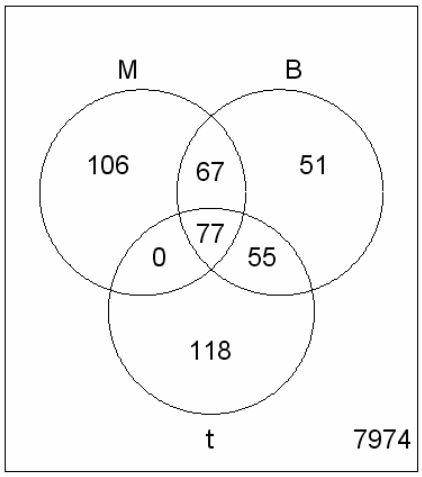
11.6 Gráficos. Comparación de criterios de ordenamientos de genes

Los criterios habituales para ordenar los genes respecto de la evidencia que presentan de estar diferencialmente expresados entre 2 tejidos son M, t, B (Speed 2003 pág 55-61). Utilizaremos los datos de swirl para compararlos.

```
targets <- readTargets("SwirlSample.txt")
RG <- read.maimages(targets$Names, source="spot")
RG$genes <- readGAL("fish.gal")
RG$printer <- getLayout(RG$genes)
RGrma <- backgroundCorrect(RG, method="rma")
MA <- normalizeWithinArrays(RGrma)
fit1 <- lmFit(MA, design=c(-1,1,-1,1))
fit <- eBayes(fit1)
```

Definimos tres grupos de genes:

- grupo 1: con los 250 valores mayores de |M|
- grupo 2: con los 250 valores mayores de B
- grupo 3: con los 250 valores mayores de |t|

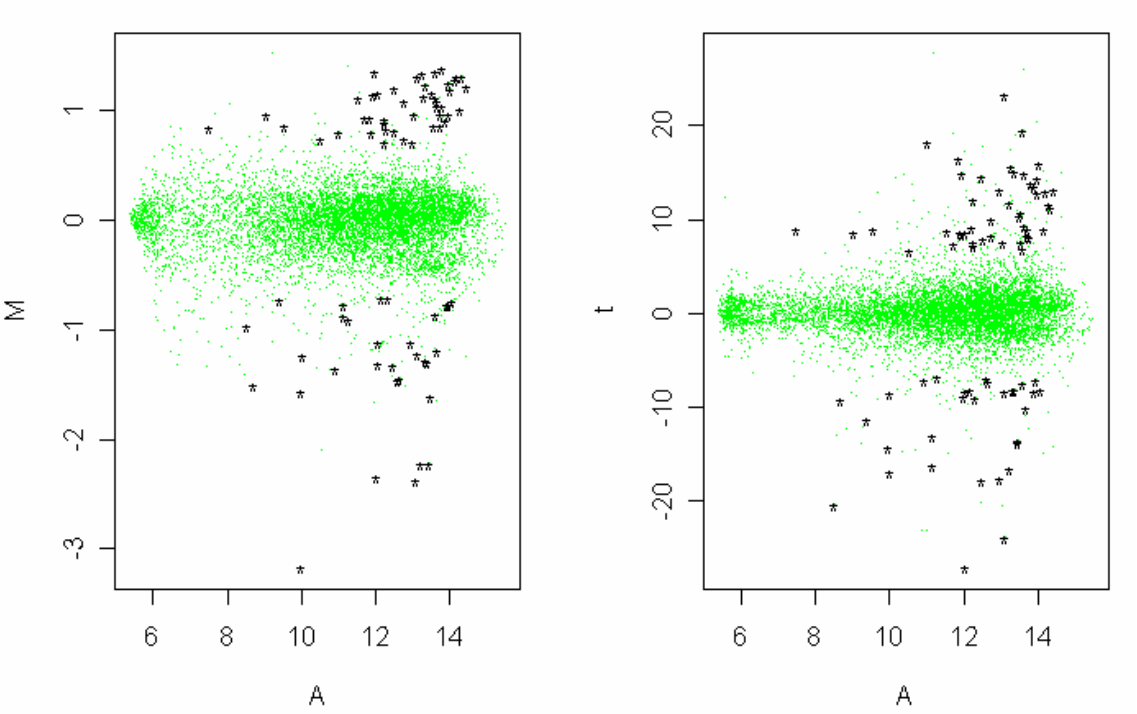


El Diagrama de Venn de los tres grupos fue obtenido mediante las siguientes instrucciones

```
ordinary.t <- fit$coef / fit$stdev.unscaled / fit$sigma

ordM <- order(abs(fit$coefficients),decreasing=TRUE)
ordB <- order(fit$lods,decreasing=TRUE)
ordt <- order(abs(ordinary.t),decreasing=TRUE)

vector <- rep(0,8448)

top250M <- ordM[1:250]
top250B <- ordB[1:250]
top250t <- ordt[1:250]

M <- vector
B <- vector
t <- vector

B[top250B]<-1
```

```
t[top250t]<-1
M[top250M]<-1
interseccion <- (B*t*M)/(B*t*M)

results <- cbind(M,B,t)
a <- vennCounts(results)
print(a)
vennDiagram(a)
```

