Análisis transcriptómicos de la expresión génica

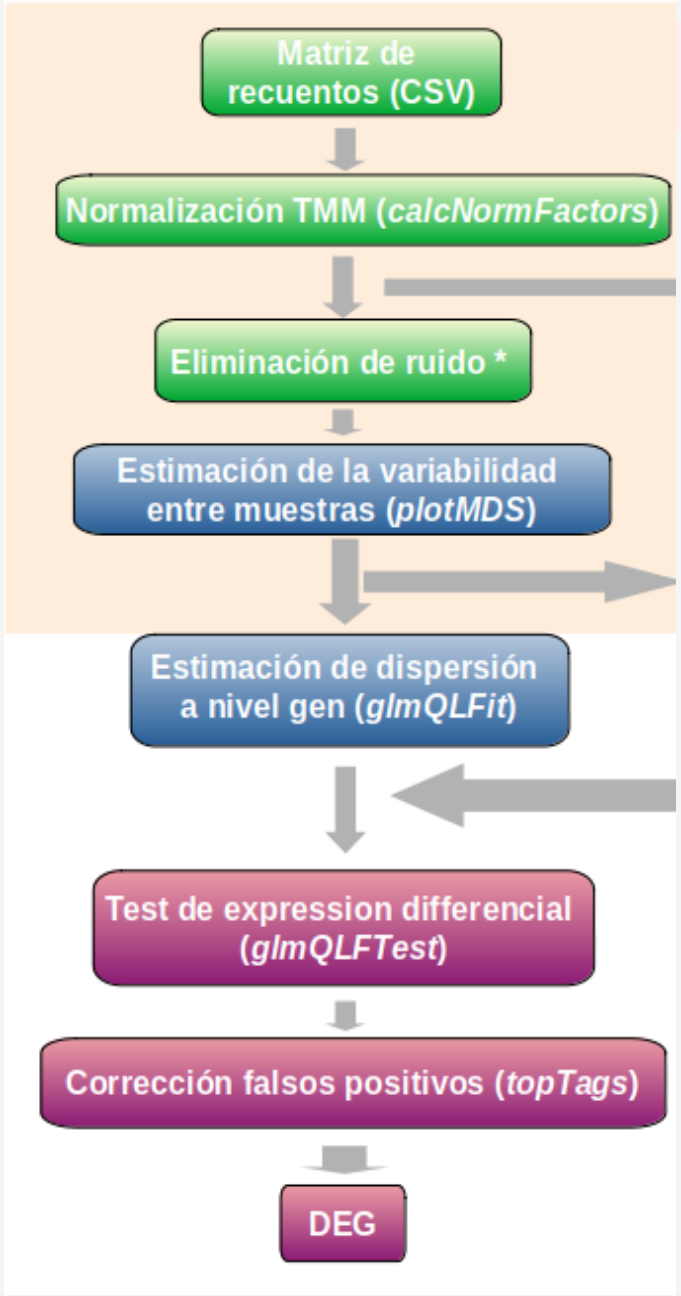
Máster Universitario en Bioinformática

Sesión 9

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

De:
Planeta Formación y Universidades

Bloque IV: Análisis estadístico de la diferencia de expresión



Sesión 8

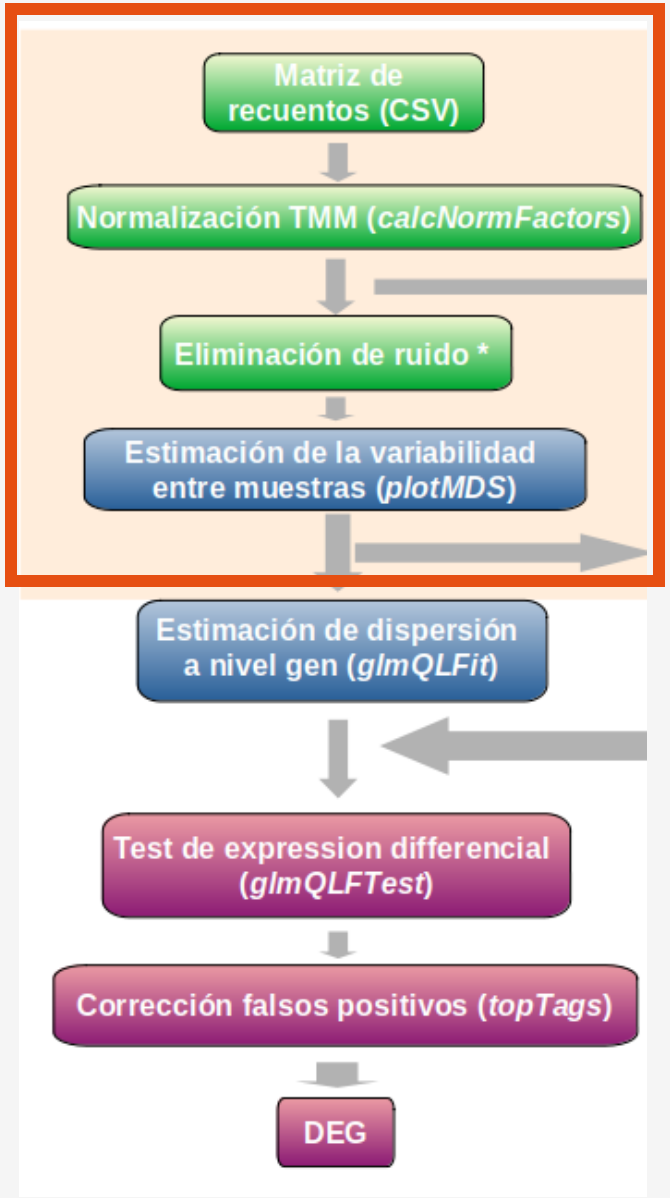
1. Conocer los principales pasos del análisis estadístico de la expresión diferencial con edgeR.

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**

Objetivos



1. Identificar y transformar la información organizada en la **matriz de recuentos.**

Librería -> Org.Mn.eg.dg

- *Transformar los identificadores ENTREZ de los genes*

Librería -> edgeR

- *Transformar los datos de conteo en un objeto DGEList*

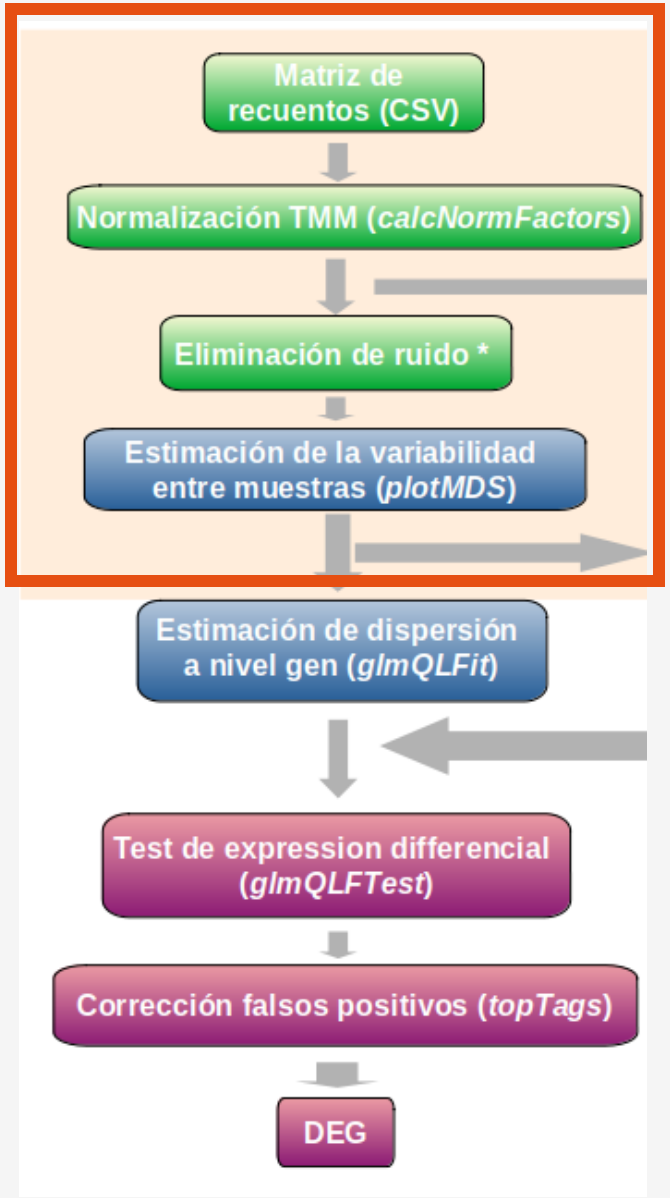
2. Preprocesamiento de la matriz de conteo

- Identificar y eliminar genes de expresión nula.
- Normalización de los datos (TMM).

