Tema 10. Minería Estadística de datos. Análisis de datos bioinformáticos

Beatriz Coronado Sanz

12 de febrero de 2019

1. Introducción

Vamos a estudiar un conjunto de datos RNA-seq reales utilizando el modelo de regresión binomial negativa. Para ello usaremos el proyecto de software libre Bioconductor y los paquetes. *DESeq* y *DESeq2*

Tras procesar los datos, realizaremos contrastes de hipótesis para detectar determinados genes y mostraremos unas gráficas con los resultados obtenidos.

2. Datos del trabajo

Usaremos datos procedentes de un experimento sobre el cultivo de células de la mosca *Drosophila melanogaster*.

Tras cargar todas las librerías necesarias, mostramos la cabecera de la tabla de conteos de nuestros datos. En esta tabla las filas corresponden a los genes y las columnas a las muestras o individuos.

	treated1fb	treated2fb	treated3fb	untreated1fb	untreated2fb	untreated3fb	untreated4fb
FBgn0000003	0	0	1	0	0	0	0
FBgn0000008	78	46	43	47	89	53	27
FBgn0000014	2	0	0	0	0	1	0
FBgn0000015	1	0	1	0	1	1	2
FBgn0000017	3187	1672	1859	2445	4615	2063	1711
FBgn0000018	369	150	176	288	383	135	174

Sus dimensiones son:

```
[1] 14470     7
```

Es necesario tener los datos en un formato adecuado para DESeqDataSet. Para realizar esto obtenemos la información de las columnas de la matriz de conteos.

```
             condition        type
treated1fb     treated single-read
treated2fb     treated  paired-end
treated3fb     treated  paired-end
untreated1fb untreated single-read
untreated2fb untreated single-read
untreated3fb untreated  paired-end
untreated4fb untreated  paired-end
```

Tenemos dos factores para cada individuo: la condición, que diferencia entre individuos no tratados e individuos tratados, y el tipo de secuenciación empleada, que puede ser "single-read" o "paired-end".

Tras esto construimos un DESeqDataSet. Para realizar el estudio queremos tener en cuenta el factor *condition*, para diferenciar entre individuos tratados e individuos no tratados. En nuestro ejemplo, tenemos 3 individuos tratados y 4 no tratados.

Aplicando la función *factor* a la columna que nos interesa en *colData*, asignamos un orden a los distintos niveles. De esta forma los individuos no tratados se consideran el nivel de control.

3. Estudios estadístico y análisis diferencial

El análisis de expresión diferencial en DESeq2 usa un modelo lineal generalizado de la forma $K_{ij} \sim NB(\mu_{ij}, \alpha_i)$ con $\mu_{ij} = s_j q_{ij} > 0, \alpha_i > 0$ donde K_{ij} es el número de conteos del gen i en la muestra j. Se sigue una distribución binomial negativa con una media ajustada μ_{ij} (producto de un factor de tamaño específicico para cada muestra s_j y un parámetro proporcional a la concentración de los fragmentos real esperada en la muestra j), y una dispersión específica para cada gen, α_i.

Como estamos interesados en encontrar aquellos genes que presenten diferentes niveles de expresión según se trate de un individuo control o un individuo tratado, el contraste que realizamos es:

$$\begin{cases} H_0 : \log_2 \frac{\mu_{iA}}{\mu_{iB}} = 0 \\ H_1 : \log_2 \frac{\mu_{iA}}{\mu_{iB}} \neq 0 \end{cases}$$

donde μ_{iA} es la media de conteos del gen i en el grupo A (individuos no tratados) y μ_{iB} la media de conteos del mismo gen pero en el grupo B (individuos tratados). A $\frac{\mu_{iA}}{\mu_{iB}}$ se le denomina *fold-change*.

Debemos realizar este contraste para cada uno de los genes de forma independiente.

