

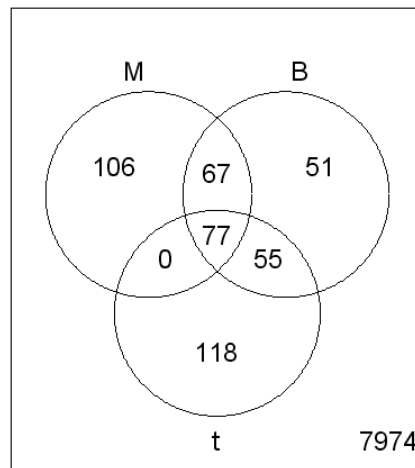
10. Gráficos. Comparación de criterios de ordenamientos de genes

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]

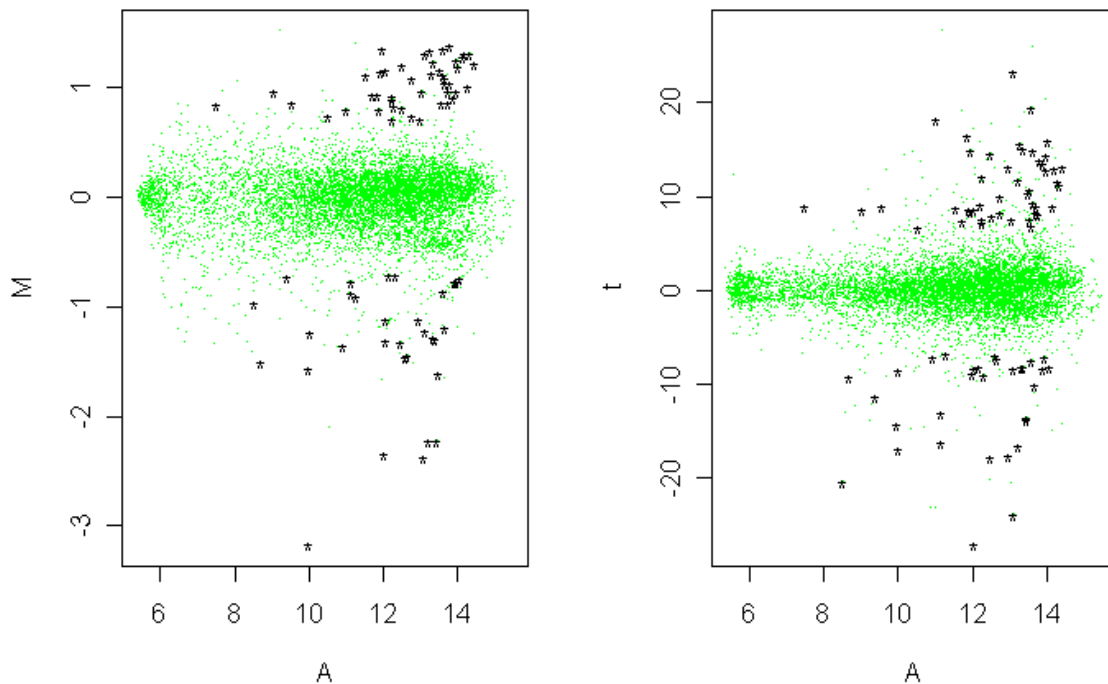
M <- vector
B <- vector
t <- vector

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="*")
```

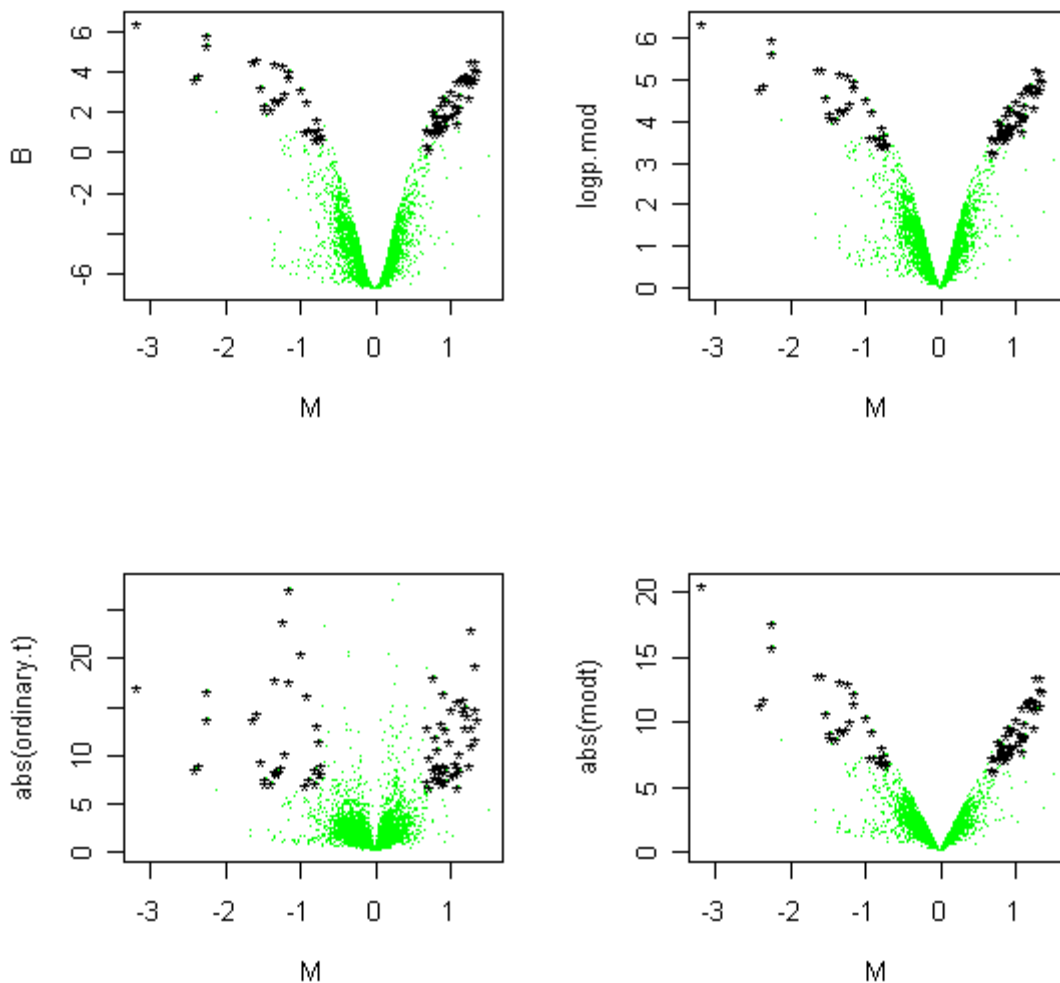
El primero de los gráficos es similar al MA plot que ya conocíamos con la diferencia que ahora tenemos un sólo gráfico para todos los arreglos juntos. M es el promedio