3. Estudio de la variabilidad muestral (plotMDS).

4. Preguntas y Respuestas

Objetivos



1 Identificar y transformar la información organizada en la **matriz de recuentos.**

Librería -> Org.Mn.eg.dg

- *Transformar los identificadores ENTREZ de los genes*

Librería -> edgeR

- *Transformar los datos de conteo en un objeto DGEList*

2 Preprocesamiento de la matriz de conteo

- Identificar y eliminar genes de expresión nula.
- Normalización de los datos (TMM).

3 Estudio de la variabilidad muestral (plotMDS).

4 Preguntas y Respuestas

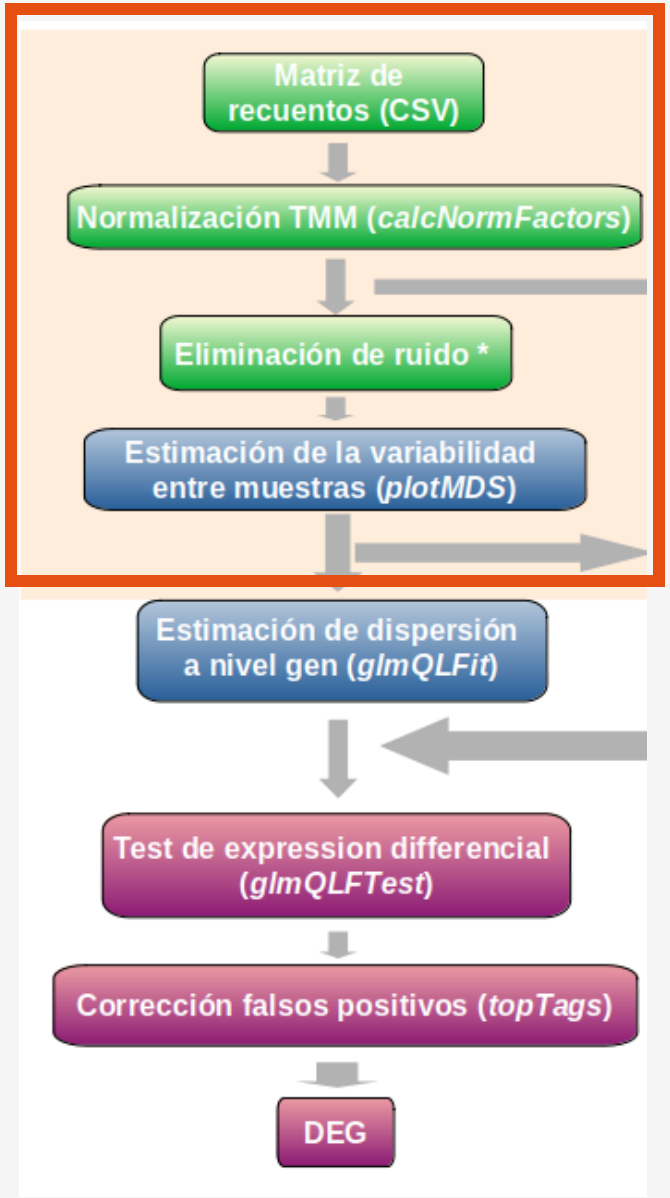
PRACTIQUEMOS

amazon WorkSpaces

Transformación de los identificadores de ENTREZ de genes

```
> library(org.Mm.eg.db)
> columns(org.Mm.eg.db)
 [1] "ACCNUM"      "ALIAS"       "ENSEMBL"     "ENSEMBLPROT"  "ENSEMBLTRANS" "ENTREZID"    "ENZYME"
 [8] "EVIDENCE"    "EVIDENCEALL" "GENENAME"    "GO"           "GOALL"        "IPI"         "MGI"
[15] "ONTOLOGY"    "ONTOLOGYALL" "PATH"        "PFAM"         "PMID"         "PROSITE"     "REFSEQ"
[22] "SYMBOL"      "UNIGENE"     "UNIPROT"

> ann <- select(org.Mm.eg.db,keys=rownames(seqdata),columns=c("ENTREZID","SYMBOL","GENENAME"))
'select()' returned 1:1 mapping between keys and columns

> head(ann)
  ENTREZID  SYMBOL                       GENENAME
1   497097    Xkr4     X-linked Kx blood group related 4
2 100503874 Gm19938              predicted gene, 19938
3 100038431 Gm10568               predicted gene 10568
4    19888     Rp1       retinitis pigmentosa 1 (human)
5    20671   Sox17 SRY (sex determining region Y)-box 17
6    27395  Mrpl15   mitochondrial ribosomal protein L15

> table(ann$ENTREZID==rownames(seqdata))
 TRUE
27179
```

Objetivos



1. Identificar y transformar la información organizada en la **matriz de recuentos.**

   Librería -> Org.Mn.eg.dg

   - *Transformar los identificadores ENTREZ de los genes*

   Librería -> edgeR

   - *Transformar los datos de conteo en un objeto DGEList*

2. Preprocesamiento de la matriz de conteo

   - Identificar y eliminar genes de expresión nula (cpm)
   - Normalización de los datos (TMM).

3. Estudio de la variabilidad muestral (plotMDS).

4. Preguntas y Respuestas

Creación de la clase de datos DGEList

EdgeR almacena los datos en un objeto simple basado en listas llamado **DGEList**

y <- **DGEList**(seqdata)

```
> y
An object of class "DGEList"
$counts
          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
27174 more rows ...

$samples
                group lib.size norm.factors
MCL1.DG luminal.virgin 23227641            1
MCL1.DH   basal.virgin 21777891            1
MCL1.DI basal.pregnant 24100765            1
MCL1.DJ basal.pregnant 22665371            1
MCL1.DK  basal.lactate 21529331            1
7 more rows ...

$genes
   ENTREZID  SYMBOL                                GENENAME
1    497097    Xkr4     X-linked Kx blood group related 4
2 100503874 Gm19938                  predicted gene, 19938
3 100038431 Gm10568                   predicted gene 10568
4     19888     Rp1        retinitis pigmentosa 1 (human)
5     20671   Sox17 SRY (sex determining region Y)-box 17
27174 more rows ...
```

El lib.size es el sumatorio de todas las filas de cada columna de muestras

- **Datos de Conteo**
- ***Información sobre las muestras***
- **Anotación genómica**

PRACTIQUEMOS

Creación de la clase de datos DGEList

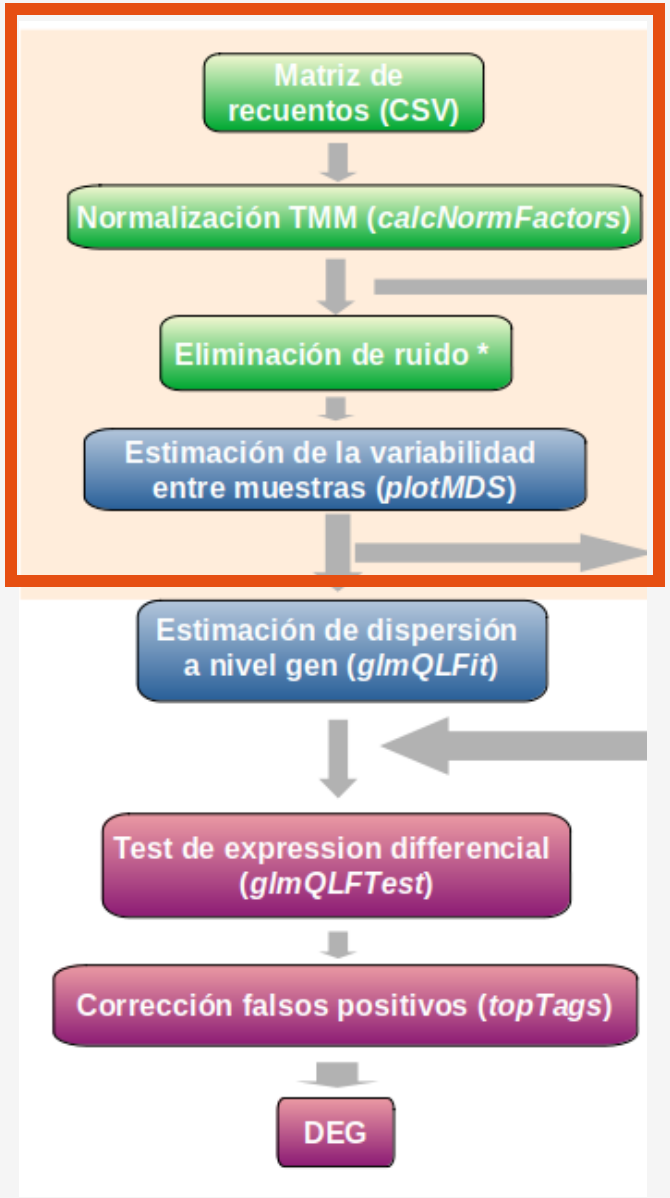
```
> # Convert counts to DGEList object
> y <- DGEList(seqdata)
> head(y)

An object of class "DGEList"
$counts
          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

$samples
        group lib.size norm.factors
MCL1.DG     1 23227641            1
MCL1.DH     1 21777891            1
MCL1.DI     1 24100765            1
MCL1.DJ     1 22665371            1
MCL1.DK     1 21529331            1
7 more rows ...

> names(y)

[1] "counts"  "samples"
```