Realizamos el análisis de expresión diferencial para nuestros datos. Se ordenan los resultados por orden creciente del p-valor ajustado. De esta forma, los primeros genes que aparecen en la tabla son los que muestran una mayor diferencia en los niveles de expresión entre los dos grupos.

```
log2 fold change (MLE): condition treated vs untreated
Wald test p-value: condition treated vs untreated
DataFrame with 6 rows and 6 columns
```

	baseMean	log2FoldChange	lfcSE
	<numeric>	<numeric>	<numeric>
FBgn0039155	453.275338598749	-4.41843095434387	0.202810428941139
FBgn0029167	2165.04449786732	-2.20328535517881	0.109970049254461
FBgn0035085	366.827879044973	-2.48223410077333	0.155683668262555
FBgn0029896	257.902702279936	-2.58137612216074	0.189388334413868
FBgn0034736	118.407382327029	-3.32697374867937	0.257868959634414
FBgn0040091	610.603484946151	-1.54669447836141	0.130704799638395

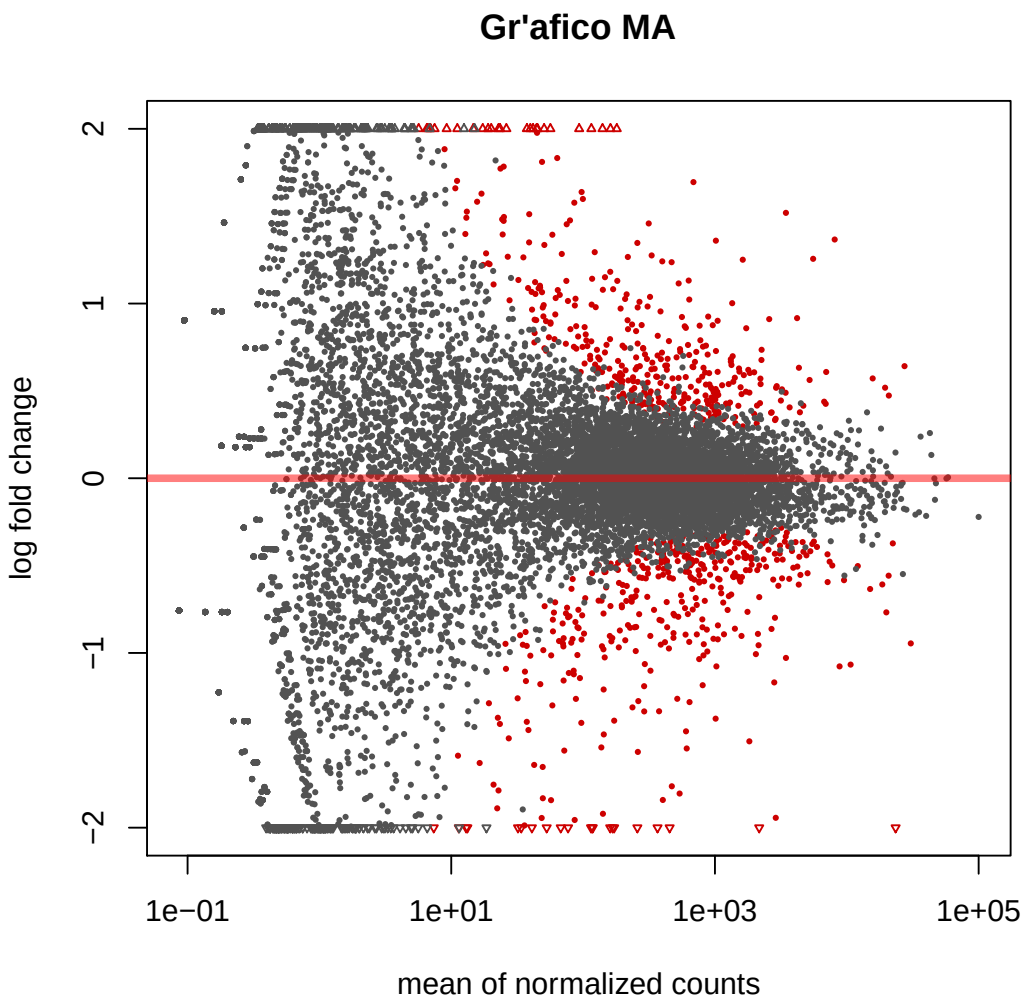
	stat	pvalue	padj
	<numeric>	<numeric>	<numeric>
FBgn0039155	-21.7860145428035	3.14893542201204e-105	2.58307172667648e-101
FBgn0029167	-20.0353220728363	2.71080044420307e-89	1.11183480218989e-85
FBgn0035085	-15.9440879603834	3.13178980573261e-57	8.56335725880821e-54
FBgn0029896	-13.6300692972973	2.65328245257746e-42	5.44121898962322e-39
FBgn0034736	-12.9018000204293	4.39704860471644e-38	7.21379794089778e-35
FBgn0040091	-11.8334941229432	2.62003589995331e-32	3.58202574788617e-29

