Metodología para el análisis e identificación de genes relacionados con cáncer de mama utilizando Machine Learning y datos de secuenciación masiva.

Jose Liñares Blanco

Máster Universitario de Bioestadística y Bioinformática

TFM “Ad-hoc”

Consultor: Melchor Sánchez Martínez

Tutor externo: Carlos Fernandez Lozano

Última actualización el 24 de Mayo de 2017

Título del trabajo:	Metodología para el análisis e identificación de genes relacionados con cáncer de mama utilizando Machine Learning y datos de secuenciación masiva
Nombre del autor:	*Jose Liñares Blanco*
Nombre del consultor/a:	*Melchor Sánchez Martínez*
Nombre del PRA:	*María Jesus Marco Galindo*
Fecha de entrega (mm/aaaa):	05/2017
Titulación::	*Máster Universitario en Bioinformática y Bioestadística*
Área del Trabajo Final:	*TFM "Ad-hoc"*
Idioma del trabajo:	*Castellano*
Palabras clave	*Machine Learning; Cáncer de mama; RNAseq*

Resumen del Trabajo:

Con el auge de las técnicas de secuenciación masiva y con la reducción tan drástica de costes que ha supuesto, cada vez existe una mayor cantidad de datos de acceso abierto para el estudio y caracterización de cada tipo de cáncer. Por lo tanto es necesaria la implementación de nuevos protocolos bioinformáticos para el análisis de estas bases de datos.

Este trabajo recoge un estudio sobre el análisis del perfil de expresión genético (RNAseq) de un total de 721 pacientes que refieren algún tipo de cáncer de mama. Además estos pacientes han sido clasificados en cuanto al status del receptor de estrógenos, dividiéndolos en 555 pacientes ER positivos y 166 negativos.

Han sido dos, los métodos utilizados. En primer lugar, a través de un protocolo estándar basado en el proyecto Bioconductor se ha realizado un análisis diferencial de la expresión genética en los pacientes según su grupo. En segundo lugar, dicha base de datos ha sido sometida a técnicas de clasificación basadas en Machine Learning (ML). Los resultados de ambas técnicas fueron comparados.

Finalmente el máximo valor, AUC=0.963, fue obtenido mediante las técnicas basadas en ML. El listado de genes relacionados con cáncer de mama difiere según la técnica utilizada, propiciando un futuro estudio de los genes obtenidos mediante técnicas de ML.

Abstract:

Thanks to advances in NGS sequencing techniques and with the drastic reduction of costs that has taken place, there is an increasing amount of open access data for the study and characterization of each type of cancer. It is therefore necessary to implement new bioinformatic protocols for the analysis of these databases.

This work includes a study on the analysis of the genetic expression profile (RNAseq) of a total of 721 patients who report some type of breast cancer. In addition, these patients have been classified by the estrogen receptor status, thus obtaining 555 ER positive and 166 negative patients.

Two methods have been used. Firstly, through a standard protocol based on the Bioconductor project, a differential analysis was performed on the genetic expression of patients according to their group. Secondly, RNAseq database has been subjected to classification techniques based on Machine Learning.

Finally the best performance value, AUC=0.963, has been obtained throgth ML approach. The list of genes related to breast cancer differs according to the technique used, favoring a future study of the genes obtained by ML techniques.

Contenido

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

1.1. Contexto y justificación del trabajo

El cáncer es ocasionado por mutaciones genéticas que provocan una desregulación en la expresión de determinados genes. Se considera una enfermedad multifactorial y enormemente heterogénea, lo que conlleva que dentro de un tipo concreto de cáncer los pacientes presenten una diferente sintomatología (Ross *et al.*, 2000).

El hecho por el que aparecen mutaciones en los genes que ocasionan el cáncer se debe a formas heredadas, creándole al individuo una alta probabilidad de riesgo de padecer cierto tipo de cáncer durante su vida, o a exposiciones a ciertos tipos de ambientes (tabaco, radiación, UV, etc.), lo que se denominada influencia epigenética.

Entre todos los genes de la célula, existen algunos que tienen una mayor probabilidad de generar la aparición de cáncer cuando se alteran. Esta alteración es fruto de una desregulación genética, que afecta en mayor medida al crecimiento y a la división celular (Koboldt *et al.*, 2012). A estos genes se les denominan *drivers* y se pueden diferenciar tres clases diferentes de *drivers* genéticos: proto-oncogenes, genes supresores del tumor y genes de reparación de DNA.

- *Proto-oncogenes*: se denominan a los genes que tienen presencia en los procesos celulares de crecimiento y división. Antes de una alteración en estos genes, se conocen con el nombre de proto-oncogen, aunque una vez han sido mutados, se denominan exclusivamente oncogenes. En esta fase, dichos genes están defectuosos o más expresados de lo normal, conduciendo a la célula a una división descontrolada o a una supervivencia celular cuando ésta no debería.
- *Genes supresores del tumor*: este tipo de genes también se relacionan con el crecimiento y división celular. Una alteración de estos genes, provoca una división descontrolada de la célula.

- *Genes de reparación del DNA*: tras las múltiples replicaciones de DNA, y la exposición a ambientes altamente mutagénicos, es vital el trabajo que hacen estos genes. Una vez alterados, aumenta la probabilidad de mutación en los demás genes, y más importante, la no reparación de las mutaciones. Esto conduce a la presencia de mutaciones en genes con más probabilidad de provocar algún tipo de tumor.

Todas estas desregulaciones genéticas en el crecimiento y la división de las células cancerígenas vienen acompañadas por unas características propias de dichas células, que difieren a las normales. Por ejemplo, las células del cáncer tienen una menor especialización en comparación con las células normales. Éstas últimas tienden a diferenciarse a medida que se dividen en varios tipos celulares. Las células del cáncer, al utilizar toda su energía en la división y el crecimiento, no presentan una diferenciación o especialización. Además, las relacionadas con el cáncer son capaces de ignorar las señales moleculares que a las células normales les indican la detección del ciclo de división celular o la muerte celular programada (apoptosis). Otra característica que presentan este tipo de células es que pueden llegar a tener un influencia sobre las células normales que están a su alrededor mediante moléculas generadas por ellas. De esta manera, por ejemplo, las células normales provocan un aporte de oxígeno, lo que les facilita su crecimiento y desarrollo. Por último, otra característica es la posibilidad de esquivar la actividad del sistema inmune

Estadísticas recientes del National Institute of Health (NIH) indican que entre los años 2012 al 2030 se espera que la incidencia de cáncer ascienda en un 50%. De 14 a 21 millones de enfermos cada año. En cuanto a las muertes, se espera un aumento del 60%, lo que provocará un aumento desde 8 a 13 millones de muertes cada año. El cáncer es considerado por lo tanto una de las principales causas de muerte en todo el mundo, con indicios a crecer en un futuro no tan lejano. Similares estadísticas reportadas por el Cancer Research UK en Enero de 2014 presentan análogos resultados. Dichas estadísticas se pueden observar en la **Figura 1.**

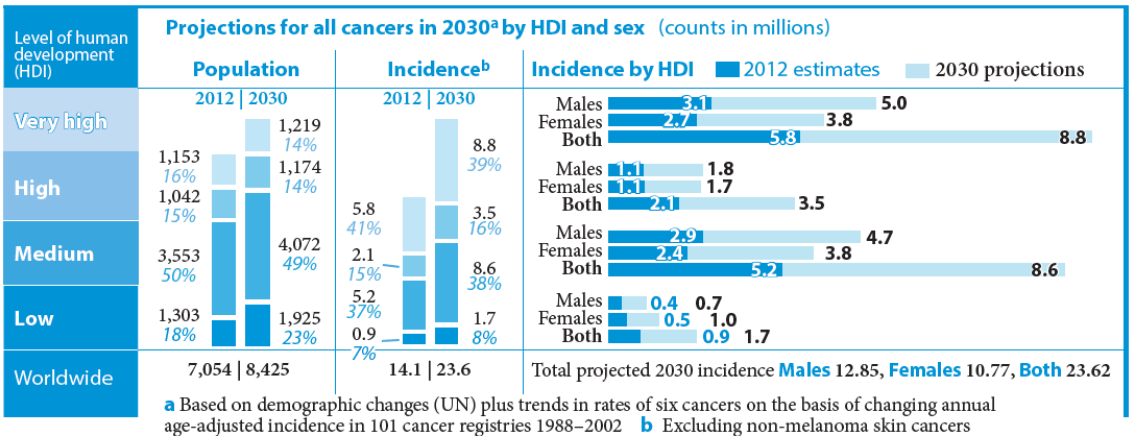


Figura 1: Estadísticas de la incidencia del cáncer en el año 2012 y proyectas al año 2030. Gráfico obtenido del Cancer research UK, publicado en Enero de 2014.

En España, según la página web http://www.infosalus.com, el tipo de tumor más frecuente es el colorrectal, considerando ambos sexos, con 41.441 nuevos casos en 2015, seguido del de próstata, pulmón, mama, vejiga, estómago, linforma no Hodgkin, páncreas, hígado y riñón. Tomando los sexos por separado, en mujeres, el tipo de cáncer más común es el de mama, seguido del colorrectal, útero, pulmón y vejiga. En hombres, el tipo más común es el de próstata, seguido del colorrectal, pulmón, vejiga y estómago.

Actualmente existen diferentes tratamientos del cáncer. Se diferencian los siguientes tipos: cirugía, terapia de radiación, quimioterapia, inmunoterapia, terapia dirigida, terapia hormonal, trasplante de células madre y medicina de precisión. Se observa que cada vez, la alternativa para el tratamiento del cáncer se basa más en terapias genéticas, las cuales presentan unos resultados esperanzadores hasta la fecha. Esto conlleva la necesidad de un mayor entendimiento y comprensión del funcionamiento genético de todos los tipos de cáncer.

Debido a las dificultades del estudio global de todos los tipos de cáncer, y a las características intrínsecas que posee esta enfermedad, es necesario focalizar el estudio en un tipo concreto. Este trabajo se ha centrado, por lo tanto, en el cáncer de mama. Éste, es uno de los más comunes con más de 1.300.000 casos y 450.000 muertes cada año en todo el mundo, datos reportados en el año 2012 (Koboldt *et al.*, 2012). Clínicamente, es una enfermedad

heterogénea, categorizada en tres grupos terapéuticos principales. Aunque a día de hoy se reconocen hasta 10 subgrupos (Curtis *et al.*, 2012). El grupo del receptor de estrógenos positivo (ER) es el más numeroso y diverso. El grupo HER2 (también llamado ERBB2) es un gran éxito clínico ya que tiene como diana terapéutica HER2. Al último grupo se le denomina triple negativo (TNBCs, por la pérdida de expresión de ER, del receptor de progesterona (PR) y HER2), también conocido como cáncer de mama basal. Éste es un grupo únicamente con opciones quimioterapéuticas, y tienen una mayor incidencia en pacientes con mutaciones en la línea germinal del gen BRCA1, o de ancestros africanos.

Con el auge de las técnicas de secuenciación NGS y con la reducción tan drástica de costes que ha supuesto, cada vez existe una mayor cantidad de datos de acceso abierto para el estudio y caracterización de cada tipo de cáncer a un nivel que engloba la mayoría de los genes presentes en el genoma humano. Todos estos datos necesitan del desarrollo de nuevas técnicas o *pipelines* distintos para identificar casos de interés.

El enfoque actualmente planteado para la caracterización genética de la mayoría de cánceres es el estudio de su perfil de expresión, englobando éste la mayor cantidad de genes posibles. Concretamente la aparición de la plataforma RNAseq facilitó el estudio y la comprensión de la transcriptómica celular. La mayoría de los laboratorios de investigación biomédica realizan técnicas de análisis de este tipo de datos de expresión mediante herramientas estadísticas convencionales. Casi todas ellas presentes en el proyecto Bioconductor, lo que propició que los experimentos fuesen altamente reproducibles en cualquier laboratorio.

Paralelamente, otros métodos, computacionalmente más potentes están siendo desarrollados y utilizados en el ámbito de la biomedicina. Entre ellos se encuentran los experimentos basados en Machine Learning (ML). Estos experimentos aportan un punto de vista estadístico más vanguardista y no dependen del contraste de hipótesis, ya que se basan únicamente en el manejo de los datos de estudio. Estas técnicas basadas en ML están obteniendo una

gran aceptación y unos grandes resultados en casi todas las ramas (Tarca *et al.*, 2007).

Es por ello que el presente trabajo se justifica en este contexto, y pretende hacer un estudio y caracterización a partir de datos de expresión genética, RNAseq, en pacientes que presentan cáncer de mama. Dentro de este tipo de cáncer, el estudio se centrará en el grupo terapéutico relacionado al *status del receptor de estrógenos (ER).* El análisis de los datos RNAseq se realizará mediante dos enfoques, uno estándar y otro más vanguardista. En cuanto al enfoque estándar, combinará información biológica (longitud de los genes, tamaño de la biblioteca genómica, etc.) con contraste de hipótesis estadísticas, todo ello sustentado en una plataforma informática presente en el proyecto Bioconductor. En segundo lugar, el enfoque más vanguardista se centrará únicamente en los datos a estudio, sin hipótesis estadísticas ni información biológica adicional. Este enfoque se efectuará mediante técnicas de ML. Ambos resultados serán comparados para percibir si el enfoque basado en ML iguala o incluso mejora al enfoque más convencional o estándar, pudiendo, de esta manera, obtener nuevas dianas genéticas terapéuticas para el futuro estudio y tratamiento del cáncer de mama.

1.2. Objetivos del trabajo

- **Objetivo 1**: Utilización de un protocolo estándar para el análisis de genes diferencialmente expresados mediante el proyecto Bioconductor
    - Comparación de los resultados de diferentes librerías para el análisis diferencial
- **Objetivo 2:** Desarrollo de un nuevo *pipeline* de análisis de datos de expresión utilizando técnicas de Machine Learning.
    - Evaluar la eficacia de diferentes aproximaciones de Feature Selection (FS) existentes para la identificación de genes con datos de RNAseq.
    - Inclusión de una batería de pruebas o benchmarking para comparar en igualdad de condiciones dichas aproximaciones.

- Selección de la mejor técnica incluyendo en el diseño experimental una fase final de comparación de técnicas basada en técnicas de contraste de hipótesis nula.

- **Objetivo 3:** Comparación de los genes DE con los genes obtenidos mediante la aproximación por ML
  - Repaso de dichos genes en las publicaciones del estado del arte.

1.3. Enfoque y métodos a seguir

Se siguió un enfoque iterativo y creciente valorando en cada momento las mejoras necesarias y la aportación de nuevas técnicas, dentro de un marco de desarrollo reutilizable y limitando el trabajo necesario para realizar la inclusión de nuevos métodos.

En la primera parte del proyecto se ha seguido un método de aproximación estándar en el análisis de la expresión genética. Dentro de las diferentes posibilidades que existen para realizar dicho protocolo se han usado las herramientas bioinformáticas más reportadas hasta la fecha para este tipo de análisis.

En la segunda parte, el diseño de los experimentos basados en ML, se realizó mediante un paquete específico del programa R, que engloba todas las herramientas necesarias para la resolución de este tipo de problemas.

1.4. Planificación del trabajo

A continuación se presentarán las tareas que se realizaron durante el proyecto así como una breve descripción de ellas. Posteriormente se presentará una tabla con su planificación temporal.

1. Búsqueda en la literatura científica: Esencial para cualquier trabajo de investigación, su duración abarcó prácticamente todo el proceso para la realización del trabajo.

2. Obtención de los datos: se han recogido de bases de datos públicas, con lo que su duración únicamente se basó en la descarga de internet.
3. Pre-procesado de los datos NGS: una vez obtenido los datos, éstos se analizaron y procesaron mediante herramientas estadísticas.
4. Aproximación convencional utilizando el diseño experimental clásico: esta tarea se realizó mediante el proyecto Bioconductor y tres librerías incluidas en él. Sirvió de precedente para el análisis de datos y su comparación con técnicas de Machine Learning.
5. Aproximación utilizando técnicas avanzadas de Machine Learning: en esta tarea se realizó una aproximación de clasificación de los datos mediante técnicas de Machine Learning.
6. Comparación entre técnicas: una vez se consiguieron los resultados de las dos aproximaciones, se compararon para poder sacar conclusiones de los dos protocolos utilizados.
7. Redacción de la memoria: se redactó la memoria haciendo un amplio recorrido sobre el tema en cuestión y presentando los resultados que se obtuvieron.
8. Preparación de la presentación ante el tribunal: una vez entregada la memoria se debe hacer una defensa de ésta ante el tribunal para ser calificada.

Se consideran estas ocho tareas las principales para la ejecución del proyecto. Aunque se pueden identificar dentro de algunas de las tareas principales, sub-tareas que también han sido de gran importancia.

- Búsqueda en la literatura científica:
    - Citar adecuadamente cada una de las ideas extraídas de la literatura científica
    - Capacidad de comprensión y análisis de la información obtenida de la literatura científica
- Aproximación convencional utilizando el diseño experimental clásico:
    - Uso adecuado de las herramientas presentes en Bioconductor.
    - Recopilación y comparación del listado de genes diferencialmente expresados obtenidos mediante los tres paquetes utilizados
- Aproximación utilizando técnicas avanzadas de Machine Learning:

    - Aprendizaje en el uso del paquete utilizado
    - Ejecución de técnicas de selección de variables
    - Diseño de tres experimentos diferentes basados en ML
    - Evaluación y selección del mejor modelo
- Comparación entre técnicas:
    - Comparación de los listados de genes entre las dos aproximaciones
    - Búsqueda en la literatura sobre los genes interesantes de estudio

Se consideraron hitos del proyecto las siguientes tres tareas:

- Aproximación convencional utilizando el diseño experimental clásico
- Aproximación utilizando técnicas avanzadas de Machine Learning
- Redacción de la memoria

A continuación se presenta en la ***Tabla 1*** la planificación temporal seguida para la ejecución de las distintas tareas. Cada uno de los números corresponde con la tarea indicada en el listado anterior. El proyecto comenzó el día 15 de Marzo de 2017 y se finalizó el día 24 de Mayo de 2017, esta última fecha es el límite de entrega que propuso la universidad.

Tabla 1: Planificación temporal de las tareas realizadas en este proyecto.

Número de la tarea	Duración	Inicio	Final
1	53 días	15/03/2017	24/05/2017
2	1 día	15/03/2017	15/03/2017
3	7 días	16/03/2017	24/03/2017
4	12 días	20/03/2017	04/04/2017
5	20 días	20/03/2017	15/04/2017
6	15 días	17/04/2017	05/05/2017
7	28 días	12/04/2017	20/05/2017
8	4 días	20/05/2017	24/05/2017

1.5. Breve sumario de productos obtenidos

Entregables para la calificación del trabajo:

- Memoria del trabajo en formato .pdf.
- Vídeo en formato .wmv de la presentación de la memoria
- Transparencias de la presentación en formato .ppt

Resultados del estudio:

- Tablas de genes obtenidos mediante las diferentes técnicas que presentan una relación tanto biológica como matemática con el *status* del receptor de estrógenos (se presentan en los distintos ANEXOS)
- Scripts en código R para el desarrollo de experimentos basados en ML
- Scripts en código R para la implementación de un protocolo estándar basado en el proyecto Bioconductor.

1.6. Breve descripción de los otros capítulos de la memoria

El trabajo se ha estructurado en ocho capítulos principales. Una vez descrita la introducción del trabajo, el segundo capítulo se centra en los fundamentos en los que se basa dicho trabajo. Se explica la técnica RNAseq, el proyecto Bioconductor así como el *workflow* utilizado para finalmente exponer los resultados. En el siguiente capítulo se hace una revisión del estado de la cuestión. Posteriormente, en el cuarto capítulo se entra en detalle en el tema principal de este proyecto. Se explica en que se basa ML, las herramientas y la metodología utilizada. El quinto capítulo se divide en los tres experimentos basados en ML que se efectuaron en este trabajo, exponiendo también sus resultados. Los últimos tres capítulos corresponden a las conclusiones extraídas, los futuros desarrollos y la bibliografía, respectivamente.

2. FUNDAMENTOS DEL TRABAJO

La variedad de datos que se están generando actualmente mediante las técnicas de secuenciación de nueva generación, necesitan de la implementación y comprobación de nuevas técnicas para el análisis de dichos datos. El objetivo de esta primera parte del proyecto trata la implementación de un protocolo estándar de análisis de datos NGS. *Grosso modo,* se pretende realizar un análisis diferencial de expresión de genes procedentes de un estudio basado en la técnica de RNAseq. El análisis se realizará mediante métodos convencionales estadísticos, utilizados en la mayoría de investigaciones de este tipo. Los resultados obtenidos tras la implementación del *pipeline* estándar será una motivación para posteriormente añadir al análisis técnicas avanzadas en ML.

Se describirá la aparición de la técnica RNAseq así como su funcionamiento. A continuación se presentarán los datos descargados del repositorio TCGA y finalmente se detallará paso a paso la metodología utilizada en esta aproximación.

2.1. Background de la técnica RNAseq

Eventos celulares tales como la replicación, la diferenciación, la división celular y otros caracteres macroscópicos como rasgos fenotípicos, morfológicos y funcionales son productos de la expresión diferencial de los genes. Especialmente, la aparición de los diferentes tipos de cáncer, son debidos a la desregulación de los mecanismos genéticos, favoreciendo una sobre o infra-expresión de determinados genes, provocando un crecimiento celular descontrolado.

El estudio de la expresión genética es fundamental para caracterizar los diferentes tipos de patologías, así como distintos estadíos de una misma enfermedad. Estos tipos de estudios se centran en la cuantificación y análisis de las moléculas de mRNA, las cuales se generan a partir de las secuencias codificantes de ADN mediante el proceso de transcripción. Las moléculas de

mRNA son las que presentan la información genética necesaria para luego producir las proteínas. Entonces, si se puede contabilizar de alguna forma el número de mRNA producidos por cada gen, se podrá generar una idea de la expresión de dicho gen ante una determinada circunstancia (patológica u ambiental). A mayor número de moléculas de mRNA producidas por un gen concreto, se considera una mayor expresión de dicho gen. Todo el conjunto de moléculas de RNA producidas a partir del proceso de transcripción es lo que se conoce como transcriptómica.

Los *microarrays* de expresión fue la técnica más utilizada en todos los laboratorios biológicos para el análisis del transcriptoma, aunque en ella existan algunas limitaciones. Éstas incluyen una posible hibridación o *cross-hibridación* de artefactos, problemas de detección basados en colorantes y restricciones de diseño que impiden o limitan seriamente la detección de patrones de *splicing* de ARN y genes previamente no mapeados (Wang, Gerstein and Snyder, 2009) (Guo *et al.*, 2013).

Aunque los *microarrays* fueron los más usados y desarrollados, existen otras técnicas para el análisis de expresión genética. Entre ellos se encuentran técnicas como el análisis de serie de la expresión genética (SAGE, por sus siglas en inglés) y métodos relacionados como *Massively Parallel Signature Sequencing (MPSS).* Una ventaja de los métodos de SAGE y relacionados a SAGE es que producen conteos digitales de la abundancia del tránscripto, en comparación con las técnicas basadas en colorantes de *arrays* (Mortazavi *et al.*, 2008).

Expressed sequence tag (EST) es otra técnica que ha sido el método central para el descubrimiento de tránscriptos de referencia. Su base reside en la secuenciación de cDNAs clonados, aunque posee restricciones cuantitativas y cualitativas impuestas en parte por las limitaciones de secuenciación que existieron hasta la fecha y su coste, así como las dificultades intrínsecas que aparecen al realizar clonación con bacterias (Mortazavi *et al.*, 2008), (Nagaraj, Gasser and Ranganathan, 2007).

Dense whole-genome tiling microarrays fueron un conjunto de técnicas desarrolladas en la primera década de los 2000 y aplicadas a los

transcriptomas como medida de expresión y descubrimiento de nuevos transcriptos. En contraste con los *arrays* de expresión, éstos pueden descubrir nuevos genes y exones, pero requieren una gran cantidad de RNA de entrada y tienen otras limitaciones que afectan a la sensibilidad, especificidad y en la detección directa de sitios de *splicing* (Gregory, Yazaki and Ecker, 2008), (Mockler and Ecker, 2005).

Gracias a los avances en las técnicas de secuenciación del ADN, a través de tecnologías de nueva generación (NGS), se ha revolucionado el campo de la genómica creando un amplio número de posibilidades para el estudio de la transcriptómica. Estas técnicas han permitido no solo generar información con altos rendimientos y a bajo costo, sino también abrir nuevos horizontes para el entendimiento detallado y global de procesos de expresión genética.

Fue a partir del año 2008, cuando se introdujo de manera formal la técnica RNAseq. Ésta técnica sería la base de los estudios de expresión genética. El artículo (Mortazavi *et al.*, 2008) introduce el concepto de RPKM, como una medida de expresión genética, donde RPKM se refiere al número de lecturas por millón de kilobases del transcripto obtenido en un gen que se está expresando. En comparación a la técnica de los microarrays, RNAseq tiene un mayor rango de dinamismo, sensibilidad y detección, englobando toda la secuencia genética (Levin *et al.*, 2010).

Es por ello que para este proyecto se ha decidido usar como base del experimento datos obtenidos a partir de la técnica RNAseq. A continuación se hará una explicación detallada sobre dicha técnica.

2.2. Modo de funcionamiento de la técnica RNAseq

RNAseq está fundamentada en el análisis del transcriptoma, donde un conjunto de mRNAs obtenidos de una muestra biológica son secuenciados mediante técnicas de secuenciación masiva, para revelar la presencia y cuantificar las seceuncas mRNA en un momento dado.

El proceso comienza con la extracción de las moléculas de mRNA a partir de las células. Se debe considerar que debido a la variabilidad biológica, los niveles de mRNA extraídos pueden diferir considerablemente entre diferentes muestras biológicas. Para el aislamiento de RNA se emplean *kits* de extracción de mRNA que aplican la captura a partir de la cola de poly(A). Estas moléculas de mRNA son fragmentadas en una longitud media de 200 nucleótidos. Una vez fragmentadas se convierten en cDNA.

El cDNA obtenido es transformado en una librería genética mediante alguna plataforma de secuenciación basada en NGS. Existen varias posibilidades a la hora de escoger la plataforma. Entre ellas se encuentran Roche/454, Solexa/Ilumina, SOLid/Life Technologies y Helicos/BioSciences. Sin embargo, dentro de las tecnologías de NGS disponibles las más usadas y conocidas son Roche/454 y Solexa/Ilumina. En este trabajo se presentará la técnica de secuenciación empleada por Solexa/Ilumina, por dos motivos principales. En primer lugar fue la técnica reportada en el artículo de (Mortazavi *et al.*, 2008) por lo que se considera la técnica pionera y original. En segundo lugar, los datos RNAseq que se obtuvieron para este trabajo fueron a partir de dicha plataforma de secuenciación.

Solexa/Ilumina se basa en el principio de amplificación en puente y el uso de marcaje por fluorescencia de nucleótidos modificados como terminadores reversibles (Metzker, 2010). El proceso se representa en la ***Figura 2*.**

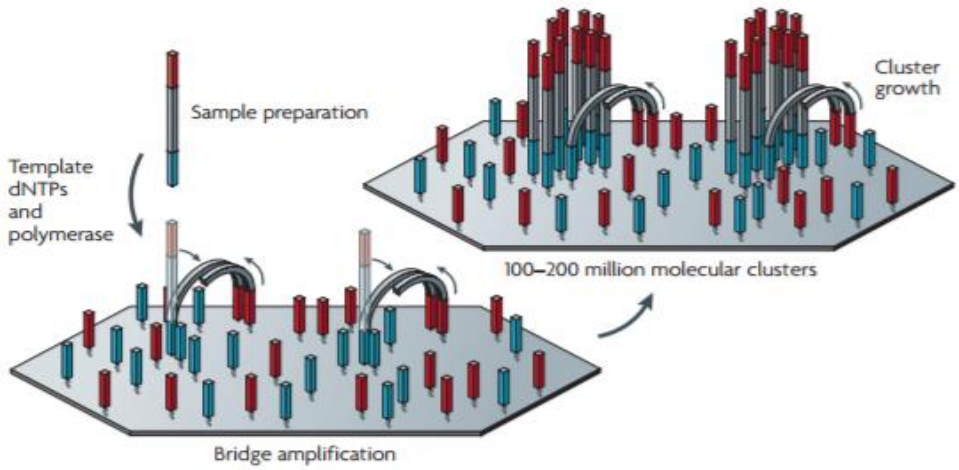


Figura 2: Técnica de secuenciación utilizada por la plataforma Solexa/Ilumina (Metzker, 2010)***.***

Las secuencias de cDNA presentan en sus extremos un adapator que se complementará específicamente a los oligonucleótidos adheridos a la placa sólida, creando una especie de puente. Estos oligonucleótidos funcionarán como cebadores de las cadenas (en sentido original o anti-sentido, dependiendo de cuál sea el extremo usado como cebador).

Una vez replicados los fragmentos, la cadena resultante, denominada amplicón permanecerá unida hasta un proceso de desnaturalización. Éstos serán usados posteriormente para formar más amplicones, tras la formación de un nuevo puente. De esta manera, se formarán *clusters* de amplicones, unos que presentan la secuencia *forward* y otros la *reverse*.

Cuando los fragmentos de cDNA ya están amplificados se realiza su secuenciación mediante nucleótidos terminadores de cadena. El proceso comienza con la adición de los cuatro nucleótidos (A, C, T, G) marcados fluorescentemente a la muestra de amplicones (Ju *et al.*, 2006). Cada vez que un nucleótido marcado se introduzca a la secuencia naciente por la polimerasa, ésta se detendrá. La secuenciación se produce por la incidencia de un láser a los nucleótidos terminadores. La emisión de luz será diferencial de acuerdo con el nucleótido incorporado (Guo *et al.*, 2008).

En concreto, los secuenciadores Ilumina contienen ocho carriles o *lanes*. Cada uno de ellos en la actualidad tiene un rendimiento de 150 millones de lecturas o *reads*. La longitud de las lecturas generadas es pequeña, del rango de 50-100 nucleótidos. Cada *lane* tiene capacidad para 24 librerías de cDNA (Garber *et al.*, 2011). Las lecturas pueden alinearse al genoma de referencia o generar un ensamblado *de novo* y, por lo tanto, hacer un conteo del número de lecturas que se alinean en un gen particular. Algunos investigadores consideran que el conteo de los datos es el productor final para un experimento RNAseq. Sin embargo, se cree necesario someter dichos datos a algún tipo de normalización (se comentará este aspecto en el apartado **2.4.2.2**).