Creación de la clase de datos DGEList

```
> # Add the group information into the DGEList
> y$samples$group <- group

> # Add gene anotation
> y$genes <- ann
```

```
> head(y,3)
An object of class "DGEList"
$counts
          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

$samples
                group lib.size norm.factors
MCL1.DG   basal.virgin 23227641            1
MCL1.DH   basal.virgin 21777891            1
MCL1.DI basal.pregnant 24100765            1
MCL1.DJ basal.pregnant 22665371            1
MCL1.DK  basal.lactate 21529331            1
7 more rows ...

$genes
   ENTREZID  SYMBOL                          GENENAME
1    497097    Xkr4 X-linked Kx blood group related 4
2 100503874 Gm19938             predicted gene, 19938
3 100038431 Gm10568              predicted gene 10568
```

Objetivos



1 Identificar y transformar la información organizada en la **matriz de recuentos.**

Librería -> Org.Mn.eg.dg

- *Transformar los identificadores ENTREZ de los genes*

Librería -> edgeR

- *Transformar los datos de conteo en un objeto DGEList*

2 Preprocesamiento de la matriz de conteo

- Identificar y eliminar genes de expresión nula
- Normalización de los datos (TMM).

3 Estudio de la variabilidad muestral (plotMDS).

4 Preguntas y Respuestas

PRACTIQUEMOS

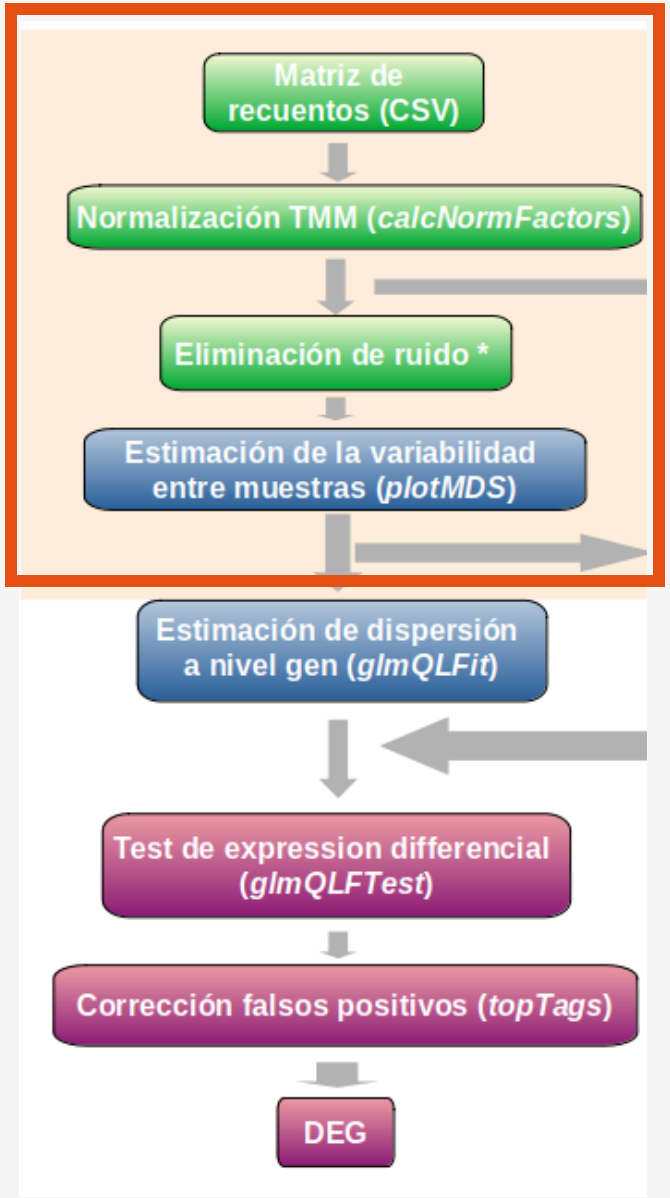
amazon WorkSpaces

Filtrado de genes

```
> keep <- filterByExpr(y)
> head(keep)
    497097 100503874 100038431    19888    20671    27395
     TRUE    FALSE    FALSE    FALSE     TRUE     TRUE
> summary(keep)
    Mode   FALSE    TRUE
     logical   11210   15969
> y <- y[keep, keep.lib.sizes=FALSE]
```

```
> head(y,3)
An object of class "DGEList"
$counts
       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
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

$samples
                group lib.size norm.factors
MCL1.DG   basal.virgin 23219195            1
MCL1.DH   basal.virgin 21769326            1
MCL1.DI basal.pregnant 24092719            1
MCL1.DJ basal.pregnant 22657703            1
MCL1.DK  basal.lactate 21522881            1
7 more rows ...

$genes
  ENTREZID SYMBOL                                GENENAME
1   497097   Xkr4     X-linked Kx blood group related 4
5    20671  Sox17 SRY (sex determining region Y)-box 17
6    27395 Mrpl15   mitochondrial ribosomal protein L15
```

Objetivos



1. Identificar y transformar la información organizada en la **matriz de recuentos.**

   Librería -> Org.Mn.eg.dg
   - *Transformar los identificadores ENTREZ de los genes*

   Librería -> edgeR
   - *Transformar los datos de conteo en un objeto DGEList*

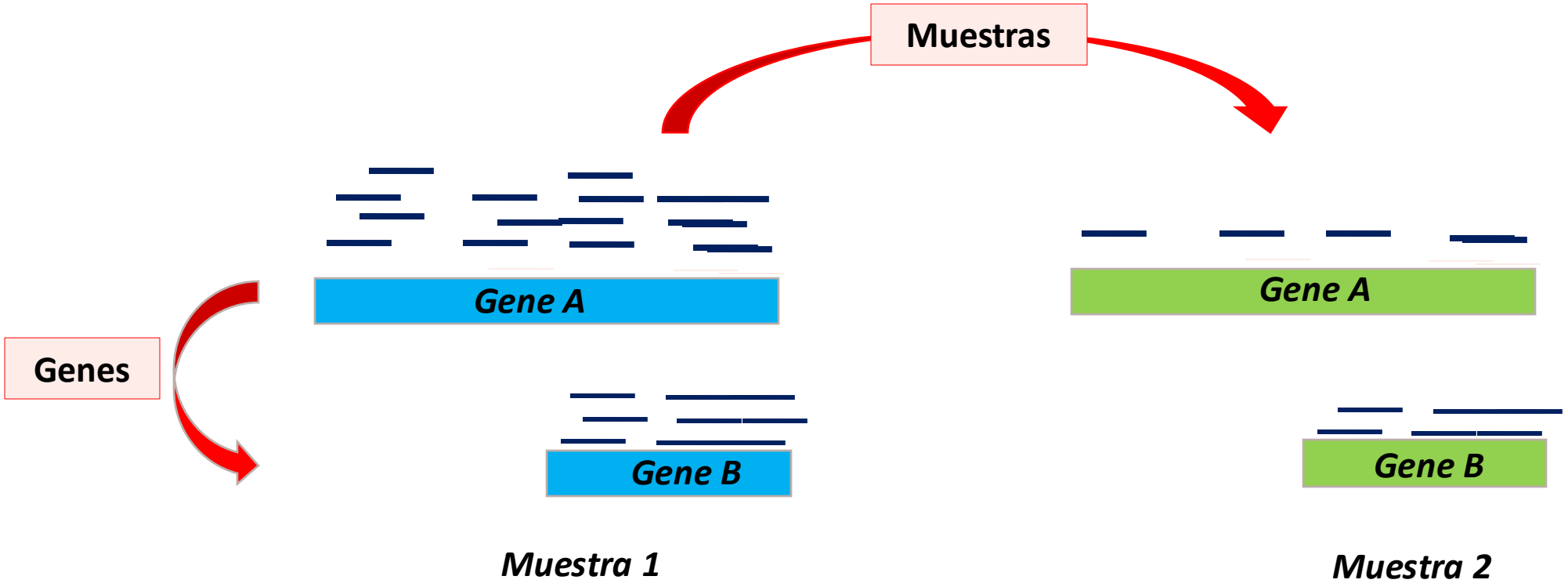
2. Preprocesamiento de la matriz de conteo
   - Identificar y eliminar genes de expresión nula
   - Normalización de los datos (TMM).

