ANÁLISIS DIFERENCIAL DE GENES UTILIZANDO EL MÉTODO SAM

ESTADÍSTICA GENÓMICA

Taller 3 Solución

Autor:
Jesús David Niño Torres
Profesora:
Liliana López Kleine

Universidad Nacional de Colombia
Facultad de Ciencias
Departamento de Estadística

Índice

1. Descripción del experimento

Título: Un enfoque multiómico integrado identifica alteraciones epigenéticas asociadas con la enfermedad de Alzheimer

Organismo: Homo sapiens

Tipo de experimento: Generación de perfiles de expresión mediante secuenciación de alto rendimiento

Resumen del experimento: Se presenta la secuencia de ARN a granel en el hipocampo humano sujeto a envejecimiento normal y enfermedad de Alzheimer.

Datos: Los datos son de RNA-seq del lóbulo temporal lateral posmortal de los cerebros afectados por la enfermedad de Alzheimer (EA), control adulto (Old) y control joven (Young). Para este trabajo se va a considerar solo dos condiciones biológicas: personas con la enfermedad de Alzheimer (AD) y personas sin la enfermedad de Alzheimer (Control).

Clase de datos: Dataframe

Dimensión de los datos: Features: 27130, Samples: 30

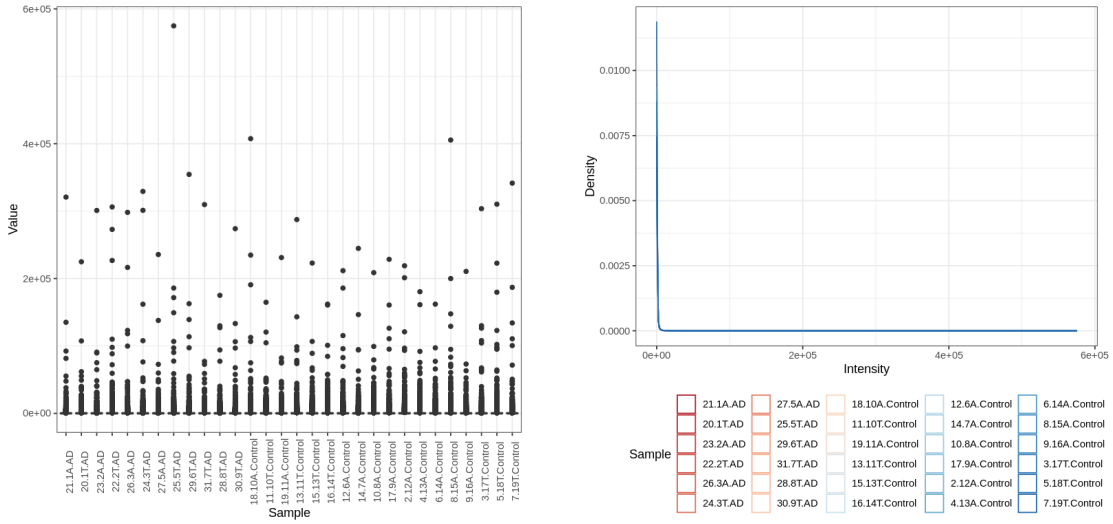
2. Procesamiento de datos

2.1. Control de calidad

Para poder detectar problemas en las muestras de RNA-seq que surgen en la mayoría de los casos por la calidad de los experimentos y que pueden interferir en un posterior análisis genómico se debe realizar el siguiente análisis descriptivo:

2.1.1. Boxplot y densidad

Para detectar las irregularidades en las muestras y supuestos necesarios para el respectivo análisis es útil realizar los boxplot y los gráficos de densidad. Los gráficos de boxplot de los datos seleccionados, Figura 1a, indica que hay bastantes datos atípicos, los datos en cada muestra son muy dispersos con una expresión mediana baja y su distribución es sesgada a la derecha. Con respecto a los gráficos de densidad, Figura 1b, se desconoce si existe bimodalidades o formadas anómalas en la estructura de la densidad de cada muestra. Por lo tanto, no se puede concluir si existen muestras defectuosas hasta que se realice la respectiva normalización de los datos.



a) Box-Whisker

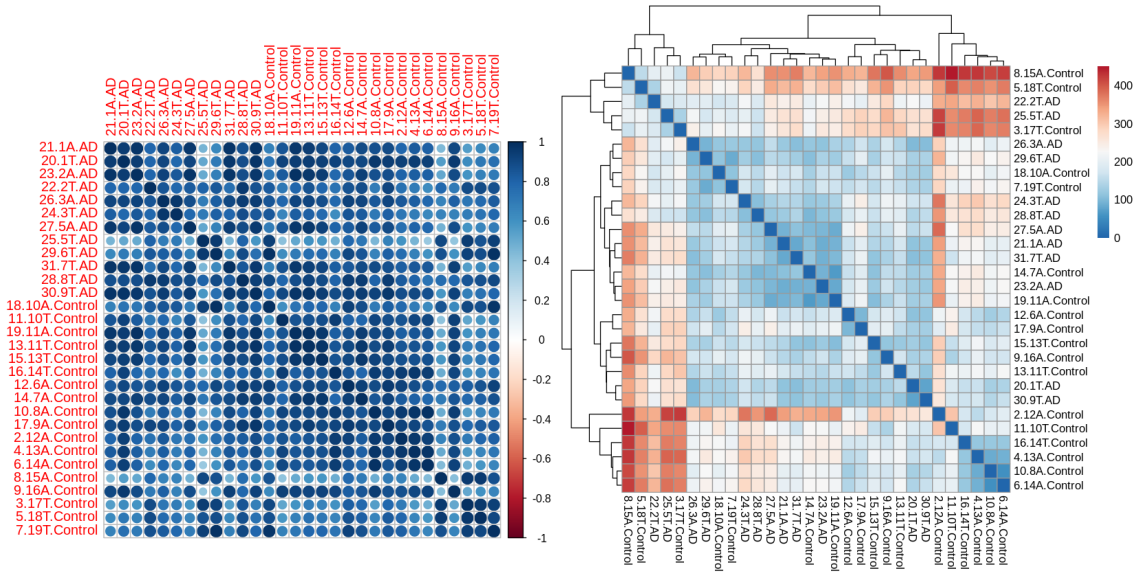
b) Densidad de valor de expresión

Figura 1

2.1.2. Correlación y Heatmap

Otros métodos apropiados para encontrar muestras defectuosas, esta vez a partir de relaciones entre variables, es el gráfico de correlación y el dendrograma. Para los datos seleccionados, del gráfico de correlación, Figura 2a, se puede deducir fácilmente que existen muestras tanto de grupo control como de tratamiento que presentan correlaciones bajas. Del grafico Heatmap (*distancias calculadas por la función dist2()*), Figura 2b, se puede observar que no hay una homogeneidad en las distancias dentro

de la gran mayoría de los grupo. Por lo tanto, se deben tener en cuenta las muestras posiblemente problemáticas detectadas en los gráficos anteriores para decidir si se eliminan luego de una posterior normalización. No se mencionan cuales son las muestras problema ya que son bastantes a considerar.