Las interpretación de las columnas de la tabla de resultados obtenida es la siguiente:

- baseMean: media de los conteos normalizados.
- lo2FoldChange: un valor nulo indica que el gen en cuestión presenta los mismos niveles de expresión para ambos grupos. Un valor no nulo indica que el gen se expresa de forma diferente en cada grupo y, según el signo de dicho valor, se ve si presenta un mayor nivel de expresión en el grupo de individuos no tratados o en el grupo de los tratados.
- p-value: p-valor para el contraste descrito
- padj: p-valor ajustado. Al realizar el constraste para cada uno de los genes podemos encontrarnos con el problema de comparaciones múltiples, es decir, al realizar un gran número de contrastes, se produce un aumento de la probabilidad de obtener falsos positivos y cometer un error de tipo I (rechazar la hipótesis nula, siendo verdadera). El programa ajusta los p-valores por el procedimiento de Benjamin-Hochberg.

Nos interesa conocer todos los genes diferencialmente expresados, es decir, aquellos genes para los que se ha rechazado la hipótesis nula para un nivel de significación $\alpha = 0,01$.

En la siguiente gráfica vemos los valores de *log2FoldChange* sobre la media de conteos. Los puntos rojos pertenecen a aquellos genes cuyo constraste ha rechazado la hipótesis nula (genes diferencialmente expresados). En este caso hay 6675 genes de este tipo.



Hasta ahora solo hemos usado el factor condición, pero también podemos crear el modelo teniendo en cuenta el tipo de secuenciaciación. El modelo que obtenemos es:

```
log2 fold change (MLE): condition treated vs untreated
Wald test p-value: condition treated vs untreated
DataFrame with 6 rows and 6 columns
                    baseMean      log2FoldChange             lfcSE
                   <numeric>           <numeric>         <numeric>
FBgn0000003 0.159468650583408   0.687317621016542     3.87108289314
FBgn0000008  52.2256775955264  0.0124835813497824 0.312794873267806
FBgn0000014 0.389708020282597   0.725641096785045  3.43795226129079
FBgn0000015 0.905358372171504  -0.659586898950503  2.12479201701311
FBgn0000017  2358.24340775372  -0.274660627261139 0.117714619266187
FBgn0000018  221.241556161604 -0.0688574674151578 0.161511724341256
                        stat             pvalue              padj
                   <numeric>          <numeric>         <numeric>
FBgn0000003  0.177551770393382  0.859075004234684                NA
FBgn0000008 0.0399098016516859  0.968165036742223 0.992489343047921
```

