tamoxifen, a common Cre recombinase inducer, on the central nervous system transcriptome in mice. The findings revealed that tamoxifen did not induce sexually divergent effects, indicating its suitability for aging studies without introducing significant transcriptomic variability.

This tutorial offers flexibility in how you approach the DE analysis:

- You have the option to conduct the DE analysis using either **DESeq2** or **limma**, and different versions of the scripts are available, from easy one to more advanced (if you are familiar with DE analysis using Bioconductor)**.**
- You can perform the DE analysis locally on your computer, either through **RStudio** or the terminal, or using a **Colab notebook** (DESeq2). With the exception of a few Python commands in the Colab notebook, the entire hands-on activity is conducted using the R programming language.

Hands-on scripts locations

You can find all the scripts and relevant publications at the following locations:

(1) At a dedicated directory on **Next Cloud**:

https://io.scicore.unibas.ch/s/kExkmzEyJm2RErz
password: 9MbT2mygZd

(2) At a dedicated repository on **GitHub**:

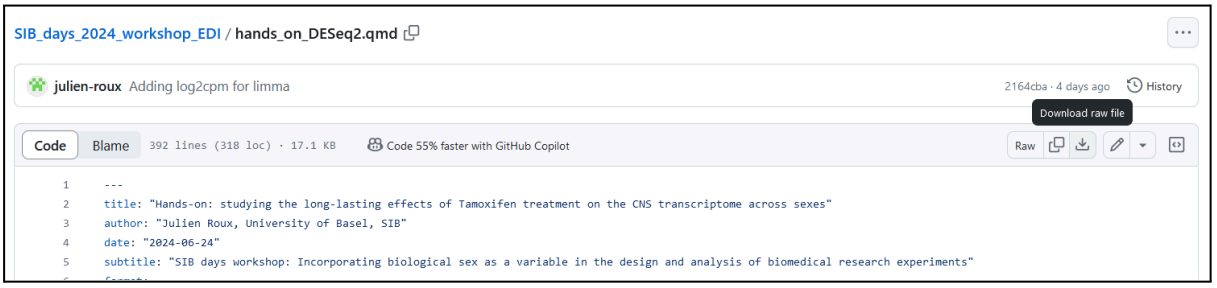
https://github.com/julien-roux/SIB_days_2024_workshop_EDI

If you have git installed in your machine, you can download the entire repo by opening a terminal, changing directory to your favorite location and type:

git clone https://github.com/julien-roux/SIB_days_2024_workshop_EDI.git

Instructions on how to install git are here.

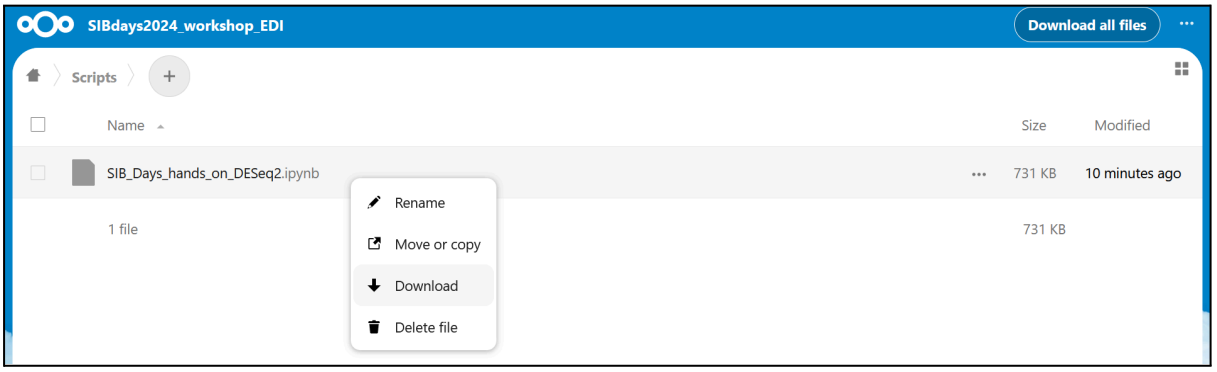
Alternatively, you can download individual scripts by clicking on them and then clicking on the “Download raw file” logo on the banner.