3. Estudio de la variabilidad muestral (plotMDS).

Necesidad de normalización de las bibliotecas de RNA-seq

La normalización de las bibliotecas de mapeos es un paso obligatorio en el DGE. Este paso es esencial para comparar la expresión de los **genes entre muestras**.

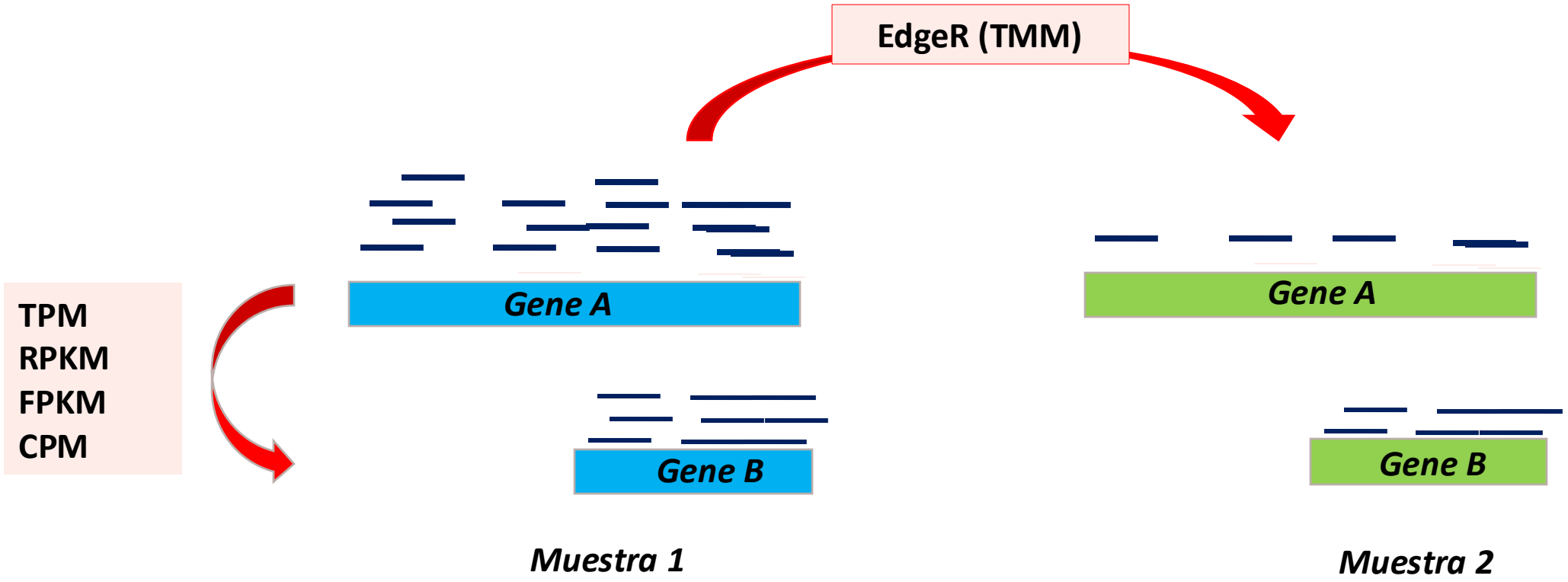
Para hacer comparaciones no sesgadas



Se va a comparar un mismo gen en diferentes muestras.

Necesidad de normalización por MUESTRA de las bibliotecas de RNA-seq

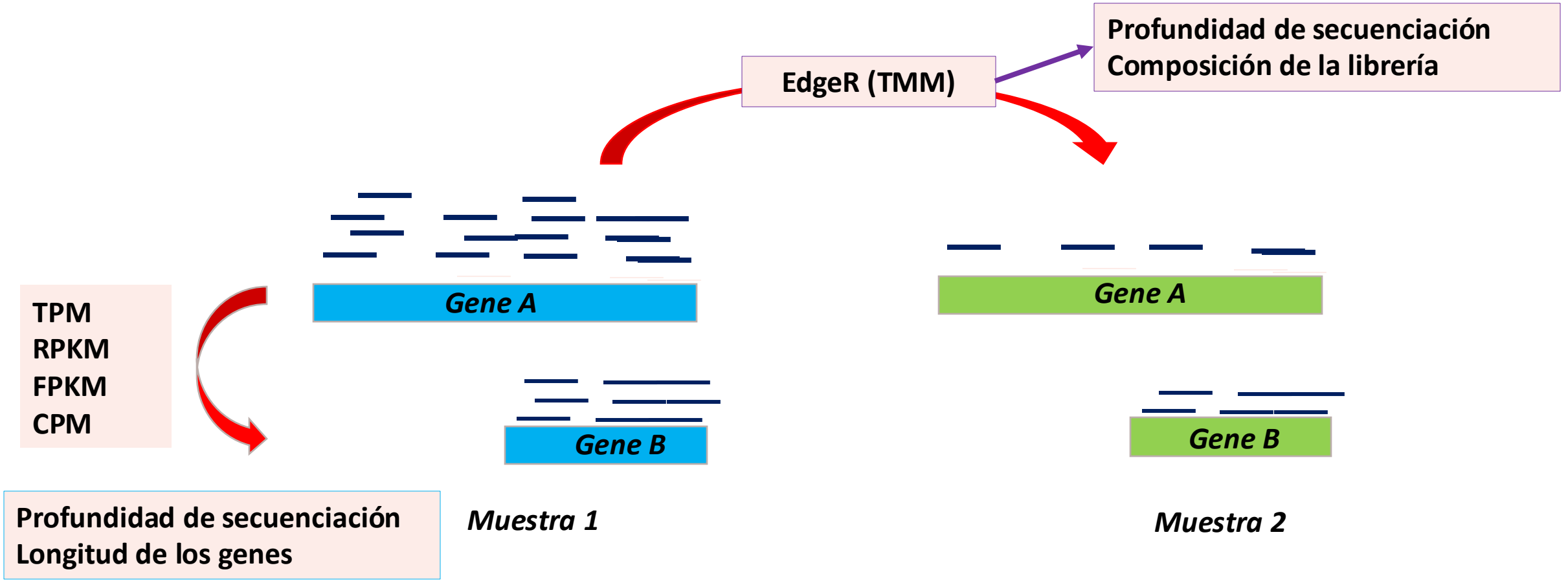
La normalización de las bibliotecas de mapeos es un paso obligatorio en el DGE. Este paso es esencial para comparar la expresión de los **genes entre muestras**.



