

PEC2-Análisis de Datos de Ultrasecuenciación

Análisis de Datos Ómicos

Alba Moya Garcés

11 de junio, 2020

Contents

1 Pipeline	1
2 Abstract	2
3 Objetivos	2
4 Material	2
4.1 Software	2
4.2 Datos	2
5 Métodos	2
5.1 Preparación del área de trabajo:	2
5.2 Instalación de paquetes en R	3
5.3 Lectura y selección de los datos	4
5.4 Control de calidad de los datos	5
5.4.1 Transformación de los datos	5
5.4.2 Gráfico de densidad	6
6 Resultados	9
7 Discusión	9
Bibliografía	9

1 Pipeline

1. Definición de los datos tal como se ha descrito en el párrafo anterior
2. Preprocesado de los datos: filtraje y normalización
3. Identificación de genes diferencialmente expresados

4. Anotación de los resultados
5. Busca de patrones de expresión y agrupación de las muestras (comparación entre las distintas comparaciones).
6. Análisis de significación biológica (“Gene Enrichment Analysis”)

2 Abstract

3 Objetivos

4 Material

El código completo para desarrollar este análisis, o cualquier otro a partir de su adaptación, puede descargarse del siguiente repositorio de *GitHub*:

<https://github.com/albamgarces/analisis-de-datos-de-RNA-seq.git>.

4.1 Software

Se realizó este análisis utilizando el lenguaje R version 3.6.3 (2020-02-29) R en la interfaz RStudio versión 1.1.456 y las librerías desarrolladas para este tipo de análisis por el proyecto Bioconductor. El programa estadístico R se puede descargar desde la página web del [proyecto CRAN](#) (The Comprehensive R Archive Network) siguiendo las indicaciones. R-Studio puede descargarse desde su página web <https://www.rstudio.com/>.

Finalmente, las librerías adicionales necesarias para llevar a cabo este análisis se obtuvieron del proyecto Bioconductor versión 3.10, el cuál se instala junto con algunos paquetes básicos mediante el siguiente código:

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

4.2 Datos

Los archivos `counts.csv` y `targets.csv` contienen la información de un estudio obtenido del repositorio del proyecto [GTEx](#) (Genotype-Tissue Expression). Encontramos los datos de expresión (RNA-seq) pertenecientes a un análisis del tiroides donde se comparan tres tipos de infiltración en 292 muestras:

- tejidos no infiltrados (NIT): 236 muestras
- infiltración focalizada (SFI): 42 muestras
- infiltración linfocitaria extensiva (ELI): 14 muestras

5 Métodos

5.1 Preparación del área de trabajo:

Para llevar a cabo el análisis, se debe gestionar una gran cantidad de archivos entre aquellos que ocupan los datos originales y los generados durante su análisis. Es por ello que siempre se debería comenzar creando las carpetas necesarias para simplificar la ruta de trabajo. Se recomienda generar una **carpeta principal**

con el nombre de nuestro proyecto en cuyo interior alojaremos una carpeta con los archivos de **datos** y otra con los **resultados** generados del análisis

Estas carpetas las genereamos rápidamente desde el explorador de archivos o la consola de cualquiera de los sistemas operativos usuales. Desde R también podemos generar estas subcarpetas mediante el siguiente código:

```
setwd(".")
dir.create("data")
dir.create("results")
```

5.2 Instalación de paquetes en R

A continuación se muestran los paquetes necesarios para este estudio que requieren instalación:

```
#para que salgan los que están en curso: (.packages())
# UNCOMMENT IF INSTALL REQUIRES
##install.packages("readr")
##install.packages("sampling")
## install.packages("knitr")
#install.packages("cluster")
# install.packages("gplots")
## install.packages("ggplot2")
# install.packages("ggrepel")
# install.packages("BiocManager")
# BiocManager::install("oligo")
# BiocManager::install("arrayQualityMetrics")
# BiocManager::install("pvca")
# BiocManager::install("pacman")
# BiocManager::install("geneplotter")
# BiocManager::install("org.Dm.eg.db")
# BiocManager::install("limma")
# BiocManager::install("genefilter")
# BiocManager::install("drosophila2.db")
# BiocManager::install("ReactomePA")
```

x

knitr
sampling
rstudioapi
stats
graphics
grDevices
utils
datasets
methods
base

5.3 Lectura y selección de los datos

Importamos los archivos proporcionados a R. El archivo `targets` contiene las 292 muestras identificadas por un número según sean provenientes de tejidos NIT (1), SFI (2) o ELI (3).

