Análisis transcriptómicos de la expresión génica

Máster Universitario en Bioinformática

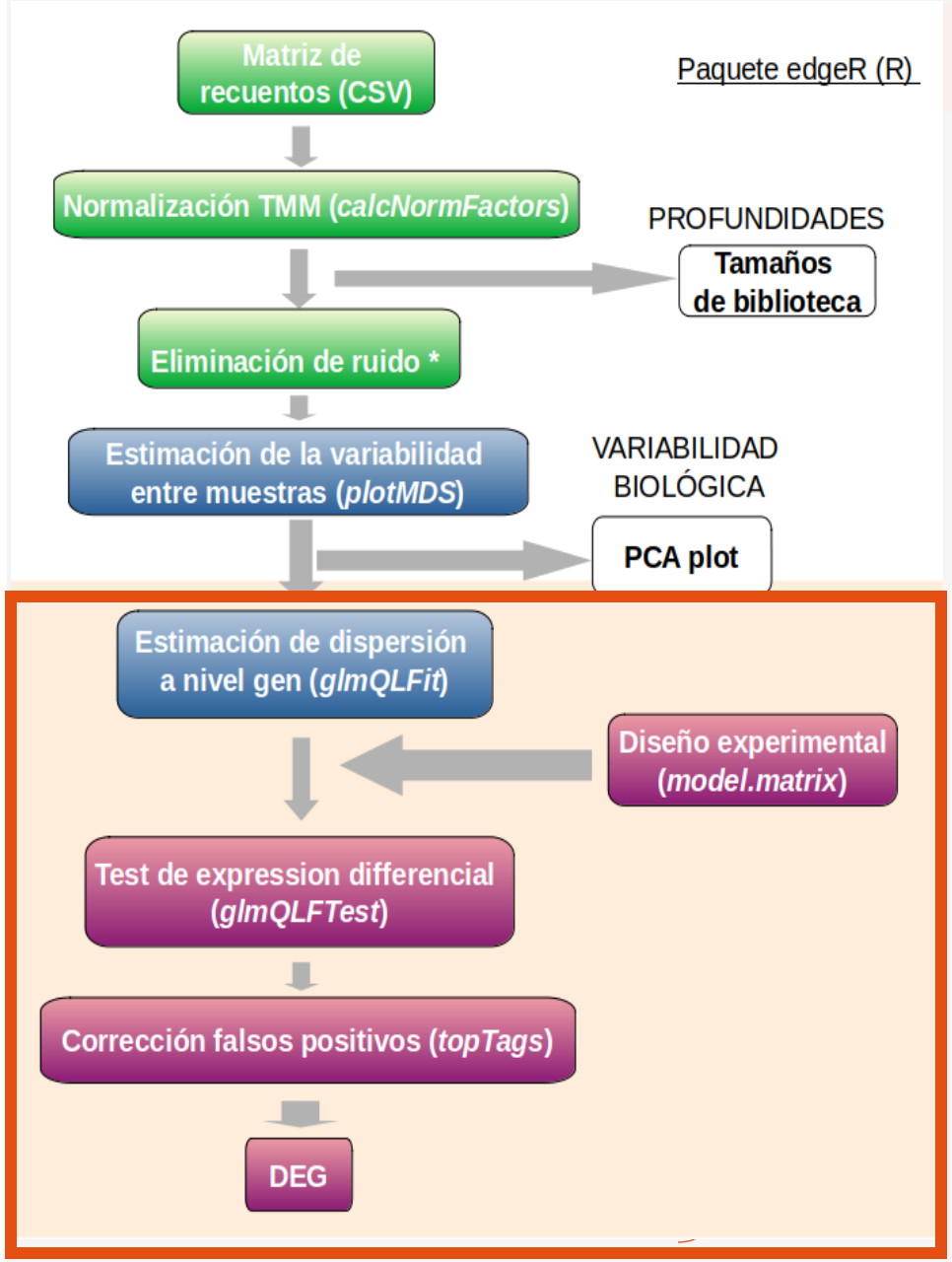
Sesión 10

Universidad Internacional de Valencia

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

De:
Planeta Formación y Universidades

Bloque IV: Análisis estadístico de la diferencia de expresión



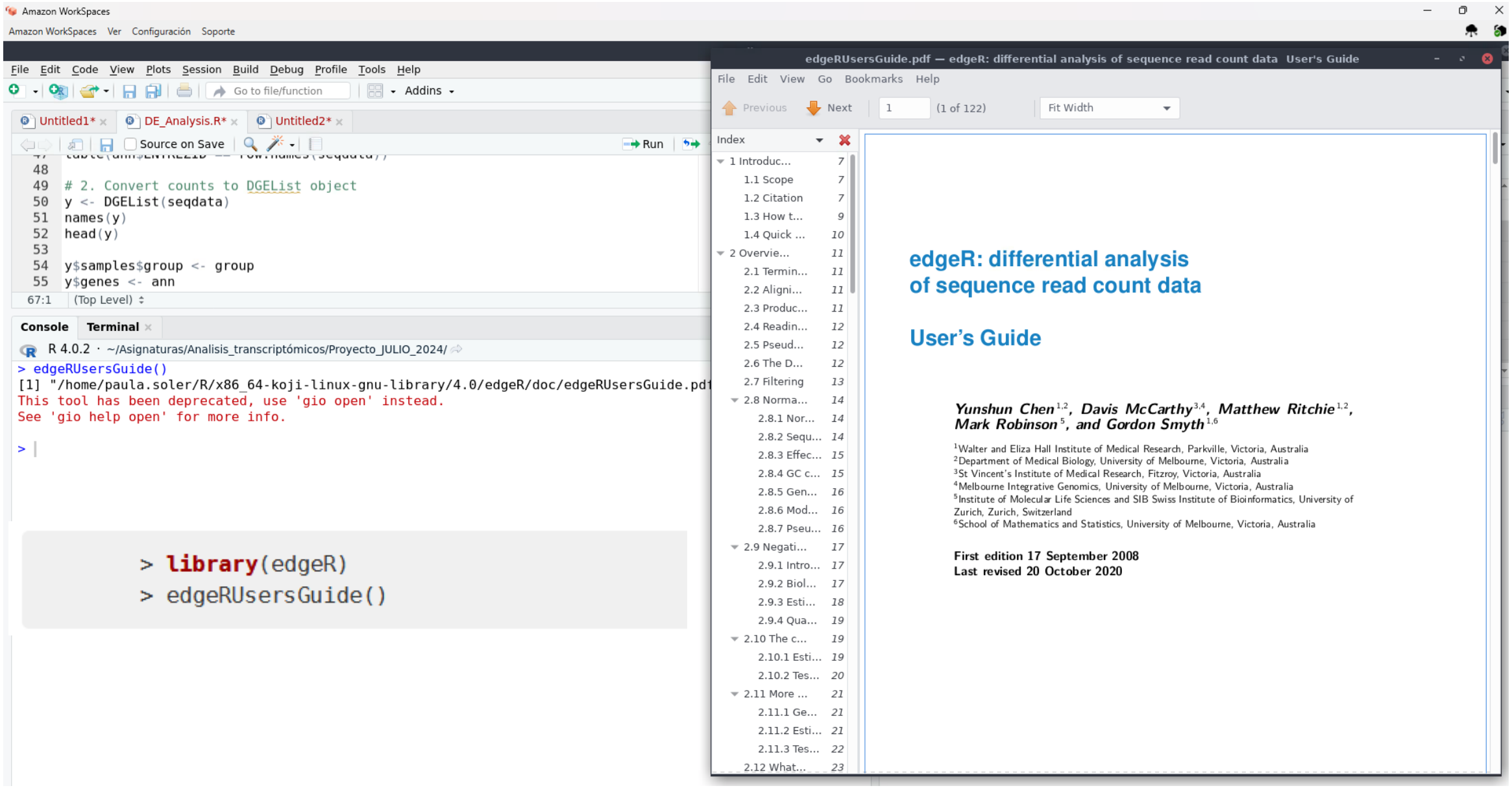
Objetivos

1. Entender y computar la estimación de la **sobredispersión** en el conjunto de datos.

2. Entender y computar el **ajuste** de modelos lineales (GLM) a nivel de gen mediante el método **quasi-likelihood**.

3. Conocer el proceso de **prueba de hipótesis nula** para el cálculo de la significancia en la diferencia de expresión.

Guía del usuario de edgeR



Matriz de diseño (*design matrix*)

Matriz de diseño

Relación de las muestras experimentales entre sí.

