Herramienta para analizar matrices de expresión génicas con machine learning

DOMINGO JAVIER RODRÍGUEZ PÉREZ

Máster de Bioinformática y Bioestadística

Desarrollo de herramientas de soporte a la ómica - Prácticas UGR-UOC

Tutor/profesor: Antonio Jesus Adsuar Gómez

Tutor prácticas UGR-UOC: Alberto Fernández Hilario

fecha: 01/2020

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Herramienta para el estudio de RNAseq con machine learning*
Nombre del autor:	*Domingo Javier Rodríguez Pérez*
Nombre del consultor/a:	*Alberto Fernández Hilario*
Nombre del PRA:	*Antonio Jesus Adsuar Gómez*
Fecha de entrega (mm/aaaa):	01/2020
Titulación::	Máster de Bioinformática y Bioestadística
Área del Trabajo Final:	*Desarrollo de herramientas de soporte a la ómica*
Idioma del trabajo:	*Castellano*
Palabras clave	*RNAseq, Ramdom forest, Feature selectión*

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

En el campo de las aplicaciones biomédicas, es tan importante obtener una alta precisión como hacer que los modelos generados sean explicables para el personal clínico. Por esta razón, es esencial aplicar técnicas inteligentes que sean capaces de aprender de manera efectiva en estos escenarios. En esta ocasión se trata de crear un software en R para proporcionar una manera sencilla de construir un análisis explicativo de la causalidad entre la expresión génica y las condiciones del paciente. El software creado está muy automatizado facilitando las entradas de datos para estudiar diferentes matrices de expresión, con un flujo lineal, con una **lectura de datos** a través del código GEO, un **preprocesamiento** en el que se facilita un contraste de hipótesis,una normalización para hacer los datos comparables entre ellos y un filtrado de genes que reduce el cálculo computacional del posterior **entrenamiento de los modelos** machine learning el cual conlleva diferentes técnicas de selección de genes para, a través de la **validación del modelo**, detectar la relación entre la expresión génica y la condición del paciente y **compartir los resultados** de los genes realmente implicados en la respuesta

Pongo a prueba esta herramienta con uno de los temas mas actuales en cuanto a diagnostico clínico, la **detección del cáncer a través de la expresión génica de las plaquetas**. Los datos se han obtenido del experimento con código GSE89843. Se obtienen **AUC** por encima del **90% con** tan solo **10 genes**, lo que supone un gran avance en este campo. El AUC se puede interpretar como la probabilidad de clasificarlos correctamente. Debido a su bajo coste por el número reducido de genes y su poca invasividad puede realizarse a modo de test preventivo y reducir su tasa de mortalidad.

Abstract (in English, 250 words or less):

In the field of biomedical applications, it is as important to obtain high precision as to make the generated models explainable to clinical staff. For this reason, it is essential to apply intelligent techniques that are able to learn effectively in these scenarios. This time it is about creating software in R to provide a simple way to construct an explanatory analysis of the causality between gene expression and patient conditions. The software created is highly automated, facilitating data entry to study different expression matrices, with a linear flow, with a reading of data through the GEO code, a preprocessing in which a hypothesis contrast is facilitated, a normalization to make the comparable data between them and a gene filtration that reduces the computational calculation of the subsequent training of machine learning models which entails different gene selection techniques to, through the validation of the model, detect the relationship between gene expression and the patient's condition and share the results of the genes really involved in the response

I test this tool with one of the most current issues in terms of clinical diagnosis, the detection of cancer through the gene expression of platelets. The data were obtained from the experiment with code GSE89843. AUC above 90% are obtained with only 10 genes, which is a great advance in this field. The AUC can be interpreted as the probability of classifying them correctly. Due to its low cost due to the reduced number of genes and its low invasiveness, it can be carried out as a preventive test and reduce its mortality rate.

Índice

Sumario

Lista de figuras

Índice de figuras

Índice de tablas

1. Introducción

1.1 Contexto y justificación del Trabajo

Las técnicas de secuenciación masiva generan una cantidad de datos tan grande que conlleva problemas relacionados con Big data. Es importante crear herramientas para analizar este tipo de datos y ponerlos a disposición de la comunidad científica. En este caso se trata de un software en R para facilitar el análisis de matrices de expresión génicas disponibles en el repositorio NCBI . Como ejemplo se ha escogido un experimento de gran importancia en la actualidad, *'Swarm Intelligence-Enhanced Detection of Non-Small-Cell Lung Cancer Using Tumor-Educated Platelets'* el cual trata de utilizar las plaquetas como biomarcador de cáncer de pulmón, al que se puede acceder a través del código 'GSE89843' en 'https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE89843'. Se utilizaran los datos del experimento para tratar el problema de la **detección del cáncer de pulmón** a través de la expresión génica contenida en las plaquetas.

Es un tema relevante ya que crear potentes herramientas de fácil uso para analizar datos ómicos, supone una gran oportunidad de estudiar problemas complejos en un simple Pcpersonal, de forma rápida y sin necesidad de un nivel alto en programación. El ejemplo que nos ocupa también es relevante, ya que es importante encontrar métodos de detección del cáncer de pulmón menos invasivo (rayos X) y mas económico que los actuales, como un simple análisis de sangre y la secuenciación de unos pocos RNAs contenidos en las plaquetas. De esta forma será mas factible hacer controles preventivos y detectar el cáncer en etapas tempranas, incluso antes de ningún síntoma, además en cada caso es importante detectar que genes están alterados en respuesta a ese cáncer para hacer un tratamiento personalizado y reducir su tasa de mortalidad

Las herramientas actuales como la tomografía computarizada de baja dosis (LDCT) combina un equipo especial de rayos X con computadoras sofisticadas para producir múltiples imágenes transversales, o fotografías del interior del cuerpo, el problema es que son invasivas de rayos X y muy caras de utilizar, que dificulta la prevención o detección de la presencia de cáncer en etapas tempranas.

Por otro lado, las muestras de biopsia de tejido se usan ampliamente para caracterizar tumores, pero están limitadas por las limitaciones en la frecuencia de muestreo y su representación incompleta de todo el tumor. Ahora, la atención se está volcando hacia biopsias líquidas mínimamente invasivas, un tipo de diagnostico con muestras de sangre. Se han publicado estudios que muestran que las plaquetas contienen 'información' de la presencia de cáncer. Por un lado, en el año 2017 se publicó el articulo bajo el título *'Swarm Intelligence-Enhanced Detection of Non-Small-Cell Lung Cancer Using Tumor-Educated Platelets'[2]* donde se llegó a un 85% de accuracy con 850 genes. Sin embargo ésta cantidad de genes implicados eleva los costes del

diagnostico y es difícil comprender el papel de tantos genes en la enfermedad, por lo que se necesita reducir esa cantidad sin perder calidad en la predicción.

En la actualidad se esta invirtiendo, por parte de varios países, en la construcción de una gran base de datos de RNAseq de aislados de plaquetas y se está investigando sobre los mejores modelos de clasificación en diferentes enfermedades.

En este trabajo se pretende desarrollar una herramienta construida en R, para analizar matrices de expresión RNAseq, que, de manera automatizada, descargue los ficheros y les de la estructura correcta para trabajar con ellos, además de facilitar la construcción del contraste de hipótesis que el científico quiera realizar, en el sentido de contrastar la expresión de genes en diferentes condiciones. Para este fin, el paquete edgeR contiene funciones muy potentes que proporciona una primera criba de genes o feature selection con las que trabajar con aprendizaje automático de manera mas precisa.

En el ejemplo que nos ocupa, se trata de utilizar las plaquetas como biomarcadores de cáncer de pulmón. Se estudiará matrices de expresión de plaquetas en varias condiciones, entre ellas cáncer de pulmón. Se pretende construir modelos de clasificación basados en random forest, con la 'ayuda' de un filtrado de genes basado en el paquete edgeR, para mejorar el rendimiento en la detección del cáncer y sus genes implicados, y así abaratar costos y aportar comprensión sobre la enfermedad, así como, la posibilidad de hacer tratamientos personalizados.

1.2 Objetivos del Trabajo

Se trata de crear un software en R para proporcionar una manera sencilla de construir un análisis explicativo de la causalidad entre la expresión génica y las condiciones del paciente (ejemplo: la condición de cáncer), con varios modelos de clasificación para optimizar la predicción y la selección de genes implicados.

Listado de objetivos:

1) Leer datos.

- Proporcionar una manera sencilla de acceder a los datos de matrices de expresión contenidos en el repositorio NCBI mediante su código GEO.

2) Preprocesar datos.

- Estructurar los datos de manera automática, además, obtener tablas y gráficos que resuman el experimento y la expresión de los diferentes grupos de pacientes.
- Control de calidad con exploración gráfica de las distribuciones de expresión.

- Facilitar la exploración de las anotaciones de cada muestra para crear los grupos de interés para hacer contraste de hipótesis.
- Filtrado de genes genes diferencialmente expresados. De esta manera se acota el número de genes que pasan al algoritmo machine learning reduciendo su calculo computacional.

3)Entrenar modelo.-

- Se ha optado por el clasificador random forest, centrando el esfuerzo en contrastar varios modelos con diferentes técnicas de selección de genes.

  - Selección de genes según paquete edgeR (FS1)

  - Selección de genes según accuracy (FS2)

  - Selección de genes según función boruta (FS3)

4) Validar modelo.

- Implementar validación cruzada k=5
- Accuracy
- AUC

5) Compartir resultados

Tabla de los genes mas importantes y el rendimiento de estos en la clasificación.

1.3 Enfoque y método seguido

Se ha optado por desarrollar una herramienta en Rstudio para estudiar matrices de expresión (RNA_seq y/o microarrays). Ya que el paquete Bioconuctor es ideal para procesar este tipo de datos y analizar su contenido.