Table 2: Fragmento de la tabla de datos `targets`

Sample_Name	Grupo_analisis	molecular_data_type	sex	Group
GTEX-111CU-0226-SM-5GZXC	1	Allele-Specific Expression	male	NIT
GTEX-111FC-1026-SM-5GZX1	1	RNA Seq (NGS)	male	NIT
GTEX-111VG-0526-SM-5N9BW	3	RNA Seq (NGS)	male	ELI
GTEX-111YS-0726-SM-5GZY8	1	Allele-Specific Expression	male	NIT
GTEX-1122O-0226-SM-5N9DA	1	RNA Seq (NGS)	female	NIT
GTEX-1128S-0126-SM-5H12S	1	Allele-Specific Expression	female	NIT

El archivo `counts` contempla las 292 muestras como variables y nos informa del número de veces que se ha detectado cada uno de los 56202 genes identificados en la primera columna.

Table 3: Fragmento de la tabla de datos `count`

	GTEX.111CU.0226.SM.5GZXC	GTEX.111FC.1026.SM.5GZX1
ENSG00000223972.4	7	0
ENSG00000227232.4	401	1064
ENSG00000243485.2	4	0
ENSG00000237613.2	2	0
ENSG00000268020.2	0	0
ENSG00000240361.1	0	1

Con el fin de simplificar el análisis, se decidió seleccionar aleatoriamente 10 muestras de cada tipo de tejido. El código para realizar esta extracción es el que se indica a continuación.

```
library(sampling)
library(knitr)
#selección de muestras en el archivo targets
set.seed(1234)
selected <- strata(targets, c("Group"), size=c(10,10,10), method = "srswor")
muestreo <- getdata(targets, selected)
muestreo <- muestreo[,1:10]
```

Table 4: Selección de diez muestras de cada tipo

	Sample_Name	molecular_data_type	sex	Group
36	GTEX-11TTK-0826-SM-5N9EG	RNA Seq (NGS)	female	NIT
107	GTEX-13O61-0226-SM-5KM52	RNA Seq (NGS)	male	NIT
129	GTEX-144GL-1226-SM-5O9A4	RNA Seq (NGS)	male	NIT
139	GTEX-14753-0926-SM-5Q5BI	RNA Seq (NGS)	male	NIT
164	GTEX-P4QS-2626-SM-2I3EV	Allele-Specific Expression	male	NIT
165	GTEX-P4QT-2626-SM-2I3FM	Allele-Specific Expression	female	NIT
172	GTEX-Q2AI-0326-SM-2I3EK	Allele-Specific Expression	male	NIT
180	GTEX-QV44-0826-SM-2S1RG	Allele-Specific Expression	male	NIT

