Clustering de microRNAs en muestras de cáncer de próstata para la detección de bio marcadores

Juan Carlos Calvo Tejedor

Máster en Ingeniería Informática

Área: Inteligencia Artificial

Samir Kanaan Izquierdo

Carles Ventura Royo

Barcelona, 28 de Diciembre de 2016

Licencia

Esta obra está sujeta a una licencia de

Reconocimiento-NoComercial-SinObraDerivada 3.0

España de Creative Commons

FICHA DEL TRABAJO FINAL

Título del trabajo:	Clustering de microRNAs en muestras de cáncer de próstata para la detección de bio marcadores
Nombre del autor:	Juan Carlos Calvo Tejedor
Nombre del consultor/a:	Samir Kanaan Izquierdo
Nombre del PRA:	Carles Ventura Royo
Fecha entrega (mm/AAAA):	28/12/2016
Titulación:	Máster en Ingeniería Informática
Área del trabajo final:	Inteligencia artificial
Idioma del trabajo:	Castellano
Palabras clave:	Cáncer de próstata, microRNA, Clustering

Agradecimientos

En primer lugar es para mi obligado agradecer a Samir, mi consultor en este proyecto, por sus importantes indicaciones para ponerme en el camino correcto en un campo del que nada sabía en un principio, y al que miro con inmenso respeto, pues inmenso es el reto.

En segundo lugar, indicar que los resultados de este proyecto se basan en datos generados por The Cancer Genome Atlas Research Network: http://cancergenome.nih.gov/.

A mi familia por el apoyo, el cariño y la atención en todo lo que hago.

A mi pareja, por sufrir las horas en compañía y recorrer juntos el camino.

Gracias, esto también os pertenece.

“We can only see a short distance ahead, but

we can see plenty there that needs to be done.”

Alan M. Turing (1912-1954).

Resumen

El Cáncer de Próstata (PCa) es el cuarto tipo de cáncer más común en el mundo. Su heterogeneidad complica mucho el desarrollo de marcadores fiables para la diagnosis y prognosis, y así evitar las biopsias y tratamientos innecesarios. Los microRNAs son un tipo de RNA no codificante que han demostrado tener un papel importante en la aparición y progresión del cáncer. Estos genes muestran un gran potencial como bio marcadores no invasivos para el diagnóstico y la monitorización del cáncer, para predecir la agresividad del tumor, la respuesta al tratamiento e incluso como objetivos terapéuticos.

El objetivo de este proyecto es analizar datos de expresión de microRNAs para detectar patrones de expresión diferenciados en muestras tumorales respecto a las normales. Para ello, se han obtenido datos pertenecientes a muestras de PCa de The Cancer Genome Atlas (TCGA). Mediante edgeR se ha realizado un análisis de expresión diferencial y se han eliminado aquellos microRNAs que no están diferencialmente expresados. Después se ha aplicado clustering para explorar el conjunto de datos. En concreto se ha utilizado K-means para explorar la separación entre los dos grupos de muestras y Clustering jerárquico para visualizar las agrupaciones que forman los microRNAs.

Como resultado se han identificado numerosos microRNAs des regulados, coincidiendo muchos de ellos con los previamente reportados en la literatura. Se ha comprobado que hay patrones claramente diferenciados entre los grupos de muestras, pero es necesario profundizar en las funciones regulatorias de los microRNAs diferenciados para destapar su potencial para el diagnóstico.

Abstract

Prostate Cancer (PCa) is the fourth most common cancer type worldwide. Its heterogeneity complicates the discovery of reliable markers for the diagnosis and prognosis, thus avoiding unnecessary biopsy and treatment. MicroRNAs are a class of non-coding RNA that play a key role in the development and progression of cancer. These genes have a potential as non-invasive biomarkers for the diagnosis and monitoring of cancer, to predict tumor aggressiveness, the outcome of the treatment and even as therapeutic targets.

The aim of this project is to analyze microRNA expression data in order to detect different expression patterns in tumor samples versus normal samples. PCa samples has been downloaded from The Cancer Genome Atlas (TCGA). A differential expression analysis has been performed with edgeR in order to remove those microRNAs that are not differentially expressed. After that, clustering has been applied to explore the data set. K-means has been used to explore the separation between the two classes of samples, and hierarchical clustering to visualize the correlations between microRNAs.

Through the analysis several altered microRNAs have been identified. Among them, many had previously been reported in the literature. Clearly differenced patterns have been identified between the two classes of samples, but is still necessary to investigate the regulatory functions of the differently expressed microRNAs to uncover their diagnostic potential.

Índice

Índice de figuras

1. Introducción

1.1. Contexto y justificación del proyecto

Cáncer es el término que define un amplio conjunto de más de 200 enfermedades, que se desarrollan a consecuencia de mutaciones en genes que regulan el crecimiento celular y dan lugar a una multiplicación descontrolada de células cancerosas. Por sus características es una enfermedad muy compleja y heterogénea ya que se puede producir en cualquier parte del organismo y expandirse. Actualmente fallecen 8.2 millones de personas al año a causa de algún tipo de cáncer, lo que sitúa dicha enfermedad como una de las principales causas de muerte a nivel mundial. A pesar de los avances conseguidos hasta el momento las previsiones son pesimistas: según la Organización Mundial de la Salud (OMS), el número de nuevos casos aumentará un 70 % en las dos próximas décadas. La detección de estas enfermedades en una fase temprana es fundamental, ya que aumenta notablemente las expectativas del tratamiento y recuperación del paciente. A fin de mejorar la calidad del diagnóstico y el tratamiento, es necesario desarrollar nuevos métodos no invasivos de detección y monitorización que permitan reducir las biopsias innecesarias y el tratamiento en exceso. En los últimos años la comunidad investigadora ha centrado su atención en algunos tipos concretos de RNA que ejercen funciones reguladoras (como por ejemplo los microRNA o los RNA largos no codificantes). Estos RNAs han sido asociados con la aparición y progresión del cáncer y otras enfermedades, y se considera que tienen un potencial uso como bio marcadores del cáncer. Para poder estudiar el rol que ejercen en la regulación de la expresión genética y sus interacciones con otros RNAs es necesario disponer de datos adecuados para su análisis mediante técnicas de Machine Learning. Recientemente ha ido aumentando el volumen de datos biomédicos disponibles gracias al desarrollo de nuevas técnicas de secuenciación, y esto a su vez ha incrementado la necesidad de desarrollar nuevas herramientas de procesamiento de secuencias. La investigación genética necesita nuevos modelos computacionales que faciliten la extracción de conocimiento y ayuden a comprender los complejos procesos celulares que regulan la aparición y desarrollo de estas enfermedades. Los métodos de Machine Learning son ampliamente utilizados en tareas como la predicción o el descubrimiento de subtipos de cáncer. Así mismo, estos métodos también pueden ser utilizados para el procesamiento de datos genéticos obtenidos a partir de ensayos biomédicos. En el presente proyecto se analizará un conjunto de

muestras de niveles de expresión de microRNA con el fin de identificar la posible correlación entre los niveles de expresión de microRNAs y la aparición y/o progresión de un tipo específico de cáncer.

1.2. Descripción y objetivos del trabajo

La finalidad de este proyecto es introducir al alumno en la investigación del cáncer mediante la aplicación de métodos de Machine Learning a un conjunto de datos de expresión genética. Para ello se analizarán los niveles de expresión de microRNAs en muestras de cáncer de próstata. El objetivo del análisis propuesto es identificar los posibles microRNAs que están relacionados de algún modo con la enfermedad y pueden ser potenciales bio marcadores. A continuación, se detallan los objetivos generales del proyecto:

1. Introducción a la biología del cáncer, en concreto al papel del RNA en la expresión genética.
2. Introducción a los métodos y herramientas relacionadas con la bio computación.
3. Entender el rol del RNA en la expresión genética y su importancia como bio marcador del cáncer.
4. Explorar distintos métodos de Machine Learning relacionados con el clustering.
5. Entender los distintos tipos de datos de expresión genética e identificar fuentes de datos accesibles.
6. Analizar el problema planteado y seleccionar los métodos adecuados.

Como objetivos concretos se establecen los siguientes:

1. Aplicar los métodos adecuados para analizar la correlación entre la expresión de los microRNAs y el desarrollo de cáncer de próstata.
2. Identificar los microRNAs con potencial para servir como bio marcadores del cáncer de próstata.

1.3. Enfoque y método seguido

La elección de un lenguaje de programación para el desarrollo de este proyecto se centra en los dos lenguajes por excelencia en la comunidad científica: R y Python. Aunque ambos ofrecen excelentes características, sus diferencias hacen que sean más apropiados dependiendo del tipo de tarea a realizar.

A continuación, se introducen dichos lenguajes y se exponen sus principales pros y contras para su uso en el presente proyecto.

1.3.1. R

R es un lenguaje de programación open source creado en 1995 con el propósito de ofrecer una herramienta de uso intuitivo para el análisis de datos estadísticos y modelos gráficos. Es ampliamente usado por investigadores y científicos de datos. Aunque en un principio fue usado principalmente por la comunidad científica, en los últimos años se ha abierto paso en la industria debido en gran parte al auge del Big Data y el Machine Learning. R cuenta con su propio IDE (RStudio), que permite trabajar con un entorno gráfico intuitivo. R cuenta con una gran comunidad y numerosos paquetes que ofrecen multitud de funcionalidades adicionales. En el ámbito de este proyecto son especialmente destacables algunos paquetes ofrecidos por Bioconductor (https://www.bioconductor.org) como edgeR, DeSeq o Limma, que ofrecen implementaciones de diversos métodos específicos de análisis y procesado para este tipo de datos.

1.3.2. Python

Python es un lenguaje de propósito general creado en 1991 que destaca por su flexibilidad y sencillez. Python cuenta con IPython (Jupiter), un Notebook muy utilizado en la comunidad científica que forma parte del paquete SciPy, y que ofrece potentes funcionalidades extra. También existen diversas librerías para la computación científica como NumPy, SciPy, Pandas, Matplotlib o Sklearn. Además Python es un lenguaje con el que los desarrolladores están más acostumbrados y con el que resulta sencillo obtener prototipos funcionales rápidamente.

1.3.3. Elección

A pesar de que Python es un lenguaje con el que me siento familiarizado, para este proyecto el uso de R resulta más adecuado. R destaca sobre todo por la visualización de datos y su ecosistema de paquetes que ofrece implementaciones de los algoritmos más comunes. En el ámbito de la investigación bio informática es el lenguaje más utilizado, y ha dado como resultado multitud de paquetes específicos para

trabajar con datos genéticos. R facilitará enormemente tareas relacionadas con la obtención y procesado de datos gracias a paquetes que permiten trabajar con repositorios de datos biológicos e implementaciones de los principales algoritmos que se usarán. Como desventajas frente a Python cabe destacar que es más lento a la hora de ejecutar procesos con conjuntos de datos extensos, aunque no debería ser un problema en este proyecto. Por otro lado, será necesario dedicar algún tiempo al aprendizaje del lenguaje y sus paquetes.

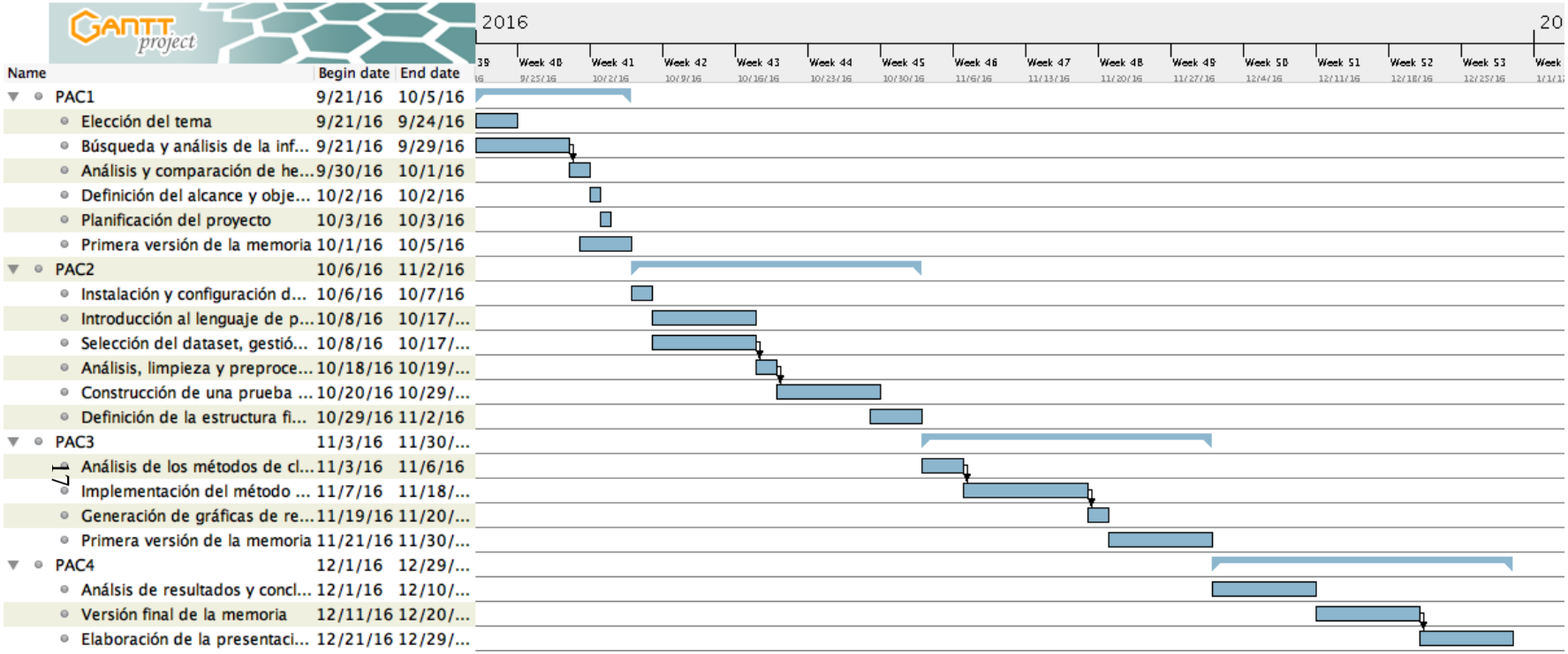
Respecto a las versiones, se utilizará la versión 3.3.1 de R y la versión 0.99.903 del IDE RStudio.

