Pipeline: miRNA

María Araceli Hernández Betancor

24 de junio, 2020

Índice

DATASET: GSE54578

Una vez instalados los paquetes de Bioconductor necesarios para el análisis de datos de *microarray* de miRNA, se desacarga el *dataset* GSE54578 de la página web Pubmed (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE54578). El estudio analiza la expresión de miRNA de todo el genoma en sangre de 15 casos de esquizofrenia (SZ) de inicio temprano (EOS) y 15 controles sanos. Los microarrays de muestras detectaron un total de 1070 miRNAs. Se analiza la expresión de miRNA en 15 muestras de pacientes con esquizofrenia y 15 controles sanos para explorar la alteración de los miRNA en la esquizofrenia, siguiendo el flujo de trabajo propuesto por *Emilio Mastriani et al.* Se utiliza la plataforma GPL16016 (Exiqon miRCURY LNA microRNA array). Se puede realizar la descarga directamente mediante el enlace, o a través del paquete *"GEOquery"* y la función *"getGEO"*. Finalmente se guarda el conjunto de datos en *gset*, para su posterior procesamiento y análisis.

```
library("GEOquery")
gset<-getGEO("GSE54578", GSEMatrix=TRUE, AnnotGPL=FALSE)
if(length(gset)>1) idx <-grep("GPL1606", attr(gset,"-names")) else idx<-1
gset<-gset[[idx]]
```

PREPROCESAMIENTO

Los datos de expresión miRNA pueden tener valores perdidos "NA" y las columnas son nominadas como "GSM". Se puede mostrar la estructura de los datos, y extraer inicialmente los regitros con valores perdidos y renombrar las columnas para facilitar su lectura (SCHIZ, CTRL).

```
head(exprs(gset)) ## Se visualiza la presencia de valores perdidos
```

```
      GSM1319258  GSM1319259 GSM1319260 GSM1319261 GSM1319262 GSM1319263 GSM1319264
4610   0.4464805  0.59599354  1.3857880 -0.1584685 -3.5064605  0.2568499  0.6547877
4700   0.1128415 -0.08805618  1.0623719  0.6509808 -4.1818383  0.8375980  0.2771786
5730  -5.1601770 -4.24792751 -4.2286244 -4.0231896 -3.1186445 -3.9096360 -3.1292830
6880  -8.2045709 -7.66296487 -8.5360526 -8.3630395 -5.9259995 -7.7169908 -6.9872641
9938   0.2132814  0.48169323  1.1406094 -0.0927443 -0.9373147  0.2006795  0.4189410
10138 -1.6499823  0.35940280  0.5524704 -0.8592139 -3.4535116 -1.2155517 -0.6766512
      GSM1319265 GSM1319266 GSM1319267 GSM1319268 GSM1319269 GSM1319270 GSM1319271
4610  -1.2473593 -0.7183438 -0.4821903  -1.260937  -1.172444  -2.015682 -0.3828823
4700  -2.7915300 -2.6722474 -2.0842263  -3.190398  -2.815288  -3.443919 -2.3795388
5730  -4.2801079 -4.3127050 -5.2438829  -4.503375  -4.242185  -5.803461 -5.7193888
6880  -6.9879272 -7.5256988 -7.3390403  -8.295933  -7.970106  -7.724026 -8.7193890
9938  -0.9984854 -0.6609498 -0.9021394  -1.425568  -1.407863  -1.565597 -1.1213363
10138 -2.6047989 -2.1777755 -1.9127754  -1.971431  -2.182203  -4.381634 -1.4156081
      GSM1319272 GSM1319273 GSM1319274  GSM1319275 GSM1319276 GSM1319277 GSM1319278
4610   -1.157132  0.7779281  0.5872010   0.9637209  -3.012142  -1.805000 -0.9869923
4700   -2.425711  0.2163341 -1.1337002   0.6316126  -3.824641  -2.374745 -2.6215475
5730   -4.459432 -4.6762777 -5.7459544  -7.8127122  -6.824641  -9.479780 -6.1193255
6880   -8.083922 -8.5507466 -8.0089886 -10.3976743         NA         NA -8.2892506
9938   -1.270141  0.3041216  0.4479816   0.6917758  -2.699485  -1.235020 -1.1327459
10138  -2.320157 -0.1512908  0.1739021   0.2818055  -4.531859  -3.116376 -1.9643199
      GSM1319279 GSM1319280 GSM1319281 GSM1319282 GSM1319283 GSM1319284 GSM1319285
4610   1.0405705  -5.457443  -1.204358  -1.698946  -3.264817  -4.565913  -1.844769
4700  -0.2676708  -3.631472  -3.512192  -3.073581  -4.194114  -4.797239  -3.930160
5730  -5.6724254  -7.042406  -5.594654  -5.370563  -5.869679  -6.320801  -5.117787
```

```
6880  -9.1318571  -8.916874  -8.682117         NA -10.039604  -8.320801         NA
9938   0.7959210  -4.857981  -1.927229  -1.632567  -2.481184  -3.366604  -0.725470
10138  0.3559831         NA  -3.016781  -3.300174  -4.410248  -5.513446  -3.097888
      GSM1319286 GSM1319287
4610   -2.439081  2.1339084
4700   -2.738306  4.4644310
5730   -6.109328         NA
6880  -10.809768         NA
9938   -1.670217  1.6198764
10138  -3.068301  0.5110349
```