a) Matriz de correlación b) Heatmap

Figura 2

2.1.3. SD vs Mean

El gráfico de las desviaciones estándar de las filas frente a los promedios de las filas, permite verificar visualmente si existe una dependencia de la desviación estándar frente a la media. La línea roja representa el estimador de la mediana corriente. Si no hay dependencia, entonces la línea debe ser aproximadamente horizontal. Para el caso de los datos seleccionados, el gráfico Figura 3 indica que no es claro si existe una dependencia de la desviación estándar frente a la media.

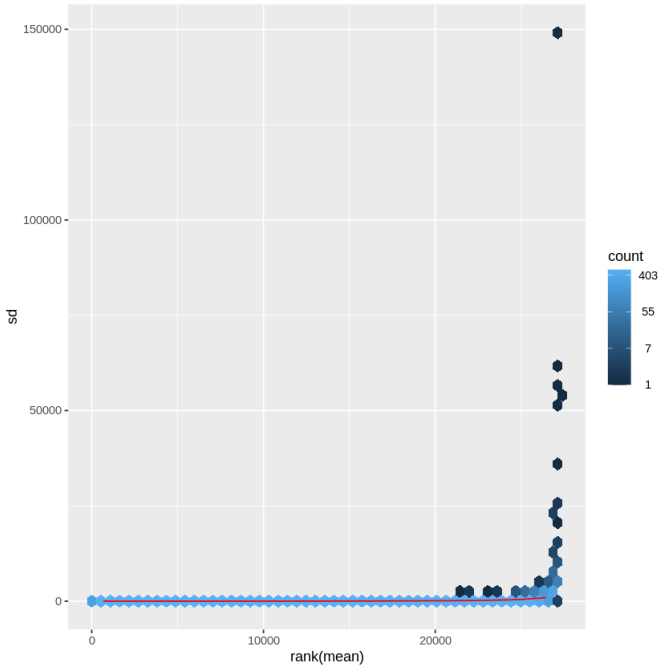


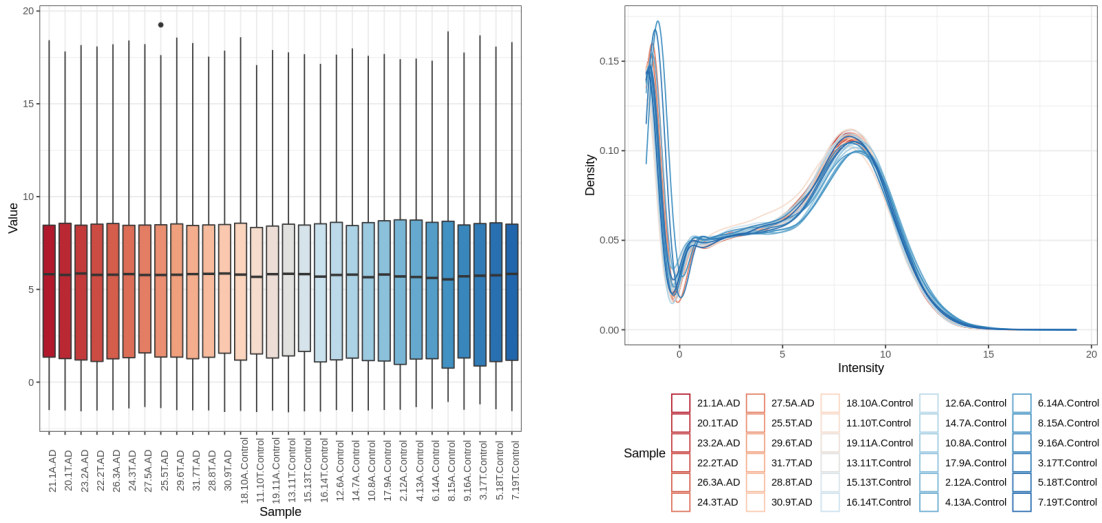
Figura 3: Desviaciones estándar de las filas frente a las medias de las filas

2.2. Normalización

La normalización es uno de los procedimientos de transformación de datos más importantes para poder llevar acabo un correcto análisis de datos genómicos. Para este trabajo se realiza la normalización con el fin de detectar genes expresados diferencialmente (DEG) de RNA-seq por el metodo SAM. Para la normalización de los datos se utiliza el método VSN (normalización estabilizadora de varianza), el cual transforma los datos de tal manera que la varianza permanece casi constante en todo el espectro de intensidad.

2.2.1. Boxplot y densidad

Al realizar los boxplot, Figura 4a, y los gráficos de densidad, Figura 4b, con los datos normalizados, se puede observar que existe un solo dato atípico en la muestra *25.5T.AD*. Las muestras se distribuyen aproximadamente igual y están sesgadas a la derecha con una expresión mediana baja. Sin embargo, las curvas de densidad poseen una estructura bimodalidad. Lo anterior sugiere una filtración de genes de baja expresión.