2.3. Obtención de los datos RNAseq para el estudio de cáncer de mama

Los datos RNAseq sobre los que se basó este trabajo se descargaron del repositorio **The Cancer Genome Atlas Project**, TCGA (www.cancergenome.nih.gov). TCGA es un proyecto masivo, comprehensivo y colaborativo para catalogar los datos genómicos del mayor número posibles de cáncer. La colaboración es entre el National Cancer Institute (NCI) y el National Human Genome Research Institute (NHGRI) y 27 institutos y centros del National Institute of Health (NIH). Se pretende obtener una información genómica a partir de diferentes perspectivas, como pueden ser modelos de expresión, detección de *drivers* genéticos o *copy number aberrations.* Gracias a este proyecto se han generado mapas multidimensionales de los cambios a nivel genómico en 33 tipos de cáncer. El conjunto de datos del TCGA comprende más de 2 petabytes de datos genómicos, que describen tejidos tumorales y tejidos normales de más de 11.000 pacientes. Una de las restricciones más comunes en los estudios relacionados al cáncer es el tamaño de las muestra a analizar. La mayoría de laboratorios, debido a los grandes costos que supone, no son capaces de obtener un número tan grande de muestras para sus estudios, lo que elevaría sustancialmente la calidad del trabajo. Los datos obtenidos por el proyecto TCGA están disponibles públicamente, por lo que ayuda enormemente a la comunidad científica a la hora de realizar estudios masivos de los determinados tipos de cáncer.

El TCGA no sólo genera los datos, sino que también los analiza, siendo pionero en el estudio de varios tipos de cáncer. En concreto, para el cáncer de mama, destaca la caracterización genética y molecular de los principales subtipos, el descubrimiento de la relación existente entre el subtipo basal (triple negativo) y el cáncer de ovario y el desarrollo de tratamientos comunes y eficientes sobre ambos tipos de cáncer (Perou *et al.*, 2000), (Ciriello *et al.*, 2015).

Entre todos los enfoques del estudio del cáncer, los datos RNAseq es la principal forma de estudio que sostiene el TCGA, presentando una gran colección de este tipo de datos de expresión. El enlace de descarga de los

datos para este trabajo fue http://firebrowse.org/?cohort=BRCA&download_dialog=true, versión 2016_01_28 de BRCA.

La base de datos consta de 20532 filas que corresponden a los *loci* mapeados mediante la técnica RNAseq, representados en notación EntrezID. Por otra parte, la presenta 757 columnas, correspondientes a pacientes que refieren cáncer de mama. En la ***Figura 3*** se muestran las primeras filas y columnas de la matriz de datos descargada del TGCA. Cada uno de los valores se corresponde con el conteo de las lecturas en bruto sin normalizar.

También se dispone de una base de datos con información fenotípica de los pacientes. Compuesta por 757 filas (pacientes) y 85 columnas. Cada una de estas columnas corresponde a una variable fenotípica diferente. La variable que clasifica los pacientes según su estatus del receptor de estrógenos en positivos o negativos, será la referencia para el desarrollo del trabajo. Se divide en: 166 pacientes 'Negative', 555 'Positive', 2 'Indeterminate', 29 'Not Performed', 5 'Performed but Not Available'. Los pacientes negativos no presentan una sobreexpresión del receptor de estrógenos, mientras que los positivos si que presentan una sobreexpresión. Los pacientes con estatus no disponible serán eliminados.

	TCGA-A1-A0SB	TCGA-A1-A0SD	TCGA-A1-A0SE	TCGA-A1-A0SF	TCGA-A1-A0SG	TCGA-A1-A0SH	TCGA-A1-A0SI	TCGA-A1-A0SJ	TCGA-A1-A0SK	TCGA-A1-A0SM	TCGA-A1-A0SN	TCGA-A1-A0SO	TCGA-A1-A0SP	TCGA-A1-A0SQ	TCGA-A2-A04N	TCGA-A2-A04P	TCGA-A2-A04Q	TCGA-A2-A04R	TCGA-A2-A04T	TCGA-A2-A04U	TCGA-A2-A04V	TCGA-A2-A04W
100130426	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	2	0	2	0	0	0
100133144	115	134	119	67	69	61	57	86	458	110	97	160	129	36	103	81	103	73	233	64	139	62
100134869	59	73	72	50	47	55	30	40	289	124	38	189	71	6	71	29	38	86	170	8	53	23
10357	269	175	392	225	182	321	218	420	6193	375	220	2063	666	72	246	544	599	404	991	1064	443	191
10431	1921	3676	5835	3108	3186	4656	2960	3584	6853	5269	2392	4399	3787	2329	3461	12723	8435	4846	4324	15721	4749	5816
136542	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
155060	1007	453	655	1038	969	732	640	505	1369	686	483	1422	1020	330	812	1462	2257	1114	1383	402	818	628
26823	9	18	13	1	8	7	5	7	17	13	5	20	22	8	12	13	16	14	15	23	15	9
280660	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
317712	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	3	0	0	0	0	0	0
340602	5	4	1	0	0	2	1	0	15	0	0	17	1	0	0	0	3	0	0	0	1	2
388795	8	6	8	9	15	8	7	13	5	7	3	20	14	5	10	20	10	1	5	42	10	58
390284	1	2	12	2	2	13	8	7	7	7	9	10	6	3	1	1	16	4	12	3	17	7
391343	0	2	2	0	10	0	8	0	0	4	0	0	0	0	2	0	0	4	1	0	0	0
391714	4	2	2	0	0	0	0	2	1	0	0	0	2	0	0	0	0	4	0	1	4	0
404770	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
441362	0	2	2	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	1	4	3	0

Figura 3: Matriz de expresión descargada del repositorio del TGCA. Únicamente se muestran las primeras filas y columnas. En total la matriz está compuesta por 757 columnas que representa a los pacientes y 20253 filas, correspondientes a los genes.

A continuación se muestra el *workflow* seguido en R/Bioconductor en el cuál se consiguió un listado de genes DE entre los dos grupos de pacientes, negativos y positivos, según el *status* del receptor de estrógenos.

2.4. *Workflow* para el análisis de datos RNAseq

El primer paso del *workflow* fue la obtención de la base datos RNAseq, comentada en el apartado anterior. En segundo lugar se le aplicó distintos tipos de normalización a los datos. A continuación se obtuvo un subgrupo de genes con una relación directa o indirecta con el cáncer de mama. Y finalmente se realizó un análisis de expresión diferencial de dichos genes a partir de paquetes presentes en el proyecto Bioconductor.

En los siguientes apartados se hará una explicación de las ventajas y posibilidades que ofrece el proyecto Bioconductor al análisis de datos ómicos y posteriormente se describirá la metodología utilizada en esta aproximación.

2.4.1. Bioconductor

Bioconductor comenzó en el año 2001 con el objetivo de desarrollar e integrar software para el análisis estadístico de datos genómicos de alto rendimiento. Es un proyecto de código abierto que consta actualmente de 1383 paquetes que aportan unas características definidas al proyecto (Gentleman *et al.*, 2004).

El uso de R como lenguaje estadístico aporta ciertas ventajas al ser un lenguaje interpretado de alto nivel que es fácil y rápido para el prototipado de nuevos métodos computacionales. Presenta un sistema bien establecido para el empaquetado combinado de software y documentación, así como un marco orientado a objetos para abordar la diversidad y complejidad de la biología computacional y los problemas bioinformáticos. Ofrece la posibilidad de acceder de forma online a datos bioinformáticos de vanguardia y tiene la capacidad de representar visualmente los modelos generados.

Cada paquete de Bioconductor posee, lo que se denomina *vignettes,* una documentación que describe la función del paquete, para conseguir una mayor reproducibilidad de la investigación.

Otra característica es la variedad de herramientas de la que dispone para la realización de experimentos con diferentes métodos estadísticos y la capacidad de generar gráficos. Los paquetes para el análisis de datos están dirigidos a *arrays* de oligonucleotidos, análisis de secuencias, citometría de flujo, y otros datos genómicos de alto rendimiento, entre ellos los datos RNAseq. Además, esta característica se ve complementada a la función que ejercen los paquetes propios de R, proporcionando un gran número de técnicas gráficas y estadísticas que incluyen métodos lineales, no lineales, análisis de *clúster*, predicción, remuestreo y análisis de supervivencia entre otros.

Finalmente, es posible la asociación de los datos genómicos, en tiempo real, con los metadatos biológicos presentes en las páginas web tales como GenBank, Entrez *genes* y Pubmed. Estos datos, que se pueden importar directamente a R, se denominan datos de anotación (Huber *et al.*, 2015).

2.4.2. Metodología para el análisis de datos RNAseq mediante Bioconductor

A continuación se describirán los procesos seguidos en Bioconductor para realizar el análisis de genes DE entre los dos grupos (positivos y negativos) en cuanto al *status del receptor de estrógenos.* Las fases serán: importación y procesado de los datos, normalización, obtención de los genes asociados a cáncer de mama y análisis diferencial.

2.4.2.1. Importación y procesado de los datos

Se importaron a R los datos obtenidos del repositorio TGCA. Los datos fueron descargados en formato .RData y su importación creó un objeto ExpressionSet.

Este objeto presenta los dos tipos de datos anteriormente comentados. Por una parte la base de datos de expresión, donde se observa el conteo de cada gen para cada paciente. Dicha matriz tiene 20532 x 757 dimensiones, que corresponden a los genes y a los pacientes, respectivamente. Por otra parte, distintas variables fenotípicas de los pacientes.

Importados los datos del estudio, se importaron los datos de anotación en la versión ensembl63 *(http://www.ensembl.org/index.html*), presentes en la plataforma Bioconductor. Esta matriz se corresponde con todos los genes humanos reportados hasta el momento (en este caso hasta la versión 63 de ensembl, aunque existen actualizaciones). Gracias a ello, se pueden comparar los genes que se obtuvieron mediante la plataforma TGCA y los que están presentes en la versión ensembl63. De esta manera se descartan los genes que no sean humanos, propiciados por contaminación, o genes que pudiesen estar mal anotados. Tras realizar la intersección entre las dos matrices, se obtuvo un total de 18257 genes. Se eliminaron 2275 genes de la matriz de expresión que no estaban representados en los datos de anotación ensembl63. La matriz obtenida posee ahora unas dimensiones 18257 x 757.

Debido a que la variable categórica que indica el *status del receptor de estrógenos* está compuesta por más de dos grupos (positivos y negativos), se eliminaron todos los pacientes que no presentaban dichas clases. Se redujo la dimensionalidad de la matriz a 18257 x 721.

2.4.2.2. Normalización

Aunque se ha realizado un procesado previo, es necesario someter los datos a algún tipo de normalización. Habitualmente los datos no presentan una relación lineal respecto al número original de transcriptos, por dos razones. Primero, los transcriptos largos generan más lecturas debido simplemente a su longitud. Segundo, porque un simple gen típicamente codifica para múltiples transcriptos con diferentes longitudes (la expresión de un gen en términos del número de conteos esperados puede permanecer constante incluso mientras el número de transcriptos que se producen a partir de ese gen cambia). Por lo tanto, si se

desea comparar los resultados de dos o más secuenciaciones, es necesaria una normalización de los resultados (Bullard *et al.*, 2009).

La mayoría de estos análisis presentan una dificultad a la hora de comparar las variables en igualdad de condiciones debido a que éstas contienen escalas o tamaños que difieren significativamente La función de la normalización es hacer un re-escalado lineal de las variables para poder observar los patrones de manera más clara. Con estas técnicas, las variables de entrada se tratan de manera independiente, de forma que las nuevas variables generadas tengan media igual a cero y desviación típica igual a uno. (Yang *et al.*, 2002).

La dificultad se presenta a la hora de elegir la técnica que mejor represente los datos. Existen multitud de técnicas de normalización, muchas de ellas presentes en paquetes de R. Elegir cuál es la mejor técnica únicamente es posible realizando la normalización sobre los datos de estudios, y obtener los resultados finales del análisis, ya que cada normalización no es general y no funciona de manera adecuada para todos los tipos de datos. Estudios realizados en los últimos años desarrollaron y compararon nuevos métodos de normalización específicos a datos RNAseq (Bullard *et al.*, 2009),(Dillies *et al.*, 2013). Las técnicas utilizadas en este trabajo fueron: RPKM, TMM y CQN.

2.4.2.2.1. RPKM

La técnica *Reads per kilo base per million mapped reads* (RPKM) fue inicialmente introducida para facilitar las comparaciones entre genes dentro de una muestra. RPKM re-escala el conteo de genes para corregir las diferencias en el tamaño de la biblioteca genómica y la longitud del gen.

Se puede dividir el proceso en tres pasos. En primer lugar se cuenta el total de lecturas en una muestra y se divide dicho número entre un millón (éste sería el factor de escala “por millón”). A continuación divide los conteos de la lectura por el factor de escala “por millón”. De esta forma se normaliza la profundidad de la secuenciación, obteniendo lecturas por millón (RPM). Finalmente los valores RPM se dividen por la longitud del gen, en kilobases, dando lugar a lo

que se conoce como RPKM (*reads per kilobase per million)* (Dillies *et al.*, 2013).

EQ 1

$$RPKM = \frac{numReads}{\frac{geneLength}{1000} * \frac{totalNumReads}{1000000}}$$

2.4.2.2.2. TMM

La técnica *Trimmed Mean of M-values* (TMM) se basa en la hipótesis de que la mayoría de los genes no están diferencialmente expresados.

Si se define Y_{gk} como el conteo por gen g observado en la biblioteca k, μ_{gk} como el nivel de expresión verdadero y desconocido (número de transcriptos), L_g como la longitud del gen g y N_k como el total de lecturas en la biblioteca k. El conteo de lecturas se obtendrá de la siguiente forma.

EQ 2

$$E[Y_{gk}] = \frac{\mu_{gk} * L_g}{S_k} N_k$$

EQ 3

$$donde\ S_k = \sum_{g=1}^{G} \mu_{gk} * L_{g_i}$$

S_k representa el total de la salida RNA de la muestra. El problema reside en el en que mientras Nk es conocido, Sk es desconocido y puede variar drásticamente de una muestra a otra, dependiendo de la composición de mRNA de cada paciente. Los pacientes que presenten una salida mucho mayor de mRNA provocarán el inframuestreo de muchos genes, en relación con las otras muestras. (Robinson and Oshlack, 2010).

Por lo tanto, el objetivo de esta técnica de normalización es reducir el sesgo producido cuando la distribución de conteos por tránscritos a lo largo de la muestra es diferente para distintas muestras (Dillies *et al.*, 2013)..

En primer lugar se realiza una reducción de la expresión de cada gen por muestra a través de un valor *log-fold-change* para cada muestra

EQ 4

$$M_g = log_2 \frac{Y_{gk} / N_k}{Y_{gk}' / N_k'}$$

Y un valor absoluto de la expresión genética

EQ 5

$$A_g = \frac{1}{2} log_2 \left(Y_{gk} / N_k * Y_{gk}' / N_k' \right) para\ Y_{g.} \neq 0$$

Finalmente, el factor de normalización TMM es calculado para cada muestra k usando de referencia las muetras r (para más detalles (Robinson and Oshlack, 2010)).

EQ 6

$$log_2\left(TMM_k^{(r)}\right) = \frac{\sum_{g \in G^*} w_{gk}^r M_{gk}^r}{\sum_{g \in G^*} w_{gk}^r}$$

2.4.2.2.3. CQN

La técnica *Conditional Quantil Normlization* (CQN) fue propuesto por (Hansen, Irizarry and Wu, 2012). Asume que los datos de RNAseq están afectados en gran medida por sesgos y errores sistemáticos, con un efecto sobre la muestra que no es común a todos los genes.

CQN se basa en un modelo estadístico para corregir el sesgo y las distorsiones distributivas. Se denota el logaritmo de la expresión genética del gen g en la muestra i con $\theta_{g,i}$, considerada una variable aleatoria. Las *p* covariaciones se

consideraran la causa del error sistemático y se denotan mediante $X_g = (X_{g,1}, \ldots, X_{g,p})'$. Ejemplo de covariaciones englobadas en *p* son, el contenido de GC, la longitud del gen y el mapeado del gen, definido como el porcentaje de mapeo exclusivo de lecturas de un gen. Por lo tanto, el modelo de lecturas observadas $Y_{g,i}$, para el gen *g* en la muestra *i*, se escribe:

EQ 7

$$Y_{g,i}|\mu_{g,i} \sim Poisson(\mu_{g,i})$$

EQ 8

$$donde, \mu_{g,i} = exp\left\{hi(\theta_{g,i}) + \sum_{j=1}^{p} f_{i,j}(X_{g,j}) + \log(m_i)\right\}$$

hi es una función no decreciente para tener en cuenta que la distribución de las lecturas de los genes son distribuciones no lineales. El $f_{i,j}$ se cuenta para los sesgos sistemáticos dependientes de la muestra. En cuanto m_i es la profundidad de secuenciación en millones.

Una vez descrito el modelo, la salida de la normalización para cada gen g en una muestra i se define como (para más información sobre cómo funciona el algoritmo, consulten el artículo (Hansen, Irizarry and Wu, 2012):

EQ 9

$$t_{g,i} = exp[\log(Y_{g,i} + 1) - \log(m_i) - \hat{h}^{-1}\{\log(Y_{g,i} + 1) - \hat{f}_{i,j}(X_{g,j})\}]$$

Por lo tanto, este método no solo considera la longitud del gen y el tamaño de la librería, sino que también considera el conteo de GC. Los autores se centraron en el sesgo del contenido en guanina y citocina específico (GC) de la muestra, que resultó en la confusión de los valores de contenido GC y en los valores *log foldchange*. De esta forma se elimina el sesgo producido por el contenido de GC específico de la muestra asumiendo que la distribución de los valores de expresión verdadera no depende del contenido de GC.

2.4.2.2.4. Evaluación gráfica de los distintos métodos de normalización

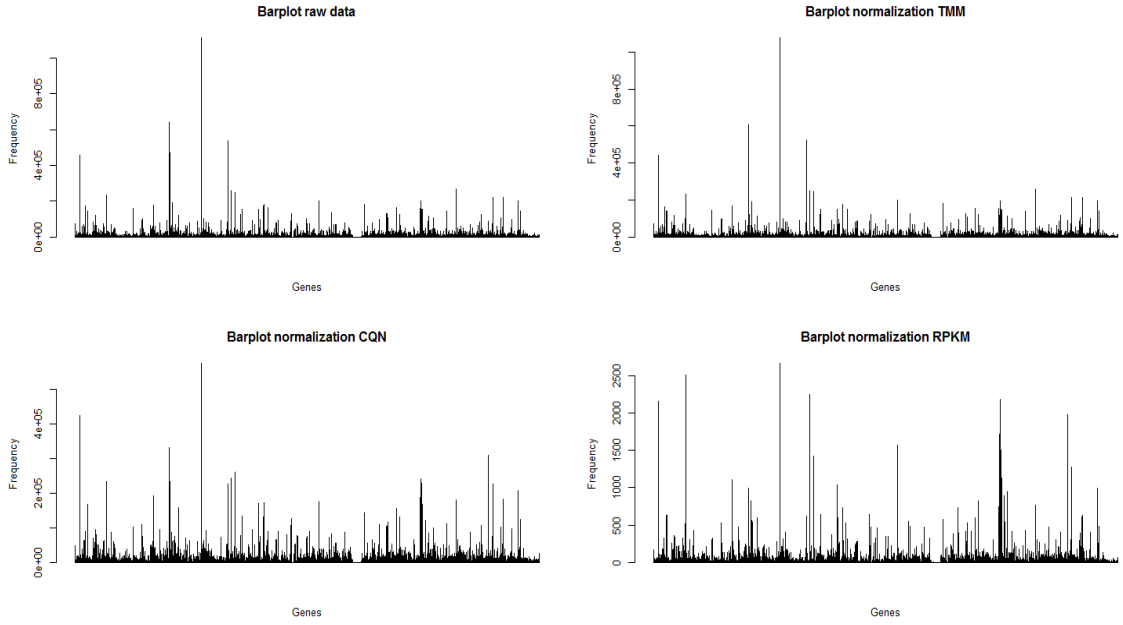


Figura 4: Gráficos de barras de la expresión media de cada gen. El primer gráfico corresponde a los datos en crudo, mientras que los otros tres, a los distintos métodos de normalización utilizados.

Para la evaluación de los distintos tipos de normalización se realizaron diferentes gráficas. En la ***Figura 4*** se observa el cambio en la media de la expresión de cada gen según el método de normalización. La diferencia de la expresión media de cada gen disminuye según se le aplica la normalización. En la gráfica se detecta un mayor cambio en este aspecto en la normalización CQN y RPKM. En cuanto a la TMM no se observa mucha diferencia con los datos en crudo.

Una forma más precisa de comparar los tipos de normalización es mediante gráficos MA-plot. Éstos son gráficos representan el ratio de la intensidad logarítmica (M-values) frente a la media de la intensidad logarítmica (A-values). Dichos gráficos son utilizados comúnmente para ilustrar la dependencia de las intensidades en datos de alto rendimiento. Usando MA-plots, se muestra la distribución de M-values comparándolo entre dos muestras después de su normalización. En una perfecta distribución de los datos normalizados la intensidad de los ratios logarítmicos M deben estar distribuidos en torno a cero a lo largo de todos los valores de intensidad A. En las ***Figura 5*** y ***Figura 6*** se

muestran los MA-plots, uno para los datos en crudo, y los otros tres para los datos normalizados.

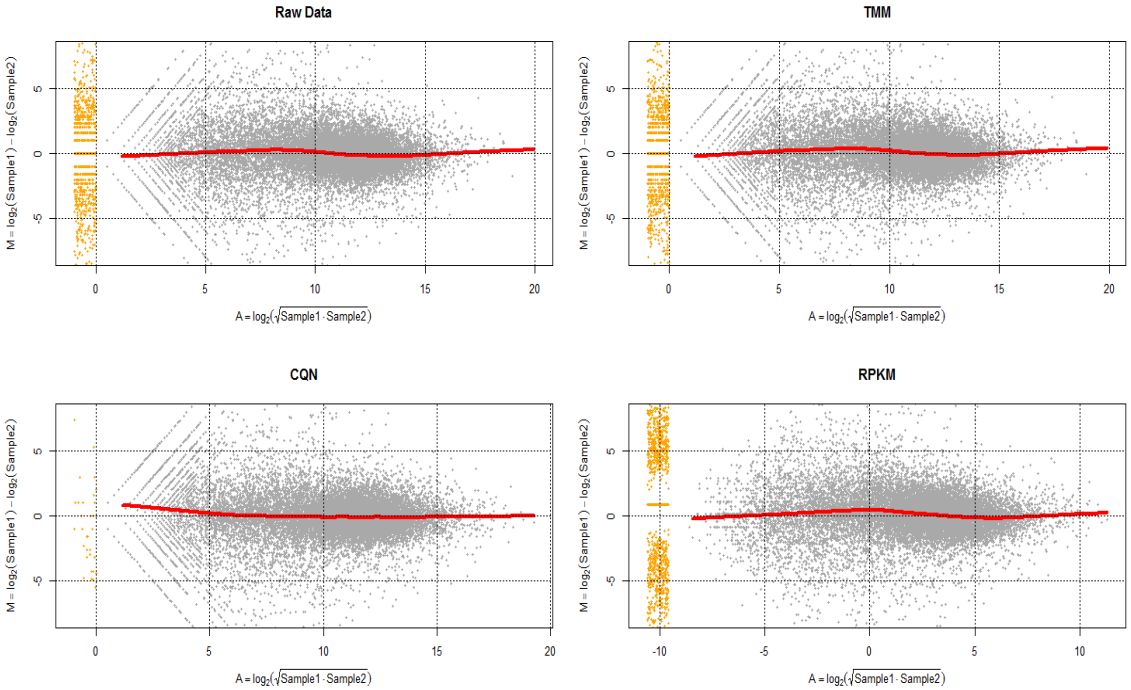


Figura 5: MA-plots para comparar la distribución de los datos normalizados en cuanto a los datos en crudo. Se compara la distribución de la expresión entre las muestras 1 y 2.

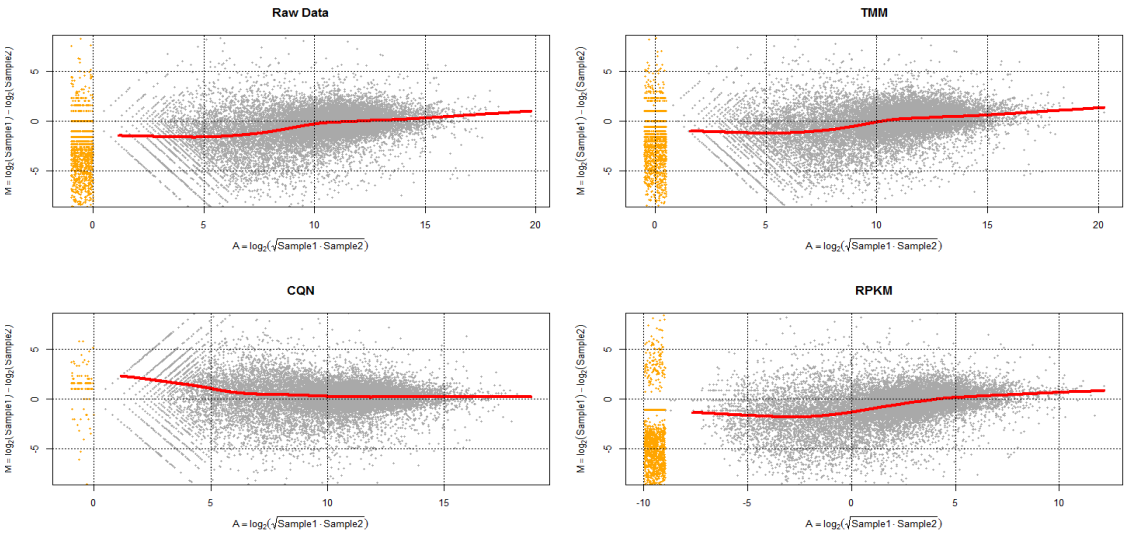


Figura 6: MA-plots para comparar la distribución de los datos normalizados en cuanto a los datos en crudo. Se compara la distribución de la expresión entre las muestras 30 y 700.

Se representan dos conjuntos de MA-plots para observar las diferencias en el efecto de la normalización sobre dos muestras distribuidas de forma deseada

(***Figura 5***) frente a dos muestras muy mal distribuidas en los datos en crudo (***Figura 6***). En lo que respecta a la ***Figura 5***, se observa que las tres normalizaciones actúan de forma correcta, distribuyendo la recta a lo largo de la horizontal. De todas formas, se detectan desviaciones de los datos en el extremo izquierdo de la normalización CQN, aunque donde existe una mayor densidad de datos, éstos se distribuyen correctamente en las cuatro gráficas.

En cuanto a la ***Figura 6*****,** la distribución de los datos en crudo es muy poco deseable, con lo que en este caso sí que es imprescindible una normalización. Se observa que CQN es el método que mejor normaliza los datos, aunque exista desviación en la parte izquierda de la gráfica. Pero en cuanto a la zona con mayor densidad de resultados, esta técnica es la que mejor resultados obtiene. Tanto TMM como RPKM presentan una desviación demasiado marcada en la zona con mayor densidad de datos.

Si se pudiesen comparar todas las muestras dos a dos, el resultado obtenido tendería más hacia la ***Figura 6*** que hacía la ***Figura 5***, con lo que someter los datos a un tipo de normalización es muy conveniente en este tipo de estudios. En este apartado del trabajo, se utilizará únicamente la normalización CQN, ya que fue considerado el método que mejor normalizó los datos. Más adelante, en el apartado de Machine Learning, se recuperarán las otras normalizaciones para poder trabajar con ellas.

2.4.2.3. Obtención de los genes asociados a cáncer de mama

El análisis sobre todo el conjunto de genes obtenidos a partir de la plataforma RNAseq (18257) reportaría un resultado que no tendría sentido biológicamente hablando.

La variabilidad de la expresión genética en los individuos no viene dada únicamente por una característica, como puede ser la sobreexpresión del receptor de estrógenos. Esta variabilidad se produce por todo un conjunto de variables que repercuten en el entorno del individuo. Por ejemplo, una persona, aún teniendo cáncer de mama, puede también tener una alteración metabólica

que genere una sobreexpresión de un determinado gen. A la hora de categorizar a dicho paciente, se le clasifica como un HER2 positivo. Cuando se le hace su perfil de expresión, resultará que un gen que nada tiene que ver con dicho receptor ha dado como resultado que está sobreexresado, en relación a otro paciente HER2 negativo. En este caso, se consideraría el gen relacionado al cáncer de mama, aunque su relación sea nula y realmente es debido a la alteración metabólica. Este es un ejemplo muy simple, ya que a la hora de analizar estos perfiles de expresión, se tienen en cuenta estas consideraciones, y en los resultados reportados siempre existe una corrección estadística. Aún así, dicha corrección no llega a ser tan eficiente cuando se está comparando un número tan grande de genes.

Es crucial, por lo tanto, extraer un subgrupo de genes del y trabajar sobre genes que a *priori* tengan sospechas una posible relación ante una patología o estadío de la enfermedad.

Tras los incesantes estudios que se realizan a diario para la categorización genética de las enfermedades, en concreto para el cáncer, existen una multitud de plataformas web públicas, con listados de los genes relacionados a cada tipo de cáncer. En la ***Tabla 2*** se muestran cinco plataformas donde se pueden consultar las características genéticas de las enfermedades, así como los genes relacionados a ellas.

Tabla 2: Repositorio de genes públicos de donde se podrían sacar información acerca de enfermedades genéticas.

Repositorio	Descripción	Enlace
OMIM	Catálogo online de genes humanos y enfermedades genéticas	https://www.omim.org/
COSMIC	Catálogo de mutaciones somáticas en cáncer. La mayor fuente para explorar el impacto de mutaciones somática en cáncer.	http://cancer.sanger.ac.uk/cosmic
DISGENET	Plataforma que integra información sobre enfermedades asociadas a genes de varios repositorios de datos públicos y literatura.	http://www.disgenet.org/web/DisGeNET
INTOGEN	Repositorio para facilitar el estudio computacional del cáncer desde una perspectiva genómica.	https://www.intogen.org
WIKIPATHWAYS	Base de datos abierta de rutas biológicas mantenida por y para la comunidad científica	http://www.wikipathways.org/index.php/WikiPathways

Para este proyecto se ha decidido realizar un listado de genes relacionadas al cáncer de mama a partir del repositorio de INTOGEN y del repositorio WIKIPATHWAYS. En los dos próximos apartados se explicará cada una de estas plataformas.

2.4.2.3.1. Repositorio INTOGEN

Intogen es un sistema innovador para el análisis e integración de una gran cantidad de datos genómicos del cáncer. Los resultados generados pueden consultarse en la web www.intogen.org. En el artículo de (Gonzalez-Perez *et al.*, 2013) publicado en Nature Methods se describe este nuevo recurso poniéndolo a disposición de la comunidad científica.