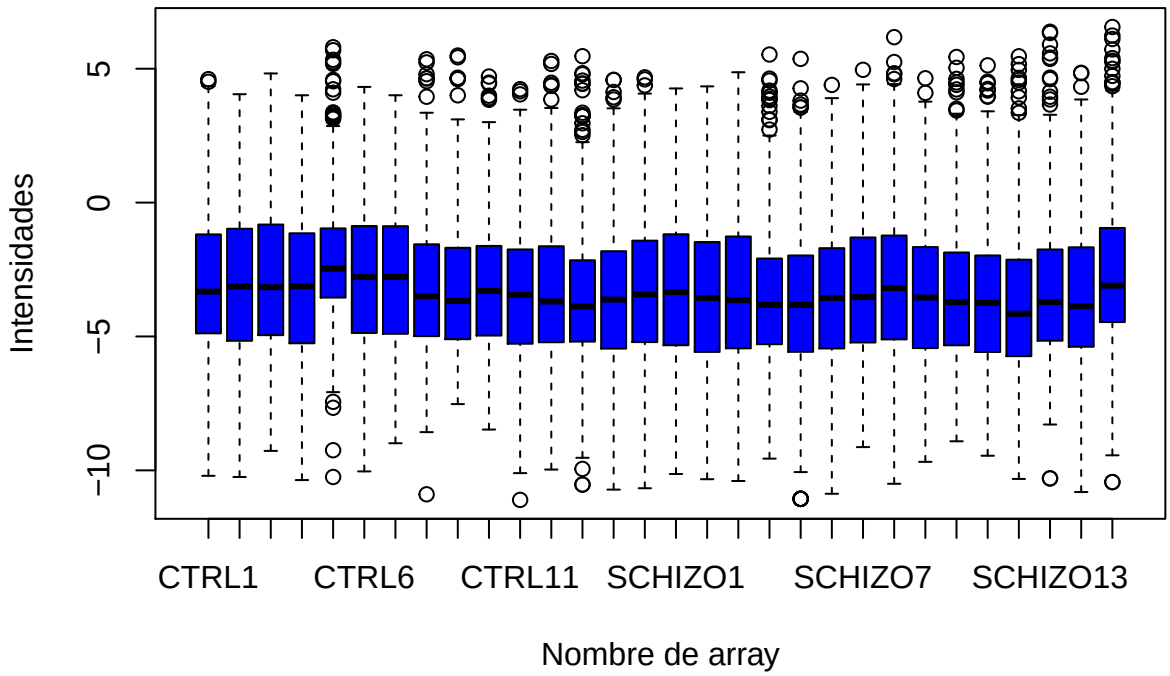
```
rmv<-which(apply(exprs(gset),1,function(x) any (is.na(x))))
```

```
gset<-gset[-rmv,]
sampleNames(gset)<-c("CTRL1","CTRL2","CTRL3","CTRL4","CTRL5","CTRL6","CTRL7","CTRL8","CTRL9","CTRL10","C
gsms<-"00000000000000011111111111111"
sml<-c()
for(i in 1:nchar(gsms)) {sml[i]<-substr(gsms,i,i)}
head(exprs(gset))
```

```
          CTRL1       CTRL2     CTRL3      CTRL4      CTRL5      CTRL6      CTRL7
4610   0.4464805  0.59599354  1.385788 -0.1584685 -3.5064605  0.2568499  0.6547877
4700   0.1128415 -0.08805618  1.062372  0.6509808 -4.1818383  0.8375980  0.2771786
9938   0.2132814  0.48169323  1.140609 -0.0927443 -0.9373147  0.2006795  0.4189410
10306 -1.2102177 -2.81329928 -1.749457 -2.3518124 -3.5474878 -1.8101003 -1.9265681
10919 -0.0159823  0.72576985  1.249236  0.3234609 -0.9625253  0.3053769  0.4071987
10923  0.8370880  1.44512378  2.581850  1.7047308 -0.4166203  1.9412206  1.4793223
          CTRL8      CTRL9     CTRL10    CTRL11    CTRL12    CTRL13     CTRL14    CTRL15
4610  -1.2473593 -0.7183438 -0.4821903 -1.260937 -1.172444 -2.015682 -0.3828823 -1.157132
4700  -2.7915300 -2.6722474 -2.0842263 -3.190398 -2.815288 -3.443919 -2.3795388 -2.425711
9938  -0.9984854 -0.6609498 -0.9021394 -1.425568 -1.407863 -1.565597 -1.1213363 -1.270141
10306 -4.9175379 -4.4615684 -4.3722070 -5.643856 -5.415517 -6.007819 -3.9512045 -4.968445