1.4. Planificación del trabajo

A continuación, se detalla la tabla de hitos, con los principales entregables del proyecto y el desglose de tareas para cada fase:

Fase	Hito	Duración	Inicio	Fin
Iniciación	**PAC1**	**15**	**21/09/16**	**05/10/16**
	Elección del tema del TFM	4	21/09/16	24/09/16
	Búsqueda y análisis de información	9	21/09/16	28/09/16
	Análisis y comparación de herramientas	2	29/09/16	30/09/16
	Definición de alcance y objetivos	1	01/10/16	01/10/16
	Planificación del proyecto	1	02/10/16	02/10/16
	Primera versión de la memoria	5	01/10/16	05/10/16
Preparación	**PAC2**	**28**	**06/10/16**	**02/11/16**
	Instalación y configuración del entorno de trabajo	2	06/10/16	07/10/16
	Introducción al lenguaje de programación y librerías	10	08/10/16	17/10/16
	Selección del dataset, gestión de permisos y obtención	10	08/10/16	17/10/16
	Análisis y limpieza del dataset	2	18/10/16	19/10/16
	Construcción de una prueba de concepto	10	20/10/16	29/10/16
	Definición de la estructura final de la memoria	5	29/10/16	02/11/16
Ejecución	**PAC3**	**28**	**03/11/16**	**30/11/16**
	Análisis de los diferentes métodos de clustering	4	03/11/16	06/11/16
	Implementación del método de clustering seleccionado	12	07/11/16	18/11/16
	Generación de gráficas de resultados	2	19/11/16	20/11/16
	Primera versión de la memoria	10	21/11/16	30/11/16
Cierre	**PAC4**	**28**	**01/12/16**	**28/12/16**
	Análisis de resultados y conclusiones	10	01/12/16	10/12/16
	Versión final de la memoria	10	11/12/16	20/12/16
	Elaboración de la presentación en vídeo	7	21/12/16	28/12/16

A continuación, se muestra el diagrama de Gantt correspondiente al a planificación de la tabla anterior:


GANTT project
2016
20
Week 40
Week 41
Week 42
Week 43
Week 44
Week 45
Week 46
Week 47
Week 48
Week 49
Week 50
Week 51
Week 52
Week 53
Name
Begin date
End date
PAC1 9/21/16 10/5/16
Elección del tema 9/21/16 9/24/16
Búsqueda y análisis de la inf... 9/21/16 9/29/16
Análisis y comparación de he... 9/30/16 10/1/16
Definición del alcance y obje... 10/2/16 10/2/16
Planificación del proyecto 10/3/16 10/3/16
Primera versión de la memoria 10/1/16 10/5/16
PAC2 10/6/16 11/2/16
Instalación y configuración d... 10/6/16 10/7/16
Introducción al lenguaje de p... 10/8/16 10/17/...
Selección del dataset, gestió... 10/8/16 10/17/...
Análisis, limpieza y preproce... 10/18/16 10/19/...
Construcción de una prueba ... 10/20/16 10/29/...
Definición de la estructura fi... 10/29/16 11/2/16
PAC3 11/3/16 11/30/...
Análisis de los métodos de cl... 11/3/16 11/6/16
Implementación del método ... 11/7/16 11/18/...
Generación de gráficas de re... 11/19/16 11/20/...
Primera versión de la memoria 11/21/16 11/30/...
PAC4 12/1/16 12/29/...
Análsis de resultados y concl... 12/1/16 12/10/...
Versión final de la memoria 12/11/16 12/20/...
Elaboración de la presentaci... 12/21/16 12/29/...

1.5. Productos obtenidos

Este proyecto no tiene como finalidad generar un producto concreto, si no la aplicación de determinados métodos para poder extraer conclusiones acerca de los microRNAs estudiados en el conjunto analizado. El resultado de este proyecto será un conjunto de microRNAs con potencial para distinguir entre muestras normales y tumorales y, por lo tanto, candidatos como bio marcadores para la detección del PCa.

1.6. Estructura del documento

Estado del arte Se introducirán los conceptos básicos relacionados con los microRNAs y su importancia como potenciales bio marcadores del cáncer. También se introducirán los métodos de Machine Learning que se emplean para la detección de microRNAs que juegan un papel relevante en la enfermedad.

Cáncer de próstata y microRNAs En este apartado se introducirán dos de los conceptos principales entorno a los cuales se centra este proyecto. Por un lado se explicará el cáncer de próstata, los síntomas y los métodos de diagnóstico y tratamiento. Por otro lado se introducirán los microRNAs y el rol que ejercen en este tipo de cáncer.

Algoritmos y métodos utilizados Aquí se presentarán los algoritmos y métodos utilizados en este proyecto para realizar el análisis propuesto. Se presentarán los algoritmos de clustering y el paquete edgeR utilizado en el análisis de genes diferencialmente expresados (DEGs).

Materiales y métodos En este apartado se explicará en detalle el proceso llevado a cabo incluyendo secciones de código, así como las decisiones tomadas en cada uno de ellos.

Resultados obtenidos En este apartado se presentarán todos los resultados obtenidos del análisis llevado a cabo en el proyecto.

Discusión En esta sección de carácter más crítico se inicia la conclusión del proyecto. Se discutirán los resultados y se compararán con otros trabajos de la literatura. Se valorarán los aspectos positivos y negativos resultantes del proyecto y se destacarán los puntos que puede ser interesante profundizar.

Conclusiones y futuros proyectos Por último se expondrá una síntesis de lo que ha significado el presente proyecto. Se realizará una valoración de los objetivos establecidos incialmente, se remarcarán los aprendizajes obtenidos en el desarrollo del mismo tanto a nivel personal como del trabajo realizado, y se trazarán los caminos que se abren en el horizonte fruto de este proceso.

2. Estado del arte

El campo de la bio medicina está uniendo sus fuerzas con el Machine Learning para conseguir diagnósticos más precoces y mejorar los tratamientos de enfermedades complejas como el cáncer. Cada vez se genera un mayor volumen de datos y se están haciendo esfuerzos para que sean accesibles por la comunidad científica.

En los últimos años se ha puesto de manifiesto la importancia de los datos de expresión genética, en concreto de RNA no codificante, ya que parecen ser la clave para mejorar la detección de enfermedades con causas genéticas, e incluso se considera que estos RNAs pueden ser objetivos terapéuticos para mejorar el resultado de los tratamientos. Sin embargo, aun se desconocen muchas de sus funciones y predecirlas es una tarea muy compleja que requiere grandes cantidades de datos adecuados y nuevos métodos que contemplen sus características específicas.

Se han puesto en marcha diversos proyectos de investigación con un doble fin: por un lado, descubrir las mutaciones genéticas inplicadas en el cancer y, por otro lado, centralizar datos genéticos y hacerlos accesibles públicamente para fomentar la investigación bio informática. Algunos ejemplos son The Cancer Genome Atlas que se inició en 2005 a fin de catalogar mutaciones genéticas y cuyos datos son utilizados en este proyecto.

Los llamados métodos de la siguiente generación (Next-Generation Sequencing o NGS) han permitido reducir los costes y a la vez procesar en paralelo miles de genes. Las mediciones resultantes de estos experimentos permiten el posterior análisis para, por ejemplo, detectar genes con diferentes patrones de expresión que permitan diferenciar las células normales de las tumorales o descubrir nuevos subtipos de un tipo de tumor, lo que permite diferenciar el tratamiento. Han sido los análisis de estos datos los que han permitido revelar la importancia de los microRNAs (un tipo de RNA no codificante) en el desarrollo de enfermedades como el cáncer.

En el artículo [21] de 2008 se realiza un estudio comparativo de diversos métodos de clustering, así como diferentes medidas de similitud y métodos de normalización de datos. Debido a su facilidad de configuración, el clustering jerárquico es el método prefererido por los bioinformáticos, lo que resulta en una cierta resistencia a apostar por nuevos métodos. El formato más común para visualizar el resultado

del clustering jerárquico es el dendrograma, una representación en forma de árbol que permite ver de una forma muy simple las relaciones entre los distintos genes. Los nuevos métodos en cambio son más complejos pero se consiguen adaptar mejor a las características de los datos de expresión genética.

En el artículo [2] de 2016 se analiza un amplio conjunto de artículos publicados entre 2007 y 2016 sobre microRNAs para diagnóstico de PCa y se recogen muchos microRNAs que se han reportado como desregulados en distintas investigaciones. Estos microRNAs reportados servirán como referencia para comparar los resultados que se obtengan en el análisis que se llevará a cabo.

Los microRNAs tienen funciones importantes en procesos celulares clave y la investigación ha asociado su desregulación con la aparición y desarrollo de algunas enfermedades. En el contexto del cáncer, éstos pueden actuar como supresores o como oncogenes, e incluso también tienen un papel importante en la metástasis. Este proyecto se centra en aplicar técnicas de clustering en muestras de niveles de expresión de microRNAs para identificar perfiles de expresión característicos que permitan diferenciar las muestras normales de las tumorales.

3. Cáncer de próstata y microRNAs

3.1. Cáncer de próstata

3.1.1. Definición y síntomas

La próstata es un órgano del tamaño de una nuez que forma parte del sistema reproductor masculino y que se encuentra ubicado en la salida de la vejiga alrededor de la uretra. Su función es segregar un fluido que forma parte del semen junto a fluido de la vesícula y los espermatozoides. El cáncer de próstata (PCa abreviado en inglés) aparece cuando las células se multiplican de forma descontrolada en este órgano, y afecta a hombres mayoritariamente de avanzada edad. En datos de 2012, es el tipo de cáncer más común en Europa en hombres y el tercero entre todos los tipos de cáncer. A nivel mundial, es el segundo con más incidencia en hombres y el cuarto entre todos los tipos. Este tipo de cáncer tiene una mayor incidencia en países desarrollados como EEUU y el norte de Europa y menos en Asia.

Desde la década de 1970 se ha ido incrementando la supervivencia a este cáncer hasta superar el 90 % el primer año, casi el 85 % a los 5 años y casi el 84 % a los 10 años (datos de Reino Unido). El índice de supervivencia en PCa va aumentando con la edad, y disminuye a partir de los 70 años (siendo aun más pronunciada la caída a partir de los 80 años). Aun así, este tipo de tumor tiende a crecer lentamente y muchos hombres en edad avanzada no mueren debido a esta enfermedad.

Los principales factores de riesgo del PCa son la edad, la etnia y las condiciones genéticas (el riesgo aumenta significativamente si hay familiares de primer o segundo grado que han padecido la enfermedad). A día de hoy no hay evidencias suficientemente claras de otros factores que aumenten la probabilidad de sufrir esta enfermedad, y por lo tanto no hay unas pautas claras para su prevención.

En cuanto a los síntomas de este tipo de tumor, a menudo no suele presentar ninguno en la fase inicial. Algunos de los síntomas en una fase más avanzada son orinar con mayor frecuencia, flujo de orina débil o interrumpido o la aparición de sangre en la orina o el semen. El problema es que algunos de estos síntomas se pueden dar en hombres con condiciones benignas, como la Hiperplasia Benigna de Próstata (BPH por sus siglas en inglés), que se debe al aumento del tamaño de la próstata debido a la edad, y puede provocar problemas urinarios. Por otro lado, cuando el tumor se encuentra aun más avanzado los síntomas pueden ser la pérdida de peso sin motivo aparente, dolor en espalda o caderas y fatiga.

3.1.2. Métodos de diagnóstico y tratamiento

Pruebas preliminares: Estas pruebas se llevan a cabo o bien de forma rutinaria a partir de cierta edad, o bien si el paciente ha detectado algunos de los síntomas que puedan indicar la posibilidad de la enfermedad.

- DRE: El examen rectal consiste en introducir un dedo a través del recto para palpar la próstata y determinar si hay algo irregular en ella. Este examen lo lleva a cabo el médico y se recomienda a partir de los 40 años.

- Test de PSA: El Antígeno Prostático Específico (PSA por sus siglas en inglés) es una proteína que sintetizan exclusivamente las células de la próstata (tanto las normales como las tumorales) y que se encuentra en la sangre en cantidades muy reducidas, aunque aumentan con enfermedades de la próstata. El problema del PSA es que se puede encontrar también en niveles elevados en hombres con otras condiciones benignas como BPH, inflamación o infección, y por tanto hay un alto ratio de falsos positivos. La discusión acerca de si es indicado medir el nivel de PSA en hombres sin síntomas de PCa no está resuelta. Aun así el PSA también es utilizado después de la cirugía o el tratamiento para evaluar el resultado.

Pruebas de confirmación: Estas pruebas se llevan a cabo si a partir de las pruebas anteriormente expuestas se encuentran indicios de un posible PCa.

- PCA3: El gen PCA3 se muestra sobre expresado en hombres con PCa y, a diferencia de PSA, este sí es específico del PCa ya que solo lo producen las células tumorales de la próstata. Este gen se puede detectar en la orina y es utilizado para ayudar a decidir si es necesaria la biopsia cuando hay indicios de PCa.

- TRUS: Es un tipo de ecografía transrectal de ultrasonidos que se suele realizar al mismo tiempo que la biopsia.

- Biopsia: Mediante una biopsia se extrae una muestra de tejido tumoral de la próstata. Esta muestra se examina mediante microscopio y sirve para realizar el diagnóstico.

- MRI: En esta prueba se combina la ecografía TRUS y una imagen de resonancia magnética (MRI) para mejorar la detección de las zonas que es necesario evaluar mediante biopsia.

El problema que presentan estos indicadores es que no aportan información precisa sobre la respuesta del paciente al tratamiento. De hecho, el test de PSA ha supuesto una mejora en la detección del PCa pero también conlleva que se realicen biopsias innecesarias o que pacientes reciban tratamiento sin ser necesario.

A continuación, se introducen dos escalas usadas para clasificar el PCa. Por un lado, la escala de Gleason, que categoriza el tumor y, por otro lado, el método TNM que asigna un estadio dependiendo de la expansión del tumor.

Gleason Score (GS): Es una medida en la escala numérica de 2 a 10 (en la actualidad se suele usar la escala de 6 a 10) de la diferencia de una célula tumoral versus una célula normal. Cuanto más alto es el valor indica que el tejido tumoral es más diferente del normal y por lo tanto el cáncer es más agresivo y hay más probabilidad de que se extienda.

Puntuación	Descripción
2 a 6	Tumor con crecimiento lento y poca agresividad.
7	Agresividad intermedia.
8 a 10	Agresividad alta.

Cuadro 1: Descripción de las diferentes puntuaciones de la escala de Gleason.

Método TNM: TNM son las siglas de Tumor, Nodos Linfáticos y Metástasis. Es un sistema usado por los médicos para describir el estadio del tumor, habiendo del estadio 0 (no hay tumor) al estadio 4. En el siguiente cuadro se muestran los estados y una breve descripción de los mismos. Es importante destacar que cada una de las partes (T, N y M) tiene su propia escala de clasificación. Se pueden encontrar los estadios de cada parte detallados en el siguiente enlace:

http://www.cancerresearchuk.org/about-cancer/prostate-cancer/stages/tnm-staging.

Estadio	Descripción
I	El tumor está en una fase temprana de desarrollo. Solo se encuentra en la próstata y no es detectable mediante DRE ni pruebas de imagen.
II	En este estadio el tumor podría ser detectado mediante DRE.
III	El tumor se ha extendido a tejidos cercanos e incluso a la vesícula.
IV	El tumor se ha extendido a cualquier otra parte del cuerpo.
Recurrente	El cáncer reaparece después del tratamiento.

Cuadro 2: Descripción de los diferentes estadios del cáncer de próstata.

En caso de confirmarse la presencia de PCa, la decisión de llevar a cabo el tratamiento, y qué tipo de tratamiento se adecúa mejor depende de varios factores. Obviamente, el estadio en que se encuentre el PCa será un factor determinante, pero también entran en juego la edad o el estado de salud del paciente. En algunos casos se puede incluso decidir monitorizar de forma activa la evolución del PCa y no tratarlo hasta que avance, ya que el tratamiento produce efectos secundarios como la impoo. Por otro lado, cuando el tumor es resistente a los tratamientos estándar se considera incurable.