Función

model.matrix (~ 0 + **group**)

- **+ group**: Es una **variable categórica** que contiene los diferentes grupos que se están comparando.
- **~**: Es el **símbolo de fórmula** en R, que especifica que estamos definiendo una relación entre variables.
- **0**: Indica que **no se incluirá el intercepto** en la matriz de diseño.

```
> table(group)
group
   basal.lactate   basal.pregnant     basal.virgin  luminal.lactate
               2               2               2               2
luminal.pregnant   luminal.virgin
               2               2
```

Matriz de diseño (*design matrix*)

Matriz de diseño

Relación de las muestras experimentales entre sí.

Función

model.matrix (~ 0 + **group**)

- **+ group**: Es una **variable categórica** que contiene los diferentes grupos que se están comparando.
- **~**: Es el **símbolo de fórmula** en R, que especifica que estamos definiendo una relación entre variables.
- **0**: Indica que **no se incluirá el intercepto** en la matriz de diseño.

Si hubiera una muestra control entonces si que habría que poner esa muestra como intercepto, para que las muestras se compararan con el control

```
> design <- model.matrix(~0+group)
> colnames(design) <- levels(group)
> design
   B.lactating B.pregnant B.virgin L.lactating L.pregnant L.virgin
1            0          0        1           0          0         0
2            0          0        1           0          0         0
3            0          1        0           0          0         0
4            0          1        0           0          0         0
5            1          0        0           0          0         0
6            1          0        0           0          0         0
7            0          0        0           0          0         1
8            0          0        0           0          0         1
9            0          0        0           0          1         0
10           0          0        0           0          1         0
11           0          0        0           1          0         0
12           0          0        0           1          0         0
attr(,"assign")
[1] 1 1 1 1 1 1
attr(,"contrasts")
attr(,"contrasts")$group
[1] "contr.treatment"
```

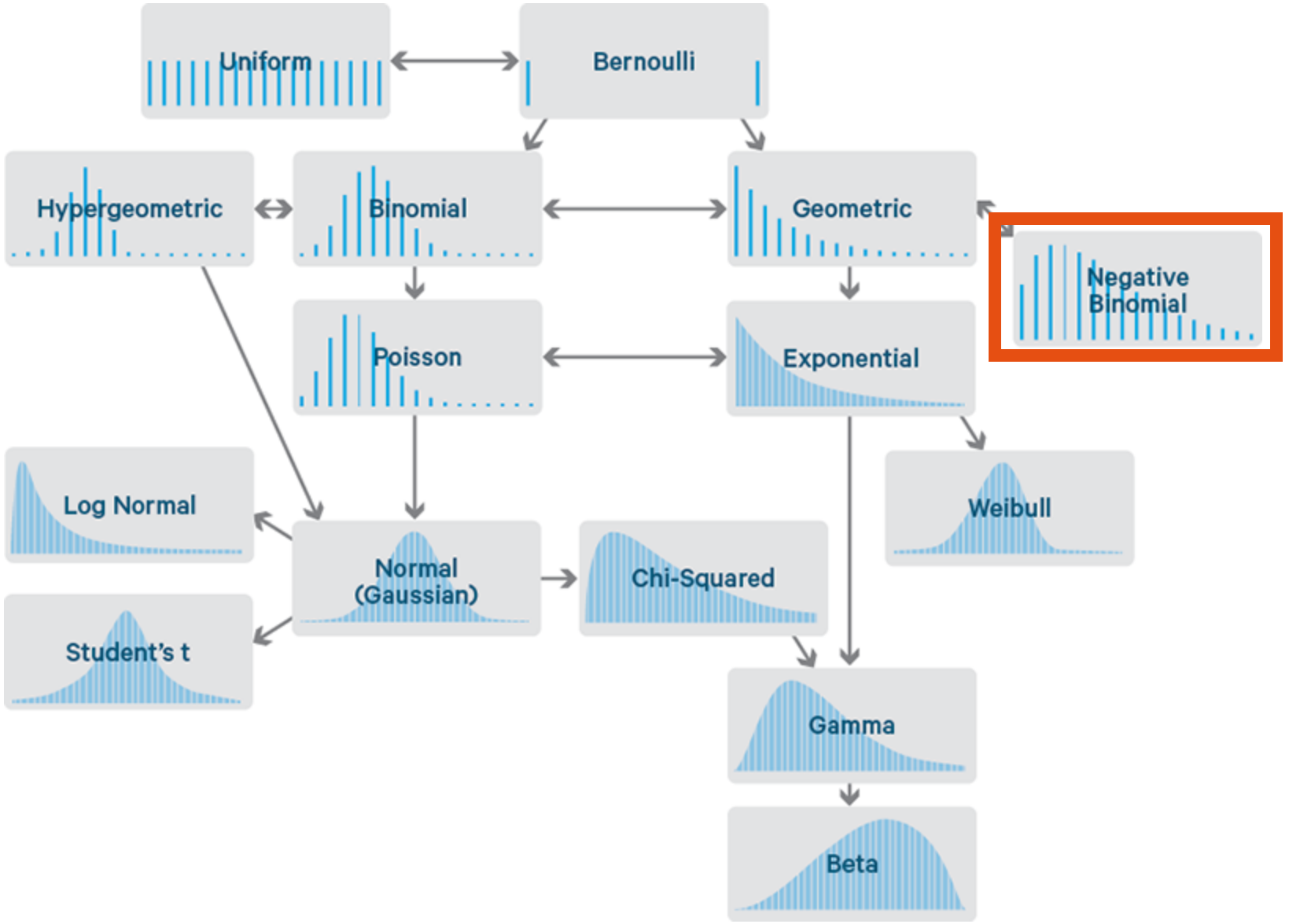
PRACTIQUEMOS

amazon WorkSpaces

Creación de la matriz de diseño

```
# Libraries
library(ggplot2)
library(edgeR)
library(statmod)


# install.packages("statmod")


# 1. Design matrix model
design <- model.matrix(~ 0 + group)
colnames(design) <- levels(group)
design
```

Distribuciones de datos



La alternativa: Distribución binominal negativa

$$Y_{gi} \sim NB(M_i\, p_{gj},\, \phi_g)$$

M, hace referencia al tamaño de la librería que tenemos por i (en cada una de nuestras muestras)

P, hace referencia a la abundancia relativa que tenemos para cada gen en los distintos grupos experimentales

