Análisis transcriptómicos de la expresión génica

Máster Universitario en Bioinformática

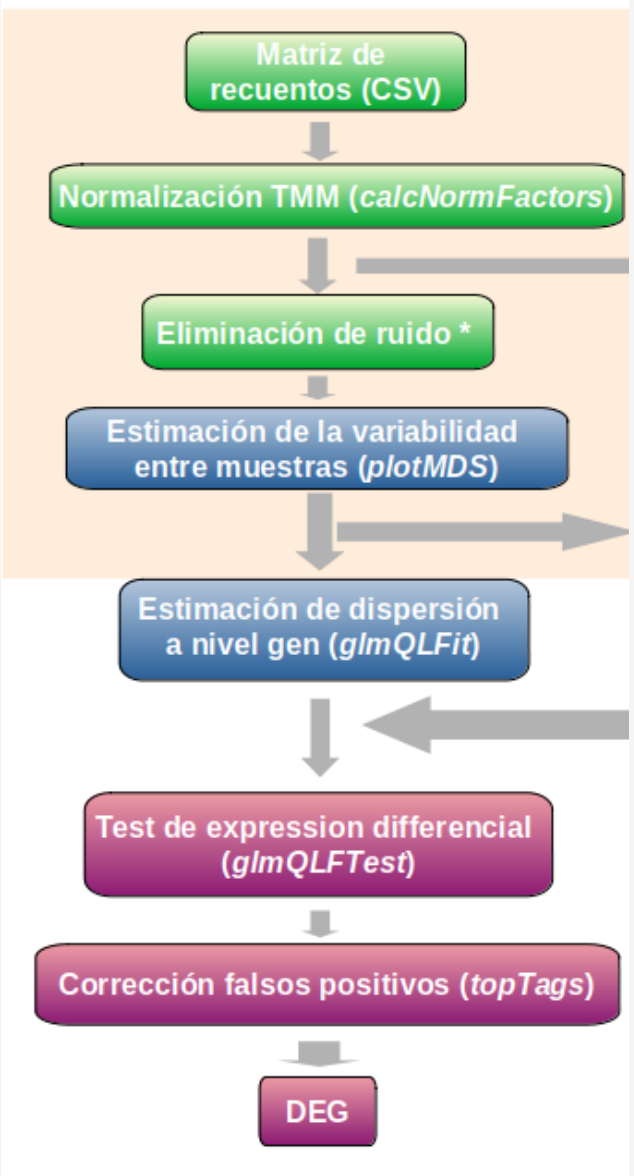
Sesión 8

Dra. Paula Soler Vila
paula.solerv@professor.universidadviu.com

De:
Planeta Formación y Universidades

Bloque IV: Análisis estadístico de la diferencia de expresión

Objetivos



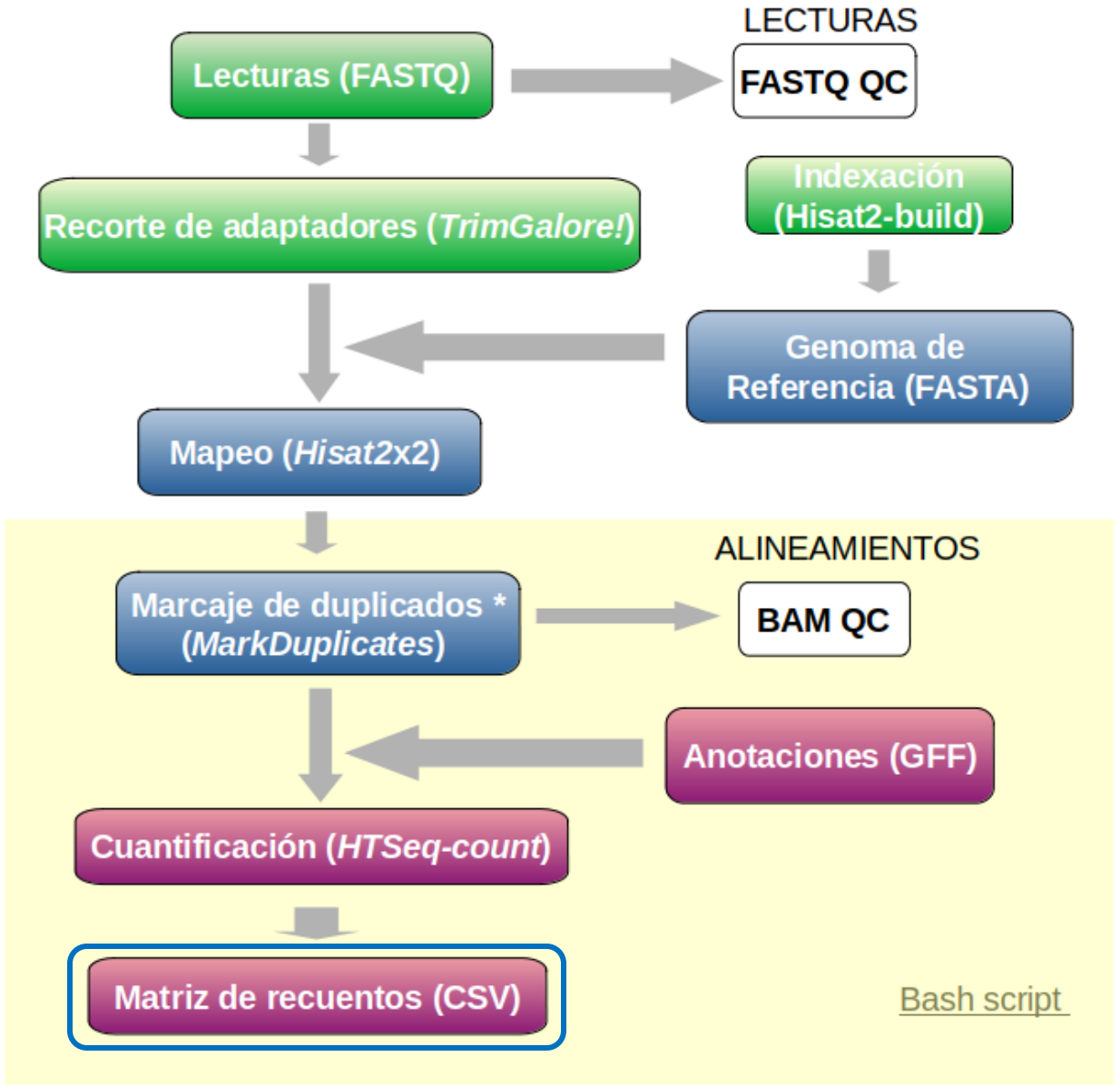
1. Conocer los principales pasos del análisis estadístico de la expresión diferencial con **edgeR**.

2. Saber identificar la información organizada en los archivos del **diseño experimental** y la **matriz de recuentos**.

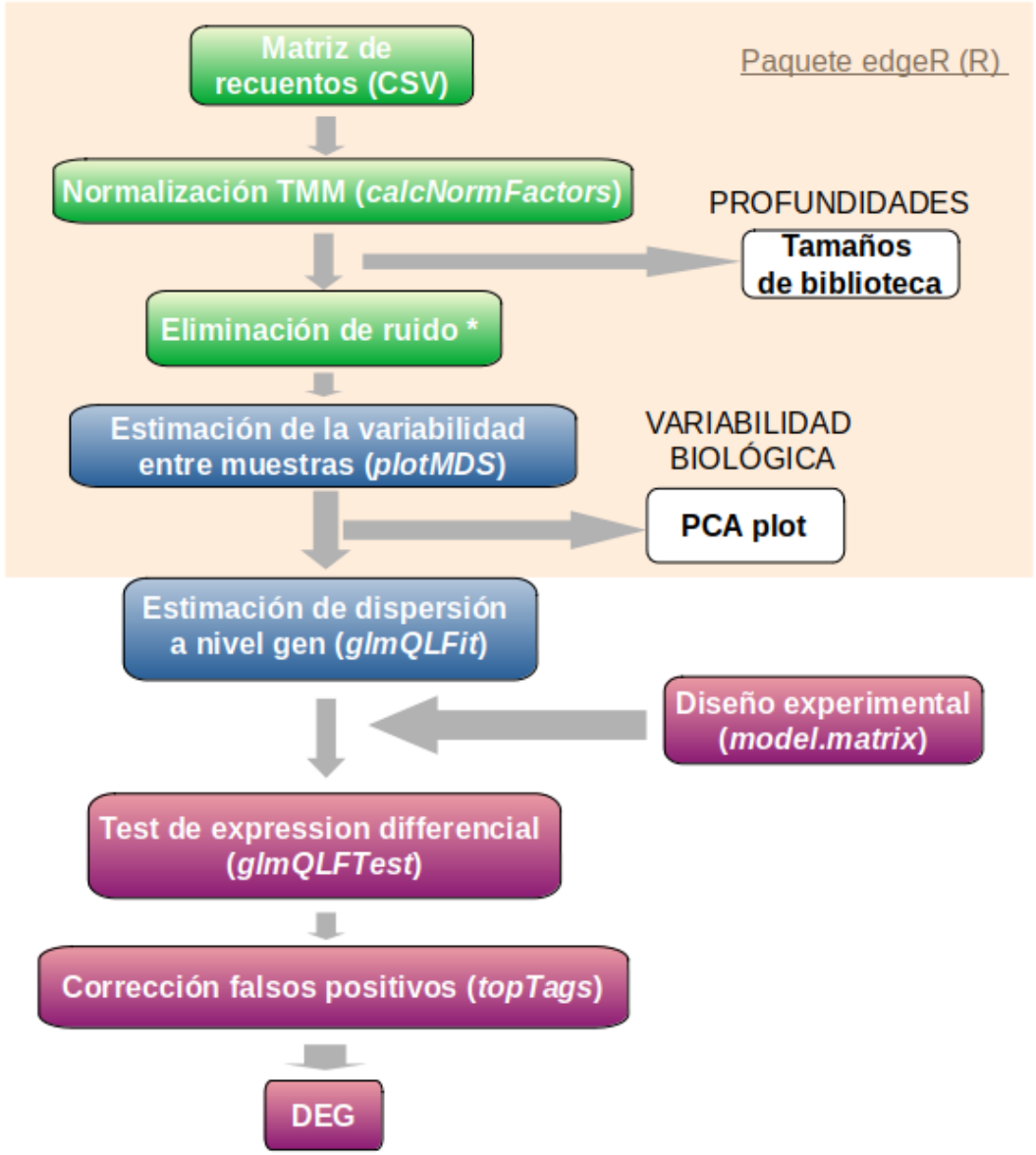
3. Entender y conocer la distribución de los datos de RNA-seq.