A continuación se enumeran los tratamientos más comunes del PCa:

- Cirugía: Mediante la cirugía (que puede ser de distintos tipos) a menudo se extirpa parte o la totalidad de la próstata y las vesiculas, y en algunos casos los nodos linfáticos de la pelvis.
- Quimioterapia: Consiste en administrar fármacos directamente al corriente sanguíneo para matar las células cancerosas. Los efectos secundarios durante el tratamiento son severos y diversos. Algunos pueden continuar o reaparecer después del tratamiento.
- Radioterapia: Se usan rayos de alta potencia para matar las células cancerosas y hay diversos tipos. Al igual que la quimioterapia también provoca efectos secundarios severos diversos.

3.2. MicroRNAs

Los microRNAs (también abreviados como miRNA) son un tipo de moléculas de RNA no codificante (ncRNA) de corta longitud (22 nucleótidos de media) y de cadena simple, cuyas funciones son la silenciación de RNA mediante la regulación negativa de genes diana (target en inglés), y la regulación de la expresión de genes diana interfiriendo en el proceso postranscripcional. La regulación de la expresión génica es un conjunto de mecanismos celulares para incrementar o decrementar la producción de

proteínas o RNA. Un microRNA puede regular (generalmente inhibir) la expresión de múltiples RNA mensajeros (mRNA) uniéndose a las regiones complementarias de éste y, a su vez, un mRNA puede ser diana de múltiples microRNAs, lo que conlleva que la red de regulación sea muy compleja.

El descubrimiento de este tipo de ncRNA es reciente (Lee, 1993), de ahí que todavía no se conozcan todas sus funciones aunque se ha demostrado su implicación en procesos biológicos muy importantes como la proliferación, diferenciación o migración celular. También se conoce que la desregulación de microRNAs tiene implicaciones en algunas enfermedades, con lo que se ha abierto un horizonte en el estudio de los microRNAs para la diagnosis, prognosis y tratamiento de enfermedades con base genética.

3.2.1. Importancia de los microRNAs en el cáncer

Los tratamientos para combatir el cáncer son demasiado genéricos en la actualidad. El futuro está en los tratamientos personalizados de forma que se consiga mayor efectividad en la respuesta del paciente y a su vez se minimicen los efectos secundarios. En este sentido, los estudios acerca de los microRNAs han demostrado que tienen capacidad para personalizar los tratamientos e incluso el desarrollo de nuevos fármacos. Por otro lado los microRNAs circulatorios son muy estables en fluidos como la orina y, por lo tanto, pueden ser marcadores no invasivos. En la actualidad se han encontrado muchas relaciones entre microRNAs y la patología del PCa, como por ejemplo la recaída después del tratamiento o la metástasis. Se distinguen varios grupos de microRNAs involucrados en el cáncer dependiendo de su rol:

- Promotores: Llamados oncomirs, están relacionados con la aparición y progresión del cáncer.
- Metastamirs: Grupo de microRNAs asociados con la agresividad del tumor y la metástasis.
- Supresores: Intervienen en procesos que inhiben el desarrollo del cáncer (control del ciclo de vida de la célula por ejemplo).

4. Algoritmos y métodos utilizados

4.1. Reducción de la dimensionalidad

En múltiples fuentes como por ejemplo en [4] se aconseja utilizar métodos de reducción de la dimensionalidad previo al clustering para obtener un mejor resultado. Esto es debido a que la mayoría de genes no aportan información útil que permita distinguir entre muestras normales y muestras tumorales. En el caso de datos de expresión genética, en este paso se pueden seleccionar los genes que están diferencialmente expresados (DEG) descartando así los que no aportan variabilidad al conjunto. Aunque la detección de DEG puede ser el fin del análisis también puede ser utilizado como paso previo al clustering, lo que permite eliminar genes que distorsionan los patrones que diferencian las muestras.
A continuación se presenta el paquete edgeR que se usará para realizar el análisis de genes diferencialmente expresados (DEG) en R.

4.1.1. EdgeR

Para el propósito anteriormente mencionado se ha utilizado edgeR, un popular paquete de Bioconductor que implementa métodos estadísticos para detectar cambios en los niveles de expresión en diferentes condiciones y así identificar genes diferencialmente expresados (DEG). Su elección se debe a las múltiples referencias en la literatura que lo sitúan como el paquete que ofrece mejor resultado con datos de RNA-Seq.
EdgeR toma como entrada una matriz de expresión con lecturas sin normalizar y con los identificadores de los genes como nombres de las filas, y los identificadores de las muestras como nombres de las columnas. Para llevar a cabo el análisis modela internamente los datos con una distribución binomial negativa. El motivo de su uso es que los datos generados con tecnología RNA-Seq son lecturas y por lo tanto son discretos, a diferencia de los datos generados por la tecnología de microarrays que son continuos. Estas lecturas pueden ser transformadas en una distribución continua pero se han desarrollado otras aproximaciones específicas para trabajar con este tipo de datos. Además, el uso de esta distribución en vez de Poisson es debido a que los datos de RNA-Seq tienen una varianza mayor a la media (este efecto se llama sobre dispersión) y en la binomial negativa se puede parametrizar la dispersión.

Un aspecto a tener en consideración previo al análisis es que es necesario filtrar aquellos genes que presentan lecturas bajas o que no están expresados. Por otro lado, este paquete usa por defecto el método Trimmed mean of M-values (TMM) para calcular los factores de normalización, aunque también se pueden seleccionar otros como Upper-quartile. Mediante estos factores se minimiza el log-FC en las lecturas de la mayoría de los genes entre las muestras para eliminar el efecto que producen genes con un número muy elevado de lecturas, dejando poco rango dinámico en la secuenciación de otros genes.
Algunos conceptos importantes en el proceso de edgeR:

- Library size: Es la suma total de las lecturas de todos los genes para una muestra.
- Effective library size: El resultado de multiplicar la library size por el factor de normalización.
- Exact test: edgeR utiliza una idea paralela al Test Exacto de Fisher para encontrar diferencias en la media de expresión de los genes entre dos clases mediante un análisis estadístico.

Aunque edgeR es el paquete más usado también existen otros paquetes alternativos para realizar este tipo de análisis. En la literatura se pueden encontrar múltiples referencias a DeSeq y BaySeq entre otros. Por último, las instrucciones de instalación del paquete edgeR se pueden encontrar en el Anexo 1.2.

4.2. Distancia Euclídea

La distancia Euclídea es la distancia entre dos puntos en el espacio Euclídeo. En el campo del Machine Learning, esta distancia es comúnmente utilizada como medida de similitud entre observaciones en algunos tipos de algoritmos como por ejemplo los de clasificación o los de clustering. En el ámbito de este proyecto, esta distancia permitirá determinar la similitud entre los distintos microRNAs a través de las muestras. Por lo tanto, (x, y) serían dos microRNAS, y n el número de muestras. La distancia Euclídea entre dos puntos (x, y) en un espacio de n dimensiones es una generalización del Teorema de Pitágoras y viene definida por la siguiente ecuación:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

Un problema a tener en cuenta con el uso de la distancia Euclídea es que es sensible a distintas escalas en las variables y, por lo tanto, es necesario estandarizar los datos al trabajar con esta métrica. Como alternativa se podría considerar la correlación de Pearson, la cual no se ve afectada por las escalas. Sin embargo la elección de la distancia Euclídea en este proyecto se debe a dos motivos. Por un lado, el algoritmo K-means utiliza esta métrica para determinar la distancia entre las observaciones y los centroides. Por otro lado, en la literatura se ha contrastado que, a pesar de no haber significativas diferencias en los resultados utilizando distancia Eucídea o la correlación Pearson, la primera es más indicada para datos provenientes de experimentos RNA-Seq. En cambio la correlación de Pearson funciona mejor con datos que provienen de tecnología Affymetrix.

4.3. Algoritmos de clustering

La técnica de clustering comprende diferentes algoritmos cuya finalidad es agrupar un conjunto de observaciones en diferentes clusters en base a la similitud entre ellas para revelar agrupaciones en los datos. Esta técnica sirve para explorar los datos y descubrir o hipotetizar asociaciones entre ellos. En el contexto de este proyecto se aplicará clustering para revelar patrones de expresión diferenciados entre las células normales y las tumorales que puedan servir para diferenciarlas y, por lo tanto, tengan potencial para diagnosticar la enfermedad. A continuación se introducen los dos algoritmos aplicados en este proyecto para el análisis de los datos de expresión de microRNAs.

4.3.1. K-means

K-means es un algoritmo iterativo que se enmarca en la rama del aprendizaje no supervisado o clustering. El objetivo es encontrar la mejor división de un conjunto de n observaciones (cada una de ellas con su conjunto de features o variables) en k grupos en base a la similitud entre ellas. Esto lo consigue minimizando la suma de las distancias entre las observaciones asignadas a un cluster y el centroide del mismo. El resultado debe satisfacer dos condiciones: por un lado, los clusters creados deben ser homogéneos (las observaciones de un mismo cluster deben ser similares entre sí), y por otro lado debe haber separación entre los clusters (los elementos de diferentes clusters son poco similares entre sí).
El algoritmo toma como entrada un conjunto de n observaciones y un número k (tal que $1 \leq k \leq n$) que

indica el número de clusters en que se ha de dividir el conjunto. Como salida produce un vector con los centroides y otro de etiquetas o labels con la asignación de un cluster a cada observación.
Algunos aspectos a tener en cuenta con este algoritmo son:

- Diferentes resultados en base a los centroides iniciales (elegir varios y tomar la mejor solución).
- Es necesario establecer un máximo de iteraciones ya que a veces puede no converger.
- Utiliza la distancia Euclídea para medir la distancia entre observaciones y centroides.

Algoritmo 1 Pasos del algoritmo de clustering *K-means*.

```
Entrada: X = {x1, x2 ... xn}
         k (número de clusters)
Salida:  C = {c1, c2 ... ck} Conjunto de centroides
         L = {l1, l2 ... ln} Labels (clusters) asignados a cada x
1. Seleccionar k observaciones como centroides de forma aleatória
2. Iterar hasta converger (los centroides se mantienen estables)
  2.1. Calcular las distancias entre todas las observaciones y cada
       centroide. Asignar cada observación al centroide más cercano.
  2.2. Recalcular la posición de los centroides con la media de las
       observaciones asignadas a su cluster.
```

Debido a su eficiencia *K-means* puede ser utilizado incluso con grandes conjuntos de datos. Es por ello que fuera del ámbito de este proyecto también es ampliamente utilizado para resolver otro tipo de problemas como la segmentación de imágenes, la clasificación de textos o la construcción de recomendadores. Existen otros algoritmos alternativos como por ejemplo *PAM* (Partitioning Around Medoids) o *SOM* (Self organizing Maps).

4.3.2. Clustering jerárquico

El clustering jerárquico es una alternativa a *K-means* para identificar agrupaciones en los datos y explorar posibles correlaciones. Este tipo de algoritmo se diferencia de *K-means* por su filosofía jerárquica en vez de particional, y por no necesitar que se especifique el número de clusters en que se agruparán los datos. Este algoritmo se puede aplicar con una estrategia aglomerativa (bottom-up) o divisiva (top-down). El

enfoque aglomerativo es el más común y el que se ha aplicado en este proyecto, por ello solo se detallará este enfoque.

En el algoritmo aglomerativo parte de un estado inicial en el que cada observación se considera un cluster representando una hoja del árbol, y en cada paso los clusters con menor distancia se van uniendo formando una jerarquía hasta que se han unido todas las observaciones en un cluster (el de más alto nivel llamado raíz). El resultado del algoritmo se puede visualizar mediante un dendrograma. En esta representación la altura de un cluster indica la distancia o similitud de los dos clusters que lo forman.

Algoritmo 2 Pasos del algoritmo de clustering jerárquico aglomerativo.

```
Entrada: X = {x1, x2 ... xn}
1. En el estado inicial hay n clusters (n = nº de observaciones).
2. Iterar hasta unir todos las observaciones.
  2.1 Unir los dos clusters con menor distancia entre ellos.
```

Hay varios métodos de enlace (linkage methods) para calcular la distancia o similitud entre clusters. En este proyecto se han utilizado *complete* y *ward* que son los de uso más extendido por ofrecer mejores resultados, aunque se presentan también algunos de los más comunes:

- Simple: Se toma la distancia entre los elementos más cercanos de los dos clusters (valor mínimo).
- Completo: Se toma la distancia entre los elementos más lejanos de los dos clusters (valor máximo).
- Media: Se toma la distancia media entre todos los elementos de los dos clusters (valor medio).
- Ward: Minimiza el total de la varianza entre los clusters.

El uso de este tipo de algoritmo es muy común en datos genéticos por su sencillez y representación visual. Por contra este tipo de algoritmo suele dar peor resultado que otros ya que, en cada paso la división o unión se basa en un criterio *greedy* o codicioso, es decir, selecciona la mejor opción en cada paso sin tener en cuenta resultados futuros.

5. Materiales y métodos

5.1. Adquisición de muestras en bases de datos públicas

Para el desarrollo de este proyecto es necesario un tipo de datos concreto: muestras de niveles de expresión de microRNAs. Se ha identificado un repositorio público de datos en el que hay muestras de cáncer de próstata. En The Cancer Genome Atlas (TCGA) se ha identificado un proyecto de cáncer de próstata que dispone de muestras de niveles de expresión de microRNA. TCGA tiene dos niveles de acceso a los datos, uno abierto y otro controlado. Los datos de acceso abierto pueden ser descargar y usados libremente, mientras que para tener acceso a los datos controlados se ha de cumplimentar y enviar una petición de acceso a datos (DACO) que debe ser aprobada por la organización y permite el acceso a ellos durante un año.

La privacidad respecto a este tipo de datos reside en que no contengan información única y se pueda identificar al paciente a través de ellos. TCGA categoriza los datos en 4 niveles dependiendo del tipo de dato, plataforma y centro. Se puede encontrar la información detallada en el siguiente enlace: https://wiki.nci.nih.gov/display/TCGA/Access+Tiers

miRNA Sequencing

Data Subtype	Cancer Types Applicable	Data Type Name	Level 1	Level 2	Level 3	Important Metadata	How to Retrieve Data Files
miRNA sequence (available at the Cancer Genomics Hub)	All except GBM	n/a	miRNA sequence for each participant's tumor sample File type: binary alignment file (.bam) (Controlled-access)	n/a	n/a	Experimental protocol, including primer information, is contained in the metadata **.xml file** associated with each .bam file on CGHub	See CGHub site
miRNA	All except GBM	miRNASeq	miRNA sequence for each participant's tumor sample - see above	n/a	The calculated expression for all reads aligning to a particular miRNA, per sample File type: tab-delimited (.txt)	Experimental protocol, including calculation methods, is included in the DESCRIPTION file of the **MAGE-TAB** archive	Data Matrix & Bulk Download: Select 'miRNASeq' For Data Type File Search: Select 'miRNA Expression' for Data Category
Isoform	All except GBM	miRNASeq	miRNA sequence for each participant's tumor sample - see above	n/a	The calculated expression for each individual miRNA sequence isoform observed, per sample File type: tab-delimited (.txt)	Experimental protocol, including calculation methods, is included in the DESCRIPTION file of the **MAGE-TAB** archive	Data Matrix & Bulk Download: Select 'miRNASeq' For Data Type File Search: Select 'miRNA Expression' for Data Category

Figura 1: Niveles de los datos de microRNAs de TCGA. Los datos marcados en rojo son de acceso controlado (Level 1). En el subtipo miRNA en la columna Level 3 se enmarcan los datos de expresión de microRNAs (Extraído de TCGA: Data Levels and Types).