Como los experimentos de matrices de expresión miden la expresión génica en varias condiciones es importante facilitar los contrastes, por ejemplo: ‘tratado-no tratado’, ‘cáncer de pulmón-otros tipos de cáncer’…. Posteriormente se hace un filtrado de genes en base a ese contraste con los diferencialmente expresaos, para este fin se ha optado por el paquete edgeR.

Tras esta criba de genes importantes, se aplican métodos de aprendizaje automático, concretamente random forest, ya que es con el que mejor resultados se obtienen en artículos anteriores, para detectar condiciones a través de la expresión génica, y a la vez, reducir el número de genes a los realmente relevantes para facilitar la comprensión.

1.4 Planificación del Trabajo

1.4.1 Recursos necesarios para realizar el trabajo:

- El proyecto se realizará en Rstudio
- Es necesario un Pc personal con conexión a internet

1.4.2 Tareas a realizar:

1) Leer datos.

- Vincular base de datos.- facilitar un input con el código GEO para ver los archivos disponibles en el repositorio NCBI
- Recopilación y exploración de los datos.- Se trabaja con grandes cantidades de datos en varios ficheros, por lo que se debe crear una manera sencilla de acceder a ellos, descargarlos y darles la estructura necesaria, ademas de las tablas y gráficos que resuman el experimento.

2) Prepocesamiento de datos

- Selección de atributo clasificador.- En los experimentos de matrices de expresión génica se miden las expresiones en diferentes condiciones como “tratado-no tratado” o mas mas condiciones (por ejemplo: “cáncer-sanos-otras enfermedades… “). En esta herramienta se facilita la elección de un clasificador dicotomico, es decir, que diferencia una condición con el resto de condiciones (cáncer-nocancer)
- Transformaciones de los datos.- De formato y estadísticos (eliminar ruido de fondo, normalización…)
- Control de calidad.- Descripción de datos (grupos, distribuciones…)mediante graficas que faciliten detectar anomalías en las distribuciones de algún grupo
- Contrastes de hipótesis.- Como se ha comentado se hará un clasificador dicotomico, por lo que es necesario facilitar un input para seleccionar los grupos que se quieren contrastar
- Selección de genes diferencialmente expresados.- Selección previa al machine learning, de los genes mas significativos que se expresan de manera diferente en ambas condiciones por ejemplo 150 genes.

3) Entrenar modelos

- Feature selection, nueva selección de genes para seleccionar los genes mas importantes. Se comprobaran 3 métodos:

    Fs1.- selección de genes según paquete edgeR

    Fs2.- selección de genes según accuracy

    Fs3.- Selección de genes según función boruta

- Implementación de validación cruzada.
- Algoritmo machine learning

4) Validar modelo.-

- Comprobación del rendimiento por validación cruzada

5) Compartir resultados

- Elección del modelo clasificador, a través de feature selection vemos los genes que cambian de expresión y su importancia en la clasificación
- Comprobación del rendimiento en los diferentes grupos de pacientes (si hay mas de 2) puede que la clasificación sean diferentes
- Tabla informativa de genes mas relevantes

1.4.3 Diagrama de Gantt:

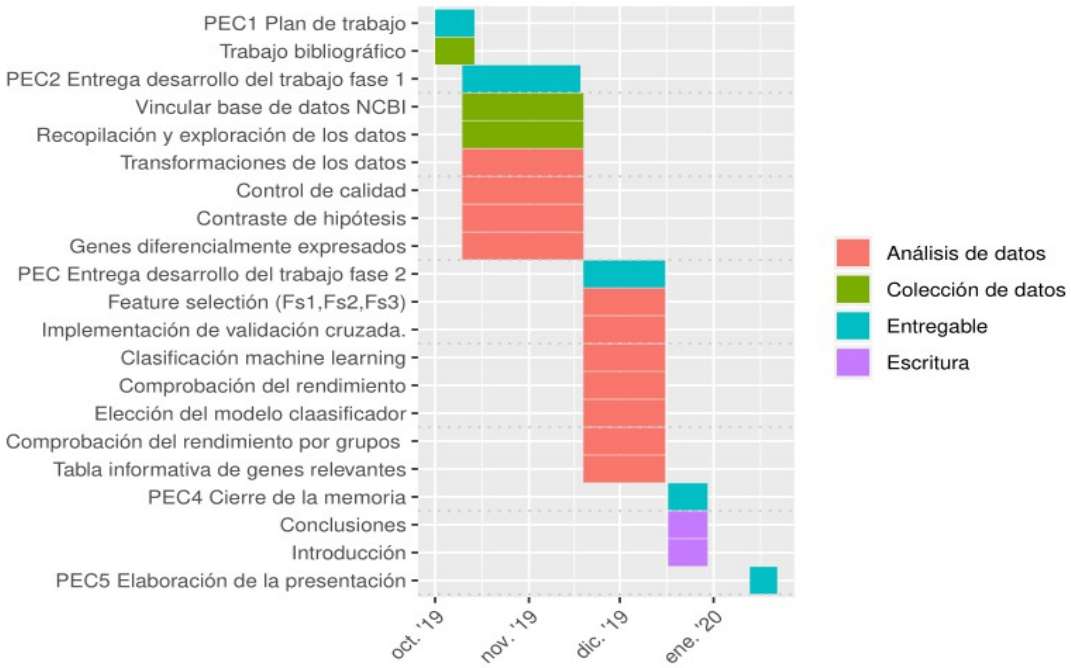

Figura 1: Diagrama de Gantt

1.5 Breve sumario de productos obtenidos

Al tratarse de un software automatizado en la medida de lo posible, considero la entrada de datos (inputs) como un producto importante que sirve para estudiar diferentes conjuntos de datos de expresión génica.

- Input datos, para facilitar la obtención de los datos mediante su código en el repositorio NCBI.
- Tabla informativa de los datos (título del experimento, tipos de datos…)

- Input atributo o condición clasificador para facilitar el contraste (ejemplo: cáncer-control)
- Gráficos informativos de la distribución de los datos (Diagrama de cajas e histograma)
- Gráficos de diferencias de expresión entre grupos (plotSmear y volcano)
- Input del cutoff del contraste, importante para seleccionar el número de genes diferecialmente expresados para el posterior entrenamiento (feature selection según paquete edgeR)
- Gráficos y tablas para la evaluación del rendimiento en la clasificación mediante aprendizaje automático (matriz de confusión, curva ROC, AUC)
- Input de la selección del clasificador
- Tabla informativa de la clasificación en los diferentes grupos de pacientes. Por ejemplo en el experimento que nos ocupa es importante ver la predicción de cáncer en pacientes sanos o con otras enfermedades inflamatorias o incluso en pacientes con metástasis.
- Tabla informativa de los genes implicados en la clasificación

1.6 Breve descripción de los otros capítulos de la memoria

En el resto de capítulos se hará una explicación de los procesos que se llevan a cabo en el software realizado. El cual, tiene un flujo lineal siguiendo el formato de articulo científico, ademas está muy automatizado e incorpora los input necesarios para adaptarlo a diferentes estudios de expresión génica.

Se ha escogido como ejemplo un experimento de gran importancia en la actualidad, *'Swarm Intelligence-Enhanced Detection of Non-Small-Cell Lung Cancer Using Tumor-Educated Platelets'* el cual trata de utilizar las plaquetas como biomarcador de cáncer de pulmón, al que se puede acceder a través del código 'GSE89843'

2. Resto de capítulos

2.1 Leer datos

2.1.1 Vincular con el repositorio NCBI (input código GEO)

La función getGEOSuppFiles() utiliza como parámetro de entrada el código identificativo Gene Expression Omnibus (GEO) el cual se encuentra en la pagina 'http://www.ncbi.nlm.nih.gov/geo/' , y devuelve los archivos disponibles en el repositorio NCBI. En este caso se utiliza el código 'GSE89843'

2.1.2 Recopilación y exploración de los datos

2.1.2.1 Obtención matriz de expresión génica

Las matrices de expresión génica RNAseq constan de una serie de muestras en columnas y una gran serie de genes dispuestas en filas y en cada celda se almacena el nivel de expresión génica.

2.1.2.2 Obtención tabla de anotaciones del RNA seq

La tabla de anotaciones es un archivo con información sobre el experimento donde se obtuvo la matriz de expresión, es muy extenso, donde se puede destacar el diseño del experimento, los grupos en los que se distribuyen las muestras así como el titulo y autores del experimento.

2.1.2.3 Información relevantes de los datos RNAseq

La siguiente información es un producto del programa, obtenido desde la base de datos NCBI, en su idioma original, inglés, sin modificaciones.

Título
Swarm intelligence-enhanced detection of non-small cell lung cancer using tumor-educated platelets
Autores
"Myron,…,Best"
Resumen
We report RNA-sequencing data of 779 blood platelet samples, including 402 tumor-educated platelet (TEP) samples collected from patients with non-small cell lung cancer (NSCLC). In addition, we report RNA-sequencing data of blood platelets isolated from 377 individuals without reported cancer, but not excluding individuals with inflammatory conditions. This dataset highlights the ability of TEP RNA-based 'liquid biopsy' diagnostics in patients with NSCLC.
Web de descarga
"https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE89843"
Diseño experimental del RNA seq
"Blood platelets were isolated from whole blood in purple-cap BD Vacutainers containing the EDTA anti-coagulant. The cells and aggregates were removed by centrifugation at room temperature for 20 minutes at 120g, resulting in platelet-rich plasma. The platelets were isolated from the platelet-rich plasma by centrifugation at room temperature for 20 minutes at 360g. The platelet pellet was collected in 30 Î¼l RNAlater (Life Technologies), incubated overnight at 4Â°C and frozen at -80Â°C for further use. Frozen platelets were thawed on ice and total RNA was isolated using the mirVana RNA isolation kit (Life Technologies) according to the manufacturersâ\200\231protocol. Complementary purification of small RNAs was included in the isolation procedure by addition of miRNA homogenate (Life Technologies). Total RNA was dissolved in 30 Î¼l Elution Buffer (Life Technologies) and RNA quality and quantity was measured using Bioanalyzer 2100 with RNA 6000 Picochip (Agilent)."