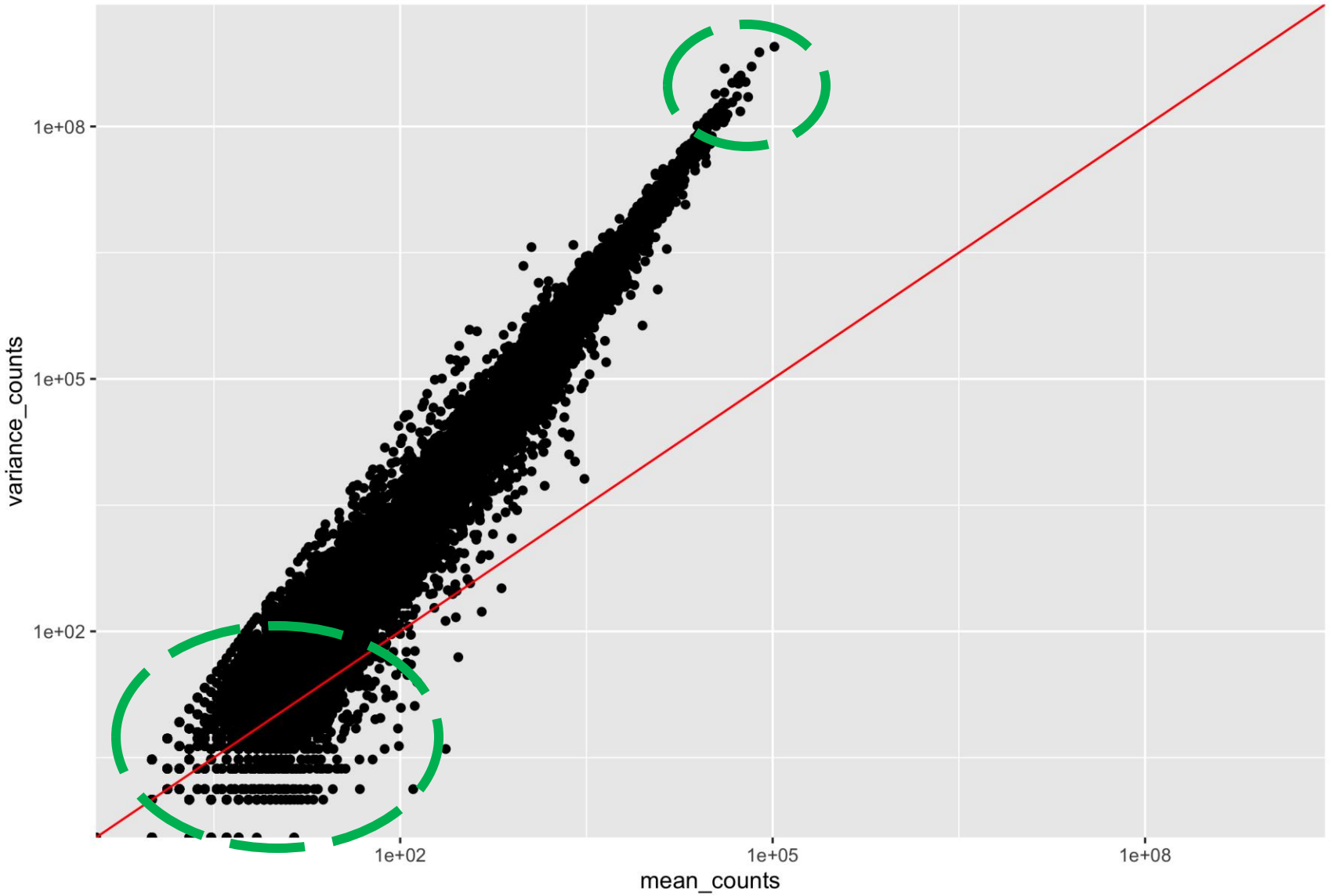
- El tamaño de la librería
- La composición o abundancia relativa de cada gen

Estimación de la sobredispersión

VARIANZA

VARIABILIDAD

Estimación de la dispersión



¿Pero cómo utiliza la dispersión EdgeR?

La Sobredispersión en Datos RNA-seq:

- Los datos de RNA-seq presentan una variabilidad mayor de la esperada bajo una distribución de Poisson (varianza > media).
- Esta sobredispersión se debe a factores biológicos y técnicos.

La Solución de EdgeR: Distribución Binomial Negativa (NB)

Estimación del Coeficiente de Variabilidad Biológica (BCV):

- El BCV representa la variabilidad biológica inherente a los datos.
- Se estima a partir de la variabilidad observada entre las réplicas biológicas.

EdgeR considera tres tipos de dispersión:

- **Common Dispersion:** Una dispersión común para todos los genes.
- **Trended Dispersion:** La dispersión varía según la media de expresión.
- **Tagwise Dispersion:** Cada gen tiene su propia dispersión.

Modelo Lineal Generalizado (GLM): Se utiliza un GLM para modelar la expresión de cada gen individualmente.

Tipos de dispersión

Dispersion measures:

VARIANCE
"**Common dispersion**"

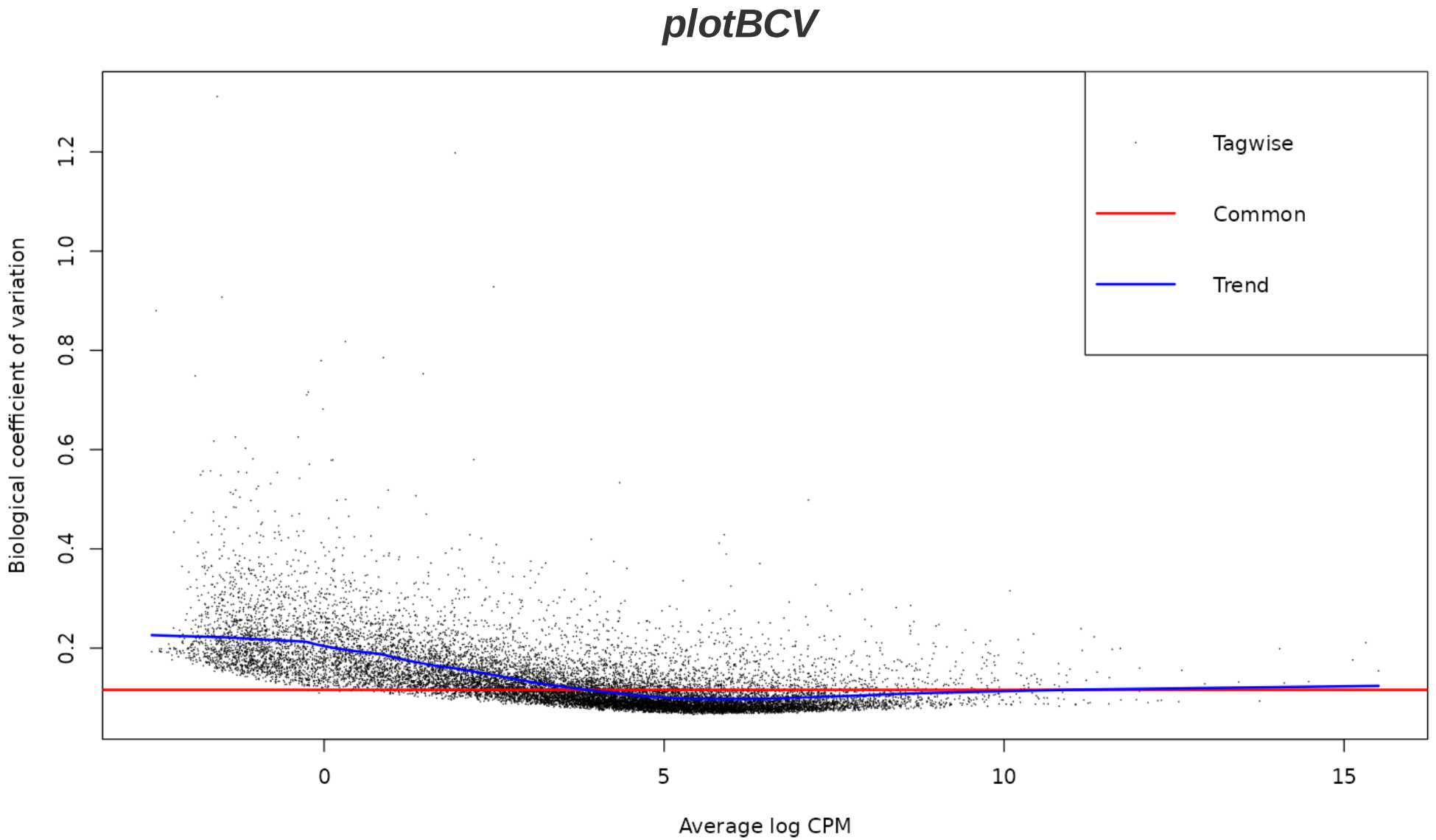
MEDIA DESVIATION (MD)
"**Trended dispersion**"

STANDARD DEVIATION (SD)
"**Tagwise dispersion**"

Observations of a NB distribution:

	C1	C2	C3	C4	E1	E2	E3	E4
gen_1	679	991	404	332	434	688	1204	693
gen_2	1676	777	1842	954	2144	1244	1305	627
gen_3	1278	648	815	517	1515	529	349	850
gen_4	908	339	1190	2022	784	877	1066	1094
gen_5	2152	1369	1099	278	922	1139	1285	496
gen_6	1002	1635	1410	874	973	1429	1427	912
gen_7	894	708	950	657	412	581	639	1373
gen_8	1628	397	1709	832	1454	313	1389	674
gen_9	605	793	1589	1320	1391	676	637	867
gen_10	465	1275	1248	943	658	1291	645	960

Tipos de dispersión



Cuanto más bajo el valor de variabilidad biológica más similitud tendrán las muestras

Ajuste de modelos GLM gen a gen.