Sin embargo, para el análisis que se llevará a cabo en este proyecto sólo son necesarios los datos de las medidas de expresión, que se enmarcan en el nivel 3 y son de carácter abierto tal como se puede ver el la segunda fila de la imagen, por lo cual no será necesaria una autorización ya que no incluyen información sensible que pueda vulnerar la privacidad de los pacientes.

Los distintos sets de datos y sus características pueden ser explorados a través del portal de datos (Genomic Data Commons) del National Cancer Institute. Los datos de acceso abierto se pueden descargar a través de este portal de forma manual, aunque una opción más interesante es utilizar alguno de los paquetes de R para descargar y trabajar con datos de TCGA. En este caso, entre las opciones disponibles se destaca *TCGA2STATS* [8], un paquete de uso muy simple y que ofrece las funcionalidades necesarias para descargar los datos en R y trabajar con ellos.

En concreto el data set que se utilizará es el PRAD-US que corresponde a datos de pacientes de EEUU con cáncer de próstata (Prostate Adenocarcinoma). Entre los datos disponibles se encuentran muestras de niveles de expresión de microRNAs obtenidos mediante la tecnología RNA-Seq.

5.2. Características y exploración del conjunto de datos

Mediante el paquete *TCGA2STAT* se ha descargado el conjunto de datos perteneciente a cáncer de próstata (PRAD).

```
library( "TCGA2STAT" )
prad <- getTCGA( disease = "PRAD", data.type = "miRNASeq" )
```

Una vez descargados los datos se ha utilizado la función *TumorNormalMatch* para obtener únicamente aquellas muestras que tienen un muestra normal adyacente. Una vez obtenidas se han unido en un único conjunto.

```
prad.tumNormal <- TumorNormalMatch( prad$dat )
data <- merge( prad.tumNormal$primary.tumor, prad.tumNormal$normal,
        by="row.names", all=T )
```

Una vez cargado el dataset tenemos lo que se llama matriz de expresión genética. Cada fila corresponde a un gen concreto, cada columna a una muestra, y cada celda indica el número de lecturas de un gen en una muestra (nivel de expresión). En este caso se han descargado los datos de expresión sin normalizar (raw counts) en lugar de los datos normalizados en lecturas por millón (reads per million o RPM) debido a que así lo requiere edgeR. Se muestra a continuación un subconjunto con las primeras filas y columnas de la matriz:

microRNA	TCGA-CH-5761.x	TCGA-CH-5767.x	TCGA-CH-5768.x	TCGA-CH-5769.x
hsa_let_7a_1	54002	21567	49048	31576
hsa_let_7a_2	107116	42837	97508	63120
hsa_let_7a_3	54196	21415	49161	31617
hsa_let_7b	50045	34483	48713	34718
hsa_let_7c	40065	19438	47141	19010

Cuadro 3: Primeras filas y columnas de la matriz de expresión con los valores de lecturas.

Se describen a continuación las características principales de este dataset:

Nº de muestras	Nº de muestras emparejadas	Nº de microRNAs	Variables clínicas
546	52	1046	Si

Cuadro 4: Resumen de las principales características del conjunto de datos de TCGA.

De los datos clínicos se han filtrado únicamente los correspondientes a los 52 pacientes seleccionados (los que tienen muestra de tejido normal adyacentes), y de las 21 variables clínicas se han descartado las que no aportan información, teniendo así un conjunto de 7 variables que se muestran en la siguiente tabla que podrían ser usados para analizar posibles correlaciones entre la expresión de los microRNAs y las variables clínicas (aunque no han sido utilizadas en este proyecto):

Muestra	Age	FollowUp(days)	Tstage	ResidualTumor	G.Score	PSA	daysToPSA
CH-5761	61	28	t3b	r0	9	8.49	39
CH-5767	66	458	t2c	r0	7	0.04	377
CH-5768	72	731	t3a	r0	6	0.1	634
CH-5769	48	62	t3b	r1	9	11.16	60
EJ-7115	65	2687	t3a	r0	7	0.1	204

Cuadro 5: Primeras filas de la matriz de variables clínicas de las muestras.

A continuación se muestra una visualización de las diferencias en la distribución entre muestras (en log2):

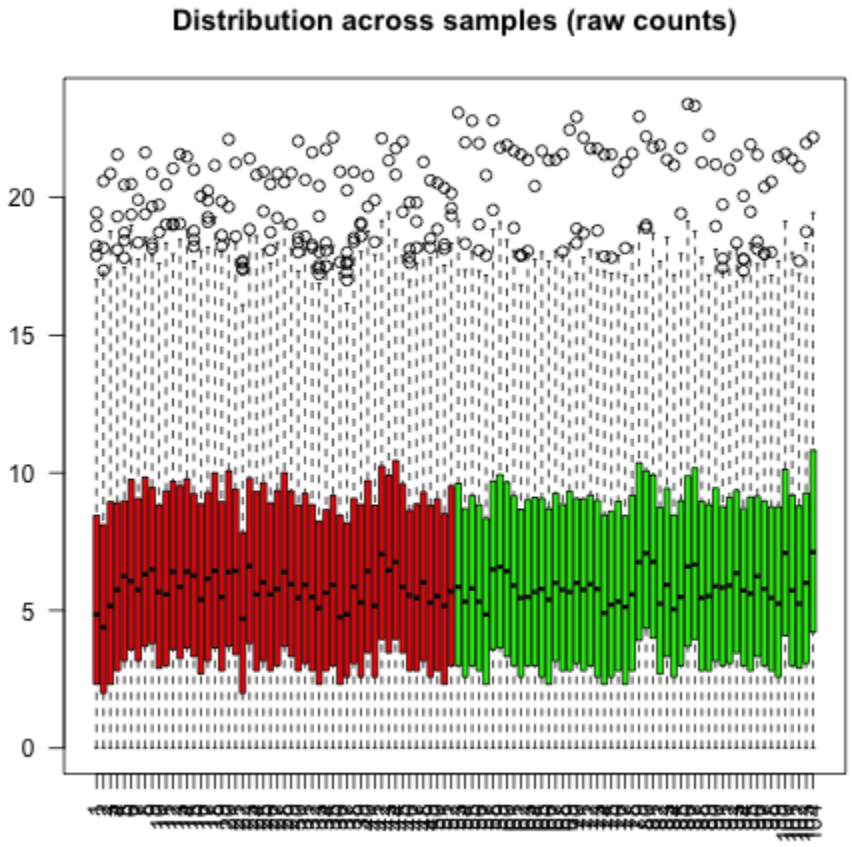


Figura 2: Distribución de lecturas en las muestras (valores sin normalizar en log2). De color rojo las muestras de tumor y de color verde las normales adyacentes.

5.3. Pre procesado de los datos

5.3.1. Valores ausentes

El conjunto de datos no presenta valores ausente. Sin embargo, una característica de los datos obtenidos en los experimentos con tecnología RNA-Seq es que pueden presentar muchos ceros. Por un lado, en el presente proyecto se pretende detectar microRNAs que puedan servir como bio marcadores para distinguir entre muestras normales y tumorales, por lo tanto interesa que sean genes que se mantengan estables en las muestras. Por otro lado para detectar genes diferencialmente expresados mediante *edgeR* es necesario eliminar genes que presenten muchos valores bajos. En el tutorial de *edgeR [12]* se propone

como criterio mantener genes con al menos 1 lectura por millón en al menos 3 muestras, aunque esto puede depender del número de muestras del conjunto.

En este caso, el criterio que se ha decidido seguir es algo más estricto: mantener aquellos genes que presentan ceros en menos de 10 muestras (que representa aproximadamente el 10% del tamaño del conjunto).

```
dge_list <- dge_list[ rowSums(cpm(dge_list) > 1) >= 10, ]
```

Como se ha visto anteriormente este conjunto presenta 1046 genes inicialmente, de los cuales se han conservado 359 después del filtrar aplicando el criterio mencionado.

5.3.2. Normalización

El paquete *edgeR* requiere lecturas sin normalizar (raw counts) para realizar el análisis ya que internamente modela los datos con una distribución binomial negativa. Una vez identificados los genes diferencialmente expresados se ha utilizado la función *equalizeLibSizes* para obtener las pseudo lecturas normalizadas, igualando así las profundidades de las muestras (library size) antes de filtrar los microRNAs identificados.

```
dge_list <- equalizeLibSizes( dge_list )
```

A continuación se muestra la distribución de los datos normalizados.

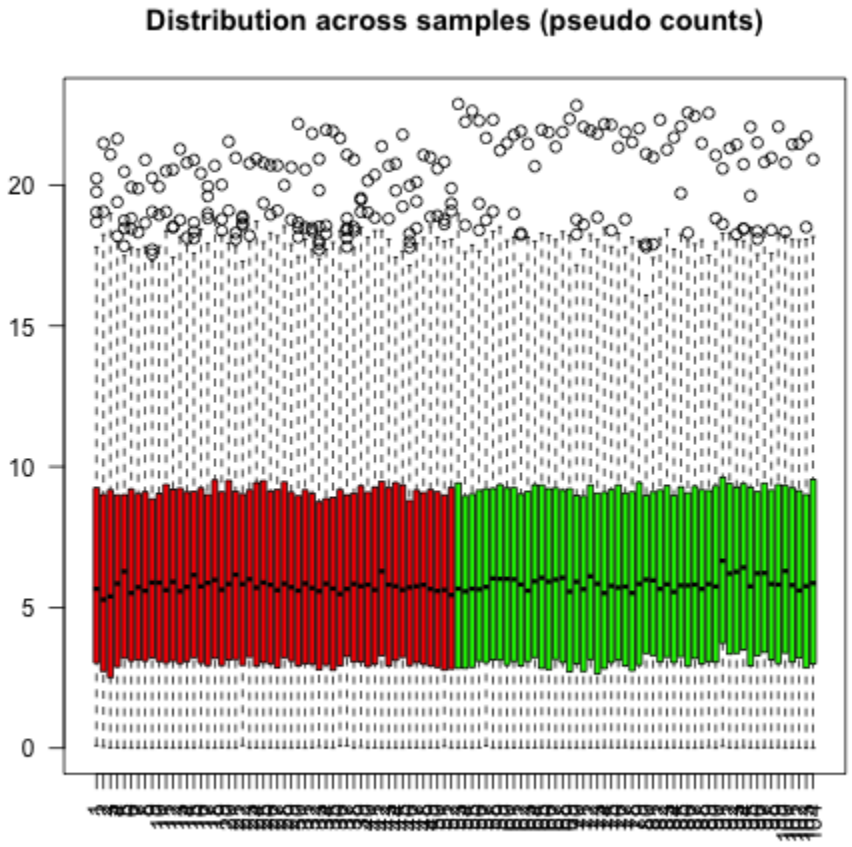


Figura 3: Distribución de lecturas en las muestras (valores normalizados en log2). De rojo las muestras de tumor y de color verde las normales adyacentes.

Además, debido a que se usará la distancia Euclídea como medida de similitud en el posterior análisis mediante clustering, es necesario centrar y escalar los datos para que no tengan mayor importancia aquellos genes con un rango dinámica más alto. Para ello se ha utilizado la función *scale* de R, que aplica normalización estándar (z-score) restando a una observación la media de las observaciones (μ) y dividiendo por la desviación estándar (σ):

$$z = \frac{X - \mu}{\sigma} \tag{2}$$

La función scale se ha aplicado a la matriz con los microRNAs diferencialmente expresados como co-

lumnas y las muestras como filas, de forma que se ha centrado y escalado la expresión de cada microRNA respecto a las muestras.

```
normCounts <- scale( t(normCounts) )
```

5.4. Selección de genes

Para reducir la dimensionalidad del conjunto de datos (número de microRNAs) se ha realizado un análisis de microRNAs diferencialmente expresados, para el cual se ha utilizado el paquete edgeR. De esta forma se han identificado los que presentan una expresión diferenciada entre los dos grupos y se han eliminado todos los demás. Para ello se han calculado los factores de normalización mediante la función *calcNormFactors* y se han estimado las dispersiones *common* y *tagwise* mediante *estimateCommonDisp* y *estimateTagwiseDisp*.

```
dge_list <- calcNormFactors( dge_list )
dge_list <- estimateCommonDisp( dge_list )
dge_list <- estimateTagwiseDisp( dge_list )
```

En la siguiente figura se puede ver la separación entre los dos grupos (tumor y normal) en un gráfico de tipo Multi-Dimensional Scaling (MDS) de 2 dimensiones generado mediante la función *plotMDS* del paquete *limma*.

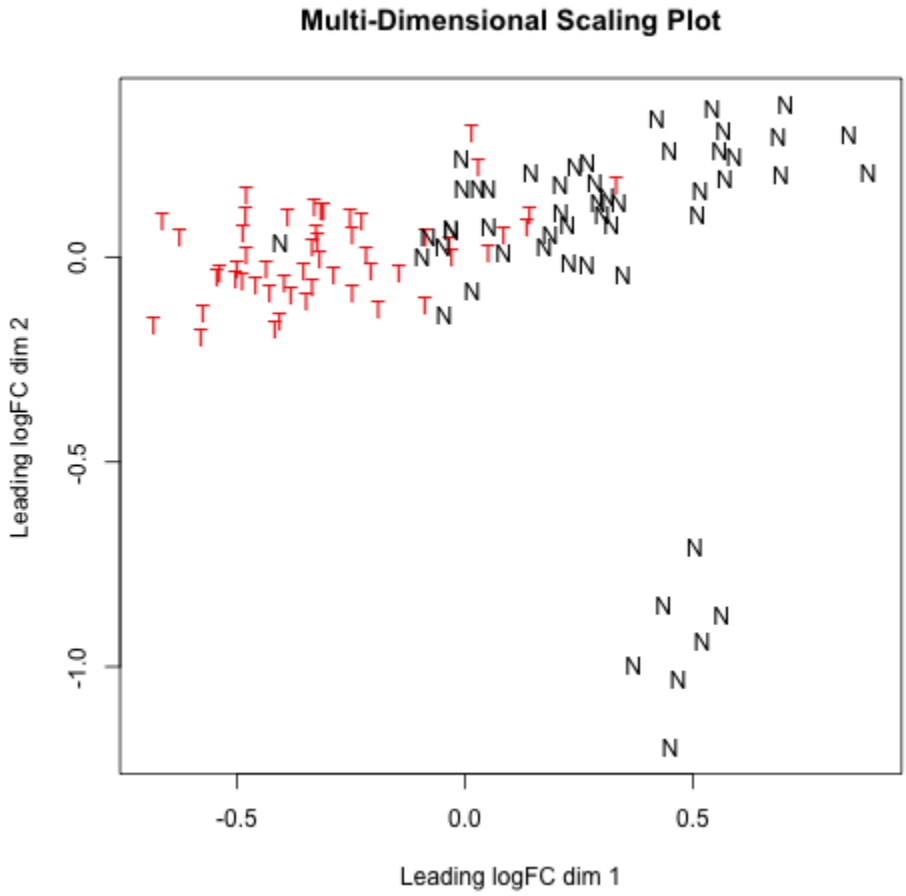


Figura 4: Proyección en dos dimensiones de las distancias entre los valores logFC de las muestras.