Tabla 1: Información relevante del RNAseq escogido

2.2 Preprocesar datos

2.2.1 Estructurar RNAseq

Los experimentos de matrices de expresión génica miden la expresión en diferentes condiciones, para medir el efecto de la causalidad entre la expresión y las condiciones, se debe seleccionar el atributo que contiene las condiciones interesantes así como la condiciones a contrastar, Para este fin se muestra la tabla de anotaciones fenotipicas y la selección de los grupos de contraste

Tabla de anotaciones fenotipicas

Aquí se facilita la tabla de anotaciones, importante para diseñar nuestro experimento, en este ejemplo hay 40 atributos para cada una de las 779 muestras, con información como el tipo de célula (plaquetas en este caso) o la fecha de publicación. Entre estos datos hay que seleccionar el atributo que divide el dataset en los grupos de interés que queremos contrastar, es decir, en 'cancer-no cancer' o 'tratados-no tratados'… , en este ejemplo el atributo con los grupos fenotípicos interesantes es "disease:ch1", los cuales se resumen:

Condición del paciente	Número de muestras
Chronic Pancreatitis	5
Healthy Control	231
Epilepsy	21
Multiple Sclerosis	58
Non-significant Atherosclerosis	12
nonCancer	6
NSCLC	402
Stable Angina Pectoris	4
Pulmonary Hypertension	34
Unstable Angina Pectoris	6

Tabla 2: distribución de grupos de pacientes

Se observa que los datos no están balanceados, es decir, hay pocas muestras de algunos grupos de pacientes

2.2.1.1 Contraste de hipótesis (Input condición de interés)

Aquí se facilita la elección del grupo de pacientes que queremos diferenciar o contrastar con el resto, en este caso se selecciona la condición de cáncer de pulmón 'NSCLC'

2.2.1.2 Target

Código GEO	Grupo de pacientes	Grupo de contraste
GSM2390709: 1	NSCLC : 402	NSCLC: 402
GSM2390710: 1	Healthy Control :231	Control:377
Other :778	Other : 146	

Tabla 3: Target o tabla con información

El Target contiene la información necesaria para realizar la clasificación entre cancér y no cancer en los diferentes grupo de pacientes. Cada gen tiene un código identificativo para localizarlo en la matriz de expresión, pertenece a un grupo de pacientes con diferentes enfermedades y por último, pertenece al grupo que se pretende contrastar en este caso cáncer de pulmón NSCLC.

2.2.2 Ajustes de los datos

2.2.2.1 Formato, objeto DGEList (edgeR)

Se trabajará con paquete edgeR, el cual es un paquete escrito para R que realiza expresión diferencial genética utilizando datos de conteos bajo un modelo binomial negativo. Cada paquete tiene sus propias clases o condiciones para poder trabajar con él, este paquete necesita la clase DGEList para poder realizar cualquier análisis,el cual almacena la matriz de conteos y un data.frame con información de cada paciente, principalmente sobre la condición (cáncer-no cancer) a la que pertenece y el nivel de expresión.

2.2.2.2 Eliminar ruido de fondo

Es muy importante, antes de comenzar con el análisis diferencial de los genes, eliminar aquellos que no tienen conteos, o que apenas tienen. Pues de antemano se sabe que estos genes no se van a expresar de forma distinta en las diferentes condiciones que se estén estudiando.

2.2.2.3 Normalización

Previamente al análisis, se deben de normalizar los datos, sin embargo en este contexto normalizar tiene un significado distinto al que es usual en estadística. En RNAseq con la normalización se busca minimizar el ruido técnico introducido en los datos durante el proceso de secuenciación con el fin de volverlos comparables entre si, no se pretende transformarlos para que sigan una distribución normal, que es lo que se suele entender como normalización en términos estadísticos. El cálculo de dichos factores de normalización se realiza por la media truncada de los valores (trimmed mean of M values, TMM) entre cada par de pacientes. Obtenemos factores de normalización cercanos a 1, lo cual nos indica que no existen grandes diferencias entre la composición de los distintos pacientes.

2.2.3 Control de calidad

Mediante la visualización gráfica de los datos de expresión génica en cada grupo, se puede observar su distribución. Un RNAseq de calidad conlleva una distribución normal

Se muestran las distribuciones por grupos con gráficos boxplot e histogramas.

2.2.3.1 Boxplot

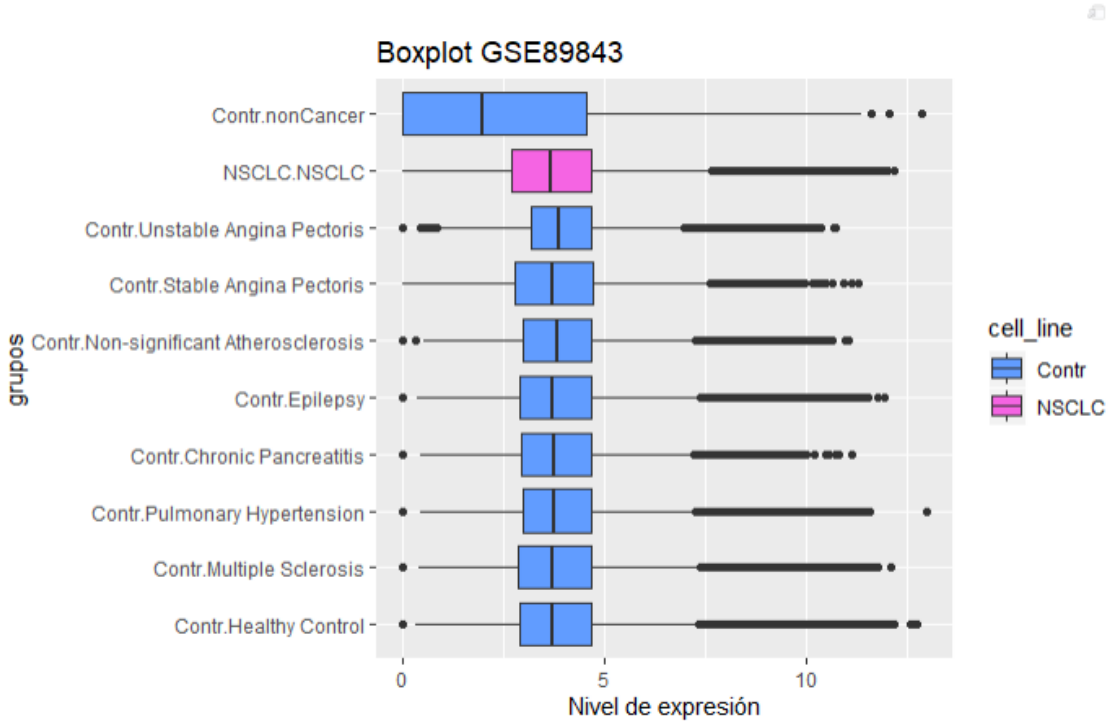

Figura 2: Boxplot de la distribución de la expresión genética por grupos

Se observa una distribución anómala en el grupo “noncancer”, Se puede ver su distribución con mas detalle en el siguiente histofgama

2.2.3.2 Histograma

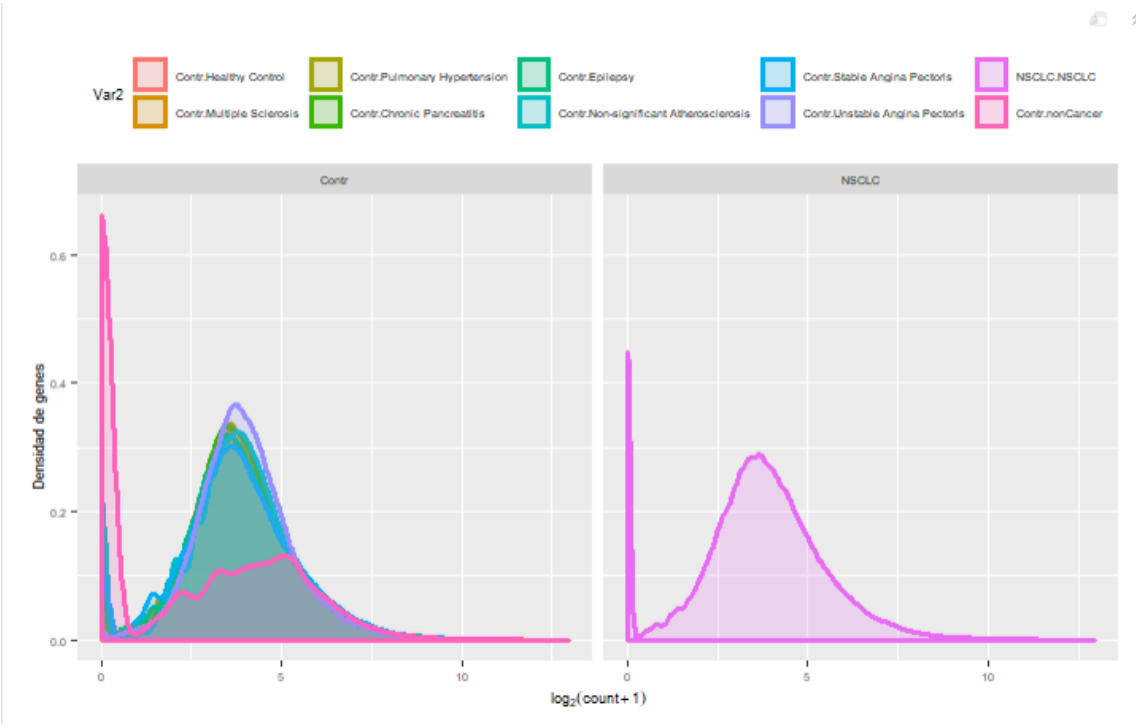

Figura 3: Histograma de la distribución de la expresión génica por grupos de pacientes.