Intogen permite recopilar, analizar e integrar los datos de experimentos genómicos que estudian diferentes tipos de alteraciones en distintos tipo de tumores humanos. La versión actual del sistema está compuesta por casi 800 experimentos independientes.

De esta manera, se obtuvo un listado de los genes somáticos más probables a ser mutados en pacientes con cáncer de mama. Se consiguió una lista de un total de 184 genes.

2.4.2.3.2. Repositorio WIKIPATHWAYS

WikiPathways es una plataforma abierta y colaborativa dedicada a la custodia de rutas biológicas. Así, WikiPathways presenta un nuevo modelo de bases de datos de rutas que mejora y complementa los esfuerzos en curso, como KEGG, Reactome y Pathway Commons. Basándose en el mismo software de MediaWiki que potencia Wikipedia, se añade una herramienta de edición de vías gráficas personalizadas y bases de datos integradas que cubren los principales sistemas de genes, proteínas y pequeñas moléculas.

De esta forma se obtiene una visión intuitiva de la miríada de interacciones que subyacen a los procesos biológicos. Una ruta de señalización típica, por ejemplo, puede representar eventos de unión al receptor, complejos de

proteínas, reacciones de fosforilación, translocaciones y regulación transcripcional, con sólo un conjunto de símbolos y líneas. El obtener todos estos componentes, en forma de genes, para cualquier tipo de patología, característica o ambiente celular es fundamental para extraer la mayor cantidad de información necesaria. Por ello, se consideró de gran utilidad añadir a la lista, los genes que participan en las rutas biológicas, en este caso, en el cáncer de mama. La mayoría de estos genes no se vieron de momento relacionados con este tipo de cáncer, aunque su desregulación conllevaría probablemente a una alteración del ciclo celular que propiciaría una catástrofe tanto genética como metabólica (Kutmon *et al.*, 2016).

La ruta seleccionada engloba las proteínas más importantes del cáncer de mama. Una vez seleccionadas las proteínas más importantes mediante el *score* Rp de la web Connectivity-Maps (C-Maps), Wikipathways determina las rutas más importantes envueltos en cáncer de mama gracias al Human Pathway Database (HPD). Estas rutas fueron anotadas y se determinó las interacciones proteína-proteína en esta ruta biológica seleccionada. Para finalmente extraer los 155 genes que intervienen en la ruta del cáncer de mama.

2.4.2.3.3. Unión de los genes obtenidos en los dos repositorios

Se obtuvo un listado final de 308 genes a partir de los dos repositorios. De este conjunto de genes, 301 fueron identificados en la base de datos. Esto es debido posiblemente a que los genes restantes no tuviesen un suficiente número de conteos.

De esta forma se busca un menor subconjunto de genes que previamente han sido reportados como *drivers* y que comparten vías celulares. Si los resultados obtenidos en este proyecto son esperanzadores para una futura investigación, la posibilidad de añadir más genes al estudio puede ser perfectamente viable. Al ser el objetivo final de este proyecto la elaboración de un *pipeline* basado en ML y su posterior comparación con los resultados obtenidos en la aproximación

estándar, el listado de genes obtenidos se considera suficiente para observar el rendimiento del modelo y las perspectivas futuras.

Una vez definido el subgrupo de genes con los que trabajar, el siguiente paso es realizar el análisis diferencial de expresión. En el siguiente apartado se describe el objetivo de este procedimiento así como la metodología utilizada.

2.4.2.4. Análisis diferencial

Los análisis de expresión diferencial sirven para identificar los genes con niveles de mRNA que fueron diferentes a través de los grupos experimentales (en este caso según el *status del receptor de estrógenos).*

El principal objetivo de este apartado es la obtención de un listado de genes que se expresan de una manera diferente entre pacientes ER positivos y negativos. En este momento se tiene un listado de genes bastante amplio, en el que se han recopilado la mayoría de genes que intervienen directa o indirectamente en este tipo de cáncer. De esta forma, es posible obtener genes DE que hasta la fecha no se habían tenido en cuenta en la caracterización del grupo de cáncer de mama HER2.

Para obtener el listado de genes DE este trabajo se basó en la estimación de la significancia de la diferencia de expresión entre los dos grupos y dicho valor fue corregido mediante *multiple testing* obteniendo así el p valor ajustado. El nivel de significación utilizado por lo tanto fue de $\alpha = 0.001$. De esta forma la probabilidad de cometer el error estadístico de tipo I es prácticamente nulo, y los resultados obtenidos tienen muy pocas posibilidades de presentar algún falso positivo.

Existe una gran variedad de paquetes para el análisis de datos de expresión con formato RNAseq. Dentro del proyecto Bioconductor se encuentran también una variedad de ellos, cada uno con sus ventajas y limitaciones. En este estudio se ha decidido usar tres paquetes para el análisis de este tipo de datos. A continuación se explicará la función de cada uno de ellos, y se destacarán las diferencias que existen entre ellos.

2.4.2.4.1. Paquete edgeR

El paquete edgeR fue diseñado para el análisis de datos de expresión basados en conteos. Aunque inicialmente fue desarrollado para el análisis en serie de expresión genética (SAGE por sus siglas en inglés), este método es igualmente aplicable para tecnologías como RNAseq (Anders *et al.*, 2013).

Este paquete realiza un conteo de los datos usando un modelo Poisson sobredispersado y un procedimiento empírico de Bayes para moderar el grado de sobredispersión de los genes. Se asume que los los datos tengan una distribución binomial negativa (NB),

EQ 10

$$Y_{gi} \sim NB(M_i p_{gi}, \varphi_g)$$

para el gen *g* y la muestra *i.* Aquí, M_i es el tamaño de la librería (total número de lecturas), φ_g coeficiente de variación entre muestras biológicas y p_{gi} es la abundancia relativa del gen *g* en un grupo experimental *j* al cual pertenece la muestra *i* (Robinson, McCarthy and Smyth, 2009).

Por lo tanto, *edgeR* estima la dispersión genética por la probabilidad del máximo condicional sobre el conteo total para cada gen. Además, un procedimiento *Bayesiano* empírico es usado para disminuir la dispersión hacia un valor consenso. Finalmente, la expresión diferencial es evaluada para cada gen usando un test análogo al test exacto de Fisher pero adaptados a datos sobredispersados.

2.4.2.4.2. Paquete DEseq

Este método es muy similar al método planteado por el paquete edgeR ya que asume que los perfiles de expresión de los datos RNAseq siguen una distribución binomial negativa y necesita información de los genes para estimar primero un parámetro de dispersión común. Luego, para cada gen, se estima su dispersión genética y se modera hacia la común. La forma en que esta moderación tiene lugar depende del método y sus parámetros de configuración. Es aquí donde los dos métodos difieren de manera significativa ya que DEseq coge el valor máximo de la dispersión estimada del individuo y la dispersión de

la tendencia media. Al contrario que edgeR que modera las estimaciones de dispersión a nivel de la característica hacia una media tendencial según la relación de la media de dispersión. En la práctica, esto conlleva a que DEseq es menos poderoso, mientras que edgeR es más sensible a los *outliers* (Anders *et al.*, 2013).

2.4.2.4.3. Paquete tweeDEseq

Esta librería se desarrolló al observar que cuando se replican ampliamente los perfiles de expresión producidos por los experimentos RNAseq se obtienen patrones de distribución muy diversos. Se detectaron distribuciones como la inversa de Poisson-Gauss o Polya-Aeppli que diferían de las distribuciones que se utilizan para el análisis de este tipo de datos, como es la distribución binomial negativa o la distribución de Poisson. Por lo tanto la implementación de técnicas bioinformáticas que presentaran familias de distribuciones de datos más generales, abarcaría un mayor número de distribuciones. Por lo que se presentó una familia más general de las distribuciones de datos de conteo denominada Poisson-Tweedie (Esnaola *et al.*, 2013).

La flexibilidad de la familia Poisson-Tweedie permite un ajuste directo de las características emergentes de los grandes perfiles de expresión, tales como las colas pesadas o la inflación cero, sin necesidad de alterar ningún parámetro de la configuración. Aunque en este trabajo no se pretende definir estas librerías mediante una jerga matemática más formal, para mayor interés, tanto la distribución como el paquete se explica en el artículo de (Esnaola *et al.*, 2013).

2.5. Discusión y resultados de la aproximación estándar

Tras la búsqueda en los repositorios genéticos de Intogen y Wikipathways, el listado final de los 301 genes que para este proyecto se consideran que tienen o pueden llegar a tener alguna relación con el cáncer de mama se presenta en el ANEXO III.

Tabla 3: Resumen de los resultados del análisis diferencial utilizando los tres paquetes.

Librería usada	Genes DE	Presentes en el listado Intogen	Presentes en el listado Wikipathways	Tipo de distribución del paquete
edgeR	161	91	81	Binomial negativa
DEseq	123	69	62	Binomial negativa
tweeDEseq	143	82	72	Tweedie-Poisson
Comunes	116	66	58	

Como se puede ver en la ***Tabla 3***, existen diferencias entre los tres métodos. Al usar edgeR se obtuvieron 161 genes DE, mientras que si se usa DEseq se obtuvieron 123 genes DE y con tweeDEseq, 143. Probablemente, edgeR ha declarado de forma significativa genes que no están involucrados con los procesos de cáncer de mama del grupo HER2 postivo/negativo. Por otra parte, las técnicas DEseq y tweeDEseq, probablemente, han representado mejor los datos. La idea ahora sería hacer un análisis de enriquecimiento, es decir, obtener cuál fue la técnica que mayor número de genes relacionados al grupo HER2 postivo/negativo detectó como DE. Esto no es posible ya que no existe una lista con todos los genes relacionados con este grupo. Se han reportado diversos genes que caracterizan a dicho grupo, pero hacer un análisis de enriquecimiento con una lista muy pequeña no sería muy adecuado. Por lo tanto se ha optado por hacer una comparación entre los resultados de las tres técnicas y sacar un listado con los genes comunes identificados por todas las técnicas.

Los genes comunes obtenidos fueron un total de 116. Se considera así, que las tres formas de análisis diferenciales usados en este proyecto han conseguido resultados bastante parecidos. De este total de genes, 66 estaban presentes en la lista de Intogen mientras que 58 estaban presentes en la lista de Wikipathways. Se detectaron 8 genes presentes en las dos listas. Existe aproximadamente un 50% de genes que pertenecen a cada una de los repositorios de donde se obtuvieron los genes. Esto es un resultado esperanzador a la hora de obtener algún gen que no se haya relacionado hasta la fecha con este grupo de cáncer de mama, ya que la mayoría de genes obtenidos mediante Wikipathways no se habían relacionado con esta patología

en la literatura científica. La ***Tabla 4*** muestra los genes que las tres pruebas detectaron como genes DE.

Tabla 4: Genes DE detectados mediante las tres técnicas

hgnc_symbol	description
RAD50	RAD50 double strand break repair protein [Source:HGNC Symbol;Acc:HGNC:9816]
AKAP9	A-kinase anchoring protein 9 [Source:HGNC Symbol;Acc:HGNC:379]
RBM5	RNA binding motif protein 5 [Source:HGNC Symbol;Acc:HGNC:9902]
CDK7	cyclin dependent kinase 7 [Source:HGNC Symbol;Acc:HGNC:1778]
CDKN1B	cyclin dependent kinase inhibitor 1B [Source:HGNC Symbol;Acc:HGNC:1785]
NDRG1	N-myc downstream regulated 1 [Source:HGNC Symbol;Acc:HGNC:7679]
CARM1	coactivator associated arginine methyltransferase 1 [Source:HGNC Symbol;Acc:HGNC:23393]
RPP38	ribonuclease P/MRP subunit p38 [Source:HGNC Symbol;Acc:HGNC:30329]
ARFGEF2	ADP ribosylation factor guanine nucleotide exchange factor 2 [Source:HGNC Symbol;Acc:HGNC:15853]
NOXA1	NADPH oxidase activator 1 [Source:HGNC Symbol;Acc:HGNC:10668]
STIP1	stress induced phosphoprotein 1 [Source:HGNC Symbol;Acc:HGNC:11387]
CHEK1	checkpoint kinase 1 [Source:HGNC Symbol;Acc:HGNC:1925]
CHEK2	checkpoint kinase 2 [Source:HGNC Symbol;Acc:HGNC:16627]
CLTC	clathrin heavy chain [Source:HGNC Symbol;Acc:HGNC:2092]
CSNK1G3	casein kinase 1 gamma 3 [Source:HGNC Symbol;Acc:HGNC:2456]
E2F1	E2F transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:3113]
ARID2	AT-rich interaction domain 2 [Source:HGNC Symbol;Acc:HGNC:18037]
EIF4G1	eukaryotic translation initiation factor 4 gamma 1 [Source:HGNC Symbol;Acc:HGNC:3296]
ELF1	E74 like ETS transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:3316]
ESR1	estrogen receptor 1 [Source:HGNC Symbol;Acc:HGNC:3467]
RASGEF1A	RasGEF domain family member 1A [Source:HGNC Symbol;Acc:HGNC:24246]
CCT5	chaperonin containing TCP1 subunit 5 [Source:HGNC Symbol;Acc:HGNC:1618]
CLASP2	cytoplasmic linker associated protein 2 [Source:HGNC Symbol;Acc:HGNC:17078]
FLT3	fms related tyrosine kinase 3 [Source:HGNC Symbol;Acc:HGNC:3765]
ACSL6	acyl-CoA synthetase long-chain family member 6 [Source:HGNC Symbol;Acc:HGNC:16496]
FMR1	fragile X mental retardation 1 [Source:HGNC Symbol;Acc:HGNC:3775]

ZMYND8	zinc finger MYND-type containing 8 [Source:HGNC Symbol;Acc:HGNC:9397]
RALGAPA1	Ral GTPase activating protein catalytic alpha subunit 1 [Source:HGNC Symbol;Acc:HGNC:17770]
RASGRP3	RAS guanyl releasing protein 3 [Source:HGNC Symbol;Acc:HGNC:14545]
ASPM	abnormal spindle microtubule assembly [Source:HGNC Symbol;Acc:HGNC:19048]
GATA3	GATA binding protein 3 [Source:HGNC Symbol;Acc:HGNC:4172]
GDI1	GDP dissociation inhibitor 1 [Source:HGNC Symbol;Acc:HGNC:4226]
FOXP1	forkhead box P1 [Source:HGNC Symbol;Acc:HGNC:3823]
AFF4	AF4/FMR2 family member 4 [Source:HGNC Symbol;Acc:HGNC:17869]
ARHGAP35	Rho GTPase activating protein 35 [Source:HGNC Symbol;Acc:HGNC:4591]
MSH6	mutS homolog 6 [Source:HGNC Symbol;Acc:HGNC:7329]
ANXA1	annexin A1 [Source:HGNC Symbol;Acc:HGNC:533]
HCFC1	host cell factor C1 [Source:HGNC Symbol;Acc:HGNC:4839]
HLA-A	major histocompatibility complex, class I, A [Source:HGNC Symbol;Acc:HGNC:4931]
FOXA1	forkhead box A1 [Source:HGNC Symbol;Acc:HGNC:5021]
IRS1	insulin receptor substrate 1 [Source:HGNC Symbol;Acc:HGNC:6125]
AR	androgen receptor [Source:HGNC Symbol;Acc:HGNC:644]
LCP1	lymphocyte cytosolic protein 1 [Source:HGNC Symbol;Acc:HGNC:6528]
LRP6	LDL receptor related protein 6 [Source:HGNC Symbol;Acc:HGNC:6698]
SMAD2	SMAD family member 2 [Source:HGNC Symbol;Acc:HGNC:6768]
MAX	MYC associated factor X [Source:HGNC Symbol;Acc:HGNC:6913]
MDM2	MDM2 proto-oncogene [Source:HGNC Symbol;Acc:HGNC:6973]
MAP3K1	mitogen-activated protein kinase kinase kinase 1 [Source:HGNC Symbol;Acc:HGNC:6848]
MMP1	matrix metallopeptidase 1 [Source:HGNC Symbol;Acc:HGNC:7155]
MSH2	mutS homolog 2 [Source:HGNC Symbol;Acc:HGNC:7325]
MYB	MYB proto-oncogene, transcription factor [Source:HGNC Symbol;Acc:HGNC:7545]
MYC	v-myc avian myelocytomatosis viral oncogene homolog [Source:HGNC Symbol;Acc:HGNC:7553]
MYH9	myosin heavy chain 9 [Source:HGNC Symbol;Acc:HGNC:7579]
NAB1	NGFI-A binding protein 1 [Source:HGNC Symbol;Acc:HGNC:7626]
NF2	neurofibromin 2 [Source:HGNC Symbol;Acc:HGNC:7773]
NOTCH1	notch 1 [Source:HGNC Symbol;Acc:HGNC:7881]
NRAS	neuroblastoma RAS viral oncogene homolog [Source:HGNC

	Symbol;Acc:HGNC:7989]
NR4A2	nuclear receptor subfamily 4 group A member 2 [Source:HGNC Symbol;Acc:HGNC:7981]
ODC1	ornithine decarboxylase 1 [Source:HGNC Symbol;Acc:HGNC:8109]
PCSK6	proprotein convertase subtilisin/kexin type 6 [Source:HGNC Symbol;Acc:HGNC:8569]
PIK3R2	phosphoinositide-3-kinase regulatory subunit 2 [Source:HGNC Symbol;Acc:HGNC:8980]
PLK1	polo like kinase 1 [Source:HGNC Symbol;Acc:HGNC:9077]
PML	promyelocytic leukemia [Source:HGNC Symbol;Acc:HGNC:9113]
BNC2	basonuclin 2 [Source:HGNC Symbol;Acc:HGNC:30988]
BCOR	BCL6 corepressor [Source:HGNC Symbol;Acc:HGNC:20893]
FBXW7	F-box and WD repeat domain containing 7 [Source:HGNC Symbol;Acc:HGNC:16712]
DHTKD1	dehydrogenase E1 and transketolase domain containing 1 [Source:HGNC Symbol;Acc:HGNC:23537]
PRKAR1A	protein kinase cAMP-dependent type I regulatory subunit alpha [Source:HGNC Symbol;Acc:HGNC:9388]
BACH1	BTB domain and CNC homolog 1 [Source:HGNC Symbol;Acc:HGNC:935]
ZMIZ1	zinc finger MIZ-type containing 1 [Source:HGNC Symbol;Acc:HGNC:16493]
PPP4R3B	protein phosphatase 4 regulatory subunit 3B [Source:HGNC Symbol;Acc:HGNC:29267]
PTEN	phosphatase and tensin homolog [Source:HGNC Symbol;Acc:HGNC:9588]
MKL1	megakaryoblastic leukemia (translocation) 1 [Source:HGNC Symbol;Acc:HGNC:14334]
BAK1	BCL2 antagonist/killer 1 [Source:HGNC Symbol;Acc:HGNC:949]
CCNB1IP1	cyclin B1 interacting protein 1 [Source:HGNC Symbol;Acc:HGNC:19437]
RAD51	RAD51 recombinase [Source:HGNC Symbol;Acc:HGNC:9817]
RB1	RB transcriptional corepressor 1 [Source:HGNC Symbol;Acc:HGNC:9884]
CCND1	cyclin D1 [Source:HGNC Symbol;Acc:HGNC:1582]
BCL2	BCL2, apoptosis regulator [Source:HGNC Symbol;Acc:HGNC:990]
RFC4	replication factor C subunit 4 [Source:HGNC Symbol;Acc:HGNC:9972]
RHEB	Ras homolog enriched in brain [Source:HGNC Symbol;Acc:HGNC:10011]
RPL5	ribosomal protein L5 [Source:HGNC Symbol;Acc:HGNC:10360]
BID	BH3 interacting domain death agonist [Source:HGNC Symbol;Acc:HGNC:1050]
CLSPN	claspin [Source:HGNC Symbol;Acc:HGNC:19715]
BLM	Bloom syndrome RecQ like helicase [Source:HGNC Symbol;Acc:HGNC:1058]
MAP2K4	mitogen-activated protein kinase kinase 4 [Source:HGNC Symbol;Acc:HGNC:6844]

CERK	ceramide kinase [Source:HGNC Symbol;Acc:HGNC:19256]
SMARCA4	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4 [Source:HGNC Symbol;Acc:HGNC:11100]
SOS1	SOS Ras/Rac guanine nucleotide exchange factor 1 [Source:HGNC Symbol;Acc:HGNC:11187]
SP1	Sp1 transcription factor [Source:HGNC Symbol;Acc:HGNC:11205]
BRCA2	BRCA2, DNA repair associated [Source:HGNC Symbol;Acc:HGNC:1101]
STAT1	signal transducer and activator of transcription 1 [Source:HGNC Symbol;Acc:HGNC:11362]
AURKA	aurora kinase A [Source:HGNC Symbol;Acc:HGNC:11393]
TBX3	T-box 3 [Source:HGNC Symbol;Acc:HGNC:11602]
TCF7L2	transcription factor 7 like 2 [Source:HGNC Symbol;Acc:HGNC:11641]
TFDP1	transcription factor Dp-1 [Source:HGNC Symbol;Acc:HGNC:11749]
VEGFA	vascular endothelial growth factor A [Source:HGNC Symbol;Acc:HGNC:12680]
CAD	carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase [Source:HGNC Symbol;Acc:HGNC:1424]
AHNAK	AHNAK nucleoprotein [Source:HGNC Symbol;Acc:HGNC:347]
ZNF655	zinc finger protein 655 [Source:HGNC Symbol;Acc:HGNC:30899]
DCAKD	dephospho-CoA kinase domain containing [Source:HGNC Symbol;Acc:HGNC:26238]
NUP85	nucleoporin 85 [Source:HGNC Symbol;Acc:HGNC:8734]
SVEP1	sushi, von Willebrand factor type A, EGF and pentraxin domain containing 1 [Source:HGNC Symbol;Acc:HGNC:15985]
CEP290	centrosomal protein 290 [Source:HGNC Symbol;Acc:HGNC:29021]
FOSL1	FOS like 1, AP-1 transcription factor subunit [Source:HGNC Symbol;Acc:HGNC:13718]
PHF6	PHD finger protein 6 [Source:HGNC Symbol;Acc:HGNC:18145]
RAD54L	RAD54-like (S. cerevisiae) [Source:HGNC Symbol;Acc:HGNC:9826]
CUL1	cullin 1 [Source:HGNC Symbol;Acc:HGNC:2551]
RUNX1	runt related transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:10471]
CBFB	core-binding factor beta subunit [Source:HGNC Symbol;Acc:HGNC:1539]
ALKBH1	alkB homolog 1, histone H2A dioxygenase [Source:HGNC Symbol;Acc:HGNC:17911]
KALRN	kalirin, RhoGEF kinase [Source:HGNC Symbol;Acc:HGNC:4814]
ACVR1B	activin A receptor type 1B [Source:HGNC Symbol;Acc:HGNC:172]
CDC25A	cell division cycle 25A [Source:HGNC Symbol;Acc:HGNC:1725]
CDC25B	cell division cycle 25B [Source:HGNC Symbol;Acc:HGNC:1726]
CDH1	cadherin 1 [Source:HGNC Symbol;Acc:HGNC:1748]

3. ESTADO DE LA CUESTION

El proyecto Bioconductor fue publicado en la literatura científica en el año 2004 (Gentleman *et al.*, 2004). El objetivo principal era su uso como una herramienta bioinformática para el análisis y desarrollo de datos de alto rendimiento. Es en 2008 cuando se citó por primera vez la técnica RNAseq, la cual fue rápidamente analizada por paquetes contenidos en Bionconductor.

El paquete edgeR fue publicado en el año 2009 (Robinson, McCarthy and Smyth, 2009), en cuanto el paquete DEseq fue desarrollado en el 2010 (Anders and Huber, 2010), mientras que tweeDEseq en el 2013 (Esnaola *et al.*, 2013).

Existen otros paquetes con análogas funciones que fueron desarrollados. Por ejemplo, NBPSeq, DEGseq o TSPM. Numerosos artículos han sido publicados para realizar una evaluación de los distintos paquetes mediante datos procedentes de la plataforma RNAseq (Guo *et al.*, 2013), (Anders *et al.*, 2013). Convirtiendo este tipo de métodos como los más usados para el análisis de datos de expresión.

Por otra parte, la historia entre la relación de la biología y el campo de ML es larga y compleja. Una técnica temprana para ML denominada *percepton,* constituyó un intento de modelar el comportamiento neuronal real, y el campo de diseño de la red neuronal artificial (RNA) surgió de este intento. Los primeros trabajos sobre la identificación de las secuencias de iniciación de la traducción emplearon el *percepton* para definir los criterios para los sitios de iniciación en *Escherichia coli.* Otras arquitecturas de RNAs como la teoría de resonancia adaptativa y neocognitron se inspiraron en la organización del sistema nervioso visual (Tarca *et al.*, 2007). En los años transcurridos, la flexibilidad de las técnicas de aprendizaje automático ha crecido junto con marcos matemáticos para medir su fiabilidad, y es natural esperar que los métodos de ML vayan mejorando la eficiencia del descubrimiento y comprensión en el volumen y complejidad de los datos biológicos.

A finales de los años 80 surgieron los primeros intentos de predicción con RNAs (red de neuronas artificiales) que consiguieron un aumento de un 20%

en el acierto de la predicción con respecto a los métodos estadísticos de estudio (Qian and Sejnowski, 1988).

A partir de ese momento, empezaron a desarrollarse experimentos que utilizan *Support Vector Machines* (SVM), llegando a ser una de las técnicas más utilizadas para la predicción de procesos biológicos (Guyon *et al.*, 2002).

Actualmente los desarrollos que se están haciendo en este campo apuntan a una medicina de precisión, lo que significa la aplicación de los biomarcadores de los datos ómicos y el genotipado de los pacientes para una mayor y apropiada terapia dirigida indiviualmente a cada paciente.

En cuanto al cáncer de mama, el trabajo de (Yamamoto *et al.*, 2017) obtiene un nivel de exactitud del 90.9% en la clasificación de los distintos tipos de tumores según las diferencias del microambiente mioepitelial. Otro trabajo reportado fue el de (Valkonen *et al.*, 2017) que consiguieron un valor AUC=0.97-0.98 para la detección de áreas metásticas usando el algoritmo Random Forest. Para la predicción de agonista del receptor de estrógenos, el trabajo de (Niu *et al.*, 2016) obtuvo un rango en el valor AUC entre 77.10% a 88.34% usando algoritmos como Naive Bayes, Random Forest, k-NN y SVM. Además, estudios como el de (Mungle *et al.*, 2017) proponen un enfoque basado en ML para la automatización de *score* del ER a partir de imágenes inmunohistoquímicas.

También son muchos los trabajos que se centran en datos genéticos del cáncer para la aplicación de ML. El trabajo de (Pepke and Ver Steeg, 2017) propone un método de *clustering* para identificar nuevos subgrupos de pacientes con tumores a partir de perfiles de expresión. En el artículo publicado por (Sundaramurthy and Eghbalnia, 2015) han aplicado algoritmos de ML para identificar enfermedades relacionadas con los genes a partir de datos RNAseq y *microarrays.* Finalmente, (Stupnikov, Glazko and Emmert-Streib, 2015) han usado técnicas de selección de variables para detectar importantes genes en pacientes de cáncer de mama triple negativos.

Toda esta revisión del estado del arte apunta a un auge de las técnicas basadas en ML para la mejora en el análisis, la predicción y el diagnóstico de los diferentes tipos de cáncer. Además, los resultados reportados por los

distintos estudios en los últimos años están siendo de gran calidad, con valores en el rendimiento de los modelos mayores al noventa por ciento.

Los resultados y la metodología presente en este proyecto no se han visto publicados anteriormente en la literatura científica. Es por ello que se considera un estudio vanguardista y de gran interés en este campo actualmente tan desarrollado.

4. DESARROLLO DEL *WORKFLOW* BASADO EN MACHINE LEARNING

Como se ha comentado anteriormente, los resultados obtenidos tras la aplicación de un protocolo basado en Bioconductor sirvieron de motivación para emplear un *pipeline* de análisis de datos mediante técnicas avanzadas en Machine Learning. En este capítulo se explicará todo el proceso seguido para este tipo de análisis, explicando cada proceso de la metodología de forma detallada.

El capítulo consta de dos grandes apartados. En el primero se hará una introducción del concepto ML y las posibilidades que genera, sobre todo en el campo de la biología. En el segundo, se realizará una explicación extensa sobre la metodología utilizada para los experimentos basados en ML..

4.1. Introducción a Machine Learning

Se denomina Machine Learning al campo de estudio interesado en el desarrollo de algoritmos computacionales capaces de transformar los datos en acciones inteligentes. Este campo fue originado en un entorno envuelto simultáneamente por la disponibilidad de datos, los métodos estadísticos y un poder computacional suficientemente rápido.

Una definición no formal de ML, propuesta por el informático Tom M. Mitchell, afirma que una máquina aprende cada vez que es capaz de utilizar una experiencia, de manera que su rendimiento es mejor en experiencias similares en el futuro.

El uso de algoritmos basados en ML tiene una gran diversidad en su aplicación. Se encuentran aplicaciones de ML, por ejemplo, en la identificación de mensajes *spam* de nuestro *e-mail,* en el autopilotaje de *drones* y coches, en la identificación de caras en *Facebook* y en la reducción de transacciones de tarjetas de créditos fraudulentas.

Concretamente los métodos basados en ML han sido aplicados en una amplia variedad de problemas de genómica y proteómica. Por ejemplo, en proteómica, ML puede ser usado para aprender como reconocer péptidos con actividad anti-angiogénica. En primer lugar los investigadores en ML desarrollan un algoritmo que conducirá a un aprendizaje exitoso. En segundo lugar, el algoritmo es provisto por una gran serie de secuencias peptídicas, clasificadas según su actividad. El algoritmo procesa esas secuencias etiquetadas y las almacena en un modelo. En tercer lugar, secuencias nuevas que no están etiquetadas son aportadas al algoritmo, y éste usa el modelo para predecir las etiquetas de cada secuencia nueva. Si el aprendizaje es exitoso, entonces la mayoría de las etiquetas predichas serán correctas (Tarca *et al.*, 2007).

A parte del ejemplo comentado, los algoritmos de ML han sido entrenados para identificar sitios de *splicing,* promotores, *enhancers,* la posición de los nucleosomas, anotación de elementos genéticos, etc. Además de los patrones reconocidos en las secuencias de DNA, ML puede funcionar con una entrada de datos generados por otras plataformas, tales como datos de expresión de *microarrays* o RNAseq, ensayos de accesibilidad de la cromatina como DNase-seq, MNase-seq y FAIRE, o datos ChIP-seq. En cuanto a los datos de expresión, éstos pueden ser usados, por ejemplo, para la distinción de fenotipos o la identificación de potentes biomarcadores dentro de la enfermedad. Un uso bastante extenso de ML dentro del campo de la biomedicina no se corresponde al nivel genómico, sino al reconocimiento de patrones en imágenes médicas (radiográficas, TACs, resonancias, etc.), al *clustering* de pacientes con dolencias comunes o la clasificación en el diagnóstico médico (Zhang and Rajapakse, 2008), (Libbrecht and Noble, 2015).