Como se puede observar no hay una separación clara entre los grupos, ya que hay varias muestras que se mezclan entre ellos. También se aprecia un grupo de 7 muestras normales claramente separadas que forman su propio grupo.

Para realizar el análisis de la diferencia en la expresión se ha utilizado la función *exactTest* de *edgeR*. Esta función detecta las diferencias en la media de expresión entre dos grupos, en este caso tumor y normal, normalizando internamente los datos mediante la función *equalizeLibSizes*.

```
groups <- factor( c(rep('T', 52), rep('N', 52)) )
res <- exactTest( dge_list, pair = levels(groups) )
deg_tags <- topTags( res, n = nrow(dge_list$counts) )$table
```

A continuación se muestra un gráfico en el que se pueden ver en color rojo los microRNAs diferencialmente expresados y en negro los que muestra una expresión normal. Las lineas de color azul marcan el umbral (threshold) de 2-logFC:

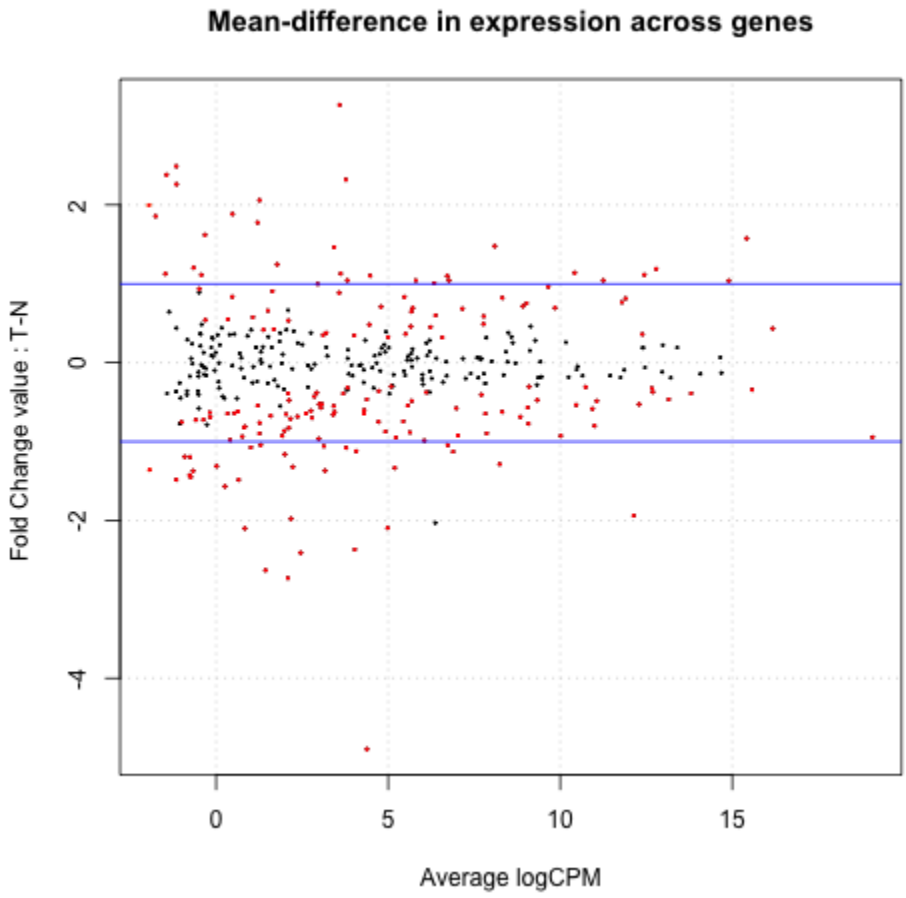


Figura 5: Dispersión de los valores. En rojo los genes que muestran un valor de expresión diferenciado (DEG) entre los dos grupos (Normal y Tumor). En negro los genes que no muestran diferencias estadísticamente significativas entre los grupos.

5.5. Clustering

Se ha aplicado el algoritmo *K-means* al conjunto para comprobar el grado de separación de las muestras. Para ello se ha utilizado la función *k-means* del paquete *stats* y la función *randIndex* del paquete *flexclust* para obtener una medida del nivel de ajuste de la agrupación creada por el algoritmo a la agrupación real

de las muestras (normal y tumor).

Se han realizado diversas pruebas con los parámetros FDR (False Discovery Rate) y logFC (log Fold Change) para seleccionar los microRNAs diferencialmente expresados. Se han ajustado para obtener el mejor resultado en el posterior agrupamiento con *K-means*. También se ha analizado la capacidad para diferenciar las muestras de los microRNAs sobre regulados y los sub regulados de forma independiente. El mejor resultado obtenido en el agrupamiento ha sido de 0.71 (mediante *randIndex*) seleccionando los microRNAs con un FDR menor a 0.01 (175 microRNAs). Sin embargo, manteniendo únicamente los 61 microRNAs que presentan además un logFC mayor a 2 (30 sobre regulados y 31 sub regulados) se obtiene un randIndex de 0.68, lo cual no supone una pérdida demasiado elevada y si una clara disminución de la dimensionalidad (la diferencia es de una muestra más agrupada incorrectamente).

Por lo tanto, el proceso se ha realizado con los datos de expresión de los microRNAs DE que se han detectado en el análisis anterior, seleccionando los que cumplen el criterio anteriormente expuesto (FRD < 0.01 y abs(logFC) > 1).

```
fdr <- 0.01 # set the FDR threshold
upreg <- rownames( deg_tags[ deg_tags$FDR < fdr & deg_tags$logFC > 1, ] )
downreg <- rownames( deg_tags[ deg_tags$FDR < fdr & deg_tags$logFC < -1, ] )
```

Como paso previo al clustering se ha utilizado la función *NbClust* del paquete *NbClust* [11] para estimar el número óptimo de grupos del conjunto.

```
library(NbClust)
res <- NbClust( normCounts, distance = "euclidean", min.nc=2, max.nc=5,
method = "complete", index = "ch")
res$Best.nc
```

Mediante el método *complete* se obtiene como mejor resultado 3 clusters.

```
Number_clusters  Value_Index
3.0000           13.5357
```

En cambio utilizando el método de enlace *ward.D2* el mejor resultado son 2 clusters.

```
Number_clusters  Value_Index
2.0000           26.6305
```

Se ha decidido realizar el clustering con *K-means* indicando 2 clusters ($k = 2$) ya que a priori se conoce que ésta es la distribución de los datos. Ya que el resultado del algoritmo depende de las condiciones iniciales, se ha indicado un valor de 20 en el parámetro *nstart* para que se elijan 20 conjuntos diferentes de forma aleatoria.

```
km <- kmeans( normCounts, 2, nstart = 20 )
table <- table( groups, km$cluster )
table
library( "flexclust" )
randIndex( table )
```

El agrupamiento da como resultado el anteriormente citado *randIndex* de 0.68 y la siguiente distribución de las muestras (la asignación a los grupos 1 y 2 puede variar entre ejecuciones):

Tipo	Grupo 1	Grupo 2
Tumor	6	46
Normal	49	3

Cuadro 6: Resultado de la ejecución del algoritmo de clustering Kmeans únicamente con los genes diferencialmente expresados. 6 muestras correspondientes a tumor han sido incorrectamente asignadas al grupo 1, y 3 muestras normales se han asignado incorrectamente al grupo 2.

Conociendo el número de clusters se ha aplicado clustering jerárquico mediante la función *hclust* del paquete *stats* (incorporado en R), utilizando como medida de similitud la distancia Euclidea y como criterios de enlace *ward.D2* y *complete*. El motivo para descartar otro criterios de enlace como *single* o *average* es que en la literatura se recomienda (en lo que se refiere a clustering con datos de expresión genética) no usar el método *single* por sus pobres resultados, y el criterio *complete* ofrece mejores resultados que *average* en la mayoría de casos [20].

```
groups <- factor( c(rep('T', 52), rep('N', 52)) )
d <- dist( normCounts, method = "euclidean" )
hc_ward <- hclust( d, method="ward.D2" )
```

También se ha aplicado clustering jerárquico para analizar las relaciones entre los distintos microRNAs.

```
mirnaGroups <- c( rep("U", 30), rep("D", 31) )
d <- dist( t(normCounts), method = "euclidean" )
hc_ward <- hclust( d, method="ward.D2" )
```

6. Resultados obtenidos

Los 30 microRNAs más sobre regulados y los 31 más sub regulados que se han detectado en el análisis y sus principales valores se muestran en la siguiente tabla. Estos son los microRNAs que tienen un FDR menor a 0.01 y al menos 2-logFC.

Sub regulados				**Sobre regulados**			
microRNA	logFC	PValue	FDR	microRNA	logFC	PValue	FDR
hsa-mir-891a	-4,8922	3,96E-21	2,03E-19	hsa-mir-1269	3,2562	7,95E-07	3,32E-06
hsa-mir-184	-2,7038	9,73E-13	1,03E-11	hsa-mir-449a	2,6514	4,55E-07	2,02E-06
hsa-mir-490	-2,5893	2,79E-08	1,50E-07	hsa-mir-153-2	2,3211	1,06E-32	1,91E-30
hsa-mir-1251	-2,3954	1,16E-08	6,81E-08	hsa-mir-1275	2,2985	4,26E-11	3,65E-10
hsa-mir-187	-2,3581	1,09E-13	1,40E-12	hsa-mir-3651	2,2290	2,79E-15	5,86E-14
hsa-mir-873	-2,2298	1,58E-11	1,42E-10	hsa-mir-615	2,1293	3,65E-10	2,62E-09
hsa-mir-204	-2,0907	2,48E-17	8,10E-16	hsa-mir-3648	2,0446	3,38E-16	8,09E-15
hsa-mir-323b	-2,0520	5,72E-09	3,67E-08	hsa-mir-153-1	1,8426	1,95E-14	3,20E-13
hsa-mir-23c	-1,9810	5,22E-15	9,87E-14	hsa-mir-3074	1,7394	8,26E-23	5,93E-21
hsa-mir-10a	-1,9390	2,16E-13	2,50E-12	hsa-mir-1304	1,6674	2,35E-06	8,77E-06
hsa-mir-135b	-1,5206	7,74E-08	3,97E-07	hsa-mir-3687	1,5762	1,12E-07	5,46E-07
hsa-mir-1258	-1,4335	1,63E-09	1,10E-08	**hsa-mir-375**	1,5744	5,06E-11	4,22E-10
hsa-mir-488	-1,3834	0,0001	0,0003	hsa-mir-92a-1	1,4768	1,99E-28	2,38E-26
hsa-mir-514-2	-1,3710	7,75E-05	0,0002	**hsa-mir-96**	1,4651	2,94E-15	5,86E-14
hsa-mir-133a-2	-1,3662	1,96E-14	3,20E-13	hsa-mir-3653	1,2510	2,26E-14	3,46E-13
hsa-mir-514-3	-1,3511	5,80E-05	0,0002	**hsa-mir-182**	1,1883	5,09E-09	3,32E-08
hsa-mir-133b	-1,3319	3,13E-14	4,30E-13	hsa-mir-592	1,1536	9,72E-05	0,0003
hsa-mir-675	-1,3059	3,70E-05	0,0001	hsa-mir-93	1,1378	1,90E-35	6,84E-33
hsa-mir-514-1	-1,2993	0,0001	0,0004	hsa-mir-190b	1,1320	0,0002	0,0004
hsa-mir-221	-1,2865	5,71E-21	2,56E-19	hsa-mir-19a	1,1295	1,82E-12	1,76E-11
hsa-mir-934	-1,2277	6,84E-07	2,96E-06	**hsa-mir-200c**	1,1137	2,81E-16	7,61E-15
hsa-mir-508	-1,1492	0,0004	0,0010	hsa-mir-3607	1,1046	1,96E-07	9,38E-07
hsa-mir-2114	-1,1398	0,0006	0,0015	hsa-mir-561	1,0993	8,71E-08	4,36E-07
hsa-mir-451	-1,1281	8,22E-07	3,35E-06	**hsa-mir-20a**	1,0983	3,75E-24	3,37E-22
hsa-mir-889	-1,1137	1,50E-13	1,86E-12	hsa-mir-20b	1,0413	7,68E-10	5,30E-09
hsa-mir-760	-1,1103	1,05E-06	4,18E-06	**hsa-mir-183**	1,0413	1,97E-06	7,61E-06
hsa-mir-652	-1,0739	3,10E-07	1,41E-06	hsa-mir-210	1,0411	2,92E-05	8,81E-05
hsa-mir-378c	-1,0552	2,10E-13	2,50E-12	hsa-mir-708	1,0402	1,54E-15	3,45E-14
hsa-mir-944	-1,0540	7,29E-06	2,52E-05	**hsa-mir-148a**	1,0371	3,23E-14	4,30E-13
hsa-mir-542	-1,0456	1,05E-10	8,38E-10	hsa-mir-19b-2	1,0067	1,86E-16	5,56E-15
hsa-mir-412	-1,0344	0,0015	0,0032				

Cuadro 7: Resumen de los DEG y sus principales valores (logFC, p-value, FDR). A la izquierda se presentan los sub regulados y a la derecha los sobre regulados. Marcados en negrita aquellos microRNAs reportados en algún artículo científico.

Aquellos microRNAs con un nivel de expresión diferencialmente mayor entre los dos grupos de muestras (sobre regulados) han demostrado tener una mayor capacidad para diferenciarlas que aquellos que muestran un nivel de expresión diferencialmente menor (sub regulados). Así, únicamente con los 30 microRNAs sobre regulados se ha obtenido un valor de 0.56 mediante *randIndex*, agrupando correctamente

50 de las 52 muestras normales, y 41 de las 52 tumorales mediante el algoritmo *K-means*. En cambio, los sub regulados se han mostrado ineficaces para recuperar la estructura de los dos grupos.

Tipo	Grupo 1	Grupo 2
Tumor	11	41
Normal	50	2

Cuadro 8: Resultado de la ejecución de K-means únicamente con los genes sobre regulados. 11 muestras correspondientes a tumor han sido incorrectamente asignadas al grupo 1, y 2 muestras normales se han asignado incorrectamente al grupo 2.

Para realizar el clustering jerárquico se han utilizado los criterios de enlace *ward.D2* y *complete*, ya que son los dos que han ofrecido un resultado que se ajusta mejor a la realidad del conjunto de datos. La agrupación parece más homogénea y compacta con *ward.D2*, ya que con *complete* hay muestras que presentan una correlación menor, e incluso alguna muestra parece no correlacionar prácticamente nada con las demás (por ejemplo la primera muestra de tumor a la izquierda en la Figura 6 forma su propio cluster). Se muestran a continuación los dendrogramas con los dos métodos especificados.