Se observa una distribución anómala en el grupo “noncancer” el resto del dataset sigue una distribución normal con excepción de un pico en los genes que no se expresan o se expresan muy poco.

2.2.4 Filtrado de genes diferencialmente expresados

2.2.4.1 Contraste de hipótesis.

- Hipótesis nula (H=0) las expresiones son iguales en ambos grupos
- Hipótesis alternativa (H=1) las expresiones difieren en ambos grupos

2.2.4.2 Estimación de la dispersión

Usando la variable de agrupación que hemos creamos(cáncer-no cancer). debemos estimar las dispersiones de cada gen con la función estimateDisp {edgeR}, necesario para poder usar la medida de variabilidad en las siguientes pruebas pruebas.

Descripción de la Función estimateDisp {edgeR}:

Estima las dispersiones binomiales negativas comunes, con tendencia y marcadas según las probabilidades empíricas de Bayes, El cual, maximiza la probabilidad binomial negativa para dar la estimación de las dispersiones comunes.

2.2.4.3 Cálculo de expresión diferencial

Calcula las diferencias en las medias entre dos grupos de conteos de expresión con la función exactTest {edgeR}

2.2.4.4 Genes infra y sobre expresados

Se procede a clasificar los genes diferencialmente expresados atendiendo a los siguientes criterios:

- El P-valor ajustado, éste se debe ajustar según el nivel de significación que se crea oportuno, en este cutoff= $5*10^{-6}$.
- El log fold change, en el que cada unidad corresponde a un cambio del 100% del nivel de expresión.
- Y el logaritmo en base 2 de la media de los cpm (conteos por millón) de cada gen (logCPM)

Estos Tests estadísticos se realizan con la función decideTests {edgeR}.

NSCLC-Control	
Down	285
NotSig	4104
Up	333

Tabla 4: Genes diferencialmente expresados

Se observan 333 genes que tienden a sobre expresarse y 285 genes que tienden a infra expresarse bajo la condición de cáncer

2.2.4.5 Gráficos del contraste

Una vez hecho el filtrado de genes que tienden a expresarse de manera diferente bajo una condición, es conveniente visualizar las graficas que muestran las diferencias de expresión

2.2.4.5.1 Gráfico plotSmear

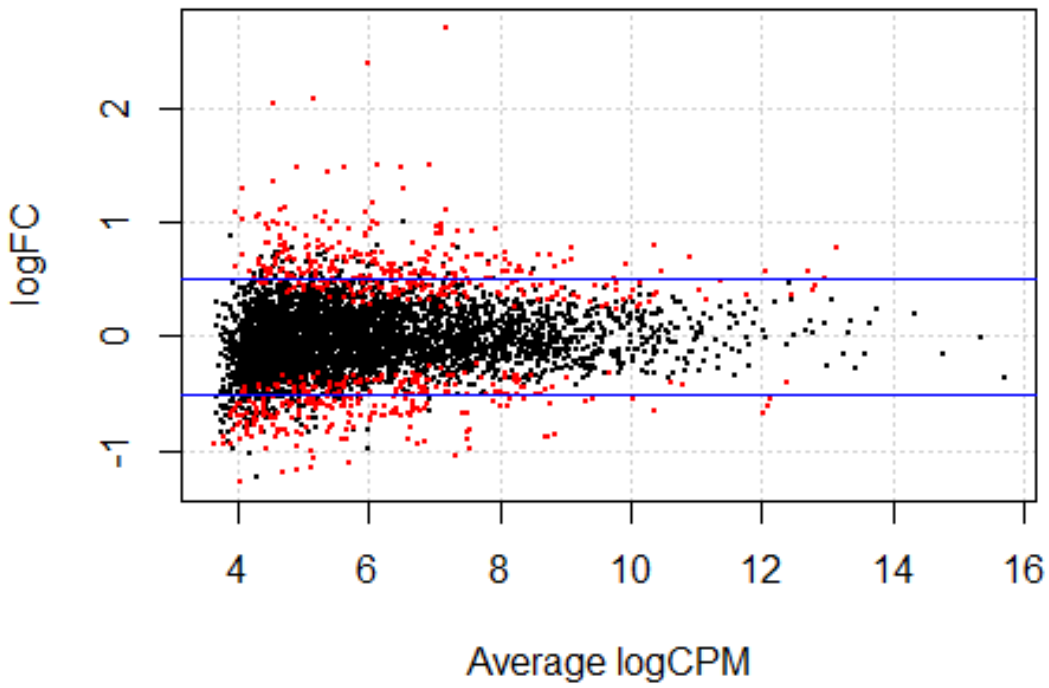

Figura 4: PlotSmear gráfico de diferencias de expresión bajo cáncer de pulmón

Cada gen está representado por un punto, en el eje Y se observan los distintos logFC en relación con el logCPM en el eje x, resaltando en color rojo aquellos transcritos considerados diferencialmente expresados, las líneas azules indican el logFC 0,5. Dicho logFC representa un cambio en la proporción de reads entre condiciones (nsclc-no cáncer) equivalente a 5, es decir, el tránscrito al que corresponde dicho logFC se expresa 5 veces más en una de las dos condiciones.

2.2.4.5.2 Gráfico volcano

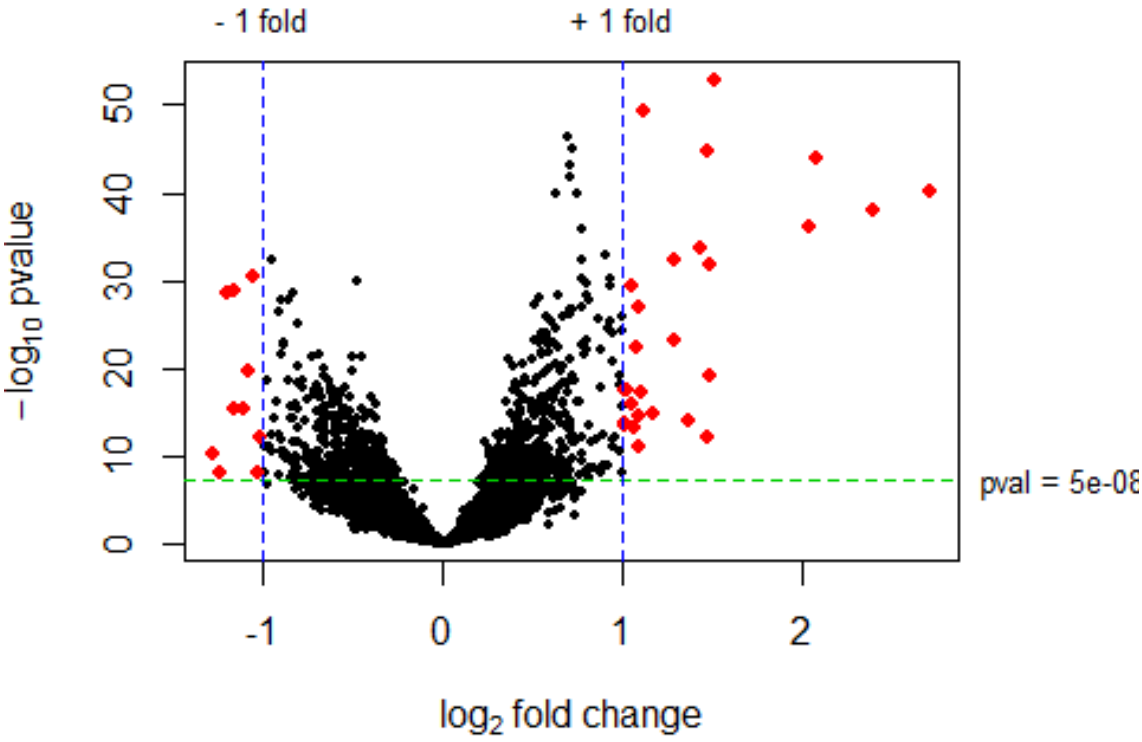

Figura 5: Volcano, gráfico de diferencias de expresión bajo cáncer de pulmón

En éste gráfico el eje X informa sobre el log fold-change de cada gen, y el eje Y mide el menos logaritmo en base 10 del p-valor asociado a cada gen. Únicamente los genes cuyo log fold-change es mayor que el umbral fijado (> 1), sufren un cambio de expresión significativo y están activados(puntos rojos del lado izquierdo); si el fold-change del gen es menor que el valor negativo del umbral fijado (< -1), el cambio es significativo y el gen está inhibido (puntos rojos en el lado izquierdo). Por otro lado, también es necesario aplicar la restricción que impone el p-valor. El cambio en expresión de un gen únicamente será significativo si el p-valor de este gen se sitúa por encima del p-valor umbral fijado anteriormente .

2.3 Entrenar y validar modelos

En esta sección se pretende detectar la condición de cáncer de pulmón (NSCLC) a través de la expresión génica en los diferentes grupos de pacientes, no solo en pacientes con cáncer de pulmón y sanos, sino también el los grupos de pacientes con otras patologías, además se calcula el porcentaje de acierto en los diferentes grupos.

2.3.1 Implementación del Cross Validation

Para aplicar la Validación Cruzada vamos a hacer uso de la función createFolds del paquete caret, y luego entrenaremos cada modelo sobre k = 5 subconjuntos.