https://www.reneshbedre.com/blog/expression_units.html

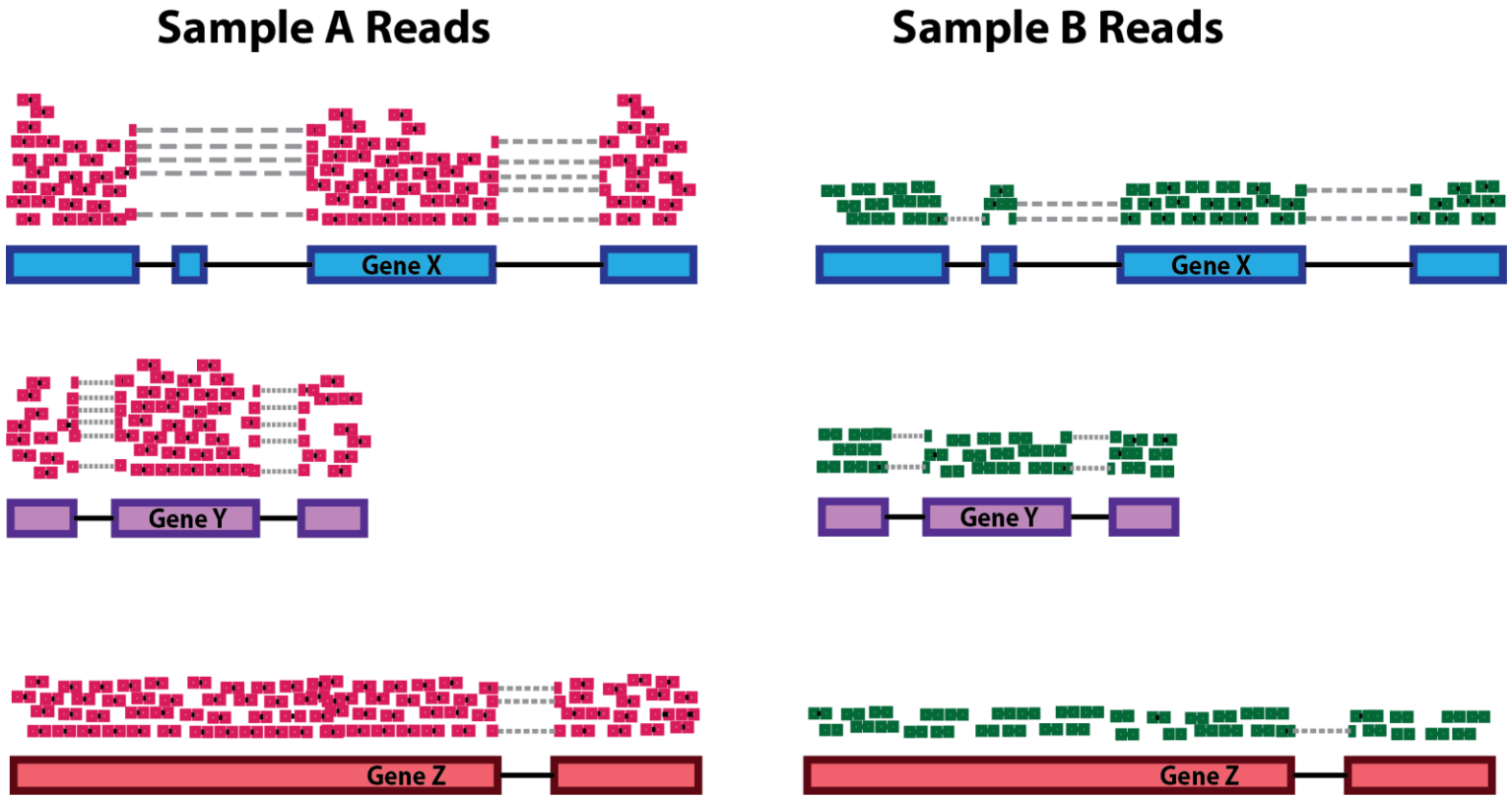
Necesidad de normalización por MUESTRA de las bibliotecas de RNA-seq

La normalización de las bibliotecas de mapeos es un paso obligatorio en el DGE. Este paso es esencial para comparar la expresión de los **genes entre muestras**.



Profundidad de secuenciación (*library size*)

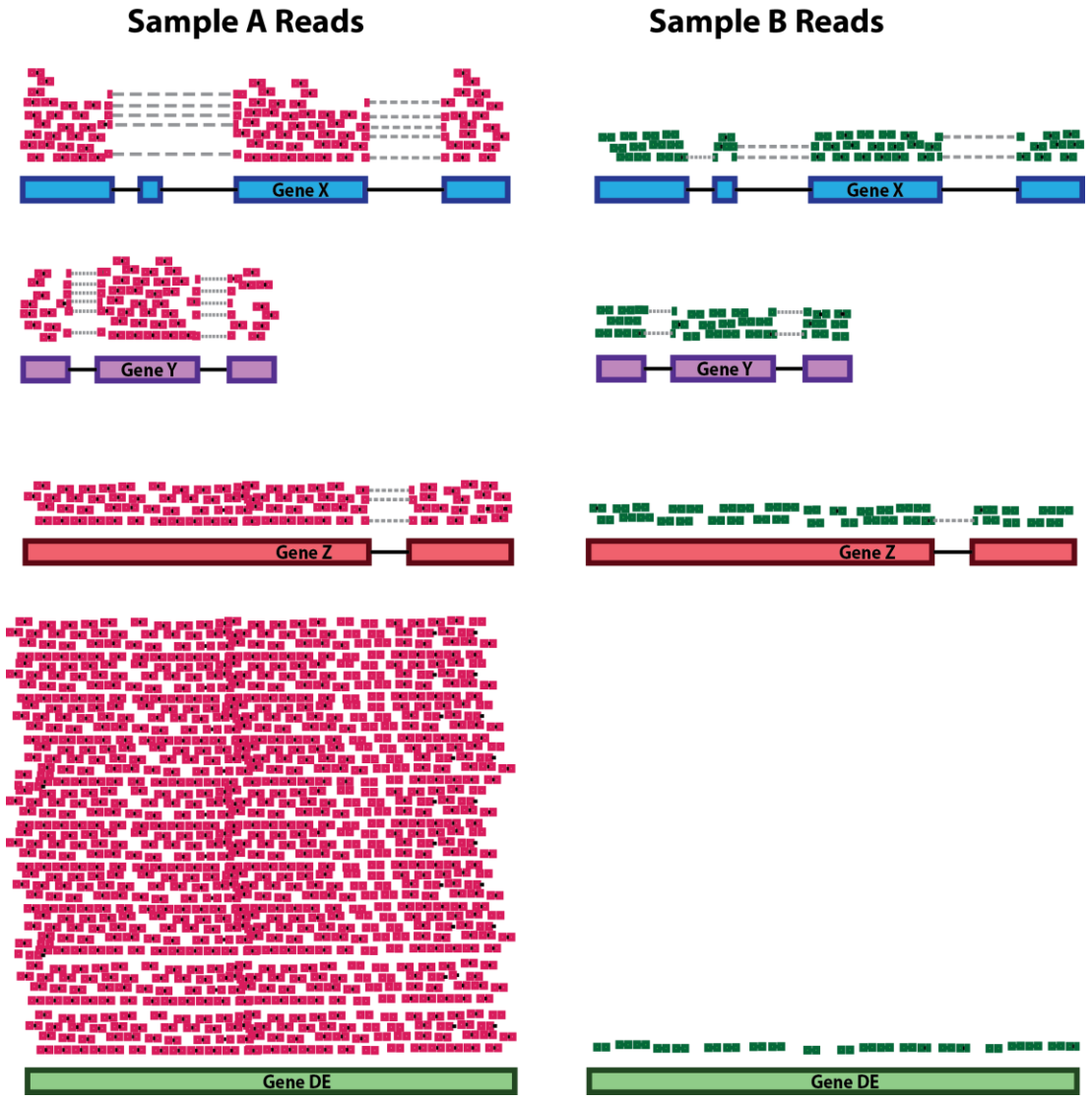
Número total de lecturas recogidas de cada transcriptoma y está definida por la muestra inicial de RNA



Composición de la librería (*library composition*)

- El RNA-seq mide la expresión relativa de cada gen en una muestra dada.

- Un pequeño número de genes se expresa de manera muy alta en algunas muestras y monopoliza la secuenciación dejando a los genes restantes parecer falsamente infra-regulados.
    - Aumentando el número de **falsos positivos**



Normalización TMM (*Trimmed mean of M values*)

- Función **calcNormFactors(y)**

- Calcula un conjunto de factores de normalización, uno para cada muestra, para eliminar los sesgos de composición entre bibliotecas.

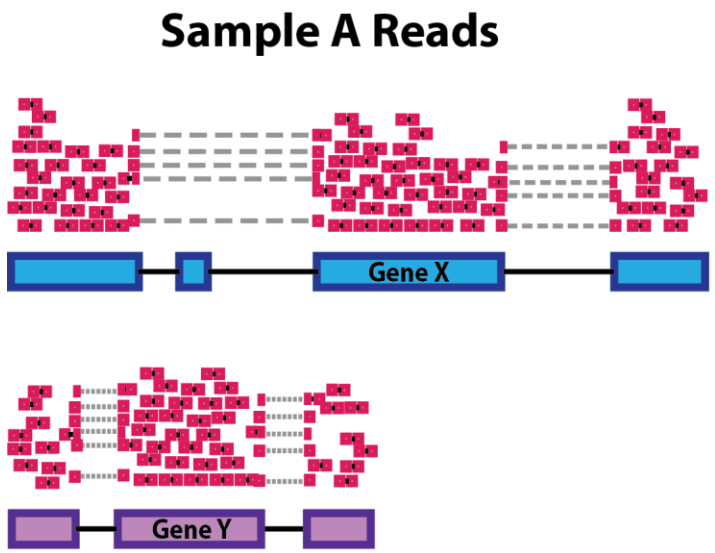
- El producto de estos factores y los tamaños de biblioteca define el **tamaño de biblioteca efectivo**, que reemplaza el tamaño de biblioteca original en todos los análisis posteriores.