10919 -1.6878034 -1.0812485 -0.9467228 -1.206955 -1.533394 -2.597691 -0.8291245 -1.043176
10923 -1.6540264 -1.1506593 -1.1856033 -1.562191 -1.820359 -2.763197 -1.5394797 -1.616317
         SCHIZO1    SCHIZO2    SCHIZO3   SCHIZO4   SCHIZO5    SCHIZO6    SCHIZO7
4610   0.7779281  0.5872010  0.9637209 -3.012142 -1.805000 -0.9869923  1.0405705
4700   0.2163341 -1.1337002  0.6316126 -3.824641 -2.374745 -2.6215475 -0.2676708
9938   0.3041216  0.4479816  0.6917758 -2.699485 -1.235020 -1.1327459  0.7959210
10306 -2.0271848 -3.3422322 -2.2378033 -5.937115 -6.420887 -4.9915699 -2.7481527
10919  0.8790086  0.7605185  1.2124274 -2.711940 -1.191299 -1.1720403  1.0567319
10923  1.4981941  0.7492344  2.2620983 -2.833686 -1.081749 -1.1548241  1.6359137
        SCHIZO8   SCHIZO9  SCHIZO10  SCHIZO11  SCHIZO12   SCHIZO13  SCHIZO14  SCHIZO15
4610  -5.457443 -1.204358 -1.698946 -3.264817 -4.565913 -1.8447689 -2.439081 2.1339084
4700  -3.631472 -3.512192 -3.073581 -4.194114 -4.797239 -3.9301604 -2.738306 4.4644310
9938  -4.857981 -1.927229 -1.632567 -2.481184 -3.366604 -0.7254700 -1.670217 1.6198764
10306 -6.414374 -6.360189 -5.799406 -7.039605 -8.320801 -5.6438562 -5.350336 1.7642502
10919 -4.479469 -1.602632 -1.885136 -2.722192 -3.606555 -0.9930916 -2.255179 3.2705074
10923 -4.273018 -1.659749 -2.261441 -3.079603 -3.579334 -1.5129253 -2.585766 0.3956792
```