2.3.2 Filtrado de genes, Feature selection 1 (FS1)

Anteriormente se ha hecho un test estadístico con la función decideTests de edgeR el cual da como resultado una lista de los genes diferencialmente expresados bajo la condición nsclc.

En este apartado se crea unas participaciones con diferentes números de genes con mayor diferencia de expresión **según el paquete edgeR,** no interactuá con un clasificador, esta selección la llamaré FS1.

Nº de genes por defecto= (5,10,15,20,30,40,50,100,150)

2.3.2.1 Random forest

El clasificador elegido es random forest ya que estudios anteriores han demostrado buenos resultados, Los hiperparametros escogidos vienen por defecto.

2.3.2.2 Matriz de confusión

5 genes		
Prediction\ Reference	Control	NSCLC
Control	252	125
NSCLC	134	268
10 genes		
Prediction\ Reference	Control	NSCLC
Control	274	103
NSCLC	113	289
15 genes		
Prediction\ Reference	Control	NSCLC
Control	278	99
NSCLC	102	300
20 genes		
Prediction\ Reference	Control	NSCLC
Control	298	79
NSCLC	77	325
30 genes		
Prediction\ Reference	Control	NSCLC
Control	315	62
NSCLC	73	329
40 genes		
Prediction\ Reference	Control	NSCLC
Control	322	55
NSCLC	64	338
50 genes		

Prediction\ Reference	Control	NSCLC
Control	320	57
NSCLC	68	334
100 genes		
Prediction\ Reference	Control	NSCLC
Control	315	62
NSCLC	64	338
150 genes		
Prediction\ Reference	Control	NSCLC
Control	319	58
NSCLC	67	335

Tabla 5: Matriz de confusión en la predicción de cáncer con selección de genes por paquete edgeR (Fs1)

2.3.2.3 Accuracy

Al hacer cross validation con 5 particiones o folds se genera 5 accuracy por cada experimento:

	5 genes	10 genes	15 genes	20 genes	30 genes	40 genes	50 genes	100 genes	150 genes
Fold1	0.6903	0.6903	0.7742	0.8129	0.8774	0.871	0.8774	0.8903	0.8839
Fold2	0.6688	0.6815	0.7325	0.7771	0.7962	0.8153	0.8217	0.8344	0.8471
Fold3	0.6731	0.7756	0.7692	0.8333	0.8333	0.8397	0.8526	0.8269	0.8205
Fold4	0.7179	0.7564	0.7244	0.8141	0.8333	0.8333	0.859	0.8526	0.859
Fold5	0.6	0.6774	0.7226	0.7806	0.7806	0.8065	0.8129	0.7935	0.8194
Medias	0.670	0.716	0.745	0.804	0.824	0.833	0.845	0.840	0.846

Tabla 6: Accuracys con selección de genes por paquete edgeR (Fs1)

Boxplot accuracys Fs1

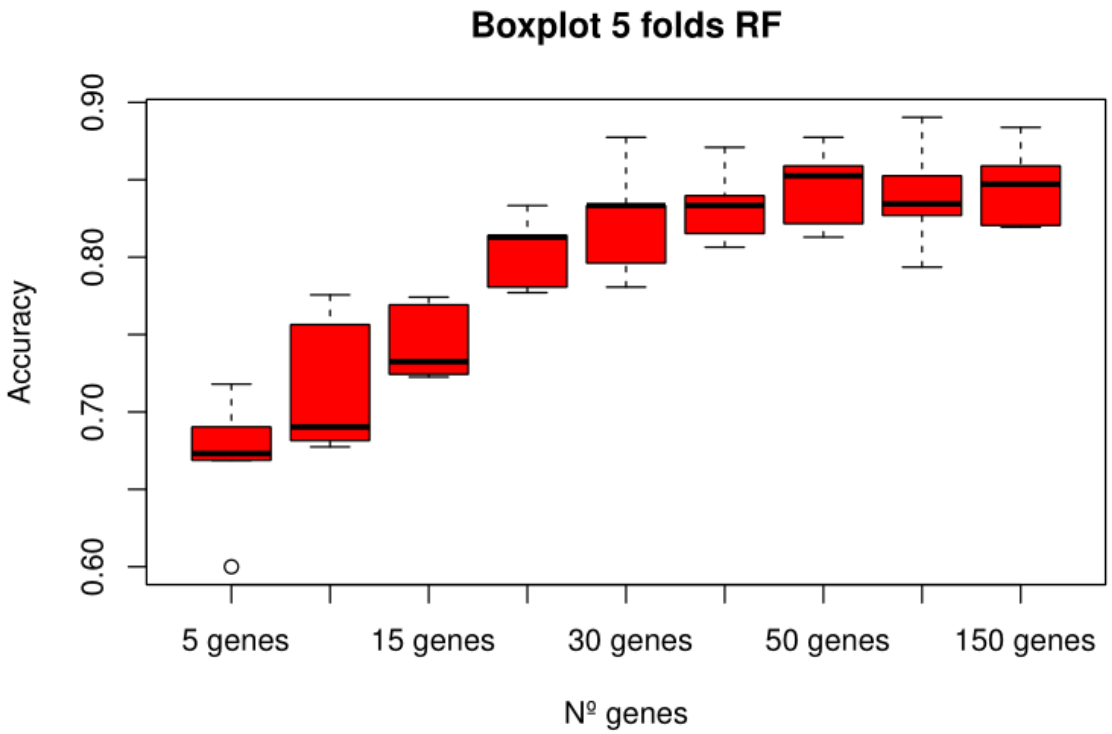

Figura 6: boxplot del accuracy de la clasificación random forest con la selección de genes basados en el paquete edgeR (Fs1)

2.3.2.3 CurvaROC y AUC

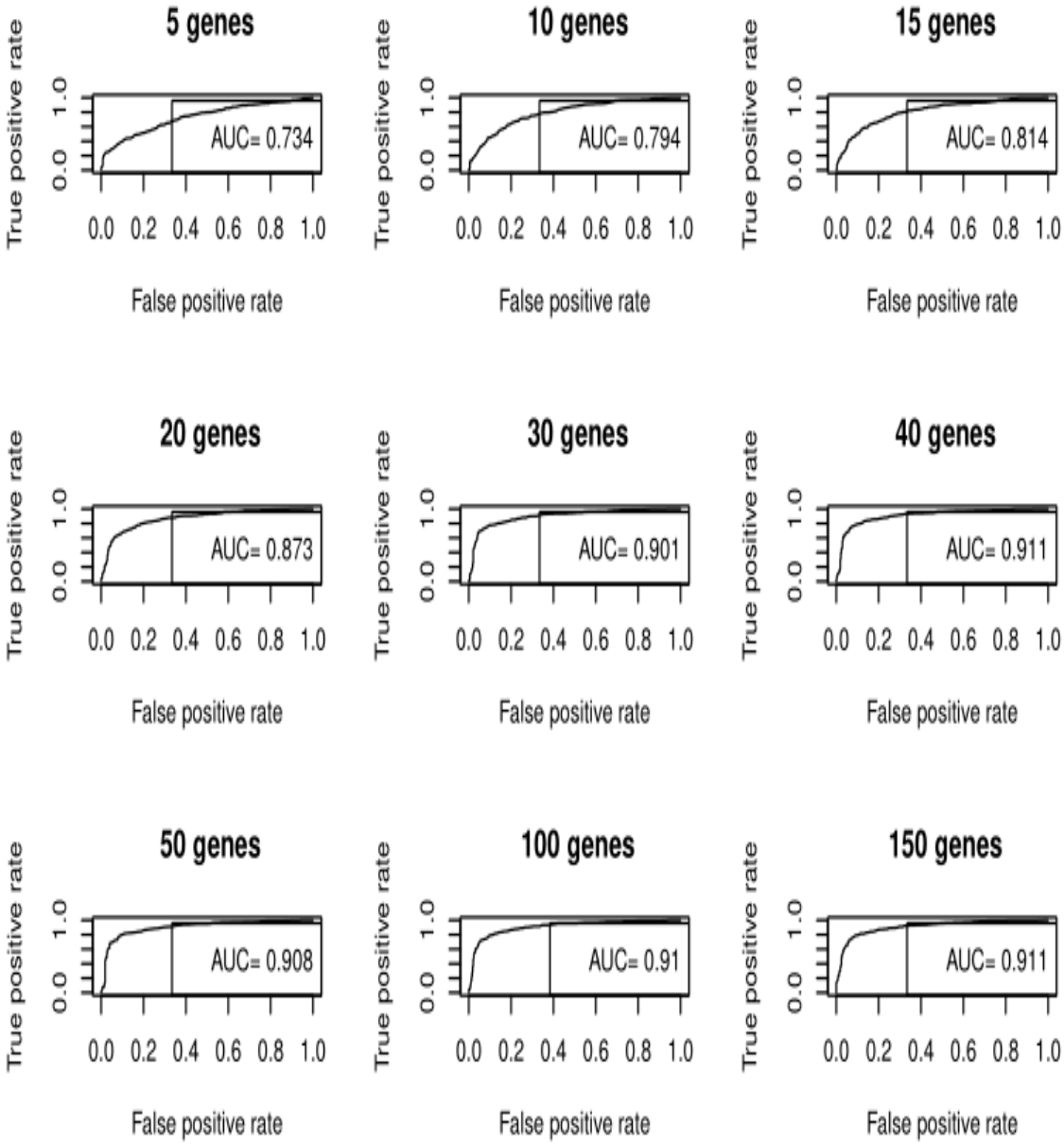

Figura 7: Curva ROC y AUC con selección de genes por paquete edgeR

2.3.3 Feature selection según accuracy (Fs2).

En esta segunda selección de genes utilizo el accuracy o importancia de los genes en la clasificación random forest.