Los 77 spots que tienen simultáneamente valores grandes de |M|, |t| y B son los principales candidatos a estar expresados diferencialmente. Se destaca la ausencia de puntos en la intersección de M y t con el complemento de B, es decir que los spots que tienen valores grandes de M y t necesariamente están en B. Esto suele ocurrir.



Los gráficos anteriores fueron obtenidos mediante las siguientes instrucciones

```
M <- fit$coefficients
A <- fit$Amean
par(mfrow=c(1,2))
plot(A,M,pch='.', col="green")
points(A*interseccion,M*interseccion, pch="*")
plot(A, ordinary.t, ylab="t",pch=".",col="green")
points(A*interseccion,ordinary.t*interseccion , pch="*")
```

El primero de los gráficos es similar al MA plot que ya conocíamos con la diferencia que ahora tenemos un sólo gráfico para todos los arreglos juntos. M es el promedio de todas las M_i = log (R/G) de cada uno de arreglos, en el que se ha tenido en cuenta los dye-swaps.

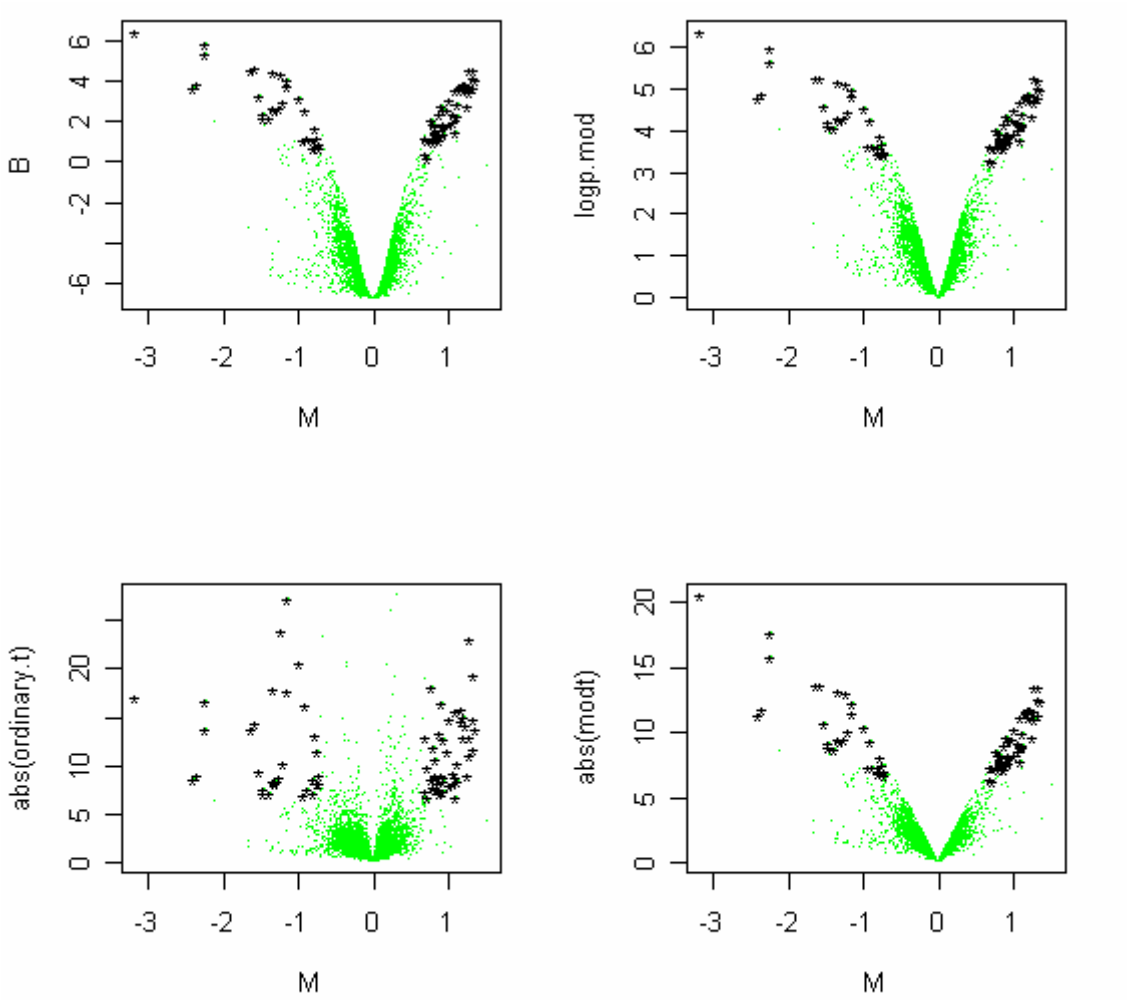
Los puntos correspondientes a la intersección de los tres grupos aparecen con un asterisco (*). Claramente se ve que son candidatos a estar expresados diferencialmente.