Todos los ejemplos nombrados aportan una idea del potencial que puede llegar a tener ML en el campo de la medicina y la investigación biológica. Aunque para cada problema existe un tipo de específico de algoritmo que utilizar. A continuación se hace una breve explicación de los principales tipos de aprendizaje automático.

Existen dos tipos principales de aprendizaje automático: el aprendizaje supervisado y el aprendizaje no supervisado.

- **Aprendizaje supervisado**: en este caso se dispone de *x* variables de entrada y una variable *Y* de salida, por lo que se usa una función que aprenda a clasificar o mapear la entrada en la salida.

EQ 11

$$Y = f(x)$$

El objetivo es encontrar una función que aplicada a unos nuevos datos (*x*), los clasifique en las clases predeterminadas de *(Y)*.

Los datos de entrada se encuentran ya clasificados, por lo que el algoritmo genera predicciones sobre los datos de entrenamiento de forma iterada. Se detiene el funcionamiento del algoritmo cuando se obtiene un nivel aceptable de rendimiento.

Cuando la variable de salida es categórica, es decir, está divida en distintas clases (*p.ej.,* positivo/negativo, enfermo/sano) son los denominados problemas de clasificación. Cuando esta variable es un valor real (p.ej., peso, tamaño) se denominan problemas de regresión.

Entre los algoritmos de aprendizaje supervisado, los más utilizados son: *linear regression* (para regresión), *Random Forest* (para ambos problemas) y *Support vector machines* (para clasificación).

- **Aprendizaje no supervisado**: en este tipo de aprendizaje se tiene un conjunto de *N* observaciones $(x_1, x_2, \dots, x_N)$ de un vector aleatorio *X* con función de probabilidad Pr(X). El objetivo es inferir directamente las propiedades de esta función de densidad sin la ayuda de un supervisor, generando las respuestas correctas o el grado de error de cada observación. Por lo que en estos problemas no se suministran categorías inicialmente, es decir, no es una tarea de clasificación sino de organización para descubrir la estructura de los datos para visualizarlos o comprimirlos mejor. Se trata de sacar algún significado a los datos, no conocido anteriormente.

  Dentro de este tipo de algoritmos, se pueden diferenciar dos categorías: *clustering y* asociación. Un algoritmo de *clustering* pretende discernir la agrupación o grupos inherentes de los datos (p.ej., agrupar los clientes según su comportamiento de compra). Los de asociación pretenden

descubrir las reglas que describen la mayoría de los datos (p.ej. los pacientes que poseen dicho síntoma X también tienden a presentar otro síntoma Y).

Los principales algoritmos de aprendizaje no supervisado son *k-means* (para *clustering*) y algoritmos *a priori* (para asociación).

En este trabajo se han utilizado dos algoritmos de clasificación. Es decir, son algoritmos basados en un modelo predictivo. Los modelo predictivos necesitan instrucciones claras sobre lo que ellos necesitan aprender y como tienen planeado aprenderlo. El proceso de entrenamiento es supervisado. En la ***Figura 7*** se muestra un esquema sencillo de cómo funciona un algoritmo de clasificación.

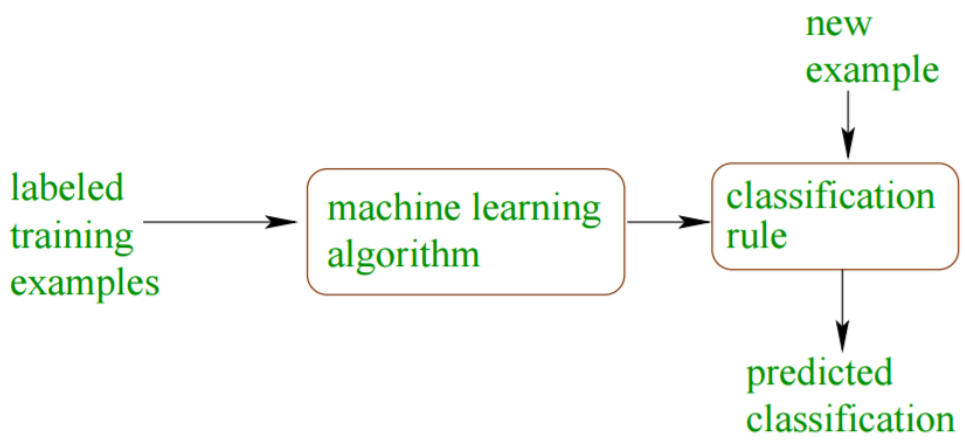


Figura 7: Esquema del funcionamiento de un algoritmo de clasificación basado en ML

El objetivo de este apartado del proyecto fue el entrenamiento de algoritmos basados en ML que permitan la clasificación de los datos RNAseq entre los dos grupos de pacientes a estudio. La matriz de datos RNAseq fue usada para entrenar un modelo Random Forest y un modelo PamR que clasifiquen a los pacientes en *status* negativo o *status* positivo. Refiriéndose al estado del receptor de estrógeno.

HERRAMIENTAS NECESARIAS PARA EL EXPERIMENTO DE ML

- **Paquete de R**: para la aplicación de todas las técnicas de ML necesarias en este proyecto se utilizó un paquete implementado en R, denominado *mlr.* En la página web se pueden observar todas las funcionalidades que te permite este paquete (https://mlr-org.github.io/mlr-tutorial/release/html/index.html). A modo de resumen, el paquete *mlr* aporta una infraestructura genérica, orientada a objetos y extensible para clasificación, regresión, análisis de supervivencia y *clustering* en lenguaje R. Genera una ventaja a la hora de escribir código, ya que ofrece un lenguaje limpio, fácil de usar y flexible para este tipo de experimentos. Está compuesto por una interfaz de más de 160 algoritmos básicos (denominados *learners)* e incluye meta-algoritmos y técnicas de selección de modelos para mejorar y extender la funcionalidad de las bases de los *learners (ej. Hyperparameter tuning, feature selection y ensemble construction)* (Bischl *et al.*, 2016).

  De forma sencilla, para realizar un experimento con este paquete es necesario previamente fijar y definir algunos objetos. En primero lugar es necesario definir las *tasks* con las que vas a operar, es decir los datasets que se usarán en el experimento. En segundo lugar es necesario definir el *learner*, es decir, el algoritmo de ML que se usará. Finalmente, para obtener una evaluación del modelo, es necesario definir el tipo de remuestreo que hará de los datos.

- ***Benchmarking**:* En un experimento *Benchamarking* diferentes métodos de aprendizaje son aplicados a una o varias bases de datos para comparar y hacer un *ranking* de los algoritmos respecto a una o más medidas de rendimiento. Como ventaja se tiene que todos los algoritmos son entrenados bajo las mismas condiciones. Además, la posibilidad de poder entrenar los algoritmos sobre una gran cantidad de *datasets* distintos permite una posterior comparación, así como una facilidad del diseño del experimento.

4.2. Metodología seguida para los experimentos basados en ML

Para la ejecución de cualquier experimento basado en ML es necesario primeramente definir la estrategia y la metodología usada. Este proyecto se ha apoyado en los pasos seguidos en la metodología utilizada en el trabajo de (Fernandez-Lozano, Gestal, Cristian R Munteanu, *et al.*, 2016), aunque para adecuarla a los objetivos principales, se cambiaron algunos procedimientos. La metodología que se seguirá durante todo el capítulo restante se refleja en la ***Figura 8***.

Figura 8: Metodología seguida para los experimentos basados en Machine Learning

4.2.1. Obtención de los datos

Como se indicó anteriormente los datos fueron descargados del repositorio TCGA. Son datos de expresión genética obtenidos a partir de la técnica de RNA-seq (explicado en el apartado 2.3). Los datos genéticos pertenecen a 757 pacientes que refieren cáncer de mama. Para más información de los datos (variables, datos fenotípicos…), consulten la página web del TGCA (https://www.cancer.gov/).

4.2.2. Pre-procesado de los datos

La mayor parte del procesado se realizó de la misma manera que para las aproximaciones anteriores. A modo de recordatorio, mencionar que se obtienen datos genéticos (18257 genes) de 721 individuos, tras la eliminación de genes contaminados o no pertenecientes a humanos gracias a los datos de anotación de *ensembl63.* Además, se eliminaron pacientes que no tenían un *status del receptor de estrógenos* ni positivo ni negativo.

No se detectó ningún valor NA *(Not Available).* Posteriormente se realizó el estudio de las variables NZV (*near zero variance).* En la ***Figura 4*** se representa la expresión media de cada gen para los 721 pacientes analizados, así como la expresión media después de cada tipo de normalización. Se detectaron 770 genes que tenían una expresión media inferior a la unidad. De todas formas, se decidió mantenerlos en el estudio para obtener una mayor representatividad de los datos.

A dicha matriz se le hizo pasar por tres métodos de normalización: RPKM, CQN y TMM. También explicados anteriormente. Por lo tanto, se obtuvieron 3 matrices de igual dimensiones (18257 x 721) pero distintamente normalizadas.

Además, se obtuvo un subconjunto de genes mediante la búsqueda en bases de datos públicas de genes relacionados a cáncer de mama. Se crearon dos listados, uno a partir de la web Intogen y otra a partir de Wikipathways. El total de genes relacionados con dicho cáncer fue de 308. Realizando la intersección de este listado de genes con las tres matrices de datos normalizados, se

obtuvieron a mayores, otras tres matrices con dimensiones 301 x 721. Se observa por lo tanto, que 7 genes obtenidos mediante las bases de datos públicas no estaban representados en nuestro estudio de RNAseq.

Se añadió la variable categórica que clasifica los pacientes en negativos o positivos según el *status del receptor de estrógenos.* Recuerden que este factor estaba compuesto por 555 pacientes positivos y 166 pacientes negativos.

Finalmente, tras el procesado de los datos, se obtuvieron un total de 6 matrices. Tres de ellas con la representación de todos los genes a estudio (18257), mientras que las otras tres, únicamente con los genes relacionados con cáncer de mama. Estas seis matrices serán la base para los experimentos realizados mediante ML.

4.2.3. Aprendizaje del modelo

Este es el paso más importante en inteligencia computacional. En primer lugar, es necesario un modelo de referencia. En este caso nuestro modelo fue sacado de la base de datos TGCA como ya se ha comentado. Los 721 pacientes presentes en la base de datos estaban clasificados según el *status* del receptor de estrógenos (positivo / negativo). Posteriormente se realizaron diversos experimentos. En primer lugar una reducción de la dimensionalidad mediante técnicas de FS. Luego, un entrenamiento del modelo y su evaluación mediante CV.

4.2.3.1. Feature Selection

En cualquier experimento basado en ML el investigador debe decidir cuáles son los datos que deben introducirse como entrada al algoritmo. Por lo tanto, este paso conlleva la necesidad de un conocimiento previo de los datos, para proporcionar sólo las variables que tengan una mayor relevancia. Este proceso normalmente es complicado de realizar, debido a la gran dimensionalidad de los datos y al desconocimiento que se tiene de ellos (Guyon, Elisseeff and De, 2003).

Para solucionar dicha complicación, se desarrollaron diferentes métodos estadísticos multivariantes para la reducción de la dimensionalidad del problema. Por ejemplo, los análisis de componentes principales (PCA) y el análisis discriminante lineal (LDA). Sin embargo, para este proyecto se ha decidido utilizar la técnica de FS ya que, en contraposición a las mencionadas anteriormente, ésta no alteran la representación original de las variables, si no que únicamente seleccionan un subgrupo de ellas. Este hecho es importante a la hora de trabajar con datos biológicos (Liu, Li and Wong, 2002).

Existen tres motivaciones distintas para realizar un proceso de selección de variables. En primer lugar si se quiere identificar un pequeño número de características que produzcan la mejor clasificación posible. En segundo lugar si se pretende usar el clasificador para entender las propiedades subyacentes de los datos. Y en tercer lugar, si el objetivo únicamente es entrenar el clasificador con la mayor exactitud posible (Liu, Li and Wong, 2002), (Degroeve *et al.*, 2002).. El enfoque seguido en este trabajo ha sido el del primer punto. En este caso, se ha querido obtener un subconjunto de *features/genes* que presenten un rendimiento en la clasificación igual o mayor que al usar todas las variables.

Los objetivos finales de las técnicas de FS, son: impedir el sobre-ajuste y mejorar el rendimiento del modelo, generar modelos más rápidos y más rentables y obtener una profundidad de información sobre los procesos subyacentes que generaron los datos. Estas ventajas generan una capa de complejidad adicional al modelo (Yahya *et al.*, 2011). En la ***Figura 9*** se presenta el método general para la selección de variables.

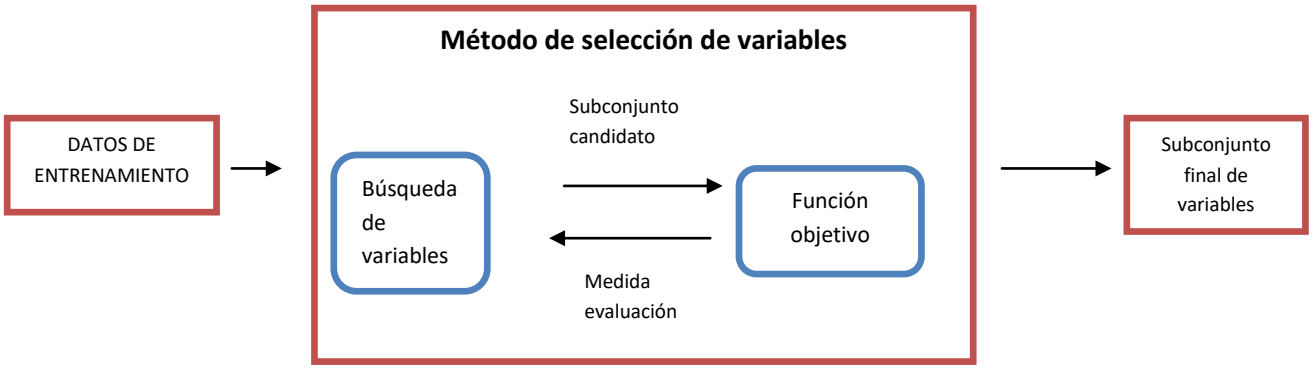


Figura 9: Esquema general de funcionamiento de un método de selección de variables.

Dentro del contexto de clasificación, las técnicas de FS se pueden organizar en tres categorías, representadas en la ***Figura 10,*** dependiendo de cómo combinan la búsqueda de características con la construcción del modelo de clasificación: *Filter, Wrapper, Embedded* (Saeys, Inza and Larrañaga, 2007).

- Los métodos de filtrado evalúan la relevancia de las características únicamente centrándose en las propiedades intrínsecas de los datos. En la mayoría de los casos se calcula un valor de la relevancia de las variables mediante test estadísticos como *T.test* o *Wilcoxon.test* y los que tengan una menor puntuación son eliminados. Finalmente, este conjunto de features se presenta como los datos de entrada en el algoritmo de clasificación. Existen varias ventajas relacionadas a este tipo de técnica. Tienen la capacidad de escalar fácilmente los datos de alta dimensión, son rápidos y simples computacionalmente y además son independientes del algoritmo de clasificación, es decir, que una vez generado el subconjunto de *features* pueden ser utilizados como entrada para cualquier tipo de algoritmo de clasificación. Por contra, ignoran la interacción con el clasificador, ya que la búsqueda en el subespacio de features está separado de la búsqueda en el espacio de hipótesis, es decir, cada feature es considerada por separado, por lo tanto ignoran las dependencias que puedan existir entre las variables, lo que puede llevar a una peor clasificación en comparación con las otra técnicas. Para solucionar este problema, se han generado técnicas de filtrado multivariante para integrar las dependencias en el modelo.
- En cuanto a los métodos *wrapped,* estos incorporan el modelo de la búsqueda de hipótesis dentro de la búsqueda del subconjunto de *features*. De esta forma, se define el procedimiento de búsqueda en el espacio del subconjunto de *features* posibles, y se generan varios subconjuntos de *features* para evaluarlos. La evaluación de un subconjunto específico de *features* se obtiene mediante entrenamiento y *testing* de un modelo de clasificación específico. Para buscar el espacio de todos los subconjuntos de *features*, un algoritmo de búsqueda es entonces 'envuelto' (*wrapped* en inglés) alrededor del modelo de clasificación. Sin embargo como el espacio de búsqueda de los

subconjuntos de *features* crece exponencialmente con el número de variables, se usan métodos de búsqueda heurísticos para guiar la búsqueda a un subconjunto óptimo. Las ventajas que tiene este método es la interacción entre la búsqueda del subconjunto de *features* y la selección del modelo, y tener en cuenta la dependencia de las variables. Como desventajas, existe un gran riesgo de sobrentrenamiento (*overfitting)* además de que computacionalmente es muy intenso.

- Por último, las técnicas *embedded* realiza la búsqueda del subconjunto óptimo de *features* dentro de la construcción del clasificador. Se considera una búsqueda en un espacio combinado sobre el subconjunto de *features* e hipótesis. Como el método anterior, es específico del algoritmo de aprendizaje. Como ventajas se considera la interacción con el modelo de casficación y además es computacionalmente menos intenso que el método *wrapped.*

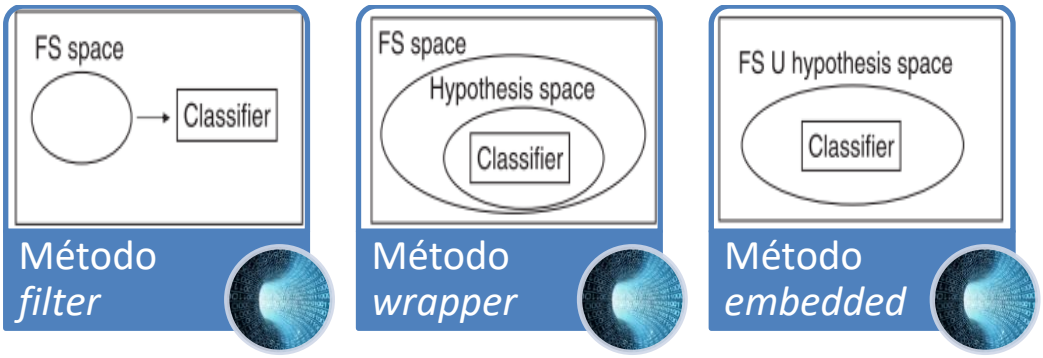


Figura 10: Representación de las tres grandes clases de FS en problemas de clasificación (Saeys, Inza and Larrañaga, 2007).

Tras repasar el funcionamiento de los tres métodos de FS, para este trabajo se ha decidido usar la técnica de filtrado univariante, que aunque parezca a priori la menos poderosa, posee grandes ventajas debido a la alta dimensionalidad del problema, al ser rápida y eficiente. Además, la salida obtenida es fácil, intuitiva y cumple los objetivos dentro del campo biológico para validar posteriormente los resultados mediante técnicas de laboratorio, o con el fin de explorar búsquedas bibliográficas.

El diseño del experimento seguido en las técnicas de FS fue sencillo para este proyecto. Se realizaron técnicas de FS para dos experimentos de ML.

El primero fue una selección de *features* de entre los 18257 genes originales a sub-conjuntos de tamaños de: 50, 100, 200, 250, 300, 350 *features.* El objetivo es identificar los genes que tengan una mayor correlación el subtipo ER.

El segundo conjunto de datos a los que se sometieron a técnicas de FS fueron las bases de datos que presentan únicamente los genes relacionados a cáncer de mama. De entre los 301 genes, se realizaron sub-conjuntos de tamaños: 50, 100, 200, 250, 300, 350. El objetivo es la búsqueda de genes que mejor clasifiquen a los pacientes, y su posterior comparación con los 116 genes DE.

4.2.3.2. Algoritmos de Machine Learning

Existen numerosos algoritmos de clasificación basados en ML. En este proyecto se ha decidido utilizar concretamente dos de los más reconocidos en el estado del arte por sus buenos resultados en general y con datos genéticos en particular: Random Forest y pamR.

4.2.3.2.1. Random Forest

Antes de describir el funcionamiento del algoritmo *Random Forest* es necesario comprender bien lo que es un algoritmo basado en árboles de decisión. Ya que se podría considerar *Random Forest* como un conjunto de árboles de decisión.

En primer lugar, los árboles de decisión son algoritmos supervisados que se usan principalmente para problemas de clasificación. Este tipo de algoritmo clasifica los elementos de los datos planteando una serie de preguntas sobre las características asociadas a los elementos. Para visualizar el concepto de manera más sencilla, imagínese el lector la raíz de un árbol real. En este caso la raíz representa todos los datos. A medida que el árbol crece, el tallo se va ramificando en ramas cada vez más pequeñas. Finalmente, en la copa del árbol existen una gran cantidad de pequeñas raíces, pero todas más o menos del mismo tamaño. Cada una de las ramas finales es un subconjunto pequeño de los datos etiquetados en alguna clase.

Los árboles de decisión están formados por nodos, vectores, flechas y etiquetas. Cada pregunta (etiqueta) está presente en un nodo, y cada nodo interno apunta a un nodo hijo para cada posible respuesta a su pregunta. Las preguntas poseen una jerarquía, codificada como un árbol. El tipo de preguntas con las que se ramifican los datos pueden ser categóricas, numéricas o probabilísticas (Kingsford and Salzberg, 2008).

Aunque la utilización de árboles de decisión simples pueden llegar a ser excelentes clasificadores, el aumento de la exactitud del modelo puede conseguirse como resultado de una colección de árboles de decisión. *Random Forest* es considerado como uno de los mejores modelos para clasificación.

Random Forest fue desarrollado por (Breiman, 2001), y trata de la ejecución de varios árboles de decisión diferentes. De esta manera, es posible combinar las predicciones del conjunto resultante de árboles de decisión y tomar la predicción más común. El mantenimiento de un conjunto de buenas hipótesis, en lugar de comprometer a un solo árbol, reduce la probabilidad de que un nuevo ejemplo sea clasificado erróneamente al ser asignado en la clase equivocada por muchos árboles. El proceso de remuestreo de los datos y el promedio de los distintos modelos conjuntamente se denomina *bagging* conduciendo a una mitigación tanto de los modelos con sesgo como los modelos con alta varianza (Svetnik *et al.*, 2003).

Fue elegido para este trabajo ya que es un modelo de uso general que realiza bien la mayoría de los problemas. Es bueno manejando datos ruidosos o no disponibles. Selecciona únicamente las características más importantes. Y sobre todo, porque se puede utilizar en datos con un gran número de funciones y ejemplos (Diaz-Uriarte and Alvarez de Andres, 2006).

4.2.3.2.2. pamR

El algoritmo PamR es un paquete de R que genera funciones para la clasificación de datos de expresión de genes por el método *nearest shrunken centroids (NSC).* El método fue desarrollado por (Tibshirani *et al.*, 2002) de la

universidad de Stanford. Además de clasificar los datos realiza un tipo de FS interna.

Se define x_{ij} como la expresión de los genes *i* = 1, 2,…*p* y las muestras *j*= 1, 2,… *n*. Se presentan las clases 1, 2,…*K*, y se considera a C_k como el índice de la muestra n_k en la clase *k*. El iésimo componente del centroide para la clase *k* es $\bar{x}_{ik} = \sum_{j \in C_k} {}^{x_{ij}}/_{n_k}$, el valor de expresión media in la clase *k* para el gen *i*; el iésimo componente de la media general del centroide es $\bar{x}_i = \sum_{j=1}^{n} {}^{x_{ij}}/_{n}$ (Tibshirani *et al.*, 2002).

Es decir, el algoritmo pretende encoger (*shrink,* en inglés) los centroides clasificados hacia los centroides generales después de una estandarización por la desviación típica dentro de cada clase para cada gen. El efecto de esta estandarización da pesos altos a los genes que tienen una expresión estable dentro de las muestras de la misma clase. Dicha estandarización es inherente a otros métodos estadísticos comunes, tales como el análisis discriminante lineal (Tibshirani *et al.*, 2002).

Se define,

EQ 12

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k * (s_i + s_0)}$$

Donde s_i es la desviación estándar dentro de la clase agrupada para el gen *i*:

EQ 13

$$s_i^2 = \frac{1}{n-K} \sum_k \sum_{j \in C_k} (x_{ij} - \bar{\bar{x}}_{ik})^2$$

Y $m_k = \sqrt{{}^{1}/_{n_k} + {}^{1}/_{n}}$ hace que el m_k*s_i sea igual al error estándar estimado del numerador en d_{ik}. En el denominador, el valor s_0 es una constante positiva (con el mismo valor para todos los genes), incluida para contrarrestar la posibilidad de que valores d_{ik} surjan por casualidad a partir de genes con bajos niveles de

expresión. Se pone s_0 igual al valor mediano del s_i sobre el conjunto de genes. De esta forma, d_{ik} es una estadística *t* para el gen *i*, comparando la clase *k* al centroide general. Se reescribe la Eq. 12 como:

EQ 14

$$\bar{x}_{ik} = \bar{x}_i + m_k(s_i + s_0)d_{ik}$$

El método encoge cada d_{ik} hacia cero, proporcionando un d'_{ik} y produciendo centroides reducidos o prototipos (Tibshirani *et al.*, 2002).

EQ 15

$$\bar{x}'_{ik} = \bar{x}_i + m_k(s_i + s_0)d'_{ik}$$

La contracción realizada se denomina umbral suave (*soft thresholding*, en inglés): cada d_{ik} se reduce por una cantidad Δ en un valor absoluto y se declara cero si su valor absoluto es menor que cero. Algebraicamente el *soft thresholding* se define por:

EQ 16

$$d'_{ik} = sign(d_{ik})(|d_{ik}| - \Delta)_+$$

Debido a que muchos de los valores de $\bar{x}_{ik}$ serán ruidosos y cerca de la media general $\bar{x}_i$, el umbral suave generalmente produce estimaciones más fiables de las medias verdaderas (Tibshirani *et al.*, 2002).

Este método elimina muchos de los genes de las clases predictoras a medida que el parámetro Δ incrementa. Específicamente, si para un gen *i*, d_{ik} es contraído a cero para todas las clases *k*, entonces el centroide para el gen *i* es $\bar{x}_i$, el mismo para todas las clases. Por lo tanto, el gen *i* no contribuye en el cálculo de los centroides contraídos. La elección de Δ se realiza mediante cross-validation.

NSC es uno de los métodos de clasificación más usados para datos de alta dimensionalidad, tal como datos de *microarray* o datos procedentes de NGS. De forma más resumida y más clara, el método consiste en el cálculo de un centroide estandarizado para cada clase. La clasificación centroide más cercana toma el perfil de expresión génica de una nueva muestra y la compara

con cada uno de estos centroides de clase. La clase cuyo centroide de mayor proximidad, es la clase prevista para esa nueva muestra. Se “contrae” cada uno de los centroides de clase hacia el centroide general para todas las clases en una cantidad que se llama el umbral. Después de encoger los centroides, la nueva muestra se clasifica por la regla del centroide más cercano, pero usando los centroides de la clase reducida (Pardo and Sberveglieri, 2008).

Esta reducción tiene dos ventajas: 1) puede hacer el clasificador más exacto reduciendo el efecto de los genes ruidosos, 2) hace la selección automática del gen (Tibshirani *et al.*, 2003).

4.2.3.3. Cross-validation

CV es un conjunto de técnicas usadas para la evaluación y selección del modelo. El principal objetivo es compensar el ratio de error, aparentemente optimista. Dicho ratio es el porcentaje de las observaciones mal clasificadas. Este número tiende a ser optimista porque los datos que son clasificados son los mismos datos que se usan para entrenar el modelo. Al tratarse de datos conocidos por el modelo, se produce un resultado optimista con respecto al que se obtendría con datos totalmente desconocidos.

La idea principal detrás de CV es separar los datos una o varias veces, para estimar el riesgo de cada algoritmo. Parte de los datos (muestra de entrenamiento) son usados para el entrenamiento del algoritmo, mientras que los datos restantes (muestra de validación) son usados para estimar el riesgo de cada algoritmo. Esta parte depende del tipo de CV utilizado. Por ejemplo, si se usa un 10 *fold* los datos son divididos en diez partes proporcionales. Cada una de estas partes va a ser usado como test y los nueve restantes servirán como grupo *train.* Cuando cada parte ha sido utilizada como test, el modelo de CV selecciona el algoritmo con el menor riesgo de estimación. Debido a que la muestra de entrenamiento y la muestra de validación son independientes, CV evita el sobreentreamiento del modelo (Syed, 2011).

Existen dos objetivos principales cuando CV es usado en el proceso del aprendizaje del modelo. Primero, para medir el grado de generalización del

modelo durante la fase de entrenamiento, evaluando el rendimiento del modelo y estimando su rendimiento con datos desconocidos. Segundo, comparar dos o más algoritmos entrenados bajo las mismas condiciones y con los mismos *datasets.* Para dicha comparación es necesaria la obtención de una medida que represente la exactitud y el rendimiento del modelo. Existen medidas bien conocidas como la *accuracy* (acc), o la medida F, aunque para este trabajo se decidió usar el área bajo la curva ROC (AUC, por sus siglas en inglés).

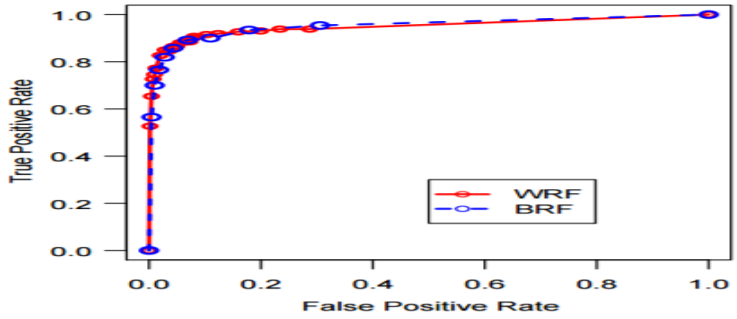


Figura 11: Gráfica que representa una curva ROC. Todo el área debajo de cada curva se corresponde con la medida AUC, usada en este proyecto para evaluar el rendimiento de los modelos (Chen, Liaw and Breiman, 2004)***.***

La curva ROC (*Receiver Operating Characteristics)* es una representación del ratio de verdaderos positivos frente al ratio de falsos positivos. Por lo que la curva ROC se puede usar para generar estadísticos que resumen la efectividad del clasificador. Entre ellos se encuentra el índice AUC, interpretado como la probabilidad de que una instancia aleatoria y uniformemente representado se clasifique antes que una instancia negativa aleatoria y uniformemente representado. En la ***Figura 11*** se muestra un ejemplo de curva ROC y el índice AUC. Se considera un valor AUC mayor a 0.75 como 'bueno', y un valor 0.9 'excelente'.