Se comprueba gráficamente la intensidades de sonda para buscar posibles *outliers*, que podrían ser excluidos de un análisis posterior. Se observa la uniformidad de la intensidad de la señal a través de la función *"boxplot()"*, con escasa variabilidad entre los arrays.

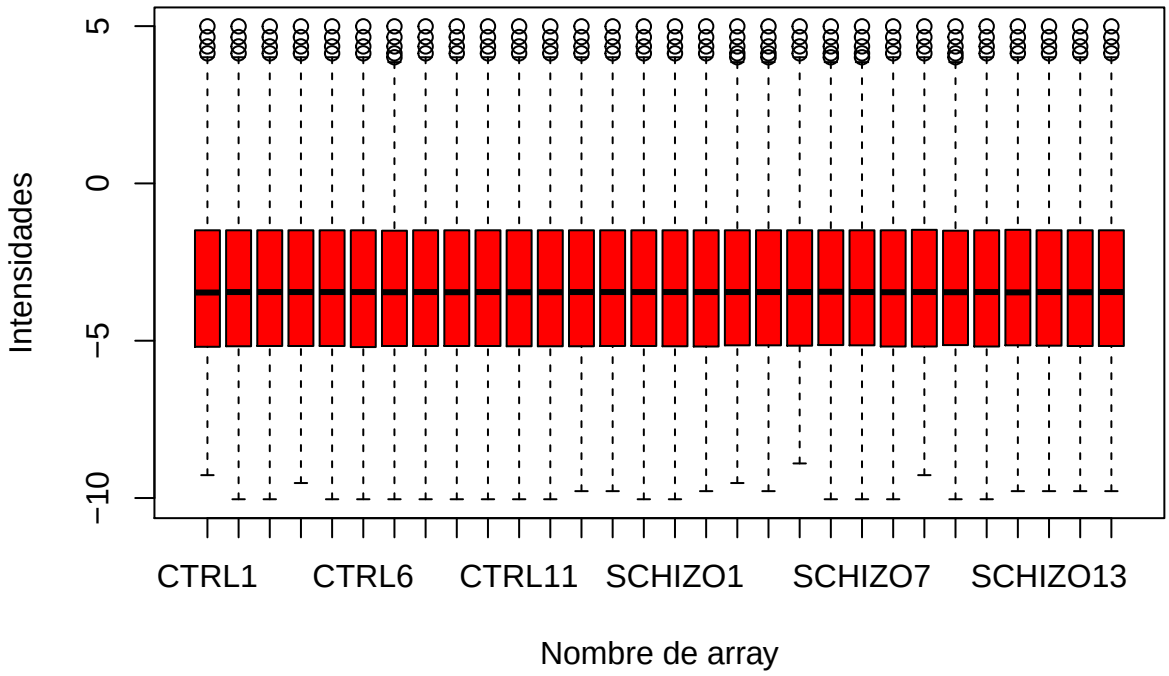
```
ex<-exprs(gset)
boxplot(ex, which="pm", ylab="Intensidades", xlab="Nombre de array", col="blue")
```



NORMALIZACIÓN

Tras renombrar los *arrays* y filtrar los datos de posibles sesgos experimentales, se continua con la normalización de los datos para evitar la variabilidad de origen no biológico. Existen diferentes paquetes para realizar la normalización ("ExiMir" y la función *"NormiR"* para los *microarrays* de dos colores, "affy" para los *arrays* de Affymetrix). En este caso se ha utilizado el paquete "limma" y la función *"normalizeBetweenArrays"* que permite una *quantile normalization*. Se muestra gráficamente con la función *"boxplot"* el efecto de la normalización de los datos. Se realiza una transformación log2 de los valores de expresión normalizados para favorecer la distribución gaussiana.

```
library("limma")
ex_norm<-normalizeBetweenArrays(ex)
qu<-as.numeric(quantile(ex,c(0.,0.25,0.5,0.75,0.99,1.0),na.rm=T))
filt<-(qu[5]>100 || (qu[6]-qu[1]>50 && qu[2]>0 || (qu[2]>0 && qu[2]<1 && qu[4]>1 && qu[4]<2)))
if(filt){ex_norm[which(ex<=0)]<-NaN; exprs(gset)<-log2(ex_norm)}
boxplot(ex_norm, which="pm", ylab="Intensidades", xlab="Nombre de array", col="red")
```



Los datos normalizados se pueden comparar entre grupos mediante *T Test*, para medir si la diferencia de expresión es significativa entre ambos grupos (*p-valores* más bajos). Esta comparación de genes entre grupos se realiza de manera múltiple y simultánea. Para la comparación se utiliza la función *eBayes* del paquete *"limma"*. Mediante esta función se calcula a través de una moderación empírica de Bayes de los errores estándar hacia un valor común, y dado un ajuste de modelo lineal de *microarrays*, estadísticos t moderados, estadístico F moderado y probabilidades logarítmicas de expresión diferencial.