	Sample_Name	molecular_data_type	sex	Group
190	GTEX-RNOR-0926-SM-2TF56	RNA Seq (NGS)	female	NIT
209	GTEX-T8EM-0226-SM-3DB7C	RNA Seq (NGS)	male	NIT
29	GTEX-11NV4-0626-SM-5N9BR	RNA Seq (NGS)	male	ELI
100	GTEX-13NZ9-1126-SM-5MR37	RNA Seq (NGS)	male	ELI
146	GTEX-14ABY-0926-SM-5Q5DY	Allele-Specific Expression	male	ELI
147	GTEX-14AS3-0226-SM-5Q5B6	RNA Seq (NGS)	female	ELI
149	GTEX-14BMU-0226-SM-5S2QA	Allele-Specific Expression	female	ELI
167	GTEX-PLZ4-1226-SM-2I5FE	RNA Seq (NGS)	female	ELI
186	GTEX-R55G-0726-SM-2TC6J	RNA Seq (NGS)	female	ELI
211	GTEX-TMMY-0826-SM-33HB9	Allele-Specific Expression	female	ELI
251	GTEX-YFC4-2626-SM-5P9FQ	Allele-Specific Expression	female	ELI
253	GTEX-YJ89-0726-SM-5P9F7	RNA Seq (NGS)	male	ELI
14	GTEX-11DXY-0426-SM-5H12R	RNA Seq (NGS)	male	SFI
21	GTEX-11EQ8-0826-SM-5N9FG	Allele-Specific Expression	male	SFI
22	GTEX-11EQ9-0626-SM-5A5K1	RNA Seq (NGS)	male	SFI
23	GTEX-11GS4-0826-SM-5986J	RNA Seq (NGS)	male	SFI
90	GTEX-13FXS-0726-SM-5LZXJ	RNA Seq (NGS)	male	SFI
98	GTEX-13NYC-2426-SM-5MR3K	RNA Seq (NGS)	male	SFI
185	GTEX-R55E-0826-SM-2TC5M	Allele-Specific Expression	male	SFI
199	GTEX-S341-0226-SM-5S2VG	RNA Seq (NGS)	female	SFI
224	GTEX-WYVS-0326-SM-3NM9V	RNA Seq (NGS)	female	SFI
261	GTEX-ZE7O-1126-SM-57WC8	Allele-Specific Expression	female	SFI

```
#selección de muestras en el archivo counts
count_subset <- counts[, c(muestreo$ID_unit)]
```

Table 5: Fragmento tabla counts de muestras seleccionadas

	NIT1	NIT2	NIT3
ENSG00000223972.4	3	2	3
ENSG00000227232.4	1241	624	913
ENSG00000243485.2	1	1	1
ENSG00000237613.2	2	1	0
ENSG00000268020.2	0	2	1

Con los datos de conteo de estas 30 muestras seleccionadas continuaremos el resto del análisis.

5.4 Control de calidad de los datos

La evaluación de la calidad de los datos es esencial para cualquier tipo de análisis de datos. Esto nos permitirá detectar errores técnicos que puedan manifestarse en la toma de datos para poder detectar los genes expresados diferencialmente.

5.4.1 Transformación de los datos

Para poder explorar y visualizar los datos en su conjunto, es esencial tener todos los valores en una escala comparable. Transformaremos los datos a escala \log_2 para normalizar las distribuciones. Se utiliza este método de transformación porque facilita su conversión a la escala original.

Tal y como recomienda González (Toulouse, 2014), sumaremos 1 a cada registro de conteo para evitar los valores 0 de algunas condiciones:

$$y = \log_2(K + 1)$$

5.4.2 Gráfico de densidad

El gráfico de densidad de Kernel nos permitirá hacernos una idea de la similitud entre las diferentes distribuciones.