Una parte interesante de este estudio fue el uso de un remuestreo anidado. En primer lugar se realizó un CV interior (*inner)*, únicamente para seleccionar la mejor combinación de parámetros para cada algoritmo (también denominado *tunning)*. Posteriormente se realizó un CV externo (*outer)* para asegurarnos de que los modelos no fueron sobre-entrenados y que el mejor valor no fue encontrado por una combinación particular de los parámetros/observaciones y

un número elevado de repeticiones. Estos dos tipos de CV se explicarán con más detalle a continuación.

4.2.3.3.1. Cross-validation interno para la selección del mejor conjunto de hiperparámetros

El CV interno es necesario para escoger los parámetros más adecuados de los algoritmos. Una vez elegido el tipo de CV, el modelo se entrenará con todos los posibles parámetros hasta dar con la mejor combinación de ellos, siempre sobre los mismos datos. El que mejor resultado obtenga, será el utilizado en el CV externo. Dicha estrategia, para seleccionar los hiperparámetros de cada algoritmo también se denomina *tuning.*

En este proyecto se utilizó un *holdout*. Éste método es el tipo más simple de *cross-validation.* Los datos son separados en dos conjuntos, llamados *train* y *test*. La función de aproximación únicamente usa el conjunto de *train.* Posteriormente se le pide a la función que haga una predicción sobre los valores de salida de los datos del conjunto de pruebas. Los errores cometidos son acumulados para dar la media absoluta de los errores en el conjunto del *test,* lo cual es usado para evaluar el modelo.

4.2.3.3.1.1. Selección de los parámetros para Random Forest

Los parámetros sometidos a *tunning* del algoritmo Random Forest fueron ‘mtry’, ‘ntree’ y ‘nodesize’. En cuanto al parámetro ‘mtry’ se refiere al número de variables aleatoriamente muestreadas en cada división de los datos. El rango de parámetros elegidos fue desde 5 hasta 13. En relación al parámetro ‘ntree’, éste se fijo en un valor de 1000. Por lo tanto, el número de árboles de decisión aleatorios estará fijado en 1000. De esta manera, el concepto de *bagging* se realizará de una manera significativa. Por último, el parámetro ‘nodesize’ se sometió a un intervalo entre 1 y 5. Este parámetro representa el tamaño mínimo de los nodos terminales, es decir, la profundidad de los árboles. Un

valor bajo genera un gran crecimiento del árbol, mejorando de esta manera la exactitud de las predicciones.

4.2.3.3.1.2. Selección de los parámetros para pamR

En cuanto al algoritmo de pamR, los parámetros sometidos a *tunning* fueron "threshold" y "threshold.scale". Se seleccionaron los mejores rangos de valores posibles utilizando el uso de pamR para ello. Esto permitió buscar el conjunto de X parámetros de *threshold* y de Y parámetros de *threshold.scale* que mejor resultados obtenía. Todos los valores estaban en el intervalo entre 0.05 y 2. Los siguientes pasos del experimento fueron iguales al *inner* del Random Forest.

4.2.3.3.2. Cross-validation exterior

La utilización de un modelo nivel de validación externo sirve para evaluar el rendimiento general del modelo.

Se optó ejecutar un 5 *Repeated 10-fold cross-validation.* Este tipo de CV divide aleatoriamente los datos en *k* bloques de aproximadamente el mismo tamaño. Cada uno de estos bloques se deja de lado como *test* y los *k-1* bloques restantes se usan como *train.* El bloque retenido se predice y estas predicciones se resumen en algún tipo de medida de rendimiento (por ejemplo, precisión, error cuadrático medio de la raíz (RMSE), área bajo la curva ROC, etc.). Las *k* estimaciones de rendimiento se promedian para obtener la estimación global remuestreada. Los datos para este trabajo por lo tanto se dividieron en 10 bloques, y se repitió la validación cruzada 5 veces, obteniendo un total de 50 remuestreos para cada base de datos.

4.2.4. Selección del mejor modelo

En los anteriores apartados se ha evaluado el rendimiento de varios modelos sobre distintos *datasets*. Con los resultados obtenidos mediante esas técnicas y con el objetivo de determinar si el rendimiento de un algoritmo es estadísticamente mejor que los otros, es necesario realizar algún tipo de test basado en hipótesis nulas.

Para el contraste de hipótesis se pueden usar tanto test paramétricos como no paramétricos. Para la utilización de los primeros, es necesario comprobar algunas características de los datos, tales como: normalidad, independencia y heterocedasticidad (García *et al.*, 2010). Estas tres características no hacen referencia a la distribución de los datos de entrada, si no a la distribución del rendimiento de los datos usados como entrada en los diferentes experimentos.

En cuanto a la independencia, ésta se refiere, en términos estadísticos, que la alteración de una variable dentro de un determinado modelo no afecta a la probabilidad de las otras variables. En problemas de ML, los datos usados para el entrenamiento de cada algoritmo son separados aleatoriamente mediante el proceso de CV, con lo que se cumple la condición de independencia.

Una distribución normal es cuando la distribución de frecuencia de sus valores describe una curva con forma de campana que es simétrica respecto a su media (Ghasemi and Zahediasl, 2012). A la hora de comprobar esta condición existen diferentes test estadísticos, entre ellos, el más conocido y utilizado es el test de Shapiro-Wilk (Shapiro and Wilk, 1965).

Finalmente, la violación de la hipótesis de igualdad de varianzas, o heterocedasticidad debe ser comprobada usando, por ejemplo, un test de Levene (Levene, 1960).

Si las condiciones de independencia, heterocedasticidad y normalidad son contrastadas, se requieren test paramétricos como el T.test o el ANOVA. Por el contrario, si estas asunciones no son validadas, es necesario el uso de tests no paramétricos, como son el test de Wilcoxon o Friedman (García *et al.*, 2010).

Aunque los test paramétricos son más potentes a priori, no deben ser usados cuando las tres condiciones no son conocidas totalmente. En ese caso, es mejor emplear un test no paramétrico, diseñado específicamente para dichas condiciones, y además los resultados obtenidos serán más precisos y ajustados a las características intrínsecas de los datos (Fernandez-Lozano, Gestal, Cristian R. Munteanu, *et al.*, 2016). Por lo tanto, para este trabajo se han utilizado test no paramétricos debido al desconocimiento de la distribución de los *datasets* usados en el enteramiento de los algoritmos. La elección de un test respecto a otro radica en el número de comparaciones que se hagan, como se observa en la ***Tabla 5***. Para cada experimento realizado en ML la comprobación entre modelos se reduce a dos (Random Forest y pamR). Es por ello que se optó por test de Wilcoxon (Fernandez-Lozano, Gestal, Cristian R. Munteanu, *et al.*, 2016).

Tabla 5: Clasificación de los test paramétricos y no paramétricos según el número de modelos comparados.

Número de modelos	Test paramétrico	Test no paramétrico
n=2	T-test	Wilcoxon
n>2	ANOVA test	Friedman

Una vez la hipótesis nula se rechace mediante los test estadísticos, es conveniente un procedimiento *post hoc* para realizar la comparación con ajuste múltiple y hacer una corrección al p-valor obtenido (García *et al.*, 2010).

Obtenido el p-valor ajustado se pueden dar dos posibilidades. Que no exista significancia estadística, o que si exista significancia estadística. En el primer caso la elección del mejor modelo ser realizaría según la simplicidad o tiempo de computación, dentro de los modelos ganadores de la prueba de hipótesis nula. En la ***Figura 12*** se puede observar el *workflow* planteado en el artículo de (Fernandez-Lozano, Gestal, Cristian R Munteanu, *et al.*, 2016) y que se utilizará para este proyecto.

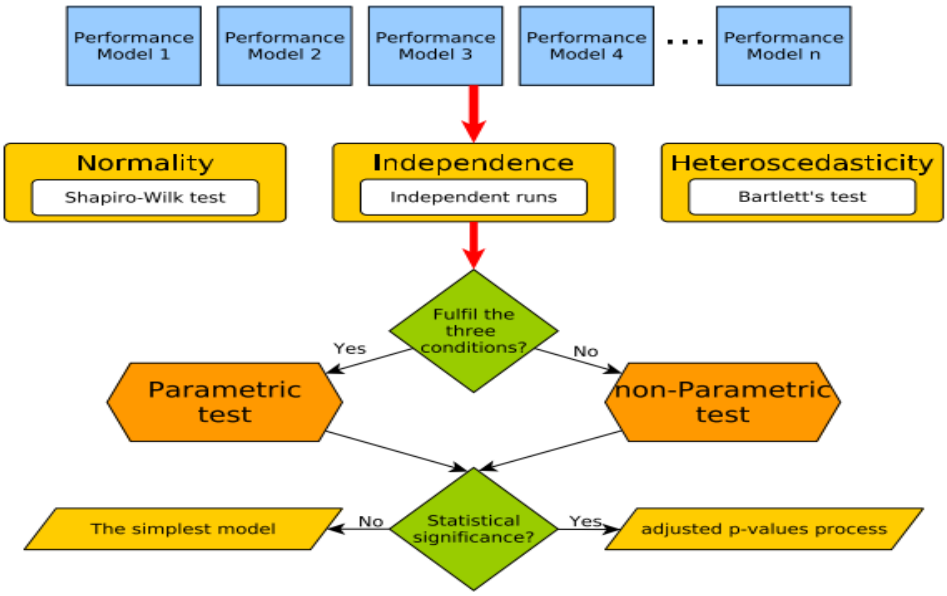


Figura 12: Representación de la metodología a seguir para la selección del mejor modelo. Figura extraída del artículo de (Fernandez-Lozano, Gestal, Cristian R Munteanu, *et al.*, 2016)***.***

5. PRUEBAS Y RESULTADOS

Se realizaron 3 experimentos basados en ML y siguiendo la metodología anteriormente explicada. El objetivo común de los tres experimentos fue la clasificación de los pacientes en los grupos pertenecientes a la variable categórica (*status del receptor de estrógenos).* Además, cada experimento tendrá de partida un tipo diferente de matrices de datos y para cada una de ellas se les podrá someter, o no, a técnicas de FS. A continuación se describirán los experimentos así como los resultados obtenidos en cada uno.

5.1. Experimento 1 basado en Machine Learning

Para este experimento se han entrenado los dos modelos elegidos (Random Forest y pamR) mediante las tres bases de datos (cada una con un tipo de normalización) que presentan únicamente los 301 genes obtenidos a partir de Wikipathways e Intogen. De esta manera se obtuvo una medida (AUC) del rendimiento de la clasificación. El valor AUC de este primer experimento proporciona una aproximación en cuanto a la calidad de los 301 genes para la clasificación de los pacientes en base a su estatus del receptor de estrógenos. Un valor alto será indicador de que el análisis de estos genes podría ser suficiente para la categorización de este tipo de pacientes. Además, se podrá inferir cual fue el tipo de normalización y el algoritmo que mejor resultado reportó.

En la ***Figura 13*** se mapea en un *scatterplot* el valor del rendimiento de los modelos según el tipo de normalización. Se observa un alto resultado en los tres tipos de normalización, aunque en RPKM se observa un mejor rendimiento, llegando a alcanzar el valor AUC=0.959. Para las tres bases de datos Random Forest fue el modelo que mejor rendimiento obtuvo. Las distribuciones de los resultados de cada modelo, representados en la ***Figura 14***, refuerza la hipótesis de que el mejor modelo es Random Forest, ya que presenta mayor uniformidad de los resultados y unas colas más cortas que pamR.

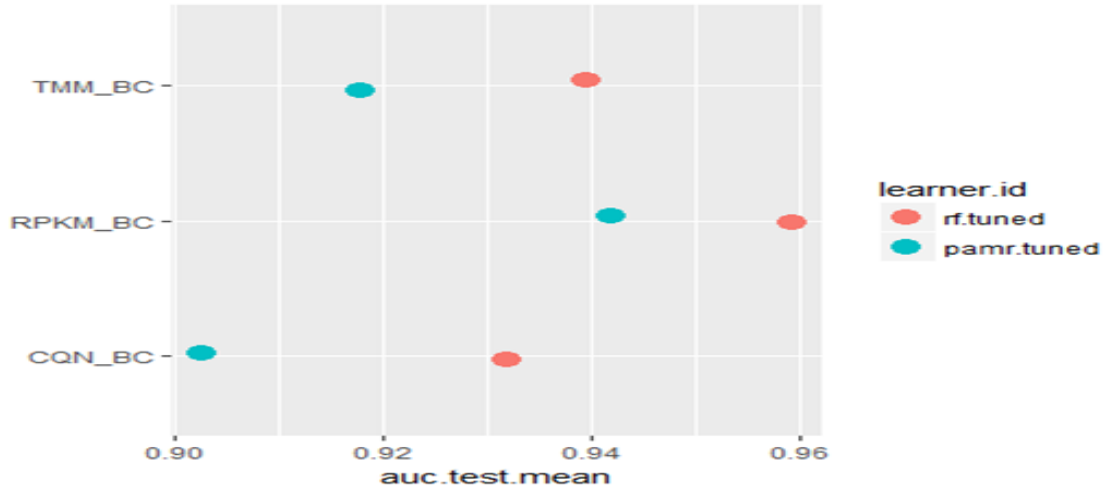


Figura 13: Gráfica de los resultados obtenidos mediante el experimento benchmark. El eje vertical representa a los datasets con los 301 genes de la Tabla 1 y con el tipo específico de normalización. Los resultados se dan en valor de AUC.

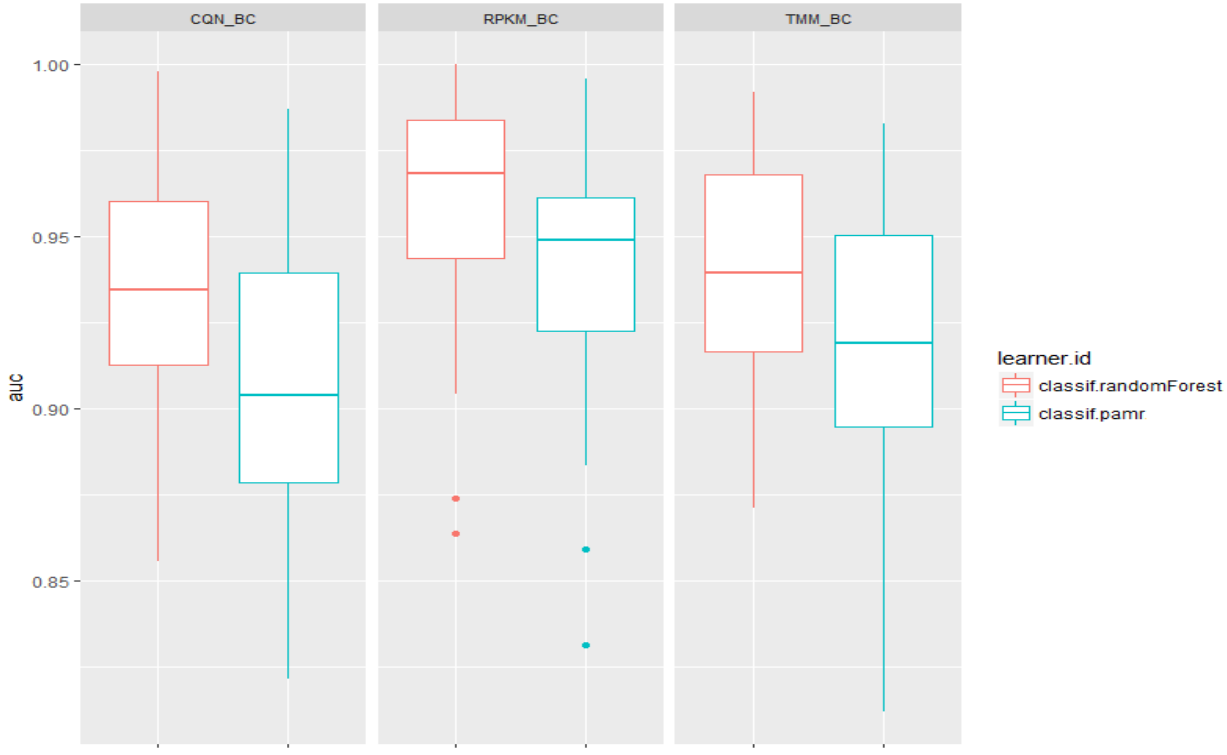


Figura 14: Representación en forma de diagrama de cajas el rendimiento de los modelos, entrenados a partir de las bases de datos que presentaban únicamente los genes relacionados con el cáncer de mama.

En cuanto a la selección del mejor modelo, se realizó en primer lugar un test no paramétrico de Wilcoxon. Éste reportó un p-valor=4.495e-06. Debido a la gran significación del resultado, y el refuerzo de éste en base a las gráficas comentadas, se decidió no realizar un test de comparación múltiple. Se confirma por lo tanto que existe una evidencia significativa entre los dos modelos, por lo que Random Forest presenta un mejor rendimiento que pamR.

5.2. Experimento 2 basado en Machine Learning

Una vez obtenida la primera aproximación y el rendimiento del modelo utilizando únicamente los genes con relación biológica al cáncer de mama, en este experimento se pretende retirar las asunciones biológicas y únicamente trabajar con asunciones matemáticas sobre los valores de expresión de los genes. Para ello se entrenaron los modelos de clasificación con el total de genes obtenidos de la técnica RNAseq (18257). Debido a la gran cantidad de variables predictoras, esta base de datos se sometió a técnicas de filtrado de FS y así, las salidas generadas fueron utilizadas para el entrenamiento de los modelos. Tras el ránking de variables obtenidos a partir de FS, se eligieron subgrupos de genes de cantidad 50, 100, 200, 250, 300 y 350. Por lo tanto se obtuvieron 18 bases de datos a entrenar por los algoritmos.

En la ***Figura 15*** se representa el valor AUC de cada algoritmo con sus datos de expresión. Se observa que Random Forest vuelve a ser el algoritmo que mejor clasifica los datos para todas las bases de datos y que RPKM fue el tipo de normalización que mejores resultados obtuvo en la clasificación, en concreto la matriz con 100 *features* seleccionadas, con un valor AUC=0.963, mejorando el resultado del anterior experimento. La ***Figura 16*** refuerza la hipótesis de que Random Forest presenta un rendimiento mejor, al observar una mayor uniformidad de los resultados y una menor cantidad de *outliers.*

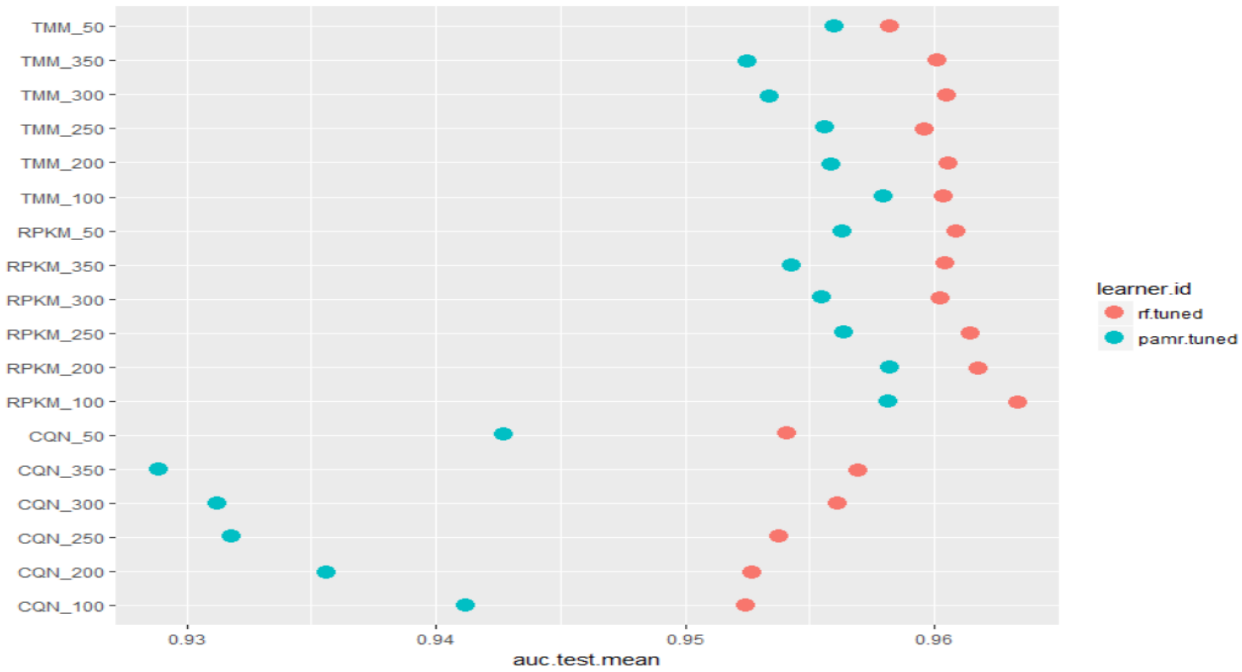


Figura 15: Resultado de la técnica Benchmarking del experimento 2 de ML. En el eje vertical representan las bases de datos utilizadas. Las letras corresponden al tipo de normalización, mientras que los números al total de features seleccionadas

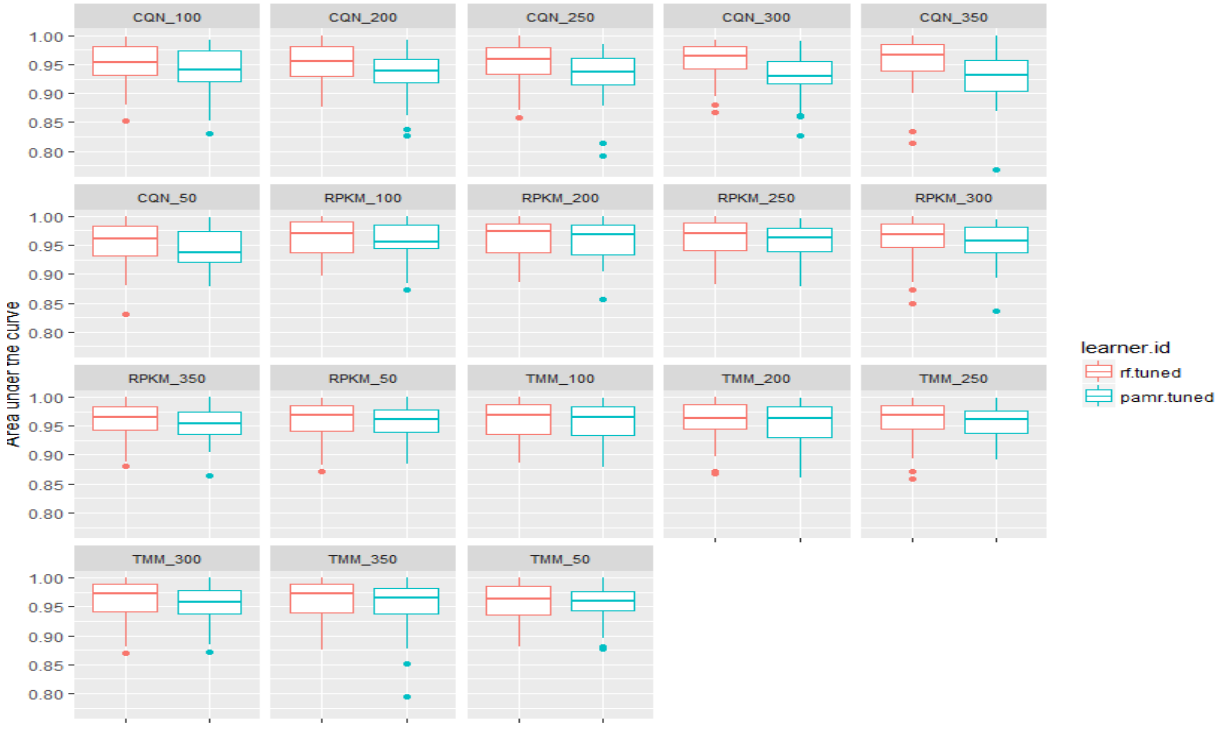


Figura 16: Representación en diagrama de cajas el rendimiento de los modelos del experimento 2 basado en ML.

En cuanto a la selección del mejor modelo, se obtuvo un p-valor= 6.828e-11 mediante el test de Wilcoxon. Concluyendo de esta manera que el modelo basado en Random Forest es estadísticamente mejor que el modelo basado en pamR. Como en el anterior experimento, el resultado es estadísticamente significativo como para necesitar un test de comparación múltiple.

Debido a la mejora de resultado tras las técnicas únicamente basadas en ML y FS, sin ofrecer información biológica al algoritmo, es conveniente saber cual fue el subgrupo de genes seleccionados que reportaron el mejor resultado (ANEXO VII). Además es interesante la comprobación de los genes obtenidos mediante esta técnica con respecto a los genes en común de los tres métodos del proyecto R/Bioconductor. Además, se hizo la comparación entre el subgrupo de genes del ANEXO VII y los genes DE obtenidos a partir del protocolo basado en Bioconductor (***Tabla 4***) y se detectaron únicamente tres genes en común: ESR1, GATA3 y FOXA1.

- **ESR1**: es un factor de transcripción activado por ligando compuesto por varios dominios importantes para la unión de la hormona, unión del DNA y activación de la transcripción. En tejido normal de mama, juega un papel crucial en su desarrollo. Es un marcador importante para el grupo ER positivo (Tomita *et al.*, 2009). ESR1 también conduce a un crecimiento en la mayoría de los cánceres de mama, uniéndose a elementos regulatorios e induciendo eventos transcripcionales que promueven el crecimiento del tumor. Diferencias en la unión de ESR1 a distintos *enhancers,* contribuye a perfiles distintos de expresión y resultados clínicos observados en pacientes con cáncer de mama (Theodorou *et al.*, 2013).
- **GATA3**: miembro de la familia de reguladores transcripcionales involucrado en la determinación de la identidad celular. Su expresión está asociada con la especificación celular del sistema inmune. En la glándula mamaria GATA3 se expresa mayoritariamente en la células epiteliales luminales diferenciadas, revistiendo las estructuras ductales de la mama. También se asociada íntimamente con la función e identidad de dichas células. Estudios indican que la pérdida de

expresión de GATA3 está asociado con tipos de tumor propensos al crecimiento invasivo y mal pronóstico (Takaku, Grimm and Wade, 2015).

- **FOXA1**: la familia FOXA son factores de transcripción que regulan la estructura de la cromatina y la expresión genética. En tejidos normales de mama, FOXA1 está involucrado en el desarrollo mamario, mientras que en cáncer de mama su expresión promueve un fenotipo luminal y se relaciona con un buen pronóstico (Hurtado *et al.*, 2011).

En cáncer de mama, GATA3 y FOXA1 son genes que definen un tumor de mama ESR1+, provocando una transcripción mediada por ESR1 y un crecimiento celular. El trabajo presentado por (Theodorou *et al.*, 2013) sugiere que GATA3 media la unión de ESR1 y FOXA1 en los elementos ERE incrustrados en al cromatina. Además GATA3 puede actuar *upstream* de FOXA1 mediante la unión de ESR1 por la modulación de la composición del *enhancer.*

Se detecta por lo tanto un alto nivel de relación de estos tres genes con el grupo de pacientes ER, dentro del cáncer de mama. Además, estos tres genes actúan dentro de un mismo *pathway* y existe un alto grado de dependencia entre ellos.

5.3. Experimento 3 basado en Machine Learning

En este experimento se pretende la mejora de los resultados del primer experimento aplicándole técnicas de FS de filtrado. Para ello, a partir de los 301 genes relacionados con cáncer de mama (ANEXO III) se obtuvieron subgrupos de cantidades 50, 100, 150, 200, 250 y 301.

Los resultados de cada modelo se presentan en la ***Figura 17*** según el valor AUC. El mejor modelo fue RPKM_150 entrenado por el algoritmo Random Forest. El valor AUC reportado fue de 0.962. En la ***Figura 18***, Random Forest presenta una mayor uniformidad de los resultados.

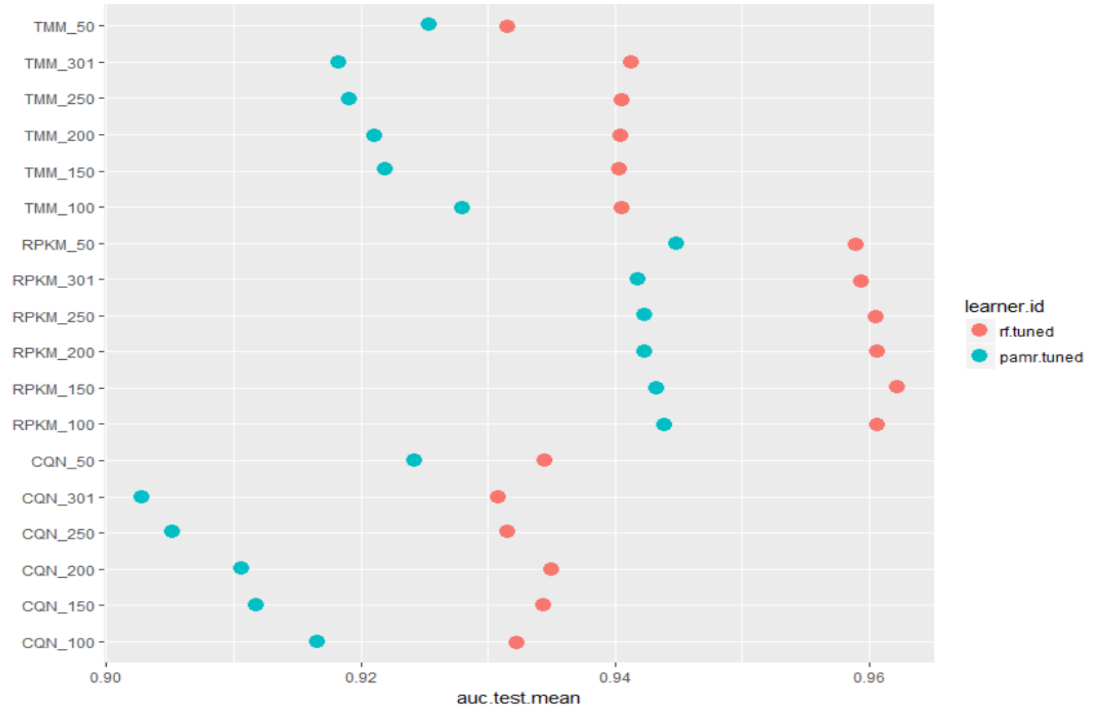


Figura 17: Resultados obtenidos tras el entrenamiento de los dos algoritmos mediante el experimento Benchmark. Los valores son AUC. En el eje vertical se representan los datasets. Las letras indican el tipo de normalización, mientras que los números las features seleccionadas.