4. Entender las ventajas de la eliminación de genes de baja y nula expresión.

Flujo de trabajo del análisis de datos de *RNA-seq* (NGS)



Flujo de trabajo del análisis estadístico de la expresión génica



Flujo de trabajo del análisis estadístico de la expresión génica

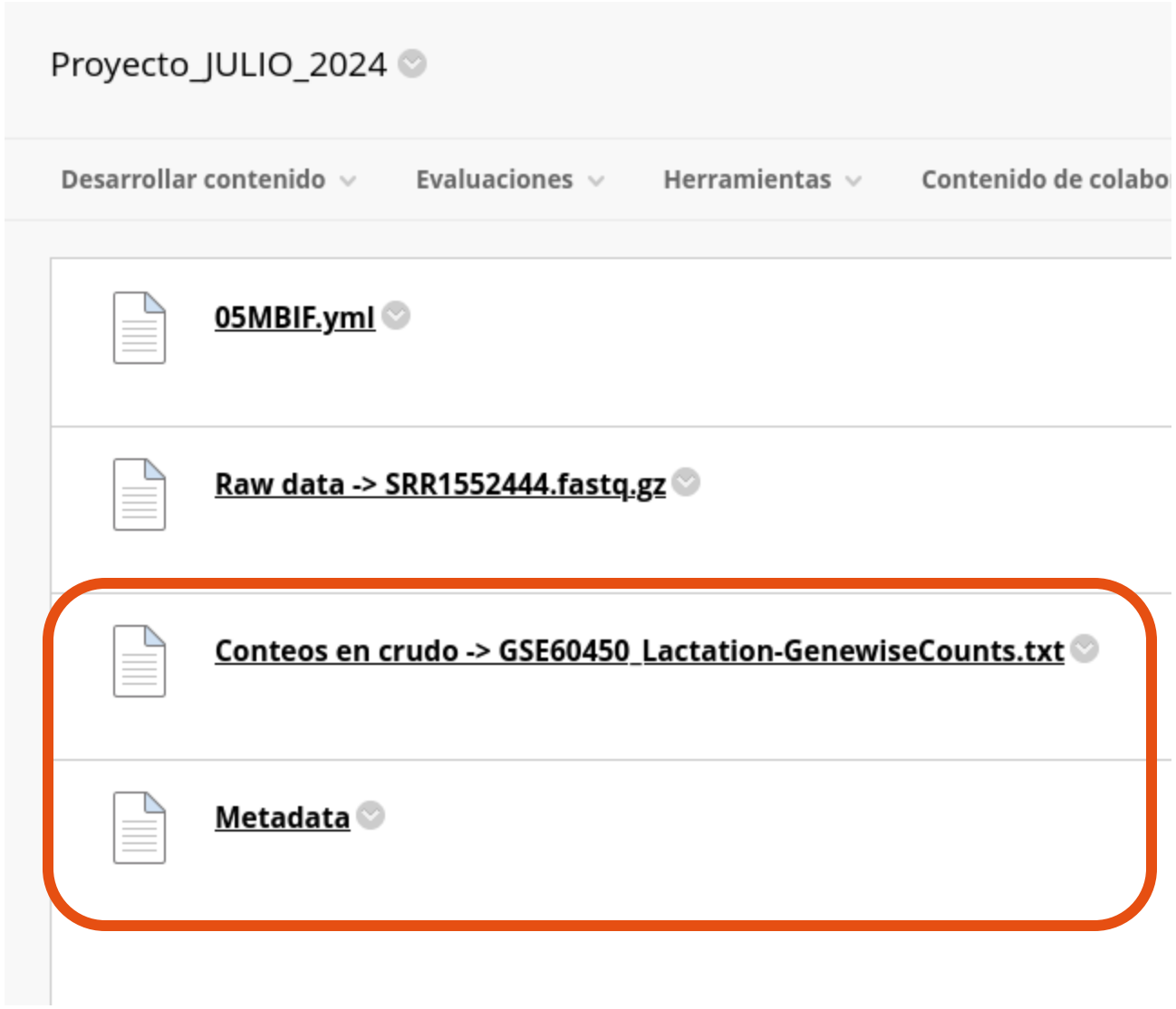
Tabla final que reúne los valores de expresión de cada gen en cada una de las muestras.

samples: want to see if differences across condition are significant (w.r.t. biological and technical variation)

features (e.g. genes)

	SRR1039508	SRR1039509	SRR1039512	SRR1039513	SRR1039516
ENSG00000000003	679	448	873	408	1138
ENSG00000000005	0	0	0	0	0
ENSG00000000419	467	515	621	365	587
ENSG00000000457	260	211	263	164	245
ENSG00000000460	60	55	40	35	78

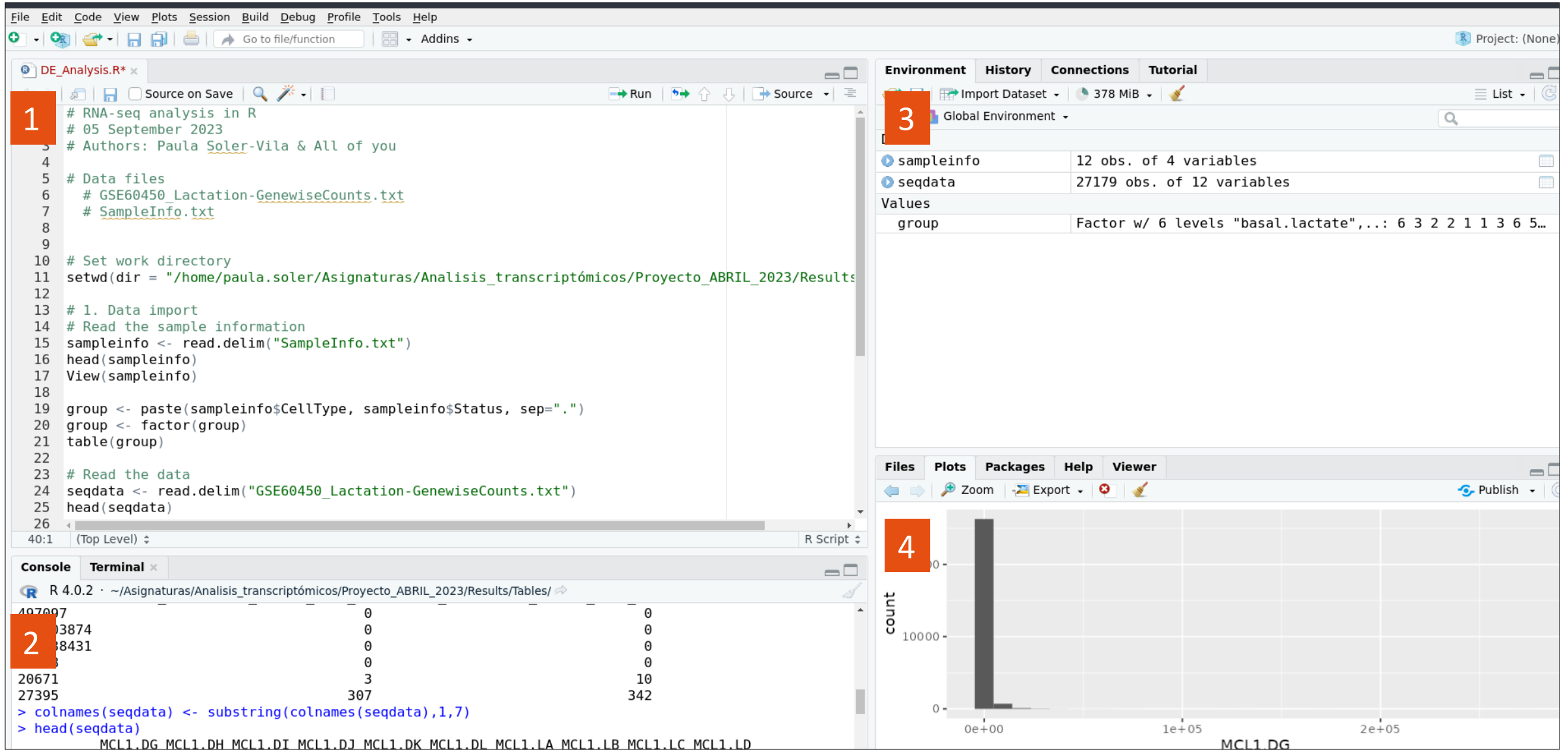
Matriz de recuentos y metadatos



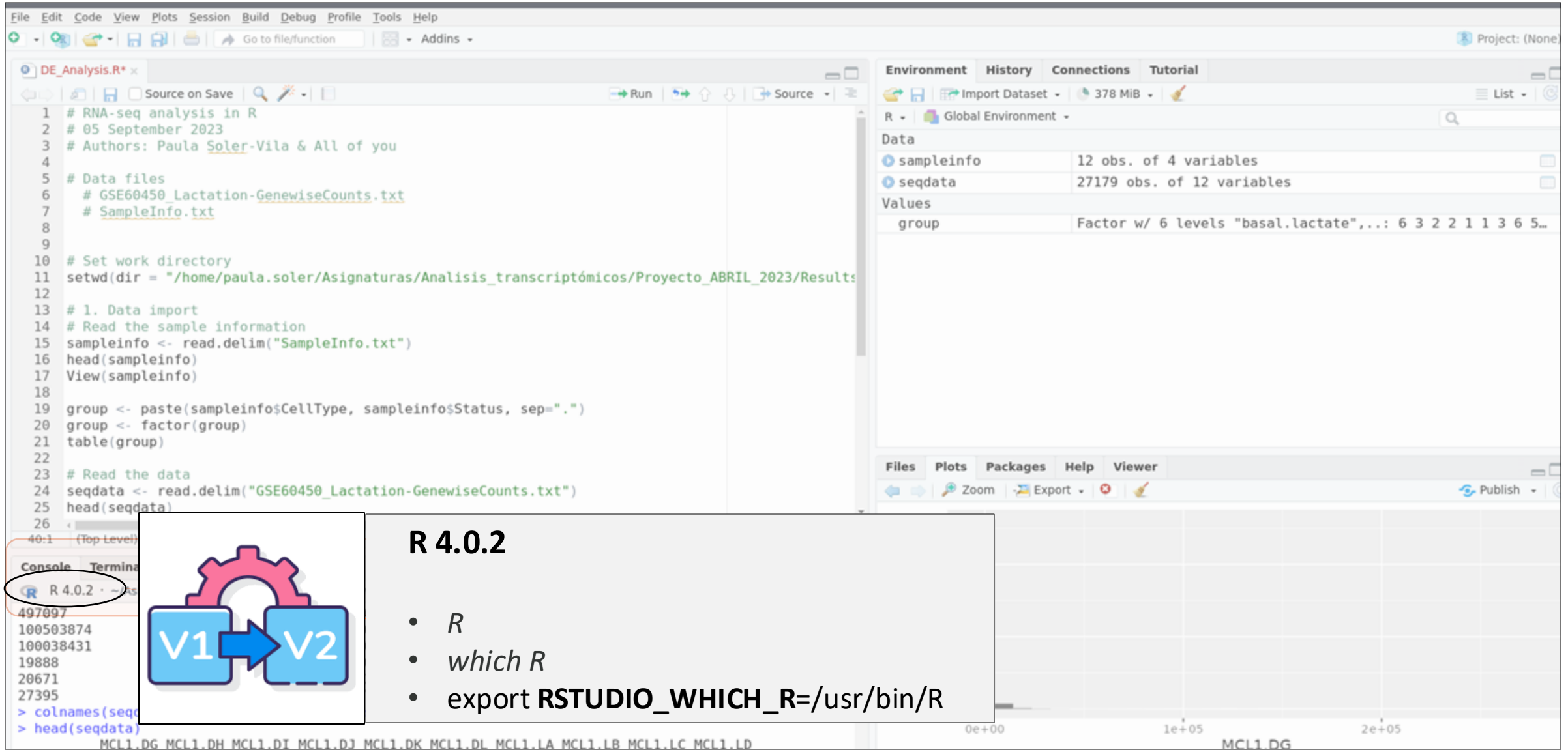
Botón derecho -> Guardar enlace como...

```
(05MBIF) Results]$ tree
.
├── GSE60450_Lactation-GenewiseCounts.txt
└── SRR1552444_counts.tsv
```