Observe el acuerdo que en general se observa en los dos gráficos respecto a los puntos destacados con (*).

Es claro que algunos puntos están bien separados de la nube. Esto suele ocurrir. Los genes correspondientes a estos puntos tienen mayores posibilidades de estar expresados diferencialmente. Speed (2003) recomienda este enfoque informal de identificación de tales genes. En muchos casos esta evidencia es tan sólida como cualquier otra que se pueda obtener.

Ejercicio: Obtener gráficos de dispersión como los dos anteriores en los que se distingan los puntos por grupo del Diagrama de Venn.

Volcano Plots



Nuevamente destacamos con (*) los puntos de la intersección de los tres criterios. Los gráficos fueron obtenidos mediante las siguientes instrucciones

```
par(mfrow=c(2,2))

B <- fit$lods
plot(M,B,pch='.',col="green")
points(M*interseccion,B*interseccion, pch="*")

logp.mod <- -log10(fit$p.value)
```

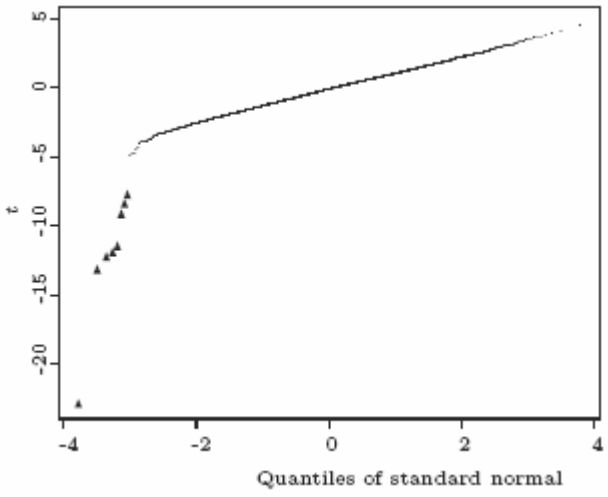
```
plot(M,logp.mod,pch='.',col="green")
points(M*interseccion,logp.mod*interseccion, pch="*")

plot(M,abs(ordinary.t),pch='.',col="green")
points(M*interseccion,abs(ordinary.t)*interseccion, pch="*")

modt <- fit$t
plot(M,abs(modt),pch='.',col="green")
points(M*interseccion,abs(modt)*interseccion, pch="*")
```

Gráficos Cuantil-Cuantil - Q-Q plots

Otro enfoque gráfico para distinguir potenciales genes expresados diferencialmente es examinar Q-Q plots para determinados estadísticos. Estos gráficos son útiles para presentar los miles de $\overline{M}$ o estadísticos t que interesa estudiar en un experimento típico de microarreglos. La distribución Normal Estándar es la referencia natural, y cuantas más réplicas entren en los promedios o en un estadístico tipo t, tanto más esperamos que se parezcan a una muestra de una distribución Normal. Con sólo cuatro arreglos, como en el experimento de Swirl, para los cuales tenemos ocho log(intensidad) que entran en nuestros promedios no podemos esperar y tampoco obtenemos una recta contra la distribución Normal. Sin embargo el gráfico puede ser valioso para indicar en qué medida los estadísticos t extremos difieren de la mayoría.