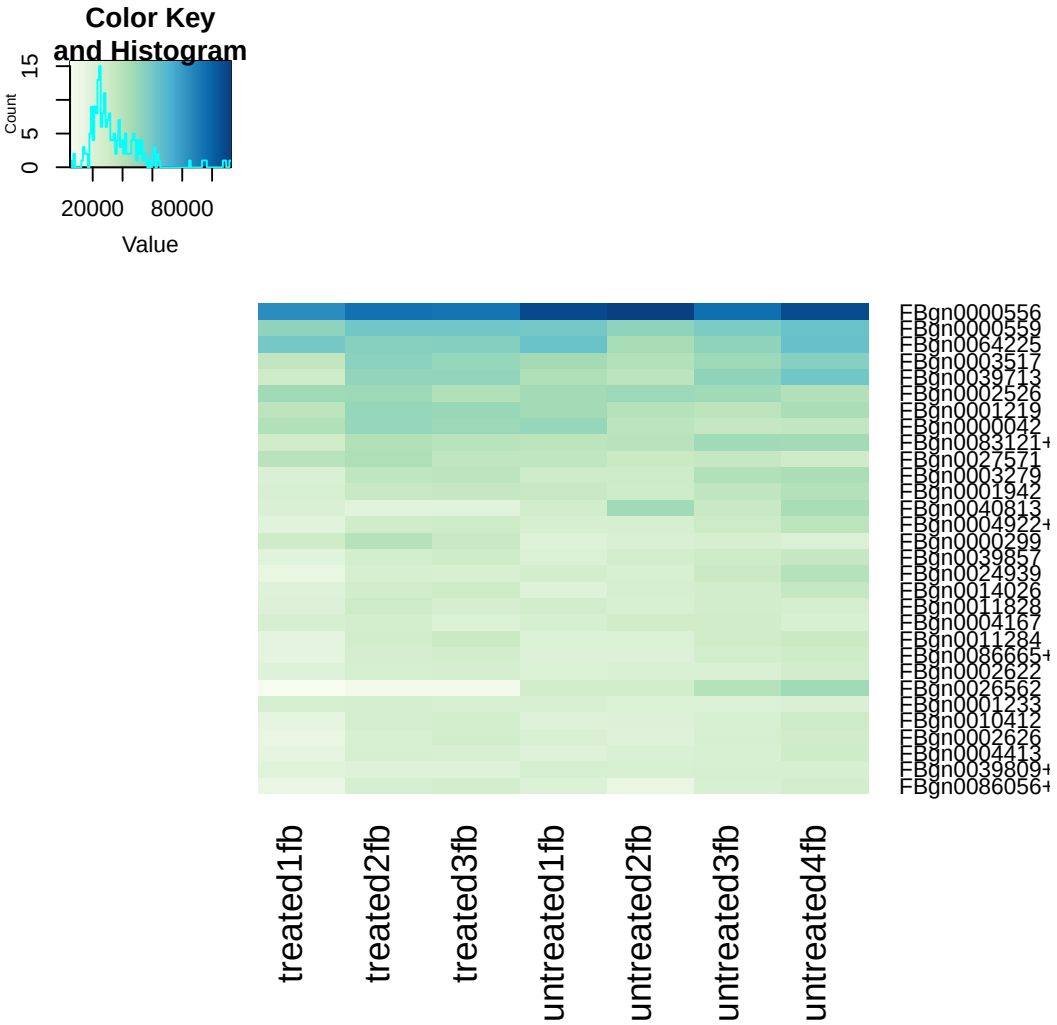
```
FBgn0000014   0.211067822248527   0.832834348712066                  NA
FBgn0000015  -0.310424217367734   0.756238380329955                  NA
FBgn0000017   -2.33327541620002 0.0196336948558276 0.138952871310778
FBgn0000018   -0.42633107717722   0.669866615819642 0.908520986767696
```

4. Evaluación de la calidad de los datos

Queremos encontrar genes diferencialmente expresados y tenemos que ver muestras cuyo tratamiento experimental ha sufrido alguna anomalía que puede transformar los datos obtenidos en perjudiciales. Para ver esto realizamos un control de calidad.