2.3.3.1 Random forest fs2

Se usa de nuevo el algoritmo random forest para clasificar nsclc con los genes seleccionado

2.3.3.2 Accuracy

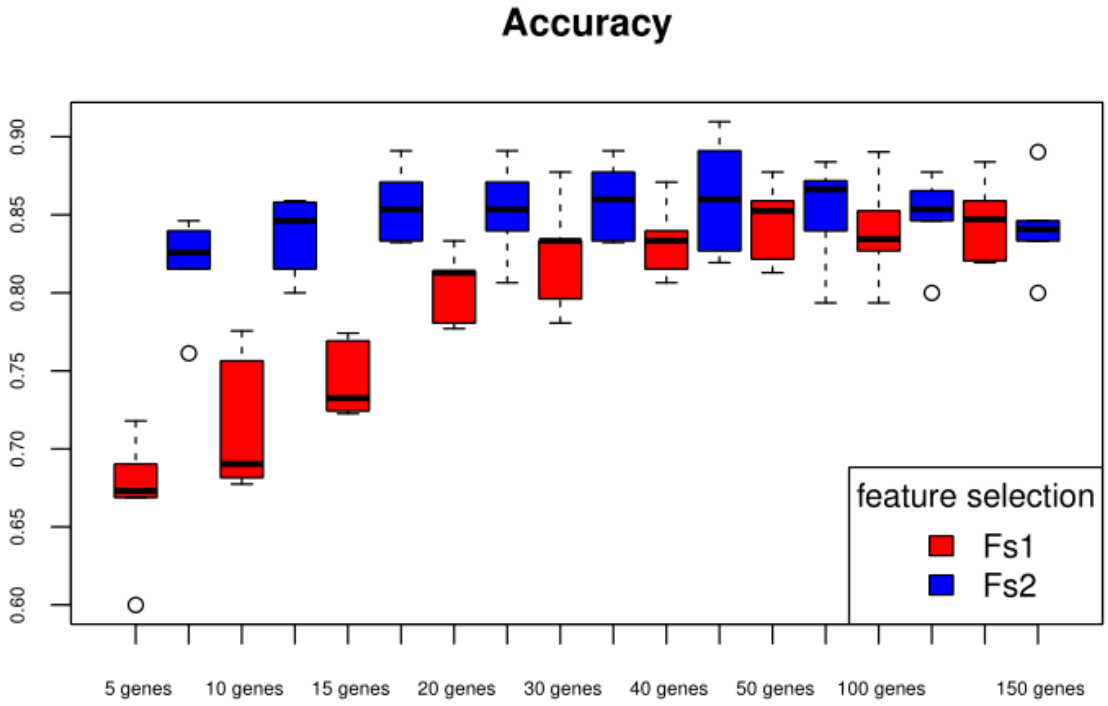

Figura 8: Boxplot accuracys (Fs2)

Boxplot accuracys del clasificador random forest con la selección de genes por edgeR en rojo (Fs1) y selección de genes teniendo en cuenta el clasificador según accuracys, en azul (Fs2)

2.3.3.3 Curva ROC y AUC Fs2

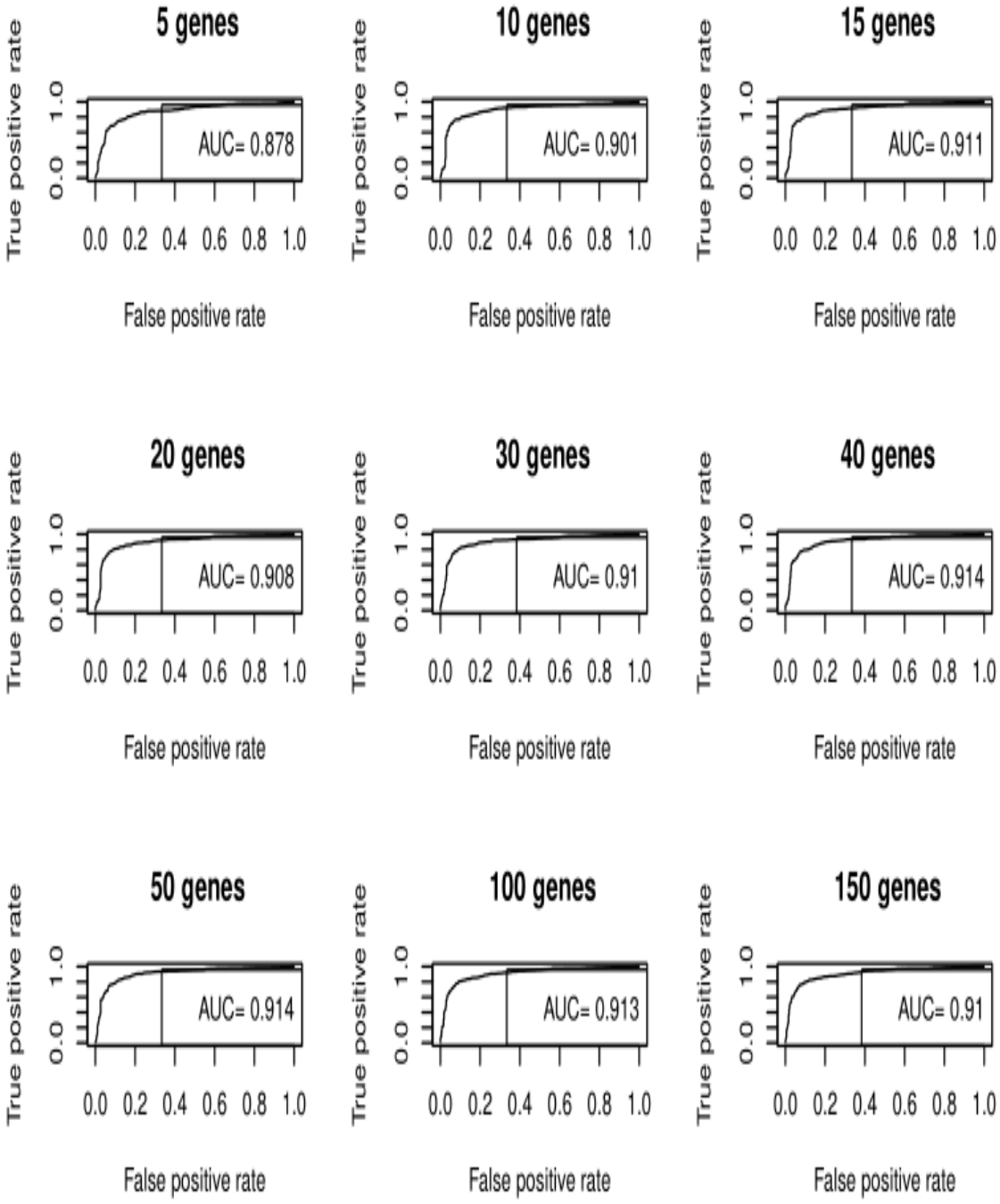

Figura 9: Curva ROC y AUC con selección de genes según accuracys (Fs2)

2.3.4 Feature selection boruta (Fs3)

Se hace una nueva selección de genes con mas importancia en el clasificador, ahora con la función boruta() . Sin entrar en muchos detalles, básicamente es como la anterior, pero es mas sofisticado, aquí se hacen unas duplicaciones de los datos para darle mas robustez a la importancia de los genes.

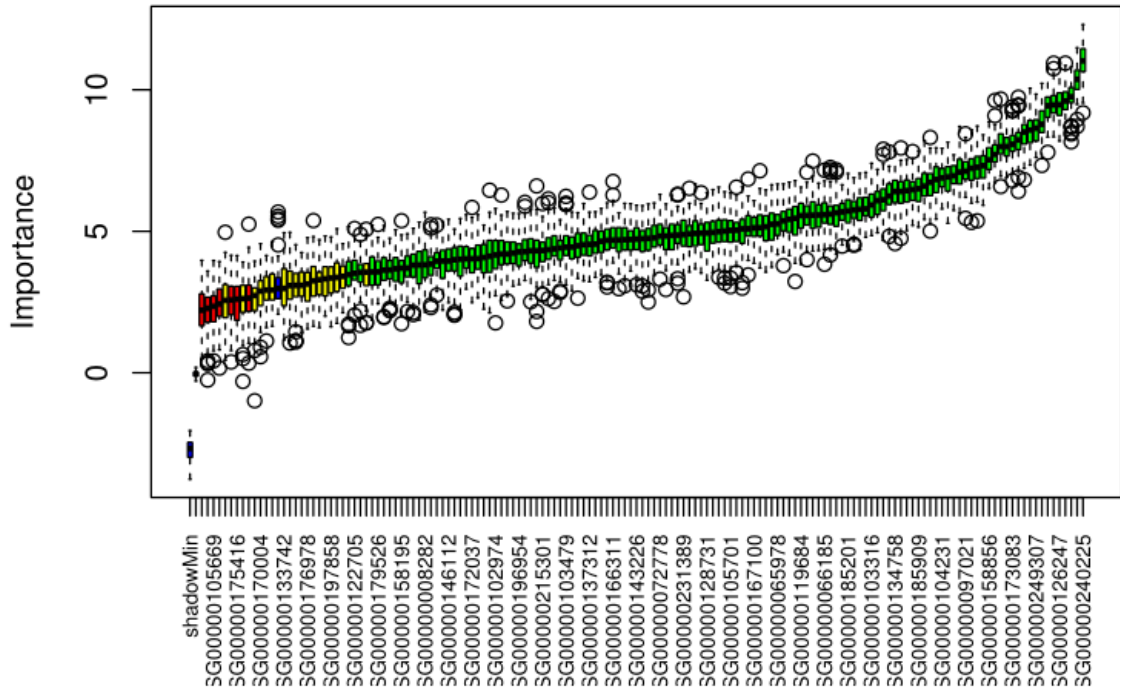

Figura 10: boxplot de la importancia de los genes en la clasificación , según la función boruta