En general, los Q-Q plots se utilizan para validar si los datos provienen de una distribución particular ó si dos conjuntos de datos tienen la misma distribución. Sin embargo en este contexto se utilizan como ayuda visual de identificación de genes con estadísticos inusuales. Proveen una manera informal de corrección de comparaciones múltiples y los puntos que se desvían en forma notoria de la relación lineal son candidatos a corresponder a genes cuyos niveles de expresión difieren entre los dos grupos.

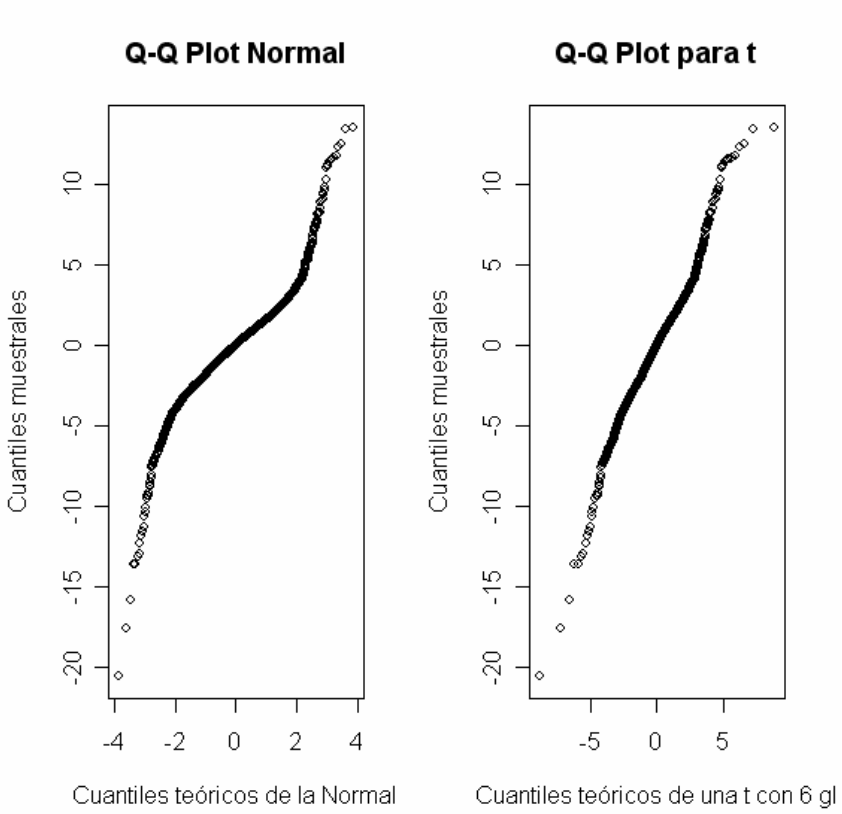
El siguiente ejemplo corresponde al experimento ApoAI que vimos en la práctica 8. El Q-Q plot (Dudoit et al., 2002) muestra los valores del t= *estadístico t- deWelch* (test para diferencia de medias basado en muestras independientes) en función de los cuantiles teóricos de la Normal.



Se destacan ocho genes (triángulos llenos) cuyos estadísticos *t* se desvían en forma marcada de la relación lineal que presentan los demás. Los ocho genes tienen

estadísticos *t* negativos sugiriendo una regulación hacia abajo (down-regulation) de dichos genes en los ratones knock-out comparados con los controles.

Como en el ejemplo anterior, algunas veces es posible determinar donde terminan los outliers y comienza la masa de estadísticos desafortunadamente este no es el caso de nuestro ejemplo de Swirl.



También comparamos con los cuantiles de la distribución t con 6 grados de libertad.

```
par(mfrow=c(1,2))
qqnorm(modt,main="",xlab="Cuantiles teóricos de la Normal",
ylab="Cuantiles muestrales")
title("Q-Q Plot Normal")
#A <- qqplot(modt,qt(ppoints(modt),6))
plot(A[[2]],A[[1]], xlab="Cuantiles teóricos de una t con 6
gl",ylab="Cuantiles muestrales")
title("Q-Q Plot para t")
```

12. Algunas consideraciones sobre diseño de experimentos de microarreglos de dos canales

12.1 Matriz diseño en el modelo lineal

En el párrafo anterior y en el práctico 8 hemos utilizado modelos lineales para obtener los log ratios estimados de las intensidades de expresión de los genes entre dos grupos. Los modelos lineales permiten, a través de la especificación de su matriz de diseño, unificar el tratamiento de los tests de dos grupos, comparación directa (una muestra, Sección 11.3), comparación indirecta (dos muestras independientes, Sección 11.4) y diseños más complejos. Hemos, en todos los ejemplos del práctico 8, construido la

matriz de diseño a mano. Veremos a continuación una forma de obtener la matriz de diseño utilizando la función **`modelMatrix`**