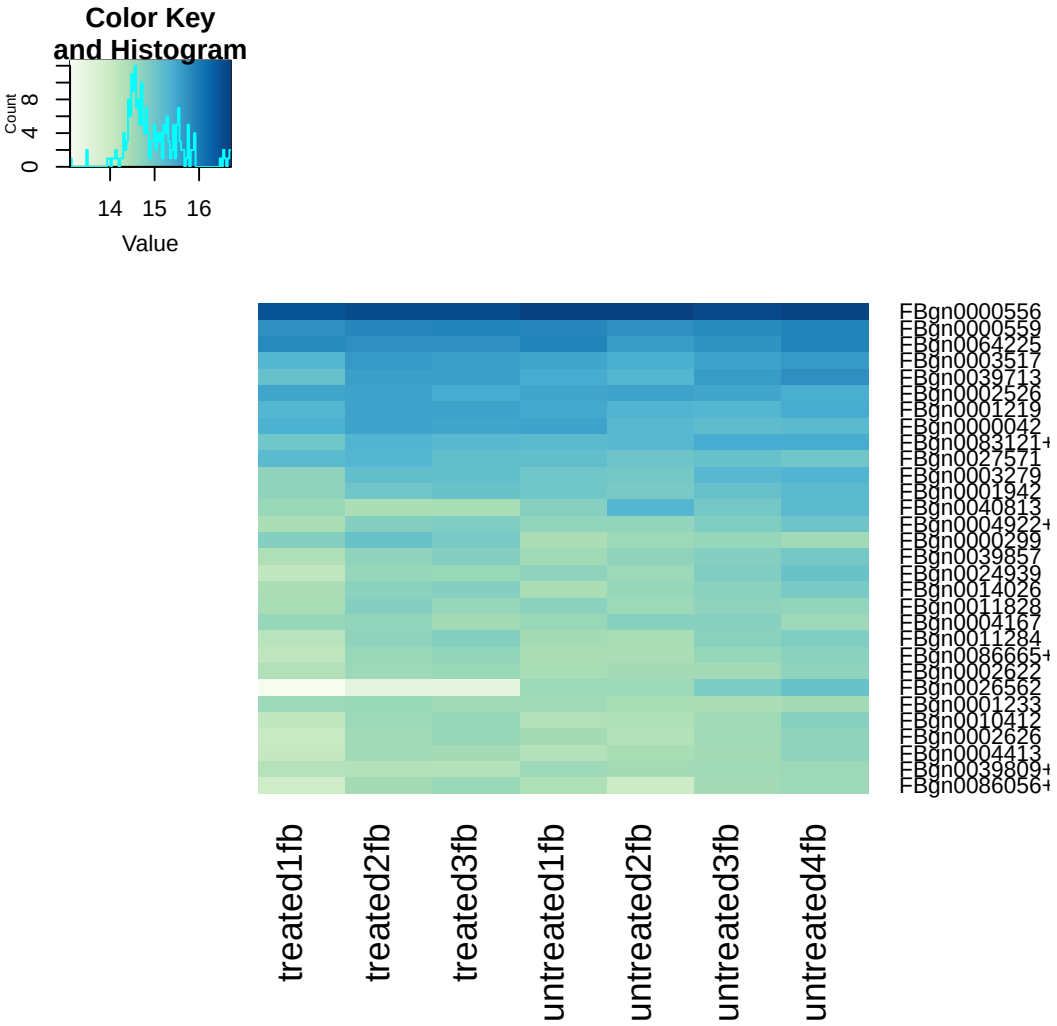
4.1. Mapa de calor de la tabla de conteos

Se puede crear un mapa de calor a partir de la tabla de conteos, los que nos proporciona información sobre los genes que muestran un mayor nivel de expresión en el experimento. El mapa de calor de los conteos brutos es:



Vemos los conteos de los 30 genes con mayor nivel de expresión. Las filas representan a los genes y las columnas a los individuos. A mayor valor de conteo, más intensidad de color.

En el análisis diferencial trabajamos con los conteos brutos y usamos distribuciones discretas, pero para análisis posteriores puede ser útil trabajar con una transformación de los datos. El mapa de calor para los datos transformados mediante la función *rlogTransformation* es:



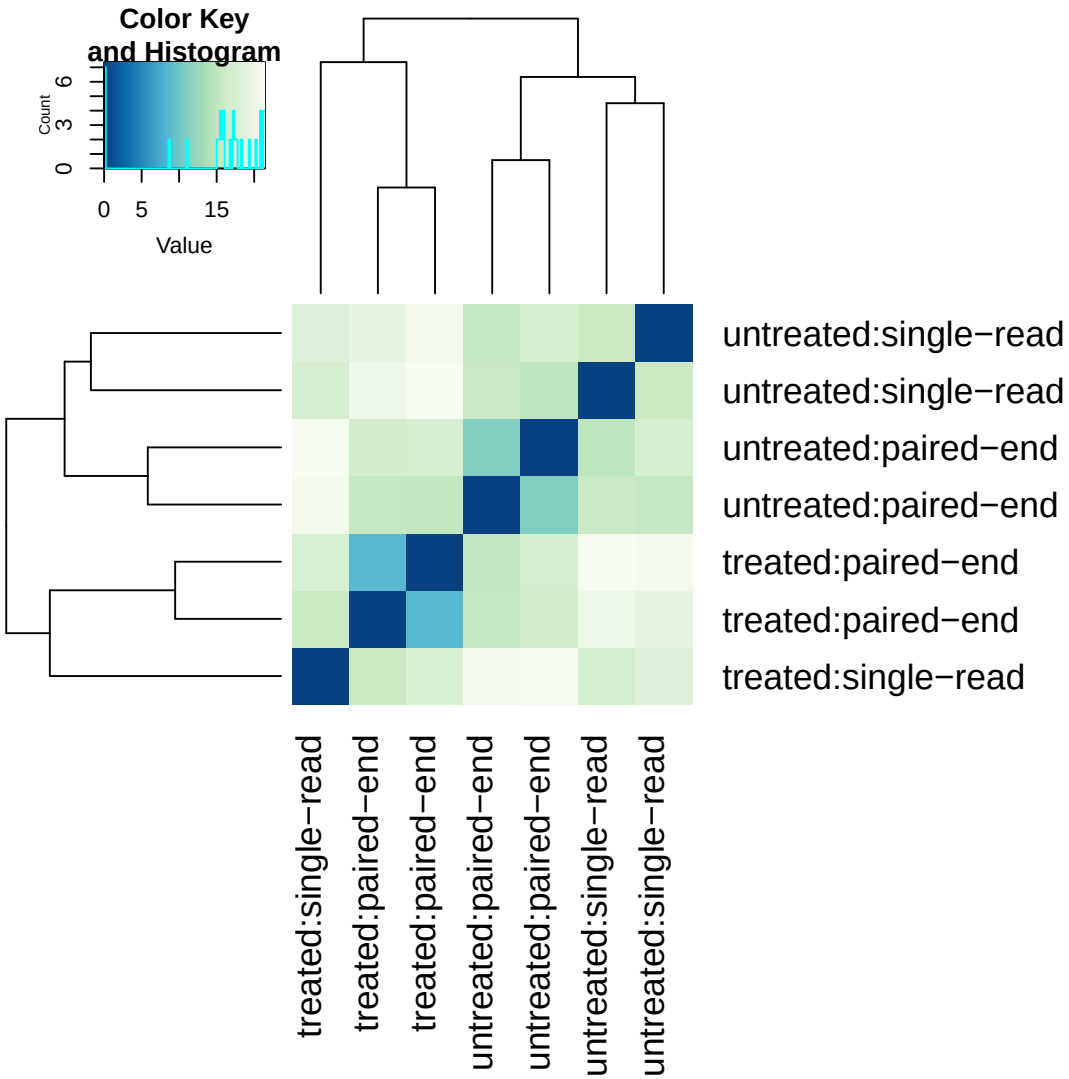
4.2. Mapa de calor de las distancias entre muestras

Podemos realizar un análisis de conglomerados (cluster) sobre los datos transformados. Para ello tenemos que aplicar la función *dist* a la traspuesta de la matriz de conteos transformados para obtener las distancias euclídeas entre las muestras. La matriz de distancias resultante es:

	treated1fb	treated2fb	treated3fb	untreated1fb	untreated2fb
treated1fb	0.00000	16.065480	17.784211	18.24555	17.30651
treated2fb	16.06548	0.000000	8.731307	19.33040	20.18860
treated3fb	17.78421	8.731307	0.000000	20.81854	21.33274
untreated1fb	18.24555	19.330396	20.818541	0.00000	15.88771

```
untreated2fb   17.30651  20.188596  21.332739     15.88771      0.00000
untreated3fb   21.43401  16.774194  17.160940     17.50276     15.04006
untreated4fb   20.94888  15.602292  15.406478     15.58571     15.79327
             untreated3fb untreated4fb
treated1fb       21.43401     20.94888
treated2fb       16.77419     15.60229
treated3fb       17.16094     15.40648
untreated1fb     17.50276     15.58571
untreated2fb     15.04006     15.79327
untreated3fb      0.00000     11.02415
untreated4fb     11.02415      0.00000
```

A partir de la matriz de distancias podemos crear el mapa de calor. Los resultados obtenidos en dicho mapa deben coincidir con los que observamos en la matriz de distancias: los individuos más alejados en la matriz de distancias deben ser los más alejados en el mapa de calor. Los mismo ocurre con los individuos más cercanos. El mapa de calor resultante es:



Los colores más fuertes indican mayor similitud entre los datos. Podemos observar también los valores de las ditancias y cuántos pares de muestras presentan ese valor.

4.3. Componentes principales de las muestras

Estudiar las componentes principales de las muestras sirve para ver el efecto toal de las covariables, así como la posible existencia de efectos lotes. En la siguiente gráfica se representan la primera y segunda componente principal de nuestros datos: