JOSÉ FRANCISCO BARREIRO GONZÁLEZ

ANÁLISIS DE DATOS ÓMICOS

UOC ABRIL 2020

PEC 1

ANÁLISIS DE MICROARRAYS

https://github.com/jbarreiro20uoc/barreiro_uoc.git

Abstract.

El estudio que usaré en esta PEC está relacionado con el medio ambiente por mi formación como ambientalista, en el cual se investiga la relación entre una baja concentración de CO_2 y la fotorrespiración en plantas del género Arabidopsis thaliana, ya que la acumulación de especies reactivas del oxígeno (ROS) a nivel intracelular inducen cambios en el transcriptoma, provocando estrés celular y muerte celular.

Objetivos.

La fotorrespiración es un proceso complejo que tiene lugar en el mesófilo de las hojas de las plantas, cuando hay luz y una elevada concentración de O_2. Se trata de un mecanismo de protección de la planta frente a condiciones adversas (sequía, elevadas intensidades de luz, etc.), en el cual se derrocha mucha energía. El objetivo del estudio es analizar los cambios transcripcionales provocados por la acumulación de ROS y la producción de estrés oxidativo, que pueden provocar muerte celular.

Materiales y métodos.

Materiales

He utilizado 6 archivos CEL, con 3 réplicas, con lo datos de la exposición de las plantas a baja concentración de CO_2 (100 ppm) y a concentración normal (400 ppm), siendo estas últimas el grupo de control. EL tipo de array utilizado es [ATH1-121501] Affymetrix Arabidopsis ATH1 Genome Array.

El estudio consiste en un perfilado de expresiones por array.

Métodos

El procedimiento realizado, “pipeline”, consiste en:

1. Identificar los grupos muestrales.
2. Exploración de los datos.
3. Control de calidad de los datos crudos.
4. Normalización de datos
5. Filtraje no específico
6. Identificación de genes diferencialmente expresados
7. Anotación de resultados
8. Comparación entre réplicas
9. Análisis de significación biológica

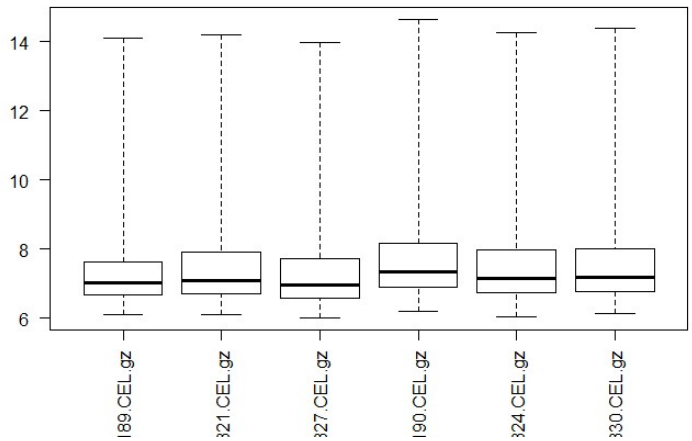
1. Los grupos muestrales son 6, tres archivos con datos de plantas expuestas a bajas concentraciones de CO_2 y tres archivos con datos de control.

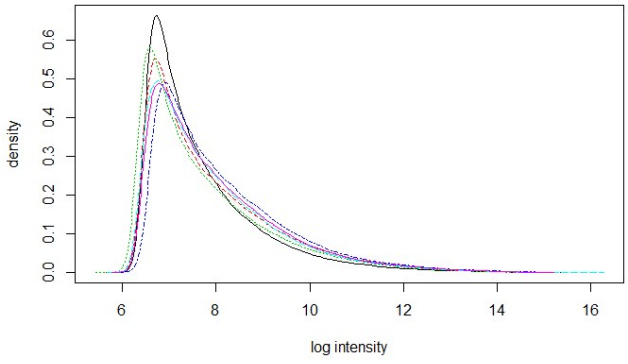
Las plantas son ejemplares de Arabidopsis thaliana de 6 semanas, expuestas durante 24 horas a bajas concentraciones de CO_2.

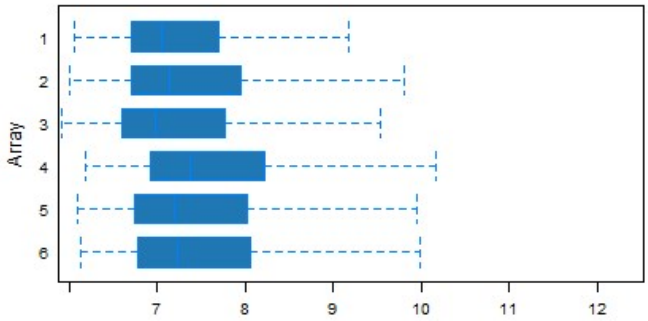
Arabidopsis thaliana fue la primera planta cuyo genoma se secuenció por completo, una tarea completada en diciembre del 2000.

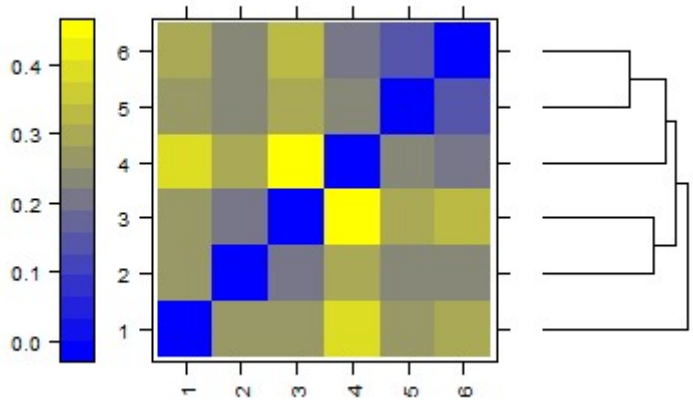
---

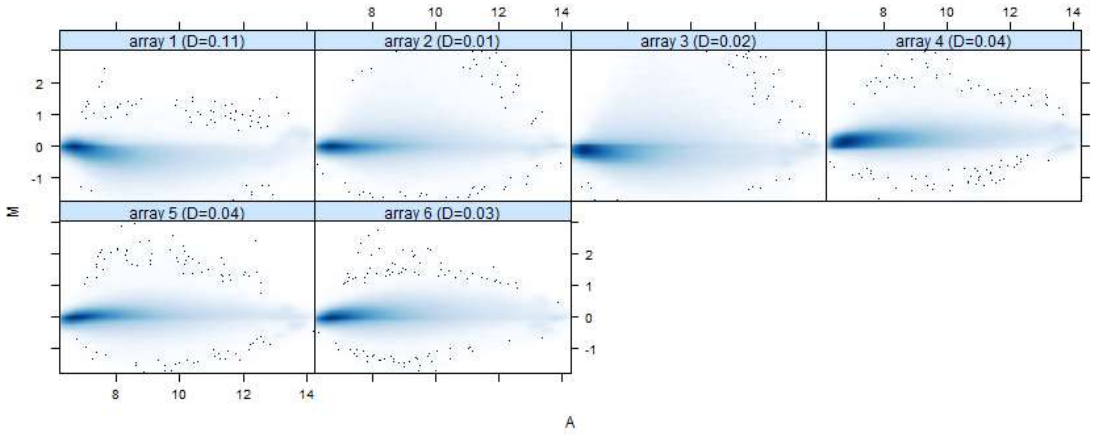
2. Exploración de los datos. Lo hacemos visualizando unos gráficos de los datos. He utilizado el paquete arrayQualityMetrics, que realiza un número elevado de gráficos, de los cuales he elegido los más representativos para simplificar el estudio

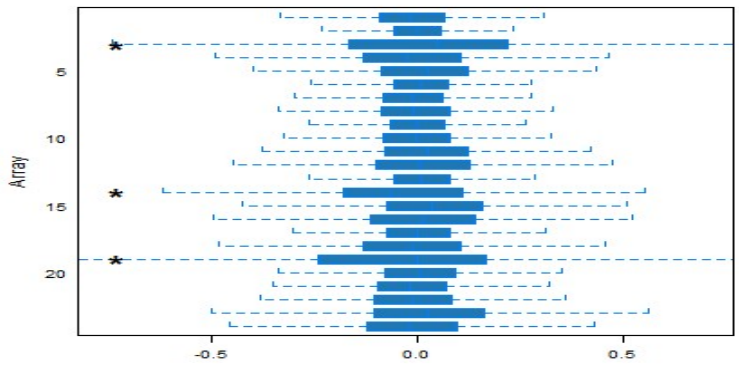





Array
1
2
3
4
5
6
7
8
9
10
11
12

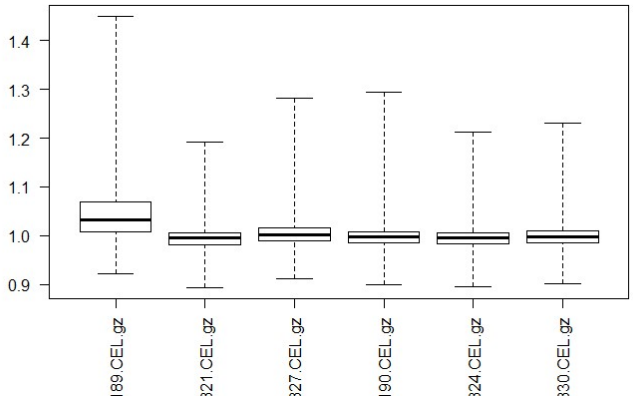

0.4
0.3
0.2
0.1
0.0
6
5
4
3
2
1
1
2
3
4
5
6


array 1 (D=0.11)
array 2 (D=0.01)
array 3 (D=0.02)
array 4 (D=0.04)
array 5 (D=0.04)
array 6 (D=0.03)
M
A
8
10
12
14
2
1
0
-1

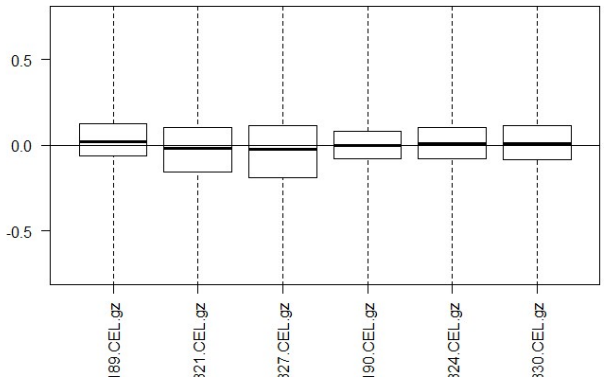

Array
5
10
15
20
-0.5
0.0
0.5

3. Control de calidad de los datos crudos. Vemos unos gráficos para comprobar si hay algún array problemático. He utilizado el paquete affyPLM.

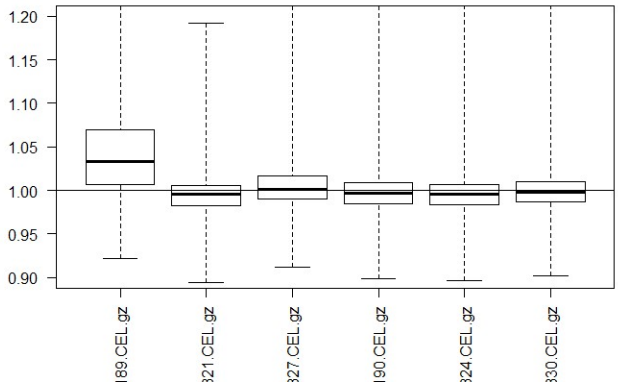
   Vemos que el primer array presenta algún problema



Expresiones relativas



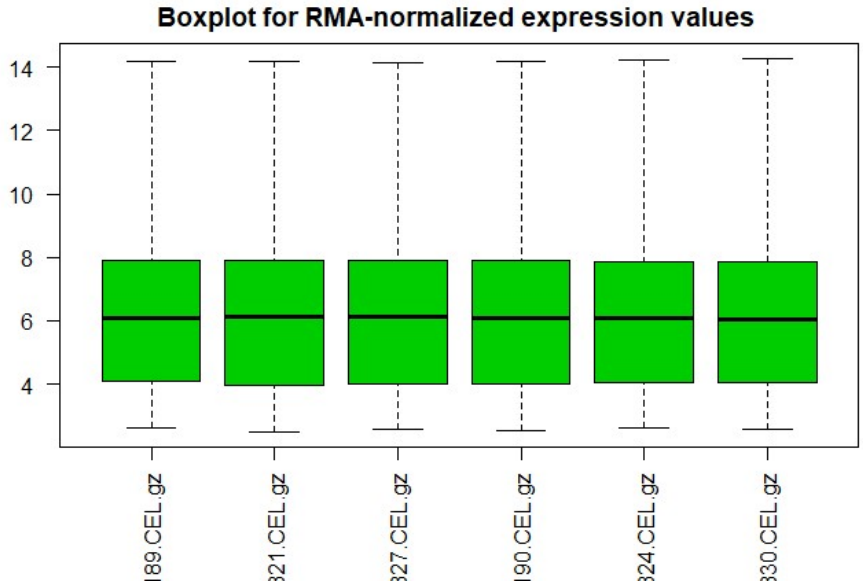
Errores estandarizados



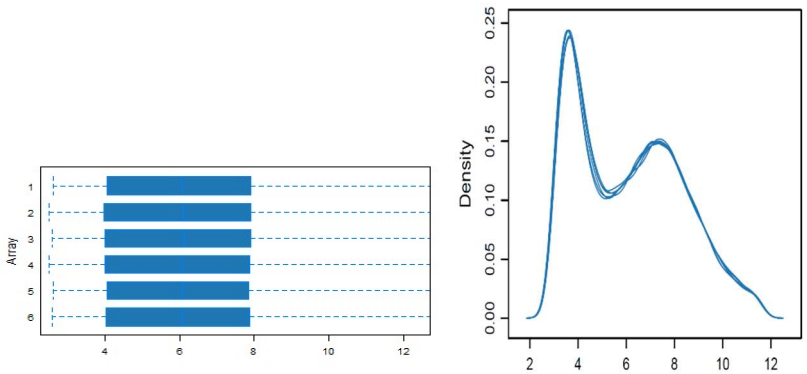
En el último gráfico, de errores estandarizados (NUSE), observamos una pequeña desviación en el primer array, pero no se descarta porque estos gráficos son exploratorios.

4. Normalización de los datos. He utilizado el método RMA, que sigue estos pasos:
    - Ajusta el ruido de fondo (background)
    - Toma logaritmos en base 2 de cada intensidad
    - Normalización por cuantiles del paso 2
    - Realiza una *median polish* para estimar las intensidades de cada gen.

Realizamos un gráfico para comprobar el ajuste, es decir, que los valores de los arrays sean comparables entre sí.



Podemos hacer un control de calidad de los datos normalizados con la función arrayQualityMetrics.



5. Filtraje no específico. Con este método filtramos los genes que varían poco entre condiciones. He utilizado la función nsFilter.
   Hemos de buscar las anotaciones correspondientes al microarray utilizado en la página de bioconductor. Obtenemos los genes filtrados.

```
$numDupsRemoved
[1] 174

$numRemoved.ACCNUM
[1] 2475

$numLowVar
[1] 10075

$feature.exclude
[1] 12
```

6. Identificación de genes diferencialmente expresados.

El primer paso consiste en crear la matriz de diseño y la de contrastes.

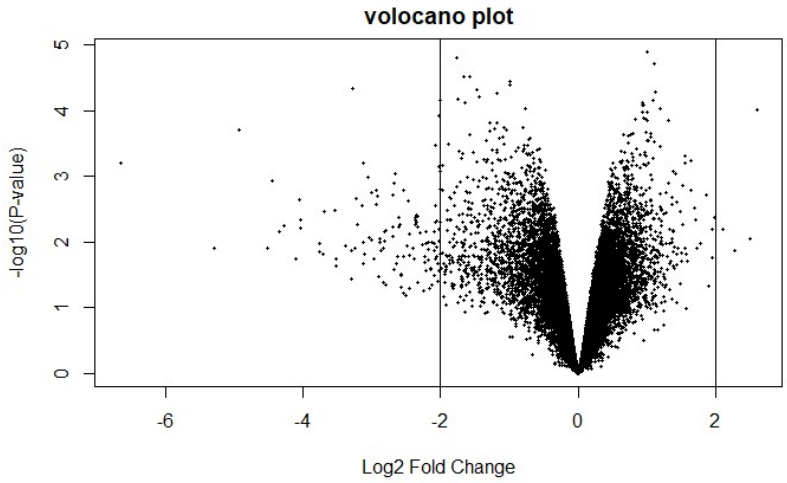
```
                           low1 low2 low3 amb1 amb2 amb3
GSM2113630_hyb6189.CEL.gz    1    0    0    0    0    0
GSM2113631_hyb5821.CEL.gz    0    1    0    0    0    0
GSM2113632_hyb5827.CEL.gz    0    0    1    0    0    0
GSM2113633_hyb6190.CEL.gz    0    0    0    1    0    0
GSM2113634_hyb5824.CEL.gz    0    0    0    0    1    0
GSM2113635_hyb5830.CEL.gz    0    0    0    0    0    1
attr(,"assign")
[1] 1 1 1 1 1 1
attr(,"contrasts")
attr(,"contrasts")$lev
[1] "contr.treatment"
```

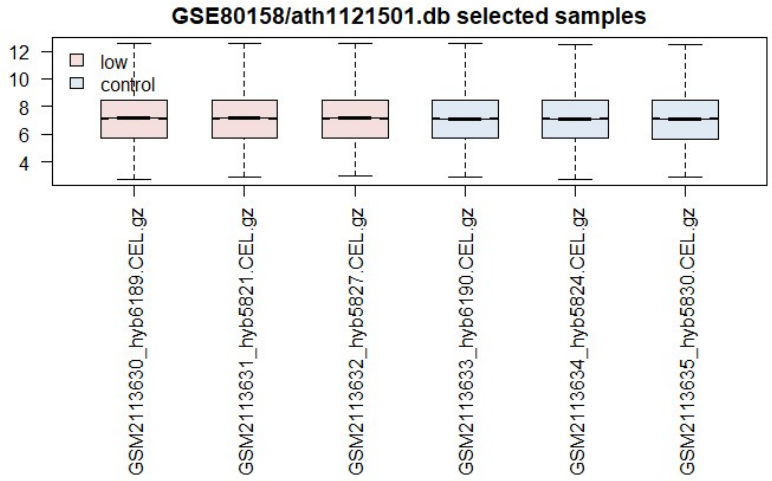
```
      Contrasts
Levels dif_1 dif_2 dif_3
  low1     1     0     0
  low2     0     1     0
  low3     0     0     1
  amb1    -1     0     0
  amb2     0    -1     0
  amb3     0     0    -1
```

En el paquete limma se incluyen las funciones necesarias para determinar los genes más diferencialmente expresados.