R / Rstudio (IDE)



Versión de R



R 4.0.2

- *R*
- *which R*
- export **RSTUDIO_WHICH_R**=/usr/bin/R

PRACTIQUEMOS

amazon WorkSpaces

Adecuando la matriz de metadatos

```
# RNA-seq analysis in R
# 25 July 2024
# Authors: Paula Soler-Vila & All of you
# Data files
  # GSE60450_Lactation-GenewiseCounts.txt
  # Metadata.txt


# Set work directory
setwd(dir = "Analisis_transcriptómicos/Proyecto_JULIO_2024/


# 1. Data import
# Read the sample information
sampleinfo <- read.delim(file = "Metadata")
head(sampleinfo)
View(sampleinfo)


group <- paste(sampleinfo$CellType, sampleinfo$Status, sep=".")
group <- factor(group)
table(group)
```

Adecuando la matriz de recuentos

```
# Read the data
seqdata <- read.delim("Results/GSE60450_Lactation-GenewiseCounts.txt")
head(seqdata)

seqdata <- read.delim("Results/GSE60450_Lactation-GenewiseCounts.txt", row.names = "EntrezGeneID")
head(seqdata)

seqdata <- seqdata[,2:ncol(seqdata)]
dim(seqdata)
head(seqdata)

colnames(seqdata) <- substr(colnames(seqdata),1,7)
head(seqdata)

           MCL1.DG MCL1.DH MCL1.DI MCL1.DJ MCL1.DK MCL1.DL MCL1.LA MCL1.LB MCL1.LC MCL1.LD MCL1.LE MCL1.LF
497097      438    300     65     237    354    287      0      0      0      0      0      0
100503874     1      0      1      1      0      4      0      0      0      0      0      0
100038431     0      0      0      0      0      0      0      0      0      0      0      0
19888         1      1      0      0      0      0     10      3     10      2      0      0
20671       106    182     82    105     43     82     16     25     18      8      3     10
```

Leyendo los datos de conteo

Si los datos de conteo están contenidos en un solo archivo de texto delimitado por tabulaciones o separado por comas con varias columnas, una para cada muestra, entonces el método más simple suele ser leer el archivo en R utilizando **read.delim**.

Si los recuentos de diferentes muestras se almacenan en archivos separados, entonces los archivos deben leerse separados y unirlos. La función **readDGE** dentro de la librería **EdgeR** se proporciona para hacer esto.

	MCL1.DG	MCL1.DH	MCL1.DI	MCL1.DJ	MCL1.DK	MCL1.DL	MCL1.LA	MCL1.LB	MCL1.LC	MCL1.LD	MCL1.LE	MCL1.LF
497097	438	300	65	237	354	287	0	0	0	0	0	0
100503874	1	0	1	1	0	4	0	0	0	0	0	0
100038431	0	0	0	0	0	0	0	0	0	0	0	0
19888	1	1	0	0	0	0	10	3	10	2	0	0
20671	106	182	82	105	43	82	16	25	18	8	3	10
27395	309	234	337	300	290	270	560	464	489	328	307	342
18777	652	515	948	935	928	791	826	862	668	646	544	581
100503730	0	1	0	0	0	0	0	1	2	0	2	2
21399	1604	1495	1721	1317	1159	1066	1334	1258	1068	926	508	500
58175	4	2	14	4	2	2	170	165	138	60	27	15
108664	769	752	1062	987	995	903	1381	1430	1762	1570	1330	1296

Entrez ID

Añadiendo anotación génica

Gene | Gene | 497097[uid]

Create RSS | Save search | Advanced

Full Report | Send to:

Showing Current items.

Xkr4 X-linked Kx blood group related 4 [*Mus musculus* (house mouse)]

Gene ID: 497097, updated on 5-Aug-2023

Download Datasets

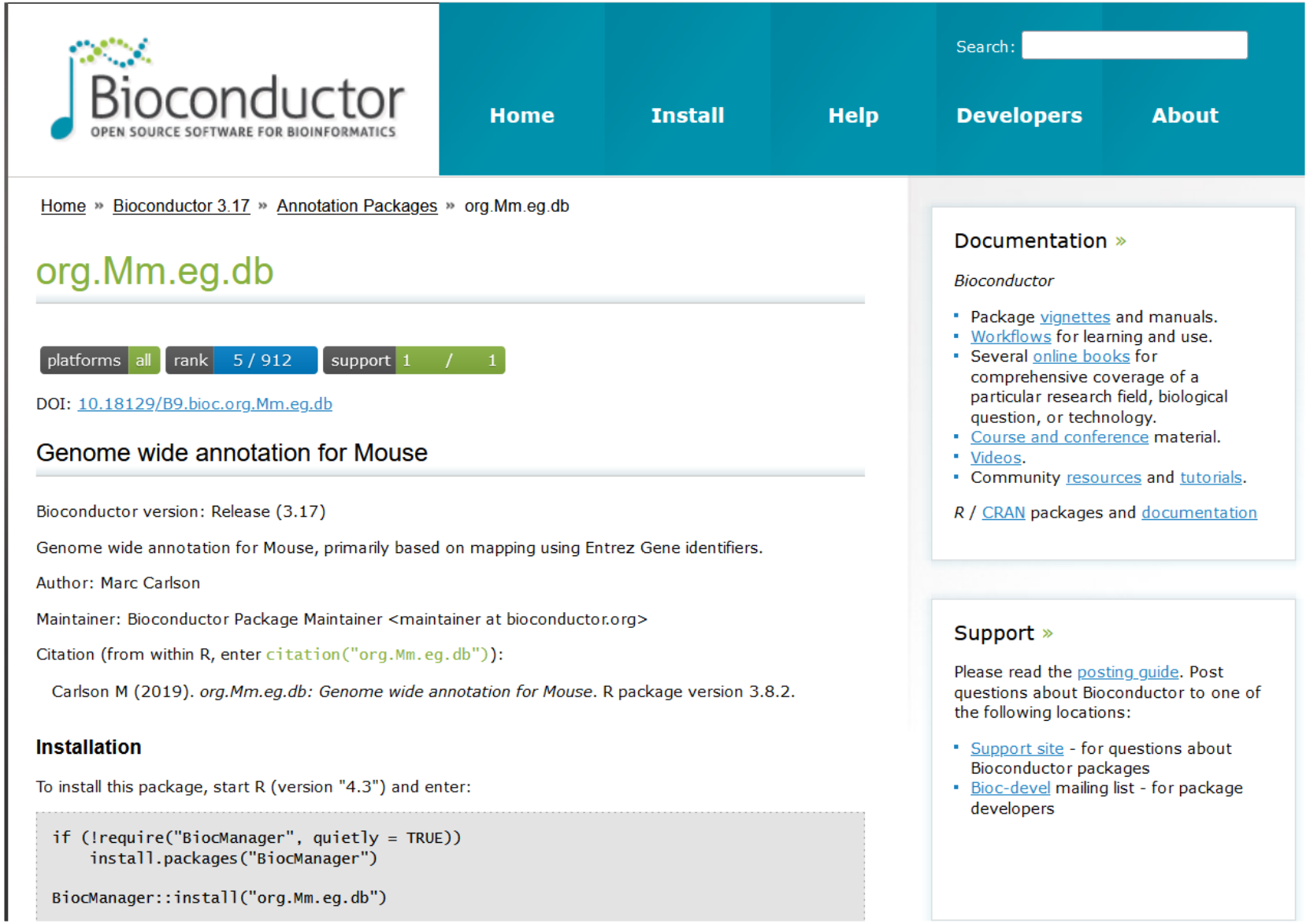
Summary

Official Symbol	Xkr4 provided by MGI
Official Full Name	X-linked Kx blood group related 4 provided by MGI
Primary source	MGI:MGI:3528744
See related	Ensembl:ENSMUSG00000051951 AllianceGenome:MGI:3528744
Gene type	protein coding
RefSeq status	PROVISIONAL
Organism	Mus musculus
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Myomorpha; Muroidea; Muridae; Murinae; Mus; Mus
Also known as	XRG4; Gm210; mKIAA1889
Summary	Enables phospholipid scramblase activity and protein homodimerization activity. Involved in phosphatidylserine exposure on apoptotic cell surface. Located in plasma membrane. Orthologous to human XKR4 (XK related 4). [provided by Alliance of Genome Resources, Apr 2022]
Expression	Biased expression in CNS E18 (RPKM 4.4), frontal lobe adult (RPKM 2.6) and 5 other tissues See more
Orthologs	human all

NEW **Try the new Gene table**

Try the new Transcript table

Añadiendo anotación génica -> org.Mn.eg.dg package



Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home | Install | Help | Developers | About

Search:

Home » Bioconductor 3.17 » Annotation Packages » org.Mm.eg.db

org.Mm.eg.db

platforms all | rank 5 / 912 | support 1 / 1

DOI: 10.18129/B9.bioc.org.Mm.eg.db

Genome wide annotation for Mouse

Bioconductor version: Release (3.17)

Genome wide annotation for Mouse, primarily based on mapping using Entrez Gene identifiers.

Author: Marc Carlson

Maintainer: Bioconductor Package Maintainer <maintainer at bioconductor.org>

Citation (from within R, enter `citation("org.Mm.eg.db")`):

Carlson M (2019). *org.Mm.eg.db: Genome wide annotation for Mouse*. R package version 3.8.2.

Installation

To install this package, start R (version "4.3") and enter:

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

BiocManager::install("org.Mm.eg.db")
```

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
- Bioc-devel mailing list - for package developers

Instalación de org.Mn.eg.dg package

```
# Instalación

if (!require("BiocManager", quietly = TRUE))

  install.packages("BiocManager")

BiocManager::install("org.Mm.eg.db")
```

```
> packageVersion("org.Mm.eg.db")
[1] ‘3.12.0’
```

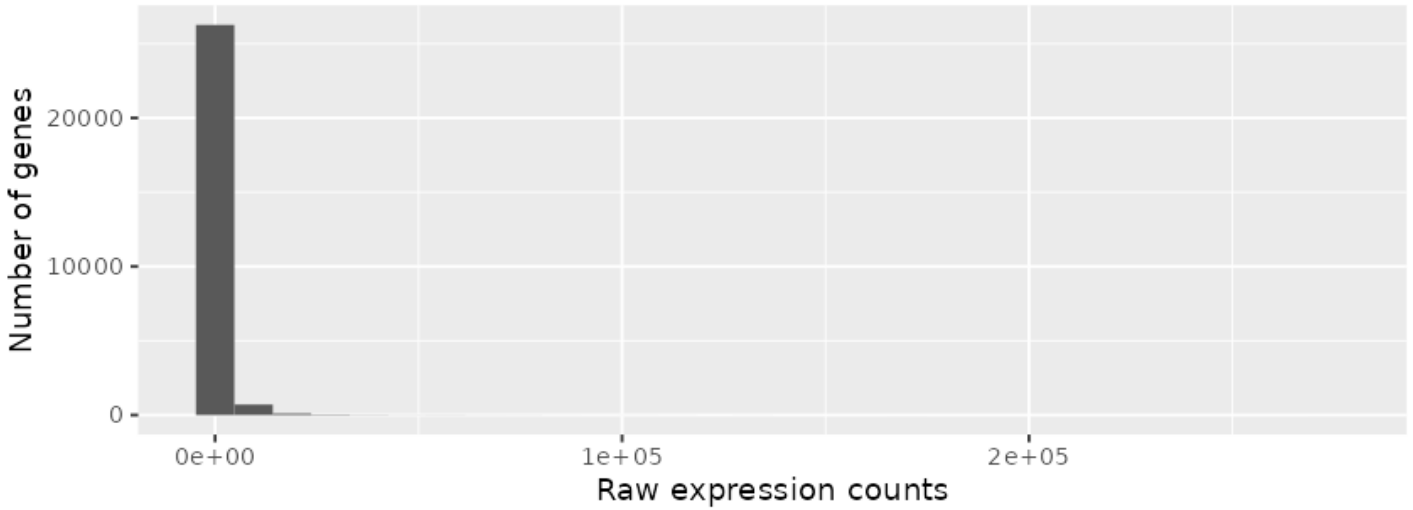
```
library(org.Mm.eg.db)
columns(org.Mm.eg.db)

 [1] "ACCNUM"      "ALIAS"       "ENSEMBL"     "ENSEMBLPROT"  "ENSEMBLTRANS"
"ENTREZID"    "ENZYME"      "EVIDENCE"    "EVIDENCEALL"
[10]
"GENENAME"    "GO"          "GOALL"       "IPI"         "MGI"         "ONTOLOGY"    "ONTOLOGYALL"  "PATH"        "PFAM"
[19] "PMID"        "PROSITE"     "REFSEQ"      "SYMBOL"      "UNIGENE"     "UNIPROT"
```

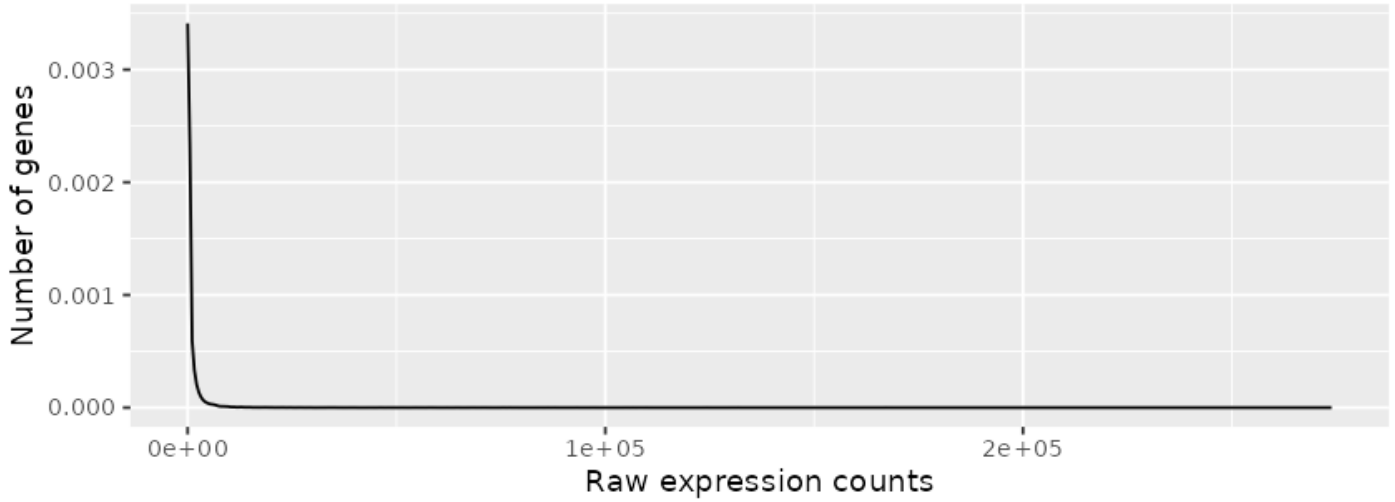
Leyendo los datos de conteo

	MCL1.DG	MCL1.DH	MCL1.DI	MCL1.DJ	MCL1.DK	MCL1.DL	MCL1.LA	MCL1.LB	MCL1.LC	MCL1.LD	MCL1.LE	MCL1.LF
497097	438	300	65	237	354	287	0	0	0	0	0	0
100503874	1	0	1	1	0	4	0	0	0	0	0	0
100038431	0	0	0	0	0	0	0	0	0	0	0	0
19888	1	1	0	0	0	0	10	3	10	2	0	0
20671	106	182	82	105	43	82	16	25	18	8	3	10
27395	309	234	337	300	290	270	560	464	489	328	307	342
18777	652	515	948	935	928	791	826	862	668	646	544	581
100503730	0	1	0	0	0	0	0	1	2	0	2	2
21399	1604	1495	1721	1317	1159	1066	1334	1258	1068	926	508	500
58175	4	2	14	4	2	2	170	165	138	60	27	15
108664	769	752	1062	987	995	903	1381	1430	1762	1570	1330	1296

```
> max(seqdata)
[1] 2887969
> min(seqdata)
[1] 0
```

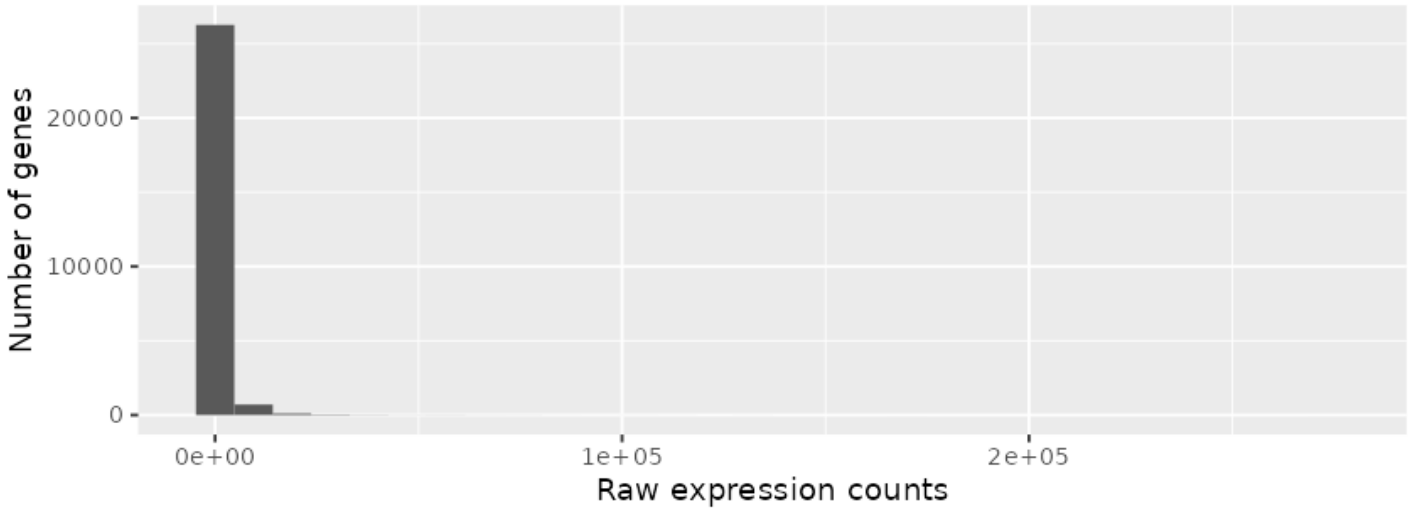
Distribución de los datos de RNA-seq





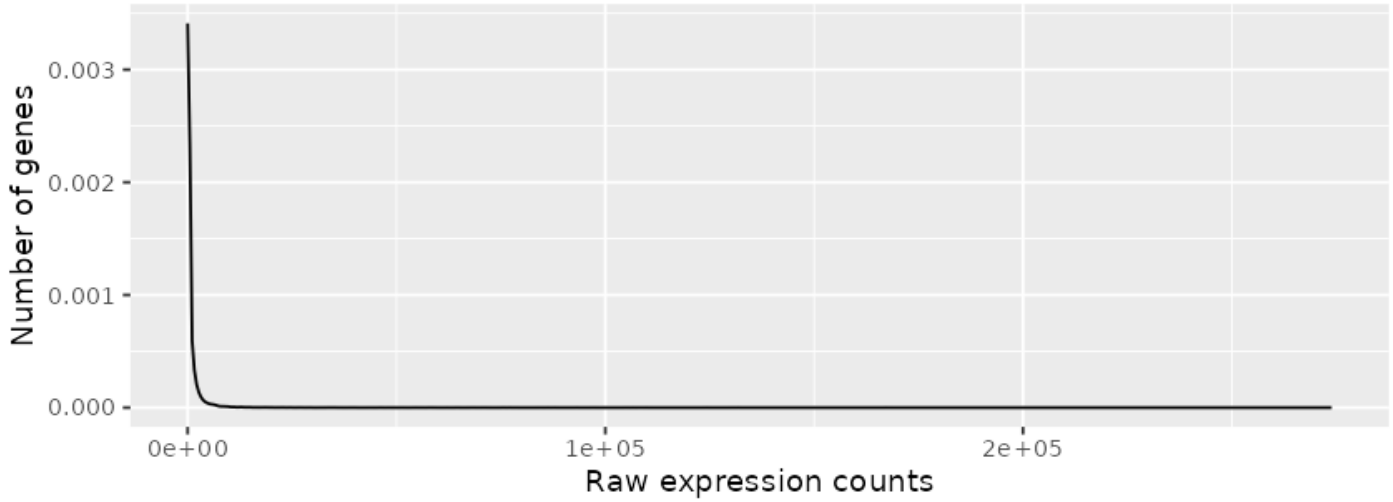