de todas las $M_i = \log(R/G)$ de cada uno de arreglos, en el que se ha tenido en cuenta los dye-swaps.

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="*")

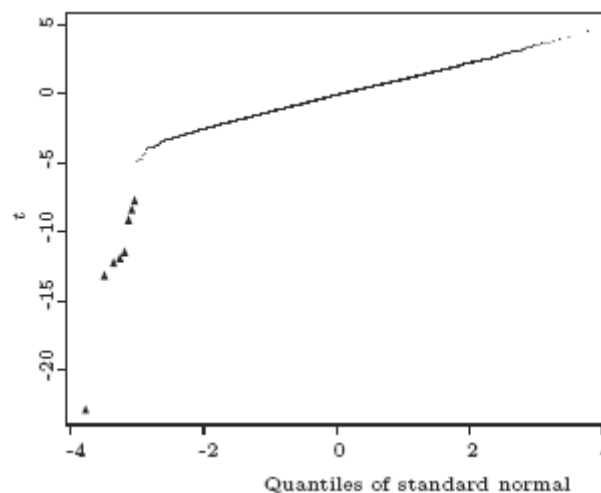
modt <- fit$t
plot(M,abs(modt),pch='.',col="green")
points(M*interseccion,abs(modt)*interseccion, pch="*")
```

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 $\bar{M}$ 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 $\log(\text{intensidad})$ 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.

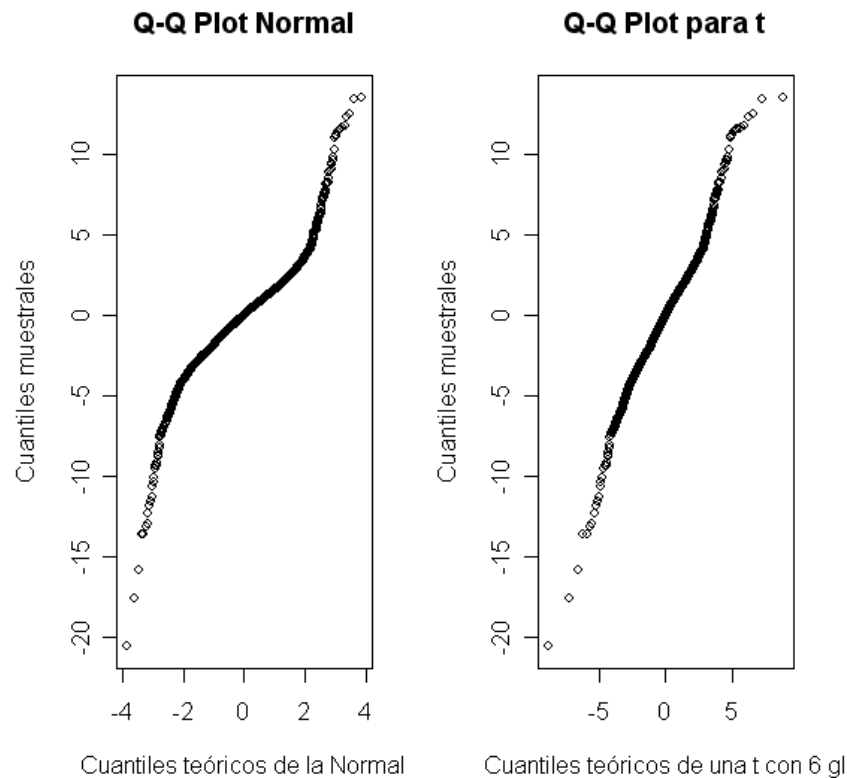
En general, los Q-Q plots se utilizan para validar si los datos provienen de una distribución particular ó si dos conjuntos de datos tienen la misma distribución. Sin embargo en este contexto se utilizan como ayuda visual de identificación de genes con estadísticos inusuales. Proveen una manera informal 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 un 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.

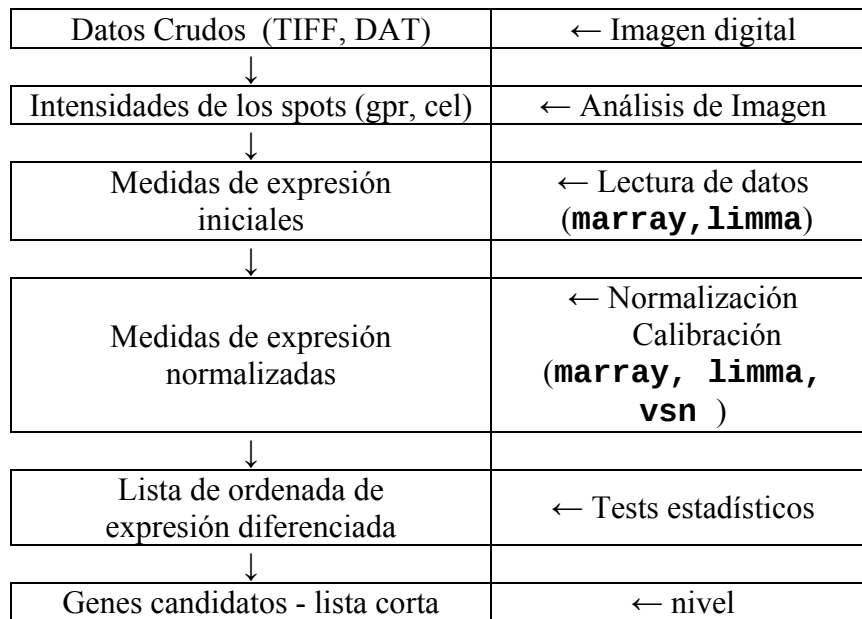
Como en el ejemplo anterior, algunas veces es posible determinar donde terminan los outliers y comienza la masa de estadísticos desafortunadamente este no es el caso de nuestro ejemplo de Swirl.



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")
```

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



Bioconductor tiene unos 350 paquetes para procesamiento de datos de microarreglos.