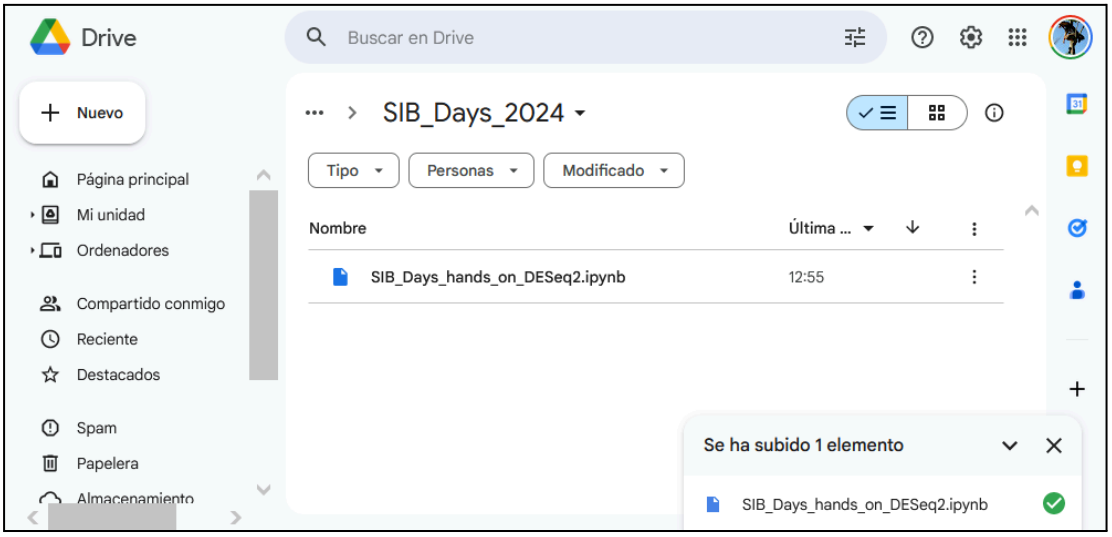
Using the Colab notebook

Google's Colaboratory (short, Colab) is a cloud-based Jupyter Notebook interface created by Google. If you have a personal Google account, you can create your own Colab notebooks on your Google Drive by clicking on *New* → *More* → *Google Colaboratory*. **For this hands-on, we have already prepared a Colab notebook that you can use following these steps:**

(1) Download the hands-on Colab notebook (“*SIB_Days_hands_on_DESeq2.ipynb*”) from either Github or Next Cloud (see previous section).



(2) Place the Colab notebook in your preferred location within your personal Google Drive (for example, a new directory called *SIB_Days_2024*).

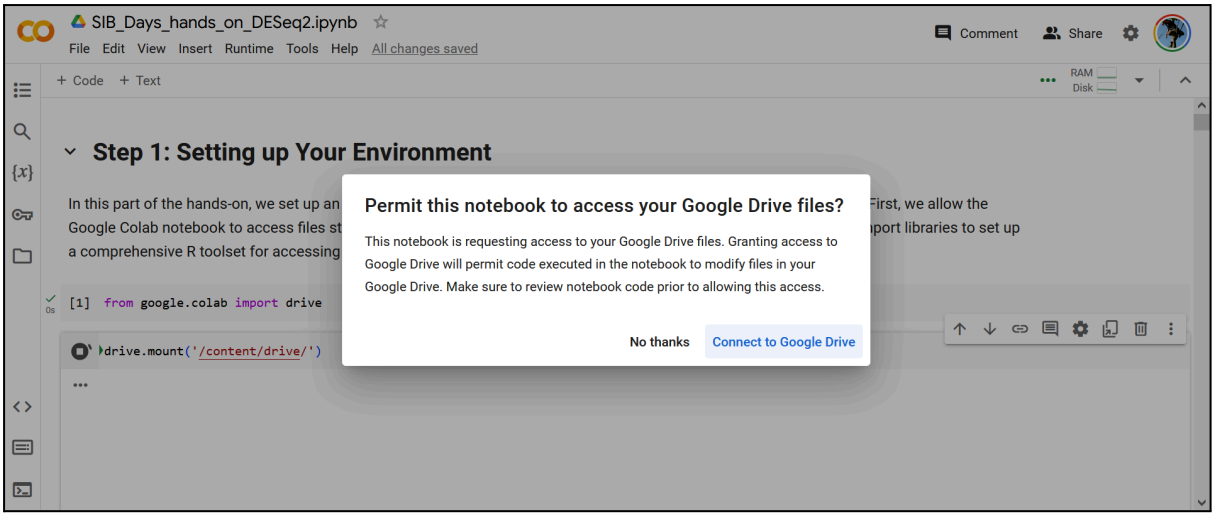


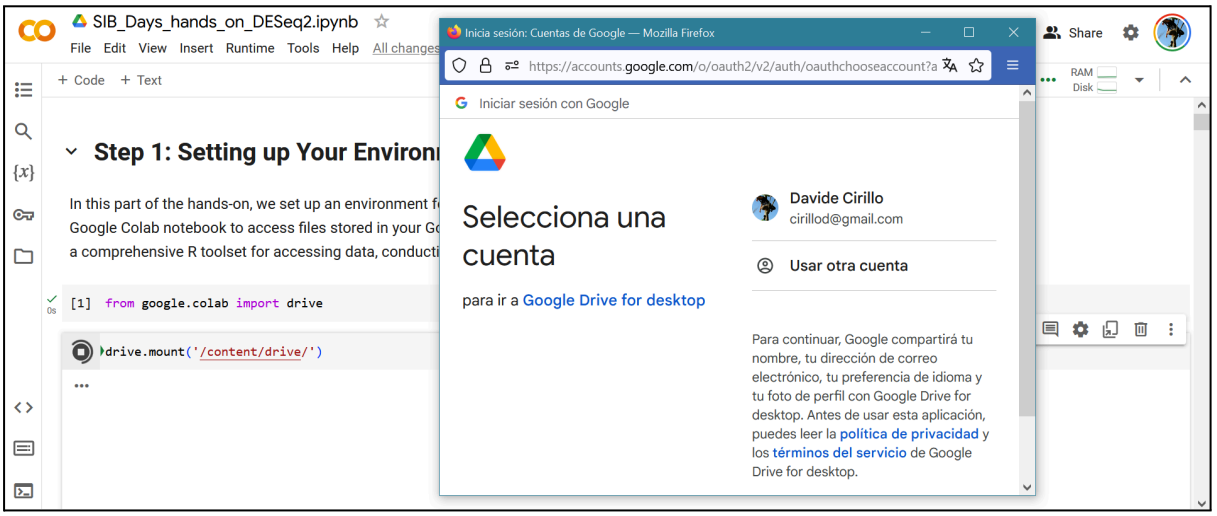
IMPORTANT: If you do not want to use your personal Google Drive, you can follow the hands-on with the instructors or create a new account just for this activity (please be aware that the creation of a new Google account requires authentication with a mobile phone number).

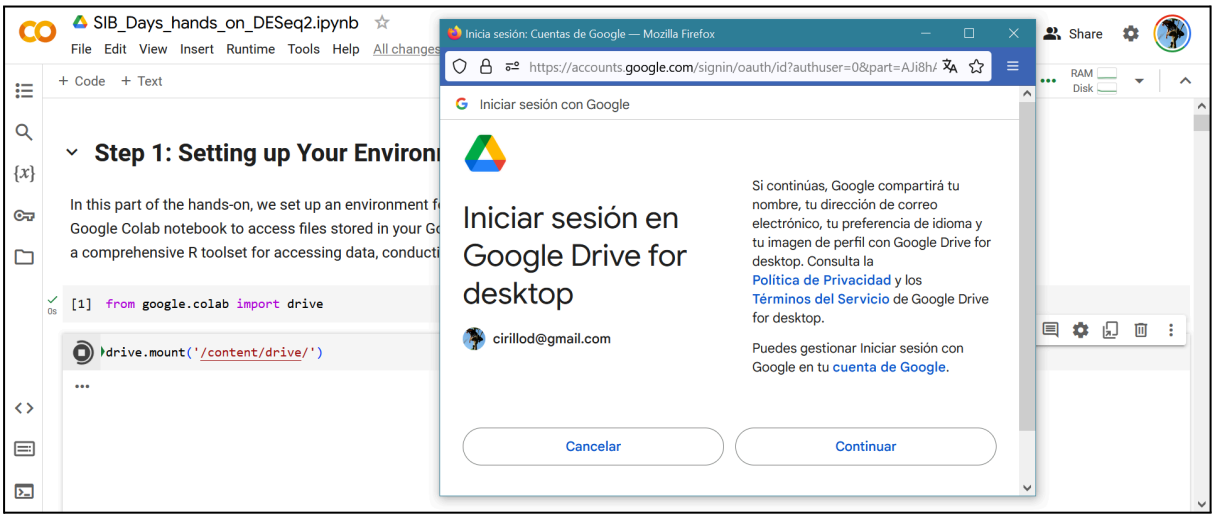
Additionally, please note that Google Drive and Google Colab have different storage systems: Google Drive is for storing files online, while **Google Colab's disk is a temporary file system used during your Colab session**. You can remove the Colab notebook from your Google Drive at any time.

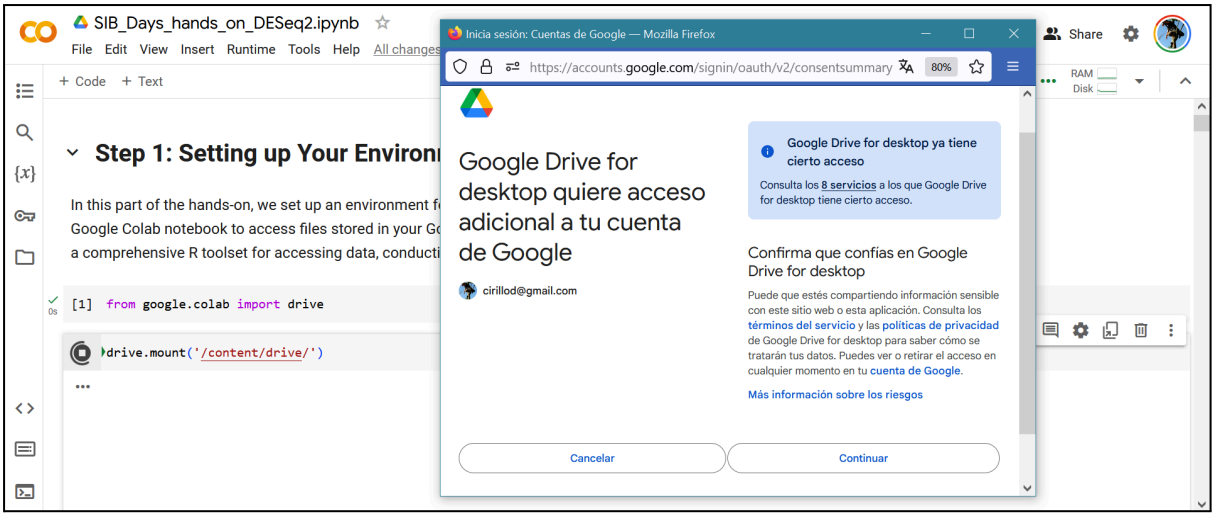
(3) To open the Colab notebook, simply click on it. Run each cell sequentially by either clicking the play icon on the left or pressing Ctrl+Enter. Please take into account the following aspects regarding the initial step, "**Step 1: Setting Up Your Environment**":

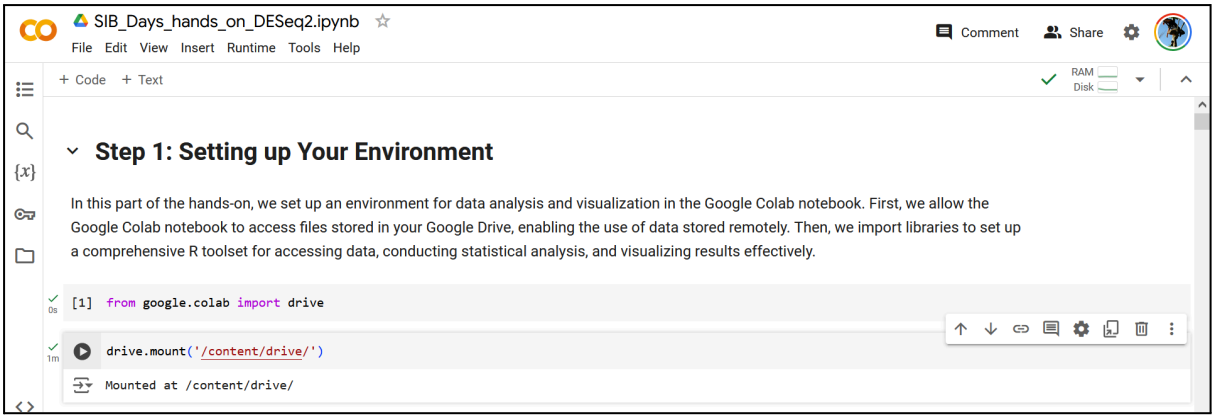
(a) When you mount the Colab session content to your Google Drive (drive.mount('/content/drive/')), the notebook will prompt you to grant permission for access to your Google Drive files. You'll need to select your account and authorize this access. This step is necessary for various reasons, such as saving plots or other notebook-generated content directly to your Google Drive. You can confidently proceed with granting these permissions for this tutorial. However, if you prefer not to grant access, you can participate the hands-on with the instructor or other attendees instead.