```
#install.packages(ggplot2)


library(ggplot2)
ggplot(seqdata) +
geom_histogram(aes(x=MCL1.DG)) +
# geom_density
  xlab("Raw expression counts") +
  ylab("Number of genes")
```

Distribución de los datos de RNA-seq



1 **Rango dinámico amplio**

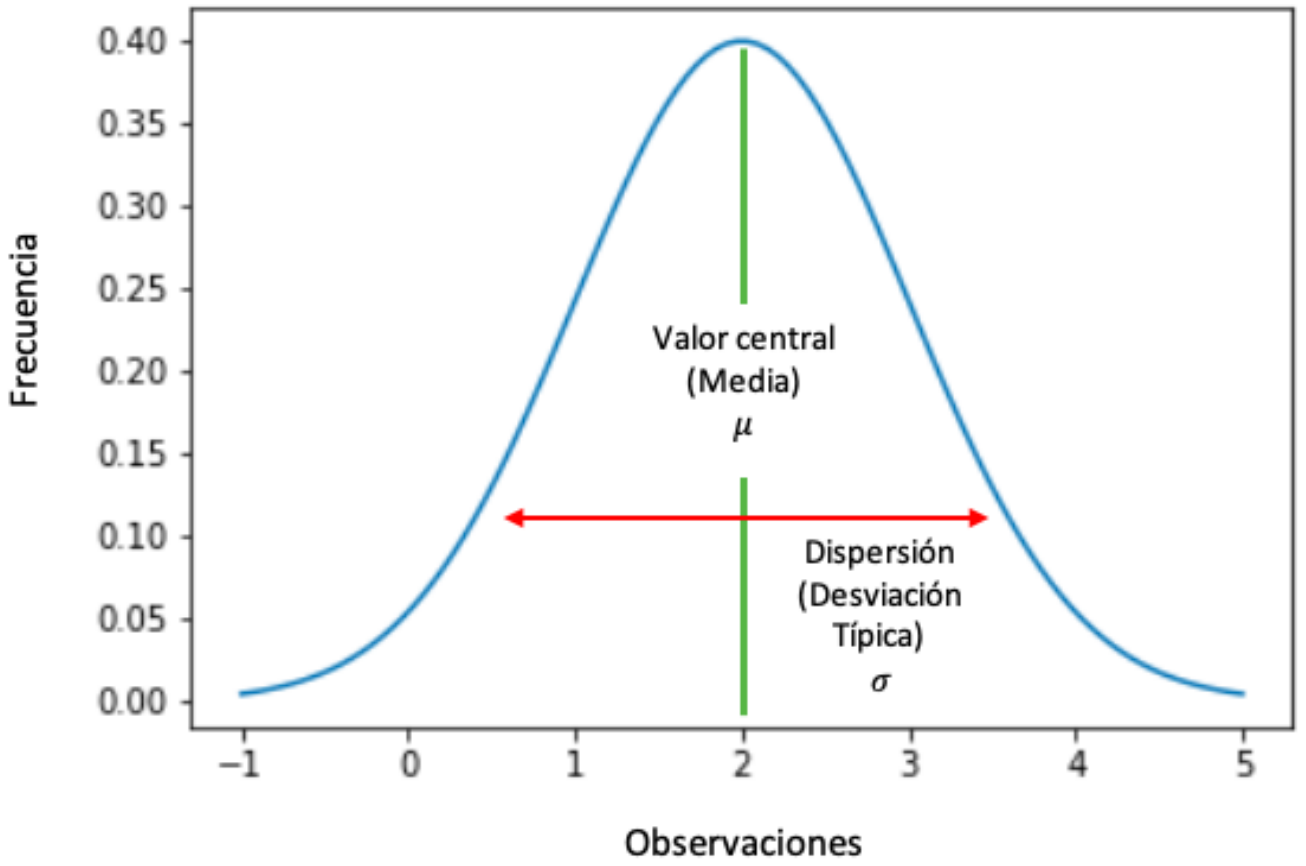
2 **Elevada proporción de valores bajos**



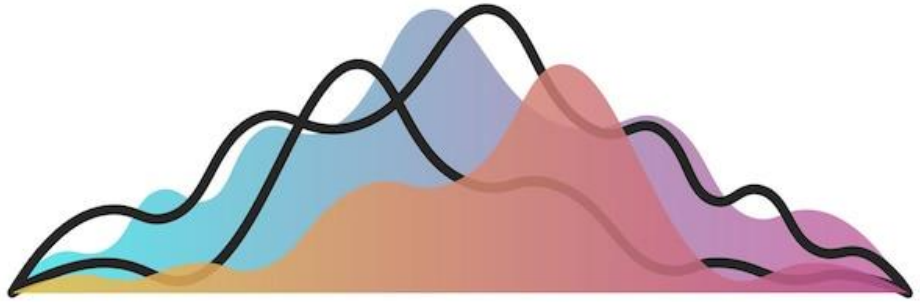
3 **Distribución sesgada a la derecha**

La distribución normal

- Distribución Gaussiana o campana de Gauss
- Forma de campana simétrica
- Variables continuas
- La media (μ) representa el valor central o promedio de la distribución
- La desviación estándar (σ) mide la dispersión o la variabilidad de los datos.



Modelado de datos de conteo



Distribución binomial

- Basado en eventos **discretos.**
- La probabilidad de que ocurra un evento es relativamente **alta**.
- Se aplica cuando hay una cantidad **finita** de posibilidades.

Distribución de Poisson

- Basado en eventos **discretos.**
- La probabilidad de que ocurra un evento es **baja**.
- Empleado cuando el número de casos es muy grande.

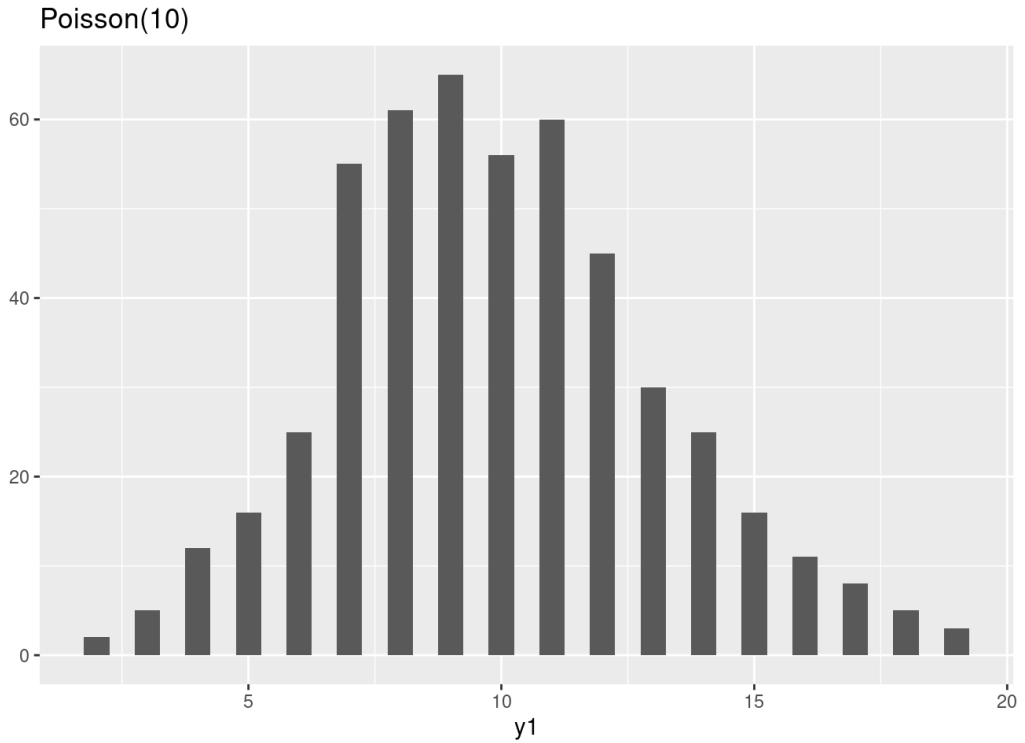
¿Por qué la distribución de Poisson NO funciona?

- La distribución de Poisson tiene un único parámetro, **lambda (λ),** que representa la tasa promedio de ocurrencia de eventos en el intervalo de tiempo o área considerada.

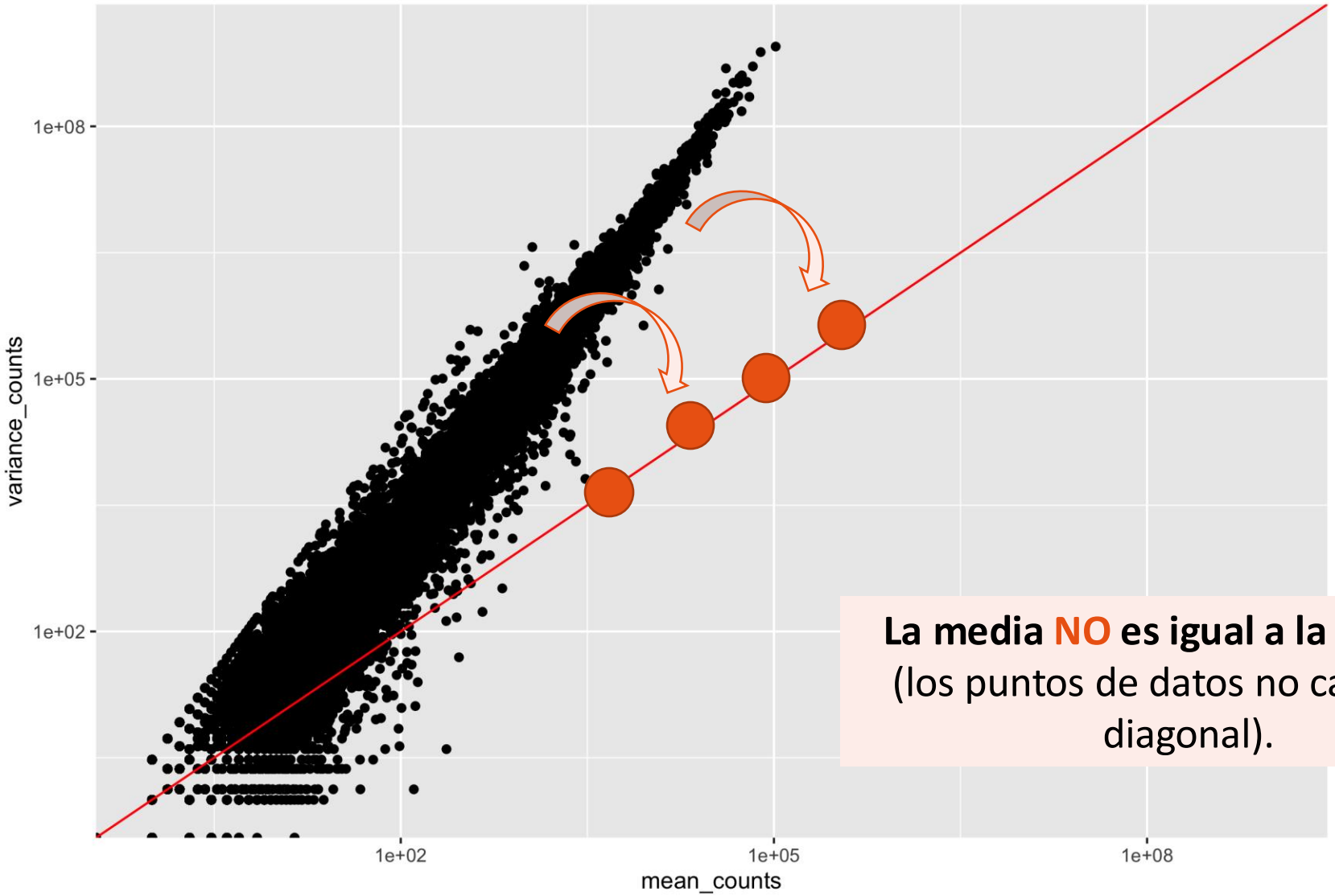
- **Media y varianza iguales:** Tanto la media como la varianza en una distribución de Poisson son iguales al valor de lambda.

  $(\mu = \sigma^2 = \lambda)$.

λ = media = varianza

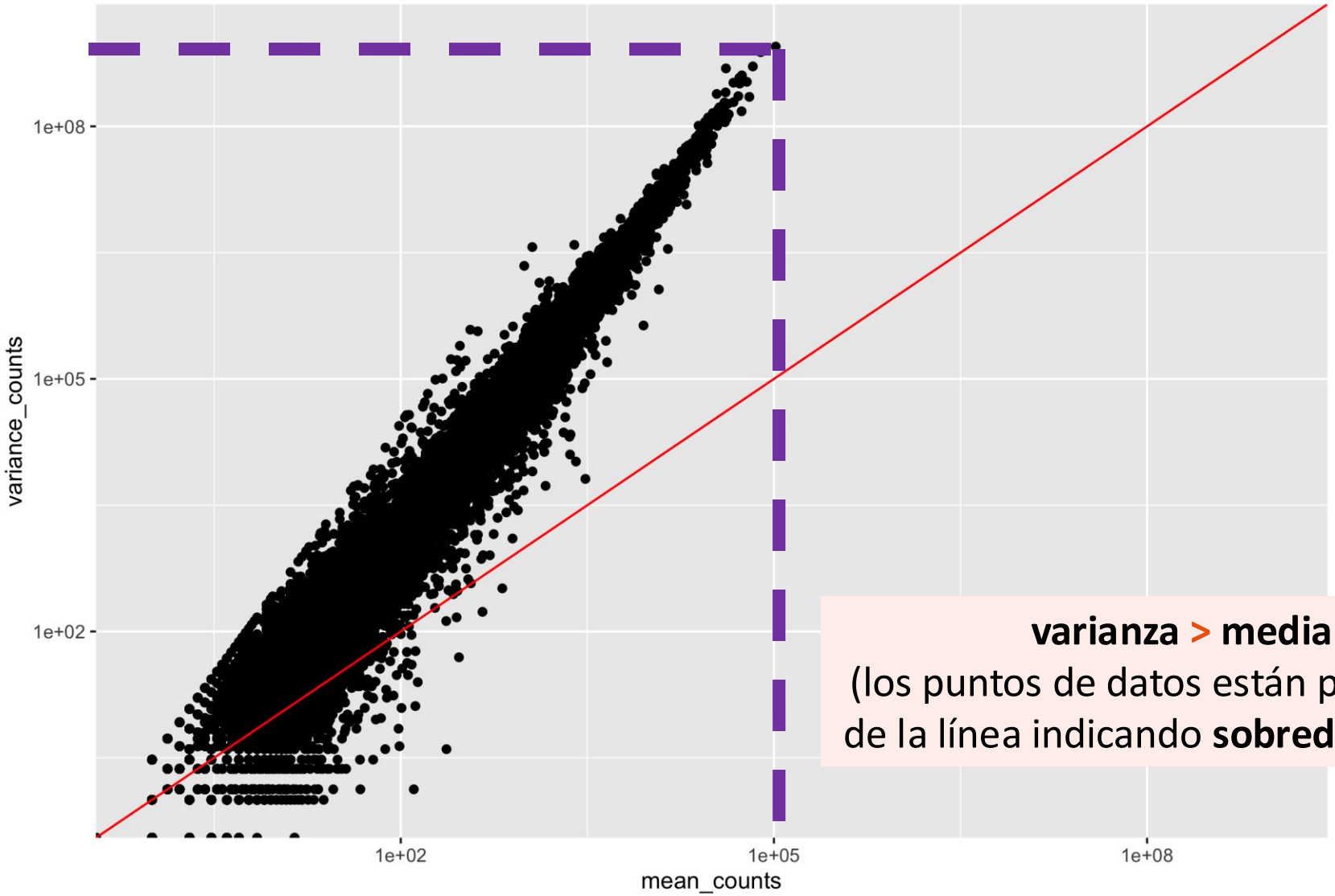


¿Qué empleamos para los datos de recuento de RNA-seq?



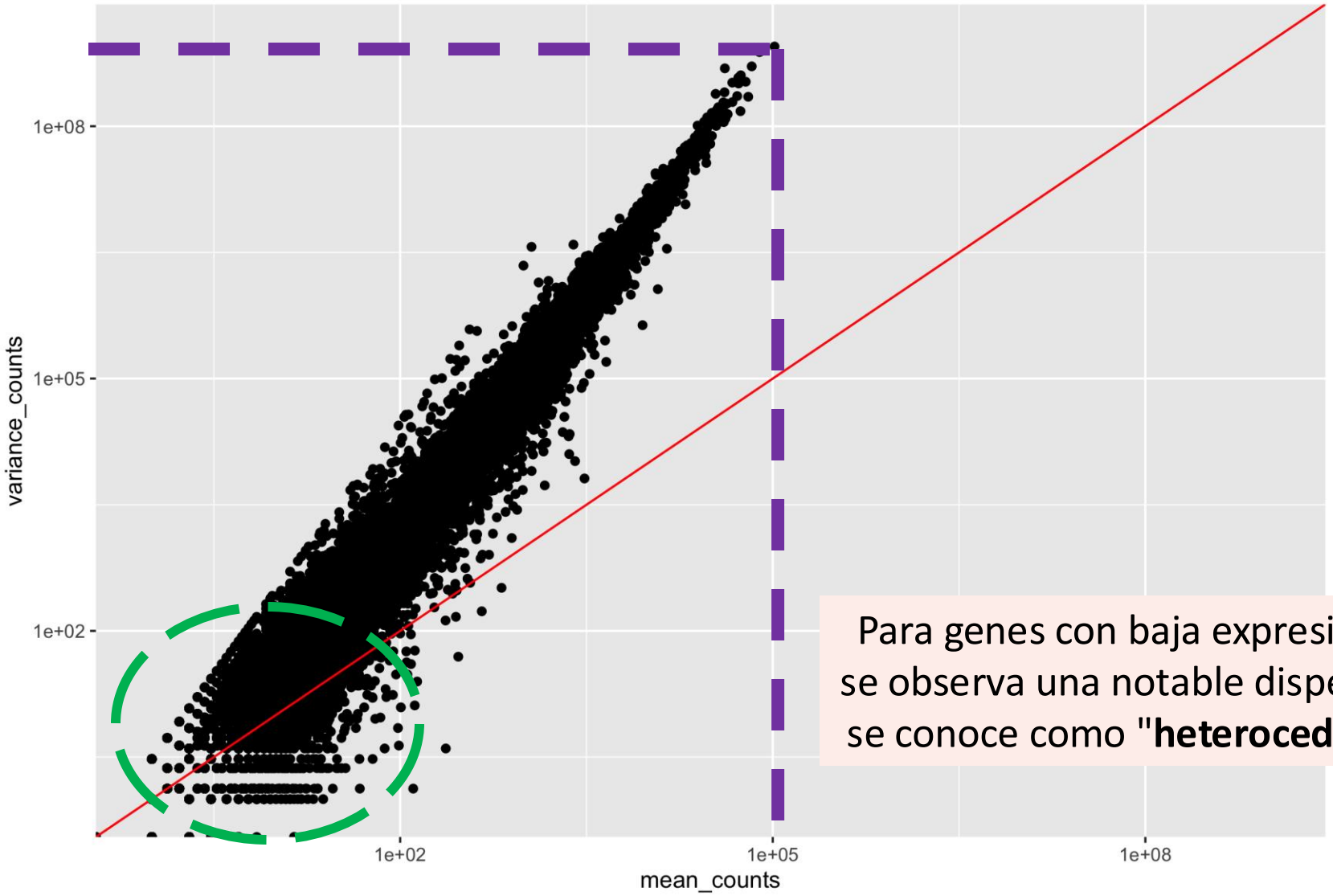
La media NO es igual a la varianza (los puntos de datos no caen en la diagonal).