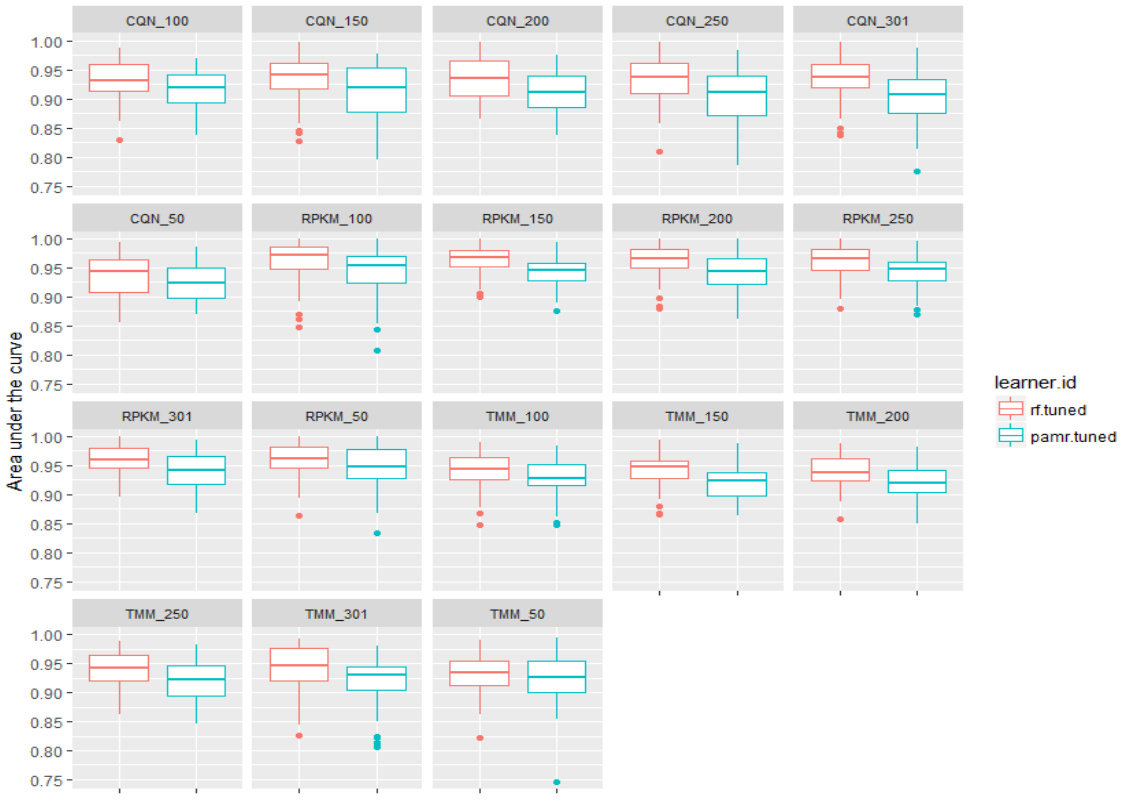


Figura 18: Representación en diagramas de cajas, del rendimiento de cada modelo entrenado en el experimento 3 basado en ML.

Se realizó la evaluación del mejor modelo mediante test de Wilcoxon que reportó un p-valor=2.22e-16. Los resultados conducen a una diferencia significativa entre el modelo Random Forest y el modelo pamR, sin necesidad de realizar el test de comparación múltiple.

Se obtuvo una mejora en cuanto al resultado obtenido en el primer experimento, aunque éste no supero al segundo experimento. El siguiente paso fue comparar las 150 *features* seleccionadas con los 116 genes DE obtenidos mediante el método basado en Bioconductor. Se identificaron 84 genes iguales en ambos listados.

No obstante, es interesante saber si existen algunas diferentes significativas a la hora de clasificar los pacientes en relación con los genes que no coinciden en ambas listas. Es decir, si los 32 no coincidentes aportan un mayor rendimiento al modelo. Para hacer esta comprobación se realizó a mayores un nuevo experimento de ML (se puede considerar como un subapartado del experimento 3). En este, se ha entrenado un algoritmo Random Forest (porque se observó que reportaba generalmente mejores resultados que pamR) y se seleccionaron, de la matriz normalizada RPKM (también debido a los resultados previos obtenidos) los 116 genes DE representados en la ***Tabla 4*****.** Finalmente, el resultado obtenido tras el entrenamiento fue un valor AUC = 0.960, lo que sugiere que los 116 genes DE generan una buena descripción para el *status* del receptor de estrógenos, aunque para obtener el mejor resultado se deben añadir otros genes. El resumen de los resultados de todos los experimentos se presenta en la **Tabla 6.**

Tabla 6: Resumen de los resultados obtenidos en los experimentos basados en ML.

Experimento ML	Mejor modelo	AUC
1	RPKM_BC	0.959
2	RPKM_100	0.963
3	RPKM_150	0.962
3.1	RPKM_116	0.960

6. CONCLUSIONES

Tras el análisis de datos genéticos de expresión RNAseq mediante dos protocolos: uno basado en Bioconductor y otro en técnicas de ML, se ha llegado a las siguientes conslusiones.

La implentación del protocolo basado en Bioconductor detectó 116 comunes a los tres paquetes utilizados para el análisis diferencial de genes entre los dos grupos de pacientes (ER+ y ER-). Lo que indica una gran coincidencia entre los paquetes. Estos genes sirvieron de base para la posterior comparación con los resultados obtenidos a través de las técnicas de ML y FS.

Mediante los experimentos basados en ML, los cuales pretendían generar un modelo para clasificar a los pacientes según su *status* del receptor de estrógenos, a través del perfil de expresión genética, se obtuvo como mejor rendimiento del modelo un valor AUC = 0.963. La base de datos con el que se entrenó dicho modelo procedió, sorprendentemente, de un subgrupo de genes del total, a partir de técnicas de FS de filtrado. Lo que supone que los genes han sido escogidos sin ninguna asunción biológica. Los demás experimentos obtuvieron también muy buenos resultados, reportando todos unos valores AUC mayores a 0.95. Esto indica que tanto los 301 genes relacionados con cáncer de mama, como los 116 genes DE son buenos descriptores de *status* del receptor de estrógenos. Además, a partir de los experimentos de ML se detectó que el mejor clasificador fue Random Forest y el mejor tipo de normalización fue RPKM.

La comparación de los 116 genes DE reportados por el protocolo de Bioconductor, y los genes que mejor clasificación generaron en los experimentos de ML (ANEXO VII) presentó únicamente tres coincidencias. Sin embargo, estos tres genes están estrechamente relacionados con el subtipo de cáncer de mama ER y presentan una función dentro de la misma ruta genética. Este hecho genera el interés de realizar un estudio más a fondo de los 97 genes obtenidos mediante ML, como posibles marcadores o dianas terapéuticas para pacientes ER.

7. FUTUROS DESARROLLOS

Debido a los resultados y a las conclusiones obtenidas en este proyecto existen futuros desarrollos que implementar a partir del presente trabajo.

En primer lugar un enfoque diferente en cuanto a los experimentos basados en ML. Se obtuvo en este proyecto una gran clasificación de los grupos de pacientes a estudio únicamente con los genes descargados de bases públicas que son considerados *drivers.* Además, cuando el conjunto total de genes se sometió a técnicas de FS, se reportó un resultado incluso mejor. Cuando se hizo la comparación entre estos dos resultados se identificaron tres coincidencias que correspondían a tres genes (ESR1, FOXA1 y GATA3) que actuaban de forma coordinada en el mismo *pathway.* Esto lleva a pensar en la posibilidad de entrenar los algoritmos basándose en los distintos *pathways* a los que pertenecen los genes DE. La forma de hacerlo mediante el proyecto Bioconductor y luego aplicándole técnicas de ML es sencilla, y los resultados obtenidos podrían ser de gran interés.

En segundo lugar, sería interesante el estudio general de los genes reportados en este trabajo. Los resultados en la clasificación son esperanzadores, y podría existir alguna relación genética subyacente en el cáncer de mama que aún no se ha encontrado, y que las técnicas basadas en ML pueden discernir únicamente basándose en asunciones matemáticas.

Todas estas ideas para futuros desarrollos pueden ir acompañadas con la creación de alguna aplicación que sea capaz de hacer un análisis bioinformático de los datos RNAseq y que presente una interfaz fácil y accesible al público, en la que todo el mundo pueda acceder. Esto sería interesante para empresas pequeñas o laboratorios biológicos. De esta forma, todo el trabajo bioinformático estaría disponible en un solo *click.* Así, el estudio y las conclusiones subyacentes de los análisis de expresión genética mejorarían en rapidez y calidad.

8. BIBLIOGRAFÍA

1. Anders, S. and Huber, W. (2010) 'Differential expression analysis for sequence count data', *Genome Biology*, 11, p. R106. doi: 10.1186/gb-2010-11-10-r106.
2. Anders, S., McCarthy, D., Chen, Y., Okoniewski, M., Smyth, G., Huber, W. and Robinson, M. (2013) 'Count-based differential expression analysis of RNA sequencing data using R and Bioconductor', *Nature Protocols*, 8(9), pp. 1765–1786. doi: 10.1038/nprot.2013.099.
3. Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G. and Jones, Z. M. (2016) 'mlr: Machine Learning in R', *Journal of Machine Learning Research*, 17, pp. 1–5.
4. Breiman, L. (2001) 'Random forests', *Machine Learning*. doi: 10.1023/A:1010933404324.
5. Bullard, J. H., Purdom, E., Hansen, K. D. and Dudoit, S. (2009) 'Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments', *U.C. Berkeley Div. Biostat. Pap. Ser.*, 11(1), p. 94. doi: 10.1186/1471-2105-11-94.
6. Chen, C., Liaw, A. and Breiman, L. (2004) 'Using random forest to learn imbalanced data', *University of California, Berkeley*, (1999), pp. 1–12. doi: ley.edu/sites/default/files/tech-reports/666.pdf.
7. Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., Bowlby, R., Shen, H., Hayat, S., Fieldhouse, R., Lester, S. C., Tse, G. M. K., Factor, R. E., Collins, L. C., Allison, K. H., Chen, Y. Y., Jensen, K., Johnson, N. B., Oesterreich, S., Mills, G. B., Cherniack, A. D., Robertson, G., Benz, C., Sander, C., Laird, P. W., Hoadley, K. A., King, T. A., Akbani, R., Auman, J. T., Balasundaram, M., Balu, S., Barr, T., Benz, S., Berrios, M., Beroukhim, R., Bodenheimer, T., Boice, L., Bootwalla, M. S., Bowen, J., Brooks, D., Chin, L., Cho, J., Chudamani, S., Davidsen, T., Demchok, J. A., Dennison, J. B., Ding, L., Felau, I., Ferguson, M. L., Frazer, S., Gabriel, S. B., Gao, J. J., Gastier-Foster, J. M., Gehlenborg, N., Gerken, M., Getz, G., Gibson, W. J., Hayes, D. N., Heiman, D. I., Holbrook, A., Holt, R. A., Hoyle, A. P., Hu, H., Huang, M., Hutter, C. M., Hwang, E. S., Jefferys, S. R., Jones, S. J. M., Ju, Z., Kim, J., Lai, P. H., Lawrence, M. S., Leraas, K. M., Lichtenberg, T. M., Lin, P., Ling, S., Liu, J., Liu, W., Lolla, L., Lu, Y., Ma, Y., Maglinte, D. T., Mardis, E., Marks, J., Marra, M. A., McAllister, C., Meng, S., Meyerson, M., Moore, R. A., Mose, L. E., Mungall, A. J., Murray, B. A., Naresh, R., Noble, M. S., Olopade, O., Parker, J. S., Pihl, T., Saksena, G., Schumacher, S. E., Shaw, K. R. M., Ramirez, N. C., Rathmell, W. K., Roach, J., Robertson, A. G., Schein, J. E., Schultz, N., Sheth, M., Shi, Y., Shih, J., Shelley, C. S., Shriver, C., Simons, J. V., Sofia, H. J., Soloway, M. G., Sougnez, C., Sun, C., Tarnuzzer, R., Tiezzi, D. G., Van Den Berg, D. J., Voet, D., Wan, Y., Wang, Z., Weinstein, J. N., Weisenberger, D. J., Wilson, R., Wise, L., Wiznerowicz, M., Wu, J., Wu, Y., Yang, L., Zack, T. I., Zenklusen, J. C.,

Zhang, J., Zmuda, E. and Perou, C. M. (2015) ‘Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer’, *Cell*, 163(2), pp. 506–519. doi: 10.1016/j.cell.2015.09.033.
8. Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Langerød, A., Green, A., Provenzano, E., Wishart, G., Pinder, S., Watson, P., Markowetz, F., Murphy, L., Ellis, I., Purushotham, A., Børresen-Dale, A.-L., Brenton, J. D., Tavaré, S., Caldas, C. and Aparicio, S. (2012) ‘The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.’, *Nature*, 486(7403), pp. 346–52. doi: 10.1038/nature10983.
9. Degroeve, S., De Baets, B., Van De Peer, Y. and Rou, P. (2002) ‘Feature subset selection for splice site prediction’, *BIOINFORMATICS*, 18(2), pp. 75–83. Available at: http://www.
10. Diaz-Uriarte, R. and Alvarez de Andres, S. (2006) ‘Gene selection and classification of microarray data using random forest’, *BMC Bioinformatics*, 7, p. 3. doi: 10.1186/1471-2105-7-3.
11. Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, N. S., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloë, D., Le Gall, C., Schaëffer, B., Le Crom, S., Guedj, M. and Jaffrézic, F. (2013) ‘A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis’, *Briefings in Bioinformatics*, 14(6), pp. 671–683. doi: 10.1093/bib/bbs046.
12. Esnaola, M., Puig, P., Gonzalez, D., Castelo, R. and Gonzalez, J. R. (2013) ‘A flexible count data model to fit the wide diversity of expression profiles arising from extensively replicated RNA-seq experiments’, *BMC Bioinformatics*, 14. Available at: http://www.biomedcentral.com/1471-2105/14/254.
13. Fernandez-Lozano, C., Gestal, M., Munteanu, C. R., Dorado, J. and Pazos, A. (2016) ‘A methodology for the design of experiments in computational intelligence with multiple regression models’. doi: 10.7717/peerj.2721.
14. Fernandez-Lozano, C., Gestal, M., Munteanu, C. R., Dorado, J. and Pazos, A. (2016) ‘A methodology for the design of experiments in computational intelligence with multiple regression models’, *PeerJ*, 4, p. e2721. doi: 10.7717/peerj.2721.
15. Garber, M., Grabherr, M. G., Guttman, M. and Trapnell, C. (2011) ‘Computational methods for transcriptome annotation and quantification using RNA-seq’, *Nat Methods*, 8(6), pp. 469–477. doi: 10.1038/nmeth.1613.
16. García, S., Fernández, A., Luengo, J. and Herrera, F. (2010) ‘Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power’, *Information Sciences*, 180(10). doi: 10.1016/j.ins.2009.12.010.
17. Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M.,

Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H. and Zhang, J. (2004) 'Bioconductor: open software development for computational biology and bioinformatics.', *Genome biology*, 5(10), p. R80. doi: 10.1186/gb-2004-5-10-r80.
18. Ghasemi, A. and Zahediasl, S. (2012) 'Normality tests for statistical analysis: A guide for non-statisticians', *International Journal of Endocrinology and Metabolism*, 10(2), pp. 486–489. doi: 10.5812/ijem.3505.
19. Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M. P., Jene-Sanz, A., Santos, A. and Lopez-Bigas, N. (2013) 'IntOGen-mutations identifies cancer drivers across tumor types', *Nature Methods*, 10(11), pp. 1081–1082. doi: 10.1038/nmeth.2642.
20. Gregory, B. D., Yazaki, J. and Ecker, J. R. (2008) 'Utilizing tiling microarrays for whole-genome analysis in plants', *Plant Journal*, 53(4), pp. 636–644. doi: 10.1111/j.1365-313X.2007.03320.x.
21. Guo, J., Xu, N., Li, Z., Zhang, S., Wu, J., Kim, D. H., Sano Marma, M., Meng, Q., Cao, H., Li, X., Shi, S., Yu, L., Kalachikov, S., Russo, J. J., Turro, N. J. and Ju, J. (2008) 'Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides.', *Proceedings of the National Academy of Sciences of the United States of America*, 105(27), pp. 9145–9150. doi: 10.1073/pnas.0804023105.
22. Guo, Y., Li, C.-I., Ye, F. and Shyr, Y. (2013) 'Evaluation of read count based RNAseq analysis methods', pp. 11–13. doi: 10.1186/1471-2164-14-S8-S2.
23. Guyon, I., Elisseeff, A. and De, A. M. (2003) 'An Introduction to Variable and Feature Selection', *Journal of Machine Learning Research*, 3, pp. 1157–1182.
24. Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002) 'Gene selection for cancer classification using support vector machines', *Machine Learning*. doi: 10.1023/A:1012487302797.
25. Hansen, K. D., Irizarry, R. A. and Wu, Z. (2012) 'Removing technical variability in RNA-seq data using conditional quantile normalization', *Biostatistics*, 13(2), pp. 204–216. doi: 10.1093/biostatistics/kxr054.
26. Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oles, A. K., Pages, H., Reyes, A., Shannon, P., Smyth, G. K., Tenenbaum, D., Waldron, L. and Morgan, M. (2015) 'Orchestrating high-throughput genomic analysis with Bioconductor', *Nat Methods*, 12(2), pp. 115–121. doi: 10.1038/Nmeth.3252.
27. Hurtado, A., Holmes, K. A., Ross-Innes, C. S., Schmidt, D. and Carroll, J. S. (2011) 'FOXA1 is a key determinant of estrogen receptor function and endocrine response', *Nature Genetics*, 43(1), pp. 27–33. doi: 10.1038/ng.730.
28. Ju, J., Kim, D. H., Bi, L., Meng, Q., Bai, X., Li, Z., Li, X., Marma, M. S., Shi, S., Wu, J., Edwards, J. R., Romu, A. and Turro, N. J. (2006) 'Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators.', *Proceedings of the National Academy*

of Sciences of the United States of America, 103(52), pp. 19635–19640. doi: 10.1073/pnas.0609513103.
29. Kingsford, C. and Salzberg, S. L. (2008) 'What are decision trees?', *Nature biotechnology*, 26(9), pp. 1011–1013. doi: 10.1038/nbt0908-1011.
30. Koboldt, D. C., Fulton, R. S., McLellan, M. D., Schmidt, H., Kalicki-Veizer, J., McMichael, J. F., Fulton, L. L., Dooling, D. J., Ding, L., Mardis, E. R., Wilson, R. K., Ally, A., Balasundaram, M., Butterfield, Y. S. N., Carlsen, R., Carter, C., Chu, A., Chuah, E., Chun, H.-J. E., Coope, R. J. N., Dhalla, N., Guin, R., Hirst, C., Hirst, M., Holt, R. a., Lee, D., Li, H. I., Mayo, M., Moore, R. a., Mungall, A. J., Pleasance, E., Gordon Robertson, a., Schein, J. E., Shafiei, A., Sipahimalani, P., Slobodan, J. R., Stoll, D., Tam, A., Thiessen, N., Varhol, R. J., Wye, N., Zeng, T., Zhao, Y., Birol, I., Jones, S. J. M., Marra, M. a., Cherniack, A. D., Saksena, G., Onofrio, R. C., Pho, N. H., Carter, S. L., Schumacher, S. E., Tabak, B., Hernandez, B., Gentry, J., Nguyen, H., Crenshaw, A., Ardlie, K., Beroukhim, R., Winckler, W., Getz, G., Gabriel, S. B., Meyerson, M., Chin, L., Park, P. J., Kucherlapati, R., Hoadley, K. a., Todd Auman, J., Fan, C., Turman, Y. J., Shi, Y., Li, L., Topal, M. D., He, X., Chao, H.-H., Prat, A., Silva, G. O., Iglesia, M. D., Zhao, W., Usary, J., Berg, J. S., Adams, M., Booker, J., Wu, J., Gulabani, A., Bodenheimer, T., Hoyle, A. P., Simons, J. V., Soloway, M. G., Mose, L. E., Jefferys, S. R., Balu, S., Parker, J. S., Neil Hayes, D., Perou, C. M., Malik, S., Mahurkar, S., Shen, H., Weisenberger, D. J., Triche Jr, T., Lai, P. H., Bootwalla, M. S., Maglinte, D. T., Berman, B. P., Van Den Berg, D. J., Baylin, S. B., Laird, P. W., Creighton, C. J., Donehower, L. a., Getz, G., Noble, M., Voet, D., Saksena, G., Gehlenborg, N., DiCara, D., Zhang, J., Zhang, H., Wu, C.-J., Yingchun Liu, S., Lawrence, M. S., Zou, L., Sivachenko, A., Lin, P., Stojanov, P., Jing, R., Cho, J., Sinha, R., Park, R. W., Nazaire, M.-D., Robinson, J., Thorvaldsdottir, H., Mesirov, J., Park, P. J., Chin, L., Reynolds, S., Kreisberg, R. B., Bernard, B., Bressler, R., Erkkila, T., Lin, J., Thorsson, V., Zhang, W., Shmulevich, I., Ciriello, G., Weinhold, N., Schultz, N., Gao, J., Cerami, E., Gross, B., Jacobsen, A., Sinha, R., Arman Aksoy, B., Antipin, Y., Reva, B., Shen, R., Taylor, B. S., Ladanyi, M., Sander, C., Anur, P., Spellman, P. T., Lu, Y., Liu, W., Verhaak, R. R. G., Mills, G. B., Akbani, R., Zhang, N., Broom, B. M., Casasent, T. D., Wakefield, C., Unruh, A. K., Baggerly, K., Coombes, K., Weinstein, J. N., Haussler, D., Benz, C. C., Stuart, J. M., Benz, S. C., Zhu, J., Szeto, C. C., Scott, G. K., Yau, C., Paull, E. O., Carlin, D., Wong, C., Sokolov, A., Thusberg, J., Mooney, S., Ng, S., Goldstein, T. C., Ellrott, K., Grifford, M., Wilks, C., Ma, S., Craft, B., Yan, C., Hu, Y., Meerzaman, D., Gastier-Foster, J. M., Bowen, J., Ramirez, N. C., Black, R. E., White, P., Zmuda, E. J., Frick, J., Lichtenberg, T. M., Brookens, R., George, M. M., Gerken, M. a., Harper, H. a., Leraas, K. M., Wise, L. J., Tabler, T. R., McAllister, C., Barr, T., Hart-Kothari, M., Tarvin, K., Saller, C., Sandusky, G., Mitchell, C., Iacocca, M. V., Brown, J., Rabeno, B., Czerwinski, C., Petrelli, N., Dolzhansky, O., Abramov, M., Voronina, O., Potapova, O., Marks, J. R., Suchorska, W. M., Murawa, D., Kycler, W., Ibbs, M., Korski, K., Spychała, A., Murawa, P., Brzeziński, J. J., Perz, H., Łaźniak, R., Teresiak, M., Tatka, H., Leporowska, E., Bogusz-Czerniewicz, M., Malicki, J., Mackiewicz, A., Wiznerowicz, M.,

Van Le, X., Kohl, B., Viet Tien, N., Thorp, R., Van Bang, N., Sussman, H., Duc Phu, B., Hajek, R., Phi Hung, N., Viet The Phuong, T., Quyet Thang, H., Zaki Khan, K., Penny, R., Mallery, D., Curley, E., Shelton, C., Yena, P., Ingle, J. N., Couch, F. J., Lingle, W. L., King, T. a., Maria Gonzalez-Angulo, A., Mills, G. B., Dyer, M. D., Liu, S., Meng, X., Patangan, M., Waldman, F., Stöppler, H., Kimryn Rathmell, W., Thorne, L., Huang, M., Boice, L., Hill, A., Morrison, C., Gaudioso, C., Bshara, W., Daily, K., Egea, S. C., Pegram, M. D., Gomez-Fernandez, C., Dhir, R., Bhargava, R., Brufsky, A., Shriver, C. D., Hooke, J. a., Leigh Campbell, J., Mural, R. J., Hu, H., Somiari, S., Larson, C., Deyarmin, B., Kvecher, L., Kovatich, A. J., Ellis, M. J., King, T. a., Hu, H., Couch, F. J., Mural, R. J., Stricker, T., White, K., Olopade, O., Ingle, J. N., Luo, C., Chen, Y., Marks, J. R., Waldman, F., Wiznerowicz, M., Bose, R., Chang, L.-W., Beck, A. H., Maria Gonzalez-Angulo, A., Pihl, T., Jensen, M., Sfeir, R., Kahn, A., Chu, A., Kothiyal, P., Wang, Z., Snyder, E., Pontius, J., Ayala, B., Backus, M., Walton, J., Baboud, J., Berton, D., Nicholls, M., Srinivasan, D., Raman, R., Girshik, S., Kigonya, P., Alonso, S., Sanbhadti, R., Barletta, S., Pot, D., Sheth, M., Demchok, J. a., Mills Shaw, K. R., Yang, L., Eley, G., Ferguson, M. L., Tarnuzzer, R. W., Zhang, J., Dillon, L. a. L., Buetow, K., Fielding, P., Ozenberger, B. a., Guyer, M. S., Sofia, H. J. and Palchik, J. D. (2012) 'Comprehensive molecular portraits of human breast tumours', *Nature*, 490(7418), pp. 61–70. doi: 10.1038/nature11412.

31. Kutmon, M., Riutta, A., Nunes, N., Hanspers, K., Willighagen, E. L., Bohler, A., Mèlius, J., Waagmeester, A., Sinha, S. R., Miller, R., Coort, S. L., Cirillo, E., Smeets, B., Evelo, C. T. and Pico, A. R. (2016) 'WikiPathways: Capturing the full diversity of pathway knowledge', *Nucleic Acids Research*. doi: 10.1093/nar/gkv1024.
32. Levene, H. (1960) 'Levene test for equality of variances', *Contributions to probability and statistics*, pp. 1–2. Available at: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Levene+'+s+Test+for+Equality+of+Variances#3.
33. Levin, J. Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D. A., Friedman, N., Gnirke, A. and Regev, A. (2010) 'Comprehensive comparative analysis of strand-specific RNA sequencing methods.', *Nature methods*. Nature Publishing Group, 7(9), pp. 709–15. doi: 10.1038/nmeth.1491.
34. Libbrecht, M. W. and Noble, W. S. (2015) 'Machine learning applications in genetics and genomics', *Nat Rev Genet*, 16(6), pp. 321–332. doi: 10.1038/nrg3920.
35. Liu, H., Li, J. and Wong, L. (2002) 'A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns', *Genome Informatics*, 13, pp. 51–60.
36. Metzker, M. L. (2010) 'Sequencing technologies - the next generation.', *Nature reviews. Genetics*, 11(1), pp. 31–46. doi: 10.1038/nrg2626.
37. Mockler, T. C. and Ecker, J. R. (2005) 'Applications of DNA tiling arrays for whole-genome analysis', *Genomics*, pp. 1–15. doi: 10.1016/j.ygeno.2004.10.005.
38. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. (2008) 'Mapping and quantifying mammalian transcriptomes by RNA-

Seq', *Nature Methods*, 5(7), pp. 621–628. doi: 10.1038/nmeth.1226.
39. Mungle, T., Tewary, S., Das, D. K., Arun, I., Basak, B., Agarwal, S., Ahmed, R., Chatterjee, S. and Chakraborty, C. (2017) 'MRF-ANN: A machine learning approach for automated ER scoring of breast cancer immunohistochemical images', *Journal of Microscopy*, 0(0), pp. 1–13. doi: 10.1111/jmi.12552.
40. Nagaraj, S. H., Gasser, R. B. and Ranganathan, S. (2007) 'A hitchhiker's guide to expressed sequence tag (EST) analysis', *Briefings in Bioinformatics*, pp. 6–21. doi: 10.1093/bib/bbl015.
41. Niu, A. Q., Xie, L. J., Wang, H., Zhu, B. and Wang, S. Q. (2016) 'Prediction of selective estrogen receptor beta agonist using open data and machine learning approach', *Drug Design, Development and Therapy*, 10, pp. 2323–2331. doi: 10.2147/DDDT.S110603.
42. Pardo, M. and Sberveglieri, G. (2008) 'Random forests and nearest shrunken centroids for the classification of sensor array data', *Sensors and Actuators, B: Chemical*, 131(1), pp. 93–99. doi: 10.1016/j.snb.2007.12.015.
43. Pepke, S. and Ver Steeg, G. (2017) 'Comprehensive discovery of subsample gene expression components by information explanation: therapeutic implications in cancer', *BMC Medical Genomics*. BMC Medical Genomics, 10(1), p. 12. doi: 10.1186/s12920-017-0245-6.
44. Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. a, Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. a, Fluge, O., Pergamenschikov, a, Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, a L., Brown, P. O. and Botstein, D. (2000) 'Molecular portraits of human breast tumours.', *Nature*, 406(6797), pp. 747–752. doi: 10.1038/35021093.
45. Qian, N. and Sejnowski, T. J. (1988) 'Predicting the secondary structure of globular proteins using neural network models', *J.Mol.Biol.*, 202(4), pp. 865–884. doi: 0022-2836(88)90564-5 [pii].
46. Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2009) 'edgeR: A Bioconductor package for differential expression analysis of digital gene expression data', *Bioinformatics*. doi: 10.1093/bioinformatics/btp616.
47. Robinson, M. and Oshlack, A. (2010) 'A scaling normalization method for differential expression analysis of RNA-seq data', *Genome Biology*, 11(3), p. R25. doi: 10.1186/gb-2010-11-3-r25.
48. Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., Van De Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J. C. F., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D. and Brown, P. O. (no date) 'Systematic variation in gene expression patterns in human cancer cell lines'.
49. Saeys, Y., Inza, I. and Larrañaga, P. (2007) 'A review of feature selection techniques in bioinformatics', *Bioinformatics*. doi: 10.1093/bioinformatics/btm344.
50. Shapiro, S. S. and Wilk, M. B. (1965) 'Biometrika Trust An Analysis of Variance Test for Normality (Complete Samples)', *Source: Biometrika Biometrika Trust*, 52(34), pp. 591–611. doi: 10.1093/biomet/52.3-4.591.
51. Stupnikov, A., Glazko, G. V. and Emmert-Streib, F. (2015) 'Effects of subsampling on characteristics of RNA-seq data from triple negative breast cancer patients', *Chinese Journal of Cancer*. BioMed Central,

34(10), pp. 1–12. doi: 10.1186/s40880-015-0040-8.
52. Sundaramurthy, G. and Eghbalnia, H. R. (2015) ‘A probabilistic approach for automated discovery of perturbed genes using expression data from microarray or RNA-Seq’, *Computers in Biology and Medicine*. Elsevier, 67, pp. 29–40. doi: 10.1016/j.compbiomed.2015.07.029.
53. Svetnik, V., Liaw, A., Tong, C., Christopher Culberson, J., Sheridan, R. P. and Feuston, B. P. (2003) ‘Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling’, *Journal of Chemical Information and Computer Sciences*, 43(6), pp. 1947–1958. doi: 10.1021/ci034160g.
54. Syed, A. R. (2011) ‘A Review of Cross Validation and Adaptive Model Selection’. Available at: http://scholarworks.gsu.edu/math_theses.
55. Takaku, M., Grimm, S. A. and Wade, P. A. (2015) ‘GATA3 in breast cancer: Tumor suppressor or oncogene?’, *Gene Expression*, 16(4), pp. 163–168. doi: 10.3727/105221615X14399878166113.
56. Tarca, A. L., Carey, V. J., Chen, X., Romero, R. and Drăghici, S. (2007) ‘Machine Learning and Its Applications to Biology’, *PLoS Computational Biology*, 3(6), p. e116. doi: 10.1371/journal.pcbi.0030116.
57. Theodorou, V., Stark, R., Menon, S. and Carroll, J. S. (2013) ‘GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility’, *Genome Research*, 23(1), pp. 12–22. doi: 10.1101/gr.139469.112.
58. Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002) ‘Diagnosis of multiple cancer types by shrunken centroids of gene expression.’, *Proceedings of the National Academy of Sciences of the United States of America*, 99(10), pp. 6567–72. doi: 10.1073/pnas.082099299.
59. Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2003) ‘Class prediction by nearest shrunken centroids, with applications to DNA microarrays’, *Statistical Science*, 18(1), pp. 104–117. doi: 10.1214/ss/1056397488.
60. Tomita, S., Zhang, Z., Nakano, M., Ibusuki, M., Kawazoe, T., Yamamoto, Y. and Iwase, H. (2009) ‘Estrogen receptor α gene ESR1 amplification may predict endocrine therapy responsiveness in breast cancer patients’, *Cancer Science*, 100(6), pp. 1012–1017. doi: 10.1111/j.1349-7006.2009.01145.x.
61. Valkonen, M., Kartasalo, K., Liimatainen, K., Nykter, M., Latonen, L. and Ruusuvuori, P. (2017) ‘Metastasis detection from whole slide images using local features and random forests’, *Cytometry Part A*, (8), pp. 1–11. doi: 10.1002/cyto.a.23089.
62. Wang, Z., Gerstein, M. and Snyder, M. (no date) ‘RNA-Seq: a revolutionary tool for transcriptomics’. doi: 10.1038/nrg2484.
63. Yahya, A. A., Osman, A., Ramli, A. R. and Balola, A. (2011) ‘Feature selection for high dimensional data: An evolutionary filter approach’, *Journal of Computer Science*, 7(5), pp. 800–820. doi: 10.3844/jcssp.2011.800.820.
64. Yamamoto, Y., Saito, A., Tateishi, A., Shimojo, H., Kanno, H., Tsuchiya, S., Ito, K., Cosatto, E., Graf, H. P., Moraleda, R. R., Eils, R. and Grabe, N. (2017) ‘Quantitative diagnosis of breast tumors by morphometric classification of microenvironmental myoepithelial cells using a machine learning approach’, *Scientific Reports*. Nature Publishing Group, 7(April),

p. 46732. doi: 10.1038/srep46732.
65. Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. P. (2002) ‘Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation’, *Nucleic Acids Research*, 30(4).
66. Zhang, Y. Q. and Rajapakse, J. C. (2008) *Machine Learning in Bioinformatics*, *Machine Learning in Bioinformatics*. doi: 10.1002/9780470397428.