Valores de norm.factors menores a 1 , es porque el tamaño de esa biblioteca es mucho más grande

```
y <- calcNormFactors(y)
y$samples
```

	group	lib.size	norm.factors
MCL1.DG	basal.virgin	23219195	1.2384285
MCL1.DH	basal.virgin	21769326	1.2137919
MCL1.DI	basal.pregnant	24092719	1.1248001
MCL1.DJ	basal.pregnant	22657703	1.0711828
MCL1.DK	basal.lactate	21522881	1.0362789
MCL1.DL	basal.lactate	20009184	1.0869418
MCL1.LA	luminal.virgin	20385437	1.3681957
MCL1.LB	luminal.virgin	21699830	1.3649435
MCL1.LC	luminal.pregnant	22236469	1.0039770
MCL1.LD	luminal.pregnant	21983364	0.9226086
MCL1.LE	luminal.lactate	24720123	0.5292854
MCL1.LF	luminal.lactate	24653390	0.5353884

Otros factores de Influencia



- **Contenido de GC (Contenido de Guanina y Citosina)**
  - Efecto limitado en los análisis de expresión diferencial.
  - Paquetes como *EDASeq* y *cqn* que estiman factores de corrección para ajustar los efectos del contenido de GC en cada muestra.

- **Longitud del gen**
  - Efecto limitado en los análisis de expresión diferencial.
  - Sin embargo, todavía pueden detectarse efectos específicos de la muestra en la longitud del gen.

Normalización en edgeR

Model-based normalization, no transformation

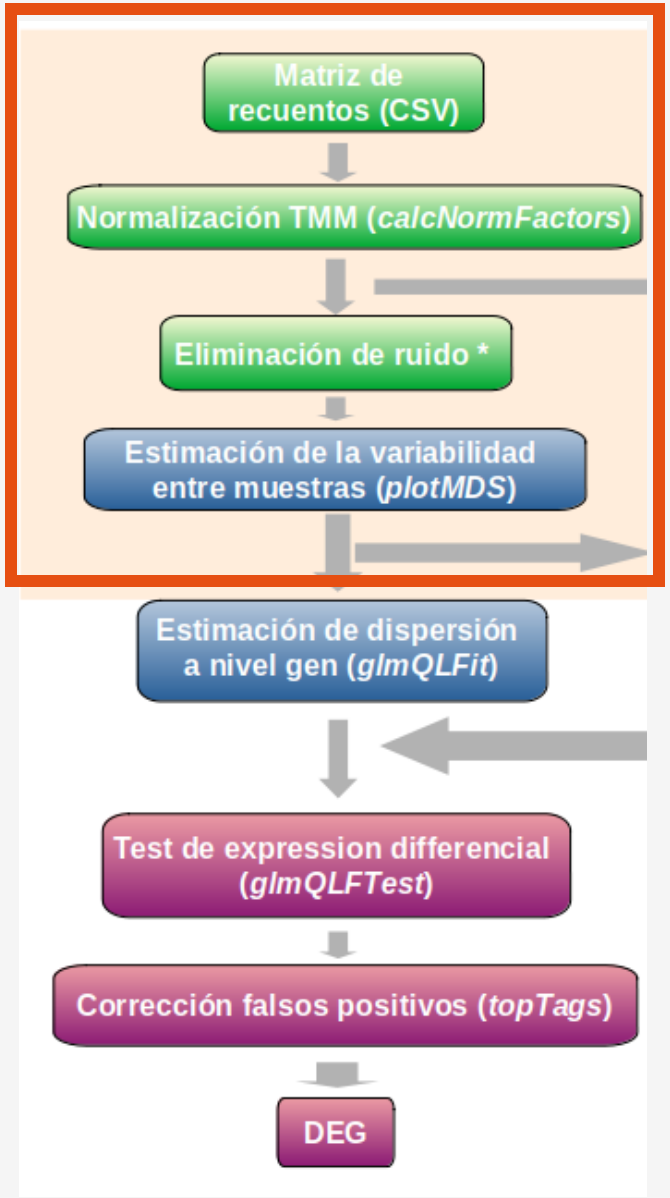
- En edgeR, la normalización se realiza mediante factores de corrección que se incorporan al modelo estadístico.
- Estos factores de corrección suelen calcularse internamente mediante funciones de edgeR, pero también es posible que el usuario los proporcione.

IMP! : La normalización se basa en factores de corrección, es una normalización basada en modelos, no se transforman los recuentos originales

Importante a Tener en Cuenta

- La normalización en edgeR se basa en modelos y no transforma los recuentos originales de lecturas.

Objetivos



1. Identificar y transformar la información organizada en la **matriz de recuentos.**

   Librería -> Org.Mn.eg.dg
   - *Transformar los identificadores ENTREZ de los genes*

   Librería -> edgeR
   - *Transformar los datos de conteo en un objeto DGEList*

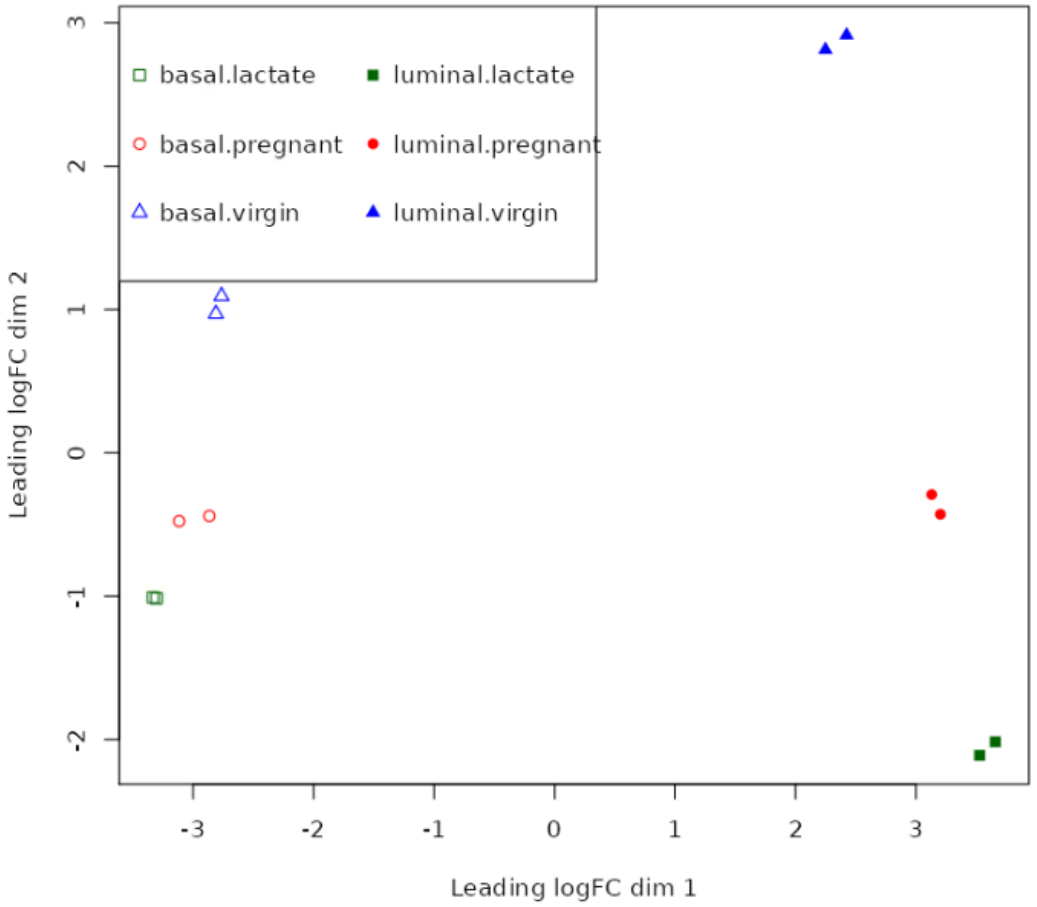
2. Preprocesamiento de la matriz de conteo
   - Identificar y eliminar genes de expresión nula (cpm)
   - Normalización de los datos (TMM).