EdgeR ajusta este nivel de dispersión para cada gen a un modelo concreto dentro de la familia de modelos generalizados negativos binomiales.

- Disminuir el nivel de falsos positivos
- Aumentar el número de DEG reales.

método quasi-likelihood (QL)

Flujo de trabajo

DGEList

```
> 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
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
21399     1604    1495    1721    1317    1159    1066    1334    1258    1068     926     508     500
15799 more rows ...

$samples
                group lib.size norm.factors
MCL1.DG   basal.virgin 23218026     1.236899
MCL1.DH   basal.virgin 21768136     1.213949
```

Model Matrix

	basal.lactate	basal.pregnant	basal.virgin	luminal.lactate	luminal.pregnant	luminal.virgin
1	0	0	1	0	0	0
2	0	0	1	0	0	0
3	0	1	0	0	0	0
4	0	1	0	0	0	0
5	1	0	0	0	0	0
6	1	0	0	0	0	0
7	0	0	0	0	0	1
8	0	0	0	0	0	1
9	0	0	0	0	1	0
10	0	0	0	0	1	0
11	0	0	0	1	0	0
12	0	0	0	1	0	0

robust va a proteger las estimaciones contra estos genes que tienen excepcionalmente una dispersión muy elevada o muy individual.

```
y <- estimateDisp(y, diseño, robust=TRUE)
```



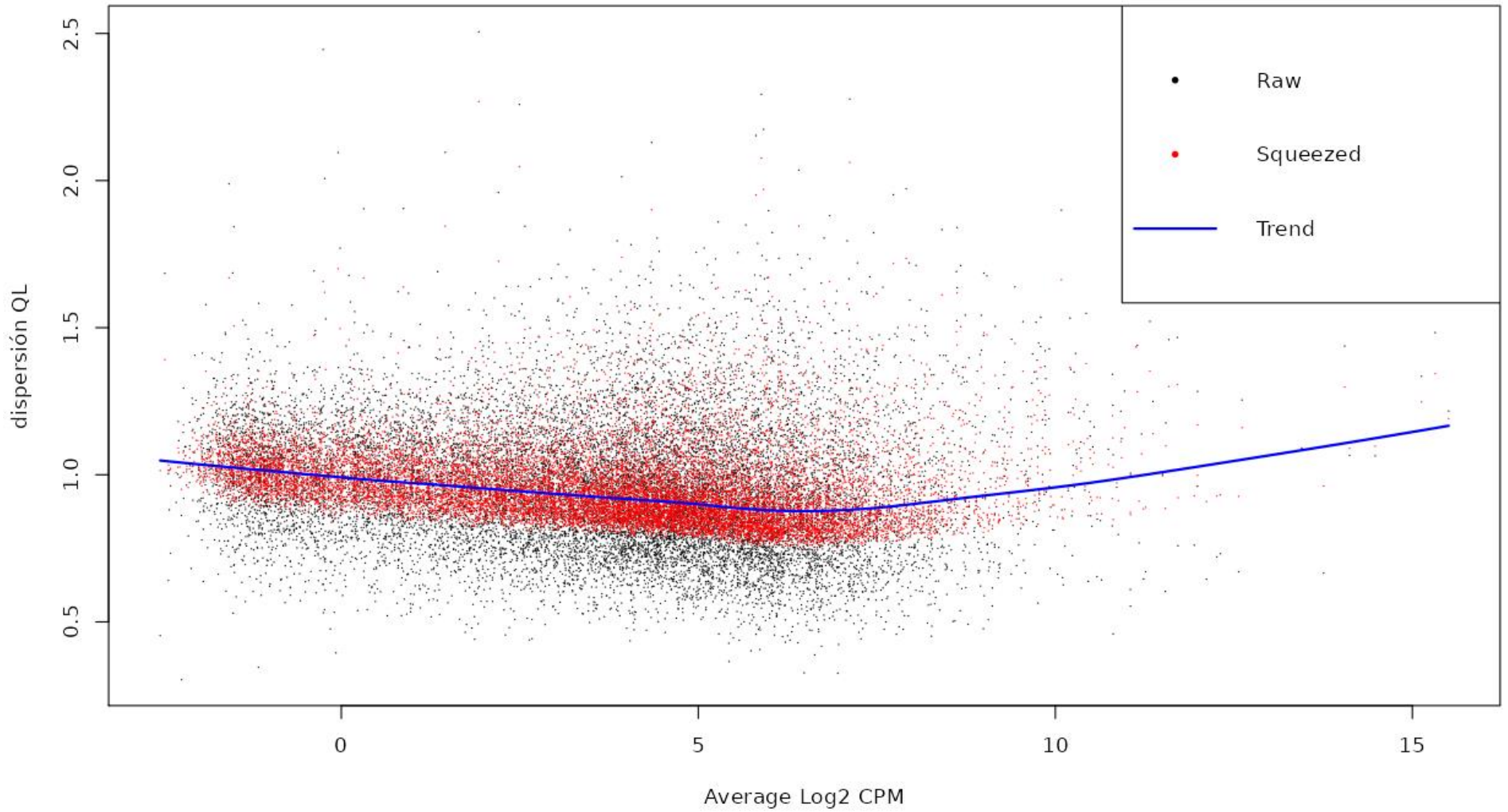
```
fit <- glmQLFit(y, diseño, robust=TRUE)
```

DGEGLM

```
fit <- glmQLFit(y, diseño, robust=TRUE)
```

fit	list [32826 x 3] (edgeR::DGE	List of length 17
coefficients	double [32826 x 3]	-14.1 -17.8 -18.8 -17.4 -17.8 -14.7 -15.1 -18.8 -17.8 -18.8 -18.8 -16.6 -14.5 -1 ...
fitted.values	double [32826 x 15]	13.420 0.204 0.000 0.408 0.204 7.343 12.840 0.195 0.000 0.390 0.195 7. ...
deviance	double [32826]	7.83 3.11 6.17 8.98 3.06 13.94 ...
method	character [1]	'oneway'
counts	integer [32826 x 15]	15 0 0 0 1 3 16 0 0 0 0 10 5 0 0 0 0 4 16 0 0 0 0 8 14 1 0 2 0 11 8 0 1 0 0 1 3 ...
unshrunk.coefficients	double [32826 x 3]	-1.41e+01 -1.83e+01 -1.00e+08 -1.76e+01 -1.83e+01 -1.47e+01 -1.51e+01 -1.00e+08 ...
df.residual	integer [32826]	12 12 12 12 12 12 ...
design	double [15 x 3]	1 1 1 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 ...
offset	double [32826 x 15] (S3: Cor	16.7 16.7 16.7 16.8 16.7 16.7 16.7 16.7 16.7 16.7 16.7 16.7 16.6 16.6 16.7 ...
dispersion	double [32826]	0.0951 0.1654 0.1654 0.1654 0.1654 0.1111 ...
prior.count	double [1]	0.125
AveLogCPM	double [32826]	-0.708 -3.101 -3.055 -3.011 -3.101 -1.596 ...