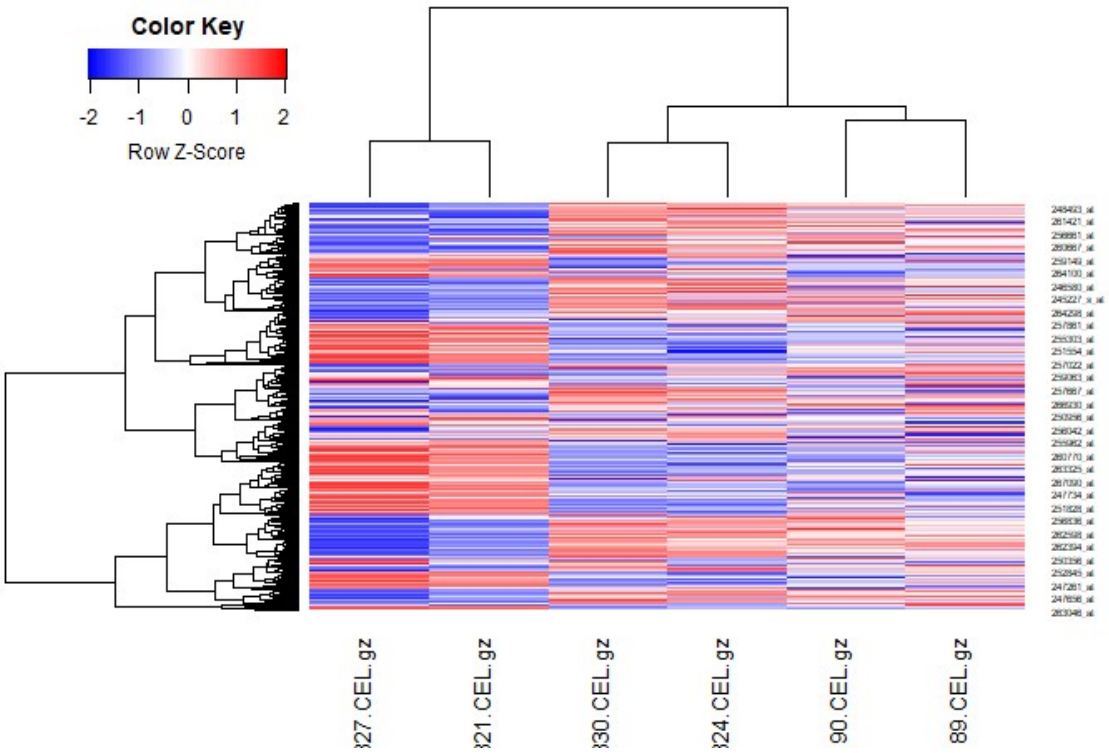
Con la función topTable generamos una lista de genes ordenados de más a menos diferencialmente ordenados.

Con un gráfico tipo volcanoplot podemos ver los resultados





A continuación, vemos el Heatmap



7. Anotación de resultados.

La siguiente relación es la cabecera(head) de los genes diferencialmente expresados

	logFC	AveExpr	t	P.Value	adj.P.Val	B
248434_at	-3.259816	5.564310	-10.292258	1.702830e-06	0.01715431	5.165705
251791_at	2.391489	5.846305	9.350228	3.962021e-06	0.01742832	4.508043
249983_at	-1.784747	4.746968	-9.064545	5.190091e-06	0.01742832	4.291168
260522_x_at	-4.734354	8.077787	-8.378484	1.021922e-05	0.02227000	3.733809
255341_at	-2.031420	5.340363	-8.301776	1.105321e-05	0.02227000	3.668109
247095_at	-1.998059	8.479458	-7.762135	1.951173e-05	0.02834783	3.185413

Con el paquete Bioconductor podemos extraer un paquete de anotaciones de los genes asociados a su identificador en la base de datos.

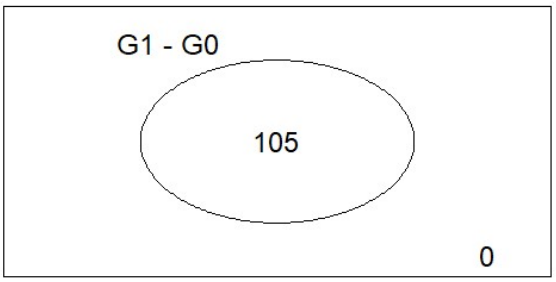
	probe_id	symbol
1	261585_at	ANAC001
2	261585_at	NAC001
3	261568_at	NGA3
4	261584_at	ASU1
5	261584_at	ATDCL1
6	261584_at	CAF
7	261584_at	DCL1
8	261584_at	EMB60
9	261584_at	EMB76

8. Análisis de significación biológica.

Con la función decideTest vemos los genes que cambian simultáneamente en más de una comparación.

```
TestResults matrix
           Contrasts

       Low ambiental
Down              83
NotSig          9969
Up                22
```



Buscamos las anotaciones Gene Ontology (GO) para asociar los genes a sus funciones:

```
261585_at ANAC001 GO:0006888 ER to Golgi vesicle-mediated t
261585_at ANAC001 GO:0007275 multicellular organism develop
261585_at ANAC001 GO:0043090 amino acid import
261585_at ANAC001 GO:0005634 nucleus
261585_at ANAC001 GO:0003700 DNA-binding transcription fact
```