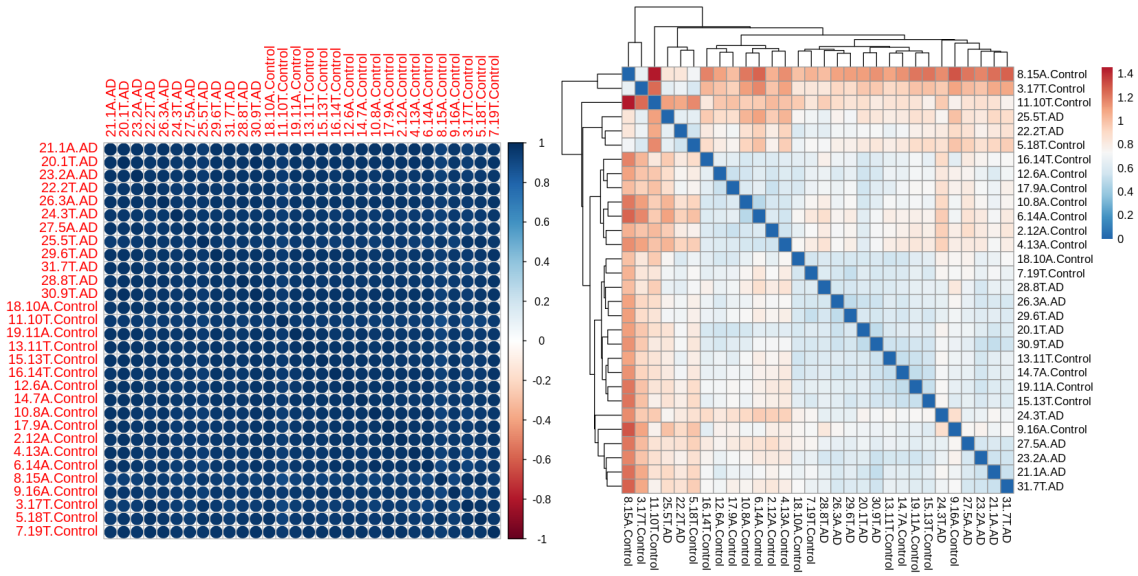


a) Box-Whisker b) Densidad de valor de expresión

Figura 4

2.2.2. Correlación y Heatmap

Al realizar el gráfico de la matriz de correlación, Figura 5a, con los datos normalizados, se puede concluir fácilmente que las muestras tanto del grupo control como del grupo tratamiento, en cuanto a la correlación, es homogénea. Con respecto al gráfico Heatmap, Figura 5b, se puede observar que mejoro la homogeneidad de distancias dentro de cada cluster.



a) Matriz de correlación b) Heatmap

Figura 5

2.2.3. SD vs Mean

La Figura 6 indica que se redujo la dependencia de la desviación estándar frente a la media en comparación con los datos sin normalizar. Sin embargo, es necesario realizar el filtro de genes de baja expresión.

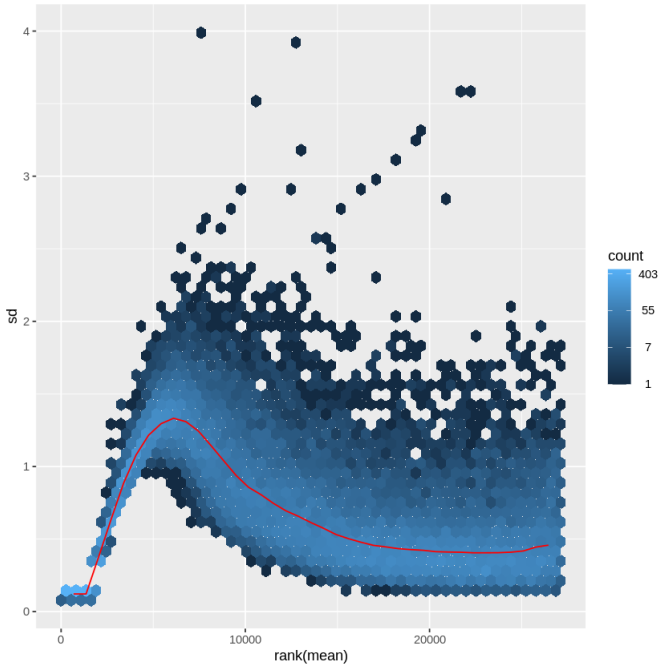


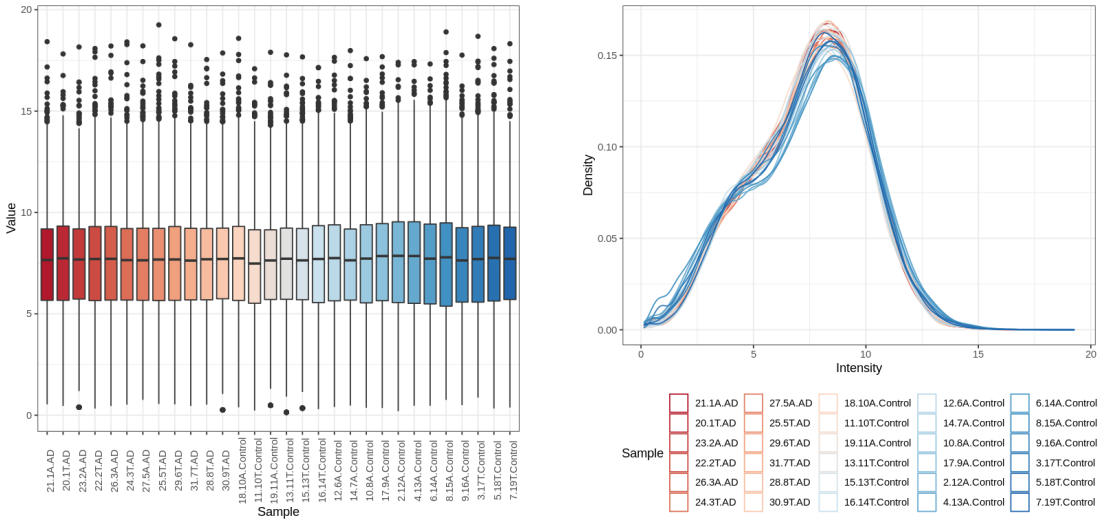
Figura 6: Desviaciones estándar de las filas frente a las medias de las filas

2.3. Filtrado de genes

Se realiza el procedimiento de filtro de genes de baja expresión para estabilizar la varianza.

2.3.1. Boxplot y densidad

Al realizar los boxplot con los datos que fueron filtrados por genes de baja expresión y normalizados, Figura 7a, se puede observar que hay mas datos atípicos en comparación al boxplot con los datos solo normalizados, Figura 4a. Sin embargo, con los datos filtrados se corrige mejor la simetría en la distribución como se puede observar en el grafico de la densidad, Figura 7b. Por lo tanto, se puede concluir que no existen muestras defectuosas.