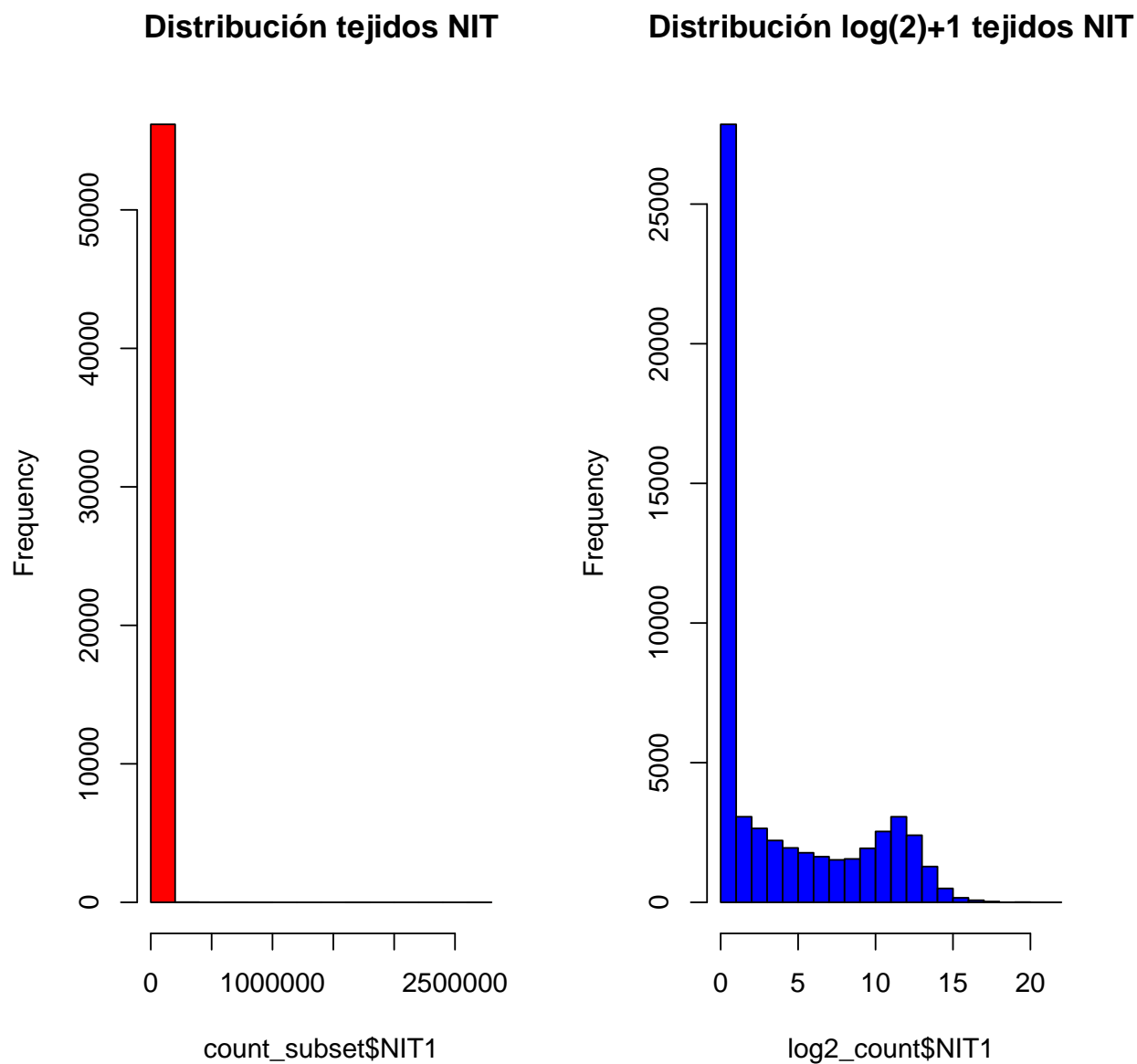


Figure 1: Distribución de las expresiones de una de las muestras de tejido NIT. Podemos ver el efecto de la transformación de los datos mediante $\log(2)+1$

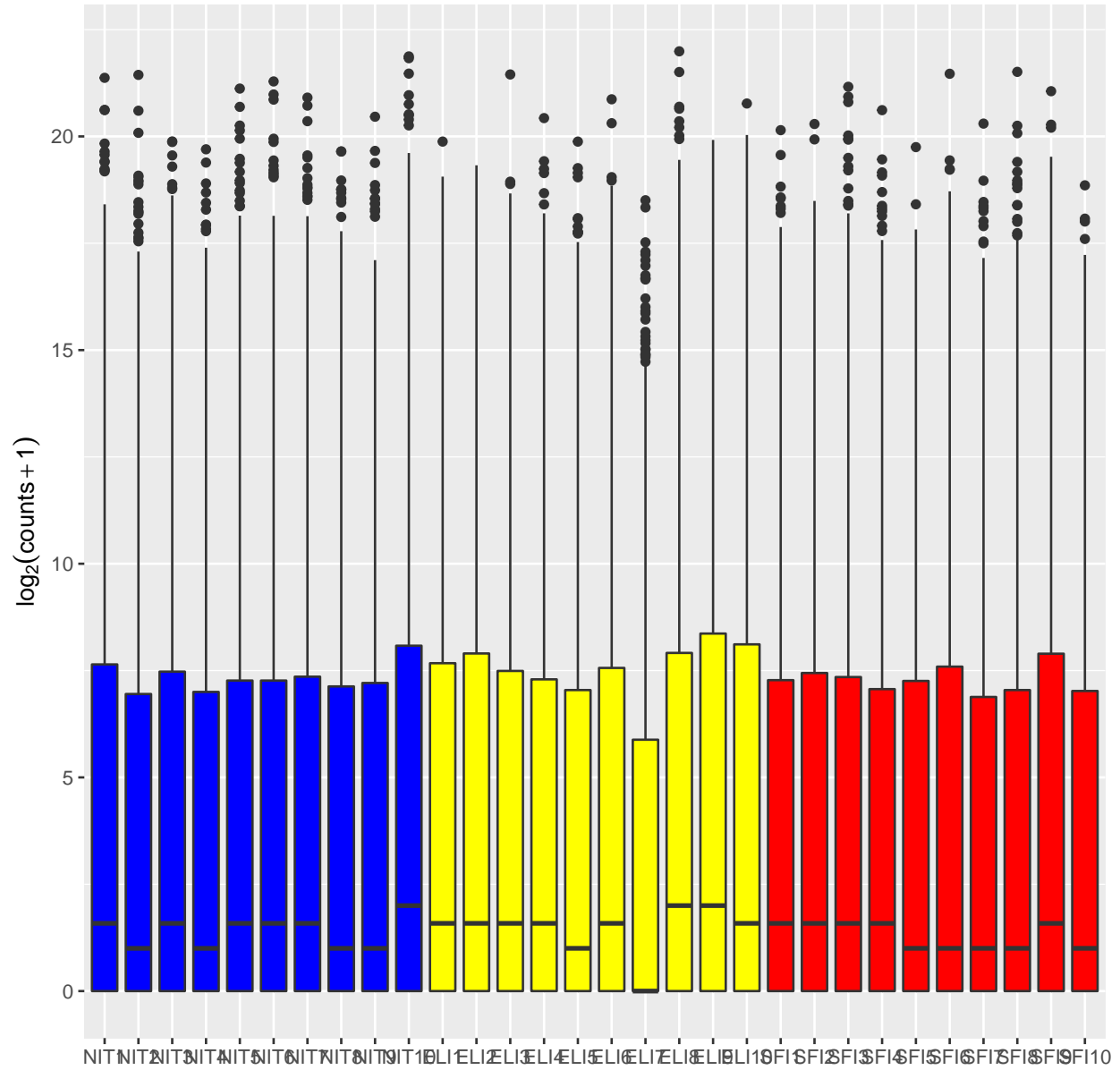


Figure 2: Diagrama de cajas que muestran la distribución de los diferentes tipos de tejido. Se puede observar cierta heterogeneidad, sobretodo en la muestra 7 del tipo tisular NIT, pero no parece que sea sistemática ni que sea producido por muestras defectuosas.