SIB_Days_hands_on_DESeq2.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Step 1: Setting up Your Environment
In this part of the hands-on, we set up an
Google Colab notebook to access files st
a comprehensive R toolset for accessing
First, we allow the
port libraries to set up
[1] from google.colab import drive
drive.mount('/content/drive/')
Permit this notebook to access your Google Drive files?
This notebook is requesting access to your Google Drive files. Granting access to Google Drive will permit code executed in the notebook to modify files in your Google Drive. Make sure to review notebook code prior to allowing this access.
No thanks
Connect to Google Drive

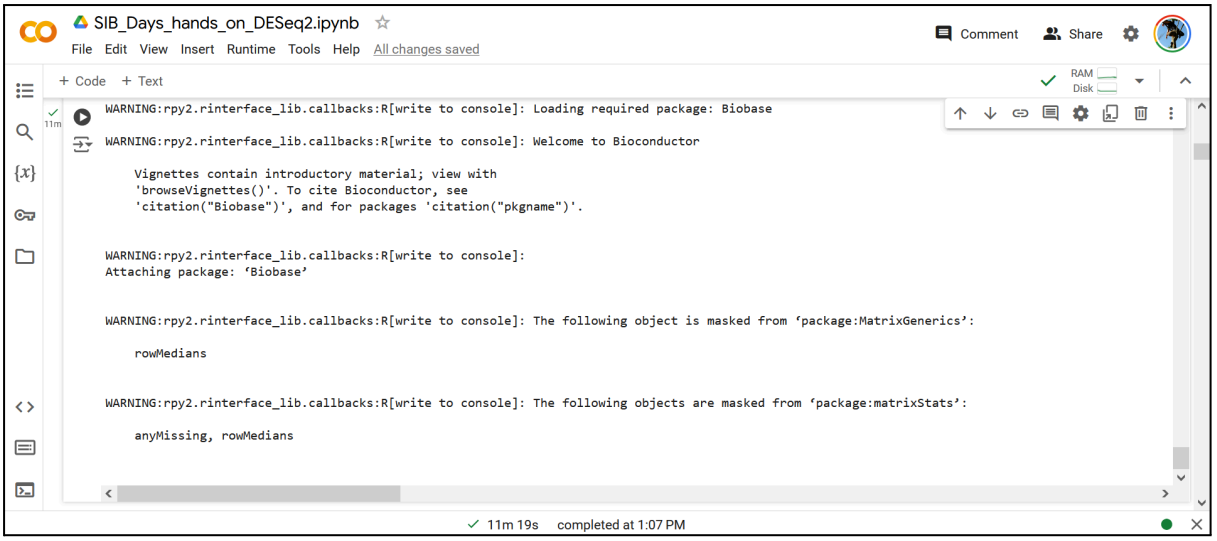

SIB_Days_hands_on_DESeq2.ipynb
File Edit View Insert Runtime Tools Help All changes
Share
+ Code + Text
RAM
Disk
Step 1: Setting up Your Environ
In this part of the hands-on, we set up an environment f
Google Colab notebook to access files stored in your Go
a comprehensive R toolset for accessing data, conducti
[1] from google.colab import drive
drive.mount('/content/drive/')
Inicia sesión: Cuentas de Google — Mozilla Firefox
https://accounts.google.com/o/oauth2/v2/auth/oauthchooseaccount?a
Iniciar sesión con Google
Selecciona una cuenta
para ir a Google Drive for desktop
Davide Cirillo
cirillod@gmail.com
Usar otra cuenta
Para continuar, Google compartirá tu nombre, tu dirección de correo electrónico, tu preferencia de idioma y tu foto de perfil con Google Drive for desktop. Antes de usar esta aplicación, puedes leer la política de privacidad y los términos del servicio de Google Drive for desktop.