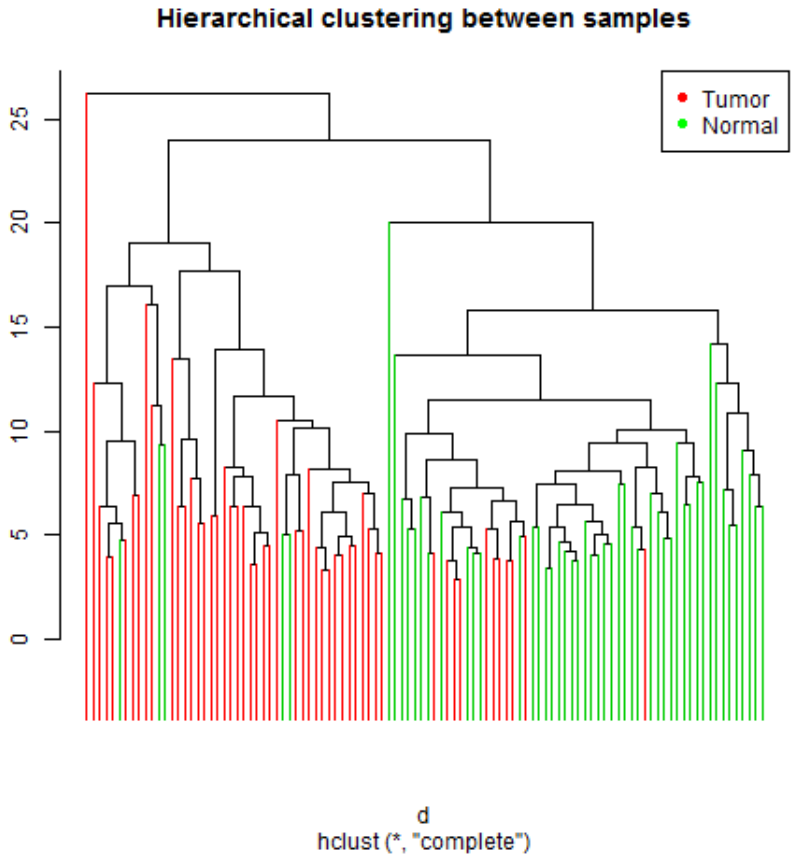


Figura 6: Dendrograma que muestra el resultado de la agrupación de las muestras entre los dos grupos mediante el método de enlace complete.

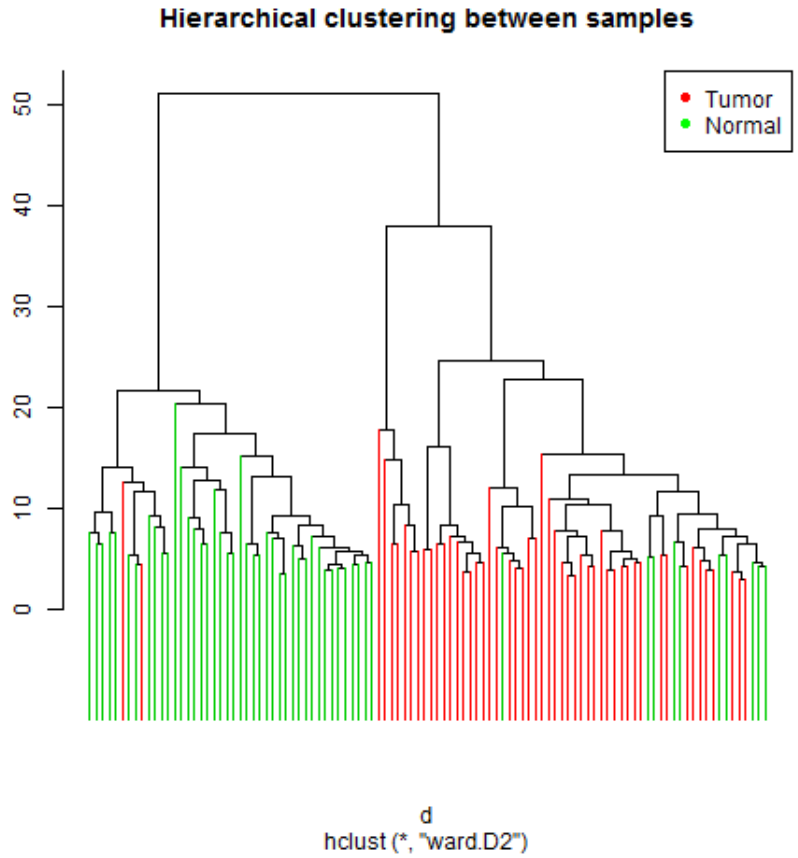


Figura 7: Dendrograma que muestra el resultado de la agrupación de las muestras entre los dos grupos mediante el método de enlace ward.D2.

Por otra parte, también se ha aplicado clustering jerárquico para agrupar los microRNAs. Primero se ha realizado sobre el conjunto de DEGs y después de forma independiente a los sobre regulados y a los sub regulados. Para los microRNAs se ha utilizado únicamente el criterio de enlace *ward.D2*. En la Figura 8 se puede ver la clara separación entre los dos grupos, y se puede apreciar también que los sobre regulados tienden a correlacionar entre ellos a un nivel inferior y de forma más homogénea respecto a los sub regulados.

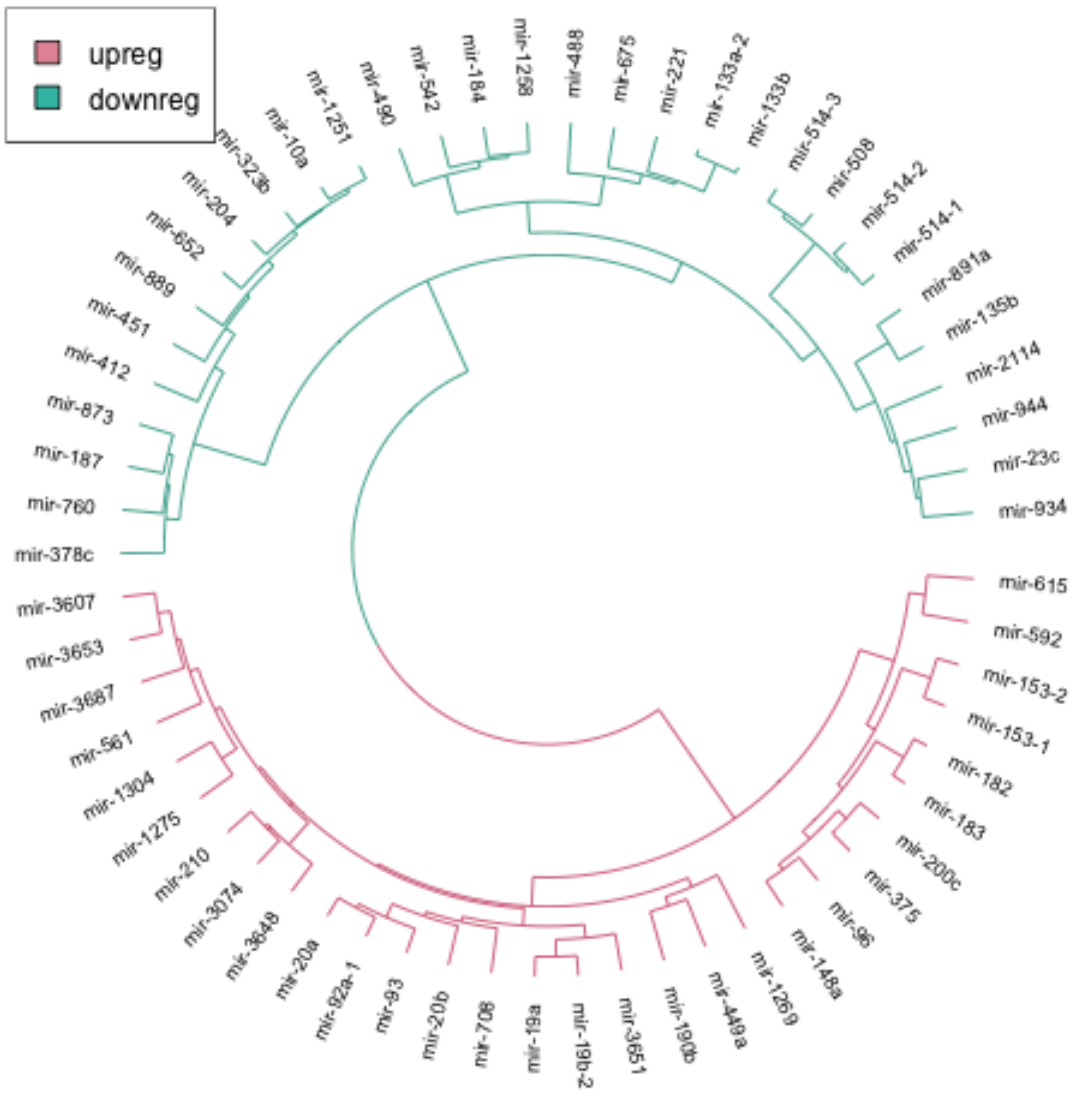


Figura 8: Dendrograma que muestra la separación entre los grupos de microRNAs sobre regulados y sub regulados.

En la Figura 9 se pueden ver las correlaciones entre los microRNAs sobre regulados. En ella se puede ver que los microRNAs que se han marcado en negrita en el Cuadro 3 (aquellos reportados en artículos por tener importancia como bio marcadores) se agrupan juntos.

Por un lado, los mir 375 y 200c muestran correlación entre ellos y se reportan con capacidad para distinguir entre tumor y normal con mayor precisión que PSA. También los mir 148a y 96 correlacionan

y parecen tener potencial para el diagnóstico. Como se puede ver estos dos grupos a su vez correlacionan entre sí formando otro cluster de más alto nivel.

Por otro lado, los mir 182 y 183 (también con potencial para diagnosticar) correlacionan juntos y se agrupan a más alto nivel con los dos grupos mencionados anteriormente. También se puede ver la correlación entre los mir 153-1 y 153-2.

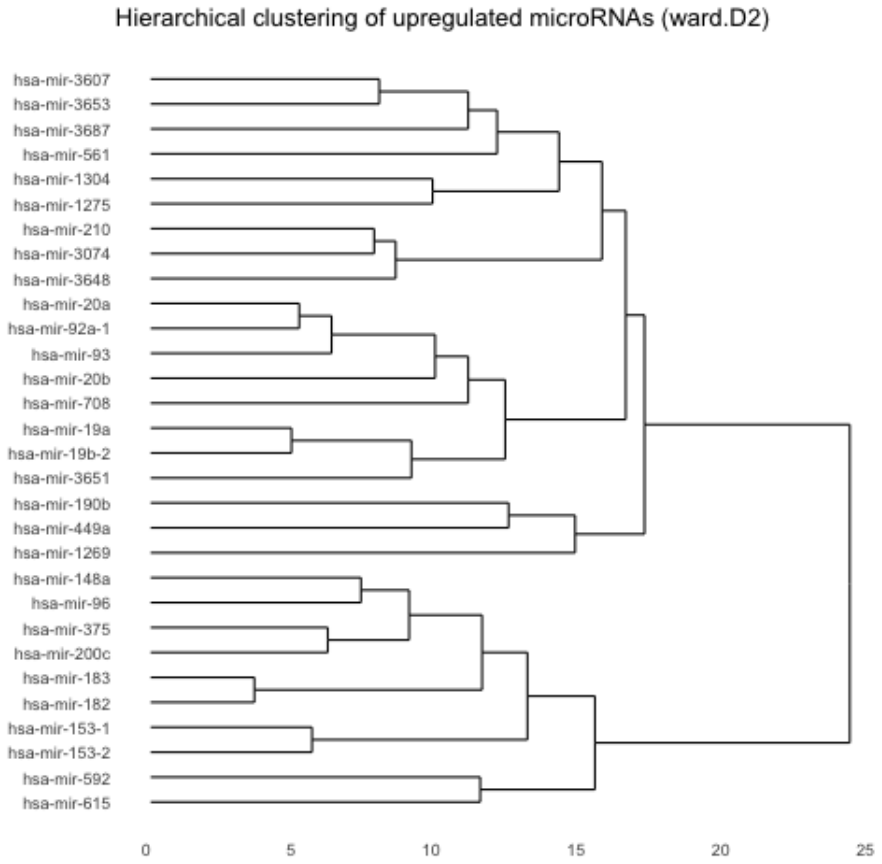


Figura 9: Dendrograma de los microRNAs sobre regulados entre las muestras de los dos grupos.

En la figura 10 se muestra el mismo dendrograma para los microRNAs sub regulados. Como se puede ver en este caso los mir-221 y 184 (también contrastados con la literatura) se encuentran en grupos relativamente cercanos.

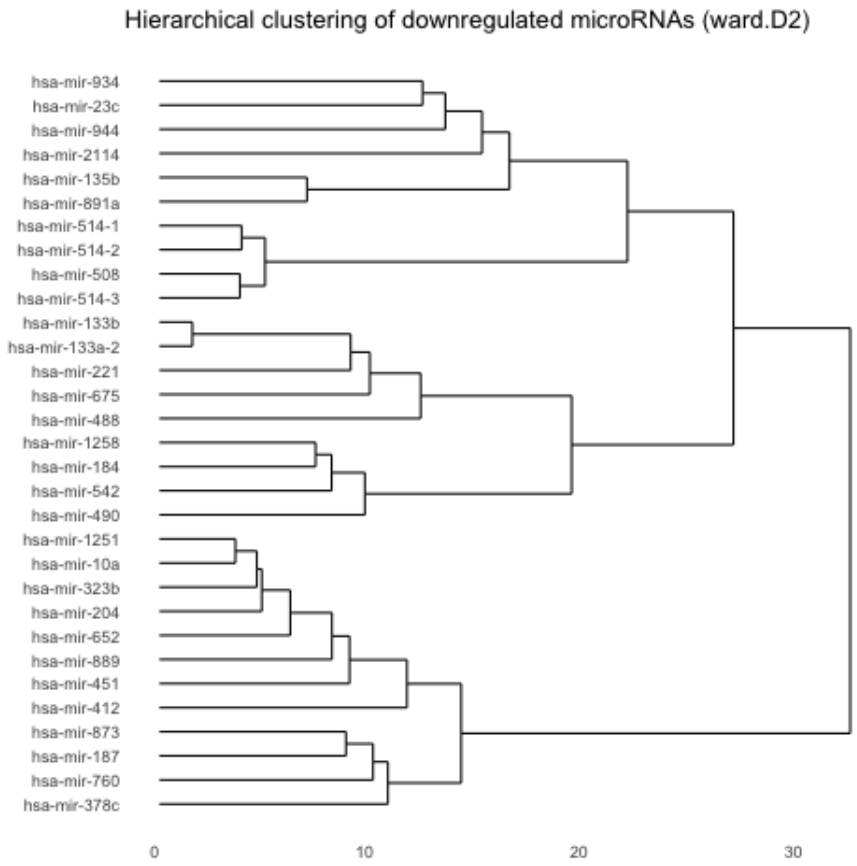


Figura 10: Dendrograma de los microRNAs sub regulados entre las muestras de los dos grupos.

Otro de los puntos interesantes es la diferencia en los patrones de expresión de los microRNAs en los grupos normal y tumor de las muestras. Para apreciar mejor las diferencias se muestran a continuación 3 gráficas de tipo heatmap: una con los microRNA sobre regulados, otra con los sub regulados y la última con los dos conjuntos.