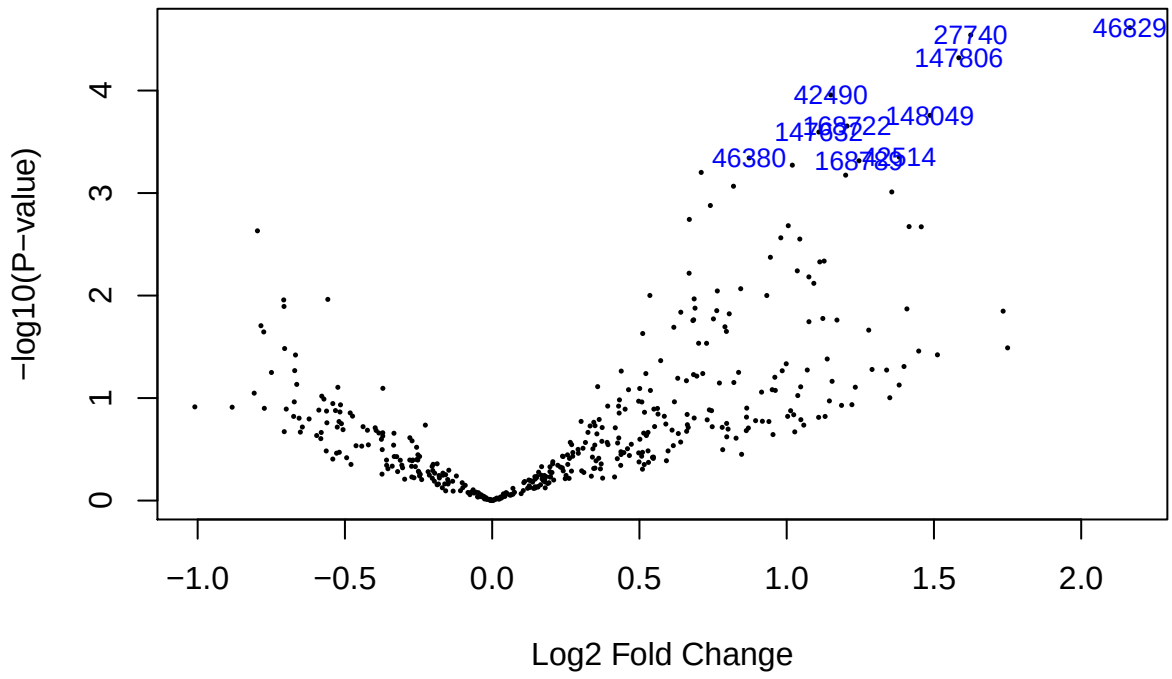
```
sml<-paste("G",sml,sep="")
fl<-as.factor(sml)
gset$description <- fl
design <- -model.matrix(~ description + 0, gset)
colnames(design) <- levels(fl)
fit <- lmFit(gset, design)
cont.matrix <- makeContrasts(G1-G0, levels=design)
fit2 <- contrasts.fit(fit, cont.matrix)
fit2 <- eBayes(fit2, 0.01)
tT <- topTable(fit2, adjust="fdr", sort.by="B", number=1000)
head(tT,10)
```

	ID	Name	miRNA_ID_LIST	Da
46829	46829	hsa-miR-664-5p	hsa-miR-664-5p	mi
27740	27740	hsa-miR-574-5p/mmu-miR-574-5p	hsa-miR-574-5p,mmu-miR-574-5p	mi
147806	147806	hsa-miR-3149	hsa-miR-3149	mi
42490	42490	hsa-miR-505-5p/mmu-miR-505-5p/rno-miR-505*	hsa-miR-505-5p,mmu-miR-505-5p,rno-miR-505*	mi
148049	148049	hsa-miR-3924	hsa-miR-3924	mi

	ID	Name	miRNA_ID_LIST	Da
168722	168722	hsa-miR-4742-3p	hsa-miR-4742-3p	mi
147632	147632	hsa-miR-4297	hsa-miR-4297	mi
42514	42514	hsa-miR-937	hsa-miR-937	mi
46380	46380	hsa-miR-1255a	hsa-miR-1255a	mi
168789	168789	hsa-miR-4686	hsa-miR-4686	mi

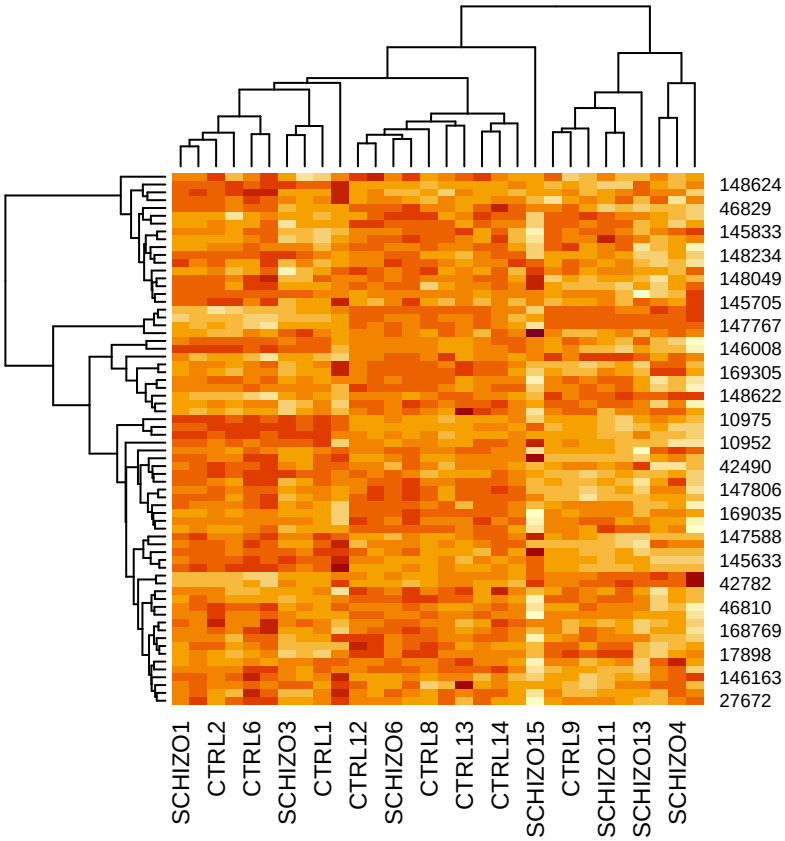
Los resultados del contraste se guardan en los objetos *"fit2"* y *"tT"*, para su posterior análisis. Es importante desde el punto de la significación, tanto el el p-valor estadístico como la amplitud del *fold change*. Estos valores pueden representarse gráficamente en un *"volcanoplot"*. En el gráfico en el eje X se representa el *fold change*, y en el eje Y los p-valores (resultados log-transformados), resaltando los conjuntos de sondas superiores. Se muestran los miRNAs diferencialmente expresados entre muestras de pacientes con esquizofrenia y controles. También se puede usar la función *"ggplot2"*.

```
volcanoplot(fit2, coef=1, highlight=10)
```



Otra opción es realizar un análisis *"clustering"* de los datos *microarray* de miRNA. Se puede visualizar un *"heatmap plot"* de un subconjunto de miRNAs con expresión diferencial significativa entre pacientes con enfermedad y controles (FDR, p-valor ajustado inferior a 0.05).

```
selected<-which(p.adjust(fit2$p.value[,1]<0.05) == 1)
esetSe1<-ex_norm[selected,]
heatmap(esetSe1)
```



DIANA DE miRNA

Esta es una particularidad que nos encontramos en el análisis de datos miRNA a diferencia de otros datos transcriptómicos, el análisis de genes diana específicos (*target*). Los miRNAs regulan la expresión de genes diana post-transcripción o traducción, resultando relevante la anotación de sus genes diana. Para la identificación de *target genes* de miRNAs, se pueden usar distintas herramientas. Para el desarrollo de este "*pipeline*" se utiliza el paquete *"SpidermiR"*. Permite obtener *target genes* validados y predichos de múltiples bases de datos o herramientas de *software* (miR2Disease, miR-Tar, mirWalk, miRTarBase, miRandola, DIANA, Pharmaco-miR, PicTar, Miranda y TargetScan). Se pueden visualizar redes de genes. Se usan los 5 miRNAs con expresión diferencial entre grupos más significativos. Las dianas de estos miRNAs se predicen con *"SpidermiRdown-load_miRNAprediction"* y son exportados a *"mirnaTar"*. Se obtiene la predicción con las herramientas Miranda, DIANA, PicTar y TargetScan. Se puede visualizar el *data frame*, se visualiza una primera columna con los nombres de miRNA y una segunda columna con el listado de genes diana. Otra opción sería descargar las dianas validadas desde miRTAR y miRwalk con la función *"SpidermiRdown-load_miRNAvalidate"*

```
library(SpidermiR)
tT[selected,]$Name[1:5]
```

```
[1] "hsa-miR-4429"    "hsa-miR-1827"    "hsa-miR-5002-5p" "hsa-miR-5187-3p"
[5] "hsa-miR-4455"
```

```
mirna<-c("hsa-miR-4429","hsa-miR-1827","hsa-miR-5002-5p","hsa-miR-5187-3p","hsa-miR-4455")
mirnaTar<-SpidermiRdownload_miRNAprediction(mirna_list=mirna)
```

```
[1] "Processing... hsa-miR-4429"
[1] "Processing... hsa-miR-1827"
[1] "Processing... hsa-miR-5002-5p"
[1] "Processing... hsa-miR-5187-3p"
[1] "Processing... hsa-miR-4455"
```

```
head(mirnaTar,10)
```