¿Qué empleamos para los datos de recuento de RNA-seq?



varianza > media
(los puntos de datos están por encima de la línea indicando **sobredispersión**).

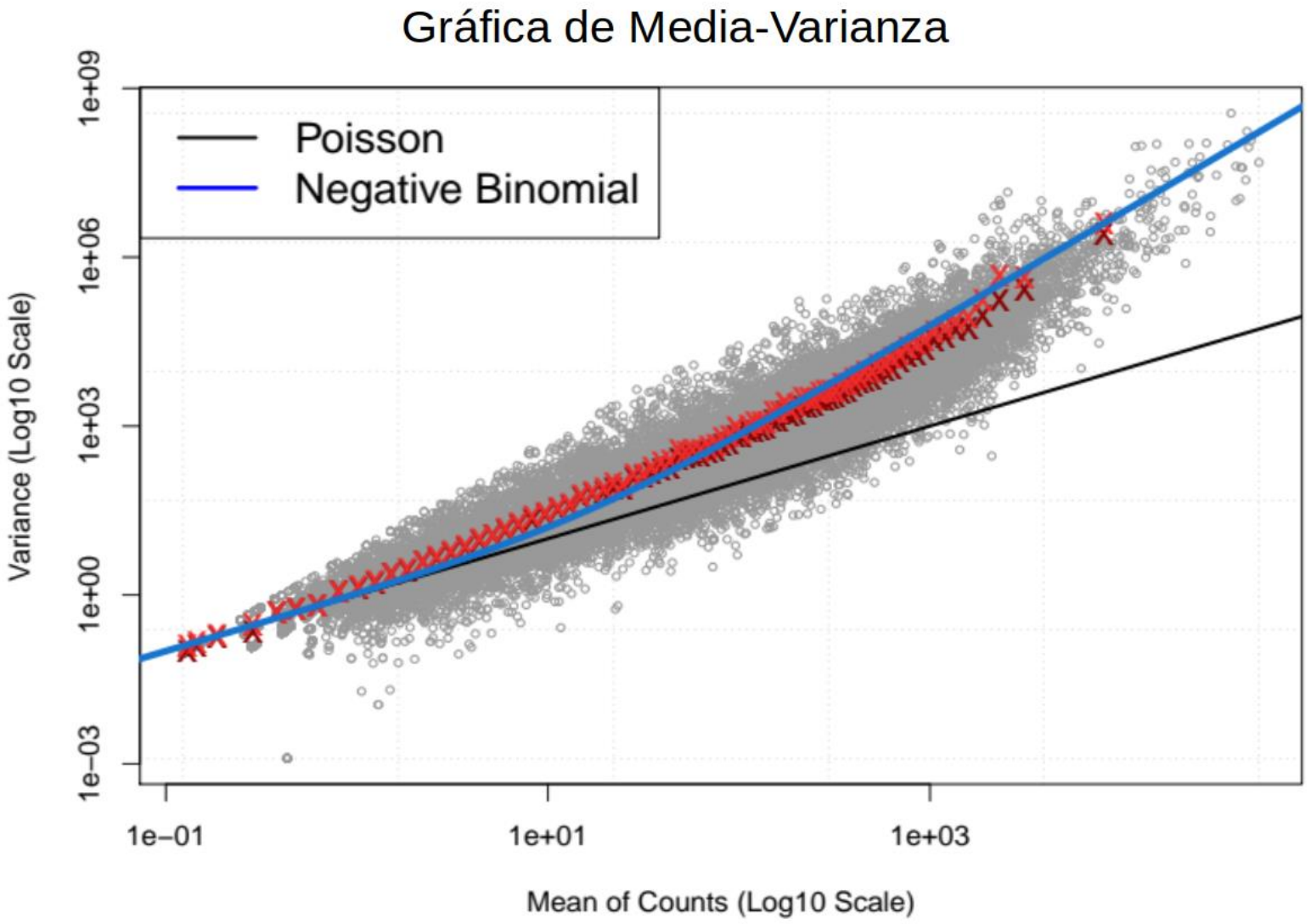
¿Qué empleamos para los datos de recuento de RNA-seq?



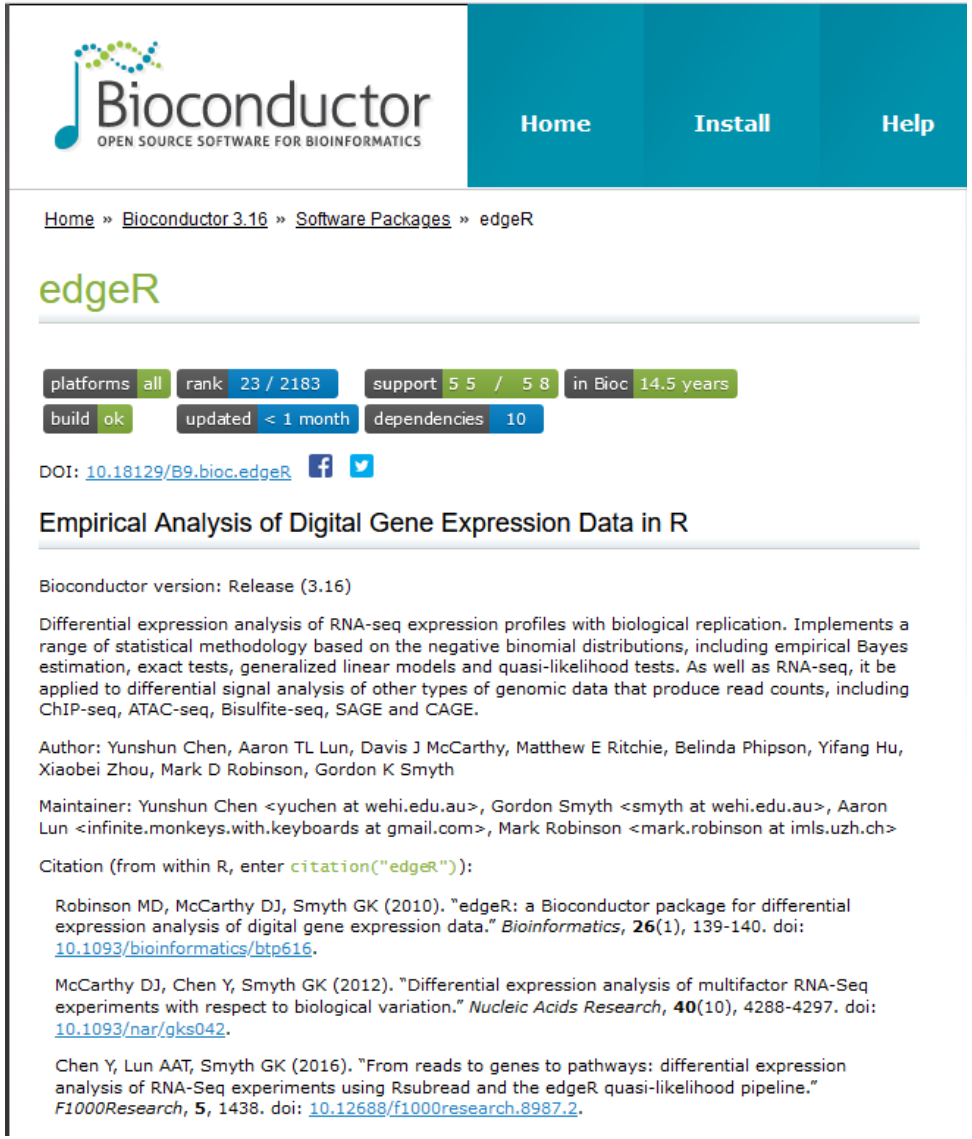
Para genes con baja expresión media, se observa una notable dispersión. Esto se conoce como "**heterocedasticidad**".

La alternativa: Distribución binominal negativa

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



Introducción a EdgeR/DESeq2



Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home Install Help

Home » Bioconductor 3.16 » Software Packages » edgeR

edgeR

platforms all | rank 23 / 2183 | support 5 5 / 5 8 | in Bioc 14.5 years
build ok | updated < 1 month | dependencies 10

DOI: 10.18129/B9.bioc.edgeR

Empirical Analysis of Digital Gene Expression Data in R

Bioconductor version: Release (3.16)

Differential expression analysis of RNA-seq expression profiles with biological replication. Implements a range of statistical methodology based on the negative binomial distributions, including empirical Bayes estimation, exact tests, generalized linear models and quasi-likelihood tests. As well as RNA-seq, it be applied to differential signal analysis of other types of genomic data that produce read counts, including ChIP-seq, ATAC-seq, Bisulfite-seq, SAGE and CAGE.

Author: Yunshun Chen, Aaron TL Lun, Davis J McCarthy, Matthew E Ritchie, Belinda Phipson, Yifang Hu, Xiaobei Zhou, Mark D Robinson, Gordon K Smyth

Maintainer: Yunshun Chen <yuchen at wehi.edu.au>, Gordon Smyth <smyth at wehi.edu.au>, Aaron Lun <infinite.monkeys.with.keyboards at gmail.com>, Mark Robinson <mark.robinson at imls.uzh.ch>