Ejemplo Swirl:

```
> targets
         Names slide.number experiment.Cy3 experiment.Cy5
1 swirl.1.spot           81          swirl      wild type
2 swirl.2.spot           82      wild type          swirl
3 swirl.3.spot           93          swirl      wild type
4 swirl.4.spot           94      wild type          swirl
       date comments
1 2001/9/20       NA
2 2001/9/20       NA
3 2001/11/8       NA
4 2001/11/8       NA

> diseño <-modelMatrix(targets, ref="wt")
Error in modelMatrix(targets, ref = "wt") :
        targets should contain columns: Cy3 and Cy5

> names(targets)
[1] "Names"          "slide.number"   "experiment.Cy3"
[4] "experiment.Cy5" "date"           "comments"

> names(targets)[3]<-"Cy3"
> names(targets)[4]<-"Cy5"
> diseño <-modelMatrix(targets, ref="wild type")
Found unique target names:
 swirl wild type
> diseño
  swirl
1    -1
2     1
3    -1
4     1
```

Ejemplo beta 7. Cambio en R el directorio de trabajo a donde están los datos de beta 7

```
>  muestras <- readTargets("TargetBeta7.txt")
>
> diseñobeta7 <-modelMatrix(targets, ref="b7 -")
Found unique target names:
 swirl wild type
Error in modelMatrix(targets, ref = "b7 -") :
        "b7 -" not among the target names found
> diseñobeta7
Error: object "diseñobeta7" not found
> diseñobeta7 <-modelMatrix(muestras, ref="b7 -")
Found unique target names:
 b7 - b7 +
>   diseñobeta7
          b7 +
6Hs.195.1    1
6Hs.168     -1
6Hs.166     -1
```

```
6Hs.187.1    1
6Hs.194      1
6Hs.243.1   -1
>
```

Ejemplo ApoAI

```
>   load("ApoAI.RData")
> summary(RG)
        Length Class      Mode
R       102144 -none-     numeric
G       102144 -none-     numeric
Rb      102144 -none-     numeric
Gb      102144 -none-     numeric
printer      4 -none-     list
genes        8 data.frame list
targets      3 data.frame list
>
> muestras2  <- RG$targets
> diseñoApo <- modelMatrix(muestras2, ref="Pool")
Found unique target names:
 ApoAI-/- C57BL/6 Pool
> diseñoApo
   ApoAI-/- C57BL/6
c1        0       1
c2        0       1
c3        0       1
c4        0       1
c5        0       1
c6        0       1
c7        0       1
c8        0       1
k1        1       0
k2        1       0
k3        1       0
k4        1       0
k5        1       0
k6        1       0
k7        1       0
k8        1       0
```

Esta matriz de diseño corresponde al siguiente sistema de ecuaciones lineales:

$$E(\log(\frac{\text{Wt}}{\text{Ref}})) = \beta_1 \qquad \text{para los arreglos } 1 \text{ a } 8$$

$$E(\log(\frac{\text{KO}}{\text{Ref}})) = \beta_2 \qquad \text{para los arreglos } 9 \text{ a } 16$$

Con esta *parametrización*, llamada *media de grupos* -group means, el contraste de interés:

el log ratio de la intensidad del gen entre el RNA del KO (**`ApoAI-/-`**) y el RNA del Wt (**`C57BL/6`**)

