Análisis comparativo de biología tumoral y de supervivencia en pacientes con adenocarcinoma con y sin historial de tabaquismo

Esther Guinart
Master Bioinformática y Bioestadística
TFM 22 - Estudio genómico del cáncer

Nombre Consultor/a
Laia Bassaganyas

Fecha Entrega
5 junio 2018

FICHA DEL TRABAJO FINAL

Título del trabajo:	Análisis comparativo de biología tumoral y de supervivencia en pacientes con adenocarcinoma pulmonar, con y sin historial de tabaquismo
Nombre del autor:	*Esther Guinart Fernández*
Nombre del consultor/a:	*Laia Bassaganyas Bars*
Nombre del PRA:	*Jose Antonio Morán Moreno*
Fecha de entrega (mm/aaaa):	05/06/2018
Titulación::	*Máster universitario en Bioinformática y bioestadística*
Área del Trabajo Final:	*Estudio genómico del cáncer*
Idioma del trabajo:	*Castellano*
Palabras clave	NSCLC, Biología tumoral, Supervivencia, historial de tabaquismo

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

El objetivo de este trabajo es estudiar la biología tumoral del adenocarcinoma pulmonar (un subtipo de cáncer de pulmón de células no pequeñas) en pacientes con y sin historial de tabaquismo.
Mucho es lo que se ha investigado y escrito sobre este tipo de tumor, sobretodo porque se trata de una de las principales causas de muerte a nivel mundial.
Se ha comprobado que uno de los orígenes más frecuentes de esta patología es el tabaquismo. Sin embargo, la incidencia de esta enfermedad va en aumento en personas sin historial de tabaquismo, primordialmente en países desarrollados, puesto que el 30% de los casos diagnosticados se dan en este tipo de entornos.
Por ello, este trabajo pretende investigar si hay diferencias entre un grupo de control con antecedentes de tabaquismo frente a otro sin historial.

- Se estudiará si la biología tumoral del adenocarcinoma pulmonar es diferente en ambos grupos de control: genes sobreexpresados y bajoregulados.
- Se realizará un análisis de supervivencia para ambos grupos.

Todos los objetivos anteriormente mencionados, se procesarán y analizarán mediante un pipeline analysis con R/Bioconductor.
Los resultados obtenidos tras esta investigación podrían ayudar a vislumbrar si los tratamientos administrados en el adenocarcinoma pulmonar sin historial de tabaquismo son válidos o deben estudiarse nuevas dianas terapéuticas y

biomarcadores dadas las características diferenciales de la patología en este grupo de pacientes.

Abstract (in English, 250 words or less):

The goal of this paper is the study of the tumor biology of the lung adenocarcinoma (a subtype of non-small cell lung cancer) in patients with and without smoking history.

A lot of research has been made with regards to this type of tumor, above all because it is one of the main causes of death worldwide.

There is evidence supporting that one of the most common sources of this cancer is the smoking history. However, this pathology is growing in patients without smoking history, primarily in developed countries, since 30% of new cases occur in such scenarios.

This is the reason why this paper intends to find out if there are differences among both groups (one with and the other without smoking history).

- We will try to find out if the tumor biology in lung adenocarcinoma is different in both control groups: overexpressed vs dowregulated genes.
- We will perform a survival analysis across groups.

Both topics above will be processed and analyzed by a pipeline analysis with R/Bioconductor.

The results obtained after this investigation may trigger a need to change the current medical treatment for adenocarcinoma patients with non-smoking history or whether new targets and biomarkers are required given the differences across groups.

Índice

Lista de figuras

Lista de tablas

1. Introducción

1.1 Contexto y justificación del trabajo

El cáncer de pulmón es el tipo cáncer más frecuente a nivel mundial y supone la primera causa de muerte en pacientes adultos [1]. Más de una quinta parte de las muertes de cáncer se deben a esta patología.
Según la histología del cáncer, existen dos grandes tipos de clasificación:

- Cáncer de pulmón de célula pequeña (SCLC): 15% de los casos.
- Cáncer de pulmón de célula no pequeña (NSCLC): 85% de los casos. Bajo este tipo, existen tres subclasificaciones más:
  - Carcinoma de célula escamosa: 25%-30% de todos los casos. Se origina en las células que recubren las vías respiratorias.
  - Adenocarcinoma (ADC): 40% de todos los casos. Se origina en las células que producen moco.
  - Carcinoma de célula grande: 10-15% de todos los casos. Se origina en células que no pertenecen a ninguno de los dos tipos anteriores.

"Evitar el consumo de tabaco es una de las medidas más eficaces para reducir la incidencia de esta enfermedad, y así se está evidenciando en los países desarrollados donde se ha invertido en campañas contra el consumo de tabaco" [2]. Sin embargo, en los últimos años se está observando que el cáncer de pulmón ha aumentado entre personas sin historial de tabaquismo. La polución es uno de los factores que incrementa el riesgo de cáncer de pulmón. De hecho, "las poblaciones urbanas donde existen altos niveles de sustancias contaminantes desarrollan cáncer de pulmón en mayor proporción que las poblaciones rurales" [1]:

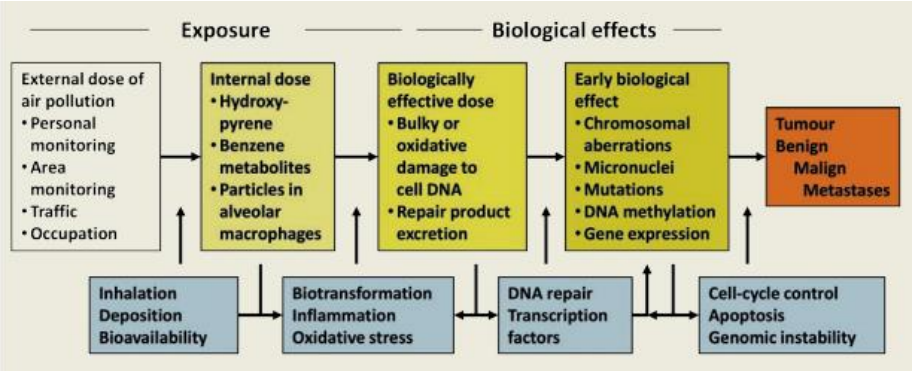


Figura 1. Marcadores biológicos de exposición y efecto de la polución atmosférica - Extraído de Stewart B. W. and Wild C.P. World Cancer Report 2014*, International Agency for Research on Cancer, Lyon, France, 2015*

El gasto sanitario que supone esta enfermedad es significativo para los gobiernos, dada la alta incidencia en la población y el gasto medio por paciente.

	France		Germany		England	
	Year 1	Year 2	Year 1	Year 2	Year 1	Year 2
Hospital in-patient	€11,667	€5916	€11,363	€6568	€5985	€1156
Hospital out-patient	€2313	€676	€1925	€1766	€1209	€834
Medicines	€3542	€321	€4488	€3805	€8593	
Other	€502	€126	€1429	€1156		
Total	€18,024	€7039	€19,205	€13,295	€15,787	€1990
2-year Total		€25,063		€32,500		€17,777

Figura 2. Gasto medio en Francia, Alemania e Inglaterra durante seguimiento de dos años. Extráido de Le Chevalier, T., Non-small cell lung cancer: the challenges of the next decade*. Frontiers in Oncology, art. 29, 1-4, vol. 1, 2011*

Este gasto está ocurriendo dentro de un contexto sanitario donde se están dando cambios sin precedentes:

- Factores demográficos: población más envejecida, más concentrada en zonas urbanas, más informada/exigente.
- Factores financieros: ahorro en gasto sanitario, necesidad de dar mejores servicios a pacientes a menor coste, presencia de más actores en la escena sanitaria con más fuerza decisoria y con intereses diferentes.

Por ello, es cada vez más importante dirigir los esfuerzos a una medicina (diagnóstico y tratamiento) personalizada, específica para la biología molecular de cada patología. Esta especialización está directamente vinculada a la eficacia del sistema sanitario, al ahorro de costes y a la satisfacción de las diferentes demandas de los actores implicados en el sistema.

Este trabajo pretende realizar un estudio sobre el adenocarcinoma, por ser el tipo más frecuente de cáncer de pulmón. Queremos analizar si este tipo de patología es diferente entre pacientes con historial y sin historial de tabaquismo Si se confirman estas diferencias, existiría la posibilidad de detectar biomarcadores y potenciales moléculas diana para diagnosticar y tratar de forma diferente a los pacientes de cada grupo.
Asimismo realizaremos un análisis de supervivencia para averiguar si el pronóstico es diferente en ambos grupos. De esta forma se podrían investigar qué covariables en cada grupo contribuyen al pronóstico diferencial.
Los estudios de este trabajo se realizan sobre una cohorte de 35 individuos.

1.2 Objetivos del Trabajo

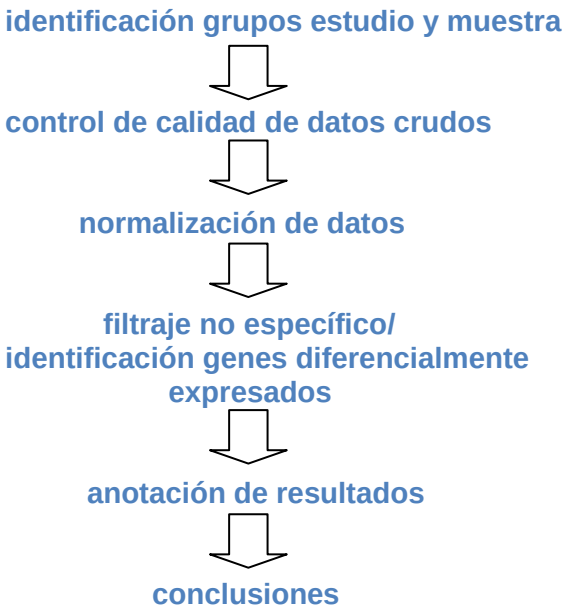
Este proyecto pretende determinar si las características tumorales del adenocarcinoma pulmonar son diferentes entre pacientes con y sin historial de tabaquismo. Para ello, se presenten conseguir los objetivos siguientes:

- Identificar si las diferentes expresiones génicas son diferentes entre ambos grupos
    - Determinar qué genes se expresan diferencialmente entre ambos grupos
    - Identificar los genes característicos en la patología de no fumadores (firmas diferentes entre grupos)

- Realizar un análisis de supervivencia para definir las diferencias de pronóstico entre ambos grupos
    - Identificar si la probabilidad de supervivencia en adenocarcinoma es mayor en pacientes sin historial de tabaquismo

1.3 Enfoque y método seguido

Para este estudio se utilizarán datos provenientes de http://www.cbioportal.org/.
Se proporcionan datos de secuenciación del exoma de adenocarcinoma tratados con pembrolizumab para una cohorte de 35 individuos.
Inicialmente se había previsto realizar el estudio sobre una cohorte de 1142 individuos y así obtener mayor poder estadístico. La complejidad de manejo de este tamaño muestral, las limitaciones temporales y los problemas de rendimiento de R nos llevaron a optar por una cohorte bastante más reducida.
La fase 1 de este trabajo consiste en identificar la expresión génica de cada grupo.
El pipeline seguirá el patrón a continuación:



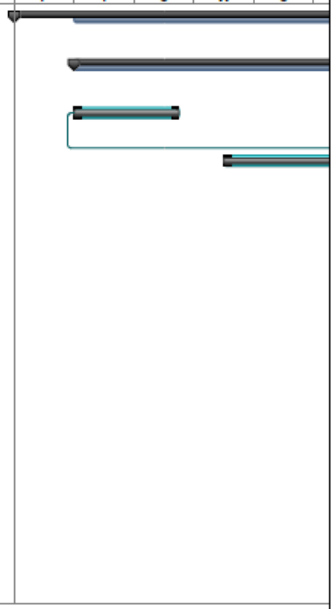
Utilizaremos el paquete de software Bioconductor para integrar metadatos biológicos durante la fase de enriquecimiento de nuestro pipeline analysis.
Se aplicará el método de la inferencia estadística en el tratamiento de los datos (contraste de hipótesis, p-valores, test t) y ANOVA.

En la fase 2 del presente trabajo, se analiza si la supervivencia entre ambos grupos es diferente en función de la variable historial de tabaquismo.
Los datos se modelarán mediante análisis de supervivencia por medio de los métodos siguientes: Kaplan Meier, Cox y exponencial de Weibull.

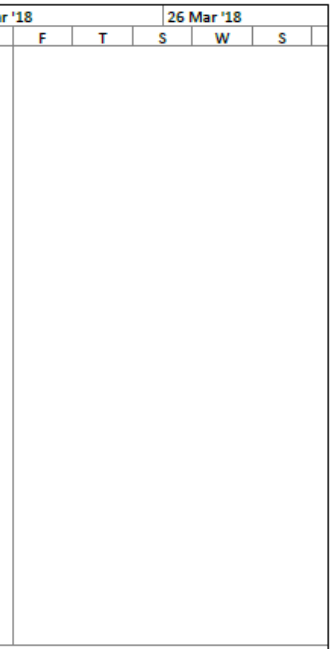
1.4 Planificación del Trabajo

El presente diagrama de Gantt muestra la planificación integral del trabajo de fin de máster. Desglosa los objetivos en tareas y les asigna una duración que finaliza en el hito de entrega de la práctica.

ID	Task Mode	Task Name	Duration	Start	Finish
1		**TFM - Características genómicas NSCLC en grupos con y sin historial tabaquismo**	**72 days**	**Fri 16/03/18**	**Mon 25/06/18**
2		**Desarrollo Fase 1 - Identificación de expresión genómica diferencial entre grupos**	**25 days**	**Tue 20/03/18**	**Mon 23/04/18**
3		Recopilación de datos previos a realización análisis	5 days	Tue 20/03/18	Mon 26/03/18
4		Estudio de los datos genómicos	6 days	Fri 30/03/18	Fri 06/04/18
5		Pipeline analysis	4 days	Sat 07/04/18	Wed 11/04/18
6		Identificación de grupos, CC datos crudos, normalización datos	2 days	Thu 12/04/18	Fri 13/04/18
7		Filtraje e identficación de genes (diferencialmente expresados > SNPs, CNV)	2 days	Sat 14/04/18	Mon 16/04/18
8		Anotación de resultados	2 days	Tue 17/04/18	Wed 18/04/18
9		Comparación entre comparaciones	2 days	Thu 19/04/18	Fri 20/04/18
10		~~Gene enrichment analysis~~	3 days	Fri 13/04/18	Tue 17/04/18
11		Conclusiones	1 day	Thu 19/04/18	Thu 19/04/18
12		Redacción final PEC 2	2 days	Fri 20/04/18	Mon 23/04/18
13		Entrega PEC 2	0 days	Mon 23/04/18	Mon 23/04/18



ID	Task Mode	Task Name	Duration	Start	Finish
14		**Desarrollo Fase 2 - Análisis supervivencia / pronóstico diferencial entre grupos**	**20 days**	**Tue 24/04/18**	**Mon 21/05/18**
15		identificacion de eventos a medir: remision gupos con trat	2 days	Tue 24/04/18	Wed 25/04/18
16		Preparación de datos: creación tablas de comparación par	3 days	Thu 26/04/18	Mon 30/04/18
17		Análisis función de supervivencia y failure rate en ambos grupos: cálculo de la funciones basicas de supervivencia: PDF, CDF, S(t), h(t)	3 days	Tue 01/05/18	Thu 03/05/18
18		Aplicación test long rank KM y evaluación de resultados	3 days	Thu 03/05/18	Mon 07/05/18
19		Aplicación modelo Cox	5 days	Tue 08/05/18	Mon 14/05/18
20		Aplicación modelo Weibull	2 days	Tue 15/05/18	Wed 16/05/18
21		Intepretacción de resultados: correlación de datos según historial tabaquismo y respuesta a tratamiento	3 days	Fri 11/05/18	Tue 15/05/18
22		Redacción final PEC 3	3 days	Wed 16/05/18	Fri 18/05/18
23		Entrega PEC 3	0 days	Mon 21/05/18	Mon 21/05/18
24		**Redacción memoria**	**11 days**	**Tue 22/05/18**	**Tue 05/06/18**
25		Redacción final capítulos PEC 2 y PEC 3	5 days	Tue 22/05/18	Mon 28/05/18



ID	Task Mode	Task Name	Duration	Start	Finish	12 Mar '18			26 Mar '18		
						M	F	T	S	W	S
26		Redacción conclusiones	3 days	Tue 29/05/18	Thu 31/05/18						
27		Redacción glosario y bibliografía	2 days	Fri 01/06/18	Mon 04/06/18						
28		Entrega PEC 4	0 days	Tue 05/06/18	Tue 05/06/18						
29		**Preparación presentación**	**6 days**	**Wed 06/06/18**	**Wed 13/06/18**						
30		Selección de contenidos relevantes	3 days	Wed 06/06/18	Fri 08/06/18						
31		Diseño PPT	2 days	Mon 11/06/18	Tue 12/06/18						
32		Entrega PEC 5	0 days	Wed 13/06/18	Wed 13/06/18						
33		**Defensa pública**	**8 days**	**Thu 14/06/18**	**Mon 25/06/18**						
34		Preparación defensa	5 days	Thu 14/06/18	Wed 20/06/18						
35		Defensa	0 days	Thu 21/06/18	Thu 21/06/18						

Asimismo se ha realizado un análisis de riesgos y se han concretado mitigaciones para cada uno de ellos:

Código	Descripción riesgo	Efectos	Mitigación del riesgo
PR1	Procesamiento ineficaz de los datos dado el gran volumen	Imposibilidad de realizar la práctica de forma eficiente	Reducir el alcance de datos para finalmente tratar una cantidad aceptable de datos que permitan cumplir los objetivos fijados
		Imposibilidad de obtener resultados fiables	
PR2	Ineficiencia en la selección óptima de servidores genómicos dada la extensa cantidad disponible	Imposibilidad de obtener resultados fiables	Uso de los servidores clásicos aprendidos durante el máster para consultar y enriquecer los datos obtenidos
PR3	Datos insuficientes necesarios para el análisis propuesto	Incumplimiento de objetivos de la práctica	Necesidad de buscar una muestra alternativa en otro portal de datos biológicos
PR4	Tiempo insuficiente para completar la práctica	Incumplimiento de objetivos de la práctica	Creación de un diagrama temporal que monitoriza el grado de cumplimiento de fechas e hitos

Tabla 1. Lista de riesgos

1.5 Breve sumario de productos obtenidos

Los entregables asociados a este trabajo de fin de máster son los siguientes:

Entregable	Nombre	Descripción
PEC 1	Plan de trabajo	Detalla los objetivos, planificación temporal de tareas y entregables previstos para el proyecto.
PEC2 **PEC 3**	Avance proyecto	Desarrolla las fases del proyecto, examina la consecución de los objetivos y de alcance.
PEC 4	Memoria	Muestra los métodos usados y los resultados obtenidos para cada uno de los objetivos desarrollados anteriormente.
PEC 5	Presentación	Presentación de los objetivos fijados, la metodología de análisis utilizada y los resultados. Un apartado final de conclusiones cerrará la presentación.
PEC 6	Defensa	Presentación de los resultados obtenidos y conclusiones

Tabla 2. Lista de entregables

1.6 Breve descripción de los otros capítulos de la memoria

Esta memoria contiene dos capítulos adicionales, destinados a presentar los resultados obtenidos al realizar los análisis de los dos objetivos generales propuestos.
Por ello, un capítulo estará focalizado sobre el estudio de la expresión génica en los dos grupos de interés y el otro capítulo describirá la comparativa de resultados en análisis de supervivencia.
Mediante el apartado *Conclusiones*, se valorarán los resultados obtenidos (posibles causas de esa patología en no fumadores y las potenciales aplicaciones de los resultados obtenidos).

2. Capítulo 1 – Análisis de expresión génica

2.1 Breve introducción a la ultra secuenciación y RNA-seq.

La ultrasecuenciación es una tecnología que permite obtener secuencias de cadenas de ADN a gran escala. Este fenómeno ha permitido reducir el coste y el tiempo de análisis de los fragmentos de ADN y ha abierto el camino a nuevas posibilidades de resolver problemas científicos.
La ultrasecuenciancián crea millones de secuencias cortas de nucleótidos, alineadas a otra referencia o genoma.
Con estos alineamientos se pueden inferir:

- niveles de expresión génica (RNA-seq),
- unión a elementos reguladores a localizaciones genómicas (ChIP-seq),
- prevalencia de variantes estructurales (SNPs, indels cortos, reorganización genómica a gran escala) [4].

Tal como hemos mencionado, “el RNA-seq permite realizar un estudio de la expresión génica de forma más precisa que otros métodos, puesto que cuantifica el producto de la expresión a la vez que obtiene información acerca de secuencias no identificadas previamente” [3].
Dado que uno de los objetivos de nuestro trabajo es identificar si hay diferencias de expresiones génicas entre grupos (fumadores/no fumadores), utilizaremos el método RNA-seq para probar esta hipótesis.

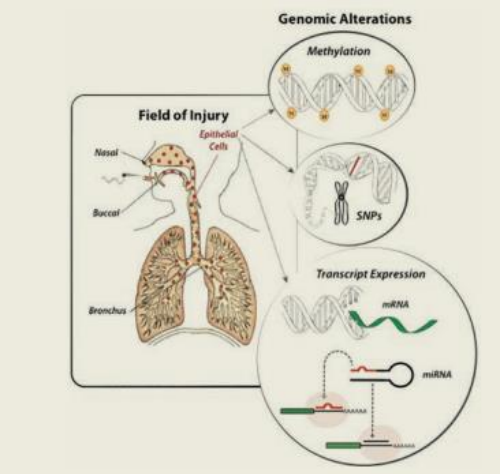


Figura 3. Zona molecular dañada [en fumadores] entre las células epiteliales y a través de las vías respiratorias. miRNA, microRNA; mRNA, messenger RNA; SNPs (single-nucleotide polymorphisms). Extraído de Stewart B. W. and Wild C.P. World Cancer Report 2014*, International Agency for Research on Cancer, Lyon, France, 2015*

2.2 Análisis - Introducción

Los datos se partida proviene de http://www.cbioportal.org/. Se extraen los datos Lung adenocarcinoa (MSKCC, 2015) de una cohorte de 35 pacientes donde se estudia la expresión de genes de muestras normales contra muestras tumorales con mutaciones somáticas de cáncer de pulmón.

Las variables clínicas sobre las que trabajaremos son las siguientes:

- Datos RNAseq:
    - t_ref_count: cantidad total de un gen particular
    - código ENSEMBL: código identificador del gen
    - Tumor_Sample_Barcode: código de la muestra secuenciada vinculada al paciente
- Datos clínicos:
    - Sample_ID: código de la muestra secuenciada vinculada al paciente
    - Smoking History: Yes/No: variable que indica si el paciente tiene o no historial como fumador.

Trabajaremos con R y con Bioconductor. Los paquetes utilizados serán:

- Lima
- EdgeR
- gPLots
- RColorBrewer
- Org.Hs.eg.db
- AnnotationsDbi

Los pasos que seguiremos en nuestro análisis pipeline de RNAseq se han detallado en el capítulo 1.3 Enfoque y método seguido.

2.3 Análisis – Preparación de datos y espacio de trabajo

En primer lugar se estudiaron los datos y las variables. Una vez seleccionados los datos con que se iba a trabajar, se configuraron el espacio de trabajo y se cargaron los paquetes de trabajo.

```
source("http://www.bioconductor.org/biocLite.R")

## Bioconductor version 3.6 (BiocInstaller 1.28.0), ?biocLite for help

biocLite()

## BioC_mirror: https://bioconductor.org

## Using Bioconductor 3.6 (BiocInstaller 1.28.0), R 3.4.4 (2018-03-
15).
```

Sólo serán necesarias tres columnas para nuestro análisis RNAseq:

- genes,
- muestras tumorales
- recuento de genes

Asimismo, se preparó el archivo de muestras vinculado al de recuento de genes por muestra.

Utilizamos un objeto countdata, que será el que, mediante la muestra, relacionará los genes y su recuento con las variables clínicas (historial de fumador / no fumador). Comprobamos que las muestras de ambos archivos coinciden en orden.

```
table(colnames(countdata)==sampleinfo$SAMPLE_ID)

##
## TRUE
##   34
```

2.4 Análisis – Filtrado de genes con baja expresión

Aplicamos un count per million (CPM), con el fin de eliminar de nuestro análisis los genes de expresión baja. De esta forma evitamos que interfiriera con las aproximaciones que se realizaron posteriormente en el análisis.

```
myCPM=cpm(countdata)

head(myCPM)

##                   AL4602 AU5884 BL3403 CA9903 CU9061 DI6359
DM123062
## ENSG00000131584 290.1434      0      0      0      0      0
0
## ENSG00000179403 414.4906      0      0      0      0      0
0
## ENSG00000197530 248.6944      0      0      0      0      0
0
## ENSG00000049239 290.1434      0      0      0      0      0
0
## ENSG00000130940 331.5925      0      0      0      0      0
0
```

```
## ENSG00000116786 290.1434      0      0      0      0      0
0
```

Utilizaremos un CPM=1, que suele ser el enfoque mínimo habitual y a partir de los datos resultantes, mantuvimos aquellos genes que presentaban un mínimos de dos *trues* por línea. Así que obtuvimos un total de 1642 genes. Estos serían los genes sobre los que proseguiríamos en el análisis.

```
keep=rowSums(thresh)>=2

counts.keep=countdata[keep,]

summary(keep)

##    Mode   FALSE    TRUE
## logical    4056    1642

dim(counts.keep)

## [1] 1642   34
```

Convertimos los recuentos obtenidos a un objeto DGEList, que guardamos en el objeto *y*, necesario para el siguiente paso dentro del pipeline.

```
y=DGEList(counts.keep)

y

## An object of class "DGEList"
## $counts
##                 AL4602 AU5884 BL3403 CA9903 CU9061 DI6359 DM123062
FR9547
## ENSG00000131584      7      0      0      0      0      0        0
0
## ENSG00000179403     10      0      0      0      0      0        0
```

2.5 Análisis - Control de calidad

Empezamos analizando la cantidad de recuentos por cada muestra de *y*.

```
y$samples$lib.size

##  [1] 12269   601  1883 12778 17910 14641  5173 25813  1391  5994
49032
## [12]  6286 14400  5076  3857 14524 12658  1005  7746  9858 34476
2178
## [23] 67717 12284 21302 13149  7206 23459   373    23  9560  4447
18116
## [34] 17122
```

También hicimos un análisis del tamaño de nuestras librerías para detectar si existían grandes diferencias entre ellas.

```
barplot(y$samples$lib.size, names=colnames(y),las=2)
```

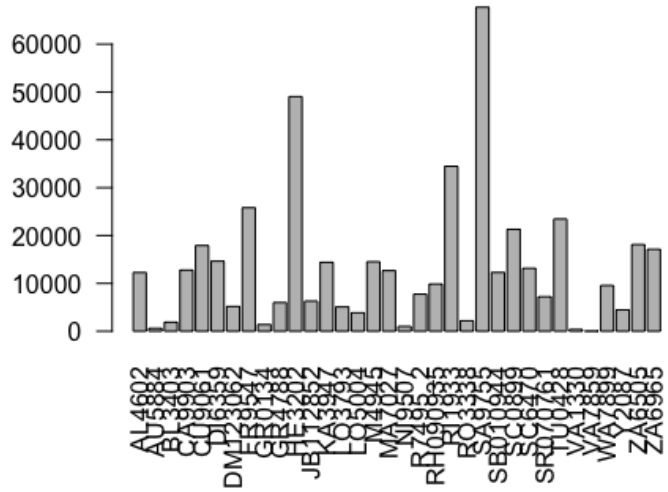


Figura 4. Tamaño de librerías

Efectivamente, la información no mostraba una distribución normal. Para observar las distribuciones de nuestros recuentos, tuvimos que aplicar un log a nuestros cálculos. Observamos nuestros datos sin normalizar con corrección log y obtuvimos una distribución con gran variedad de intensidades log y una media constante.

```
boxplot(logcounts, xlab="", ylab="Recuento Log2 por millón")
```

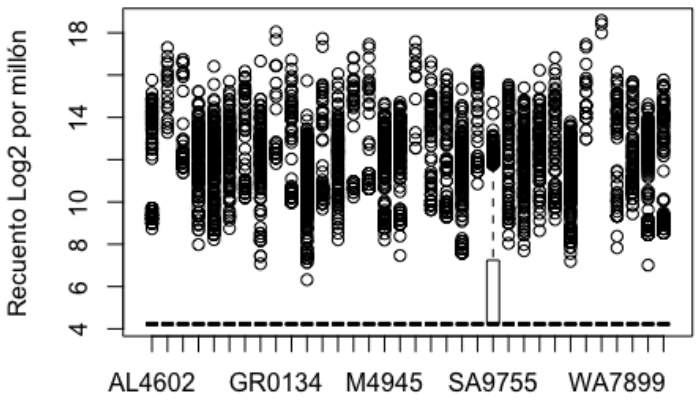


Figura 5. Distribución de datos en librerías con corrección log2

Llegado este punto, creamos un gráfico escalar multidimensional (MDS) con el fin de entender el origen de la variedad de nuestros datos. El objetivo era observar tendencias, o agrupamientos de muestras y explicar fuentes de variabilidad. Según nuestra hipótesis de diferencias entre grupos, esperábamos ver una agrupación de datos por grupos (historial de tabaquismo frente a sin historial de tabaquismo).

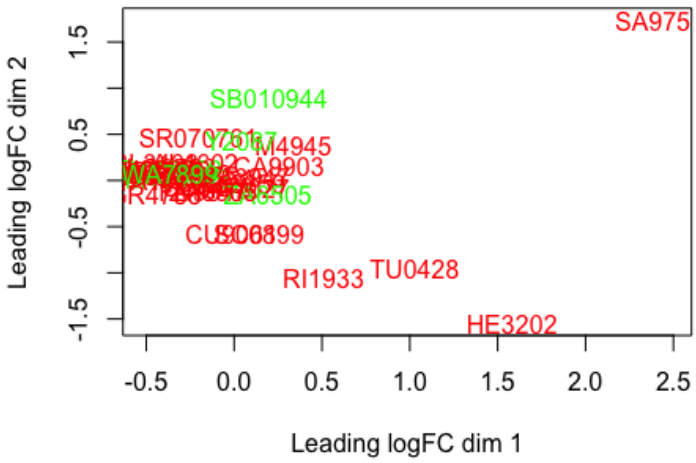


Figura 6. Visualización de agrupaciones de muestras en MDS

El resultado gráfico es una mezcla de grupos: las muestras sin historial de tabaquismo se mezclan con las otras muestras, especialmente tres comparten la zona del grafico ocupada por una gran concentración de muestras de paciente con antecedentes de tabaquismo. Son seis las muestras que se localizan fuera de área de concentración. En cinco casos, se corresponden con muestras de fumadores, constando uno de ellos (SA9755) como “outlier”. Nuestro análisis, focalizado en mostrar expresión génica diferencial entre grupos, no muestra, en este estadio del análisis, ningún tipo de variabilidad por grupo. Este hecho es un primer indicador de que la hipótesis alternativa no es correcta. Adicionalmente esta observación demuestra que existe gran variabilidad entre muestras del mismo grupo (grupo historial tabaquismo).
Seguidamente, generamos clusters jerárquicos con mapas de calor, donde analizamos las diferencias del logCPM para los 250 genes más variables.

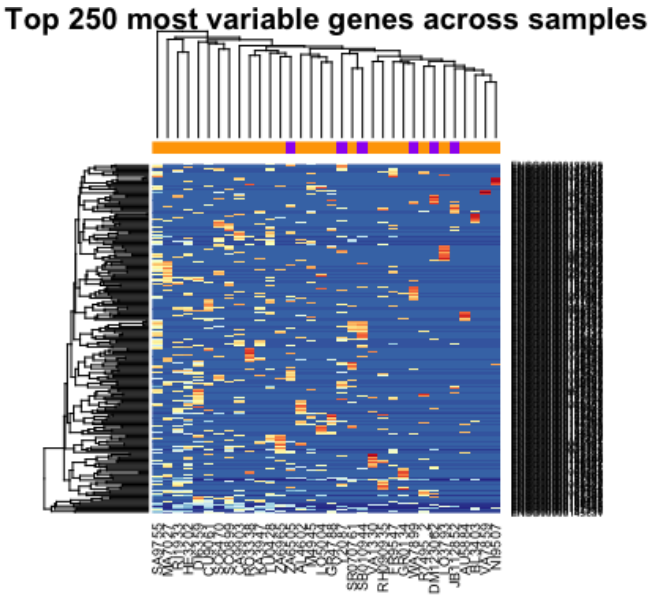


Figura 7. Mapa de calor con los 250 más sobreexpresados entre muestras

2.6 Análisis - Normalización de librerías

Tras constatar en el control de calidad que nuestros datos disponían de gran variabilidad entre grupos, y entre librerías de muestras. Se tuvo que corregir el problema de las tendencias de composición mediante un factor de normalización. Los datos contenidos en el objeto DGE se normalizaron utilizando la función calcNormFactors porque normaliza entre librerías, en lugar de hacerlo según una sola librería. Con estos factores de normalización, en los casos (>0) se escala hacia abajo ya que hay más tendencia que en otras librerías.

2.7 Análisis - Expresión diferencial

Primeramente, creamos la matriz de diseño para comparar los dos grupos bajo estudio y su relación en cuanto a la expresión génica.
Para ello, creamos la variable grupo. Sólo deseamos analizar la expresión diferencial según la variable de grupo de historial de tabaquismo (sí-no).

Creamos el objeto para realizar el análisis diferencial sin intercepto.

```
design=model.matrix(~ 0+group)
```

Cada una de las columnas de nuestra matriz de diseño indica qué muestras corresponden a cada grupo.

```
design
```

```
##    groupNO groupYES
## 1        0        1
## 2        0        1
## 3        0        1
## 4        0        1
## 5        0        1
## 6        0        1
```

```
colnames(design)=levels(group)

design
```

```
##    NO YES
## 1   0   1
## 2   0   1
## 3   0   1
## 4   0   1
## 5   0   1
## 6   0   1
## attr(,"assign")
## [1] 1 1
## attr(,"contrasts")
## attr(,"contrasts")$group
## [1] "contr.treatment"
```

Utilizamos la función voom para ajustar automáticamente las medidas de librerías utilizando la normalización calculada anteriormente. Esta transformación voom utiliza la matriz de diseño experimental y genera un objeto EList. Asimismo generamos un gráfico que nos permitió observar la tendencia

media-varianza, y así ver si existían genes realmente variables en el conjunto de nuestros datos y si filtramos los recuentos bajos correctamente.

```
par(mfrow=c(1,2))

v=voom(z,design,plot=TRUE)
```

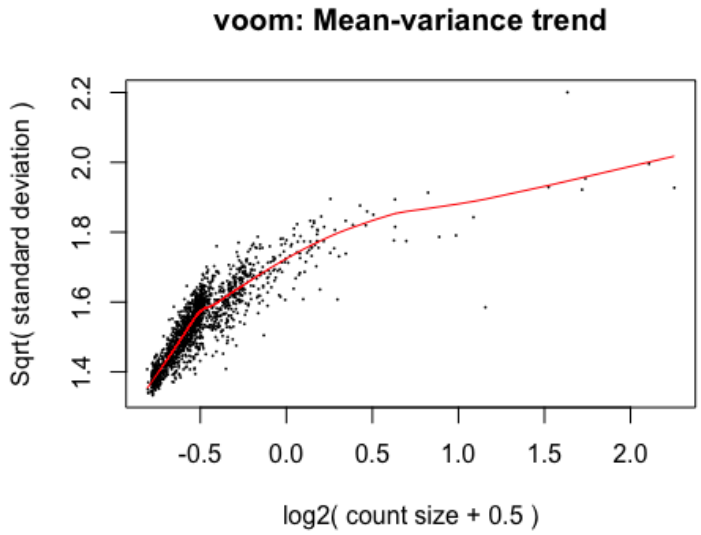


Figura 8. Tendencia media-varianza del conjunto de datos

Ahora podremos comparar los datos normalizados contra los no normalizados

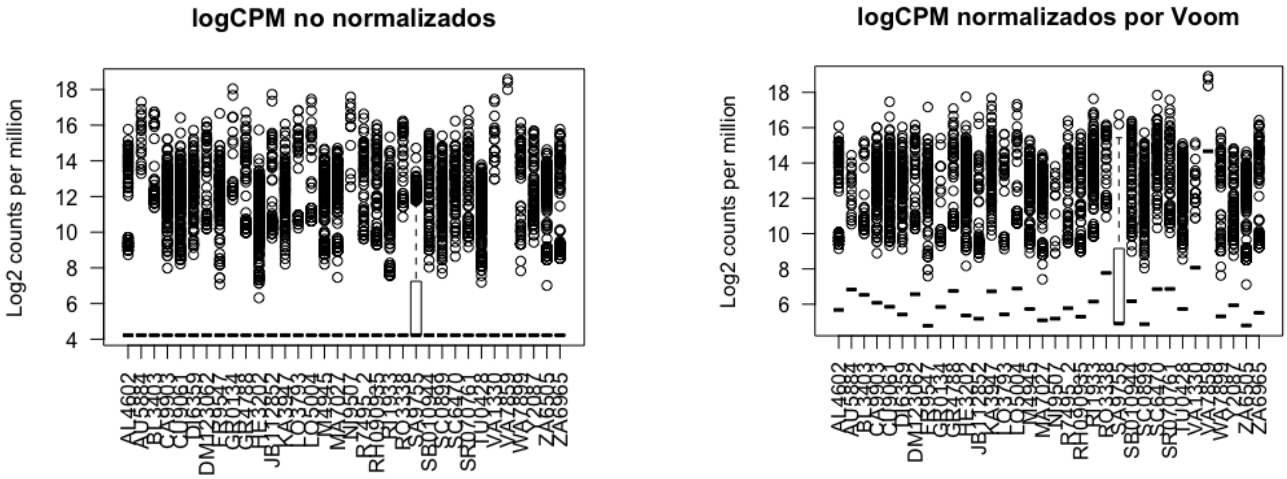


Figura 9. Datos no normalizados frente a normalizados por voom

Como resultado, las medias subieron respecto de las no normalizadas, pero seguíamos teniendo el problema de que no eran suficientemente comparables dada la variabilidad de medias, que ya no estaba en una constante.

Para explorar la diferencia de expresión hicimos el ajuste del modelo para estimar medias de grupos en función de la matriz de diseño así como varianza de genes. Para detectar diferencias entre grupos, establecimos que la hipótesis nula era la igualdad entre grupos (H0: SH-NSH=0) para cada gen.

```
fit=lmFit(v)
```

```
names(fit)
```

```
##  [1] "coefficients"     "stdev.unscaled"   "sigma"
##  [4] "df.residual"      "cov.coefficients" "pivot"
##  [7] "rank"             "Amean"            "method"
## [10] "design"
```

Aplicamos la matriz de contrastes al ajuste (para así obtener la estadística y las estimaciones paramétricas) y aplicamos el shrink de Bayes sobre las varianzas para así obtener los valores t y p. Generamos un diagrama Venn para visualizar los resultados de sobreexpresión / bajoregulación génica:

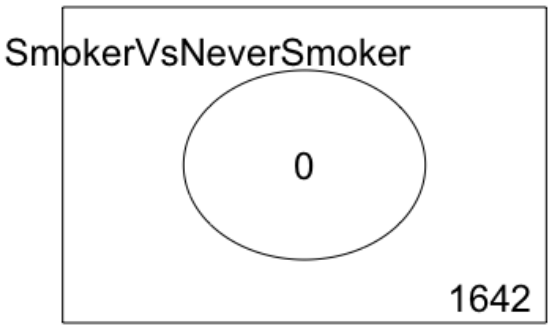


Figura 10. Diagrama Venn – Grupos de genes

Pero los grupos obtenidos no indicaban la presencia de genes diferencialmente expresados (sobreexpresión o bajoregulación). Dado este resultado, desestimamos hipótesis según la cual había diferencias génicas entre los dos grupos de la cohorte.

Sin embargo, quisimos constatar cuáles eran los10 genes más relevantes en ambos grupos.

```
topTable(fit.cont,coef="SmokerVsNeverSmoker",sort.by="p")
```

```
##                     logFC  AveExpr         t      P.Value adj.P.Val
## ENSG00000164796 -4.338128 8.933749 -3.947018 7.923365e-05 0.1301017
## ENSG00000154358 -3.334135 8.083976 -3.116203 1.832950e-03 0.2920500
## ENSG00000133703 -3.093202 7.827856 -2.949194 3.187439e-03 0.2920500
## ENSG00000141837 -3.002128 7.820985 -2.864258 4.181531e-03 0.2920500
## ENSG00000178104 -2.898789 7.374194 -2.846167 4.426616e-03 0.2920500
## ENSG00000183117 -2.914022 7.501429 -2.843360 4.465787e-03 0.2920500
## ENSG00000169876 -2.934627 7.826866 -2.798268 5.139617e-03 0.2920500
## ENSG00000134516 -2.735758 7.398169 -2.682783 7.303518e-03 0.2920500
## ENSG00000112079 -2.558351 6.790570 -2.637967 8.342931e-03 0.2920500
## ENSG00000116183 -2.726950 7.657483 -2.632274 8.484019e-03 0.2920500
##                         B
## ENSG00000164796 -3.245817
```

```
## ENSG00000154358 -3.757787
## ENSG00000133703 -3.832867
## ENSG00000141837 -3.881772
## ENSG00000178104 -3.860070
## ENSG00000183117 -3.868789
## ENSG00000169876 -3.919983
## ENSG00000134516 -3.956446
## ENSG00000112079 -3.932803
## ENSG00000116183 -3.999386
```

2.8 Análisis - Anotación de genes

Para finalizar el análisis, quisimos realizar la anotación de genes. Para ello utilizamos el paquete org.Hs.eg.db y mantuvimos las columnas de interés: Ensembl, Name y Symbol, que añadimos con la función mapIDs.

```
columns(org.Hs.eg.db)
```

```
##  [1] "ACCNUM"       "ALIAS"        "ENSEMBL"      "ENSEMBLPROT"
##  [5] "ENSEMBLTRANS" "ENTREZID"     "ENZYME"       "EVIDENCE"
##  [9] "EVIDENCEALL"  "GENENAME"     "GO"           "GOALL"
## [13] "IPI"          "MAP"          "OMIM"         "ONTOLOGY"
## [17] "ONTOLOGYALL"  "PATH"         "PFAM"         "PMID"
## [21] "PROSITE"      "REFSEQ"       "SYMBOL"       "UCSCKG"
## [25] "UNIGENE"      "UNIPROT"
```

```
fit.cont$entrez=mapIds(org.Hs.eg.db,
keys=row.names(fit.cont),column="ENTREZID",keytype="ENSEMBL",multiVals
="first")
```

```
## 'select()' returned 1:many mapping between keys and columns
```

```
fit.cont$name=mapIds(org.Hs.eg.db,
keys=row.names(fit.cont),column="GENENAME",keytype="ENSEMBL",multiVals
="first")
```

```
## 'select()' returned 1:many mapping between keys and columns
```

```
results=as.data.frame(fit.cont)
```

La última columna de nuestra matriz muestra el nombre de cada gen, el símbolo y el código Ensembl.
En esta etapa del estudio, creamos los gráficos para volver a comprobar la expresión diferencial y constatar la falta de sobreexpresados o bajoexpresados en la comparativa de grupos:

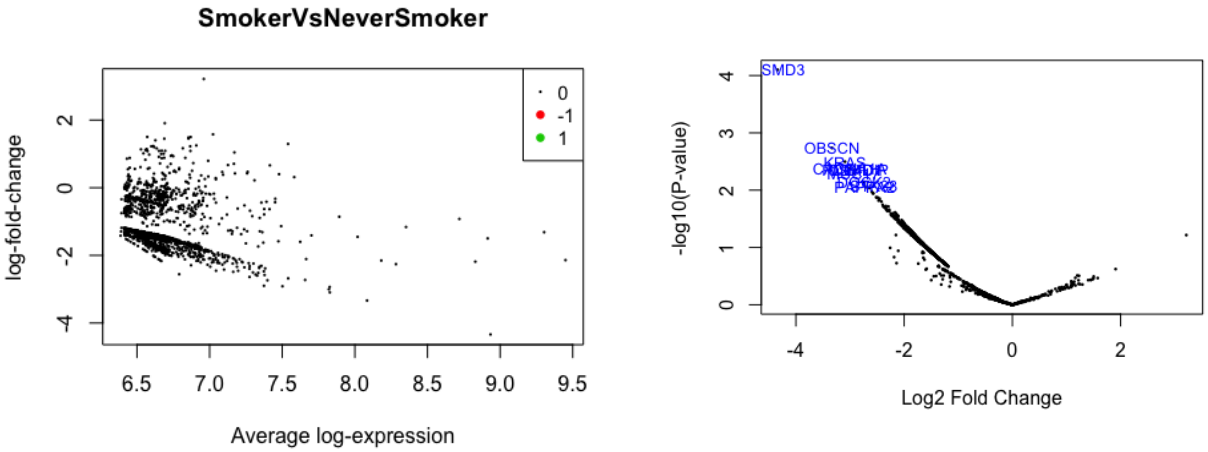


Figura 11. Genes diferencialmente expresados y diagrama tipo volcán con los 10 genes más sobreexpresados

No se realizó enriquecimento, dado que no hay DEs.

```
go=goana(fit.cont,coef="SmokerVsNeverSmoker", species="Hs")

## No DE genes

topGO(go, n=10)

## data frame with 0 columns and 0 rows
```

3. Capítulo 2 – Análisis de supervivencia entre grupos

3.1 Introducción

En este capítulo realizamos un análisis de supervivencia para medir el tiempo transcurrido hasta el evento muerte, nuestra variable de interés. El objetivo era determinar si el intervalo hasta el evento era igual o diferente entre ambos grupos (historial tabaquismo / sin historial), es decir, nuestro objetivo era comprobar si nuestra hipótesis alternativa "existen diferencias de supervivencia entre ambos grupos" era cierta. En términos estadísticos, debíamos averiguar si el tiempo de supervivencia se encontraba íntimamente ligado con la covariante "historial de tabaquismo", concretamente si la pertenencia a alguno era significativa. En el caso de ser cierto, habría indicios para afirmar que las características del adenocarcinoma en ambos grupos es diferente y esto podría abrir una línea de investigación nueva para detectar el origen de estas diferencias y la necesidad de biomarcadores y tratamientos diferentes.

Se han aplicados tres métodos diferentes (Kaplan Meier, Cox y Weibull) para realizar este análisis con el fin de dar más robustez a los resultados y las conclusiones. A continuación se define brevemente cada uno de ellos:

- Kaplan-Meier:
  Se utiliza para crear gráficos de los tiempos de supervivencia observados dentro de nuestros grupos comparándolos por medio del test log-rank
- Cox:
  Se utiliza para "medir la tasa de fallo para un determinado individuo utilizando una aproximación exponencial para las covariables (sin contar la variable tiempo") [5].
  Se trata de un modelo semiparamétrico (los parámetros de regresión (betas) son conocidas aunque no lo es la distribución resultante).
- Weibull (AFT):
  Efecto multiplicativo de las covariables (proporcional) respecto al tiempo de supervivencia.
  Modelo paramétrico.

Se trabajó con:

- archivo survival.txt.
  Este archivo contiene la tabla de supervivencia del grupo (y las covariables de interés) para el que en la PEC 2 se realizó el análisis RNAseq.
  En este archivo existen 6 individuos sin historial de tabaquismo, por lo que disponemos de una muestra desequilibrada y en el que existen 7 casos de censura (abandono del estudio antes de que éste termine).
- software R y los paquetes survival y survreg:

Los pasos que seguimos en el análisis de supervivencia fueron:

- Preparación de archivos
- Aplicación método KM
- Aplicación método Cox

    - Verificación de la asunción de riesgos proporcionales
    - Aplicación del método extendido
- Aplicación método Weibull
- Conclusiones

3.2 Análisis - Preparación de archivos y espacio de trabajo

Se prepararon los datos y se importó a R el data set perteneciente a las muestras analizadas en la Fase 1.
Eliminamos la fila 35 puesto que había datos incompletos que podrían haber causado ruido en el análisis.

```
survival=survival[-c(35), ]
```

Asimismo, cargamos los paquetes necesarios y se renombraron las columnas del data set:

```
library(survival)

library(survminer)

## Loading required package: ggplot2

## Loading required package: ggpubr

## Loading required package: magrittr
```

```
colnames(survival)=c("id","sample","note","cancer","C.subtype","age","DCB","Censoring","Event","Histology","Signature","Neoantigen","Mutation","Drug","Oncotree","PDL1","Agent","N.Cigarretes","Prior Chemo","Death","C.studies","sex","SmokingH","Exonic","TreatmentResponse")
```

En primer lugar, creamos la variable respuesta, que fue la base del análisis. Para ello, se creó un objeto Surv con las dos variables que miden el tiempo hasta la muerte. Se tomó el valor censoring=1, puesto que el interés se centra en el evento que ocurre cuando no hay censura (es decir, 1 indica la muerte del individuo).
Y creamos el objeto que fue la variable respuesta del análisis:

```
Y=Surv(survival$Death,survival$Censoring==1)
```

3.3 Análisis por Kaplan Meier

Utilizamos el método Kaplan Meier (KM) para conocer la media de supervivencia para nuestro grupo, así como la supervivencia por intervalos de confianza de 95%. Aplicamos el argumento survfit sin condicionar la variable respuesta con otras covariables, por ello ~1:

```
kmfit1=survfit(Y~1)

kmfit1

## Call: survfit(formula = Y ~ 1)
##
##       n  events  median 0.95LCL 0.95UCL
##    34.0    22.0     6.5     3.5      NA
```

El set de datos con el que trabajamos indicaba que de entre el total de observaciones (la cantidad en riesgo en el tiempo 0 y los 22 casos en los que

ocurría), la media de vida era de 6.5 meses, y dentro del intervalo de confianza del 95%, el límite inferior era de 3.5 meses.
Seguidamente observamos con la función summary las estimaciones a nivel de probabilidad para cada uno de los eventos y la cantidad de individuos en riesgo.

```
summary(survfit(Surv(survival$Death,survival$Censoring==1)~1))
```

```
## Call: survfit(formula = Surv(survival$Death, survival$Censoring ==
##     1) ~ 1)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   1.4     34       1    0.971  0.0290        0.915        1.000
##   1.8     33       2    0.912  0.0486        0.821        1.000
##   1.9     31       5    0.765  0.0727        0.635        0.921
##   2.1     26       2    0.706  0.0781        0.568        0.877
##   3.3     24       1    0.676  0.0802        0.536        0.853
##   3.4     23       1    0.647  0.0820        0.505        0.829
##   3.5     22       1    0.618  0.0833        0.474        0.805
##   4.1     20       1    0.587  0.0847        0.442        0.779
##   6.3     17       2    0.518  0.0877        0.371        0.722
##   6.5     15       1    0.483  0.0884        0.338        0.692
##   8.1     14       1    0.449  0.0885        0.305        0.661
##   8.3     13       3    0.345  0.0860        0.212        0.562
##  14.5      5       1    0.276  0.0924        0.143        0.532
```

Creamos la gráfica KM de visualización de los resultados anteriores, donde podíamos ver la supervivencia integral del grupo:

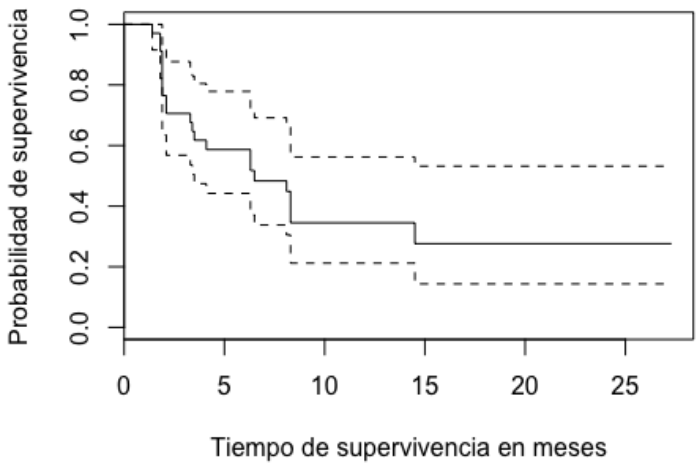


Figura 12. Tiempo de supervivencia total en el grupo con CI 95%

Posteriormente, creamos la función condicionando por la covariable de interés (historial fumador sí o no) SmokingH para ver si podíamos observar diferencias entre las curvas de supervivencia de ambas categorías de la variable: SmokingH=1 (Sí) vs. SmokingH=0 (No). Para ello tuvimos que realizar un ajuste de la variable:

```
kmfit2=survfit(Y~survival$SmokingH)
```

Observamos que el grupo de no fumadores sobrevivía hasta la semana 8.3, mientras que para entonces el data set de fumadores todavía contenía 15 observaciones dentro del grupo de riesgo. Esto sería un primer indicio para descartar la hipótesis de que las categorías de la variable no influyen sobre las probabilidades de supervivencia.

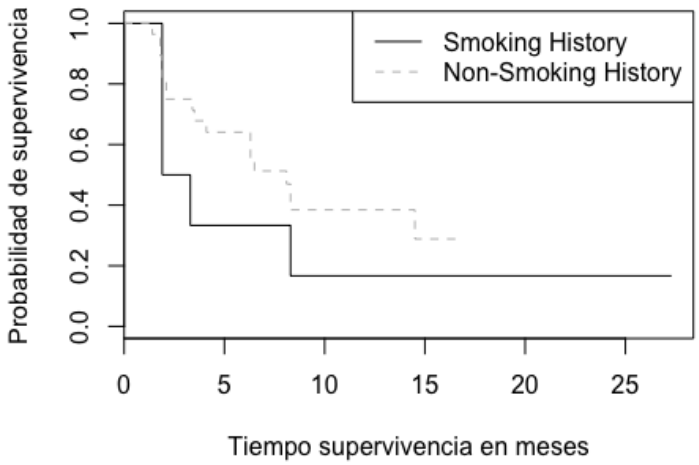


Figura 13. Tiempo de supervivencia comparado entre grupos

Existía un paralelismo entre curvas de supervivencia.

La curva también indicaba que el grupo sin historial como fumador era el grupo que más probabilidades tenía de supervivencia aunque su curva era significativamente más corta en lo que al tiempo de supervivencia se refería (probablemente dado el desequilibrio entre las observaciones de SmokingH=1 (17) vs. 0 (5).

Adicionalmente, efectuamos un test de rangos logarítmicos sobre la variable historial de tabaquismo, para log rank test. El p-valor de la estadística logarítmica no mostraba un valor significativo sobre la variable, indicio del poco efecto de la variable sobre el tiempo de supervivencia.

Ampliamos el modelo KM una covariable más, Signature, que podría ser significativa sobre la supervivencia.

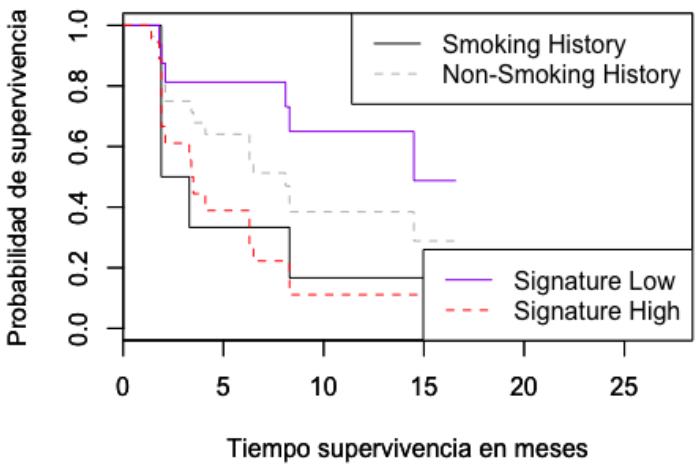


Figura 14. Tiempo de supervivencia entre grupos y tipos de firma

Añadiendo la variable Signature al modelo podríamos ver el efecto del modelo al incorporar una variable más.

```
survdiff(Surv(survival$Death,
survival$Censoring)~survival$SmokingH+survival$Signature)

## Call:
## survdiff(formula = Surv(survival$Death, survival$Censoring) ~
##     survival$SmokingH + survival$Signature)
##
##                                                                N
Observed
## survival$SmokingH=0, survival$Signature=transversion low    6
5
## survival$SmokingH=1, survival$Signature=transversion high 16
6
## survival$SmokingH=1, survival$Signature=transversion low  12
11
##                                                           Expected
## survival$SmokingH=0, survival$Signature=transversion low      3.31
## survival$SmokingH=1, survival$Signature=transversion high    12.67
## survival$SmokingH=1, survival$Signature=transversion low      6.02
##                                                           (O-E)^2/E
## survival$SmokingH=0, survival$Signature=transversion low      0.857
## survival$SmokingH=1, survival$Signature=transversion high     3.510
## survival$SmokingH=1, survival$Signature=transversion low      4.127
##                                                           (O-E)^2/V
## survival$SmokingH=0, survival$Signature=transversion low       1.10
## survival$SmokingH=1, survival$Signature=transversion high      9.19
## survival$SmokingH=1, survival$Signature=transversion low       6.35
##
##  Chisq= 9.5  on 2 degrees of freedom, p= 0.009
```

Al añadir la variable obtenemos un p-valor significativo, que requeriría mantener este nuevo coeficiente en el modelo.

Validamos ambos resultados aplicando una variante del rango logarítmico para validar el primer modelo con sólo la variable SmokingH, donde el argumento rho pondera los valores observados y esperados de forma diferente a la función anterior.

```
survdiff(Surv(Death,Censoring)~SmokingH,data=survival,rho=1)

## Call:
## survdiff(formula = Surv(Death, Censoring) ~ SmokingH, data =
survival,
##     rho = 1)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## SmokingH=0  6     3.89     2.41     0.912       1.5
## SmokingH=1 28    11.65    13.13     0.167       1.5
##
##  Chisq= 1.5  on 1 degrees of freedom, p= 0.2
```

Y obtuvimos el mismo resultado que en el primer caso, donde el p- valor de esta covariable tampoco mostraba un efecto significativo sobre la supervivencia.
Para terminar, se realizó un último test estratificando por la variable Signature:

```
survdiff(Surv(Death,Censoring)~SmokingH+strata(Signature),data=surviva
l)

## Call:
## survdiff(formula = Surv(Death, Censoring) ~ SmokingH +
strata(Signature),
##     data = survival)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## SmokingH=0  6        5     5.35   0.02323    0.0412
## SmokingH=1 28       17    16.65   0.00747    0.0412
##
##  Chisq= 0  on 1 degrees of freedom, p= 0.8
```

Y el resultado del p-valor tampoco era significativo.
Así, los análisis KM mostraron a partir de los p-valores que el efecto de la covariable bajo estudio no era significativo sobre la supervivencia y que no era necesaria incluirla en nuestros modelos.
La representación de las curvas sí que indica que las probabilidades de supervivencia para SmokingH=1 son menores que para el otro grupo, pero dado el desequilibrio entre las observaciones de las dos categorías y el tamaño de la muestra podría no ser fiable. Seguimos con el método Cox para ampliar y confirmar los primeros resultados derivados de KM.

3.4 Análisis por método Cox

En este apartado tomamos como referencia el método Cox para medir las tasas de fallo (contra la tasa de supervivencia de KM) y así ver si los resultados estaban en la misma línea que en análisis KM.
Para probar que el modelo era correcto, realizamos comprobaciones sobre la asunción de riesgos proporcionales (PH). El desarrollo de estas comprobaciones se puede examinar en el Anexo 2.
Aplicamos la función coxph para ajustar únicamente por la variable de interés.

```
coxph(Y~SmokingH,data=survival)

## Call:
## coxph(formula = Y ~ SmokingH, data = survival)
##
##            coef exp(coef) se(coef)     z    p
## SmokingH -0.575     0.562    0.512 -1.12 0.26
##
## Likelihood ratio test=1.13  on 1 df, p=0.3
## n= 34, number of events= 22
```

Obtuvimos el riesgo de fallo estimado para SmokingH=0 vs SmokingH=1 dentro de un intervalo de confianza al 95% de entre 0.2 y 1.5 meses. El p-valor obtenido no era significativo respecto del nivel 0.05 de H0.

```
summary(coxph(Y~SmokingH,data=survival))

## Call:
## coxph(formula = Y ~ SmokingH, data = survival)
##
##   n= 34, number of events= 22
##
##             coef exp(coef) se(coef)      z Pr(>|z|)
## SmokingH -0.5754    0.5625   0.5120 -1.124    0.261
##
##          exp(coef) exp(-coef) lower .95 upper .95
## SmokingH    0.5625      1.778    0.2062     1.534
##
## Concordance= 0.558  (se = 0.046 )
## Rsquare= 0.033   (max possible= 0.98 )
## Likelihood ratio test= 1.13  on 1 df,   p=0.3
## Wald test            = 1.26  on 1 df,   p=0.3
## Score (logrank) test = 1.3  on 1 df,   p=0.3
```

Las tasas de fallo eran de 0.56 para SmokingH=0 y de 1.7 para SmokingH=1. De nuevo, se confirmaba que SmokingH=0 presentaba una tasa de fallo anterior a SmokingH=1, y que dicha variable no tenía influencia sobre el tiempo de supervivencia.
En este punto, y de forma análoga al punto anterior de análisis por KM, añadimos al modelo una covariable más. La intención era valorar los efectos de una covariable adicional en el modelo:

```
modelo2=coxph(Y~SmokingH+Signature,data=survival)

modelo2

## Call:
## coxph(formula = Y ~ SmokingH + Signature, data = survival)
##
##                             coef exp(coef) se(coef)    z      p
## SmokingH                   0.208     1.232    0.554 0.38 0.7070
## Signaturetransversion low 1.465     4.328    0.527 2.78 0.0054
##
## Likelihood ratio test=9.41  on 2 df, p=0.009
## n= 34, number of events= 22
```

Los p-valores de Signature eran los significativos sobre la variable respuesta, rente a SmokingH que mostró no serlo.
Mediante la función survfit, creamos la curva de supervivencia de la función Cox con el valor SmokingH=0 y otra con el valor SmokingH=1 para observar el comportamiento de los coeficientes de regresión.

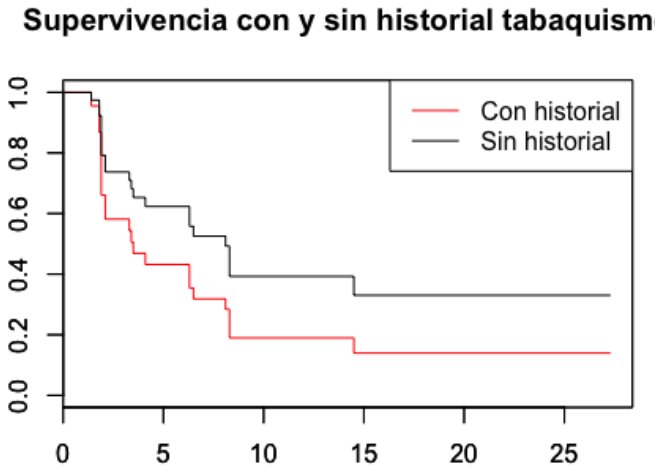


Figura 15. Tiempo de supervivencia entre grupos por método Cox

Este gráfico indicaba que estábamos utilizando un método adecuado, puesto que probaba la asunción de riesgos proporcionales al mostrar que la dos curvas de supervivencia transcurren en paralelo (excepto al inicio de la gráfica). Asimismo, este gráfico también mostraba que en las previsiones de supervivencia, existía una ligera mayor longevidad y probabilidades de supervivencia en el grupo sin historial de tabaquismo.

3.5 Análisis por modelo Weibull

Con este método exploraremos el valor que obtenemos desde un enfoque de factor de aceleración (AFT). Queríamos averiguar si alguna de las categorías de la variable de grupo tenía un efecto multiplicador sobre la supervivencia.
Para ello, utilizamos la función survreg y el argumento weibull.

```
modpar1=survreg(Surv(survival$Death,survival$Censoring==1)~SmokingH,da
ta=survival,dist="weibull")

summary(modpar1)

##
## Call:
## survreg(formula = Surv(survival$Death, survival$Censoring ==
##     1) ~ SmokingH, data = survival, dist = "weibull")
##              Value Std. Error     z       p
## (Intercept)  2.197      0.443  4.96 7.2e-07
## SmokingH     0.213      0.511  0.42    0.68
## Log(scale)  -0.022      0.173 -0.13    0.90
##
```

```
## Scale= 0.978
##
## Weibull distribution
## Loglik(model)= -74   Loglik(intercept only)= -74.1
##  Chisq= 0.17 on 1 degrees of freedom, p= 0.68
## Number of Newton-Raphson Iterations: 5
## n= 34
```

El factor de aceleración que compara SmokingH=1 contra SmokingH=0 se estima en el exp(0.213) = 1.23. Así que el tiempo medio de vida para un grupo contra el otro es de 1.23. En el caso del modelo Weibull, los p-valores que estamos observando no dan un valor significativo a esta covariable.
Estos resultados nos llevan a aceptar la falta de efecto/influencia de la variable sobre el tiempo de supervivencia SmokingH y así aceptar la H0, aunque la categoría 0 y la 1 muestran factores de aceleración diferentes, por lo que el comportamiento entre categorías de la variable no es el mismo.

4. Conclusiones

Al inicio de nuestro análisis, partíamos de una hipótesis alternativa donde creíamos que los grupos de enfermos con adenocarcinoma de pulmón mostraban diferencias en la expresión de genes en función de si habían sido fumadores o no en el pasado.

Nuestro actual análisis RNAseq procesado sobre una muestra de tamaño 34 indica que nuestra hipótesis de rechaza.

Asimismo, los resultados del análisis de supervivencia basado en el uso de tres métodos diferentes indican también que dicha hipótesis podría no ser cierta, puesto que los p-valores obtenidos en los tres métodos indican la falta de efecto de la variable sobre la supervivencia del grupo.

Es posible que efectivamente no existan diferencias entre grupos, pero existen algunas cuestiones que deben apuntarse para comprender estos resultados
Sí que es cierto que la comparación de categorías muestra, en una aproximación Cox, curvas de supervivencia diferentes, con mayor supervivencia (probabilidad) dentro del grupo sin historial.
¿Significa esto que con una muestra considerablemente mayor, podríamos obtener otros resultados que apoyen la hipótesis alternativa?

¿Hemos obtenido este resultado debida la descompensación entre tamaño de grupos (6 sin historial tabaquismo vs 28 con historial tabaquismo)?

Otra de las preguntas que generan los resultados del trabajo, es que se han estudiado las expresiones somáticas, pero ¿mostraría el estudio epigenético otros resultados entre grupos, incluso en una muestra pequeña y descompensada?

Otro elemento interesante que hemos observado y que no deberíamos ignorar es la variabilidad entre nuestras muestras dentro del propio grupo de fumadores. ¿A qué son debidas? ¿Por qué tenemos dispersión en el diagrama de MDS de control de calidad? ¿Hay alguna covariable que esté afectando a esta distribución?

Los objetivos fijados al inicio del trabajo se han cumplido: hemos desarrollado cada uno de los objetivos marcados y hemos verificado la hipótesis nula. Sin embargo, tal como plantean las preguntas de este apartado, los resultados podrían haber sido diferentes si se hubiera manejado una muestra de mayor tamaño. De hecho, inicialmente se partía de una muestra bastante más extensa de 1144 observaciones. El análisis con esa hubiera tenido mucho más poder estadístico, dada la magnitud de los datos y el potencial equilibrio entre grupos. Sin embargo, los problemas de rendimiento y de complejidad de manejo que exigía tal muestra, forzó la necesidad de eliminar alcance y seleccionar una muestra más manejable para poder desarrollar los objetivos propuestos y cumplir con los hitos marcados. Por lo tanto, es necesario resaltar que precisamente la inclusión de buffer en la planificación del proyecto, y el

seguimiento metódico de la duración de cada una de las tareas, ha permitido poder reducir el alcance para seguir cumplimiento con los hitos del proyecto.

5. Glosario

AFT: factor de aceleración (del inglés, acceleration failure time)
ADC: adenocarcinoma
CI: intervalo de confianza (del inglés, confidence interval)
DE: expresión diferencial (del inglés, differential expression)
H0: hipótesis nula
NSCLC: cáncer de célula no pequeña (del inglés, non-small cell lung cáncer)
SCLC: cáncer de célula pequeña (del inglés, small cell lung cáncer)
SNP: single nucleotide polimorphisms

6. Bibliografía

1. Stewart B. W. and Wild C.P. *World Cancer Report 2014*, International Agency for Research on Cancer, Lyon, France, 2015

2. Le Chevalier, T., *Non-small cell lung cancer: the challenges of the next decade*. Frontiers in Oncology, art. 29, 1-4, vol. 1, 2011

3. Mcguire, A., & Martin, M. & Lenz, C. & Sollano, J., *Treatment cost of non-small cell lung cancer in three European countries: Comparisons across France, Germany, and England using administrative databases*, Journal of medical economics, 18. 1-16. 10.3111/13696998.2015.1032974, 2015

4. Ruiz, M.C. & Sánchez-Pla, A. *Genómica funcional y análisis de microarrays.* PID_00192741

5. Carlson, M., Obenchain, V., Pagès, H., Shannon, P., Tenenbaum, D., Morgan, M.*, High-throughput sequence analysis with R and Bioconductor*, https://www.bioconductor.org/help/course-materials/2012/CSC2012/Bioconductor-tutorial.pdf, 2012.

6. Kleinbaum, D., Klein, M., *Survival Analysis. A self-learning text*. Third edition. Springer. 2012

7. Langevin, S., Kratzke, R., Kelsey, K., *Epigenetics of Lung Cancer,* National Institute of Health, USA, 2015

8. Ran, D. & Daye, Z.J., *Gene expression variability and the analysis of large-scale RNA-seq studies with the MDSeq*, Nucleic Acids Research, 1-17, No. 13/Vol.45, 2017

9. https://www.cancer.gov/espanol/tipos/pulmon/paciente/tratamiento-pulmon-celulas-no-pequenas-pdq, Marzo 2018

10. Jamal Hanjani, M., et al. *Tracking the Evolution of Non–Small-Cell Lung Cancer*, The New England Journal of Medicine, 2109-2121, No.22/Vol.376, 2017

11. Pazhouhandeh, M., Samiee, F., Boniadi T., *Comparative network analysis of patients with non-small cell lung cancer and smokers for representing potential Therapeutic targets*, Scientific Reports, 1-15, 7:13812, doi:10.1038/s41598-017-14195-1, 2017

12. Campbell, J., Alezandrov, A., Kim, J., et al, *Disctinct Patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas*, HHS, 48(6): 607–616., doi:10.1038/ng.3564, www.broadinstitute.org/cancer/cga/MutSig, 2016

13. Faraway, J. Linear Models with R. 2015

14. Herbst, R., Morgensztern, D., Boshoff, C., *The biology and management of non-small cell lung cancer*, Nature, 446-454, No. 553, doi:10.1038/nature25183, 2018

15. Rotow, J., Bivona, T., *Understanding and targeting resistance mechanisms in NSCLC*, Nature, 637-658, No. 17, doi:10.1038/nrc.2017.84, 2017

16. https://www.cancer.org/latest-news/why-lung-cancer-strikes-nonsmokers.html, Abril 2018

17. http://www.geneontology.org/page/go-enrichment-analysis, Abril, 2018

18. https://www.bioconductor.org/help/course-materials/2015/LearnBioconductorFeb2015/B02.1.1_RNASeqLab.html, Abril 2018

19. https://www.cancer.org/cancer/non-small-cell-lung-cancer/about/what-is-non-small-cell-lung-cancer.html, Marzo 2018

20. Beer, D., Kardia, S., Huang, C., Giordano, T., Levin, A., Misek, D., Lin, L., Chen, G., Gharib, T., Thomas, D., Lizyness, M., Kuick, R., Hayasaka, S., Taylor, J., Iannettoni, M., Orringer, M., Hanash, S., *Gene-expression profiles predict survival of patients with lung adenocarcinoma*, Nature Medicine, 2002/07/15/online, Volume 8, http://dx.doi.org/10.1038/nm733, 10.1038/nm733, https://www.nature.com/articles/nm733#supplementary-information

21.Raponi, M., Jack Yu, J., Grace Lee, G., Taylor, J., MacDonald, J., Thomas, D., Moskaluk, C., Wang, Y., and Beer, D. *Gene Expression Signatures for Predicting Prognosis of Squamous Cell and Adenocarcinomas of the Lung*, Cancer Res August 1 2006 (66) (15) 7466-7472; DOI: 10.1158/0008-5472.CAN-06-1191

22. Beane, J., Vick, J., Schembri, F., Anderlind, C., Gower, A., Campbell, J., Luo, L., Zhang, X., Xiao, J., Alekseyev, Y., Wang, S., Levy, S., Massion, P., Lenburg, M., Spira, A. *Characterizing the Impact of Smoking and Lung Cancer on the Airway Transcriptome Using RNA-Seq, Cancer Prevention Research*, 803- 817, 10.1158/1940-6207.CAPR-11-0212, 2011/06/01, http://cancerpreventionresearch.aacrjournals.org/content/4/6/803.abstract

23. Myron G.Best, Sol, N., Kooi, I., Tannous, J., Westerman, B., Rustenburg, F:, Schellen, P., Verschueren, H., Post, E., Koster, J., Yistra, B., Ameziane, N., Dorsman, J., Smit, E., Verheul, H., Noske, D., Reijneveld, J., Nilsson, J. Tannous, B., Wesseling, P. Wurdinger, T., *RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics*, Cancer Cell , Volume 28 , Issue 5 , 666 – 676, 2015

24.https://www.cancer.org/latest-news/why-lung-cancer-strikes-nonsmokers.html, American cancer society, *April 2018*

7. Anexos

Anexo 1

Preparación

Procedemos a configurar el espacio de trabajo

```
workingDir=getwd()

dataDir=file.path(workingDir,"data_TFM")

resultsDir=file.path(workingDir,"results_TFM")
```

Cargamos los paquetes de trabajo

```
source("http://www.bioconductor.org/biocLite.R")

## Bioconductor version 3.6 (BiocInstaller 1.28.0), ?biocLite for help

biocLite()

## BioC_mirror: https://bioconductor.org

## Using Bioconductor 3.6 (BiocInstaller 1.28.0), R 3.4.4 (2018-03-
15).

biocLite("edgeR")

## BioC_mirror: https://bioconductor.org

## Using Bioconductor 3.6 (BiocInstaller 1.28.0), R 3.4.4 (2018-03-
15).

## Installing package(s) 'edgeR'

##
## The downloaded binary packages are in
##
/var/folders/45/j7vkblj156d4rmml06hp_dnh0000gn/T//RtmpP2e88K/downloade
d_packages

biocLite("limma")

## BioC_mirror: https://bioconductor.org

## Using Bioconductor 3.6 (BiocInstaller 1.28.0), R 3.4.4 (2018-03-
15).

## Installing package(s) 'limma'

##
## The downloaded binary packages are in
##
/var/folders/45/j7vkblj156d4rmml06hp_dnh0000gn/T//RtmpP2e88K/downloade
d_packages

biocLite("gplots")

## BioC_mirror: https://bioconductor.org
```

```
## Using Bioconductor 3.6 (BiocInstaller 1.28.0), R 3.4.4 (2018-03-
15).

## Installing package(s) 'gplots'

##
## The downloaded binary packages are in
##
/var/folders/45/j7vkblj156d4rmml06hp_dnh0000gn/T//RtmpP2e88K/downloade
d_packages

biocLite("org.Hs.eg.db")

## BioC_mirror: https://bioconductor.org

## Using Bioconductor 3.6 (BiocInstaller 1.28.0), R 3.4.4 (2018-03-
15).

## Installing package(s) 'org.Hs.eg.db'

## installing the source package 'org.Hs.eg.db'

## Warning in install.packages(pkgs = doing, lib = lib, ...):
installation of
## package 'org.Hs.eg.db' had non-zero exit status

biocLite("Glimma")

## BioC_mirror: https://bioconductor.org

## Using Bioconductor 3.6 (BiocInstaller 1.28.0), R 3.4.4 (2018-03-
15).

## Installing package(s) 'Glimma'

##
## The downloaded binary packages are in
##
/var/folders/45/j7vkblj156d4rmml06hp_dnh0000gn/T//RtmpP2e88K/downloade
d_packages

biocLite("RColorBrewer")

## BioC_mirror: https://bioconductor.org

## Using Bioconductor 3.6 (BiocInstaller 1.28.0), R 3.4.4 (2018-03-
15).

## Installing package(s) 'RColorBrewer'

##
## The downloaded binary packages are in
##
/var/folders/45/j7vkblj156d4rmml06hp_dnh0000gn/T//RtmpP2e88K/downloade
d_packages

biocLite("AnnotationsDbi")
```

```
## BioC_mirror: https://bioconductor.org

## Using Bioconductor 3.6 (BiocInstaller 1.28.0), R 3.4.4 (2018-03-
15).

## Installing package(s) 'AnnotationsDbi'

## Warning: package 'AnnotationsDbi' is not available (for R version
3.4.4)
```

Mantenemos tres columnas de análisis RNAseq: genes, muestras tumorales, recuento de genes.

```
seqdata=read.delim("/Users/llorencrubiovives/Desktop/data_mutations_ex
tended.txt", stringsAsFactors=FALSE)

head(finalseqdata)

##           ENSEMBL t_ref_count.AL4602 t_ref_count.AU5884
t_ref_count.BL3403
## 1 ENSG00000131584                 7                 0
0
## 2 ENSG00000179403                10                 0
0
## 3 ENSG00000197530                 6                 0
0
## 4 ENSG00000049239                 7                 0
0
## 5 ENSG00000130940                 8                 0
0
## 6 ENSG00000116786                 7                 0
0
##   t_ref_count.CA9903 t_ref_count.CU9061 t_ref_count.DI6359
## 1                  0                  0                  0
## 2                  0                  0                  0
## 3                  0                  0                  0
## 4                  0                  0                  0
## 5                  0                  0                  0
## 6                  0                  0                  0
##   t_ref_count.DM123062 t_ref_count.FR9547 t_ref_count.GR0134
## 1                    0                  0                  0
## 2                    0                  0                  0
## 3                    0                  0                  0
## 4                    0                  0                  0
## 5                    0                  0                  0
## 6                    0                  0                  0
##   t_ref_count.GR4788 t_ref_count.HE3202 t_ref_count.JB112852
## 1                  0                  0                  0
## 2                  0                  0                  0
## 3                  0                  0                  0
## 4                  0                  0                  0
## 5                  0                  0                  0
## 6                  0                  0                  0
##   t_ref_count.KA3947 t_ref_count.LO3793 t_ref_count.LO5004
## 1                  0                  0                  0
## 2                  0                  0                  0
```

```
## 3                  0                  0                  0
## 4                  0                  0                  0
## 5                  0                  0                  0
## 6                 21                  0                  0
##   t_ref_count.M4945 t_ref_count.MA7027 t_ref_count.NI9507
## 1                 0                  0                  0
## 2                 9                  0                  0
## 3                 0                  0                  0
## 4                 0                  0                  0
## 5                 0                  0                  0
## 6                 0                  0                  0
##   t_ref_count.R7495_2 t_ref_count.RH090935 t_ref_count.RI1933
## 1                   0                   0                  0
## 2                   0                   0                  0
## 3                   0                   0                  0
## 4                   0                   0                  0
## 5                   0                   0                  0
## 6                   0                   0                  0
##   t_ref_count.RO3338 t_ref_count.SA9755 t_ref_count.SB010944
## 1                  0                  0                   0
## 2                  0                 16                   0
## 3                  0                  0                   0
## 4                  0                  0                   0
## 5                  0                  0                   0
## 6                  0                  0                   0
##   t_ref_count.SC0899 t_ref_count.SC6470 t_ref_count.SR070761
## 1                  0                  0                   0
## 2                  0                  0                   0
## 3                  0                  0                   0
## 4                  0                  0                   0
## 5                  0                  0                   0
## 6                  0                 11                   8
##   t_ref_count.TU0428 t_ref_count.VA1330 t_ref_count.VA7859
## 1                  0                  0                  0
## 2                  0                  0                  0
## 3                  0                  0                  0
## 4                  0                  0                  0
## 5                  0                  0                  0
## 6                  0                  0                  0
##   t_ref_count.WA7899 t_ref_count.Y2087 t_ref_count.ZA6505
## 1                  0                 9                  0
## 2                  0                 0                  0
## 3                  0                 0                  0
## 4                  0                 0                  0
## 5                  0                 7                  0
## 6                  0                 0                 10
##   t_ref_count.ZA6965
## 1                  0
## 2                  0
## 3                  0
## 4                  0
## 5                  0
## 6                  0
```

```
dim(finalseqdata)

## [1] 5698   35
```

Preparamos el archivo de muestras.

```
sampleinfo=read.delim("/Users/llorencrubiovives/Desktop/data_clinical.
txt")
```

Creamos el objeto countdata.

```
countdata=finalseqdata[,-(1)]

head(countdata)

##   t_ref_count.AL4602 t_ref_count.AU5884 t_ref_count.BL3403
## 1                  7                  0                  0
## 2                 10                  0                  0
## 3                  6                  0                  0
## 4                  7                  0                  0
## 5                  8                  0                  0
## 6                  7                  0                  0
##   t_ref_count.CA9903 t_ref_count.CU9061 t_ref_count.DI6359
## 1                  0                  0                  0
## 2                  0                  0                  0
## 3                  0                  0                  0
## 4                  0                  0                  0
## 5                  0                  0                  0
## 6                  0                  0                  0
##   t_ref_count.DM123062 t_ref_count.FR9547 t_ref_count.GR0134
## 1                    0                  0                  0
## 2                    0                  0                  0
## 3                    0                  0                  0
## 4                    0                  0                  0
## 5                    0                  0                  0
## 6                    0                  0                  0
##   t_ref_count.GR4788 t_ref_count.HE3202 t_ref_count.JB112852
## 1                  0                  0                  0
## 2                  0                  0                  0
## 3                  0                  0                  0
## 4                  0                  0                  0
## 5                  0                  0                  0
## 6                  0                  0                  0
##   t_ref_count.KA3947 t_ref_count.LO3793 t_ref_count.LO5004
## 1                  0                  0                  0
## 2                  0                  0                  0
## 3                  0                  0                  0
## 4                  0                  0                  0
## 5                  0                  0                  0
## 6                 21                  0                  0
##   t_ref_count.M4945 t_ref_count.MA7027 t_ref_count.NI9507
## 1                 0                  0                  0
## 2                 9                  0                  0
## 3                 0                  0                  0
## 4                 0                  0                  0
## 5                 0                  0                  0
## 6                 0                  0                  0
```

```
##   t_ref_count.R7495_2 t_ref_count.RH090935 t_ref_count.RI1933
## 1                   0                    0                  0
## 2                   0                    0                  0
## 3                   0                    0                  0
## 4                   0                    0                  0
## 5                   0                    0                  0
## 6                   0                    0                  0
##   t_ref_count.RO3338 t_ref_count.SA9755 t_ref_count.SB010944
## 1                  0                  0                    0
## 2                  0                 16                    0
## 3                  0                  0                    0
## 4                  0                  0                    0
## 5                  0                  0                    0
## 6                  0                  0                    0
##   t_ref_count.SC0899 t_ref_count.SC6470 t_ref_count.SR070761
## 1                  0                  0                    0
## 2                  0                  0                    0
## 3                  0                  0                    0
## 4                  0                  0                    0
## 5                  0                  0                    0
## 6                  0                 11                    8
##   t_ref_count.TU0428 t_ref_count.VA1330 t_ref_count.VA7859
## 1                  0                  0                  0
## 2                  0                  0                  0
## 3                  0                  0                  0
## 4                  0                  0                  0
## 5                  0                  0                  0
## 6                  0                  0                  0
##   t_ref_count.WA7899 t_ref_count.Y2087 t_ref_count.ZA6505
## 1                  0                 9                  0
## 2                  0                 0                  0
## 3                  0                 0                  0
## 4                  0                 0                  0
## 5                  0                 7                  0
## 6                  0                 0                 10
##   t_ref_count.ZA6965
## 1                  0
## 2                  0
## 3                  0
## 4                  0
## 5                  0
## 6                  0
```

```
colnames(countdata)=substr(colnames(countdata),start=13, stop=100)
```

```
head(countdata)
```

```
##                 AL4602 AU5884 BL3403 CA9903 CU9061 DI6359 DM123062
FR9547
## ENSG00000131584      7      0      0      0      0      0        0
0
## ENSG00000179403     10      0      0      0      0      0        0
0
## ENSG00000197530      6      0      0      0      0      0        0
0
```

```
## ENSG00000049239      7      0      0      0      0      0        0
0
## ENSG00000130940      8      0      0      0      0      0        0
0
## ENSG00000116786      7      0      0      0      0      0        0
0
##                 GR0134 GR4788 HE3202 JB112852 KA3947 LO3793 LO5004
M4945
## ENSG00000131584      0      0      0        0      0      0      0
0
## ENSG00000179403      0      0      0        0      0      0      0
9
## ENSG00000197530      0      0      0        0      0      0      0
0
## ENSG00000049239      0      0      0        0      0      0      0
0
## ENSG00000130940      0      0      0        0      0      0      0
0
## ENSG00000116786      0      0      0        0     21      0      0
0
##                 MA7027 NI9507 R7495_2 RH090935 RI1933 RO3338 SA9755
## ENSG00000131584      0      0       0        0      0      0      0
## ENSG00000179403      0      0       0        0      0      0     16
## ENSG00000197530      0      0       0        0      0      0      0
## ENSG00000049239      0      0       0        0      0      0      0
## ENSG00000130940      0      0       0        0      0      0      0
## ENSG00000116786      0      0       0        0      0      0      0
##                 SB010944 SC0899 SC6470 SR070761 TU0428 VA1330
VA7859
## ENSG00000131584        0      0      0        0      0      0
0
## ENSG00000179403        0      0      0        0      0      0
0
## ENSG00000197530        0      0      0        0      0      0
0
## ENSG00000049239        0      0      0        0      0      0
0
## ENSG00000130940        0      0      0        0      0      0
0
## ENSG00000116786        0      0     11        8      0      0
0
##                 WA7899 Y2087 ZA6505 ZA6965
## ENSG00000131584      0     9      0      0
## ENSG00000179403      0     0      0      0
## ENSG00000197530      0     0      0      0
## ENSG00000049239      0     0      0      0
## ENSG00000130940      0     7      0      0
## ENSG00000116786      0     0     10      0
```

Comprobamos que las muestras de ambos archivos coinciden en orden.

```
table(colnames(countdata)==sampleinfo$SAMPLE_ID)
```

```
##
## TRUE
##   34
```

```
library(limma)
```

```
library(edgeR)
```

```
library(Glimma)
```

Filtrado de genes con baja expresión

Aplicamos count per million (CPM).

```
myCPM=cpm(countdata)
```

```
head(myCPM)
```

```
##                   AL4602 AU5884 BL3403 CA9903 CU9061 DI6359
DM123062
## ENSG00000131584 290.1434      0      0      0      0      0
0
## ENSG00000179403 414.4906      0      0      0      0      0
0
## ENSG00000197530 248.6944      0      0      0      0      0
0
## ENSG00000049239 290.1434      0      0      0      0      0
0
## ENSG00000130940 331.5925      0      0      0      0      0
0
## ENSG00000116786 290.1434      0      0      0      0      0
0
```

Establecemos como regla de CPM=1.

```
thresh=myCPM > 1
```

```
head(thresh)
```

```
##                 AL4602 AU5884 BL3403 CA9903 CU9061 DI6359 DM123062
FR9547
## ENSG00000131584   TRUE  FALSE  FALSE  FALSE  FALSE  FALSE    FALSE
FALSE
## ENSG00000179403   TRUE  FALSE  FALSE  FALSE  FALSE  FALSE    FALSE
FALSE
## ENSG00000197530   TRUE  FALSE  FALSE  FALSE  FALSE  FALSE    FALSE
FALSE
## ENSG00000049239   TRUE  FALSE  FALSE  FALSE  FALSE  FALSE    FALSE
FALSE
## ENSG00000130940   TRUE  FALSE  FALSE  FALSE  FALSE  FALSE    FALSE
FALSE
## ENSG00000116786   TRUE  FALSE  FALSE  FALSE  FALSE  FALSE    FALSE
FALSE
##                 GR0134 GR4788 HE3202 JB112852 KA3947 LO3793 LO5004
```

```
M4945
## ENSG00000131584  FALSE  FALSE  FALSE    FALSE  FALSE  FALSE  FALSE
FALSE
## ENSG00000179403  FALSE  FALSE  FALSE    FALSE  FALSE  FALSE  FALSE
TRUE
## ENSG00000197530  FALSE  FALSE  FALSE    FALSE  FALSE  FALSE  FALSE
FALSE
## ENSG00000049239  FALSE  FALSE  FALSE    FALSE  FALSE  FALSE  FALSE
FALSE
## ENSG00000130940  FALSE  FALSE  FALSE    FALSE  FALSE  FALSE  FALSE
FALSE
## ENSG00000116786  FALSE  FALSE  FALSE    FALSE   TRUE  FALSE  FALSE
FALSE
##                 MA7027 NI9507 R7495_2 RH090935 RI1933 RO3338 SA9755
## ENSG00000131584  FALSE  FALSE   FALSE    FALSE  FALSE  FALSE  FALSE
## ENSG00000179403  FALSE  FALSE   FALSE    FALSE  FALSE  FALSE   TRUE
## ENSG00000197530  FALSE  FALSE   FALSE    FALSE  FALSE  FALSE  FALSE
## ENSG00000049239  FALSE  FALSE   FALSE    FALSE  FALSE  FALSE  FALSE
## ENSG00000130940  FALSE  FALSE   FALSE    FALSE  FALSE  FALSE  FALSE
## ENSG00000116786  FALSE  FALSE   FALSE    FALSE  FALSE  FALSE  FALSE
##                 SB010944 SC0899 SC6470 SR070761 TU0428 VA1330
VA7859
```

Miramos el número de trues en cada línea.

```
table(rowSums(thresh))
```

```
##
##    1    2    3    4    5    6    7    8    9   10   11   13   14
16   19
## 4056 1099  331  121   50   20    5    3    4    3    2    1    1
1    1
```

Conservamos genes con un mínimo de 2 TRUES por línea.

```
keep=rowSums(thresh)>=2

counts.keep=countdata[keep,]

summary(keep)
```

```
##    Mode   FALSE    TRUE
## logical    4056    1642
```

```
dim(counts.keep)
```

```
## [1] 1642   34
```

Examinamos la primera muestra.

```
plot(myCPM[,1],countdata[,1])
```

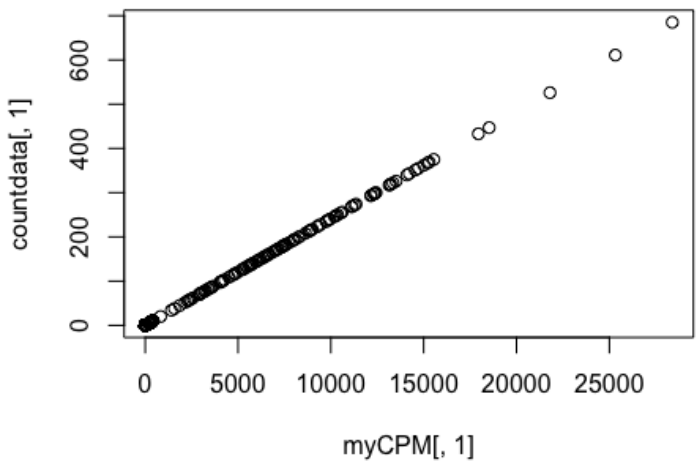


Figura 16. Filtrado por CPM en muestra 1

```
plot(myCPM[,1],countdata[,1],ylim=c(0,50),xlim=c(200,2000))
```

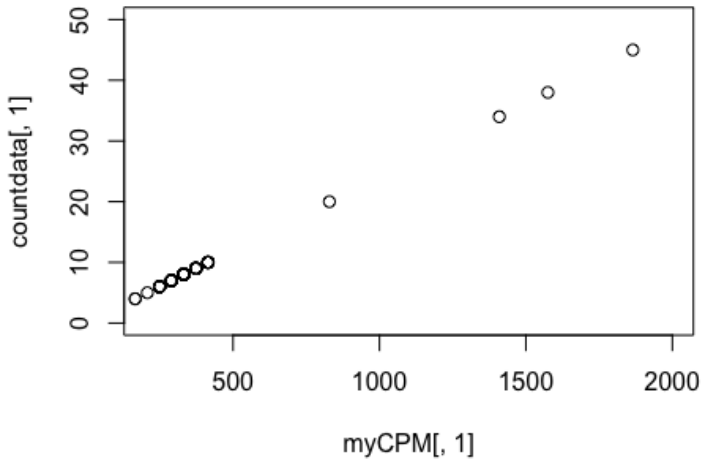


Figura 17. Distribución muestra 1 dentro del rango 500 a 2000

Volvemos a examinar la segunda muestra.

```
plot(myCPM[,2],countdata[,2])
```

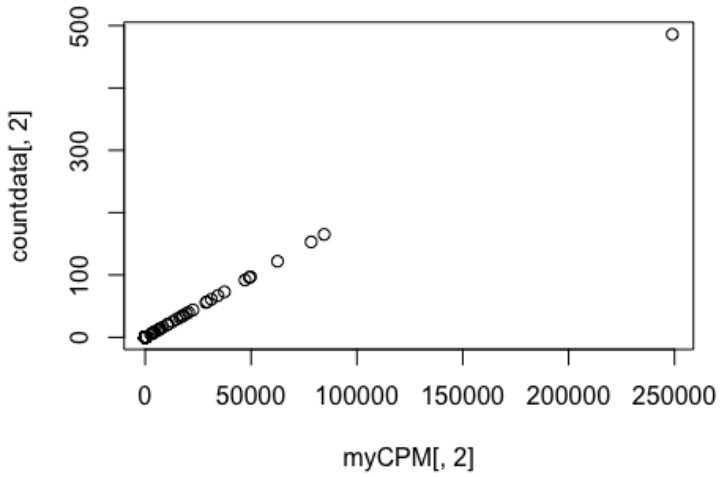


Figura 18. Filtrado por CPM en muestra 2

```
plot(myCPM[,2],countdata[,2],ylim=c(0,10),xlim=c(0,10))
```

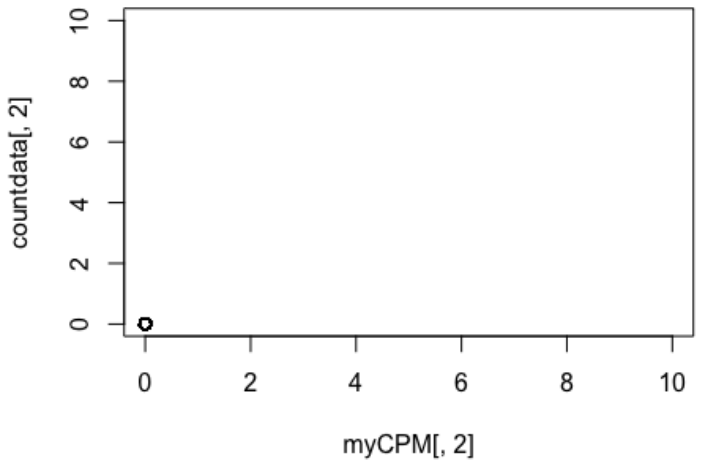


Figura 19. Distribución muestra 2 dentro del rango 500 a 2000

Convertimos recuentos a objeto DGEList.

```
y=DGEList(counts.keep)

y

## An object of class "DGEList"
## $counts
##                 AL4602 AU5884 BL3403 CA9903 CU9061 DI6359 DM123062
FR9547
## ENSG00000131584      7      0      0      0      0      0        0
0
## ENSG00000179403     10      0      0      0      0      0        0
```

Examinamos los datos guardados en y.

```
names(y)

## [1] "counts"  "samples"
```

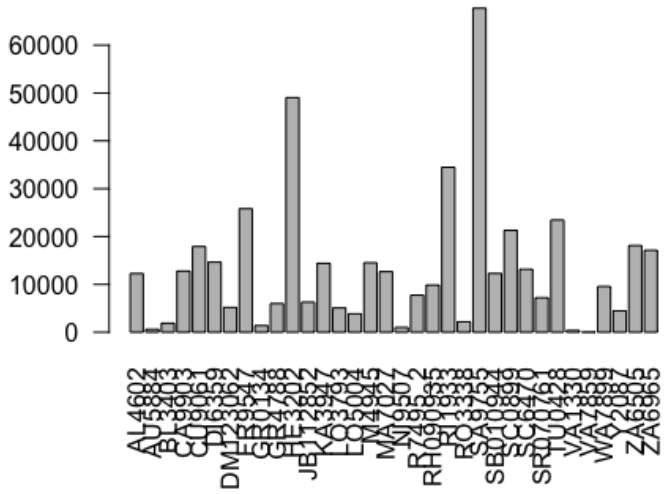
Control de calidad

Analizamos la cantidad de recuentos por cada muestra de y.

```
y$samples$lib.size

##  [1] 12269   601  1883 12778 17910 14641  5173 25813  1391  5994
49032
## [12]  6286 14400  5076  3857 14524 12658  1005  7746  9858 34476
2178
## [23] 67717 12284 21302 13149  7206 23459   373    23  9560  4447
18116
## [34] 17122
```

Analizamos el tamaño de las librerías.

```
barplot(y$samples$lib.size, names=colnames(y),las=2)
```

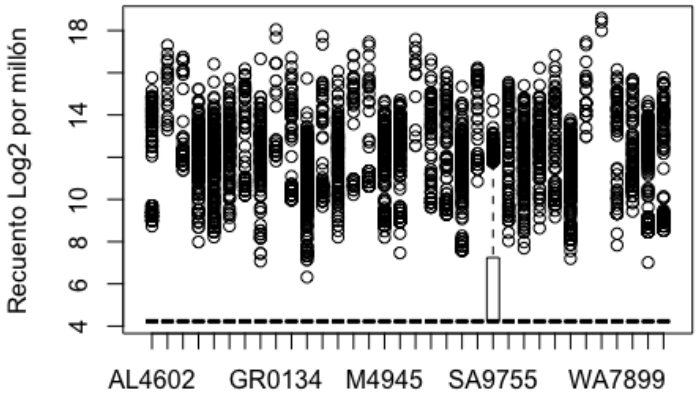


Observamos nuestros datos sin normalizar con corrección log.

```
logcounts=cpm(y,log=TRUE)
```

Plasmamos el recuento en diagrama de cajas.

```
boxplot(logcounts, xlab="", ylab="Recuento Log2 por millón")
```




```
##abline(h=median(logcounts),col="blue")

##título("Diagrama de cajas de logCPMs sin normalizar")
```

Creamos un gráfico multiescalar (MDS) para examinar la agrupación de datos.

```
library(edgeR)

plotMDS(y)
```

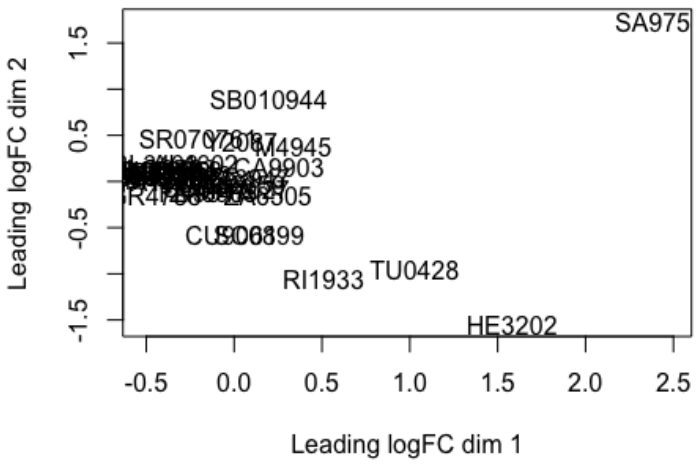
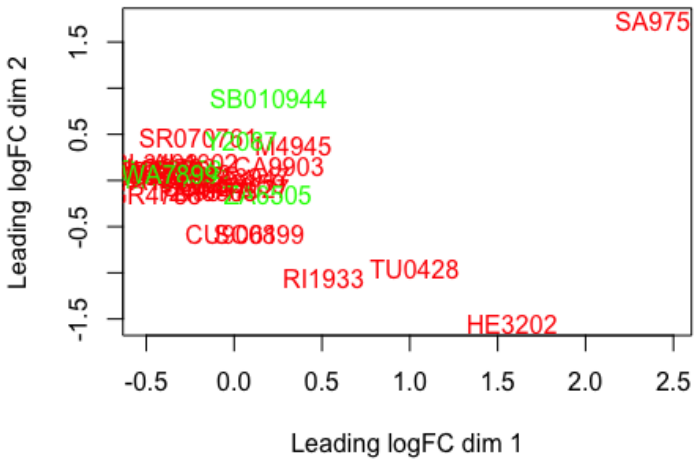


Figura 20. Distribución de datos en gráficos MDS

Procedemos a asignarles un color por grupo.

```
par(mfrow=c(1,2))

levels(sampleinfo$SMOKING_HISTORY)

## [1] "NO"  "YES"
```

```
col.cell=c("green","red")[sampleinfo$SMOKING_HISTORY]

data.frame(sampleinfo$SMOKING_HISTORY,col.cell)

##    sampleinfo.SMOKING_HISTORY col.cell
## 1                         YES      red
## 2                         YES      red
## 3                         YES      red
## 4                         YES      red
## 5                         YES      red
## 6                         YES      red
## 7                          NO    green
## 8                         YES      red
## 9                         YES      red
## 10                        YES      red
## 11                        YES      red
## 12                         NO    green
## 13                        YES      red
## 14                        YES      red
## 15                        YES      red
## 16                        YES      red
## 17                        YES      red
## 18                        YES      red
plotMDS(y,col=col.cell)
```



Generamos clusters jerárquicos con mapas de calor.

```
var_genes=apply(logcounts,1,var)

head(var_genes)

## ENSG00000131584 ENSG00000179403 ENSG00000130940 ENSG00000116786
##        2.016683        1.942270        1.934369        3.992006
## ENSG00000159363 ENSG00000142798
##        1.182668        3.934204

select_var=names(sort(var_genes,decreasing=TRUE))[1:250]
```

```
head(select_var)
```

```
## [1] "ENSG00000143631" "ENSG00000155657" "ENSG00000164796"
"ENSG00000141510"
## [5] "ENSG00000168702" "ENSG00000181143"
```

```
highly_variable_lcpm=logcounts[select_var,]
```

```
dim(highly_variable_lcpm)
```

```
## [1] 250  34
```

```
head(highly_variable_lcpm)
```

```
##                    AL4602   AU5884    BL3403    CA9903   CU9061
DI6359
## ENSG00000143631  4.225669 4.225669  4.225669  4.225669 16.41345
4.225669
## ENSG00000155657  4.225669 4.225669  4.225669 12.178617 13.50580
12.507821
## ENSG00000164796  4.225669 4.225669  4.225669 14.130849 14.05255
4.225669
```

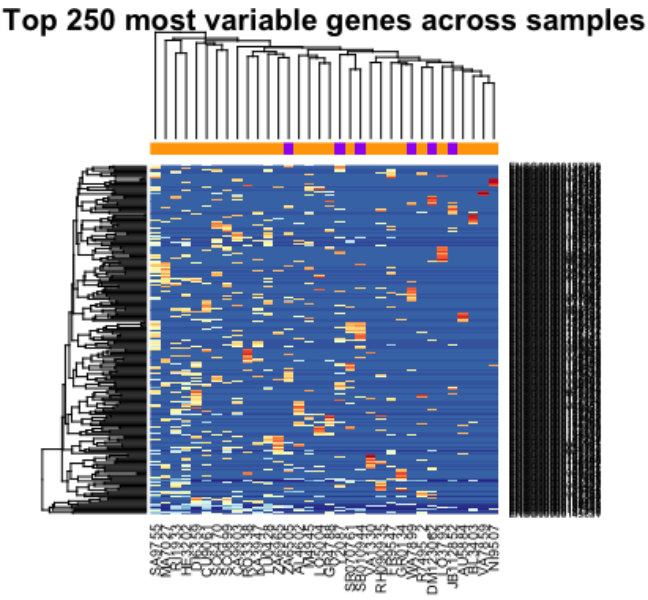
```
library(RColorBrewer)
```

```
mypalette=brewer.pal(11,"RdYlBu")
```

```
morecols=colorRampPalette(mypalette)
```

```
col.cell=c("purple","orange")[sampleinfo$SMOKING_HISTORY]
```

```
heatmap(highly_variable_lcpm,col=rev(morecols(50)),main="Top 250 most
variable genes across samples",ColSideColors=col.cell,scale="row")
```



Guardamos el mapa de genes más variables.

```
png(file="High_var_genes_heatmap.png")
```

Normalización de librerías

Utilizamos calcNormFactors para normalizar entre librerías.

```
z=calcNormFactors(y)
```

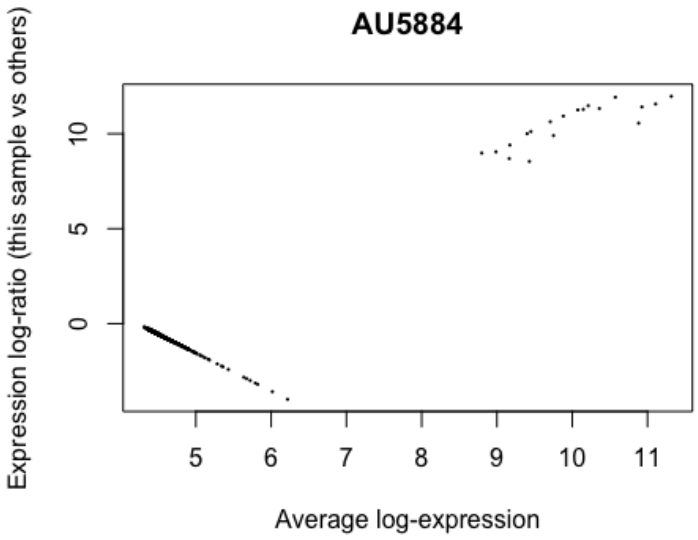
```
## Warning in max(abs(logR)): no non-missing arguments to max;
returning -Inf
```

```
z$samples
```

```
##          group lib.size norm.factors
## AL4602       1    12269    0.7949055
## AU5884       1      601    7.2983771
## BL3403       1     1883    2.8719975
## CA9903       1    12778    0.5773701
## CU9061       1    17910    0.4819198
## DI6359       1    14641    0.7974352
## DM123062     1     5173    1.0182362
## FR9547       1    25813    0.7046038
## GR0134       1     1391    6.2489408
## GR4788       1     5994    0.7744768
## HE3202       1    49032    0.2478255
## JB112852     1     6286    2.1912809
## KA3947       1    14400    0.3271162
## LO3793       1     5076    2.2903281
## LO5004       1     3857    1.0928097
## M4945        1    14524    0.6479460
## MA7027       1    12658    1.1597415
## NI9507       1     1005   13.6550653
## R7495_2      1     7746    1.1788112
## RH090935     1     9858    1.2929192
## RI1933       1    34476    0.2036825
## RO3338       1     2178    1.0496692
## SA9755       1    67717    0.2461262
## SB010944     1    12284    0.5644806
## SC0899       1    21302    0.8016483
## SC6470       1    13149    0.3276058
## SR070761     1     7206    0.5939380
## TU0428       1    23459    0.4008418
## VA1330       1      373    4.9750344
## VA7859       1       23    0.7949055
## WA7899       1     9560    1.3170096
## Y2087        1     4447    1.8294572
## ZA6505       1    18116    0.9924362
## ZA6965       1    17122    0.6379046
```

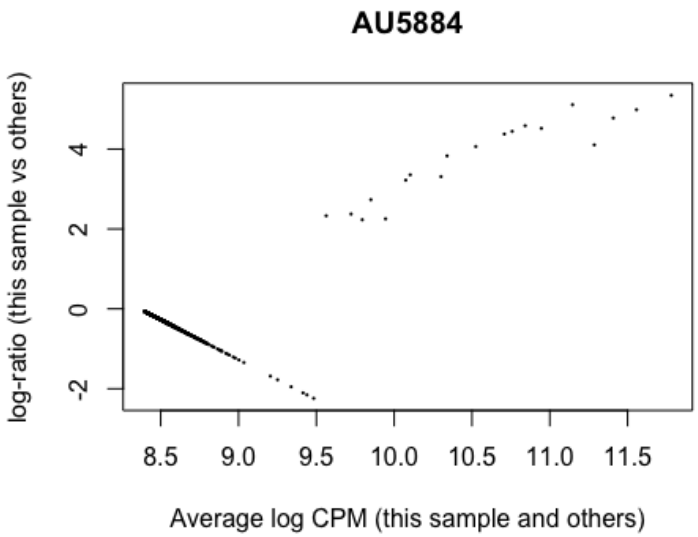
Observaremos el resultado obtenido en la muestra número 2, que muestra el mayor factor de normalización (observaremos en primer lugar el objeto logcounts (donde solo estaba normalizado por medida de la librería)).

```
par(mfrow=c(1,2))

plotMD(logcounts,column=2)
```



Ahora observamos la muestra según el objeto z.

```
par(mfrow=c(1,2))

plotMD(z,column=2)
```



Expresión diferencial

Creamos la matriz de diseño.

```
group=paste(sampleinfo$SMOKING_HISTORY)

group=factor(group)

group
```

```
##  [1] YES YES YES YES YES YES NO  YES YES YES YES NO  YES YES YES
YES YES
## [18] YES YES YES YES YES YES NO  YES YES YES YES YES YES NO  NO  NO
```

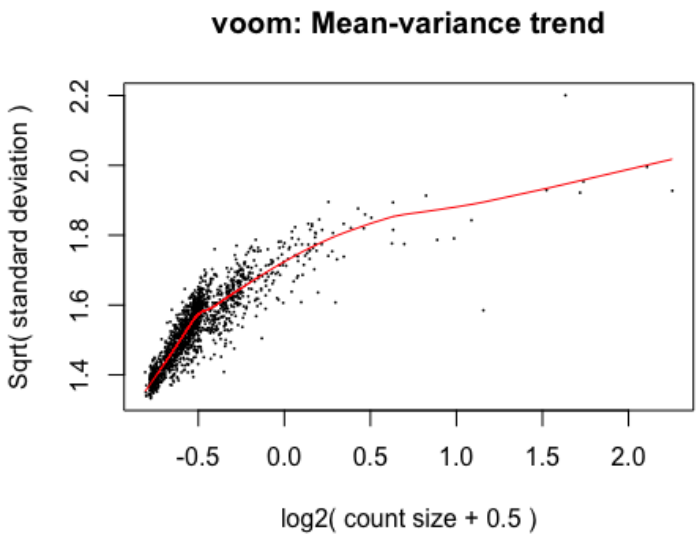
```
YES
## Levels: NO YES
```

Creamos el objeto para realizar el análisis diferencial sin intercepto.

```
design=model.matrix(~ 0+group)

design

##    groupNO groupYES
## 1        0        1
## 2        0        1
## 3        0        1
## 4        0        1
## 5        0        1
## 6        0        1
colnames(design)=levels(group)

design

##    NO YES
## 1   0   1
## 2   0   1
## 3   0   1
## 4   0   1
## 5   0   1
## 6   0   1
## attr(,"assign")
## [1] 1 1
## attr(,"contrasts")
## attr(,"contrasts")$group
## [1] "contr.treatment"
```

Utilizamos voom para ajustar automáticamente las medidas de librerías utilizando la normalización que hemos calculado anteriormente.

```
par(mfrow=c(1,2))
v=voom(z,design,plot=TRUE)
```

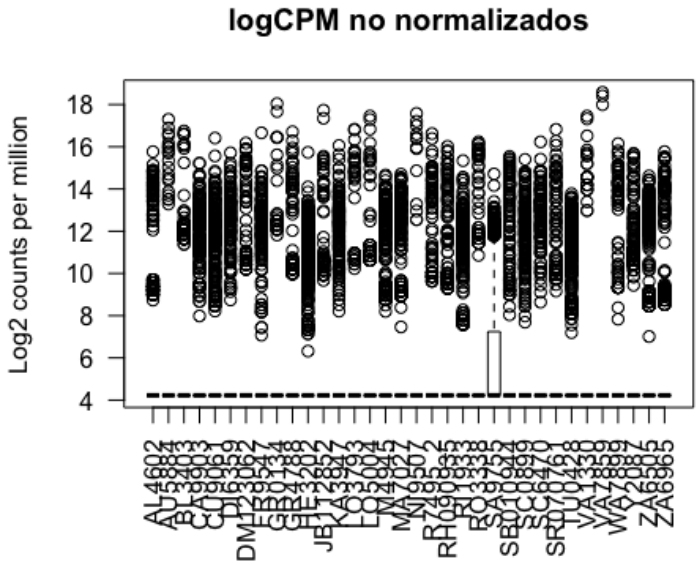


Ahora miramos qué contiene nuestro objeto voom.

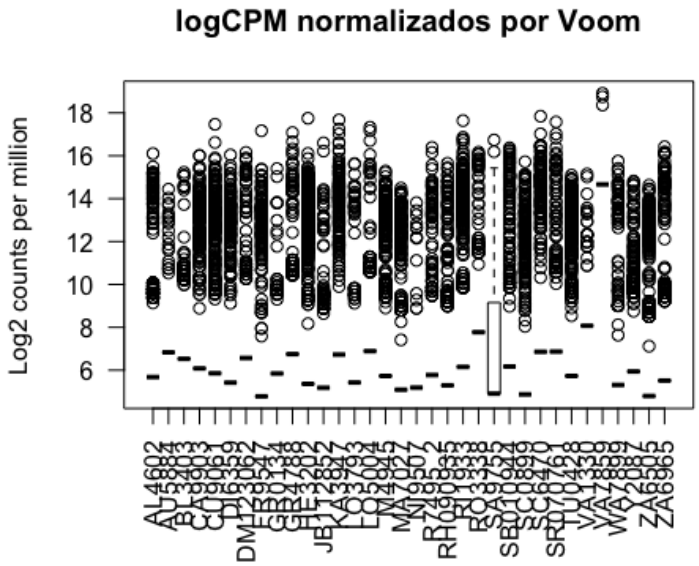
```
v

## An object of class "EList"
## $targets
##        group lib.size norm.factors
## AL4602     1 9752.696    0.7949055
## AU5884     1 4386.325    7.2983771
## BL3403     1 5407.971    2.8719975
## CA9903     1 7377.635    0.5773701
## CU9061     1 8631.184    0.4819198
## 29 more rows ...
##
## $E
##                    AL4602   AU5884  BL3403  CA9903   CU9061
DI6359
## ENSG00000131584  9.586726 6.832443 6.53043 6.08243 5.856059
5.420279
## ENSG00000179403 10.072153 6.832443 6.53043 6.08243 5.856059
5.420279
## ENSG00000130940  9.767298 6.832443 6.53043 6.08243 5.856059
5.420279
## ENSG00000116786  9.586726 6.832443 6.53043 6.08243 5.856059
5.420279
## ENSG00000159363  9.586726 6.832443 6.53043 6.08243 5.856059
5.420279
##                 DM123062   FR9547   GR0134   GR4788   HE3202
JB112852
## ENSG00000131584 6.568437 4.780795 5.845884 6.750661 5.362617
5.181763
## ENSG00000179403 6.568437 4.780795 5.845884 6.750661 5.362617 ##
1637 more rows ...
##
## $design
##   NO YES
## 1  0   1
## 2  0   1
## 3  0   1
## 4  0   1
## 5  0   1
## 29 more rows ...
```

Ahora miraremos nuestros datos normalizados mediante diagramas de cajas.

```
par(mfrow=c(1,2))

boxplot(logcounts,xlab="",ylab="Log2 counts per
million",las=2,main="logCPM no normalizados")
```




```
boxplot(v$E,xlab="",ylab="Log2 counts per million", las=2,
main="logCPM normalizados por Voom")
```



Para explorar la diferencia de expresión hacemos el ajuste del modelo.

```
fit=lmFit(v)

names(fit)

##  [1] "coefficients"     "stdev.unscaled"   "sigma"
##  [4] "df.residual"      "cov.coefficients" "pivot"
##  [7] "rank"             "Amean"            "method"
## [10] "design"
```

```
design
```

```
##    NO YES
## 1   0   1
## 2   0   1
## 3   0   1
## 4   0   1
## 5   0   1
## 6   0   1
## attr(,"assign")
## [1] 1 1
## attr(,"contrasts")
## attr(,"contrasts")$group
## [1] "contr.treatment"
```

Seguidamente miramos los resultados comparativos en nuestros 2 niveles.

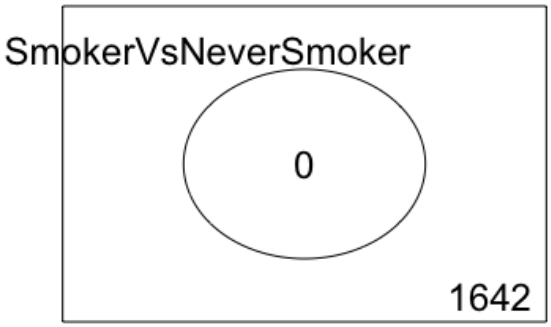
```
cont.matrix=makeContrasts(SmokerVsNeverSmoker=NO-YES, levels=design)

cont.matrix

##       Contrasts
## Levels SmokerVsNeverSmoker
##    NO                    1
##    YES                  -1
```

Aplicamos la matriz de contrastes.

```
fit.cont=contrasts.fit(fit,cont.matrix)
```

Aplicamos el shrink de Bayes.

```
fit.cont=eBayes(fit.cont)

dim(fit.cont)

## [1] 1642    1

summa.fit=decideTests(fit.cont)

summary(summa.fit)

##        SmokerVsNeverSmoker
## Down                     0
## NotSig                1642
## Up                       0

vennDiagram(summa.fit)
```



Investigamos cuáles son los 10 genes más relevantes en ambos casos.

```
topTable(fit.cont,coef="SmokerVsNeverSmoker",sort.by="p")
```

```
##                     logFC  AveExpr         t      P.Value adj.P.Val
## ENSG00000164796 -4.338128 8.933749 -3.947018 7.923365e-05 0.1301017
## ENSG00000154358 -3.334135 8.083976 -3.116203 1.832950e-03 0.2920500
## ENSG00000133703 -3.093202 7.827856 -2.949194 3.187439e-03 0.2920500
## ENSG00000141837 -3.002128 7.820985 -2.864258 4.181531e-03 0.2920500
## ENSG00000178104 -2.898789 7.374194 -2.846167 4.426616e-03 0.2920500
## ENSG00000183117 -2.914022 7.501429 -2.843360 4.465787e-03 0.2920500
## ENSG00000169876 -2.934627 7.826866 -2.798268 5.139617e-03 0.2920500
## ENSG00000134516 -2.735758 7.398169 -2.682783 7.303518e-03 0.2920500
## ENSG00000112079 -2.558351 6.790570 -2.637967 8.342931e-03 0.2920500
## ENSG00000116183 -2.726950 7.657483 -2.632274 8.484019e-03 0.2920500
##                         B
## ENSG00000164796 -3.245817
## ENSG00000154358 -3.757787
## ENSG00000133703 -3.832867
## ENSG00000141837 -3.881772
## ENSG00000178104 -3.860070
## ENSG00000183117 -3.868789
## ENSG00000169876 -3.919983
## ENSG00000134516 -3.956446
## ENSG00000112079 -3.932803
## ENSG00000116183 -3.999386
```

Anotación de genes

Añadimos la anotación de genes y grabamos los resultados.

```
library("org.Hs.eg.db")
```

```
## Loading required package: AnnotationDbi

## Loading required package: stats4

## Loading required package: BiocGenerics
```

```
## Loading required package: parallel

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB

## The following object is masked from 'package:limma':
##
##     plotMA

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, cbind, colMeans,
##     colnames, colSums, do.call, duplicated, eval, evalq, Filter,
##     Find, get, grep, grepl, intersect, is.unsorted, lapply,
##     lengths, Map, mapply, match, mget, order, paste, pmax,
##     pmax.int, pmin, pmin.int, Position, rank, rbind, Reduce,
##     rowMeans, rownames, rowSums, sapply, setdiff, sort, table,
##     tapply, union, unique, unsplit, which, which.max, which.min

## Loading required package: Biobase

## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.

## Loading required package: IRanges

## Loading required package: S4Vectors

##
## Attaching package: 'S4Vectors'

## The following object is masked from 'package:base':
##
##     expand.grid

##

columns(org.Hs.eg.db)

##  [1] "ACCNUM"       "ALIAS"        "ENSEMBL"      "ENSEMBLPROT"
##  [5] "ENSEMBLTRANS" "ENTREZID"     "ENZYME"       "EVIDENCE"
##  [9] "EVIDENCEALL"  "GENENAME"     "GO"           "GOALL"
## [13] "IPI"          "MAP"          "OMIM"         "ONTOLOGY"
## [17] "ONTOLOGYALL"  "PATH"         "PFAM"         "PMID"
```

```
## [21] "PROSITE"      "REFSEQ"       "SYMBOL"       "UCSCKG"
## [25] "UNIGENE"      "UNIPROT"
```

Seleccionamos las columnas, Ensembl, Name y Symbol.

```
tabfitcont=as.data.frame(fit.cont)

tabfitcont

##                 coefficients stdev.unscaled     sigma df.residual
Amean
## ENSG00000131584 -0.254146807      1.0581878 0.9152279          32
6.435485
## ENSG00000179403 -1.394131103      0.9603543 0.8737874          32
6.598127
## ENSG00000130940 -0.268327090      1.0443323 0.9140814          32
6.430766
## ENSG00000116786 -1.662205519      1.0835457 1.0758213          32

fit.cont$symbol=mapIds(org.Hs.eg.db, keys=rownames(fit.cont),
column="SYMBOL",keytype="ENSEMBL",multiVals="first")

## 'select()' returned 1:many mapping between keys and columns

fit.cont$entrez=mapIds(org.Hs.eg.db,
keys=row.names(fit.cont),column="ENTREZID",keytype="ENSEMBL",multiVals
="first")

## 'select()' returned 1:many mapping between keys and columns

fit.cont$name=mapIds(org.Hs.eg.db,
keys=row.names(fit.cont),column="GENENAME",keytype="ENSEMBL",multiVals
="first")

## 'select()' returned 1:many mapping between keys and columns

results=as.data.frame(fit.cont)
```

La última columna de nuestra matriz muestra el nombre de cada gen, el símbolo y el código Ensembl.

Ensembl ID	Gen	Función biológica	Función molecular
ENSG00000131584	ACAP3	Arf-GAP with coiled-coil, ANK repeat and PH domain-containing protein 3	
ENSG00000179403	VWA1	von Willebrand factor A domain-containing protein 1	
ENSG00000130940	CASZ1	cell differentiation, cellular process, nervous system development, regulation of biological process, single-multicellular organism process	
ENSG00000116786	PLEKHM2	Pleckstrin homology domain-containing family M member 2	
ENSG00000159363	ATP13A2	Probable cation-transporting ATPase 13A2	

Ensembl ID	Gen	Función biológica	Función molecular
ENSG00000142798	HSPG2	cell-cell adhesion, cell-matrix adhesion, ectoderm development, neurological system process, signal transduction,	receptor activity
ENSG00000176083	ZNF683	Tissue-resident T-cell transcription regulator protein ZNF683	
ENSG00000117713	ARID1A	AT-rich interactive domain-containing protein 1A	Wnt signaling pathway: Switched/Sucrose Non Fermentation
ENSG00000116544	DLGAP3	neurological system process	n/a
ENSG00000162624	LHX8	anatomical structure morphogenesis, cellular defense response, ectoderm development, embryo development, mesoderm development	RNA binding
ENSG00000215853	RPTN	Repetin	
ENSG00000197915	HRNR	Hornerin	
ENSG00000160783	PMF1	Polyamine-modulated factor 1	
ENSG00000143297	FCRL5	Fc receptor-like protein 5	
ENSG00000158477	CD1A	antigen processing and presentation	antigen binding, lipid binding
ENSG00000196184	OR10J1	G-protein coupled receptor signaling pathway, regulation of biological process, response to stimulus	receptor activity, signal transducer activity
ENSG00000085552	IGSF9	cell surface receptor signaling pathway	n/a
ENSG00000163531	NFASC	Neurofascin	
ENSG00000133019	CHRM3	intracellular signal transduction, neurological system process, regulation of biological process, response to stimulus, synaptic transmission	G-protein coupled receptor activity, binding, signal transducer activity
ENSG00000228198	OR2M3	regulation of biological process, response to stimulus, sensory perception of smell	receptor activity, signal transducer activity
ENSG00000216937	CCDC7	Coiled-coil domain-containing protein 7	
ENSG00000176769	TCERG1L	biosynthetic process, cellular process, nitrogen compound metabolic process	RNA polymerase II transcription factor binding transcription factor activity, transcription cofactor activity
ENSG00000152270	PDE3B	sensory perception, signal transduction, visual perception	
ENSG00000181830	SLC35C1	biosynthetic process, carbohydrate metabolic process,	transmembrane transporter activity

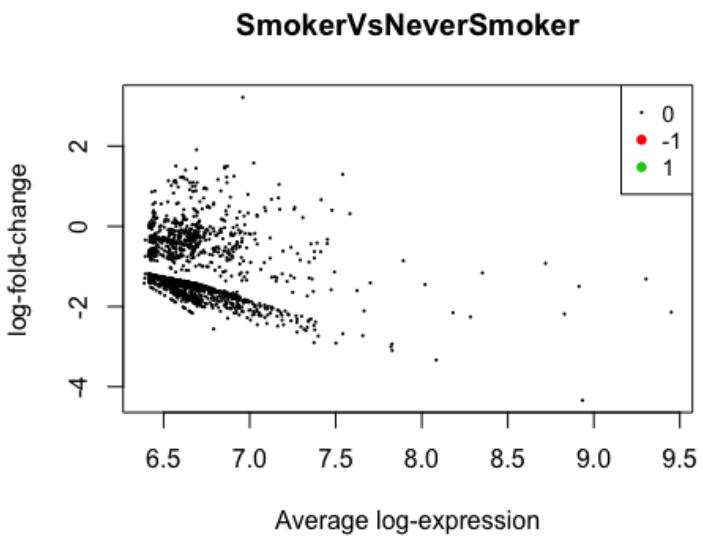
Ensembl ID	Gen	Función biológica	Función molecular
		cellular process, nucleobase-containing compound transport, protein glycosylation	
ENSG00000181939	OR4C15	G-protein coupled receptor signaling pathway, regulation of biological process, response to stimulus	receptor activity, signal transducer activity
ENSG00000150261	OR8K1	G-protein coupled receptor signaling pathway, regulation of biological process, response to stimulus, sensory perception of smell	binding, receptor activity, signal transducer activity
ENSG00000162302	RPS6KA4		PDGF signaling pathway ↳ Ribosomal protein S6 kinase, 90kD Insulin/IGF pathway-mitogen activated protein kinase kinase/MAP kinase cascade, ↳ Ribosomal protein S6 kinase, 90kD p38 MAPK pathway, ↳ mitogen- and stress-activated protein kinase 2
ENSG00000197891	SLC22A12	Solute carrier family 22 member 12	n/a
ENSG00000014216	CAPN1	proteolysis	cysteine-type peptidase activity
ENSG00000149256	TENM4	cell differentiation, cell-cell adhesion, cellular process, nervous system development, single-multicellular organism process	protein binding
ENSG00000165494	PCF11	biosynthetic process, cellular process, mRNA polyadenylation, nitrogen compound metabolic process, termination of RNA polymerase II transcription	mRNA binding, protein binding
ENSG00000133703	KRAS	G-protein coupled receptor signaling pathway, I-kappaB kinase/NF-kappaB cascade, MAPK cascade, cell adhesion, intracellular protein transport, neurological system process,	GTPase activity, protein binding

Ensembl ID	Gen	Función biológica	Función molecular
		receptor-mediated endocytosis, synaptic transmission	
ENSG00000167566	NCKAP5L	NCK associated protein 5 like	n/a
ENSG00000155974	GRIP1	glutamate receptor interacting protein 1	
ENSG00000111647	UHRF1BP1L	UHRF1 binding protein 1 like	
ENSG00000089169	RPH3A	rabphilin 3A	
ENSG00000196498	NCOR2	biosynthetic process, cellular process, nitrogen compound metabolic process, regulation of transcription from RNA, polymerase II promoter	receptor binding, transcription cofactor activity
ENSG00000102804	TSC22D1	zinc finger protein 862	n/a
ENSG00000148384	INPP5E	inositol polyphosphate-5-phosphatase E	n/a
ENSG00000175984	DENND2C	DENN domain containing 2C. SUPPRESSION OF TUMORIGENICITY 5 ST5 (PTHR15288)	
ENSG00000087266	SH3BP2	SH3 domain-binding protein 2	
ENSG00000197363	ZNF517	zinc finger protein 517	n/a

Tabla 3. Tabla de enriquecimiento genes más sobreexpresados [17]

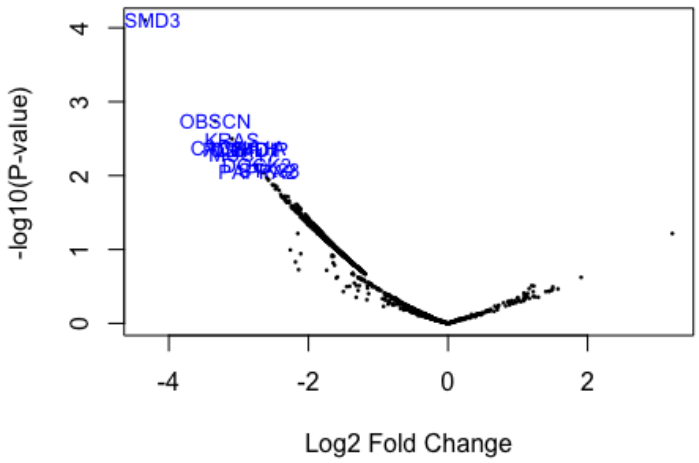
Creamos los gráficos para testear la expresión diferencial.

```
par(mfrow=c(1,2))

plotMD(fit.cont,coef=1,status=summa.fit[,], values = c(-1,1))
```



Creamos un diagrama de tipo volcán con los 10 genes con mayor expresión diferencial.

```
volcanoplot(fit.cont,coef=1,highlight = 10,names=fit.cont$symbol)
```



No realizaremos un enriquecimento, dado que no hay DEs.

```
go=goana(fit.cont,coef="SmokerVsNeverSmoker", species="Hs")

## No DE genes

topGO(go, n=10)

## data frame with 0 columns and 0 rows
```

Anexo 2

Método Cox

Ajustamos únicamente por la variable de interés.

```
coxph(Y~SmokingH,data=survival)

## Call:
## coxph(formula = Y ~ SmokingH, data = survival)
##
##             coef exp(coef) se(coef)     z    p
## SmokingH -0.575     0.562    0.512 -1.12 0.26
##
## Likelihood ratio test=1.13  on 1 df, p=0.3
## n= 34, number of events= 22
```

Así, obtenemos el riesgo de fallo estimado.

```
summary(coxph(Y~SmokingH,data=survival))

## Call:
## coxph(formula = Y ~ SmokingH, data = survival)
##
##   n= 34, number of events= 22
##
##              coef exp(coef) se(coef)      z Pr(>|z|)
## SmokingH -0.5754    0.5625   0.5120 -1.124    0.261
##
##          exp(coef) exp(-coef) lower .95 upper .95
## SmokingH    0.5625      1.778    0.2062     1.534
##
## Concordance= 0.558  (se = 0.046 )
## Rsquare= 0.033   (max possible= 0.98 )
## Likelihood ratio test= 1.13  on 1 df,   p=0.3
## Wald test            = 1.26  on 1 df,   p=0.3
## Score (logrank) test = 1.3  on 1 df,   p=0.3
```

Añadimos al modelo una covariante más.

```
modelo2=coxph(Y~SmokingH+Signature,data=survival)

modelo2

## Call:
## coxph(formula = Y ~ SmokingH + Signature, data = survival)
##
##                              coef exp(coef) se(coef)    z      p
## SmokingH                    0.208     1.232    0.554 0.38 0.7070
## Signaturetransversion low 1.465     4.328    0.527 2.78 0.0054
##
## Likelihood ratio test=9.41  on 2 df, p=0.009
## n= 34, number of events= 22
```

Proseguimos con el modelo con interacciones:

```
modelo3=coxph(Y~SmokingH+Signature+SmokingH:Signature,data=survival)

modelo3
```

```
## Call:
## coxph(formula = Y ~ SmokingH + Signature + SmokingH:Signature,
##     data = survival)
##
##                                      coef exp(coef) se(coef)    z
p
## SmokingH                            0.208     1.232    0.554 0.38
0.7070
## Signaturetransversion low           1.465     4.328    0.527 2.78
0.0054
## SmokingH:Signaturetransversion low    NA        NA    0.000   NA
NA
##
## Likelihood ratio test=9.41  on 2 df, p=0.009
## n= 34, number of events= 22
```

Valoraremos la interacción, para cerciorar las conclusiones:

```
LRT=(-2)*(modelo2$loglik[2]-modelo3$loglik[2])

LRT

## [1] 0

Pvalue=1-pchisq(LRT,2)

Pvalue

## [1] 1
```

Procedemos a evaluar la asunción de riesgos proporcionales y así valorar si el modelo escogido es el adecuado:

1. Utilizamos la función KM que se muestra a continuación:

```
plot(kmfit2,fun="cloglog",xlab="Tiempo en meses en escala
logarítmica",ylab="supervivencia log log", main="Curva log log
por historial de tabaquismo")
```

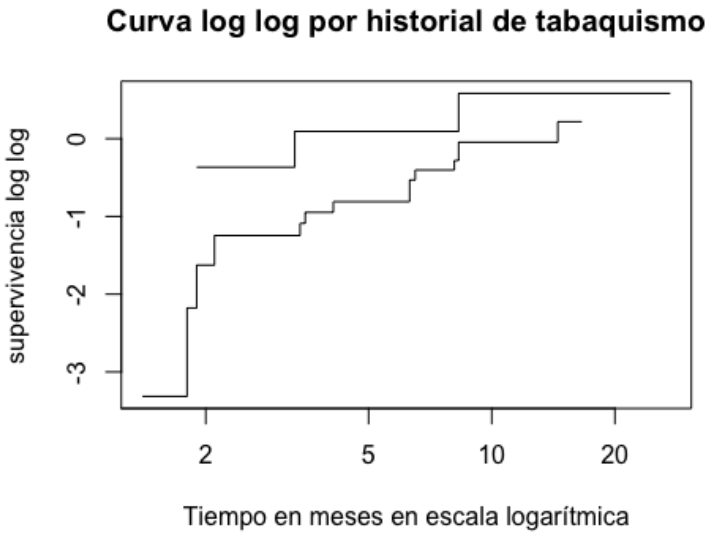


Figura 21. Comparativa KM en escala log – Fumadores / no fumadores

La asunción de riesgos proporcionales se cumple de forma relativa dentro de nuestra variable. Podría ser por el desequilibrio dentro de la muestra que estamos procesando.

Como hemos estratificado por la variable Signature, procedemos a validar la asunción PH para esa variable también:

```
plot(kmfit3,fun="cloglog",xlab="Tiempo en meses en escala
logarítmica",ylab="supervivencia log log", main="Curva log log
por signatura")
```

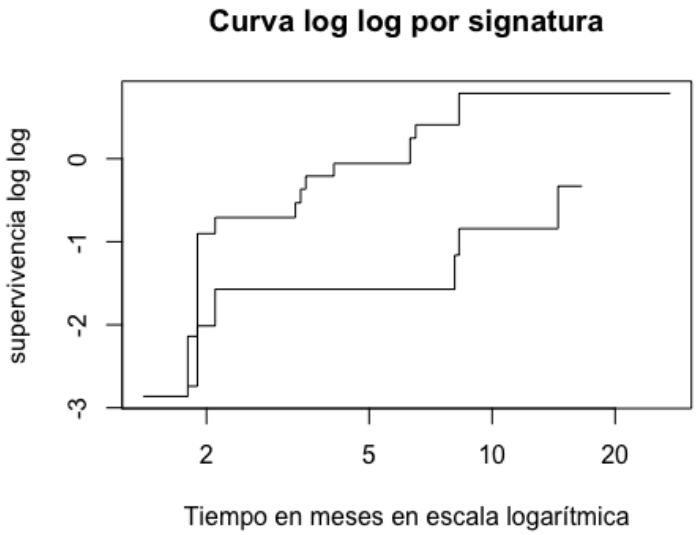


Figura 22. Comparativa KM en escala logarítima – Curvas por firma alta / baja

Hay una intersección de la variable al inicio de la gráfica, sin embargo posteriormente evolucionan ambas curvas de forma paralela, aunque con cierta distancia.
Estas observaciones no periten concluir la violación o cumplimiento de PH, por lo que probamos con otro tipo de test.

2. Método Schönfeld
   Ahora procederemos a probar la asunción mediante un test estadístico de Schönfeld y los residuales.

```
modelo1=coxph(Y~SmokingH, data=survival)
```

Para ello utilizaremos la función cox.zph para nuestros modelos reducidos y el que contiene interacciones:

```
ph1=cox.zph(modelo1, transform=rank)

ph1

##            rho chisq    p
## SmokingH 0.109 0.261 0.61

modelo2=coxph(Y~SmokingH+Signature, data=survival)

ph2=cox.zph(modelo2, transform=rank)
```

```
ph2

##                                rho chisq     p
## SmokingH                     0.226 1.187 0.276
## Signaturetransversion low 0.131 0.362 0.547
## GLOBAL                         NA 1.234 0.540

modelo3=coxph(Y~SmokingH+Signature+SmokingH:Signature,
data=survival)

ph3=cox.zph(modelo3, transform=rank)

## Warning in cor(xx, r2): the standard deviation is zero

ph3
```

```
##                                       rho chisq     p
## SmokingH                            0.226 1.187 0.276
## Signaturetransversion low           0.131 0.362 0.547
## SmokingH:Signaturetransversion low     NA   NaN   NaN
## GLOBAL                                 NA 1.234 0.745
```

En ninguno de los tres casos (simple, dos covariables, con interacciones) el test es estadísticamente significativo para ninguna de las covariables. Asimismo, el test global tampoco lo es. Podemos entonces asumir la proporcionalidad de riesgos.

La correlación entre los residuales de Schönfeld para la variable SmokingH y el tiempo de supervivencia es de 0.109 con un p-valor de 0.61. Este valor confirma que la asunción de riesgos proporcionales se cumple en nuestro modelo para la variable.

A pesar de estos valores, al representar la curva obtenemos esta forma, lejos del modelo horizontal que debería aparecer en el caso de no violar la asunción de riesgos proporcionales puesto que indicaría que los residuales son independientes del tiempo de supervivencia.

```
ggcoxzph(ph1)
```

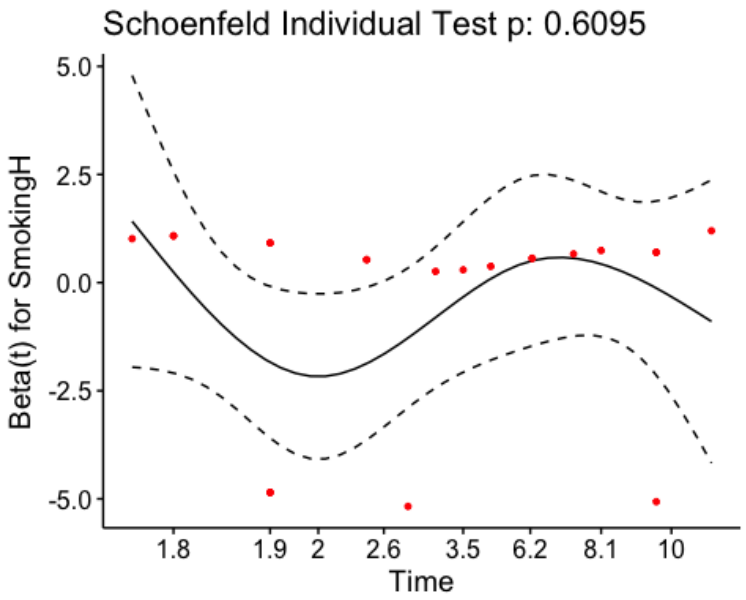


Figura 23. Gráfica Schönfeld con Beta (t) para variable fumador

```
ggcoxzph(ph2)
```

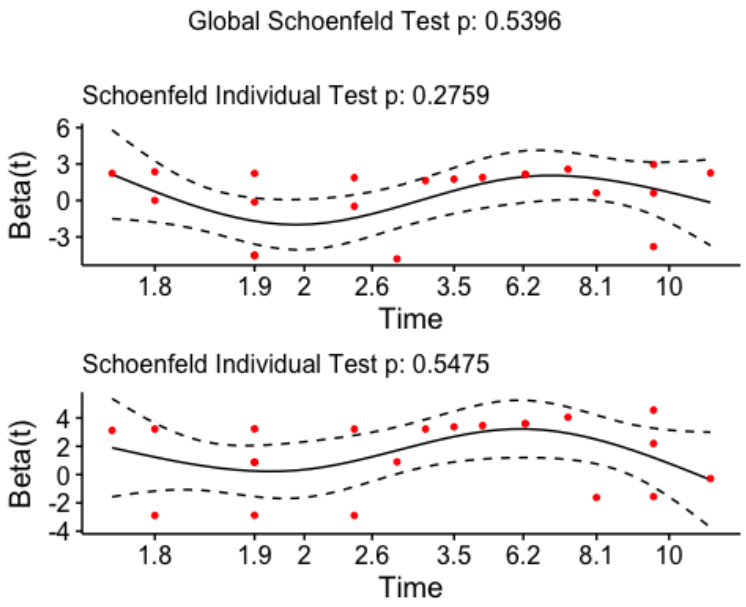


Figura 24. Gráfica Schönfeld con Beta (t) para variable fumador y firma

```
ggcoxzph(ph3)
```

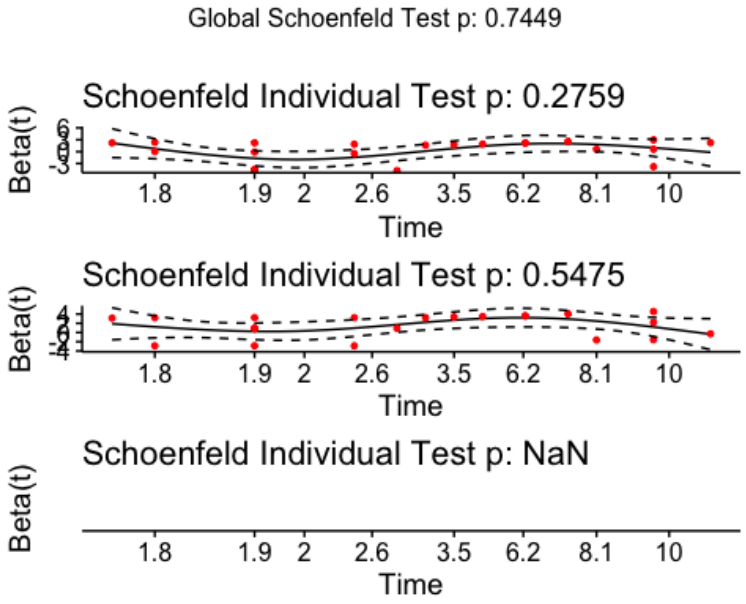


Figura 25. Gráfica Schönfeld con Beta (t) para variable fumador, firma con interacciones

En ninguno de los dos modelos anteriores las líneas son totalmente horizontales, así para los casos en que muestren una ligera curvatura, esas variables violan la asunción.

3. Decidimos realizar un modelo extendido de Cox dados los resultados de violación de la asunción de PH:

```
survival.cp=survSplit(survival,cut=survival$Death[survival$Censo
ring==1],end="Death",event="Censoring",start="start",id="ID")
```

Con este nuevo data set que contiene el formato principìo-fin, podemos modelar la variable SmokingH para hacerla dependiente del tiempo:

```
survival.cp$logtSmokingH=survival.cp$SmokingH*log(survival.cp$De
ath)
```

La nueva variable cambia a lo largo del tiempo, por lo que podemos concluir que la PH se cumple para SmokingH y podemos aceptar los resutados de nuestros análisis.

```
coxph(Surv(survival.cp$start,survival.cp$Death,survival.cp$Censo
ring)~SmokingH+logtSmokingH+cluster(ID),data=survival.cp)

## Warning in Surv(survival.cp$start, survival.cp$Death,
survival.cp
## $Censoring): Stop time must be > start time, NA created

## Call:
## coxph(formula = Surv(survival.cp$start, survival.cp$Death,
survival.cp$Censoring) ~
##     SmokingH + logtSmokingH + cluster(ID), data =
survival.cp)
##
##                coef exp(coef) se(coef) robust se     z    p
## SmokingH     -1.231     0.292    1.021     0.913 -1.35 0.18
## logtSmokingH  0.554     1.740    0.787     0.691  0.80 0.42
##
## Likelihood ratio test=1.66  on 2 df, p=0.4
## n= 289, number of events= 22
##    (194 observations deleted due to missingness)
```

Al crear el modelo extendido, el p-valor que obtenemos es de 0.4, por lo que la estadística Wald de 0.80 con p-valor 0.42 no es significante para logtSmokingH. La asunción de los riesgos proporcionales no queda violada por la variable SmokingH, y concluimos que nuestro modelo es adecuado para valorar nuestra H0.