Citation (from within R, enter `citation("edgeR")`):

Robinson MD, McCarthy DJ, Smyth GK (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics*, **26**(1), 139-140. doi: 10.1093/bioinformatics/btp616.

McCarthy DJ, Chen Y, Smyth GK (2012). "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation." *Nucleic Acids Research*, **40**(10), 4288-4297. doi: 10.1093/nar/gks042.

Chen Y, Lun AAT, Smyth GK (2016). "From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline." *F1000Research*, **5**, 1438. doi: 10.12688/f1000research.8987.2.

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home Install Help

Home » Bioconductor 3.16 » Software Packages » DESeq2

DESeq2

platforms all | rank 29 / 2183 | support 250 / 254 | in Bioc 10 years
build ok | updated < 1 month | dependencies 90

DOI: 10.18129/B9.bioc.DESeq2

Differential gene expression analysis based on the negative binomial distribution

Bioconductor version: Release (3.16)

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], RADIANT EU FP7 [fnd], NIH NHGRI [fnd], CZI [fnd]

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.

Introducción a EdgeR/DESeq2

- Ambas herramientas están dentro de la categoría de **enfoque paramétricos**
    - **Modelos lineales generalizados binomiales negativos (*dispersión*)**
- Se recomienda un mínimo de **3 réplicas biológicas** por muestra para evitar desviaciones del modelo.

EdgeR's quasi-likelihood pipeline

- Control estricto en la tasa de error
- Mejoras en la velocidad
- Refinamientos estadísticos
- Funcionamiento mejor en situaciones con bajos conteos

Instalación EdgeR

Paso 1: Falta de dependencia

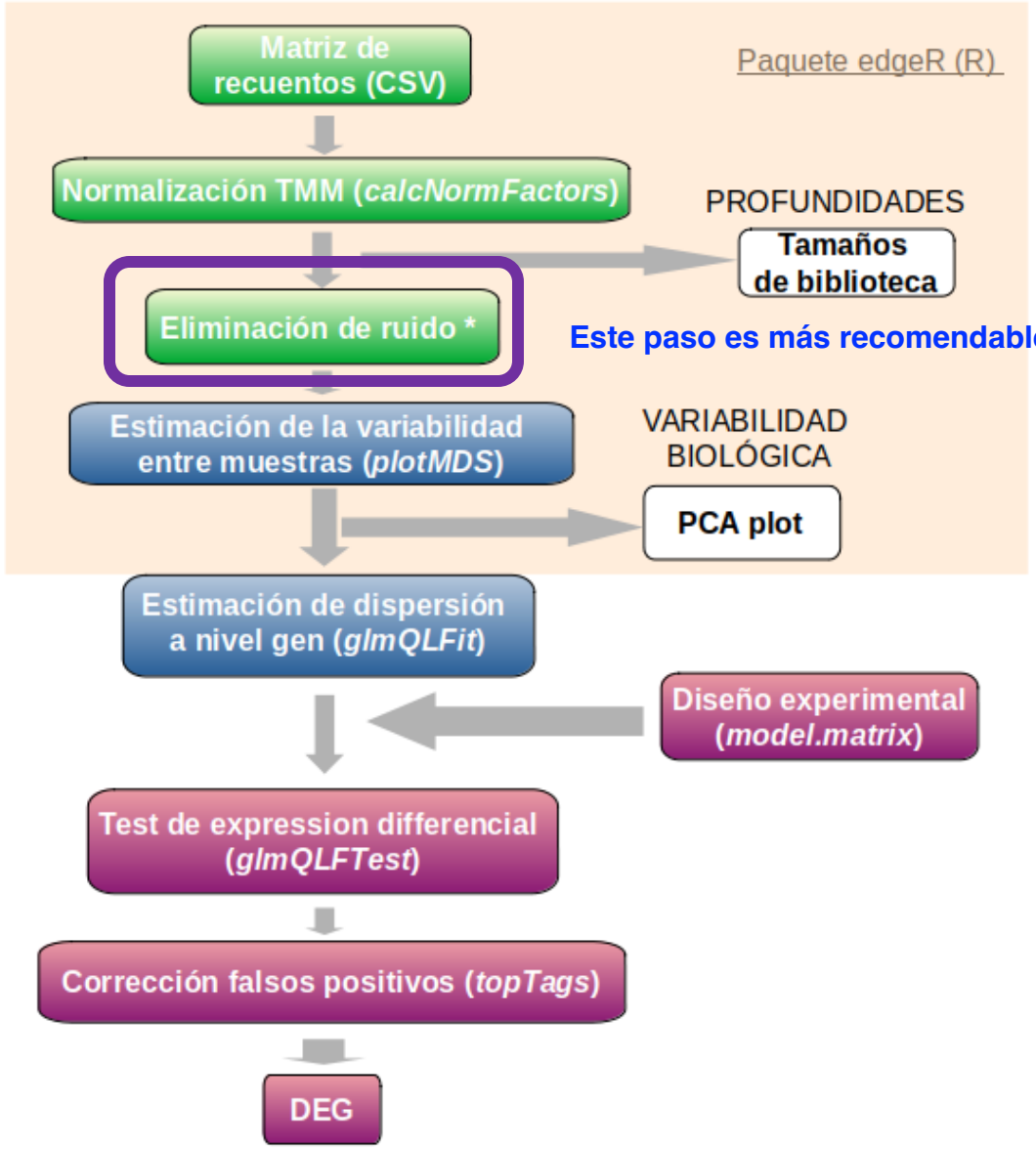
```
install.packages("https://cran.r-project.org/src/contrib/Archive/locfit/locfit_1.5-9.4.tar.gz",