plotQLDisp


dispersión QL
2.5
2.0
1.5
1.0
0.5
0
5
10
15
Average Log2 CPM
Raw
Squeezed
Trend

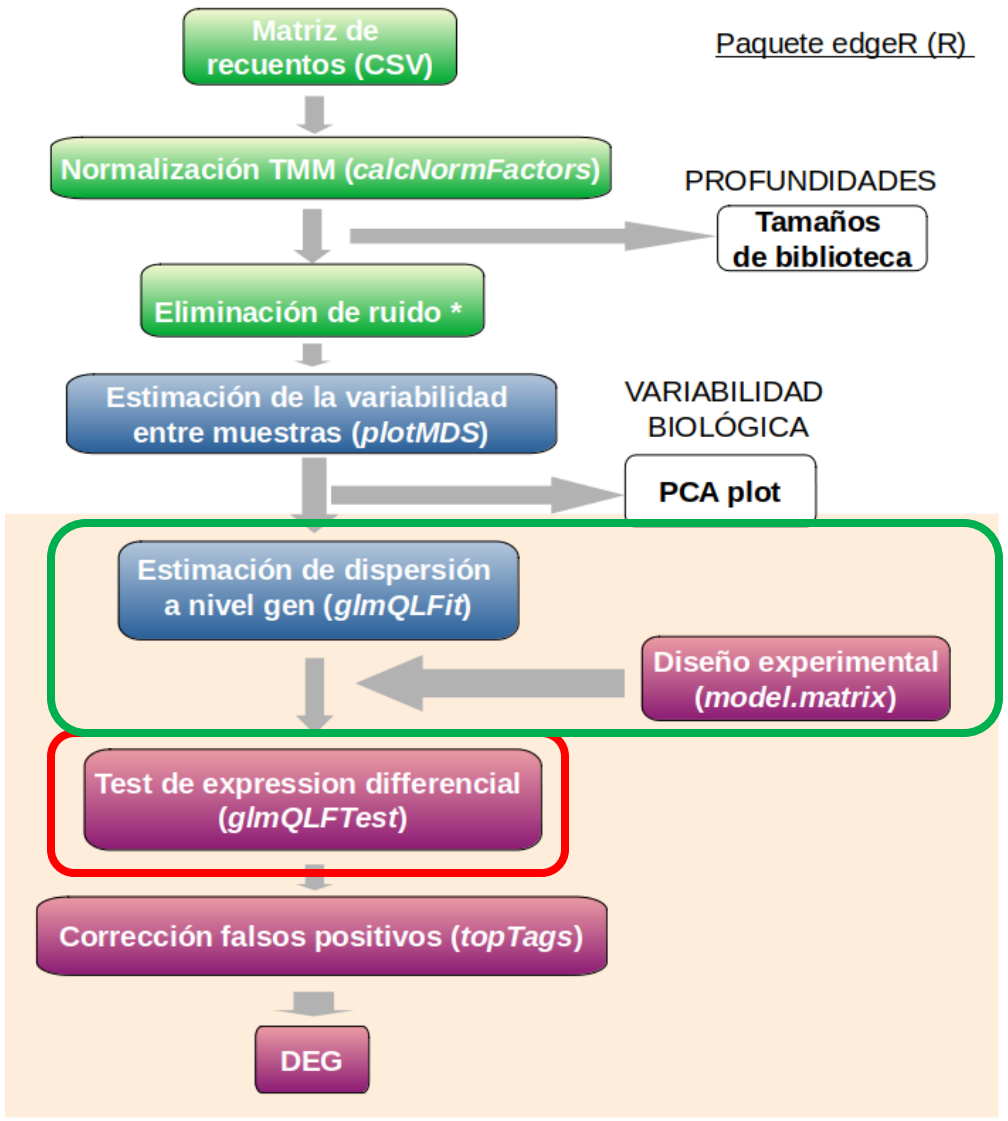
PRACTIQUEMOS

amazon WorkSpaces

Estimación de las dispersions y ajuste de modelo

```
# 2. Estimate dispersions

y <- estimateDisp(y, design, robust=TRUE)
y$common.dispersion
head(y$trended.dispersion)
head(y$tagwise.dispersion)

plotBCV(y)



# 3. GLM Model Adjustment
fit <- glmQLFit(y, design, robust=TRUE)

head(fit$counts)
head(fit$fitted.values)

plotQLDisp(fit, ylab = "dispersión QL")
```

Flujo de trabajo del análisis estadístico de la expresión génica

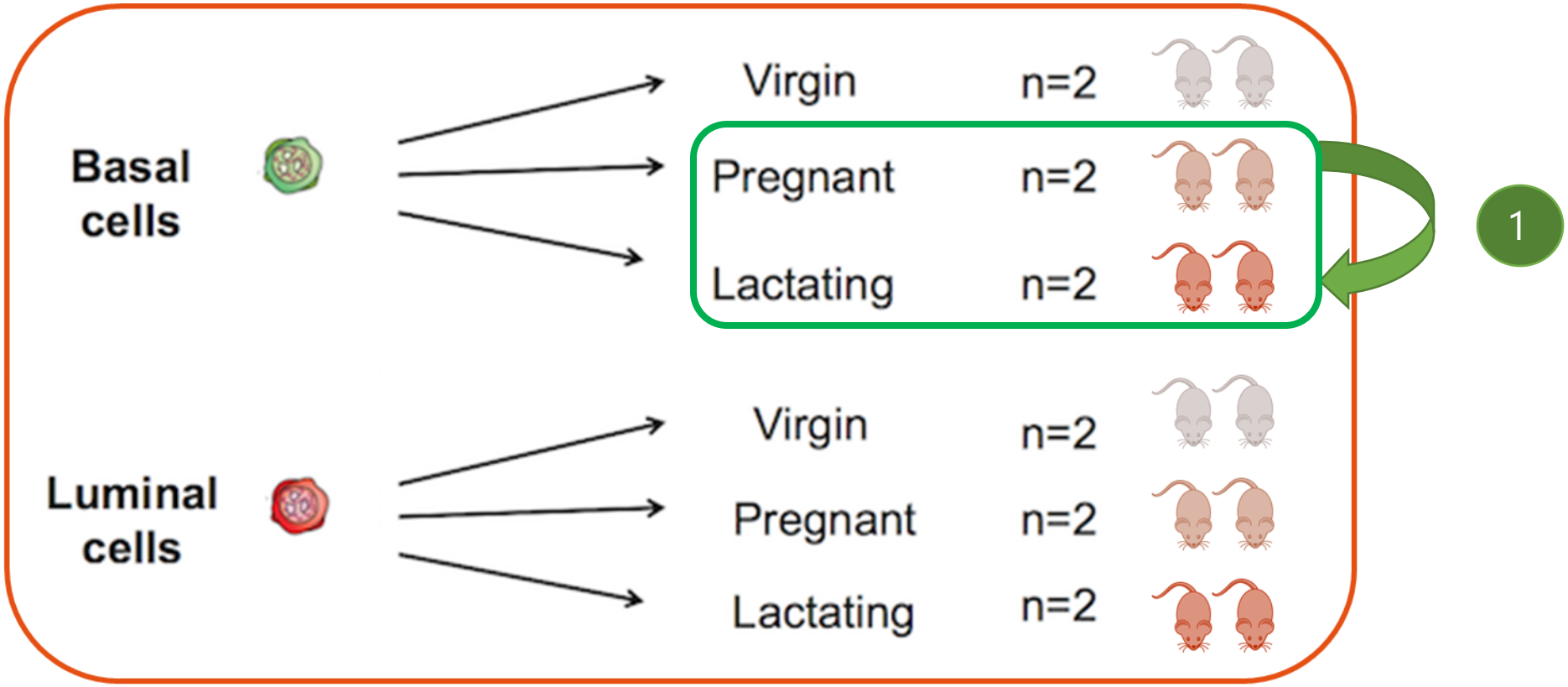


fit <- glmQLFit(y, design = design, robust= TRUE)

res <- **glmQLFTest** (**fit**, contrast)

Con Fit ajustamos y con QLFtest es la que vamos a usar para hacer esos test estadísticos , el testeo de hipótesis nula, y se quiere estudiar la probabilidad con la que un determinado gen cambie o no su expresión teniendo en cuenta qué grupos se va a comparar (esto se hace con lo de contraste)

¿Qué comparaciones vamos a realizar?