2.3.4.1 Random forest fs3

Se utiliza de nuevo random forest para los genes con mas importancia escogidos con la función boruta

2.3.4.2 Accuracys Fs3

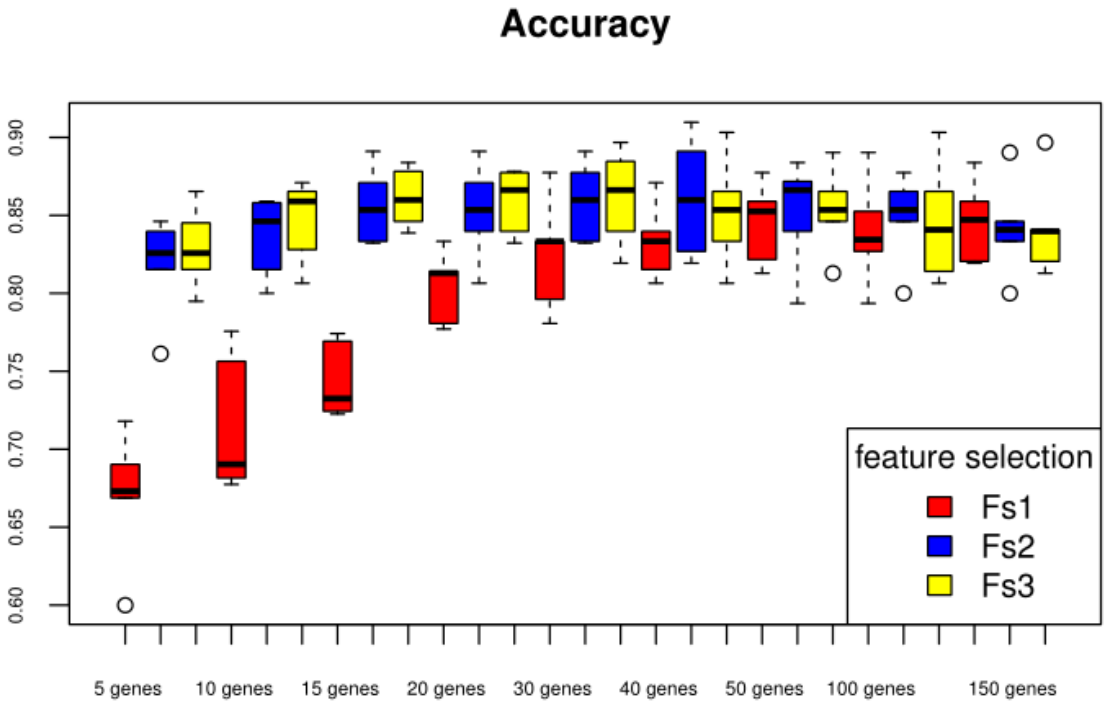

Figura 11: Boxplot de accuracys (Fs3)

Boxplot del accuracy(eje vertical) de los modelos de clasificación random forest en cuanto a la detección del cancer de pulmón con diferentes números de genes (eje horizontal). los colores corresponden a las técnicas de selección de genes. En rojo la selección de genes se ha hecho a través del paquete edgeR sin tener en cuenta el clasificador. En azul se seleccionan los genes por su nivel de importancia o accuracy. En amarillo se seleccionan los genes tambien por su importancia con el clasificador pero a través de la función boruta, con la que se consigue modelos mas robustos con menos genes.

2.3.4.2 AUC Fs3

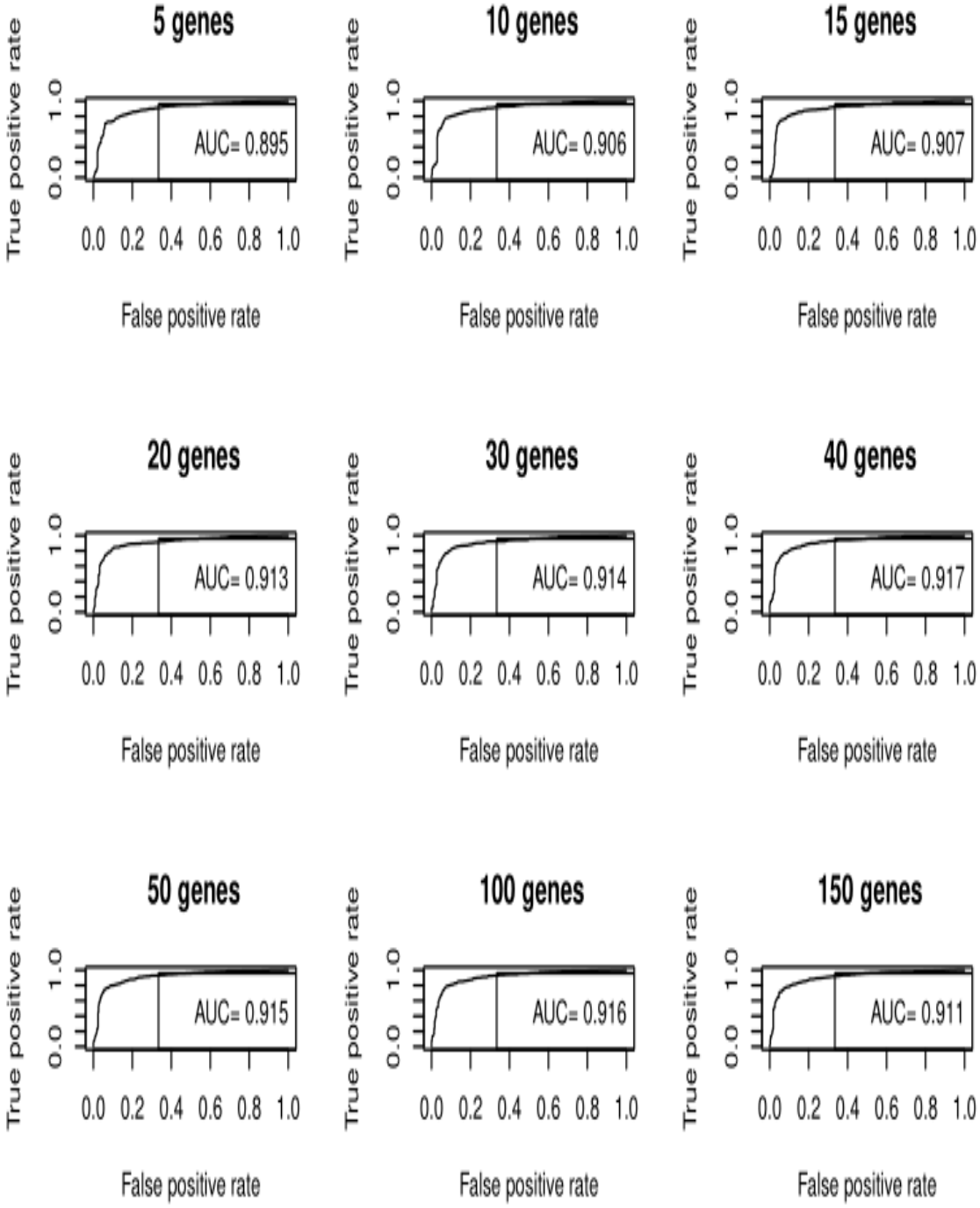

Figura 12: Curva Roc y AUC Fs3

Curva ROC y AUC en la detección del cáncer de pulmón con diferentes número de genes. La selección de genes se ha hecho en base la importancia d el clasificador mediante la función boruta se obderva una AUC cercana al 90 solo con 5 genes y tiene el máximo con 40 genes en 91,7%

2.3 Compartir resultados

2.3.1 Elección del clasificador.

El clasificador con la selección de genes por la función boruta, es la que mejor resultados devuelve, llegando a un área bajo la curva ROC del 91,7% con 40 genes y un accuracy de 86,3% con 30 genes, Se observa buenos resultados con tan solo 15 genes, con accuracy medio superior al 85% y un AUC superior al 90%

En este caso elijo el clasificador con 15 genes ya que la mejora de la clasificación con 40 genes es pequeña y 15 genes es lo bastante reducido como para sacar conclusiones biológicas.

2.3.2 Rendimiento por grupos del clasificador elegido.

Chronic Pancreatitis Accuracy 60 %		
Real\Predicción	Control	NSCLC
Control	3	2
Epilepsy Accuracy 80.9523809523809 %		
Real\Predicción	Control	NSCLC
Control	17	4
Healthy Control Accuracy 89.6103896103896 %		
Real\Predicción	Control	NSCLC
Control	207	24
Multiple Sclerosis Accuracy 93.1034482758621 %		
Real\Predicción	Control	NSCLC
Control	54	4
Non-significant Atherosclerosis, accuracy 83.3333333333333 %		
Real\Predicción	Control	NSCLC
Control	10	2
nonCancer, accuracy 33.3333333333333 %		
Real\Predicción	Control	NSCLC
Control	2	4
NSCLC, accuracy 84.3283582089552 %		
Real\Predicción	Control	NSCLC
NSCLC	63	339
Pulmonary Hypertension, accuracy 76.4705882352941 %		
Real\Predicción	Control	NSCLC
Control	26	8
Stable Angina Pectoris, accuracy 100 %		
Real\Predicción	Control	NSCLC
Control	4	0
Unstable Angina Pectoris, accuracy 66.6666666666667 %		
Real\Predicción	Control	NSCLC
Control	4	2

Tabla 7: Matriz de confusión por grupos de pacientes

Tabla informativa de la calidad del clasificador en los diferentes grupos de pacientes, se puede ver el accuracy y la cantidad de muestras bien y mal clasificadas, llama la atención que los grupos no están balanceados, con muy pocas muestras en algunos grupo donde los resultados carecen de robustez