3. Estudio de la variabilidad muestral (plotMDS).

4. Preguntas y Respuestas

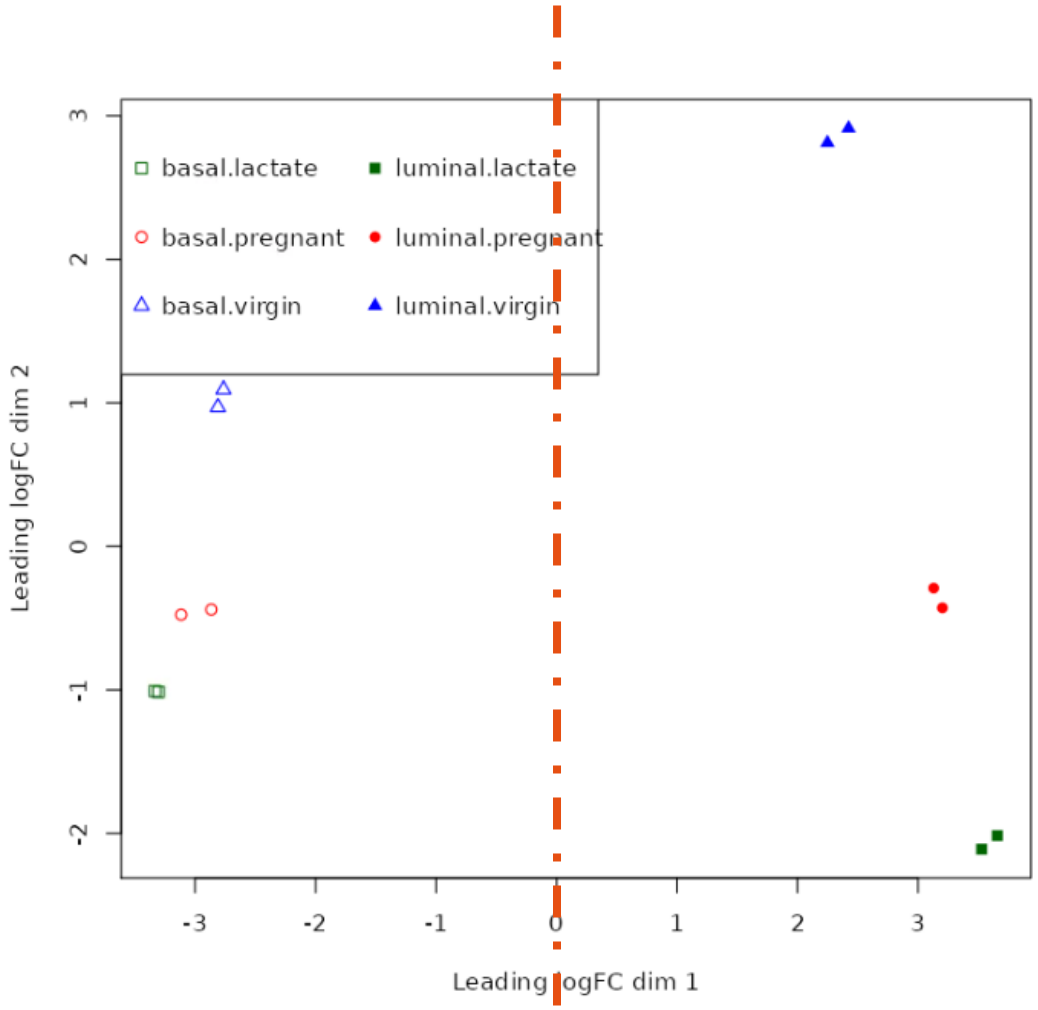
Estimación de la variabilidad biológica

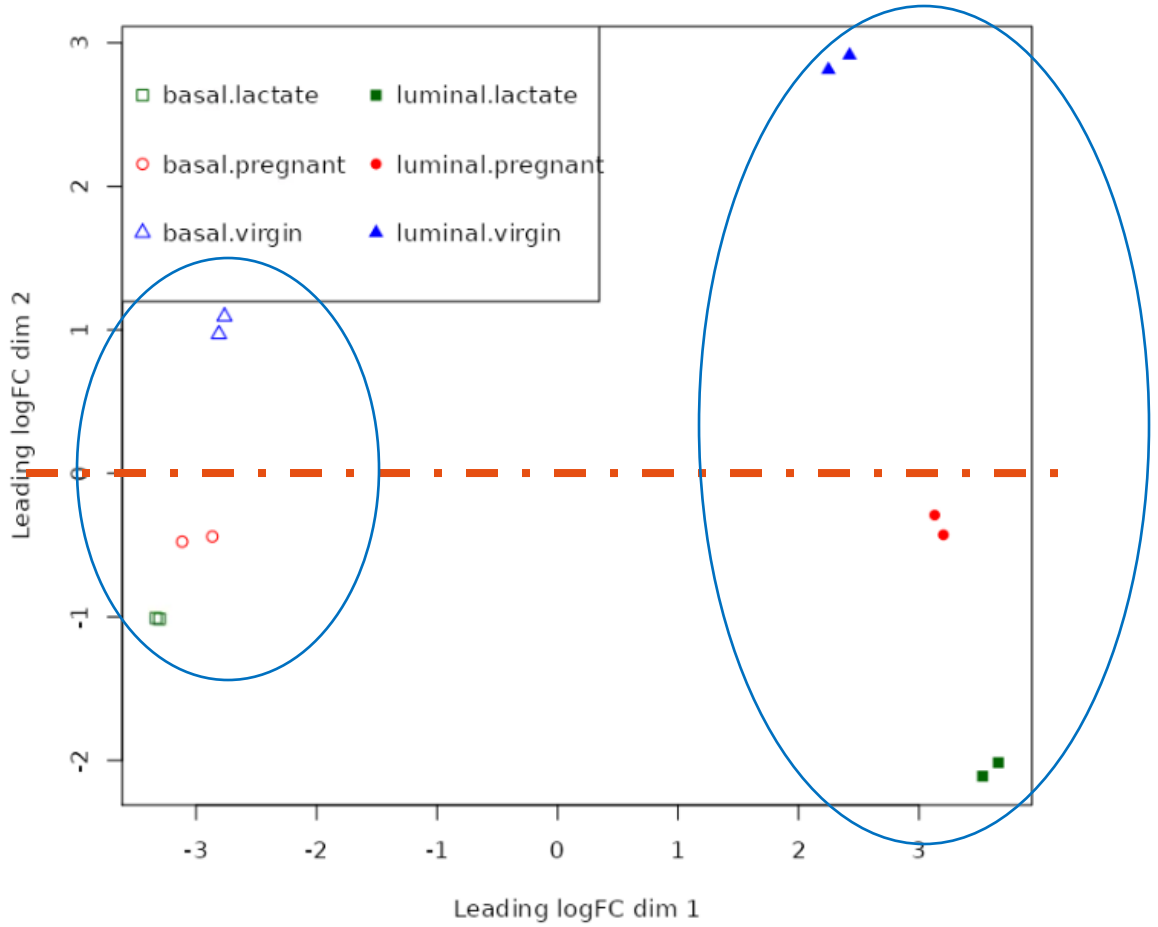
MDS -> Multi-dimensional scaling plot



- plotMDS calcula las distancias entre las muestras como las diferencias de nivel de expresión dentro de los 500 genes más dinámicos.