repos=NULL, type="source")
```

Paso 2: Instalación de la librería principal

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

    install.packages("BiocManager")

BiocManager::install("edgeR")



> library(edgeR)
```

```
> packageVersion("edgeR")
[1] ‘3.32.1’
```

Flujo de trabajo del análisis estadístico de la expresión génica



Eliminación de genes de expresión nula y baja

Un gen tiene que expresarse un mínimo antes de poder traducirse

Filtrar aquellos genes con recuentos muy bajos (cercanos a 0) en **todas** las muestras

- **BIOLÓGICO**
- **ESTADÍSTICO**

Se hacen tantos test estadísticos como genes se tengan, cuanto mayor test mayor es la probabilidad de que haya falsos positivos

- Los genes se eliminan si no es posible expresarlos en todas las muestras
- Los usuarios pueden establecer su propia definición de los genes que se expresan.
  - Recuento de 5-15 lecturas en un gen para que se considere expresado en esa biblioteca. Mantener genes con un recuento mínimo de lecturas en todas las muestras.
  - Filtrar con recuento por millón (**CPM**) en lugar de filtrar los contajes directamente.

CMP (Counts per million mapped reads)

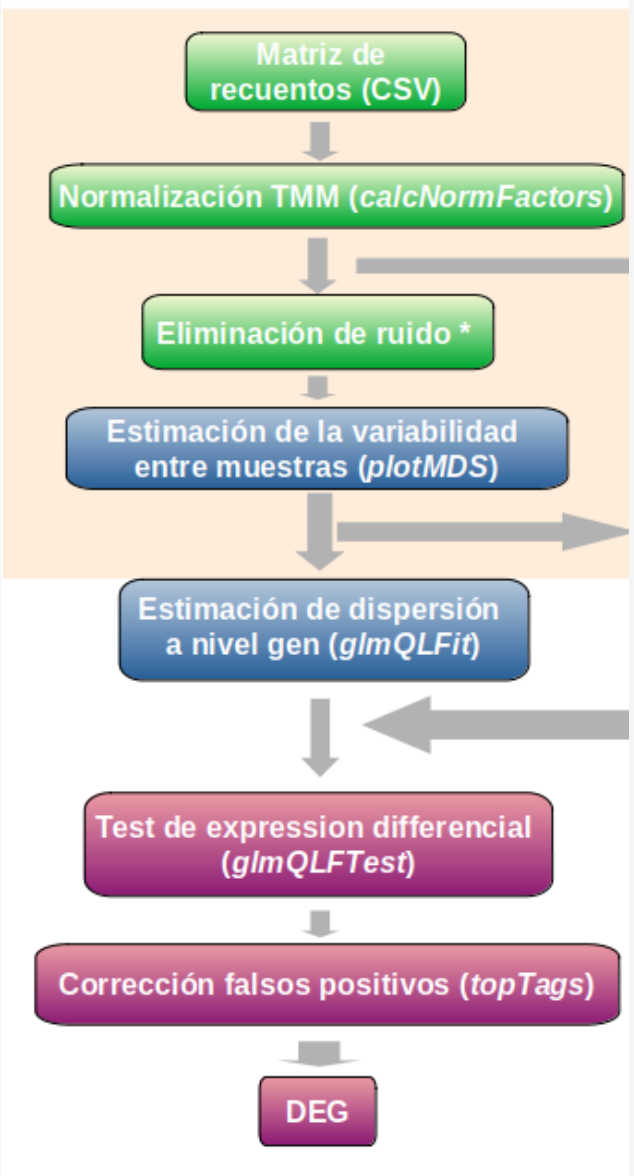
	MCL1.DG	MCL1.DH	MCL1.DI	MCL1.DJ	MCL1.DK	MCL1.DL	MCL1.LA
497097	438	300	65	237	354	287	0
100503874	1	0	1	1	0	4	0
100038431	0	0	0	0	0	0	0
19888	1	1	0	0	0	0	10
20671	106	182	82	105	43	82	16
27395	309	234	337	300	290	270	560

myCPM <- **cpm**(seqdata)

$$\mathrm{CPM} = \frac{\text{Number of reads mapped to gene} \times 10^6}{\text{Total number of mapped reads}}$$

	MCL1.DG	MCL1.DH	MCL1.DI	MCL1.DJ	MCL1.DK	MCL1.DL	MCL1.LA
497097	18.85684388	13.77543859	2.69700983	10.45648006	16.442685	14.3389690	0.0000000
100503874	0.04305215	0.00000000	0.04149246	0.04412017	0.000000	0.1998463	0.0000000
100038431	0.00000000	0.00000000	0.00000000	0.00000000	0.000000	0.0000000	0.0000000
19888	0.04305215	0.04591813	0.00000000	0.00000000	0.000000	0.0000000	0.4903857
20671	4.56352843	8.35709941	3.40238163	4.63261775	1.997275	4.0968483	0.7846171
27395	13.30311589	10.74484210	13.98295863	13.23605071	13.469996	13.4896224	27.4615975

Objetivos



1. Conocer los principales pasos del análisis estadístico de la expresión diferencial con edgeR.

2. Saber identificar la información organizada en los archivos del **diseño experimental** y la matriz de recuentos.
   - Instalación **org.Mn.eg.dg package**

3. Entender y conocer la distribución de los datos de RNA-seq.
   - Instalación **ggplot2 package**

4. Entender las ventajas de la eliminación de genes de baja y nula expresión.
   - Instalación **edgeR package**

DISPONIBILIDAD DE LA ACTIVIDAD 1

From reads to gene counts --> 1234

P1. Responde brevemente a las siguientes preguntas sobre la caracterización inicial del estudio (1 pts)

- ¿Cuántas réplicas biológicas por grupo se utilizaron en este estudio? ¿Es este un número óptimo para experimentos de RNA-seq?. Justifica tu respuesta.
- Describe el protocolo de secuenciación utilizado en este estudio.

P2. Emplea la herramienta FastQC para evaluar la calidad de las lecturas del archivo SRR1552444.fastq.gz. A continuación, utiliza la herramienta TrimGalore para eliminar adaptadores y extremos de baja calidad. Reporta los comandos utilizados y responde a las siguientes preguntas (3 pts).

- ¿Cuántas lecturas iniciales había en el archivo SRR1552444.fastq.gz y cuántas permanecen tras la utilización de TrimGalore?
- Comenta las diferencias que observas en los reportes de FastQC antes y después del filtrado (recuerda reportar los gráficos resultantes), específicamente en:
    - **Per base sequence quality**
    - **Sequence Length Distribution**
    - **Adapter Content.** ¿Qué secuencia se determina como el adaptador principal? ¿En cuántas lecturas ha encontrado dicho adaptador TrimGalore?

P3. Alinea las lecturas depuradas anteriormente sobre el genoma de referencia indexado empleando la herramienta Hisat2 con la opción (-k = 1). Reporta los comandos empleados y contesta a las siguientes preguntas. (1,5 pts)

- Después del alineamiento con Hisat2, ¿cuál fue el número total de lecturas mapeadas y no mapeadas en la muestra?
- Recuerda que uno de los parámetros de Hisat2 es el valor de -k, que indica el número máximo de alineamientos por lectura a generar. Si empleas un valor de **k = 5**, ¿Cómo afecta esto a los valores finales de alineamiento obtenidos y al tiempo de computación?

P4. Utiliza SAMtools para convertir el archivo SAM (generado con el valor de -k igual a 1) a BAM, ordenarlo por coordenadas e indexarlo. Reporta los comandos utilizados y contesta a las siguientes preguntas. (1,5 pts)

- ¿Cuántas FLAGs distintas se encuentran en el archivo **SRR1552444_hisat2.sam**? Indica cuáles son y sus cantidades.
- ¿Cuántos valores distintos de MAPQ hay en el archivo **SRR1552444_hisat2.sam**? Indícalos y cuenta cuántos son.
- ¿Cuántas letras distintas del alfabeto encontramos en la columna CIGAR del archivo **SRR1552444_hisat2.sam**? Indica cuáles son, sus cantidades y qué información proporcionan.

P5. Como hemos visto en el archivo BAM, hay una ubicación cromosómica para cada lectura asignada. Ahora que ya hemos descubierto de dónde proviene cada lectura en el genoma, necesitamos anotar dicha información. Para ello obtén el archivo de anotaciones en formato GTF y responde a las siguientes cuestiones. (1,5 pts)

- ¿Cuántos y qué programas y bases de datos se emplearon para anotar este archivo? Para responder a esta pregunta, interroga la columna **SOURCE**.
- ¿Cuáles y cuántas **FEATURES** podemos encontrar en el archivo de anotaciones?
- ¿Qué porcentaje de **GENES** en el archivo GTF están ubicados en el cromosoma X?

P6. Finalmente, ejecuta el recuento de los alineamientos sobre el archivo BAM con htseq-count. Reporta los comandos utilizados y computa sobre el archivo tsv final, el porcentaje total de lecturas asignadas, no_feature, ambiguous, too_low_aQual y not_aligned. Una vez computado estos porcentajes, muéstralos con un gráfico (usando cualquier lenguaje de programación) y comenta brevemente su resultado. (1,5 pts)

P7 OPCIONAL. A lo largo del flujo de trabajo, existen etapas que no se han abordado en clase, como la identificación y marcaje de duplicados con MarkDuplicates(Picard), la evaluación de calidad del alineamiento con RseQC, o el uso de herramientas alternativas para el recorte, filtrado, mapeado y recuento de lecturas. En esta parte opcional de la actividad (que puede aportar hasta 0,5 puntos adicionales), se brinda al estudiante la libertad de seleccionar y aplicar alguna de estas etapas adicionales. Se solicita que, además de implementar la etapa seleccionada, se comparta tanto el código desarrollado como las observaciones principales derivadas de su aplicación.

RECOMENDACIONES AGOSTO

1. **Disfruten**: Aprovechen el verano para relajarse, disfrutar del buen clima y compartir momentos especiales con sus seres queridos.
2. **Descansen**: El verano es una excelente oportunidad para descansar y recargar pilas.
3. **Practiquen un poquito de R**: Intenten revisar el manual de EdgeR, ya que será la herramienta principal que emplearemos.
4. **Realicen la actividad nº1: From reads to gene counts.** Con ello afianzarán los conceptos dados hasta el momento y se quitarán cierto peso para el final de la asignatura.

Universidad
Internacional
de Valencia

universidadviu.com

De:
Planeta Formación y Universidades