2.3.3 Tabla informativa de genes implicados.

ENSEMBL	SYMBOL	GENENAME
ENSG00000126247	CAPNS1	calpain small subunit 1
ENSG00000213465	ARL2	ADP ribosylation factor like GTPase 2
ENSG00000142089	IFITM3	interferon induced transmembrane protein 3
ENSG00000173083	HPSE	heparanase
ENSG00000163359	COL6A3	collagen type VI alpha 3 chain
ENSG00000158578	ALAS2	5'-aminolevulinate synthase 2
ENSG00000206549	PRSS50	serine protease 50
ENSG00000172757	CFL1	cofilin 1
ENSG00000137312	FLOT1	flotillin 1
ENSG00000223609	HBD	hemoglobin subunit delta
ENSG00000023191	RNH1	ribonuclease/angiogenin inhibitor 1
ENSG00000072778	ACADVL	acyl-CoA dehydrogenase very long chain
ENSG00000160446	ZDHHC12	zinc finger DHHC-type containing 12
ENSG00000133742	CA1	carbonic anhydrase 1
ENSG00000093010	COMT	catechol-O-methyltransferase

Tabla 8: Genes implicados en respuesta al cáncer

3. Conclusiones

Se ha utilizado el software en uno de los problemas mas importantes en el diagnostico médico, la detección del cáncer de pulmón. Con este fin, he seleccionado los datos del estudio [2] que utilizan las plaquetas como biomarcadores de cáncer de pulmón. En ese estudio se llegó a un accuracy del 85% y un AUC del 91% utilizando 850 genes, una cantidad de genes demasiado grande como para sacar conclusiones biológicas del papel de esos genes en la enfermedad. En esta memoria, se expone que se pueden lograr resultados similares con un umbral de entre 10 y 20 genes, que es un valor manejable por el experto final. Se ha propuesto 15 genes en los que que cambia la expresión y son responsables de la respuesta en plaquetas, al cáncer de pulmón con un área bajo la curva de mas del 91%, este área puede interpretarse como la probabilidad de que ante un par de individuos, uno con cáncer y el otro sin cáncer, la prueba los clasifique correctamente, incluso en estadios tempranos de la enfermedad. Se trata de una prueba poco invasiva a través de una biopsia liquida del torrente sanguíneo, la extracción de RNA de aislados de plaquetas y la secuenciaón de un numero pequeño de RNAs se puede realizar de forma muy económica. Utilizando esta técnica a modo de test

de diagnostico preventivo se podrían detectar la presencia de cáncer en etapas tempranas y reducir su tasa de mortalidad. Ademas surgen otras preguntas como:

- El papel general de las plaquetas en el cáncer de pulmón
- El papel de cada gen implicado en esa respuesta
- Mecanismo de trasmisión de esa respuesta en las plaquetas (sin núcleo)
- Posibilidad de tratamientos personalizados
-Utilizar las plaquetas como biomarcador de otras enfermedades (ejemplo la enfermedad EPOC) y los principales genes implicados en la respuesta.

3.1 Sobre los objetivos planteados inicialmente

Se ha comprobado que es una herramienta precisa, selectiva y dinámica, que explica la causalidad entre la expresión génica y las condiciones del paciente en diferentes RNAseq.

1) leer datos
Se ha proporcionado una manera sencilla de **acceder a los datos** de secuenciación, los cuales deben ser descargados en varios ficheros por su gran tamaño. El programa crea una carpeta donde guardar estos datos, accede a ellos a través del código en el repositorio NCBI, comprueba cuales son los archivos y los descarga solo si no han sido descargados anteriormente , además muestra información relevante del experimento de expresión descargado, como el título y el diseño experimental, construye unas gráficas de expresión que facilitan mucho la comprensión del mismo y la comprobación de un test de calidad de los RNAseq.

2) preprocesamiento de datos
El programa es capaz de **estructurar los datos** automáticamente, transformándolos en el objeto que utiliza el paquete edgeR y normalizando los datos no en el sentido de transformarlos para que sigan una normal, normalmente los RNAseq ya vienen normalizados en este sentido mediante la técnica denominada TMM (esté se puede comprobar en el apartado protocolo, en información relevante), sino en el sentido de hacerlos comparables entre ellos a través de un valor para cada muestra, este paso lo considero importante en los buenos resultados obtenidos.

Se ha facilitado la tarea de realizar los **contrastes apropiados** a cada experimento, con éste fin se facilita la tabla de anotaciones fenotipicas para elegir los grupos que interesen al investigador.

Otro punto importante es el filtrado de **genes** que tienden a estar **diferencialmente expresados.** Se realiza a través del paquete edgeR sin tener en cuenta el clasificador. De esta manera se acota mucho el número de

genes que pasan al algoritmo machine learning reduciendo su complejidad computacional y mejorando los resultados

3) Entrenar modelo.
Se crea automáticamente varios modelos de clasificador con diferentes técnicas de selección de genes para poder compararlos entre ellos. El numero de genes escogido es: (5,10,15,20,30,40,50,100,150) para cada técnica de selección (Fs1,Fs2 y Fs3).

4) Validar modelo.
Se valida a través de validación cruzada k=5, es decir, se hacen 5 particiones donde repetir el experimento, de esta manera se le da mas robustez a la clasificación mostrando la media y la variabilidad en una grafica boxplot ademas se muestra el área bajo la curva AUC de cada experimento, es decir, uno por numero de genes y selección de los mismos (9*3=27 experimentos) el investigador elige el clasificador teniendo en cuenta su rendimiento y su sencillez.

5) Compartir resultados
Una tabla de los genes importantes con su código GEO, su símbolo y su nombre facilita al experto final una especie de enfoque hacia los genes interesantes, es decir los genes que reaccionan a una droga o que están implicados en un proceso, etc.

3.2 Futuras lineas de investigación:

Desde el descubrimiento de las plaquetas (Bizzozero,1881) se ha estudiado en profundidad su papel en la hemostasia y la trombosis. Sin embargo, en las últimas décadas, la lista de procesos biológicos en los que las plaquetas juegan un papel importante sigue expandiéndose. Las plaquetas pueden ingerir moléculas de RNA durante la circulación o interacción con otros tipos de células, sus funciones pueden ser muy versátiles, al igual que sus firmas de expresión genéticas, lo que las convierte en potenciales biomarcadores del estado de salud en general.

En este trabajo se han conseguido los mejores resultados en la detección del cáncer de pulmón a través de las plaquetas con el menor número de genes hasta el momento. De acuerdo con los resultados experimentales actuales, planeo extender la investigación a mejorar la firma genética del cáncer de pulmón y aplicarlo a otras enfermedades. Sobretodo con la llegada de más conjuntos de datos relacionado con la expresión génica de las plaquetas en diferentes enfermedades.

4. Glosario

Accuracy: medida de precisión de un modelo

AUC : Area under the ROC Curve

edgeR: Empirical Analysis of Digital Gene Expression Data in R

FS: Feature selection

GEO: Repositorio Gene Expression Omnibus

NCBI: National Center for Biotechnology Information

NSCLC: Non Small-Cells Lung Cancer

TEP: del inglés, plaquetas educadas por tumor

5. Bibliografía

[1] Heitzer, Ellen, et al. "Current and future perspectives of liquid biopsies in genomics-driven oncology." *Nature Reviews Genetics* 20.2 (2019): 71-88.

[2] M. G. Best et al., “Swarm Intelligence-Enhanced Detection of Non-Small-Cell Lung Cancer Using Tumor-Educated Platelets.,” Cancer cell, vol. 32, pp. 238–252.e9, aug 2017.

[3] M. Best et al., “RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics,” Cancer Cell, vol. 28, pp. 666–676, nov 2015.

[4] Y. Saeys, I. Inza, and P. Larra˜naga, “A review of feature selection techniques in bioinformatics,” Bioinformatics, vol. 23, no. 19, pp. 2507–2517, 2007.

[5] Y. Saeys, T. Abeel, and Y. Van De Peer, “Robust feature selection using ensemble feature selection techniques,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 5212 LNAI, no. PART 2, pp. 313–325, 2008.

[6] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. De Schaetzen, R. Duque, H. Bersini, and A. Nowac, “A survey on filter techniques for feature selection in gene expression microarray analysis,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, no. 4, pp. 1106–1119, 2012.

[7] S. Liu, C. Xu, Y. Zhang, J. Liu, B. Yu, X. Liu, and M. Dehmer, “Feature selection of gene expression data for cancer classification using double rbf-kernels.,” BMC Bioinformatics, vol. 19, no. 1, pp. 396:1–396:14, 2018.

[8] P. Sajda, “Machine learning for detection and diagnosis of disease,” Annual Review of Biomedical Engineering, vol. 8, pp. 537–565, 2006.

[9] A. Fern´andez, C. J. Carmona, M. J. del Jes´us, and F. Herrera, “A pareto-based ensemble with feature and instance selection for learning from multi-class imbalanced datasets,” Int. J. Neural Syst., vol. 27, no. 6, pp. 1–21, 2017.

[10] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” Machine Learning, vol. 46, no. 1-3, pp. 389–422, 2002.

[11] I. Dom´ınguez-Vigil, A. Moreno-Mart´ınez, W. J.Y., M. Roehrl, and H. Barrera-Salda˜na, “The dawn of the liquid biopsy in the fight against cancer.,” Oncotarget, vol. 9, pp. 2912–2922, jan 2017.

[12] S. Dudoit, J. Fridlyand, and T. Speed, “Comparison of discrimination methods for the classification of tumors using gene expression data,” Journal of the American Statistical Association, vol. 97, no. 457, pp. 77–86, 2002.

[13] Zhang, Ling, et al. "Multiplatform Biomarker Identification using a Data-driven Approach Enables Single-sample Classification." *bioRxiv* (2019): 581686.