está dado por la diferencia de los coeficientes: $\beta_2 - \beta_1$ y debe especificarse al realizar el análisis.

```
 >  MA <- normalizeWithinArrays(RG)
>   fit  <- lmFit(MA,design=diseñoApo)
>   fit2 <- contrasts.fit(fit, c(-1,1))
>   fit3 <- eBayes(fit2)
>   topTable(fit3)
     GridROW GridCOL ROW COL                 NAME TYPE
2149       2       2   8   7      ApoAI,lipid-Img cDNA
540        1       2   7  15 EST,HighlysimilartoA cDNA
5356       4       2   9   1 CATECHOLO-METHYLTRAN cDNA
4139       3       3   8   2 EST,WeaklysimilartoC cDNA
1739       2       1   7  17    ApoCIII,lipid-Img cDNA
2537       2       3   7  17 ESTs,Highlysimilarto cDNA
1496       1       4  15   5                 est cDNA
4941       4       1   8   6 similartoyeaststerol cDNA
947        1       3   8   2 EST,WeaklysimilartoF cDNA
5604       4       3   1  18                     cDNA

        CLID          ACC         M        A         t
2149 1077520              3.1661645 12.46803 23.977452
540   439353              3.0485504 12.28064 12.966085
5356 1350232              1.8481659 12.92660 12.441023
4139  374370              1.0269537 12.60572 11.745772
1739  483614              0.9325824 13.73744  9.828717
2537  483614              1.0098117 13.63012  9.012156
1496  484183 genome.wustl 0.9774236 12.22891  8.998754
4941  737183              0.9549693 13.28750  7.440276
947   353292              0.5705900 10.54291  4.553917
5604  317638              0.3663573 12.71267  3.959899
          P.Value          B
2149 2.981631e-11 14.9393981
540  4.929901e-07 10.8250716
5356 6.421453e-07 10.4567996
4139 1.211754e-06  9.9284666
1739 1.565049e-05  8.1922752
2537 4.211962e-05  7.3073136
1496 4.211962e-05  7.2919773
4941 5.593459e-04  5.3136117
947  1.767001e-01  0.5630483
5604 5.269534e-01 -0.5576895
```

Veamos que obtenemos lo mismo utilizando la parametrización “tratamientos” que utilizamos en la práctica 8

```
> design <- matrix(c(rep(1,16),rep(0,8),rep(1,8)),ncol=2)
>   colnames(design) <- c("WT-Ref","KO-WT")
>   design
      WT-Ref KO-WT
 [1,]      1     0
 [2,]      1     0
 [3,]      1     0
 [4,]      1     0
 [5,]      1     0
 [6,]      1     0
 [7,]      1     0
 [8,]      1     0
```

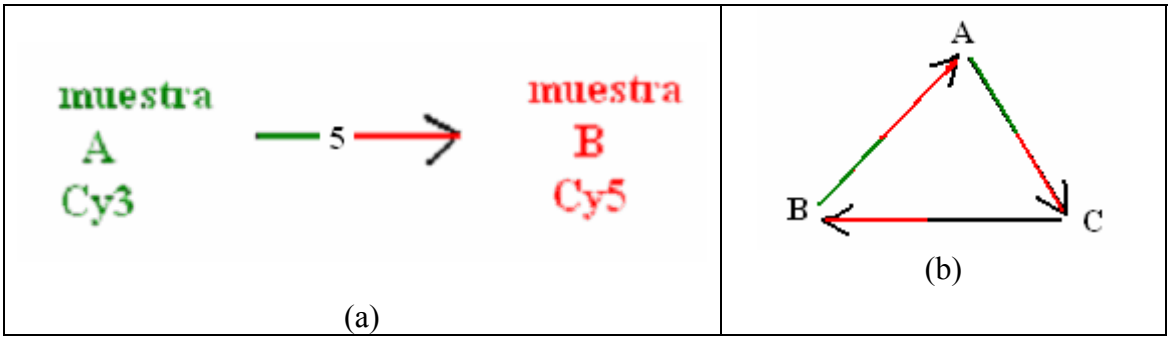
```
 [9,]      1     1
[10,]      1     1
[11,]      1     1
[12,]      1     1
[13,]      1     1
[14,]      1     1
[15,]      1     1
[16,]      1     1
>   fit <- lmFit(MA,design=design)
>   fit <- eBayes(fit)
> topTable(fit,coef="KO-WT")
     GridROW GridCOL ROW COL                  NAME TYPE
2149       2       2   8   7      ApoAI,lipid-Img cDNA
540        1       2   7  15 EST,HighlysimilartoA cDNA
5356       4       2   9   1 CATECHOLO-METHYLTRAN cDNA
4139       3       3   8   2 EST,WeaklysimilartoC cDNA
1739       2       1   7  17    ApoCIII,lipid-Img cDNA
2537       2       3   7  17 ESTs,Highlysimilarto cDNA
1496       1       4  15   5                 est cDNA
4941       4       1   8   6 similartoyeaststerol cDNA
947        1       3   8   2 EST,WeaklysimilartoF cDNA
5604       4       3   1  18                     cDNA
        CLID          ACC          M        A          t
2149 1077520              -3.1661645 12.46803 -23.977452
540   439353              -3.0485504 12.28064 -12.966085
5356 1350232              -1.8481659 12.92660 -12.441023
4139  374370              -1.0269537 12.60572 -11.745772
1739  483614              -0.9325824 13.73744  -9.828717
2537  483614              -1.0098117 13.63012  -9.012156
1496  484183 genome.wustl -0.9774236 12.22891  -8.998754
4941  737183              -0.9549693 13.28750  -7.440276
947   353292              -0.5705900 10.54291  -4.553917
5604  317638              -0.3663573 12.71267  -3.959899
         P.Value          B
2149 2.981631e-11 14.9393981
540  4.929901e-07 10.8250716
5356 6.421453e-07 10.4567996
4139 1.211754e-06  9.9284666
1739 1.565049e-05  8.1922752
2537 4.211962e-05  7.3073136
1496 4.211962e-05  7.2919773
4941 5.593459e-04  5.3136117
947  1.767001e-01  0.5630483
5604 5.269534e-01 -0.5576895
>
```