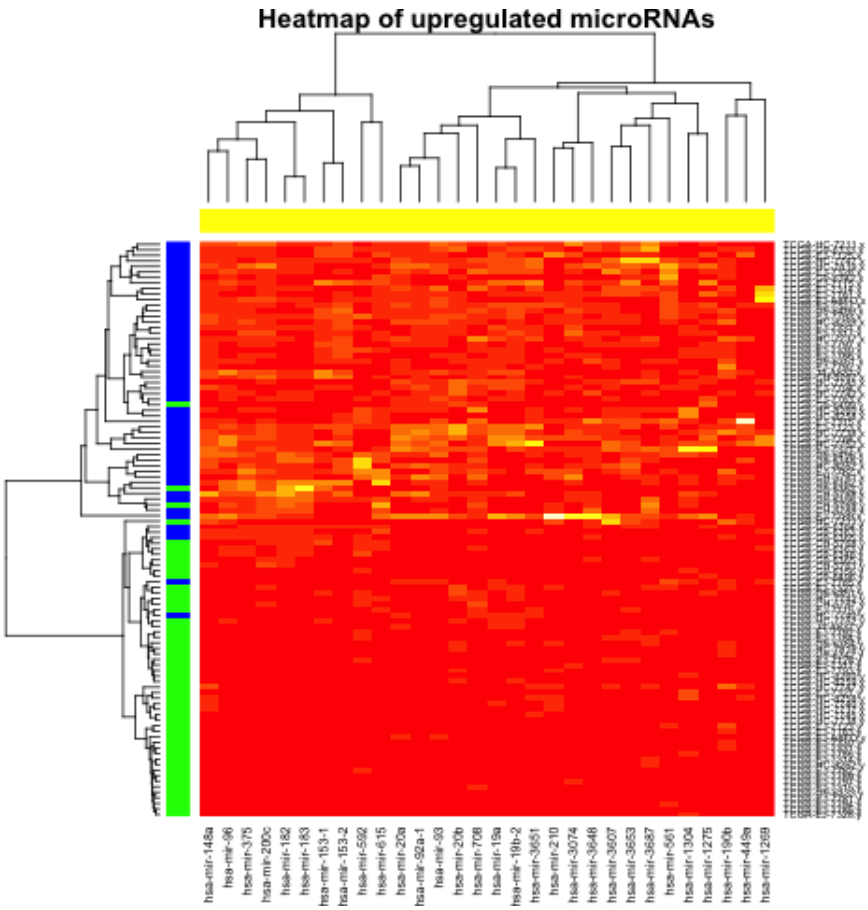


Figura 11: Heatmap de los microRNAs sobre regulados entre las muestras de los dos grupos. En la izquierda: en color azul las muestras de tumor y en color verde las normales.

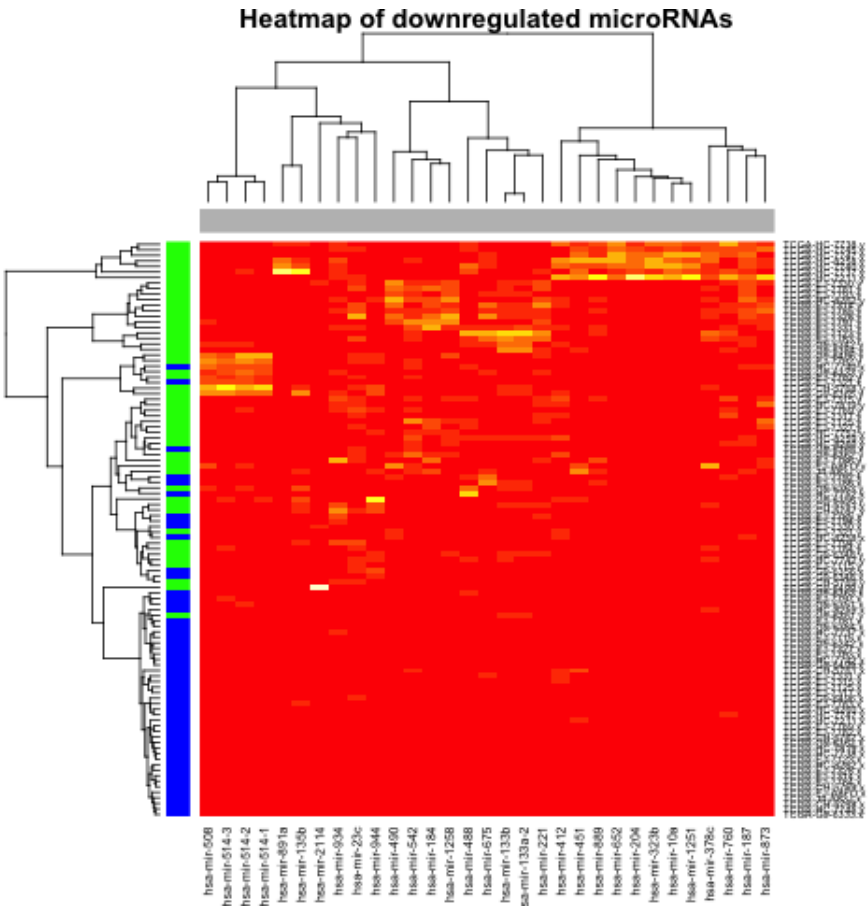


Figura 12: Heatmap de los microRNAs sub regulados entre las muestras de los dos grupos. En la izquierda: en color azul las muestras de tumor y en color verde las normales.

Comparando las dos gráficas anteriores se puede ver que los microRNAs sobre regulados muestran un patrón de expresión más consistente en las diferentes muestras y separa mejor los dos grupos de muestras (normal y tumor). En la siguiente figura se puede apreciar como estos dos conjuntos de microRNAs se expresan de forma inversa entre los grupos normal y tumor.

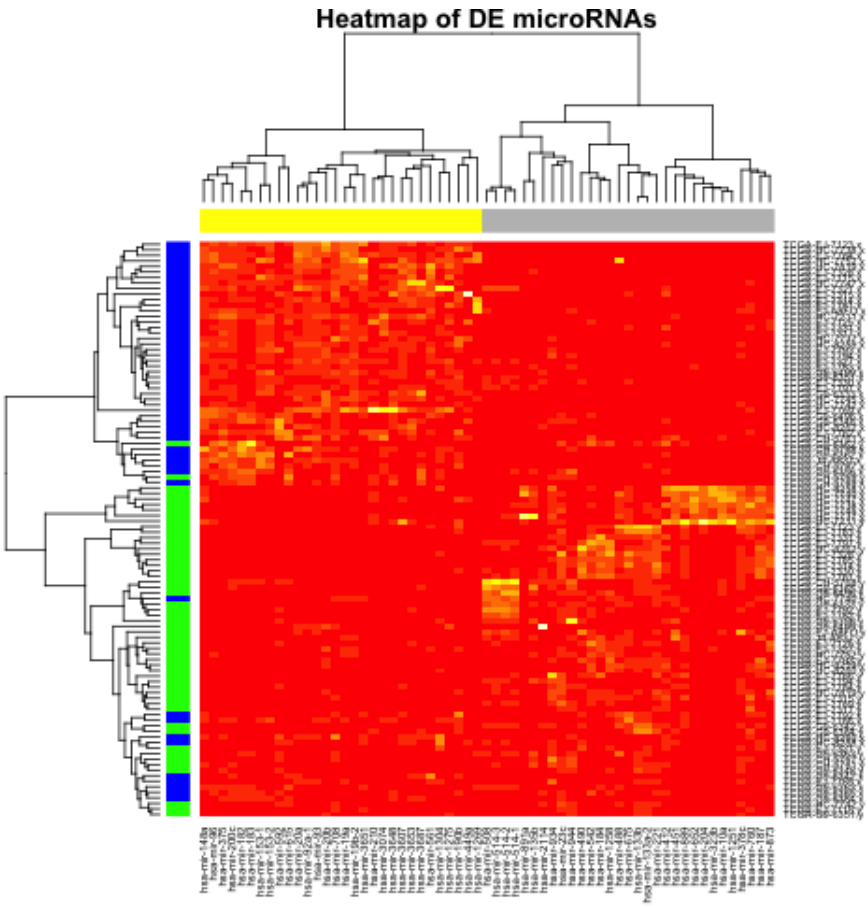


Figura 13: Heatmap de los microRNAs diferencialmente expresados entre las muestras de los dos grupos. En la parte superior: en color amarillo los genes sobre regulados y en color gris los sub regulados. En la izquierda: en color azul las muestras de tumor y en color verde las normales.

7. Discusión

Mediante el análisis llevado a cabo se han identificado microRNAs diferencialmente expresados que coinciden en gran medida con los que se reportan en varios artículos. En [5] se reportan multitud de microRNAs desregulados, de los cuales varios coinciden con los que se han identificado en este proyecto. De los sobre regulados se han encontrado como principales coincidencias los mir 148a, 96, 182, 183, 200c, 375. El mir-96 se reporta en la [29] por estar asociado con la recaída después del tratamiento, por lo tanto, tiene capacidad para predecir la prognosis. También se confirma como se puede ver en la figura 9 que los mir 375 y 182 tienen cierta tendencia a co-expresarse de forma parecida.

De los sub regulados se han encontrado coincidencias en los mir 221, 133b y 184. En [2] se reporta el mir-221 como sobre regulado, pero se indica que también ha sido reportado en otras investigaciones como sub regulado, como es el caso de este proyecto. Este microRNA está asociado con la recaída y la metástasis junto al mir-21 y al mir-145, y tienen capacidad para distinguir el riesgo del paciente. También en [29] se reporta el mir-133b. En este caso el 133b está relacionado con el deterioro de la proliferación, de ahí que en muestras de PCa se encuentre sub regulado.

Hasta aquí teniendo en cuenta únicamente los microRNAs que se ha detectado que muestran una mayor desregulación. Teniendo en cuenta microRNAs que han sido filtrados por sus valores de FDR y logFC, se encuentran varios microRNAs que se reportan en estos artículos. Se destacan los let-7a, 7b y 7c, y los mir 21, 100, 106a, 126, 130b, 141, 143, 145, 146b, 205, 222 y 330. Varios de estos microRNAs han demostrado su eficacia para diagnosticar (distinguiendo muestras de tumor de las normales) o su correlación con alguna variable como la supervivencia, agresividad del tumor, resistencia al tratamiento, etc.

Por ello, los criterios estadísticos para determinar qué genes son importantes es un punto sobre el que merece la pena reflexionar ya que, dependiendo de este criterio se obtiene un conjunto u otro de DEGs. Es necesario un consenso para determinar el mejor criterio para seleccionarlos para que éste no sea subjetivo. En este proyecto se ha puesto de manifiesto que se pueden descartar genes cuya importancia ha sido reportada en la literatura dependiendo del criterio que se seleccione para decidir cuáles muestran una diferencia significativa en su expresión.

Del análisis llevado a cabo se desprende la hipótesis de que los microRNAs sobre regulados muestran una mayor capacidad para distinguir entre los grupos normal y tumor. A pesar de que no se han encontrado evidencias de ello en los artículos revisados merece la pena contrastar esta hipótesis. Se ha profundizado más en este aspecto analizando las 7 muestras normales que en la Figura 4 no correlacionan con el resto. Se ha comprobado que eliminando estas 7 muestras varían un poco los resultados en cuanto a los microRNAs sub regulados, detectando menos con una mayor expresión diferencial e incluyendo uno reportado en la literatura que quedaba fuera del grupo de los más diferencialmente expresados, el mir-222. De este análisis se desprende la hipótesis de que estas muestras presentan niveles anormales en los microRNAs sub regulados y, por lo tanto, se considera que deberían ser eliminadas porque alteran los resultados. Se detallan a continuación los 7 identificadores de las muestras normales:

ID de la muestra
HC-7745
HC-7737
HC-7738
HC-8258
HC-7747
HC-7740
HC-7211

Cuadro 9: Identificadores de las 7 muestras cuyas muestras normales adyacentes muestran anomalías.

Por otro lado, en la literatura se ha comprobado que los datos de experimentos con RNA-Seq están altamente sesgados, en gran parte por su elevado rango dinámico. Puede ser debido a este efecto que hay muestras en este conjunto de datos que correlacionan incorrectamente. Por ejemplo, durante el clustering jerárquico de las muestras se ha comprobado que algún par de muestras se agrupan entre sí (normal y tumor). Esto podría deberse a que estas muestras por diferentes razones como puede ser el estadio del tumor, tienen más correlación con su muestra normal. Sería necesario analizar más a fondo las posibles irregularidades en el conjunto de datos para identificar a qué se deben y tomar las medidas correctivas necesarias.

Otro punto que se ha corroborado en este proyecto es la importancia de reducir la dimensionalidad del conjunto, ya que la mayoría de los microRNAs no aportan información útil que permita diferenciar entre los grupos. Para ello se ha usado el paquete edgeR del que se puede decir que ha mostrado ser muy

potente, contar con una buena documentación y ofrecer capacidades de análisis muy potentes con una gran facilidad de uso para usuarios sin conocimientos avanzados de este tipo de métodos estadísticos. Además, tal como se ha comentado anteriormente se ha detectado un gran número de coincidencias con microRNAs reportados en los artículos consultados.

Por otro lado, coincidiendo con lo que ya se conocía, se puede decir que los resultados del clustering jerárquico dependen en gran medida del método de enlace que se utiliza. En el análisis se ha podido comprobar como distintos métodos (*complete* y *ward.D2*) han producido resultados dispares. Sin embargo, el uso de esta herramienta está muy extendido porque permite explorar posibles asociaciones en los datos y a partir de la interpretación derivar ciertas hipótesis como punto de partida para otros análisis.

Por último, con los resultados obtenidos en el análisis no se puede decir que sea posible diagnosticar únicamente con microRNAs con una precisión aceptable, ya que únicamente hemos utilizado algoritmos de clustering y, además, con un conjunto reducido de muestras. Sin embargo, sí ha quedado patente que los microRNAs muestran patrones de expresión diferenciados y permiten distinguir en la mayoría de los casos las muestras normales de las tumorales. Sí se puede intuir que, tal como indica la literatura, pueden ser de gran utilidad en la monitorización de los pacientes y la predicción de la respuesta a tratamientos y la recaída, sobre todo si se combinan con otros marcadores como PSA y Gleason Score.

8. Conclusiones y futuros proyectos

En este proyecto se han usado datos de expresión de microRNAs en muestras de cáncer de próstata pertenecientes a TCGA obtenidos con tecnología RNA-Seq. Se ha llevado a cabo un análisis de expresión diferencial mediante edgeR para detectar genes diferencialmente expresados entre las muestras de los dos grupos (normal y tumor). Se ha detectado un número relevante de microRNAs cuya expresión se muestra alterada en un grupo respecto al otro. Muchos de ellos han sido reportados en artículos consultados, y algunos han sido reportados en varios artículos, confirmando por lo tanto su importancia. Posteriormente se ha aplicado clustering jerárquico para detectar asociaciones entre los microRNAs seleccionados.