Estudio de la variabilidad muestral



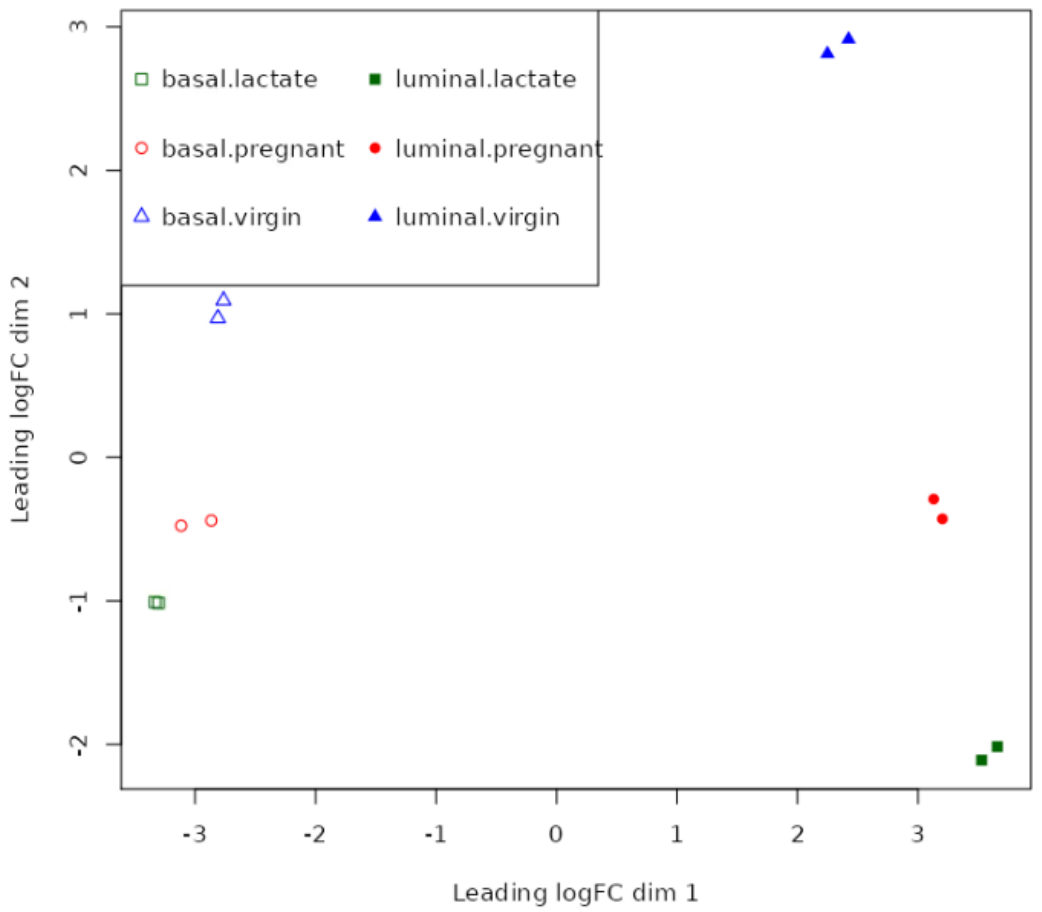


PRACTIQUEMOS

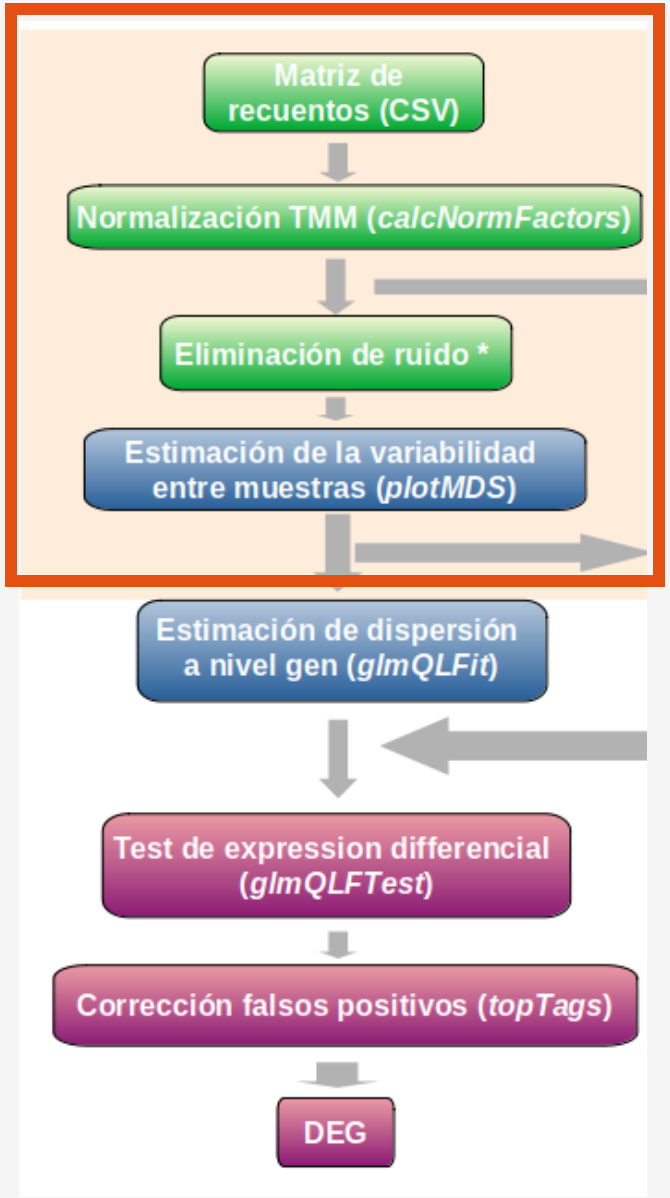
amazon WorkSpaces

Normalización + Estudio de la variabilidad muestral

```
# 4. Normalization

y <- calcNormFactors(y)



# 5. Multidimensional scaling plots

pch <- c(0,1,2,15,16,17)

colors <- rep(c("darkgreen", "red", "blue"), 2)

plotMDS(y, col=colors[group], pch=pch[group])

legend("topleft", legend=levels(group), pch=pch,

col=colors, ncol=2, cex = 0.8)
```



Objetivos



1. Identificar y transformar la información organizada en la **matriz de recuentos.**

   Librería -> Org.Mn.eg.dg
   - *Transformar los identificadores ENTREZ de los genes*

   Librería -> edgeR
   - *Transformar los datos de conteo en un objeto DGEList*

2. Preprocesamiento de la matriz de conteo
   - Identificar y eliminar genes de expresión nula (cpm)
   - Normalización de los datos (TMM).

3. Estudio de la variabilidad muestral (plotMDS).

4. Preguntas y Respuestas

PRIMER CONTACTO

BUZZ!

CONCURSO UNIVERSAL

PlayStation Portable

Preguntas y respuestas

P1. En el flujo de trabajo de datos de RNA-seq, ¿cuál de las siguientes afirmaciones es VERDADERA?:

1. Markduplicates (GATK) sirve para eliminar los adaptadores de las lecturas.
2. SAMTools elimina los adaptadores de las lecturas.
3. Hisat2 es un mapeador que puede funcionar sin genoma de referencia.
4. FastQC es una herramienta que evalúa la calidad de los datos de secuencia.

P2. Los archivos FASTQ recogen la siguiente información:

1. La secuencia FASTA de las lecturas.
2. Los alineamientos de las lecturas.
3. La calidad de base de las lecturas.
4. a y c.

P3. Acerca del flujo de trabajo con muestras de lecturas de RNA-seq:

1. Una base (A,C,G,T) de lectura con una calidad de 20 Phred tiene una probabilidad de error de 1 entre 100 designaciones de base
2. El mapeo con Hisat2 permite alinear una lectura en dos regiones exónicas separadas en el genoma.
3. El mapeo con Hisat2 necesita la indexación previa de la secuencia de referencia.
4. Todas son ciertas.

Preguntas y respuestas

P4. Acerca del contenido medio en GC de las lecturas, ¿qué es verdadero?:

1. Aumenta a medida que crece el tamaño de las lecturas en el secuenciador.
2. Es un indicador con el que identificar si hubo contaminación con RNA de otro organismo.
3. Aumenta si el contenido en adaptadores es alto.
4. Ninguna de las anteriores es cierta.

El contenido medio de GC es independiente del tamaño de las lecturas y de los adaptadores.

P5. Acerca de la calidad de las lecturas en el RNA-seq, son normales:

1. Una disminución de la calidad de la base hacia el final de la lectura.
2. Que las lecturas de secuencia única representen entre el 60 y el 80 % de las lecturas.
3. Que las primeras 12 bases de la lectura no representen un porcentaje equitativo de A,T,G,C.
4. Todas las anteriores.

Primero se eliminan los adaptadores, el orden es al revés. Es tanto para lecturas single-end como paird-end. y lo más recomentable es empezar con un trimming/filtering poco restrictivo, ser lo más laxos posibles y evaluar su impacto, porque si es muy restrictivo se están perdiendo datos

P6. Selecciona la afirmación CORRECTA sobre TrimGalore!:

1. Cuando se eliminan posiciones de baja calidad también se van a estar eliminado bases de buena calidad que pueden estar incrustadas en ella.
2. Primero se filtran las lecturas en función de su longitud y después se eliminan los adaptadores que en ellas se puedan encontrar desde el extremo 3'.
3. TrimGalore solo es aplicable para trabajar con lectura single-end o de único extremo.
4. Inicialmente siempre se recomienda realizar un trimming/filtering muy restrictivo y evaluar su impacto.

Universidad
Internacional
de Valencia

universidadviu.com

De:
Planeta Formación y Universidades