12.1 Representación gráfica de los experimentos

Los experimentos se representan mediante flechas. Por convención se pone en la cola de la flecha la muestra etiquetada verde y la roja en la punta.

***Vértices*: muestras de** mRNA;
Lados: hibridación;

Dirección: asignación de tintes del rojo al verde .



(a) (b)

La figura (a) representa un experimento con 5 hibridaciones replicadas. En cada microarreglo la muestra A ha sido etiquetada con el tinte verde (por ej. Cy3) y la muestra B con el rojo (por ejemplo Cy5).

La estructura del gráfico determina que efectos pueden ser estimados.

Dos muestras pueden ser comparadas si están unidas por un camino no dirigido (undirected path) correspondiente a dos vértices.

La precisión de los contrastes estimados depende de la cantidad de segmentos que se requieren para unir los dos vértices.

En el experimento hipotético de la figura (b), que consiste en 3 conjuntos de hibridaciones (tres microarreglos) hay 2 caminos que unen A con B; un camino de longitud 1 que une directamente A con B; el otro camino de longitud 2 une A con B vía C. Cuando estimemos la abundancia relativa de las muestras A y B el estimador de log2 (A/B) del camino de A a B es probablemente más preciso que el estimador de log2(A/B) que resulta de estimar log2 (A/C) - log2 (B/C) del camino de longitud 2 que une A con B vía C.

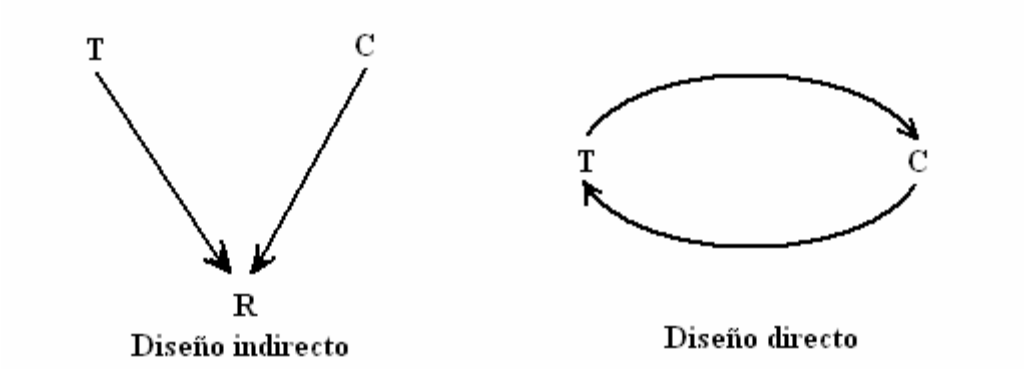
12.2 Diseños directos vs. diseños indirectos

Todas las mediciones de microarreglos de dos colores corresponden a comparaciones apareadas. Consideremos las posibles elecciones en el caso que interesa comparar dos muestras T y C. Estos experimentos incluyen la comparación de células entre

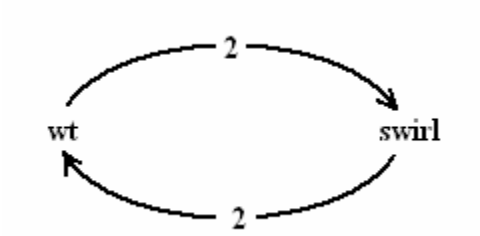
- Tratamiento - Control
- Mutante-wild type
- Tumor -normal

Comparación directa: Medimos la expresión diferencial por log2(T/C) de una única hibridación competitiva

Comparación indirecta: log2 T y log2 C se estiman en dos hibridaciones, T con una tercera muestra R y C con R'. El log-ratio log2(T/C) será reemplazado por la diferencia log2(T/R) — log2(C/R')