SIB_Days_hands_on_DESeq2.ipynb
File Edit View Insert Runtime Tools Help All changes
Share
+ Code + Text
RAM
Disk
Step 1: Setting up Your Environ
In this part of the hands-on, we set up an environment f
Google Colab notebook to access files stored in your Go
a comprehensive R toolset for accessing data, conducti
[1] from google.colab import drive
drive.mount('/content/drive/')
Inicia sesión: Cuentas de Google — Mozilla Firefox
https://accounts.google.com/signin/oauth/id?authuser=0&part=AJi8hA
Iniciar sesión con Google
Iniciar sesión en Google Drive for desktop
cirillod@gmail.com
Si continúas, Google compartirá tu nombre, tu dirección de correo electrónico, tu preferencia de idioma y tu imagen de perfil con Google Drive for desktop. Consulta la Política de Privacidad y los Términos del Servicio de Google Drive for desktop.
Puedes gestionar Iniciar sesión con Google en tu cuenta de Google.
Cancelar
Continuar





(b) Initially, you will need to install the required libraries on your assigned cloud node, which typically takes about 10 minutes. Please be patient and don't worry about any WARNINGS you may see 🙂. When prompted to update all/some/none of the packages, just type “n”.



A refresher on DE Analysis with DESeq2

DESeq2 is a powerful method for identifying genes that are differentially expressed across different conditions or treatments. Below, each step of the DESeq2 algorithm is explained in detail.

1. Estimating Size Factors

- DESeq2 calculates size factors for each sample to account for differences in library size or sequencing depth.
- The median ratio method is typically used. For each gene, the counts in each sample are divided by the geometric mean of counts across all samples.
- The size factor for each sample is the median of these ratios, which helps scale the counts to a comparable level across all samples.

$$\text{Size Factor}_i = \text{median}_j \left(\frac{K_{ij}}{(\prod_{k=1}^{n} K_{kj})^{1/n}} \right)$$

where K_{ij} is the count for gene j in sample i, and n is the number of samples.

2. Estimating Dispersions

- DESeq2 models the dispersion for each gene, which represents the variability of gene expression counts across replicates.
- Dispersion estimation involves fitting a generalized linear model (GLM) for each gene and estimating how much the counts deviate from the mean after accounting for the size factors.

- This step is crucial because RNA-Seq data often show greater variability at low counts and less variability at high counts.

$$\mathrm{Var}(K_{ij}) = \mu_{ij} + \alpha\mu_{ij}^2$$

where μ_{ij} is the expected count and α is the dispersion parameter.

3. Gene-wise Dispersion Estimates

- Initially, DESeq2 fits a dispersion parameter for each gene without borrowing information across genes.
- This individual estimate can be noisy, especially for genes with low counts or low replication.

$$\alpha_j = \frac{\sum_i \left(\frac{K_{ij}-\mu_{ij}}{\mu_{ij}^2}\right)^2}{\sum_i \frac{1}{\mu_{ij}}}$$

4. Mean-Dispersion Relationship

- DESeq2 fits a curve to the dispersion estimates across all genes to establish a mean-dispersion trend.
- This step involves smoothing the gene-wise dispersion estimates to borrow strength across genes and stabilize the estimates, especially for lowly expressed genes.
- The mean-dispersion relationship helps in shrinking the gene-wise dispersion estimates towards the trend, which improves reliability.

$$\alpha(\mu) = \frac{a}{\mu} + b$$

where $\alpha(\mu)$ represents the smoothed dispersion as a function of mean count μ, and a and b are parameters estimated from the data.

5. Final Dispersion Estimates

- DESeq2 computes the final dispersion estimates by shrinking the initial, noisy gene-wise estimates towards the mean-dispersion trend.
- This shrinkage is achieved through an empirical Bayes approach, which balances the gene-specific estimates with the overall trend.

$$\alpha_j^{\text{final}} = \frac{\alpha_j}{\tau + \alpha_j} \times \alpha_j^{\text{initial}} + \frac{\tau}{\tau + \alpha_j} \times \alpha(\mu_j)$$

where τ is a parameter that determines the extent of shrinkage.

6. Fitting Model and Testing

- DESeq2 fits a negative binomial GLM to the count data, incorporating the size factors and dispersion estimates.
- It tests for differential expression using a likelihood ratio test or the Wald test, depending on the experimental design and user preference.
- The resulting p-values are adjusted for multiple testing to control the false discovery rate (FDR).

Model Fitting

$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \alpha_j)$$

where NB denotes the negative binomial distribution.

Wald Test

$$z = \frac{\hat{\beta}_k}{\mathrm{SE}(\hat{\beta}_k)}$$

where $\hat{\beta}_k$ is the estimated coefficient and $\mathrm{SE}(\hat{\beta}_k)$ is its standard error.

References for Further Reading

- Love MI, Huber W, Anders S (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” *Genome Biology* (link)
- DESeq2 Documentation: Bioconductor DESeq2