Fu, Nai Yang, et al. "EGF-mediated induction of Mcl-1 at the switch to lactation is essential for alveolar cell survival." *Nature Cell Biology*

¿Qué comparaciones vamos a realizar?

Basal lactate (n=2)

Basal pregnant (n=2)

B.LvsP <- **makeContrasts**(basal.lactate-basal.pregnant, levels=design)

design es la matriz de diseño

```
> B.LvsP
                  Contrasts
Levels             basal.lactate - basal.pregnant
  basal.lactate                                 1
  basal.pregnant                               -1
  basal.virgin                                  0
  luminal.lactate                               0
  luminal.pregnant                              0
  luminal.virgin                                0
```

Con 1 y -1 señala los grupos que vamos a comparar, con 0 los grupos que no se van a comparar

Prueba de significancia o de hipótesis nula

res <- ***glmQLFTest***(fit, contrast= B.LvsP)

fit <- **glmQLFit**(y, design = design, robust= TRUE)

fit	list [32826 x 3] (edgeR::DGE	List of length 17
coefficients	double [32826 x 3]	-14.1 -17.8 -18.8 -17.4 -17.8 -14.7 -15.1 -18.8 -17.8 -18.8 -18.8 -16.6 -14.5 -1 ...
fitted.values	double [32826 x 15]	13.420 0.204 0.000 0.408 0.204 7.343 12.840 0.195 0.000 0.390 0.195 7. ...
deviance	double [32826]	7.83 3.11 6.17 8.98 3.06 13.94 ...
method	character [1]	'oneway'
counts	integer [32826 x 15]	15 0 0 0 1 3 16 0 0 0 0 10 5 0 0 0 0 4 16 0 0 0 0 8 14 1 0 2 0 11 8 0 1 0 0 1 3 ...
unshrunk.coefficients	double [32826 x 3]	-1.41e+01 -1.83e+01 -1.00e+08 -1.76e+01 -1.83e+01 -1.47e+01 -1.51e+01 -1.00e+08 ...
df.residual	integer [32826]	12 12 12 12 12 12 ...
design	double [15 x 3]	1 1 1 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 ...
offset	double [32826 x 15] (S3: Co	16.7 16.7 16.7 16.8 16.7 16.7 16.7 16.7 16.7 16.7 16.7 16.7 16.6 16.6 16.7 ...
dispersion	double [32826]	0.0951 0.1654 0.1654 0.1654 0.1654 0.1111 ...
prior.count	double [1]	0.125
AveLogCPM	double [32826]	-0.708 -3.101 -3.055 -3.011 -3.101 -1.596 ...

B.LvsP <- **makeContrasts**(basal.lactate-basal.pregnant, levels=design)

```
> B.LvsP
                  Contrasts
Levels             basal.lactate - basal.pregnant
  basal.lactate                                 1
  basal.pregnant                               -1
  basal.virgin                                  0
  luminal.lactate                               0
  luminal.pregnant                              0
  luminal.virgin                                0
```

Prueba de significancia: Resultados -> LogFoldChange

```
head(res$table)
dim(res$table)
[1] 15804     4
```

	logFC	logCPM	F	PValue
497097	1.26683972	2.572255	5.9734805	0.04128796
20671	-0.36599078	1.310053	0.7765410	0.39539097
27395	0.03534833	3.998092	0.0414571	0.84186948
18777	0.08527488	5.055784	0.5740522	0.46241732
21399	-0.22250269	5.629011	4.6412559	0.05093238
58175	-1.87817810	1.156086	4.4224714	0.05603412

Muestra el valor o nivel de cambio de expresión de cada gen respecto a los grupos determinados.

- Se calcula como la relación entre las medias de las réplicas de un grupo frente a las del otro.
- Escala logarítmica
- Escala simétrica

Prueba de significancia: Resultados -> LogCPM

```
head(res$table)
dim(res$table)
[1] 15804     4
```

	logFC	logCPM	F	PValue
497097	1.26683972	2.572255	5.9734805	0.04128796
20671	-0.36599078	1.310053	0.7765410	0.39539097
27395	0.03534833	3.998092	0.0414571	0.84186948
18777	0.08527488	5.055784	0.5740522	0.46241732
21399	-0.22250269	5.629011	4.6412559	0.05093238
58175	-1.87817810	1.156086	4.4224714	0.05603412

Expresión media del gen en todas las muestras

- Escala logarítmica y en conteos por millón

Prueba de significancia: Resultados -> F

```
head(res$table)
dim(res$table)
[1] 15804     4
```

	logFC	logCPM	F	PValue
497097	1.26683972	2.572255	5.9734805	0.04128796
20671	-0.36599078	1.310053	0.7765410	0.39539097
27395	0.03534833	3.998092	0.0414571	0.84186948
18777	0.08527488	5.055784	0.5740522	0.46241732
21399	-0.22250269	5.629011	4.6412559	0.05093238
58175	-1.87817810	1.156086	4.4224714	0.05603412

Estadístico F de la prueba F de cuasi-verosimilitud.
- Se asocia con la probabilidad obtenida en el test de la hipótesis nula

¿Cuántos genes diferenciales tenemos que reportar?

```
head(res$table)
dim(res$table)
[1] 15804     4
```

	logFC	logCPM	F	PValue
497097	**1.26683972**	**2.572255**	**5.9734805**	**0.04128796**
20671	-0.36599078	1.310053	0.7765410	0.39539097
27395	0.03534833	3.998092	0.0414571	0.84186948
18777	0.08527488	5.055784	0.5740522	0.46241732
21399	-0.22250269	5.629011	4.6412559	0.05093238
58175	-1.87817810	1.156086	4.4224714	0.05603412

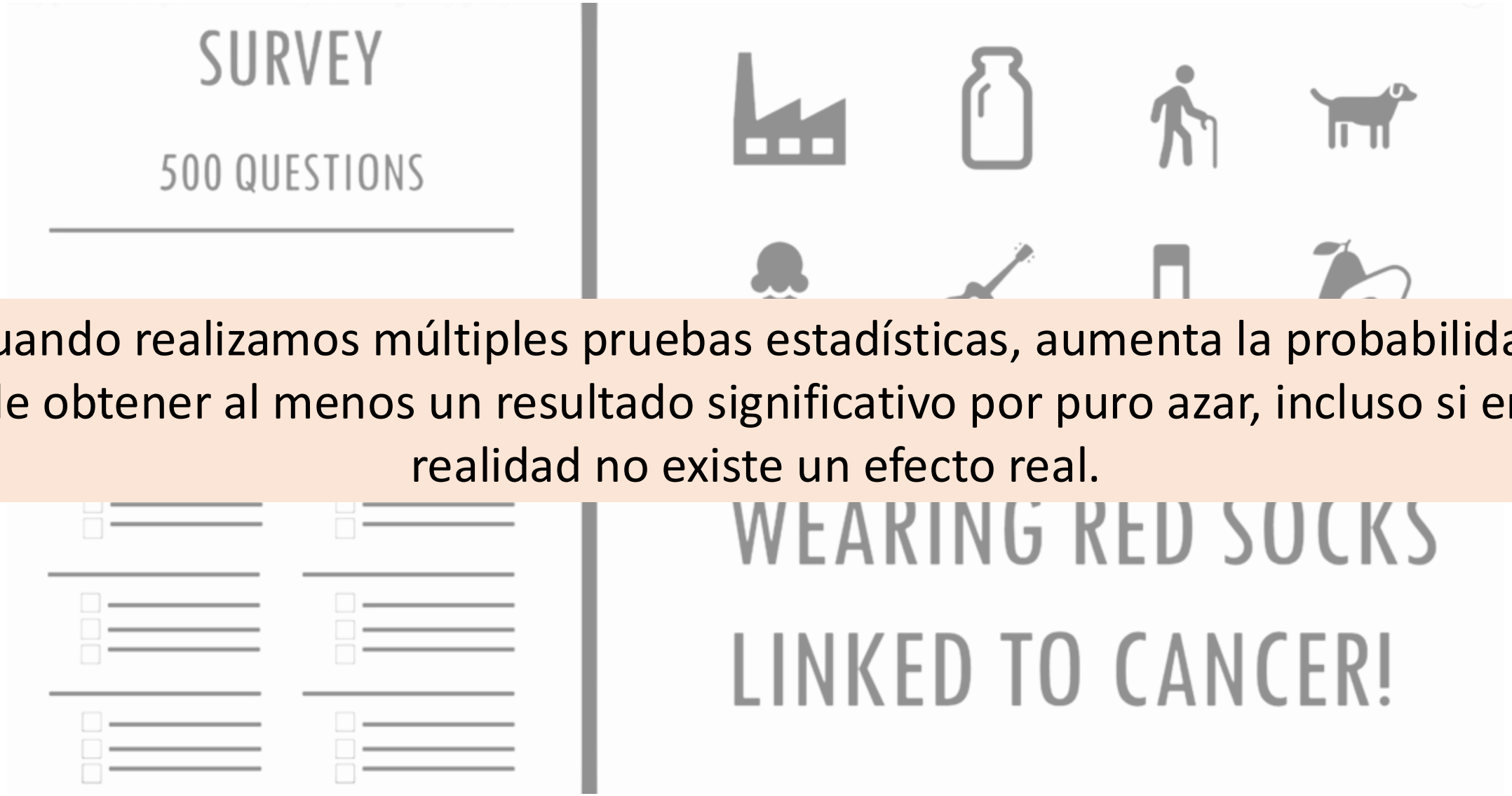
cuantos mas test estadísticos realicemos mayor es la probabilidad de tener falsos positivos