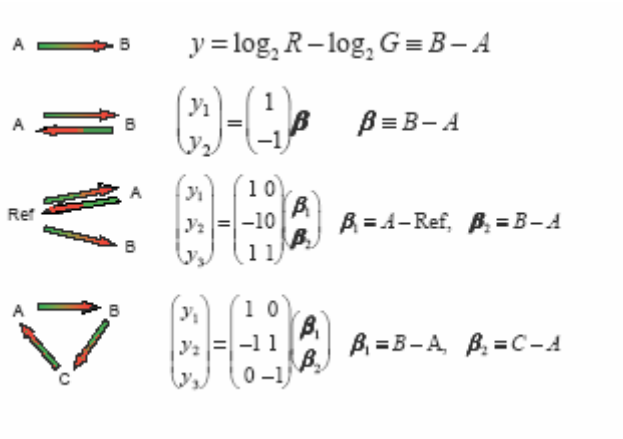


En los estudios iniciales de microarreglos (DeRisi et al., 1996; Spellman et al., 1998; Perou et al., 1999; y muchos otros) se utilizaron diseños indirectos. Posteriormente se empezaron a realizar diseños directos (Jin et al., 2001; Kerr et al., 2001; Lin et al., 2002)



La figura anterior representa el experimento de swirl en el que se realiza un par de experimentos dye swap comparando el pez cebra (zebrafish) mutante swirl y el wild-type. El número 2 en las flechas representa las replicaciones.

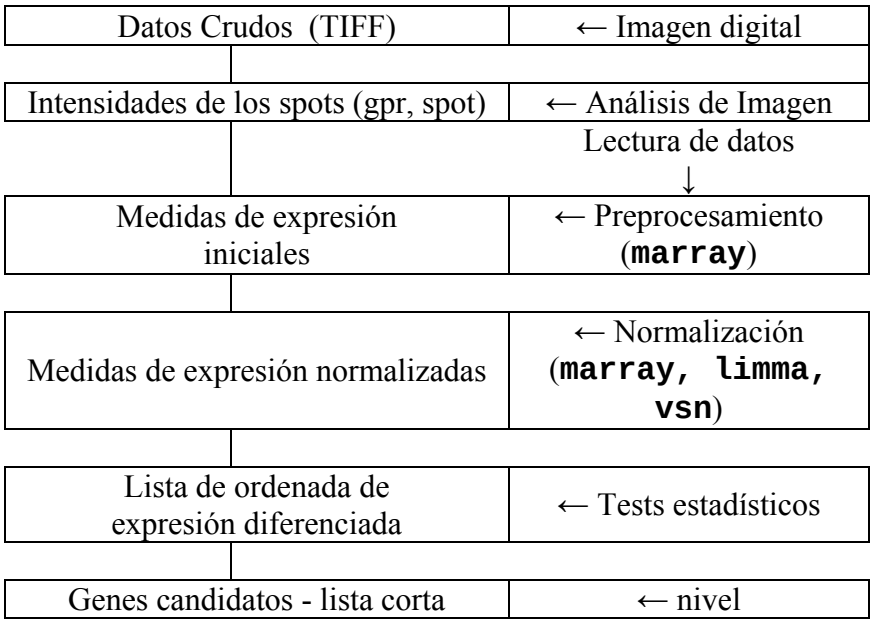
Analicemos los siguientes diseños y sus correspondientes matrices



¿qué significan los coeficientes en cada caso?

Pueden encontrar más ejemplos con distintos diseños en la Guía Limma

Hemos completado el esquema general para hallar genes diferencialmente expresados utilizando microarrays de dos canales.



Bioconductor tiene unos 180 paquetes para procesamiento de datos de microarreglos

Referencias

DeRisi, J. et al., Use of a cDNA microarray to analyse gene expression patterns in human cancer, Nature Genetics, 14:457-460, 1996.

Dudoit, S. et al., Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, Statistica Sinica, **12**: 111-139, 2002.

Kerr, M.K., P. Leiter, and G.A. Churchill, Analysis of a designed microarray experiment, in Proc. IEEE-Eurasip Nonlinear Signal and Image Processing Workshop, June 2001.

Lin, D.M. et al., A spatial map of gene expression in the olfactory bulb, submitted, 2002.

Perou, CM. et al., Distinctive gene expression patterns in human mammary epithelial cells and breast cancers, Proc. Natl. Acad. Scl, 96:9212-9217, 1999.

Spellman, RT. et al., Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization, Molecular Biol. of the Cell, 9:3273-3297, 1998.