	V1	V2
NAT1	hsa-miR-4429	NAT1
AK4	hsa-miR-4429	AK4
ALOX12	hsa-miR-4429	ALOX12
AMBN	hsa-miR-4429	AMBN
XIAP	hsa-miR-4429	XIAP
AR	hsa-miR-4429	AR
ARF1	hsa-miR-4429	ARF1
RHOG	hsa-miR-4429	RHOG
ARHGAP5	hsa-miR-4429	ARHGAP5
ZFHX3	hsa-miR-4429	ZFHX3

ANÁLISIS DE ENRIQUECIMIENTO

A través del análisis de redes se pueden observar las dianas compartidas de múltiples miRNAs, y también las interacciones y *pathways* entre genes diana. Se puede usar la herramienta *"Cytoscape"* para la construcción de una red regulatoria entre los 5 miRNAs más significativos y las dianas predichas, 50 por cada miRNA. GeneMANIA trata redes validadas y predichas entre genes de una variedad de especies y proporciona un servidor web para visualizarlo. Entre los tipos de red se incluye: colocalización, coexpresión, *pathway*, interacciones genéticas y físicas, dominios proteicos compartidos, y predicción de interacciones. *"SpidermiR"* permite descargar los datos de interacción de GeneMANIA y visualizar la red de genes. El análisis de enriquecimiento para miRNAs y dianas también debe realizarse para estudiar su significación biológica y aumentar su potencia estadística. Se utiliza el paquete *"GOstats"* para realizar el analisis de enriquecimiento GO (proceso biológico) para los genes diana predichos de los 5 miRNAs más significativos.

```
library("org.Hs.eg.db")
library("GSEABase")
library("GOstats")
mirTarget<-as.vector(mirnaTar$V2)
goAnn<-get("org.Hs.egGO")
universe<-Lkeys(goAnn)
entrezIDs<-mget(mirTarget, org.Hs.egSYMBOL2EG, ifnotfound = NA)
entrezIDs<-as.character(entrezIDs)
params<-new("GOHyperGParams",
            geneIds=entrezIDs,
            universeGeneIds=universe,
            annotation="org.Hs.eg.db",
            ontology="BP",
```

```
            pvalueCutoff=0.01,
            conditional=FALSE,
            testDirection="over")
goET<-hyperGTest(params)
```

```
library(Category)
GObp<-summary(goET)
head(GObp)
```

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0048731	0	1.593279	608.2236	804	4670	system development
GO:0007399	0	1.800407	288.6132	439	2216	nervous system development
GO:0007275	0	1.559329	680.5072	878	5225	multicellular organism development
GO:0048856	0	1.544399	735.8594	936	5650	anatomical structure development
GO:0032502	0	1.530347	786.7835	988	6041	developmental process
GO:0048523	0	1.531822	596.6322	773	4581	negative regulation of cellular process

Se realiza además análisis de enriquecimiento KEGG.

```
keggAnn<-get("org.Hs.egPATH")
universe <-Lkeys(keggAnn)
params<-new("KEGGHyperGParams",
            geneIds=entrezIDs,
            universeGeneIds=universe,
            annotation="org.Hs.eg.db",
            categoryName="KEGG",
            pvalueCutoff=0.01,
            testDirection="over")
keggET<-hyperGTest(params)
kegg<-summary(keggET)
library(Category)
head(kegg)
```

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04012	0.0000002	3.631004	11.29562	30	87	ErbB signaling pathway
05200	0.0000002	2.112076	42.32612	75	326	Pathways in cancer
04360	0.0000005	2.893085	16.74868	38	129	Axon guidance
04310	0.0000104	2.427872	19.47521	39	150	Wnt signaling pathway
04510	0.0000111	2.191508	25.96694	48	200	Focal adhesion
05215	0.0001062	2.672893	11.55529	25	89	Prostate cancer