ANEXO I: CÓDIGO DE MACHINE LEARNING SIN FEATURE SELECTION

```
setwd()
library(pamr)

library(mlr)

#Definir tasks
RPKM_task=makeClassifTask(data=RPKM_BC, target = "ER_status")
CQN_task=makeClassifTask(data=CQN_BC, target = "ER_status")
TMM_task=makeClassifTask(data=TMM_BC, target = "ER_status")

task_list<-list(RPKM_task,CQN_task,TMM_task)

#Inner
##Random Forest
psrf<-makeParamSet(
 makeDiscreteParam("mtry", values = c(5:13)),
 makeDiscreteParam("ntree", values= 1000L),
 makeDiscreteParam("nodesize", values= c(1:5))
)
ctrl<-makeTuneControlGrid()
inner<-makeResampleDesc("Holdout")

l<-makeLearner("classif.randomForest", predict.type = "prob") #prob para que den los resultados auc
lrn_rf<-makeTuneWrapper(l, resampling = inner, par.set = psrf, measures = auc, control=ctrl, show.info = T)

##PamR
pspamr<-makeParamSet(
 makeDiscreteParam("threshold", values = c(0.05:2)),
 makeDiscreteParam("threshold.scale", values=c(0.05:2))
)

l<-makeLearner("classif.pamr", predict.type = "prob") #prob para que den los resultados auc
lrn_pamr<-makeTuneWrapper(l, resampling = inner, par.set = pspamr, measures = auc, control=ctrl, show.info = T)


learners<-list( lrn_rf,lrn_pamr)


#outer
outer1=list(
 makeResampleDesc("RepCV",reps=5, folds=10, stratify = TRUE),
 makeResampleDesc("RepCV",reps=5, folds=10, stratify = TRUE),
 makeResampleDesc("RepCV",reps=5, folds=10, stratify = TRUE))

bmr= benchmark(learners, task_list, outer1, measures = list(auc,tpr,fpr), show.info = T, models=T)
```

ANEXO II: CÓDIGO DE MACHINE LEARNING CON FEATURE SELECTION

```
####EXPERIMENTO 3######  LOS 3 DATASETS SOLO CON GENES BR HACIENDOLES FS

####### FEATURE SELECTION  ########

RPKM_task=makeClassifTask(data=RPKM_BC, target = "ER_status")

CQN_task=makeClassifTask(data=CQN_BC, target = "ER_status")

TMM_task=makeClassifTask(data=TMM_BC, target = "ER_status")


#defino el t.test para hacer FS

require(stringi)  #necesario este paquete para la función stri_paste

makeFilter(

 name = "t.test",

 desc = "Calculates scores according to t-test",

 pkg = character(0L),

 supported.tasks = c("classif"),

 supported.features = c("numerics", "factors"),

 fun = function(task, nselect,...) {

  data = getTaskData(task)

  sapply(getTaskFeatureNames(task), function(feat.name) {

   f = as.formula(stri_paste(feat.name,"~", getTaskTargetNames(task)))

   t = t.test(f, data = data)

   return(unname(abs(t$statistic)))

  })

 }

)



library(Rfast)

CQN_fs50<-filterFeatures(CQN_task, method = "t.test", abs=50)

CQN_fs100<-filterFeatures(CQN_task, method = "t.test", abs=100)
```

```
CQN_fs150<-filterFeatures(CQN_task, method = "t.test", abs=150)
CQN_fs200<-filterFeatures(CQN_task, method = "t.test", abs=200)
CQN_fs250<-filterFeatures(CQN_task, method = "t.test", abs=250)
CQN_fs301<-filterFeatures(CQN_task, method = "t.test", abs=301)



RPKM_fs50<-filterFeatures(RPKM_task, method = "t.test", abs=50)
RPKM_fs100<-filterFeatures(RPKM_task, method = "t.test", abs=100)
RPKM_fs150<-filterFeatures(RPKM_task, method = "t.test", abs=150)
RPKM_fs200<-filterFeatures(RPKM_task, method = "t.test", abs=200)
RPKM_fs250<-filterFeatures(RPKM_task, method = "t.test", abs=250)
RPKM_fs301<-filterFeatures(RPKM_task, method = "t.test", abs=301)



TMM_fs50<-filterFeatures(TMM_task, method = "t.test", abs=50)
TMM_fs100<-filterFeatures(TMM_task, method = "t.test", abs=100)
TMM_fs150<-filterFeatures(TMM_task, method = "t.test", abs=150)
TMM_fs200<-filterFeatures(TMM_task, method = "t.test", abs=200)
TMM_fs250<-filterFeatures(TMM_task, method = "t.test", abs=250)
TMM_fs301<-filterFeatures(TMM_task, method = "t.test", abs=301)

CQN_fs50$task.desc$id<-'CQN_50'
CQN_fs100$task.desc$id<-'CQN_100'
CQN_fs150$task.desc$id<-'CQN_150'
CQN_fs200$task.desc$id<-'CQN_200'
CQN_fs250$task.desc$id<-'CQN_250'
CQN_fs301$task.desc$id<-'CQN_301'

RPKM_fs50$task.desc$id<-'RPKM_50'
RPKM_fs100$task.desc$id<-'RPKM_100'
RPKM_fs150$task.desc$id<-'RPKM_150'
```

```
RPKM_fs200$task.desc$id<-'RPKM_200'

RPKM_fs250$task.desc$id<-'RPKM_250'

RPKM_fs301$task.desc$id<-'RPKM_301'


TMM_fs50$task.desc$id<-'TMM_50'

TMM_fs100$task.desc$id<-'TMM_100'

TMM_fs150$task.desc$id<-'TMM_150'

TMM_fs200$task.desc$id<-'TMM_200'

TMM_fs250$task.desc$id<-'TMM_250'

TMM_fs301$task.desc$id<-'TMM_301'


FS_tasks<-
list(CQN_fs50,CQN_fs100,CQN_fs150,CQN_fs200,CQN_fs250,CQN_fs301,RPKM_
fs50,RPKM_fs100,RPKM_fs150,RPKM_fs200,RPKM_fs250,RPKM_fs301,

      TMM_fs50,TMM_fs100,TMM_fs150,TMM_fs200,TMM_fs250,TMM_fs301)


#Inner

##Random Forest


getParamSet("classif.randomForest") #para ver los hiperparámetros del
algoritmo

psrf<-makeParamSet(

 makeDiscreteParam("mtry", values = c(5:13)),

 makeDiscreteParam("ntree", values= 1000L),

 makeDiscreteParam("nodesize", values= c(1:5))

)

ctrl<-makeTuneControlGrid()

inner<-makeResampleDesc("Holdout")


l<-makeLearner("classif.randomForest", predict.type = "prob") #prob para que
den los resultados auc

lrn_rf<-makeTuneWrapper(l, resampling = inner, par.set = psrf, measures = auc,
control=ctrl, show.info = T)
```

```
##PamR

getParamSet("classif.pamr")

pspamr<-makeParamSet(

 makeDiscreteParam("threshold", values = c(0.05:2)),

 makeDiscreteParam("threshold.scale", values=c(0.05:2))

)

l<-makeLearner("classif.pamr", predict.type = "prob") #prob para que den los resultados auc

lrn_pamr<-makeTuneWrapper(l, resampling = inner, par.set = pspamr, measures = auc, control=ctrl, show.info = T)

learners<-list( lrn_rf,lrn_pamr)

outer2=list(makeResampleDesc("RepCV",reps=5, folds=10, stratify = TRUE),
makeResampleDesc("RepCV",reps=5, folds=10, stratify = TRUE),
makeResampleDesc("RepCV",reps=5, folds=10, stratify = TRUE),
makeResampleDesc("RepCV",reps=5, folds=10, stratify = TRUE),
makeResampleDesc("RepCV",reps=5, folds=10, stratify = TRUE),
makeResampleDesc("RepCV",reps=5, folds=10, stratify = TRUE),
makeResampleDesc("RepCV",reps=5, folds=10, stratify = TRUE),
makeResampleDesc("RepCV",reps=5, folds=10, stratify = TRUE),
makeResampleDesc("RepCV",reps=5, folds=10, stratify = TRUE),
makeResampleDesc("RepCV",reps=5, folds=10, stratify = TRUE),
makeResampleDesc("RepCV",reps=5, folds=10, stratify = TRUE),
makeResampleDesc("RepCV",reps=5, folds=10, stratify = TRUE),
makeResampleDesc("RepCV",reps=5, folds=10, stratify = TRUE),
makeResampleDesc("RepCV",reps=5, folds=10, stratify = TRUE),
makeResampleDesc("RepCV",reps=5, folds=10, stratify = TRUE),
makeResampleDesc("RepCV",reps=5, folds=10, stratify = TRUE),
makeResampleDesc("RepCV",reps=5, folds=10, stratify = TRUE),
makeResampleDesc("RepCV",reps=5, folds=10, stratify = TRUE),
makeResampleDesc("RepCV",reps=5, folds=10, stratify = TRUE),
makeResampleDesc("RepCV",reps=5, folds=10, stratify = TRUE))

bmr_FS_breast_cancer= benchmark(learners, FS_tasks, outer2, measures = list(auc,tpr,fpr), show.info = T, models=T)perf<-getBMRPerformances(bmr_FS_breast_cancer,as.df = TRUE)plot<-plotBMRSummary(bmr_FS_breast_cancer)
```

ANEXO III: TABLA DE GENES QUE SE RELACIONAN CON EL CÁNCER DE MAMA

hgnc_symbol	description
ABL1	ABL proto-oncogene 1, non-receptor tyrosine kinase [Source:HGNC Symbol;Acc:HGNC:76]
ACO1	aconitase 1 [Source:HGNC Symbol;Acc:HGNC:117]
ACSL6	acyl-CoA synthetase long-chain family member 6 [Source:HGNC Symbol;Acc:HGNC:16496]
ACTB	actin beta [Source:HGNC Symbol;Acc:HGNC:132]
ACVR1B	activin A receptor type 1B [Source:HGNC Symbol;Acc:HGNC:172]
AFF4	AF4/FMR2 family member 4 [Source:HGNC Symbol;Acc:HGNC:17869]
AHNAK	AHNAK nucleoprotein [Source:HGNC Symbol;Acc:HGNC:347]
AHR	aryl hydrocarbon receptor [Source:HGNC Symbol;Acc:HGNC:348]
AKAP9	A-kinase anchoring protein 9 [Source:HGNC Symbol;Acc:HGNC:379]
AKT1	AKT serine/threonine kinase 1 [Source:HGNC Symbol;Acc:HGNC:391]
ALKBH1	alkB homolog 1, histone H2A dioxygenase [Source:HGNC Symbol;Acc:HGNC:17911]
ANK3	ankyrin 3 [Source:HGNC Symbol;Acc:HGNC:494]
ANXA1	annexin A1 [Source:HGNC Symbol;Acc:HGNC:533]
APC	APC, WNT signaling pathway regulator [Source:HGNC Symbol;Acc:HGNC:583]
APOBEC3G	apolipoprotein B mRNA editing enzyme catalytic subunit 3G [Source:HGNC Symbol;Acc:HGNC:17357]
AQR	aquarius intron-binding spliceosomal factor [Source:HGNC Symbol;Acc:HGNC:29513]
AR	androgen receptor [Source:HGNC Symbol;Acc:HGNC:644]
ARAF	A-Raf proto-oncogene, serine/threonine kinase [Source:HGNC Symbol;Acc:HGNC:646]
ARFGEF2	ADP ribosylation factor guanine nucleotide exchange factor 2 [Source:HGNC Symbol;Acc:HGNC:15853]
ARHGAP35	Rho GTPase activating protein 35 [Source:HGNC Symbol;Acc:HGNC:4591]
ARID1A	AT-rich interaction domain 1A [Source:HGNC Symbol;Acc:HGNC:11110]
ARID2	AT-rich interaction domain 2 [Source:HGNC Symbol;Acc:HGNC:18037]
ARID4B	AT-rich interaction domain 4B [Source:HGNC Symbol;Acc:HGNC:15550]

ARNTL	aryl hydrocarbon receptor nuclear translocator like [Source:HGNC Symbol;Acc:HGNC:701]
ASH1L	ASH1 like histone lysine methyltransferase [Source:HGNC Symbol;Acc:HGNC:19088]
ASPM	abnormal spindle microtubule assembly [Source:HGNC Symbol;Acc:HGNC:19048]
ATF1	activating transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:783]
ATIC	5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase [Source:HGNC Symbol;Acc:HGNC:794]
ATM	ATM serine/threonine kinase [Source:HGNC Symbol;Acc:HGNC:795]
ATR	ATR serine/threonine kinase [Source:HGNC Symbol;Acc:HGNC:882]
AURKA	aurora kinase A [Source:HGNC Symbol;Acc:HGNC:11393]
BACH1	BTB domain and CNC homolog 1 [Source:HGNC Symbol;Acc:HGNC:935]
BAD	BCL2 associated agonist of cell death [Source:HGNC Symbol;Acc:HGNC:936]
BAK1	BCL2 antagonist/killer 1 [Source:HGNC Symbol;Acc:HGNC:949]
BAP1	BRCA1 associated protein 1 [Source:HGNC Symbol;Acc:HGNC:950]
BARD1	BRCA1 associated RING domain 1 [Source:HGNC Symbol;Acc:HGNC:952]
BAX	BCL2 associated X, apoptosis regulator [Source:HGNC Symbol;Acc:HGNC:959]
BCL2	BCL2, apoptosis regulator [Source:HGNC Symbol;Acc:HGNC:990]
BCOR	BCL6 corepressor [Source:HGNC Symbol;Acc:HGNC:20893]
BID	BH3 interacting domain death agonist [Source:HGNC Symbol;Acc:HGNC:1050]
BLM	Bloom syndrome RecQ like helicase [Source:HGNC Symbol;Acc:HGNC:1058]
BMPR1A	bone morphogenetic protein receptor type 1A [Source:HGNC Symbol;Acc:HGNC:1076]
BMPR2	bone morphogenetic protein receptor type 2 [Source:HGNC Symbol;Acc:HGNC:1078]
BNC2	basonuclin 2 [Source:HGNC Symbol;Acc:HGNC:30988]
BPTF	bromodomain PHD finger transcription factor [Source:HGNC Symbol;Acc:HGNC:3581]
BRAF	B-Raf proto-oncogene, serine/threonine kinase [Source:HGNC Symbol;Acc:HGNC:1097]
BRCA1	BRCA1, DNA repair associated [Source:HGNC Symbol;Acc:HGNC:1100]
BRCA2	BRCA2, DNA repair associated [Source:HGNC Symbol;Acc:HGNC:1101]

CAD	carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase [Source:HGNC Symbol;Acc:HGNC:1424]
CARM1	coactivator associated arginine methyltransferase 1 [Source:HGNC Symbol;Acc:HGNC:23393]
CASP3	caspase 3 [Source:HGNC Symbol;Acc:HGNC:1504]
CASP8	caspase 8 [Source:HGNC Symbol;Acc:HGNC:1509]
CASP9	caspase 9 [Source:HGNC Symbol;Acc:HGNC:1511]
CAST	calpastatin [Source:HGNC Symbol;Acc:HGNC:1515]
CBFB	core-binding factor beta subunit [Source:HGNC Symbol;Acc:HGNC:1539]
CCAR1	cell division cycle and apoptosis regulator 1 [Source:HGNC Symbol;Acc:HGNC:24236]
CCNB1IP1	cyclin B1 interacting protein 1 [Source:HGNC Symbol;Acc:HGNC:19437]
CCND1	cyclin D1 [Source:HGNC Symbol;Acc:HGNC:1582]
CCT5	chaperonin containing TCP1 subunit 5 [Source:HGNC Symbol;Acc:HGNC:1618]
CDC25A	cell division cycle 25A [Source:HGNC Symbol;Acc:HGNC:1725]
CDC25B	cell division cycle 25B [Source:HGNC Symbol;Acc:HGNC:1726]
CDC42	cell division cycle 42 [Source:HGNC Symbol;Acc:HGNC:1736]
CDH1	cadherin 1 [Source:HGNC Symbol;Acc:HGNC:1748]
CDK12	cyclin dependent kinase 12 [Source:HGNC Symbol;Acc:HGNC:24224]
CDK2	cyclin dependent kinase 2 [Source:HGNC Symbol;Acc:HGNC:1771]
CDK4	cyclin dependent kinase 4 [Source:HGNC Symbol;Acc:HGNC:1773]
CDK7	cyclin dependent kinase 7 [Source:HGNC Symbol;Acc:HGNC:1778]
CDKN1B	cyclin dependent kinase inhibitor 1B [Source:HGNC Symbol;Acc:HGNC:1785]
CEP290	centrosomal protein 290 [Source:HGNC Symbol;Acc:HGNC:29021]
CERK	ceramide kinase [Source:HGNC Symbol;Acc:HGNC:19256]
CHD4	chromodomain helicase DNA binding protein 4 [Source:HGNC Symbol;Acc:HGNC:1919]
CHD9	chromodomain helicase DNA binding protein 9 [Source:HGNC Symbol;Acc:HGNC:25701]
CHEK1	checkpoint kinase 1 [Source:HGNC Symbol;Acc:HGNC:1925]
CHEK2	checkpoint kinase 2 [Source:HGNC Symbol;Acc:HGNC:16627]
CHUK	conserved helix-loop-helix ubiquitous kinase [Source:HGNC Symbol;Acc:HGNC:1974]
CIC	capicua transcriptional repressor [Source:HGNC Symbol;Acc:HGNC:14214]
CLASP2	cytoplasmic linker associated protein 2 [Source:HGNC Symbol;Acc:HGNC:17078]

CLSPN	claspin [Source:HGNC Symbol;Acc:HGNC:19715]
CLTC	clathrin heavy chain [Source:HGNC Symbol;Acc:HGNC:2092]
CNOT3	CCR4-NOT transcription complex subunit 3 [Source:HGNC Symbol;Acc:HGNC:7879]
CREB1	cAMP responsive element binding protein 1 [Source:HGNC Symbol;Acc:HGNC:2345]
CSDE1	cold shock domain containing E1 [Source:HGNC Symbol;Acc:HGNC:29905]
CSNK1D	casein kinase 1 delta [Source:HGNC Symbol;Acc:HGNC:2452]
CSNK1G3	casein kinase 1 gamma 3 [Source:HGNC Symbol;Acc:HGNC:2456]
CTCF	CCCTC-binding factor [Source:HGNC Symbol;Acc:HGNC:13723]
CTNNB1	catenin beta 1 [Source:HGNC Symbol;Acc:HGNC:2514]
CUL1	cullin 1 [Source:HGNC Symbol;Acc:HGNC:2551]
CYP19A1	cytochrome P450 family 19 subfamily A member 1 [Source:HGNC Symbol;Acc:HGNC:2594]
DAG1	dystroglycan 1 [Source:HGNC Symbol;Acc:HGNC:2666]
DCAKD	dephospho-CoA kinase domain containing [Source:HGNC Symbol;Acc:HGNC:26238]
DDX3X	DEAD-box helicase 3, X-linked [Source:HGNC Symbol;Acc:HGNC:2745]
DDX5	DEAD-box helicase 5 [Source:HGNC Symbol;Acc:HGNC:2746]
DHTKD1	dehydrogenase E1 and transketolase domain containing 1 [Source:HGNC Symbol;Acc:HGNC:23537]
DHX15	DEAH-box helicase 15 [Source:HGNC Symbol;Acc:HGNC:2738]
DIS3	DIS3 homolog, exosome endoribonuclease and 3'-5' exoribonuclease [Source:HGNC Symbol;Acc:HGNC:20604]
E2F1	E2F transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:3113]
EDAR	ectodysplasin A receptor [Source:HGNC Symbol;Acc:HGNC:2895]
EGFR	epidermal growth factor receptor [Source:HGNC Symbol;Acc:HGNC:3236]
EIF1AX	eukaryotic translation initiation factor 1A, X-linked [Source:HGNC Symbol;Acc:HGNC:3250]
EIF4A2	eukaryotic translation initiation factor 4A2 [Source:HGNC Symbol;Acc:HGNC:3284]
EIF4G1	eukaryotic translation initiation factor 4 gamma 1 [Source:HGNC Symbol;Acc:HGNC:3296]
ELF1	E74 like ETS transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:3316]
EP300	E1A binding protein p300 [Source:HGNC Symbol;Acc:HGNC:3373]
ERAL1	Era like 12S mitochondrial rRNA chaperone 1 [Source:HGNC Symbol;Acc:HGNC:3424]

ERBB2	erb-b2 receptor tyrosine kinase 2 [Source:HGNC Symbol;Acc:HGNC:3430]
ERCC2	ERCC excision repair 2, TFIIH core complex helicase subunit [Source:HGNC Symbol;Acc:HGNC:3434]
ESR1	estrogen receptor 1 [Source:HGNC Symbol;Acc:HGNC:3467]
FADD	Fas associated via death domain [Source:HGNC Symbol;Acc:HGNC:3573]
FAU	FAU, ubiquitin like and ribosomal protein S30 fusion [Source:HGNC Symbol;Acc:HGNC:3597]
FBXW7	F-box and WD repeat domain containing 7 [Source:HGNC Symbol;Acc:HGNC:16712]
FER	FER tyrosine kinase [Source:HGNC Symbol;Acc:HGNC:3655]
FILIP1	filamin A interacting protein 1 [Source:HGNC Symbol;Acc:HGNC:21015]
FLT3	fms related tyrosine kinase 3 [Source:HGNC Symbol;Acc:HGNC:3765]
FMR1	fragile X mental retardation 1 [Source:HGNC Symbol;Acc:HGNC:3775]
FN1	fibronectin 1 [Source:HGNC Symbol;Acc:HGNC:3778]
FOSL1	FOS like 1, AP-1 transcription factor subunit [Source:HGNC Symbol;Acc:HGNC:13718]
FOXA1	forkhead box A1 [Source:HGNC Symbol;Acc:HGNC:5021]
FOXO1	forkhead box O1 [Source:HGNC Symbol;Acc:HGNC:3819]
FOXP1	forkhead box P1 [Source:HGNC Symbol;Acc:HGNC:3823]
FUBP1	far upstream element binding protein 1 [Source:HGNC Symbol;Acc:HGNC:4004]
FUS	FUS RNA binding protein [Source:HGNC Symbol;Acc:HGNC:4010]
G3BP2	G3BP stress granule assembly factor 2 [Source:HGNC Symbol;Acc:HGNC:30291]
GADD45A	growth arrest and DNA damage inducible alpha [Source:HGNC Symbol;Acc:HGNC:4095]
GATA3	GATA binding protein 3 [Source:HGNC Symbol;Acc:HGNC:4172]
GDI1	GDP dissociation inhibitor 1 [Source:HGNC Symbol;Acc:HGNC:4226]
GOLGA5	golgin A5 [Source:HGNC Symbol;Acc:HGNC:4428]
GPS2	G protein pathway suppressor 2 [Source:HGNC Symbol;Acc:HGNC:4550]
GRN	granulin precursor [Source:HGNC Symbol;Acc:HGNC:4601]
GSK3A	glycogen synthase kinase 3 alpha [Source:HGNC Symbol;Acc:HGNC:4616]
HCFC1	host cell factor C1 [Source:HGNC Symbol;Acc:HGNC:4839]
HDAC1	histone deacetylase 1 [Source:HGNC Symbol;Acc:HGNC:4852]

HIPK2	homeodomain interacting protein kinase 2 [Source:HGNC Symbol;Acc:HGNC:14402]
HLA-A	major histocompatibility complex, class I, A [Source:HGNC Symbol;Acc:HGNC:4931]
HLF	HLF, PAR bZIP transcription factor [Source:HGNC Symbol;Acc:HGNC:4977]
HMGCR	3-hydroxy-3-methylglutaryl-CoA reductase [Source:HGNC Symbol;Acc:HGNC:5006]
HSPA8	heat shock protein family A (Hsp70) member 8 [Source:HGNC Symbol;Acc:HGNC:5241]
IDH1	isocitrate dehydrogenase (NADP(+)) 1, cytosolic [Source:HGNC Symbol;Acc:HGNC:5382]
IMPA1	inositol monophosphatase 1 [Source:HGNC Symbol;Acc:HGNC:6050]
IRS1	insulin receptor substrate 1 [Source:HGNC Symbol;Acc:HGNC:6125]
ITPKC	inositol-trisphosphate 3-kinase C [Source:HGNC Symbol;Acc:HGNC:14897]
ITSN1	intersectin 1 [Source:HGNC Symbol;Acc:HGNC:6183]
JAK1	Janus kinase 1 [Source:HGNC Symbol;Acc:HGNC:6190]
JAKMIP1	janus kinase and microtubule interacting protein 1 [Source:HGNC Symbol;Acc:HGNC:26460]
JUN	Jun proto-oncogene, AP-1 transcription factor subunit [Source:HGNC Symbol;Acc:HGNC:6204]
KALRN	kalirin, RhoGEF kinase [Source:HGNC Symbol;Acc:HGNC:4814]
KDM5C	lysine demethylase 5C [Source:HGNC Symbol;Acc:HGNC:11114]
KEAP1	kelch like ECH associated protein 1 [Source:HGNC Symbol;Acc:HGNC:23177]
KLF4	Kruppel like factor 4 [Source:HGNC Symbol;Acc:HGNC:6348]
KRAS	KRAS proto-oncogene, GTPase [Source:HGNC Symbol;Acc:HGNC:6407]
LCP1	lymphocyte cytosolic protein 1 [Source:HGNC Symbol;Acc:HGNC:6528]
LGALS13	galectin 13 [Source:HGNC Symbol;Acc:HGNC:15449]
LRP6	LDL receptor related protein 6 [Source:HGNC Symbol;Acc:HGNC:6698]
MACF1	microtubule-actin crosslinking factor 1 [Source:HGNC Symbol;Acc:HGNC:13664]
MAP2K4	mitogen-activated protein kinase kinase 4 [Source:HGNC Symbol;Acc:HGNC:6844]
MAP3K1	mitogen-activated protein kinase kinase kinase 1 [Source:HGNC Symbol;Acc:HGNC:6848]
MAP3K13	mitogen-activated protein kinase kinase kinase 13 [Source:HGNC Symbol;Acc:HGNC:6852]