La probabilidad de que el valor estadístico sea significativo

p value < 0.05

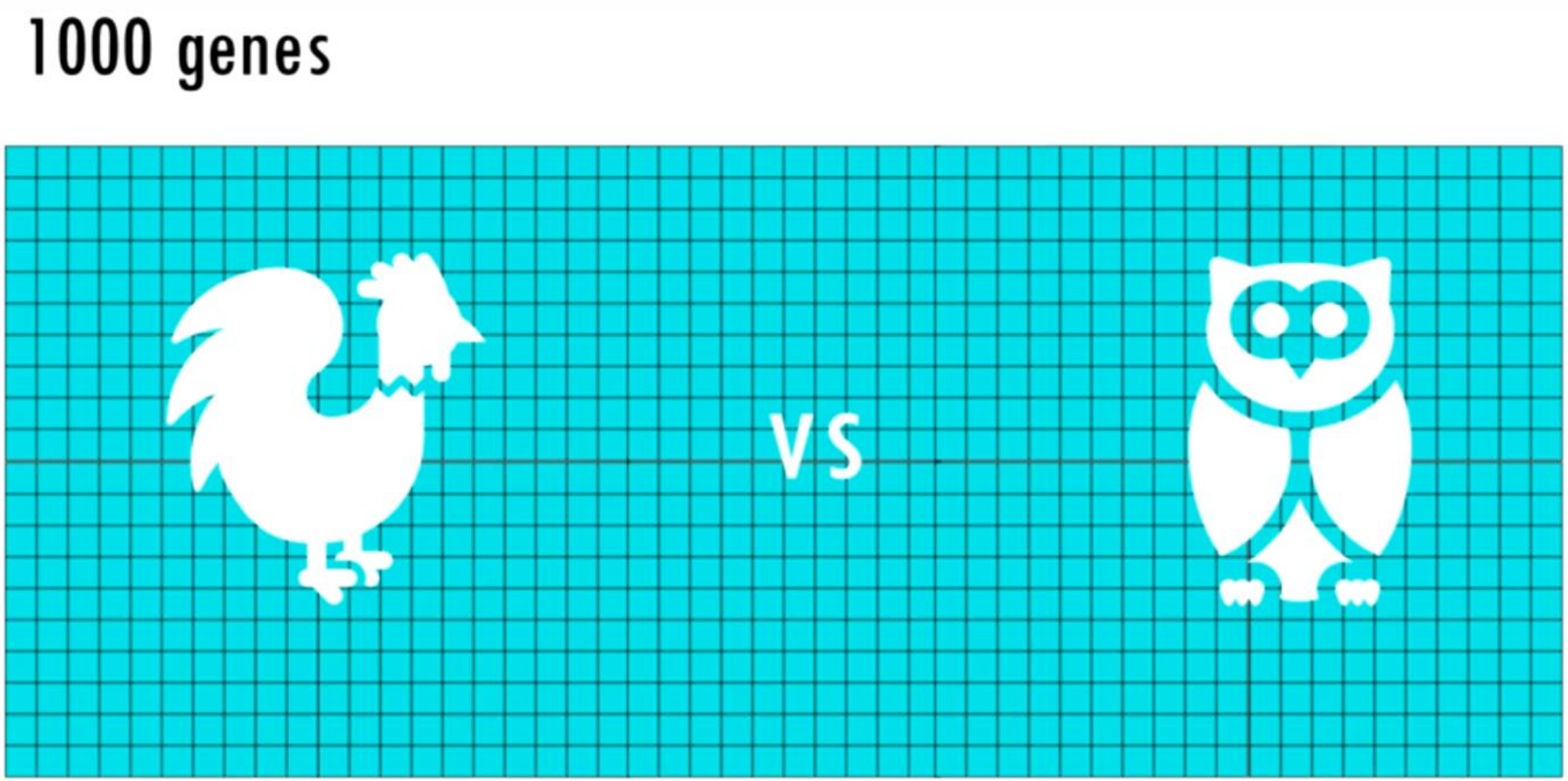
Problema del testeo múltiple (Multiple testing)

Problema del testeo múltiple (Multiple testing)