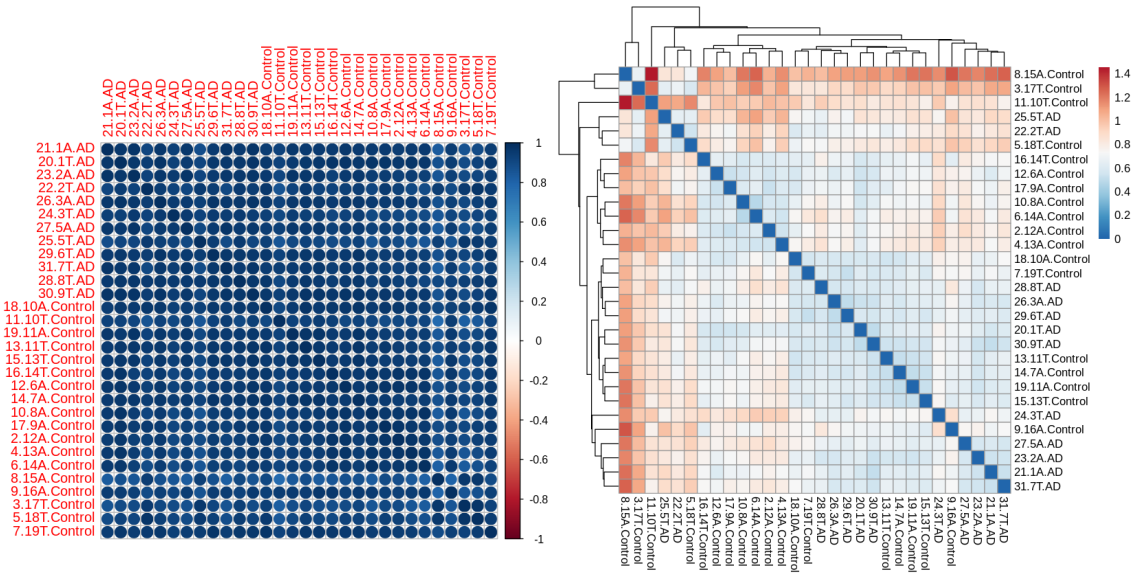


a) Box-Whisker b) Densidad de valor de expresión

Figura 7

2.3.2. Correlación y Heatmap

Al analizar el gráfico de la matriz de correlación, Figura 8a, con los datos filtrados y normalizados, se puede concluir fácilmente que las muestras tanto del grupo control como del grupo tratamiento, en cuanto a la correlación, es homogénea y superior a 0.85, con algunas pocas excepción cuya correlación varia entre 0.80 y 0.85. Respecto al gráfico Heatmap, Figura 8b, se puede observar que se mantiene la misma homogeneidad de distancias dentro de cada cluster.



a) Matriz de correlación b) Heatmap

Figura 8

2.4. SD vs Mean

Para el caso de los datos filtrados por genes de baja expresión, el diagrama SD vs Mean, Figura 9, indica que se redujo la dependencia de la desviación estándar frente a la media en comparación con los datos sin normalizar y con los datos normalizados sin filtro de genes.

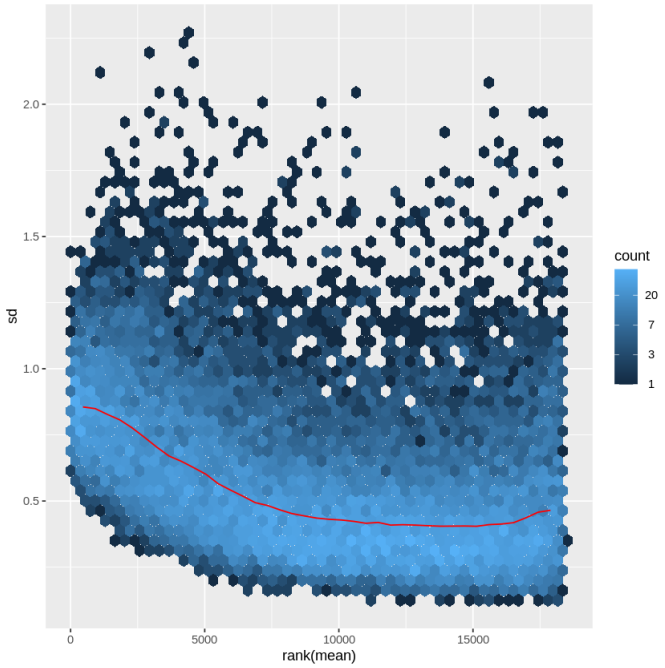


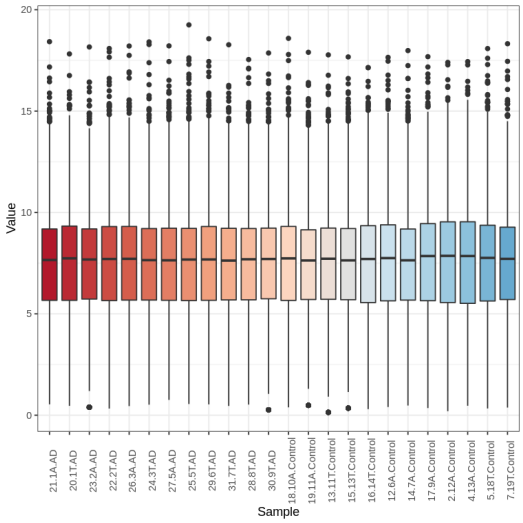
Figura 9: Desviaciones estándar de las filas frente a las medias de las filas

2.5. Filtrado de muestras

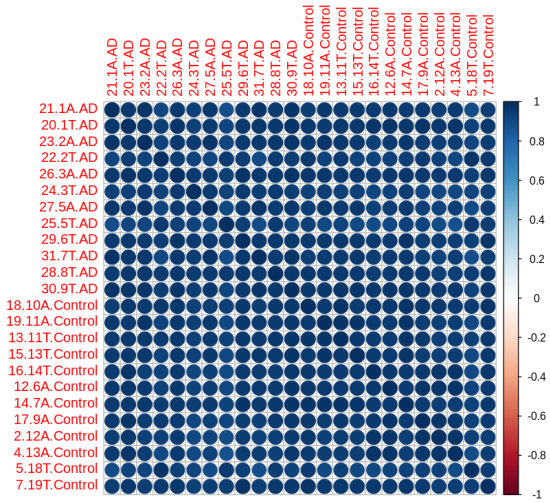
Se tomo la matriz de correlación de los datos filtrados por genes de baja expresión y posteriormente normalizados, luego se detectaron seis muestras control cuyas correlaciones eran inferior a 0.839 y se eliminaron. Esto para tener un grupo de 12 muestras tratamiento y 12 muestras control. Se debe aclarar que se eliminan las seis muestras con el fin de realizar el MA plot y utilizar el método SAM, y no porque sean muestras defectuosas.

2.5.1. Boxplot y correlación

De los gráficos con los datos de filtro de muestras, Figura 10, se infiere que la homogeneidad en cuento a la correlación de las muestras es mucho mejor. Sin embargo, la distribución de los datos se mantuvo igual.



a) Box-Whisker



b) Matriz de correlación

Figura 10

2.5.2. SD vs Mean

Del gráfico Figura 11, se concluye que la estabilidad de la varianza permanece igual que con los datos de filtrado por genes de baja expresión.

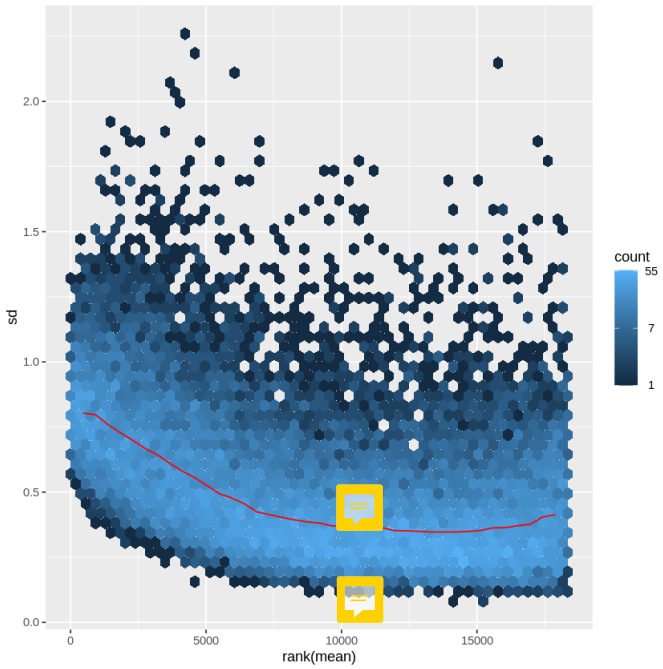


Figura 11: Desviaciones estándar de las filas frente a las medias de las filas

3. Diagrama MA plot

Los gráficos MA permiten la visualización de la variación de las relaciones de expresión génica ($M = log_2[R/G]$) en función de la intensidad de la señal promedio ($A = log_2\sqrt{[R * G]}$). Es decir, estos gráficos representa el logaritmo de los fold change frente a la expresión media entre dos condiciones. Los puntos de datos con valores extremos a lo largo del eje y representan los genes que tienen niveles de expresión altamente diferenciales (aunque no necesariamente expresados diferencialmente).

Para este trabajo se va suponer que $S = \{(x, y) : -2 > y\} \cap \{(x, y) : 2 < y\}$, donde cada punto (x, y) representa un gen y S el umbral de significancia. Al realizar el diagrama MA plot con los respecto datos, Figura 12, se detectaron 8 genes de expresión altamente diferenciales, 3 estaban regulados al alza y 5 estaban regulados a la baja.

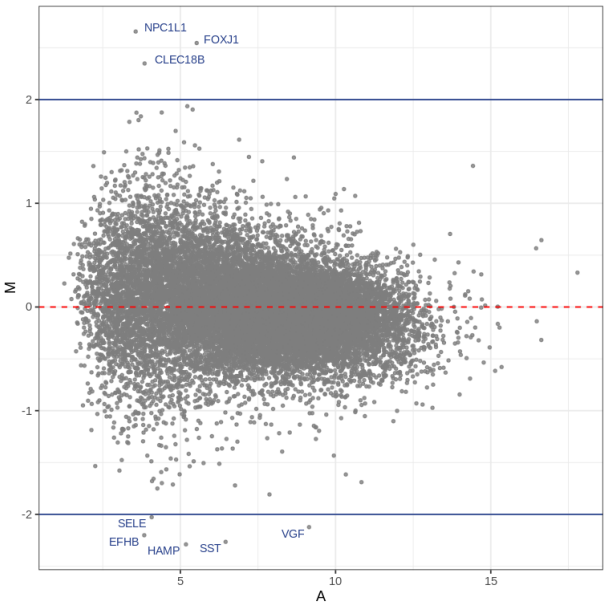


Figura 12: MA plot

El Cuadro 1 y Cuadro 2, muestra los genes de expresión altamente diferenciales con el símbolo de la proteína que codifica el gen y la función relacionada. La información fue obtenida de la base de datos de *GeneCards: The Human Gene Database*. La base *GeneCards* es una base de datos de genes humanos que proporciona información genómica, proteómica, transcriptómica, genética y funcional sobre todos los genes humanos conocidos y predichos.

Gen	Proteína	Función relacionada
NPC1L1	Q9UHC9-NPCL1_HUMAN	Transportador intracelular de colesterol 1, Enfermedad de Niemann-Pick, tipo C1
FOXJ1	Q92949-FOXJ1_HUMAN	Este gen codifica un miembro de la familia de factores de transcripción Forkhead
CLEC18B	Q6UXF7-CL18B_HUMAN	Este gen incluyen en la unión a carbohidratos

Cuadro 1: Genes regulados al alza

Gen	Proteína	Función relacionada
SELE	P16581-LYAM2_HUMAN	La proteína codificada por este gen se encuentra en las células endoteliales estimuladas por citoquinas y se cree que es responsable de la acumulación de leucocitos sanguíneos en los sitios de inflamación al mediar la adhesión de las células al revestimiento vascular
EFHB	Q8N7U6-EFHB_HUMAN	Este gen incluyen en la unión de iones de calcio
HAMP	P81172-HEPC_HUMAN	El producto codificado por este gen está implicado en el mantenimiento de la homeostasis del hierro, y es necesario para la regulación del almacenamiento de hierro en los macrófagos, y para la absorción intestinal de hierro
SST	P61278-SMS_HUMAN	La hormona somatostatina tiene formas activas de 14 aa y 28 aa que se producen por escisión alternativa de la única preproteína codificada por este gen. La somatostatina se expresa en todo el organismo e inhibe la liberación de numerosas hormonas secundarias al unirse a receptores de somatostatina acoplados a proteínas G de alta afinidad
VGF	O15240-VGF_HUMAN	Este gen se expresa específicamente en una subpoblación de células neuroendocrinas y es regulado por el factor de crecimiento nervioso

Cuadro 2: Genes regulados a la baja

4. Método SAM

SAM es un método no paramétrico, basado en la permutación, propuesto especialmente para el análisis de datos de microarrays. Calcula la tasa empírica de falsos positivos (FDR) mediante la permutación aleatoria de las etiquetas de clase. La permutación genera una distribución nula, porque se supone que la aleatoriedad elimina todos los efectos biológicos. Por lo tanto, proporciona un medio para controlar los falsos positivos bajo varios umbrales cuando se ensayan múltiples genes simultáneamente en un array.

En este trabajo se utilizo el método SAM con el fin detectar genes expresados diferencialmente (DEG), ya que es un método bastante robusto para variables continuas con dos condiciones biológicas. El Cuadro 3, muestra el resultado del análisis SAM para el caso de dos clases no emparejadas suponiendo varianzas desiguales y con un numero de permutaciones igual a 100.

Delta	**p0**	**False**	**Called**	**FDR**	**cutlow**	**cutup**	**j2**	**j1**
0.10	0.54	11745.55	14645.00	0.43	-0.58	0.25	5541.00	9245.00
0.50	0.54	3888.84	8879.00	0.24	-1.43	0.89	2767.00	12237.00
0.90	0.54	442.91	3217.00	0.07	-2.26	1.96	1112.00	16244.00
1.30	0.54	29.94	642.00	0.03	-3.15	3.06	288.00	17995.00
1.70	0.54	1.96	126.00	0.01	-4.08	3.96	52.00	18275.00
2.10	0.54	0.05	11.00	0.00	-Inf	4.93	0.00	18338.00
2.60	0.54	0.00	2.00	0.00	-Inf	6.57	0.00	18347.00
3.00	0.54	0.00	2.00	0.00	-Inf	6.57	0.00	18347.00
3.40	0.54	0.00	1.00	0.00	-Inf	7.41	0.00	18348.00
3.80	0.54	0.00	0.00	0.00	-Inf	Inf	0.00	18349.00

Cuadro 3

4.1. Valor critico suave

Se tomo un valor crítico de $\Delta = 1.3$, que corresponde a una tasa de falsos positivos (FDR) igual al 2.520 %. Con este criterio se detectaron 642 genes expresados diferencialmente, 354 estaban regulados al alza y 288 estaban regulados a la baja, el *cutlow* fue de -3.146 y *cutup* fue de 3.063, ver la Figura 13.

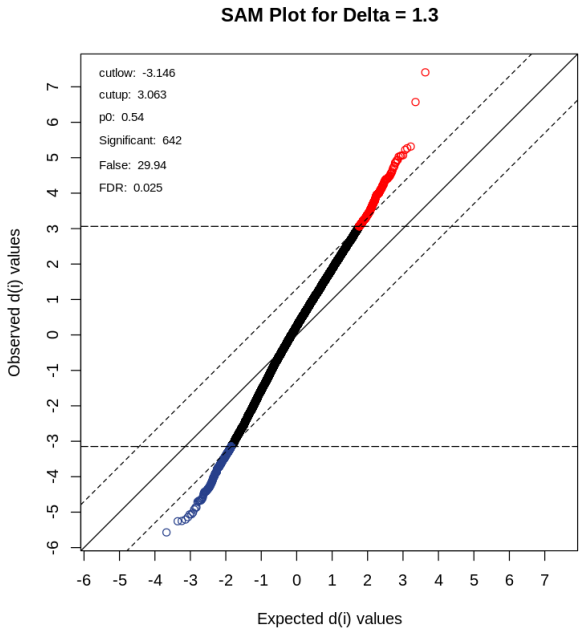


Figura 13: SAM plot

4.2. Valor critico estricto

Se tomo un valor crítico de $\Delta = 2.1$, que corresponde a una tasa de falsos positivos (FDR) igual al 0.246 %. Con este criterio se detectaron 11 genes expresados diferencialmente, 11 estaban regulados al alza y 0 estaban regulados a la baja, el *cutup* fue de 4.925, ver Figura 14.

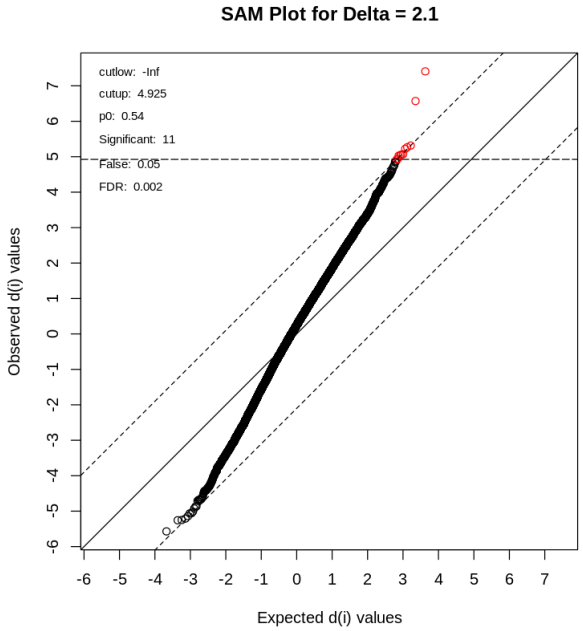


Figura 14: SAM plot

El Cuadro 4, muestra el resultado del análisis SAM para $\Delta = 2.1$ en el caso de dos clases no emparejadas suponiendo varianzas desiguales y con un numero de permutaciones igual a 100. La columna *Name* del Cuadro 4 indica los nombres de los DEGs.

Row	d.value	stdev	rawp	q.value	R.fold	Name
7328	7.41	0.2559	0	0	4.36	VGF
11356	6.57	0.2265	0	0	3.23	RPH3A
4421	5.32	0.1091	4.91e-06	0.00473	1.67	ABCA11P
15215	5.27	0.1647	6.00e-06	0.00473	2.04	DHRS11
7329	5.22	0.2439	6.54e-06	0.00473	2.70	NAT16
16351	5.07	0.2180	8.72e-06	0.00473	2.40	LOC101928238
8968	5.06	0.3077	8.72e-06	0.00473	3.28	PPEF1
3906	5.05	0.1898	8.72e-06	0.00473	2.16	FAM86HP
14636	5.03	0.2212	8.72e-06	0.00473	2.40	LYRM9
8948	4.94	0.1407	9.27e-06	0.00473	1.80	F8
15117	4.93	0.0979	1.04e-05	0.00473	1.55	DHRS7B

Cuadro 4

El Cuadro 5, muestra los primeros cinco DEGs regulados al alza con el símbolo de la proteína que codifica el gen y la función relacionada. La información fue obtenida de la base de datos de *GeneCards: The Human Gene Database*.

Gen	Proteína	Función relacionada
VGF	O15240-VGF_HUMAN	Este gen se expresa específicamente en una subpoblación de células neuroendocrinas y es regulado por el factor de crecimiento nervioso
RPH3A	Q9Y2J0-RP3A_HUMAN	Se cree que la proteína codificada por este gen es un efector de RAB3A, que es una pequeña proteína G que actúa en las últimas etapas de la exocitosis de los neurotransmisores. La proteína codificada puede estar implicada en la liberación de neurotransmisores y en el tráfico de vesículas sinápticas.
ABCA11P	Q4W5N1-ABCAB_HUMAN	Este gen incluyen la actividad transportadora y la actividad ATPasa
DHRS11	Q6UWP2-DHR11_HUMAN	Las enfermedades asociadas con DHRS11 incluyen en la orientación sexual ego-distónica y enfermedad de Alzheimer 12. Entre sus vías relacionadas se encuentran la biosíntesis de hormonas esteroides.
NAT16	Q8N8M0-NAT16_HUMAN	Las enfermedades asociadas con NAT16 incluyen la enfermedad de Von Willebrand, tipo 2 y la enfermedad de Von Willebrand, tipo 1. Este gen incluyen la actividad de la N-acetiltransferasa y la actividad del péptido alfa-N-acetiltransferasa.

Cuadro 5: Genes regulados al alza

5. Referencias

Erich Neuwirth (2014). RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Hadley Wickham (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1-20. URL http://www.jstatsoft.org/v21/i12/.

Holger Schwender (2020). siggenes: Multiple Testing using SAM and Efron's Empirical Bayes Approaches. R package version 1.64.0.

Nativio R, Lan Y, Donahue G, Sidoli S et al. An integrated multi-omics approach identifies epigenetic alterations associated with Alzheimer's disease. Nat Genet 2020 Oct;52(10):1024-1035. PMID: 32989324

R. Gentleman, V. Carey, W. Huber and F. Hahne (2021). genefilter: genefilter: methods for filtering genes from high-throughput experiments. R package version 1.72.1.

Raivo Kolde (2019). pheatmap: Pretty Heatmaps. R package version 1.0.12. https://CRAN.R-project.org/package=pheatmap

Taiyun Wei and Viliam Simko (2017). R package çorrplot": Visualization of a Correlation Matrix (Version 0.84). Available from https://github.com/taiyun/corrplot

Wolfgang Huber, Anja von Heydebreck, Holger Sueltmann, Annemarie Poustka and Martin Vingron. Variance Stabilization Applied to Microarray Data Calibration and to the Quantification of Differential Expression. Bioinformatics 18, S96-S104 (2002).

6. Código en R

```
library(ggplot2)
library(RColorBrewer)
library(genefilter)
library(vsn)
library(reshape2)
library(corrplot)
library(pheatmap)
library(siggenes)

############################## LECTURA DE DATOS ################################

RNADATA <- read.table("GSE159699_summary_count.star.txt", h=T)
```

```
# SUMMARY RNADATA
class(RNADATA)
# colnames(RNADATA)

# SUMMARY RNAMATRIX
RNAMTRX <- as.matrix(RNADATA[,2:31])
class(RNAMTRX)

# CLEANING DATA
col_names <- colnames(RNAMTRX)
col_names <- gsub("X", "", col_names)
col_names <- gsub("Young", "Control", col_names)
col_names <- gsub("Old", "Control", col_names)
colnames(RNAMTRX) <- col_names
RNAMTRX_melt <- melt(RNAMTRX)[,-1]
colnames(RNAMTRX_melt) <- c("Sample", "Value")
rownames(RNAMTRX) <- RNADATA$refGene

# COLORS
nb.cols <- dim(RNAMTRX)[2]
mycolors <- colorRampPalette(brewer.pal(8, "RdBu"))(nb.cols)

#SUMMARY
dim(RNAMTRX)
dim(RNAMTRX_melt)

############################## CONTROL DE CALIDAD ################################

# BOXPLOT
ggplot(RNAMTRX_melt, aes(Sample, Value, fill=Sample)) +
  geom_boxplot() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90),
        legend.position = "none") +
  scale_fill_manual(values = mycolors) +
  theme(plot.margin = unit(c(1,1,1,1), "cm"))

# VIOLIN
ggplot(RNAMTRX_melt, aes(Sample, Value, fill=Sample)) +
  geom_violin() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90),
        legend.position = "none") +
  scale_fill_manual(values = mycolors) +
  theme(plot.margin = unit(c(1,1,1,1), "cm"))

# DENSIDAD
ggplot(RNAMTRX_melt, aes(x=Value)) +
  theme_bw() +
```

```
  theme(legend.position="bottom") +
  geom_density(aes(group=Sample, colour=Sample)) +
  labs(x="Intensity", y="Density") +
  scale_color_manual(values = mycolors) +
  theme(plot.margin = unit(c(1,1,1,1), "cm"))

# CORRELOGRAMA
corrplot(cor(RNAMTRX))

# DENDROGRAMA
dists <- as.matrix(as.dist(dist2(RNAMTRX)))
hmcol <- rev(colorRampPalette(RColorBrewer::brewer.pal(9,
 "RdBu"))(255))
pheatmap(dists, col = (hmcol))

# SD VS MEAN
meanSdPlot(RNAMTRX, ranks=TRUE)

############################## NORMALIZACION ####################################

# NORMALIZACION
RNAMTRXnorm = justvsn(RNAMTRX)
RNAMTRXnorm_melt = melt(RNAMTRXnorm)[,-1]
colnames(RNAMTRXnorm_melt) = c("Sample", "Value")

# BOXPLOT
ggplot(RNAMTRXnorm_melt, aes(Sample, Value, fill=Sample)) +
  geom_boxplot() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90),
        legend.position = "none") +
  scale_fill_manual(values = mycolors) +
  theme(plot.margin = unit(c(1,1,1,1), "cm"))

# VIOLIN
ggplot(RNAMTRXnorm_melt, aes(Sample, Value, fill=Sample)) +
  geom_violin() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90),
        legend.position = "none") +
  scale_fill_manual(values = mycolors) +
  theme(plot.margin = unit(c(1,1,1,1), "cm"))

# DENSIDAD
ggplot(RNAMTRXnorm_melt, aes(x=Value)) +
  theme_bw() +
  theme(legend.position="bottom") +
  geom_density(aes(group=Sample, colour=Sample)) +
  labs(x="Intensity", y="Density") +
```

```
  scale_color_manual(values = mycolors) +
  theme(plot.margin = unit(c(1,1,1,1), "cm"))

# CORRELOGRAMA
corrplot(cor(RNAMTRXnorm))

# DENDROGRAMA
dists <- as.matrix(as.dist(dist2(RNAMTRXnorm)))
hmcol <- rev(colorRampPalette(RColorBrewer::brewer.pal(9, "RdBu"))(255))
pheatmap(dists, col = (hmcol))

# SD VS MEAN
meanSdPlot(RNAMTRXnorm, ranks=TRUE)

############################## FILTRO DE GENES ###################################

# FILTRO DE GENES CON BAJA EXPRESION
select = (0==rowSums(RNAMTRX<=0))
RNAMTRXnormfilt = RNAMTRXnorm[select,]
RNAMTRXnormfilt_melt = melt(RNAMTRXnormfilt)[,-1]
colnames(RNAMTRXnormfilt_melt) = c("Sample", "Value")

# BOXPLOT
ggplot(RNAMTRXnormfilt_melt, aes(Sample, Value, fill=Sample)) +
  geom_boxplot() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90),
        legend.position = "none") +
  scale_fill_manual(values = mycolors) +
  theme(plot.margin = unit(c(1,1,1,1), "cm"))

# VIOLIN
ggplot(RNAMTRXnormfilt_melt, aes(Sample, Value, fill=Sample)) +
  geom_violin() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90),
        legend.position = "none") +
  scale_fill_manual(values = mycolors) +
  theme(plot.margin = unit(c(1,1,1,1), "cm"))

# DENSIDAD
ggplot(RNAMTRXnormfilt_melt, aes(x=Value)) +
  theme_bw() +
  theme(legend.position="bottom") +
  geom_density(aes(group=Sample, colour=Sample)) +
  labs(x="Intensity", y="Density") +
  scale_color_manual(values = mycolors) +
  theme(plot.margin = unit(c(1,1,1,1), "cm"))
```

```
# CORRELOGRAMA
corrplot(cor(RNAMTRXnormfilt))

# DENDROGRAMA
dists <- as.matrix(as.dist(dist2(RNAMTRXnormfilt)))
hmcol <- rev(colorRampPalette(RColorBrewer::brewer.pal(9, "RdBu"))(255))
pheatmap(dists, col = (hmcol))

# SD VS MEAN
meanSdPlot(RNAMTRXnormfilt, ranks=TRUE)

############################## FILTRO DE MUESTRAS ##############################

# FILTRO DE MUESTRAS
vector_sample = NULL
for (i in 13:30) {
  results = all(cor(RNAMTRXnormfilt)[,i] >= 0.839)
  if (results == FALSE) {
    vector_sample[i] = i
  }
}

RNAMTRXnormfilt2 = RNAMTRXnormfilt[, -as.vector(na.omit(vector_sample))]
RNAMTRXnormfilt2_melt = melt(RNAMTRXnormfilt2)[,-1]
colnames(RNAMTRXnormfilt2_melt) = c("Sample", "Value")

# BOXPLOT
ggplot(RNAMTRXnormfilt2_melt, aes(Sample, Value, fill=Sample)) +
  geom_boxplot() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90),
        legend.position = "none") +
  scale_fill_manual(values = mycolors) +
  theme(plot.margin = unit(c(1,1,1,1), "cm"))

# VIOLIN
ggplot(RNAMTRXnormfilt2_melt, aes(Sample, Value, fill=Sample)) +
  geom_violin() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90),
        legend.position = "none") +
  scale_fill_manual(values = mycolors)+
  theme(plot.margin = unit(c(1,1,1,1), "cm"))

# DENSIDAD
ggplot(RNAMTRXnormfilt2_melt, aes(x=Value)) +
   theme_bw() +
  theme(legend.position="bottom") +
  geom_density(aes(group=Sample, colour=Sample)) +
```

```
  labs(x="Intensity", y="Density") +
  scale_color_manual(values = mycolors)+
  theme(plot.margin = unit(c(1,1,1,1), "cm"))

# CORRELOGRAMA
corrplot(cor(RNAMTRXnormfilt2))

# DENDROGRAMA
dists <- as.matrix(as.dist(dist2(RNAMTRXnormfilt2)))
hmcol <- rev(colorRampPalette(RColorBrewer::brewer.pal(9, "RdBu"))(255))
pheatmap(dists, col = (hmcol))

# SD VS MEAN
meanSdPlot(RNAMTRXnormfilt2, ranks=TRUE)

########################### GENES EXPRESADOS DIFERENCIALMENTE ##################

# FOLD CHANGE
dim(RNAMTRXnormfilt2)
AD = RNAMTRXnormfilt2[,1:12]
CONTROL = RNAMTRXnormfilt2[,13:24]
TC = AD - CONTROL
TC2 = (AD + CONTROL)/2
dim(TC)
dim(TC2)

# MA
MA_data = data.frame(A = rowMeans(TC2), M = rowMeans(TC),
Genes = rownames(RNAMTRXnormfilt2))

# MA plot
ggplot(MA_data, aes(x = A, y = M, label = Genes)) +
  theme_bw() +
  geom_point(alpha = 0.8, shape = 20, color="gray50") +
  geom_hline(yintercept=0, linetype="dashed", color = "red") +
  geom_hline(yintercept=2, color = "royalblue4") +
  geom_hline(yintercept=-2, color = "royalblue4") +
  theme(plot.margin = unit(c(1,1,1,1), "cm")) +
  geom_text(aes(label=ifelse(M > 2, as.character(Genes),'')),
   hjust=-0.2, vjust=0, size = 3, color = "royalblue4") +
  geom_text(aes(label=ifelse(M < -2, as.character(Genes),'')),
   hjust=1.2, vjust=1.3, size = 3, color = "royalblue4")

# SAM
# class(RNAMTRXnormfilt2)
# dim(RNAMTRXnormfilt2)
vector_cl = rep(c(0,1), each = 12)
sam.out <- sam(RNAMTRXnormfilt2, vector_cl, rand = 123,
  gene.names = rownames(RNAMTRXnormfilt2))
```

```
sam.out

summary(sam.out)

plot(sam.out, 2.1)

summary(sam.out, 2.1)
```