Una vez finalizado el proyecto se puede decir que la complejidad detrás del descubrimiento de bio marcadores se ha ido evidenciando a medida que avanzaba el desarrollo del mismo, ya que se puede profundizar mucho en cada una de las áreas implicadas. Por un lado, la biología molecular es un campo muy desconocido para los informáticos en general por lo tanto, hay que empezar desde la base y absorber mucho conocimiento. Por otro lado, el cáncer es una enfermedad muy compleja y cada tipo tiene sus propias características y peculiaridades. Por último, el machine learning y la estadística, aunque más familiares para los informáticos, son muy amplios y hay mucha investigación al respecto.

Por lo que respecta a los objetivos establecidos, se puede decir que en líneas generales se han cumplido. Se ha obtenido un conocimiento básico de los microRNAs, de sus funciones como reguladores de la expresión genética y el motivo por el cual son importantes como marcadores genéticos para el cáncer. Así mismo se han conocido mejor las características de una enfermedad compleja como es el PCa.

Por otro lado, hubiera sido muy interesante poder señalar algún microRNA concreto que sirva como marcador en el PCa, pero para esto es necesario otro tipo de análisis posterior que queda fuera del alcance de este proyecto, y por lo tanto se deja abierto para el futuro. Sin embargo, todos los microRNAs diferencialmente expresados que se han detectado tienen un cierto potencial por el hecho de mostrar un patrón claramente diferenciado entre los dos grupos de muestras.

En cuanto a la planificación se puede decir que en líneas generales se ha seguido lo previsto inicialmente. La parte que ha sufrido un poco de desviación ha sido la redacción de las partes de la memoria

correspondientes a la última entrega. Respecto a la metodología seguida en el desarrollo del proyecto, aunque no de una manera formal, se ha basado en la metodología Scrum para dividir las tareas generales en sub tareas más fáciles de acometer y así poder evaluar mejor el progreso. Seguir esta metodología ha sido de gran ayuda para visualizar el avance del proyecto y determinar mejor qué debe hacerse en cada momento.

El trabajo desarrollado en este proyecto es solamente una primera aproximación al análisis con datos de expresión genética. A partir de aquí se pueden abrir varios frentes para continuar adelante, o incluso se puede trabajar en la fase previa de procesamiento para obtener los datos de expresión (secuenciación, alineamiento, etc.).

Para continuar hacia adelante, como primer paso sería interesante analizar las causas por las que algunas muestras del conjunto analizado presentan niveles de expresión anormales.

Una vez aclarado este asunto, se podría llevar a cabo un análisis para descubrir correlaciones entre los niveles de expresión de los microRNAs y las variables clínicas con el objetivo de descubrir aquellos microRNAs que tienen un rol importante en el desarrollo o progresión de la enfermedad.

Una vez reducido el conjunto de microRNAs se pueden predecir sus genes diana y descubrir en que procesos biológicos intervienen estos microRNAs de forma directa o indirecta. El objetivo de este tipo de análisis es obtener un conocimiento del rol que pueden tener en el PCa.

Por último y concluyendo, me gustaría destacar que la investigación del cáncer es un campo muy vivo. Solamente en las dos últimas semanas se ha podido ver en diversos medios no especializados entre 3 y 4 noticias sobre nuevos avances al respecto. Por lo tanto, creo que hay razones para ser positivos de cara al futuro.

9. Glosario

PCa: Siglas en inglés de Cáncer de Próstata.

GS: Gleason Score o Escala de Gleason.

RNA: Siglas en inglés de Ácido Ribonucleico (ARN en español).

ncRNA: Siglas en inglés de ARN no codificante. Tipo de ARN que no contiene información para codificarse en proteínas.

MicroRNA: Cadena corta de 21 o 22 nucleótidos de media que se genera a partir de precursores codificados en el genoma.

Nucleótidos: Son las sub unidades que componen una cadena de ADN o ARN.

RNA-Seq: RNA-Sequencing es un método de secuenciación de nueva generación (NGS) para cuantificar RNA en muestras biológicas.

TCGA: The Cancer Genome Atlas. Proyecto colaborativo para catalogar mutaciones genéticas relacionadas con el cáncer.

DEG: Siglas en inglés de Genes diferencialmente expresados (Differential Expressed Genes).

DEA: Siglas en inglés de Análisis de expresión diferencial (Differential Expression Analysis).

FDR: False Discovery Rate.

log-FC: Log-Fold change.

CRAN: Comprehensive R Archive Network. Red para descargar versiones de paquetes de R.

10. Referencias

[1] Marcilio CP. de Souto, Ivan G. Costa, Daniel SA. de Araujo, Teresa B. Ludermir and Alexander Schliep: Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics* 2008, 9:497.

[2] Gloria Bertoli, Claudia Cava and Isabella Castiglioni (2016): MicroRNAs as Biomarkers for Diagnosis, Prognosis and Theranostics in Prostate Cancer. *International Journal of Molecular Science*, 22;17(3):421. [PubMed]

[3] Dong-Fu Liu, Ji-Tao Wu, Jian-Ming Wang, Qing-Zuo Liu, Zhen-Li Gao, Yun-Xiang Liu (2012): MicroRNA expression profile analysis reveals diagnostic biomarker for human prostate cancer. *Asian Pacific J Cancer Prev*, 13, 3313-3317.

[4] Vanessa M. Kvam, Peng Liu and Yaqing Si: A comparison of statistical methods for detecting differentially expressed genes from RNA-Seq data. *American Journal of Botany* 99(2): 248-256. 2012.

[5] Robinson, MD, McCarthy, DJ, Smyth, GK (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* (Oxford, England) 26: 139–140.

[6] Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, ET AL. 2004. Bioconductor: Open software de- velopment for computational biology and bioinformatics. *Genome Biology* 5: R80.

[7] R Core Team (2016). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. https://www.R-project.org/.

[8] Wan Y-W, Allen GI, Liu Z: TCGA2STAT: simple TCGA data access for integrated statistical analysis in R. *Bioinformatics* 2016, 32(6):952-954. https://CRAN.R-project.org/package=TCGA2STAT

[9] Daniela M. Witten and Robert Tibshirani (2013). sparcl: Perform sparse hierarchical clustering and sparse k-means clustering. R package version 1.0.3. https://CRAN.R-project.org/package=sparcl

[10] Friedrich Leisch. A Toolbox for K-Centroids Cluster Analysis. *Computational Statistics and Data Analysis*, 51 (2), 526-544, 2006.

[11] Malika Charrad, Nadia Ghazzali, Veronique Boiteau, Azam Niknafs (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. Journal of Statistical Software, 61(6), 1-36. http://www.jstatsoft.org/v61/i06/

[12] Sean Rudy: edgeR Tutorial: Differential Expression in RNA-Seq Data. September 26, 2011.

[13] https://cran.r-project.org, Noviembre 2016.

[14] https://www.r-bloggers.com, Noviembre 2016.

[15] Robert I. Kabacoff: R in Action, 2nd Edition. *Manning Publications*, NY, 2015.

[16] http://www.sthda.com/english/wiki/hierarchical-clustering-essentials-unsupervised-machine-learning, Noviembre 2016.

[17] http://davetang.org/muse/2013/11/12/analysing-mirna/, Octubre 2016.

[18] Samy M. Mekhail, Peter G. Yousef, Stephen W. Jackinsky, Maria Pasic, George M. Yousef. miRNA in prostate cancer: new prospects for old challenges. *EJIFCC* 2014, 25(1): 79-98. [PubMed]

[19] Benjamin L. Jackson, Anna Grabowska and Hari L. Ratan: MicroRNA in prostate cancer: functional importance and potential as circulating biomarkers. *BMC Cancer* 2014, 14:930. [PubMed]

[20] Patrik D'haeseleer (2005): How does gene expression clustering work? *Nature Biotechnology* 23, 1499-1501.

[21] Raúl Benítez, Gerard Escudero, Samir Kanaan. Inteligencia Artificial Avanzada. UOC.

[22] https://en.wikipedia.org, Diciembre 2016.

[23] https://cancergenome.nih.gov/cancersselected/prostatecancer, Diciembre 2016.

[24] http://www.cancerresearchuk.org/, Diciembre 2016.

[25] http://www.cancer.net/, Diciembre 2016.

[26] https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R, Diciembre 2016.

[27] Guo Y, Sheng Q, Li J, Ye F, Samuels DC, et al. (2013) Large Scale Comparison of Gene Expression Levels by Microarrays and RNAseq Using TCGA Data. PLoS ONE 8(8): e71462. doi:10.1371/journal.pone.0071462.

[28] https://gist.github.com/jdblischak/11384914, Diciembre 2016.

[29] Annika Fendler, Phd Thesis (2011). MiRNAs and prostate cancer: Identification, functional characterization and their potential use in medical practice. Fachbereich Biologie, Chemie, Pharmazie der Freien Universität Berlin.

[30] http://www.exiqon.com/what-are-microRNAs, Diciembre 2016.

[31] Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59, 433-460.

[32] Tal Galili (2015). dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. Bioinformatics. DOI: 10.1093/bioinformatics/btv428

11. Annexos

11.1. Configuración del entorno

11.1.1. Instalación de R y RStudio

Para instalar R solo se ha de descargar R del siguiente enlace, ejecutar el instalador y seguir las indicaciones:

https://cran.r-project.org/

Lo mismo en el caso de R-Studio:

https://www.rstudio.com/products/rstudio/download/

*Se requiere acceso a internet, tanto para la descarga de R y RStudio, como para la instalación de los diferentes paquetes.

A continuación se muestra la información de la plataforma utilizada para el desarrollo de este proyecto:

```
> sessionInfo()
R version 3.3.1 (2016-06-21)
Platform: x86_64-apple-darwin13.4.0 (64-bit)
Running under: OS X 10.11.6 (El Capitan)
```

11.1.2. Instalación de paquetes

EdgeR

Tal como se indica en el tutorial de edgeR, para su instalación se han de ejecutar las siguientes instrucciones en R:

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

Automáticamente se instalarán otras dependencias. Si aparece un mensaje preguntando si se quieren actualizar los paquetes antiguos se puede seleccionar cualquiera de las opciones (a/s/n). En el caso de este proyecto se ha optado por actualizar todos los paquetes (opción a).

La documentación de edgeR se puede encontrar en Bioconductor en el apartado documentación:

http://bioconductor.org/packages/release/bioc/html/edgeR.html

TCGA2STAT

En el siguiente enlace se puede encontrar información acerca de la instalación:

http://www.liuzlab.org/TCGA2STAT/#install-from-cran

Primero es necesario descargar el fichero comprimido del repositorio:

https://CRAN.R-project.org/package=TCGA2STAT

Después en la consola de R ejecutar el siguiente comando:

```
install.packages("TCGA2STAT_1.0.tar.gz", repos = NULL, type = "source")
```

flexclust y sparcl

Estos paquetes se pueden instalar de la forma habitual en R.

```
install.packages("flexclust", "sparcl")
```

11.2. Código

```
# install edgeR if it's not already installed
if( !require(edgeR) ) {
        source( "http://www.bioconductor.org/biocLite.R" )
        biocLite("edgeR")
}

# function to load data from tcga repository
getTCGAData <- function() {
        library( "TCGA2STAT" )
        # by default raw counts
        prad <- getTCGA( disease="PRAD", data.type="miRNASeq" )
        prad.tumNormal <- TumorNormalMatch( prad$dat )
        data <- merge( prad.tumNormal$primary.tumor,
                prad.tumNormal$normal, by="row.names", all=T )
```

```
        row.names( data ) <- data$Row.names
        data <- data[ , -1 ]
        prad <- NULL
        return( data )
}

# copy this function in the R console and hit enter before using it
getDGEList <- function( matrix, groups ) {
        library( "edgeR" )
        # miRNAs as rows, samples as cols
        dge_list <- DGEList( matrix, group = groups )
        # filter out low counts
        dge_list <- dge_list[ rowSums(cpm(dge_list) > 1) >= 10, ]
        dge_list <- calcNormFactors( dge_list )
        dge_list <- estimateCommonDisp( dge_list )
        dge_list <- estimateTagwiseDisp( dge_list )
        dge_list <- equalizeLibSizes( dge_list )
        return( dge_list )
}

getwd() # shows the current directory

# define factors for the groups (tumor and normal)
groups <- factor( c(rep('T', 52), rep('N', 52)) )

# PLACE the .rds file in the current directory before loading
data <- readRDS( "tcga_raw_tumNorm.rds" )
# it can be also downloaded from tcga
```

```
#data <- getTCGAData()

# get the edgeR object with all the calculations
dge_list <- getDGEList( data, groups )

png( "mds_countData.png" )
plotMDS( dge_list, main = "MDS Plot for Count Data",
        col = as.numeric(groups), labels = as.factor(groups) )
dev.off()

fdr <- 0.01 # set the FDR threshold
res <- exactTest( dge_list, pair = levels(groups) )
deg_tags <- topTags( res, n = nrow(getCounts(dge_list)) )$table
de <- rownames( deg_tags[ deg_tags$FDR < fdr, ] )

# extract upreg and downreg microRNAs
upreg <- rownames( deg_tags[ deg_tags$FDR < fdr & deg_tags$logFC > 1, ] )
downreg <- rownames( deg_tags[ deg_tags$FDR < fdr & deg_tags$logFC < -1, ] )
# filter out irrelevant microRNAs
normCounts <- dge_list$pseudo.counts[ c(upreg, downreg), ]
normCounts <- scale(t( normCounts )) # microRNAs as columns, samples in rows

# Clustering
if( !require("flexclust") ) {
        install.packages("flexclust")
}
if( !require("sparcl") ) {
        install.packages("sparcl")
```

```
}

# Kmeans clustering
km <- kmeans( normCounts, 2, nstart = 20 )
table <- table( groups, km$cluster )
table
library( "flexclust" )
randIndex( table )

# compute the distance matrix between samples
d <- dist( normCounts, method = "euclidean" )

library( "sparcl" )

# hierarchical clustering of samples with ward.D2 method
hc_ward <- hclust( d, method="ward.D2" )
png( "hclust_ward_tumNorm.png" )
ColorDendrogram( hc_ward, y = as.numeric(groups), branchlength = 150000,
        main = "Hierarchical clustering between samples" )
legend( "topright", c("Tumor", "Normal"), col = c("red", "green"), pch = 19 )
dev.off()

# hierarchical clustering of samples with complete method
hc_comp <- hclust( d, method="complete" ) png( "hclust_complete_tumNorm.png" )
ColorDendrogram( hc_comp, y = as.numeric(groups), branchlength = 150000,
        main = "Hierarchical clustering between samples" )
legend( "topright", c("Tumor", "Normal"), col = c("red", "green"), pch = 19 )
dev.off()
```

```
# hierarchical clustering microRNAs
genes <- c( rep(1, length(upreg)), rep(2, length(downreg)) )

# compute distance matrix between microRNAs
d <- dist( t(normCounts), method = "euclidean" )
hc_ward <- hclust( d, method="ward.D2" )

png( "mirnas_ward.png" )
plot( as.dendrogram(hc_ward),
        main = "Hierarchical clustering of microRNAs (ward.D2)" )
rect.hclust( hc_ward, k = 2, border = 2:3 )
legend( "topright", c("Upreg", "Downreg"), col = c("red", "green"), pch=15 )
dev.off()
```