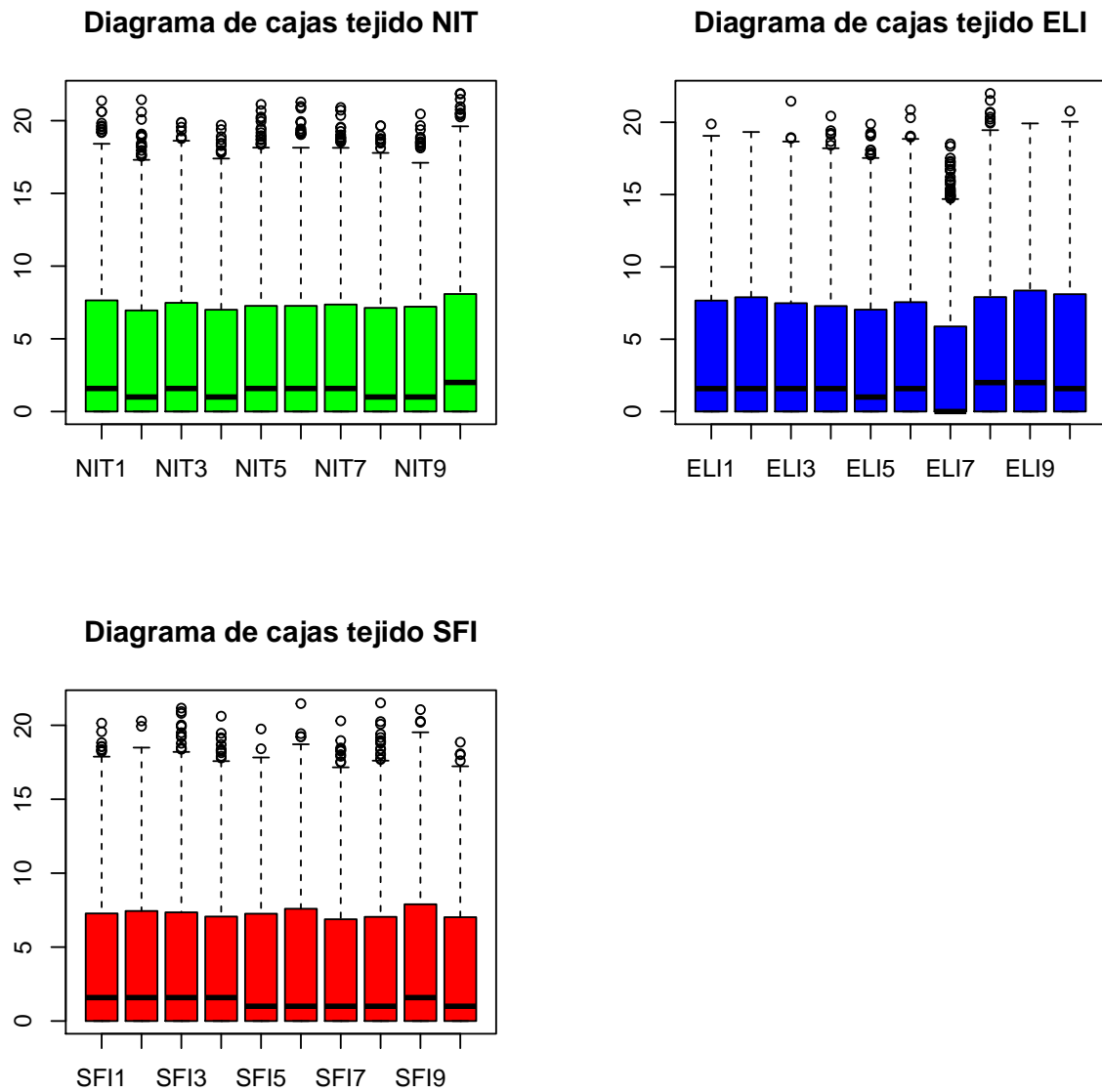


Figure 3: Diagrama de cajas que muestran la distribución de los diferentes tipos de tejido. Se puede observar cierta heterogeneidad, sobretodo en la muestra 7 del tipo tisular NIT, pero no parece que sea sistemática ni que sea producido por muestras defectuosas.

6 Resultados

7 Discusión

Bibliografía

Baumbach, Janina, Mitchell P Levesque, and Jordan W Raff. 2012. “Centrosome Loss or Amplification Does Not Dramatically Perturb Global Gene Expression in *Drosophila*.” *Biology Open* 1 (10). The Company of Biologists Ltd: 983–93.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1). Wiley Online Library: 289–300.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. “limma powers differential expression analyses for RNA-sequencing and microarray studies.” *Nucleic Acids Research* 43 (7): e47–e47. doi:[10.1093/nar/gkv007](https://doi.org/10.1093/nar/gkv007).