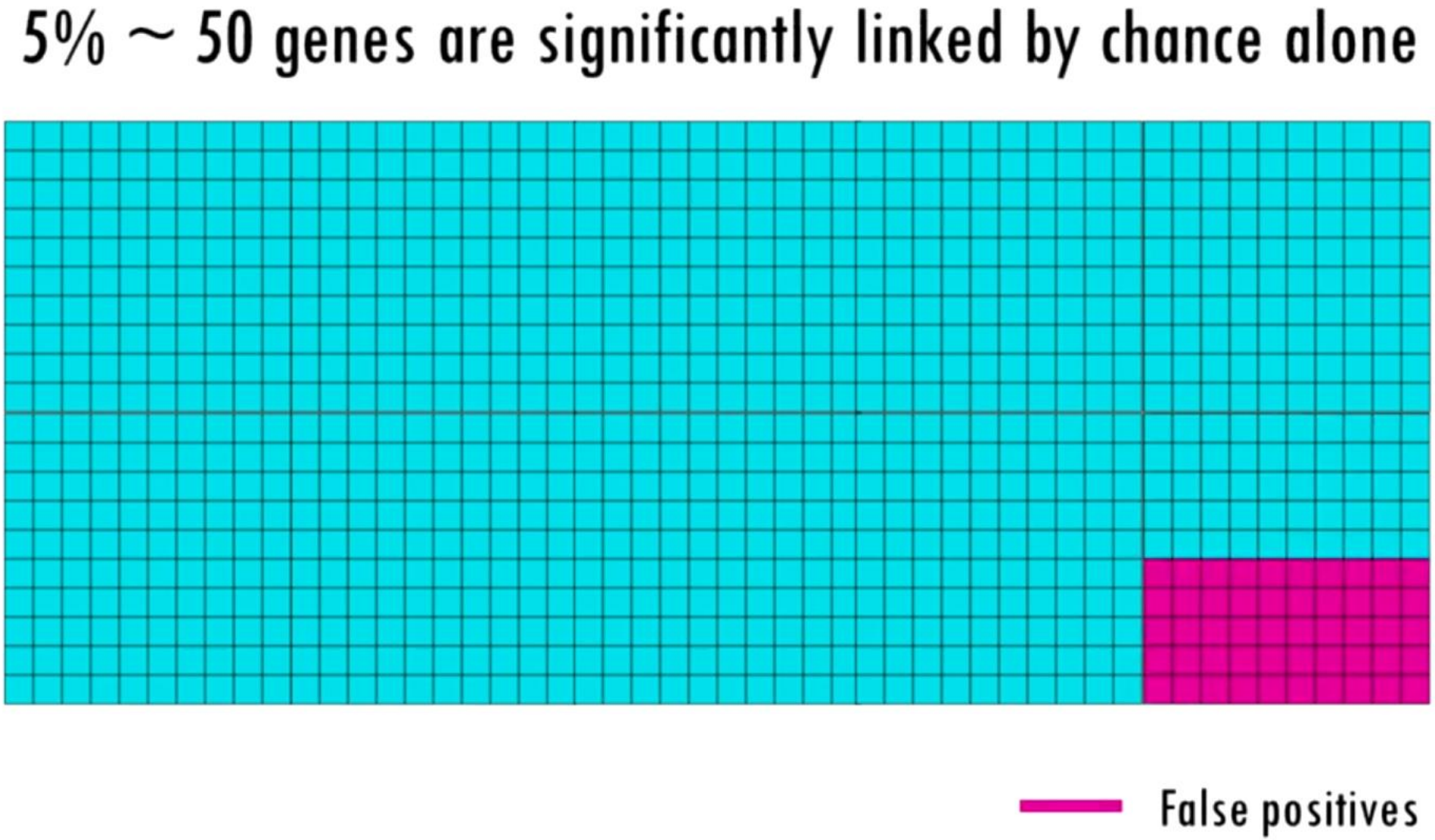


Cuando realizamos múltiples pruebas estadísticas, aumenta la probabilidad de obtener al menos un resultado significativo por puro azar, incluso si en realidad no existe un efecto real.

¿Cuántos genes diferenciales tenemos que reportar?



¿Cuántos genes diferenciales tenemos que reportar?



El método de Benjamini-Hochberg (1995)

Genes with p-value < 0.05 which are actually not significant, it was just by chance that they got p-value < 0.05

$$FDR = \frac{\text{False positives}}{\text{All significant results}}$$

All significant genes ($p\text{-value} < 0.05$)

false discovered read, va a tener en cuenta solo los genes significativos.

El método de Benjamini-Hochberg (1995)

p-values: 0.361, 0.387, 0.005, 0.009, 0.022, 0.051, 0.101, 0.019

Rank (*k*)	Sorted p-values	$p_{(k)}\frac{m}{k}$
1	0.005	0.005 x (8/1) = 0.040
2	0.009	0.009 x (8/2) = 0.036
3	0.019	0.019 x (8/3) = 0.051
4	0.022	0.022 x (8/4) = 0.044
5	0.051	0.051 x (8/5) = 0.082
6	0.101	0.101 x (8/6) = 0.135
7	0.361	0.361 x (8/7) = 0.413
8	0.387	0.387 x (8/8) = 0.387

- **P** es el p-value computado anteriormente
- **K** es el rango (n = 1 al 8)
- **M** es el total de p-values que estamos comparando (n = 8)

El método de Benjamini-Hochberg (1995)

p-values: 0.361, 0.387, 0.005, 0.009, 0.022, 0.051, 0.101, 0.019

Rank (k)	Sorted p-values	$p_{(k)} \frac{m}{k}$	BH adjusted p-value
1	0.005	0.005 x (8/1) = 0.040	0.036
2	0.009	0.009 x (8/2) = 0.036	0.036
3	0.019	0.019 x (8/3) = 0.051	0.044
4	0.022	0.022 x (8/4) = 0.044	0.044
5	0.051	0.051 x (8/5) = 0.082	0.082
6	0.101	0.101 x (8/6) = 0.135	0.135
7	0.361	0.361 x (8/7) = 0.413	0.387
8	0.387	0.387 x (8/8) = 0.387	0.387

El método de Benjamini-Hochberg (1995)

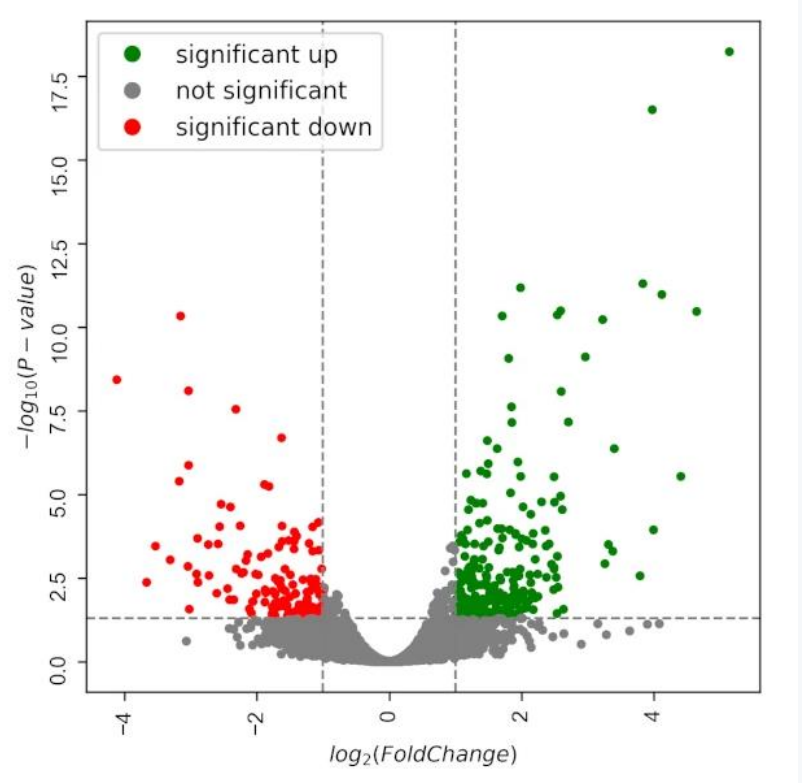
p-values: 0.361, 0.387, 0.005, 0.009, 0.022, 0.051, 0.101, 0.019

Rank (k)	Sorted p-values	$p_{(k)} \frac{m}{k}$	BH adjusted p-value	Reject?
1	0.005	0.005 x (8/1) = 0.040	0.036	YES
2	0.009	0.009 x (8/2) = 0.036	0.036	YES
3	0.019	0.019 x (8/3) = 0.051	0.044	YES
4	0.022	0.022 x (8/4) = 0.044	0.044	YES
5	0.051	0.051 x (8/5) = 0.082	0.082	NO
6	0.101	0.101 x (8/6) = 0.135	0.135	NO
7	0.361	0.361 x (8/7) = 0.413	0.387	NO
8	0.387	0.387 x (8/8) = 0.387	0.387	NO

BH adjusted p-value < 0,05

Función **topTags**

```
> topTags(res)
Coefficient:  1*B.lactating -1*B.pregnant
       Length  Symbol logFC logCPM   F   PValue      FDR
12992     765 Csn1s2b  6.09  10.18 421 4.86e-11 7.68e-07
211577   2006  Mrgprf  5.15   2.74 345 1.30e-10 8.05e-07
226101   7094    Myof  2.32   6.44 322 1.97e-10 8.05e-07
381290   8292  Atp2b4  2.14   6.14 320 2.04e-10 8.05e-07
140474  11281    Muc4 -7.17   6.05 308 2.64e-10 8.34e-07
231830   3346 Micall2 -2.25   5.18 282 4.47e-10 1.18e-06
24117    2242    Wif1 -1.82   6.76 260 7.30e-10 1.65e-06
12740    1812   Cldn4 -5.32   9.87 298 8.88e-10 1.71e-06
21953     667   Tnni2  5.75   3.86 313 9.76e-10 1.71e-06
231991   2873   Creb5  2.57   4.87 241 1.16e-09 1.83e-06
```



Recuperación de clase 06/09/2024

Universidad
Internacional
de Valencia

universidadviu.com

De:
Planeta Formación y Universidades