MAP3K7CL	MAP3K7 C-terminal like [Source:HGNC Symbol;Acc:HGNC:16457]
MAPK1	mitogen-activated protein kinase 1 [Source:HGNC Symbol;Acc:HGNC:6871]
MAX	MYC associated factor X [Source:HGNC Symbol;Acc:HGNC:6913]
MDM2	MDM2 proto-oncogene [Source:HGNC Symbol;Acc:HGNC:6973]
MECOM	MDS1 and EVI1 complex locus [Source:HGNC Symbol;Acc:HGNC:3498]
MED12	mediator complex subunit 12 [Source:HGNC Symbol;Acc:HGNC:11957]
MED23	mediator complex subunit 23 [Source:HGNC Symbol;Acc:HGNC:2372]
MED24	mediator complex subunit 24 [Source:HGNC Symbol;Acc:HGNC:22963]
MGA	MGA, MAX dimerization protein [Source:HGNC Symbol;Acc:HGNC:14010]
MIR21	microRNA 21 [Source:HGNC Symbol;Acc:HGNC:31586]
MIR29B1	microRNA 29b-1 [Source:HGNC Symbol;Acc:HGNC:31619]
MIR29B2	microRNA 29b-2 [Source:HGNC Symbol;Acc:HGNC:31620]
MKL1	megakaryoblastic leukemia (translocation) 1 [Source:HGNC Symbol;Acc:HGNC:14334]
MLH1	mutL homolog 1 [Source:HGNC Symbol;Acc:HGNC:7127]
MMP1	matrix metallopeptidase 1 [Source:HGNC Symbol;Acc:HGNC:7155]
MSH2	mutS homolog 2 [Source:HGNC Symbol;Acc:HGNC:7325]
MSH6	mutS homolog 6 [Source:HGNC Symbol;Acc:HGNC:7329]
MSR1	macrophage scavenger receptor 1 [Source:HGNC Symbol;Acc:HGNC:7376]
MTOR	mechanistic target of rapamycin [Source:HGNC Symbol;Acc:HGNC:3942]
MUC20	mucin 20, cell surface associated [Source:HGNC Symbol;Acc:HGNC:23282]
MYB	MYB proto-oncogene, transcription factor [Source:HGNC Symbol;Acc:HGNC:7545]
MYC	v-myc avian myelocytomatosis viral oncogene homolog [Source:HGNC Symbol;Acc:HGNC:7553]
MYCBP2	MYC binding protein 2, E3 ubiquitin protein ligase [Source:HGNC Symbol;Acc:HGNC:23386]
MYH11	myosin heavy chain 11 [Source:HGNC Symbol;Acc:HGNC:7569]
MYH14	myosin heavy chain 14 [Source:HGNC Symbol;Acc:HGNC:23212]
MYH9	myosin heavy chain 9 [Source:HGNC Symbol;Acc:HGNC:7579]
MYT1	myelin transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:7622]
NAB1	NGFI-A binding protein 1 [Source:HGNC Symbol;Acc:HGNC:7626]

NCOA3	nuclear receptor coactivator 3 [Source:HGNC Symbol;Acc:HGNC:7670]
NCOR1	nuclear receptor corepressor 1 [Source:HGNC Symbol;Acc:HGNC:7672]
NDRG1	N-myc downstream regulated 1 [Source:HGNC Symbol;Acc:HGNC:7679]
NF1	neurofibromin 1 [Source:HGNC Symbol;Acc:HGNC:7765]
NF2	neurofibromin 2 [Source:HGNC Symbol;Acc:HGNC:7773]
NFKB1	nuclear factor kappa B subunit 1 [Source:HGNC Symbol;Acc:HGNC:7794]
NOTCH1	notch 1 [Source:HGNC Symbol;Acc:HGNC:7881]
NOTCH2	notch 2 [Source:HGNC Symbol;Acc:HGNC:7882]
NOXA1	NADPH oxidase activator 1 [Source:HGNC Symbol;Acc:HGNC:10668]
NR4A2	nuclear receptor subfamily 4 group A member 2 [Source:HGNC Symbol;Acc:HGNC:7981]
NRAS	neuroblastoma RAS viral oncogene homolog [Source:HGNC Symbol;Acc:HGNC:7989]
NSD1	nuclear receptor binding SET domain protein 1 [Source:HGNC Symbol;Acc:HGNC:14234]
NUP107	nucleoporin 107 [Source:HGNC Symbol;Acc:HGNC:29914]
NUP85	nucleoporin 85 [Source:HGNC Symbol;Acc:HGNC:8734]
NUP98	nucleoporin 98 [Source:HGNC Symbol;Acc:HGNC:8068]
ODC1	ornithine decarboxylase 1 [Source:HGNC Symbol;Acc:HGNC:8109]
PAK1	p21 (RAC1) activated kinase 1 [Source:HGNC Symbol;Acc:HGNC:8590]
PAX5	paired box 5 [Source:HGNC Symbol;Acc:HGNC:8619]
PBRM1	polybromo 1 [Source:HGNC Symbol;Acc:HGNC:30064]
PCDH18	protocadherin 18 [Source:HGNC Symbol;Acc:HGNC:14268]
PCSK6	proprotein convertase subtilisin/kexin type 6 [Source:HGNC Symbol;Acc:HGNC:8569]
PHB	prohibitin [Source:HGNC Symbol;Acc:HGNC:8912]
PHF6	PHD finger protein 6 [Source:HGNC Symbol;Acc:HGNC:18145]
PIAS1	protein inhibitor of activated STAT 1 [Source:HGNC Symbol;Acc:HGNC:2752]
PIGR	polymeric immunoglobulin receptor [Source:HGNC Symbol;Acc:HGNC:8968]
PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha [Source:HGNC Symbol;Acc:HGNC:8975]
PIK3CB	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit beta [Source:HGNC Symbol;Acc:HGNC:8976]
PIK3R1	phosphoinositide-3-kinase regulatory subunit 1 [Source:HGNC

	Symbol;Acc:HGNC:8979]
PIK3R2	phosphoinositide-3-kinase regulatory subunit 2 [Source:HGNC Symbol;Acc:HGNC:8980]
PIK3R3	phosphoinositide-3-kinase regulatory subunit 3 [Source:HGNC Symbol;Acc:HGNC:8981]
PIP5K1A	phosphatidylinositol-4-phosphate 5-kinase type 1 alpha [Source:HGNC Symbol;Acc:HGNC:8994]
PKIA	cAMP-dependent protein kinase inhibitor alpha [Source:HGNC Symbol;Acc:HGNC:9017]
PLK1	polo like kinase 1 [Source:HGNC Symbol;Acc:HGNC:9077]
PLK3	polo like kinase 3 [Source:HGNC Symbol;Acc:HGNC:2154]
PML	promyelocytic leukemia [Source:HGNC Symbol;Acc:HGNC:9113]
POLR2B	RNA polymerase II subunit B [Source:HGNC Symbol;Acc:HGNC:9188]
PPP4R3A	protein phosphatase 4 regulatory subunit 3A [Source:HGNC Symbol;Acc:HGNC:20219]
PPP4R3B	protein phosphatase 4 regulatory subunit 3B [Source:HGNC Symbol;Acc:HGNC:29267]
PRKAR1A	protein kinase cAMP-dependent type I regulatory subunit alpha [Source:HGNC Symbol;Acc:HGNC:9388]
PRKCZ	protein kinase C zeta [Source:HGNC Symbol;Acc:HGNC:9412]
PTEN	phosphatase and tensin homolog [Source:HGNC Symbol;Acc:HGNC:9588]
PTGS1	prostaglandin-endoperoxide synthase 1 [Source:HGNC Symbol;Acc:HGNC:9604]
PTPRU	protein tyrosine phosphatase, receptor type U [Source:HGNC Symbol;Acc:HGNC:9683]
RAC1	ras-related C3 botulinum toxin substrate 1 (rho family, small GTP binding protein Rac1) [Source:HGNC Symbol;Acc:HGNC:9801]
RAD50	RAD50 double strand break repair protein [Source:HGNC Symbol;Acc:HGNC:9816]
RAD51	RAD51 recombinase [Source:HGNC Symbol;Acc:HGNC:9817]
RAD54L	RAD54-like (S. cerevisiae) [Source:HGNC Symbol;Acc:HGNC:9826]
RALA	RAS like proto-oncogene A [Source:HGNC Symbol;Acc:HGNC:9839]
RALGAPA1	Ral GTPase activating protein catalytic alpha subunit 1 [Source:HGNC Symbol;Acc:HGNC:17770]
RAP1A	RAP1A, member of RAS oncogene family [Source:HGNC Symbol;Acc:HGNC:9855]
RASGEF1A	RasGEF domain family member 1A [Source:HGNC Symbol;Acc:HGNC:24246]
RASGRP3	RAS guanyl releasing protein 3 [Source:HGNC Symbol;Acc:HGNC:14545]

RB1	RB transcriptional corepressor 1 [Source:HGNC Symbol;Acc:HGNC:9884]
RBBP7	RB binding protein 7, chromatin remodeling factor [Source:HGNC Symbol;Acc:HGNC:9890]
RBM5	RNA binding motif protein 5 [Source:HGNC Symbol;Acc:HGNC:9902]
RFC4	replication factor C subunit 4 [Source:HGNC Symbol;Acc:HGNC:9972]
RHEB	Ras homolog enriched in brain [Source:HGNC Symbol;Acc:HGNC:10011]
RHO	rhodopsin [Source:HGNC Symbol;Acc:HGNC:10012]
RPGR	retinitis pigmentosa GTPase regulator [Source:HGNC Symbol;Acc:HGNC:10295]
RPL5	ribosomal protein L5 [Source:HGNC Symbol;Acc:HGNC:10360]
RPP38	ribonuclease P/MRP subunit p38 [Source:HGNC Symbol;Acc:HGNC:30329]
RRAS	related RAS viral (r-ras) oncogene homolog [Source:HGNC Symbol;Acc:HGNC:10447]
RUNX1	runt related transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:10471]
SEC24D	SEC24 homolog D, COPII coat complex component [Source:HGNC Symbol;Acc:HGNC:10706]
SETD2	SET domain containing 2 [Source:HGNC Symbol;Acc:HGNC:18420]
SETDB1	SET domain bifurcated 1 [Source:HGNC Symbol;Acc:HGNC:10761]
SF3B1	splicing factor 3b subunit 1 [Source:HGNC Symbol;Acc:HGNC:10768]
SFPQ	splicing factor proline and glutamine rich [Source:HGNC Symbol;Acc:HGNC:10774]
SIRT1	sirtuin 1 [Source:HGNC Symbol;Acc:HGNC:14929]
SMAD1	SMAD family member 1 [Source:HGNC Symbol;Acc:HGNC:6767]
SMAD2	SMAD family member 2 [Source:HGNC Symbol;Acc:HGNC:6768]
SMAD4	SMAD family member 4 [Source:HGNC Symbol;Acc:HGNC:6770]
SMAD6	SMAD family member 6 [Source:HGNC Symbol;Acc:HGNC:6772]
SMAD7	SMAD family member 7 [Source:HGNC Symbol;Acc:HGNC:6773]
SMARCA4	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4 [Source:HGNC Symbol;Acc:HGNC:11100]
SOS1	SOS Ras/Rac guanine nucleotide exchange factor 1 [Source:HGNC Symbol;Acc:HGNC:11187]
SOS2	SOS Ras/Rho guanine nucleotide exchange factor 2 [Source:HGNC Symbol;Acc:HGNC:11188]
SP1	Sp1 transcription factor [Source:HGNC Symbol;Acc:HGNC:11205]
SPTAN1	spectrin alpha, non-erythrocytic 1 [Source:HGNC Symbol;Acc:HGNC:11273]

SRGAP1	SLIT-ROBO Rho GTPase activating protein 1 [Source:HGNC Symbol;Acc:HGNC:17382]
STAG1	stromal antigen 1 [Source:HGNC Symbol;Acc:HGNC:11354]
STAG2	stromal antigen 2 [Source:HGNC Symbol;Acc:HGNC:11355]
STAT1	signal transducer and activator of transcription 1 [Source:HGNC Symbol;Acc:HGNC:11362]
STIP1	stress induced phosphoprotein 1 [Source:HGNC Symbol;Acc:HGNC:11387]
STK11	serine/threonine kinase 11 [Source:HGNC Symbol;Acc:HGNC:11389]
STK4	serine/threonine kinase 4 [Source:HGNC Symbol;Acc:HGNC:11408]
SUZ12	SUZ12 polycomb repressive complex 2 subunit [Source:HGNC Symbol;Acc:HGNC:17101]
SVEP1	sushi, von Willebrand factor type A, EGF and pentraxin domain containing 1 [Source:HGNC Symbol;Acc:HGNC:15985]
TAB1	TGF-beta activated kinase 1 (MAP3K7) binding protein 1 [Source:HGNC Symbol;Acc:HGNC:18157]
TAF1	TATA-box binding protein associated factor 1 [Source:HGNC Symbol;Acc:HGNC:11535]
TBL1XR1	transducin beta like 1 X-linked receptor 1 [Source:HGNC Symbol;Acc:HGNC:29529]
TBX3	T-box 3 [Source:HGNC Symbol;Acc:HGNC:11602]
TCF12	transcription factor 12 [Source:HGNC Symbol;Acc:HGNC:11623]
TCF7L2	transcription factor 7 like 2 [Source:HGNC Symbol;Acc:HGNC:11641]
TFDP1	transcription factor Dp-1 [Source:HGNC Symbol;Acc:HGNC:11749]
TFPI	tissue factor pathway inhibitor [Source:HGNC Symbol;Acc:HGNC:11760]
TGFBR1	transforming growth factor beta receptor 1 [Source:HGNC Symbol;Acc:HGNC:11772]
TGFBR2	transforming growth factor beta receptor 2 [Source:HGNC Symbol;Acc:HGNC:11773]
THRAP3	thyroid hormone receptor associated protein 3 [Source:HGNC Symbol;Acc:HGNC:22964]
TNPO1	transportin 1 [Source:HGNC Symbol;Acc:HGNC:6401]
TOM1	target of myb1 membrane trafficking protein [Source:HGNC Symbol;Acc:HGNC:11982]
TP53	tumor protein p53 [Source:HGNC Symbol;Acc:HGNC:11998]
TPR	translocated promoter region, nuclear basket protein [Source:HGNC Symbol;Acc:HGNC:12017]
TRADD	TNFRSF1A associated via death domain [Source:HGNC Symbol;Acc:HGNC:12030]
TRIO	trio Rho guanine nucleotide exchange factor [Source:HGNC Symbol;Acc:HGNC:12303]

TSC1	tuberous sclerosis 1 [Source:HGNC Symbol;Acc:HGNC:12362]
TSC2	tuberous sclerosis 2 [Source:HGNC Symbol;Acc:HGNC:12363]
UBE2F	ubiquitin conjugating enzyme E2 F (putative) [Source:HGNC Symbol;Acc:HGNC:12480]
USP15	ubiquitin specific peptidase 15 [Source:HGNC Symbol;Acc:HGNC:12613]
USP16	ubiquitin specific peptidase 16 [Source:HGNC Symbol;Acc:HGNC:12614]
USP21	ubiquitin specific peptidase 21 [Source:HGNC Symbol;Acc:HGNC:12620]
USP38	ubiquitin specific peptidase 38 [Source:HGNC Symbol;Acc:HGNC:20067]
VEGFA	vascular endothelial growth factor A [Source:HGNC Symbol;Acc:HGNC:12680]
WEE1	WEE1 G2 checkpoint kinase [Source:HGNC Symbol;Acc:HGNC:12761]
XRCC3	X-ray repair cross complementing 3 [Source:HGNC Symbol;Acc:HGNC:12830]
ZFP36L1	ZFP36 ring finger protein like 1 [Source:HGNC Symbol;Acc:HGNC:1107]
ZFP36L2	ZFP36 ring finger protein like 2 [Source:HGNC Symbol;Acc:HGNC:1108]
ZMIZ1	zinc finger MIZ-type containing 1 [Source:HGNC Symbol;Acc:HGNC:16493]
ZMYND8	zinc finger MYND-type containing 8 [Source:HGNC Symbol;Acc:HGNC:9397]
ZNF655	zinc finger protein 655 [Source:HGNC Symbol;Acc:HGNC:30899]

ANEXO IV: TABLA DE GENES DIFERENCIALMENTE EXPRESADOS OBTENIDOS CON EL PAQUETE edgeR

X	hgnc_symbol	description
1	PTPRU	protein tyrosine phosphatase, receptor type U [Source:HGNC Symbol;Acc:HGNC:9683]
2	RAD50	RAD50 double strand break repair protein [Source:HGNC Symbol;Acc:HGNC:9816]
3	AKAP9	A-kinase anchoring protein 9 [Source:HGNC Symbol;Acc:HGNC:379]
4	CDK2	cyclin dependent kinase 2 [Source:HGNC Symbol;Acc:HGNC:1771]
5	RBM5	RNA binding motif protein 5 [Source:HGNC Symbol;Acc:HGNC:9902]
6	CDK7	cyclin dependent kinase 7 [Source:HGNC Symbol;Acc:HGNC:1778]
7	CDKN1B	cyclin dependent kinase inhibitor 1B [Source:HGNC Symbol;Acc:HGNC:1785]
8	NDRG1	N-myc downstream regulated 1 [Source:HGNC Symbol;Acc:HGNC:7679]
9	CARM1	coactivator associated arginine methyltransferase 1 [Source:HGNC Symbol;Acc:HGNC:23393]
10	RPP38	ribonuclease P/MRP subunit p38 [Source:HGNC Symbol;Acc:HGNC:30329]
11	ARFGEF2	ADP ribosylation factor guanine nucleotide exchange factor 2 [Source:HGNC Symbol;Acc:HGNC:15853]
12	NOXA1	NADPH oxidase activator 1 [Source:HGNC Symbol;Acc:HGNC:10668]
13	EDAR	ectodysplasin A receptor [Source:HGNC Symbol;Acc:HGNC:2895]
14	STIP1	stress induced phosphoprotein 1 [Source:HGNC Symbol;Acc:HGNC:11387]
15	CHEK1	checkpoint kinase 1 [Source:HGNC Symbol;Acc:HGNC:1925]
16	CHEK2	checkpoint kinase 2 [Source:HGNC Symbol;Acc:HGNC:16627]
17	CLTC	clathrin heavy chain [Source:HGNC Symbol;Acc:HGNC:2092]
18	CSNK1G3	casein kinase 1 gamma 3 [Source:HGNC Symbol;Acc:HGNC:2456]
19	JAKMIP1	janus kinase and microtubule interacting protein 1 [Source:HGNC Symbol;Acc:HGNC:26460]
20	CYP19A1	cytochrome P450 family 19 subfamily A member 1 [Source:HGNC Symbol;Acc:HGNC:2594]

21	E2F1	E2F transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:3113]
22	EGFR	epidermal growth factor receptor [Source:HGNC Symbol;Acc:HGNC:3236]
23	ARID2	AT-rich interaction domain 2 [Source:HGNC Symbol;Acc:HGNC:18037]
24	EIF4G1	eukaryotic translation initiation factor 4 gamma 1 [Source:HGNC Symbol;Acc:HGNC:3296]
25	ELF1	E74 like ETS transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:3316]
26	ERBB2	erb-b2 receptor tyrosine kinase 2 [Source:HGNC Symbol;Acc:HGNC:3430]
27	ESR1	estrogen receptor 1 [Source:HGNC Symbol;Acc:HGNC:3467]
28	RASGEF1A	RasGEF domain family member 1A [Source:HGNC Symbol;Acc:HGNC:24246]
29	DIS3	DIS3 homolog, exosome endoribonuclease and 3'-5' exoribonuclease [Source:HGNC Symbol;Acc:HGNC:20604]
30	CCT5	chaperonin containing TCP1 subunit 5 [Source:HGNC Symbol;Acc:HGNC:1618]
31	CLASP2	cytoplasmic linker associated protein 2 [Source:HGNC Symbol;Acc:HGNC:17078]
32	FLT3	fms related tyrosine kinase 3 [Source:HGNC Symbol;Acc:HGNC:3765]
33	ACSL6	acyl-CoA synthetase long-chain family member 6 [Source:HGNC Symbol;Acc:HGNC:16496]
34	FMR1	fragile X mental retardation 1 [Source:HGNC Symbol;Acc:HGNC:3775]
35	SIRT1	sirtuin 1 [Source:HGNC Symbol;Acc:HGNC:14929]
36	ZMYND8	zinc finger MYND-type containing 8 [Source:HGNC Symbol;Acc:HGNC:9397]
37	RALGAPA1	Ral GTPase activating protein catalytic alpha subunit 1 [Source:HGNC Symbol;Acc:HGNC:17770]
38	RASGRP3	RAS guanyl releasing protein 3 [Source:HGNC Symbol;Acc:HGNC:14545]
39	ASPM	abnormal spindle microtubule assembly [Source:HGNC Symbol;Acc:HGNC:19048]
40	GATA3	GATA binding protein 3 [Source:HGNC Symbol;Acc:HGNC:4172]
41	GDI1	GDP dissociation inhibitor 1 [Source:HGNC Symbol;Acc:HGNC:4226]
42	FOXP1	forkhead box P1 [Source:HGNC Symbol;Acc:HGNC:3823]
43	AFF4	AF4/FMR2 family member 4 [Source:HGNC Symbol;Acc:HGNC:17869]

44	ARHGAP35	Rho GTPase activating protein 35 [Source:HGNC Symbol;Acc:HGNC:4591]
45	MSH6	mutS homolog 6 [Source:HGNC Symbol;Acc:HGNC:7329]
46	ANXA1	annexin A1 [Source:HGNC Symbol;Acc:HGNC:533]
47	HCFC1	host cell factor C1 [Source:HGNC Symbol;Acc:HGNC:4839]
48	HLA-A	major histocompatibility complex, class I, A [Source:HGNC Symbol;Acc:HGNC:4931]
49	HMGCR	3-hydroxy-3-methylglutaryl-CoA reductase [Source:HGNC Symbol;Acc:HGNC:5006]
50	FOXA1	forkhead box A1 [Source:HGNC Symbol;Acc:HGNC:5021]
51	APC	APC, WNT signaling pathway regulator [Source:HGNC Symbol;Acc:HGNC:583]
52	IDH1	isocitrate dehydrogenase (NADP(+)) 1, cytosolic [Source:HGNC Symbol;Acc:HGNC:5382]
53	IRS1	insulin receptor substrate 1 [Source:HGNC Symbol;Acc:HGNC:6125]
54	AR	androgen receptor [Source:HGNC Symbol;Acc:HGNC:644]
55	ARAF	A-Raf proto-oncogene, serine/threonine kinase [Source:HGNC Symbol;Acc:HGNC:646]
56	LCP1	lymphocyte cytosolic protein 1 [Source:HGNC Symbol;Acc:HGNC:6528]
57	LRP6	LDL receptor related protein 6 [Source:HGNC Symbol;Acc:HGNC:6698]
58	SMAD2	SMAD family member 2 [Source:HGNC Symbol;Acc:HGNC:6768]
59	SMAD6	SMAD family member 6 [Source:HGNC Symbol;Acc:HGNC:6772]
60	MAX	MYC associated factor X [Source:HGNC Symbol;Acc:HGNC:6913]
61	MDM2	MDM2 proto-oncogene [Source:HGNC Symbol;Acc:HGNC:6973]
62	MAP3K1	mitogen-activated protein kinase kinase kinase 1 [Source:HGNC Symbol;Acc:HGNC:6848]
63	MLH1	mutL homolog 1 [Source:HGNC Symbol;Acc:HGNC:7127]
64	MMP1	matrix metallopeptidase 1 [Source:HGNC Symbol;Acc:HGNC:7155]
65	MSH2	mutS homolog 2 [Source:HGNC Symbol;Acc:HGNC:7325]
66	MYB	MYB proto-oncogene, transcription factor [Source:HGNC Symbol;Acc:HGNC:7545]
67	MYC	v-myc avian myelocytomatosis viral oncogene homolog [Source:HGNC Symbol;Acc:HGNC:7553]
68	MYH9	myosin heavy chain 9 [Source:HGNC Symbol;Acc:HGNC:7579]

69	MYH11	myosin heavy chain 11 [Source:HGNC Symbol;Acc:HGNC:7569]
70	MYT1	myelin transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:7622]
71	NAB1	NGFI-A binding protein 1 [Source:HGNC Symbol;Acc:HGNC:7626]
72	NF1	neurofibromin 1 [Source:HGNC Symbol;Acc:HGNC:7765]
73	NF2	neurofibromin 2 [Source:HGNC Symbol;Acc:HGNC:7773]
74	NOTCH1	notch 1 [Source:HGNC Symbol;Acc:HGNC:7881]
75	NRAS	neuroblastoma RAS viral oncogene homolog [Source:HGNC Symbol;Acc:HGNC:7989]
76	NR4A2	nuclear receptor subfamily 4 group A member 2 [Source:HGNC Symbol;Acc:HGNC:7981]
77	ODC1	ornithine decarboxylase 1 [Source:HGNC Symbol;Acc:HGNC:8109]
78	PCSK6	proprotein convertase subtilisin/kexin type 6 [Source:HGNC Symbol;Acc:HGNC:8569]
79	PAX5	paired box 5 [Source:HGNC Symbol;Acc:HGNC:8619]
80	CDK12	cyclin dependent kinase 12 [Source:HGNC Symbol;Acc:HGNC:24224]
81	PIGR	polymeric immunoglobulin receptor [Source:HGNC Symbol;Acc:HGNC:8968]
82	PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha [Source:HGNC Symbol;Acc:HGNC:8975]
83	PIK3R2	phosphoinositide-3-kinase regulatory subunit 2 [Source:HGNC Symbol;Acc:HGNC:8980]
84	PLK1	polo like kinase 1 [Source:HGNC Symbol;Acc:HGNC:9077]
85	PML	promyelocytic leukemia [Source:HGNC Symbol;Acc:HGNC:9113]
86	POLR2B	RNA polymerase II subunit B [Source:HGNC Symbol;Acc:HGNC:9188]
87	ATR	ATR serine/threonine kinase [Source:HGNC Symbol;Acc:HGNC:882]
88	BNC2	basonuclin 2 [Source:HGNC Symbol;Acc:HGNC:30988]
89	BCOR	BCL6 corepressor [Source:HGNC Symbol;Acc:HGNC:20893]
90	FBXW7	F-box and WD repeat domain containing 7 [Source:HGNC Symbol;Acc:HGNC:16712]
91	DHTKD1	dehydrogenase E1 and transketolase domain containing 1 [Source:HGNC Symbol;Acc:HGNC:23537]
92	PKIA	cAMP-dependent protein kinase inhibitor alpha [Source:HGNC Symbol;Acc:HGNC:9017]
93	PRKAR1A	protein kinase cAMP-dependent type I regulatory subunit alpha [Source:HGNC Symbol;Acc:HGNC:9388]

94	ASH1L	ASH1 like histone lysine methyltransferase [Source:HGNC Symbol;Acc:HGNC:19088]
95	BACH1	BTB domain and CNC homolog 1 [Source:HGNC Symbol;Acc:HGNC:935]
96	ZMIZ1	zinc finger MIZ-type containing 1 [Source:HGNC Symbol;Acc:HGNC:16493]
97	BAD	BCL2 associated agonist of cell death [Source:HGNC Symbol;Acc:HGNC:936]
98	PPP4R3B	protein phosphatase 4 regulatory subunit 3B [Source:HGNC Symbol;Acc:HGNC:29267]
99	PTEN	phosphatase and tensin homolog [Source:HGNC Symbol;Acc:HGNC:9588]
100	PTGS1	prostaglandin-endoperoxide synthase 1 [Source:HGNC Symbol;Acc:HGNC:9604]
101	SRGAP1	SLIT-ROBO Rho GTPase activating protein 1 [Source:HGNC Symbol;Acc:HGNC:17382]
102	MKL1	megakaryoblastic leukemia (translocation) 1 [Source:HGNC Symbol;Acc:HGNC:14334]
103	BAK1	BCL2 antagonist/killer 1 [Source:HGNC Symbol;Acc:HGNC:949]
104	CCNB1IP1	cyclin B1 interacting protein 1 [Source:HGNC Symbol;Acc:HGNC:19437]
105	BARD1	BRCA1 associated RING domain 1 [Source:HGNC Symbol;Acc:HGNC:952]
106	BAX	BCL2 associated X, apoptosis regulator [Source:HGNC Symbol;Acc:HGNC:959]
107	RAD51	RAD51 recombinase [Source:HGNC Symbol;Acc:HGNC:9817]
108	RB1	RB transcriptional corepressor 1 [Source:HGNC Symbol;Acc:HGNC:9884]
109	RBBP7	RB binding protein 7, chromatin remodeling factor [Source:HGNC Symbol;Acc:HGNC:9890]
110	CCND1	cyclin D1 [Source:HGNC Symbol;Acc:HGNC:1582]
111	BCL2	BCL2, apoptosis regulator [Source:HGNC Symbol;Acc:HGNC:990]
112	RFC4	replication factor C subunit 4 [Source:HGNC Symbol;Acc:HGNC:9972]
113	RHEB	Ras homolog enriched in brain [Source:HGNC Symbol;Acc:HGNC:10011]
114	APOBEC3G	apolipoprotein B mRNA editing enzyme catalytic subunit 3G [Source:HGNC Symbol;Acc:HGNC:17357]
115	RPGR	retinitis pigmentosa GTPase regulator [Source:HGNC Symbol;Acc:HGNC:10295]
116	RPL5	ribosomal protein L5 [Source:HGNC Symbol;Acc:HGNC:10360]
117	BID	BH3 interacting domain death agonist [Source:HGNC

		Symbol;Acc:HGNC:1050]
118	CLSPN	claspin [Source:HGNC Symbol;Acc:HGNC:19715]
119	BLM	Bloom syndrome RecQ like helicase [Source:HGNC Symbol;Acc:HGNC:1058]
120	MAP2K4	mitogen-activated protein kinase kinase 4 [Source:HGNC Symbol;Acc:HGNC:6844]
121	CERK	ceramide kinase [Source:HGNC Symbol;Acc:HGNC:19256]
122	SMARCA4	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4 [Source:HGNC Symbol;Acc:HGNC:11100]
123	SOS1	SOS Ras/Rac guanine nucleotide exchange factor 1 [Source:HGNC Symbol;Acc:HGNC:11187]
124	SP1	Sp1 transcription factor [Source:HGNC Symbol;Acc:HGNC:11205]
125	BRCA2	BRCA2, DNA repair associated [Source:HGNC Symbol;Acc:HGNC:1101]
126	STAT1	signal transducer and activator of transcription 1 [Source:HGNC Symbol;Acc:HGNC:11362]
127	AURKA	aurora kinase A [Source:HGNC Symbol;Acc:HGNC:11393]
128	TBX3	T-box 3 [Source:HGNC Symbol;Acc:HGNC:11602]
129	TCF7L2	transcription factor 7 like 2 [Source:HGNC Symbol;Acc:HGNC:11641]
130	TFDP1	transcription factor Dp-1 [Source:HGNC Symbol;Acc:HGNC:11749]
131	TSC2	tuberous sclerosis 2 [Source:HGNC Symbol;Acc:HGNC:12363]
132	VEGFA	vascular endothelial growth factor A [Source:HGNC Symbol;Acc:HGNC:12680]
133	XRCC3	X-ray repair cross complementing 3 [Source:HGNC Symbol;Acc:HGNC:12830]
134	CAD	carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase [Source:HGNC Symbol;Acc:HGNC:1424]
135	AHNAK	AHNAK nucleoprotein [Source:HGNC Symbol;Acc:HGNC:347]
136	ZNF655	zinc finger protein 655 [Source:HGNC Symbol;Acc:HGNC:30899]
137	MYH14	myosin heavy chain 14 [Source:HGNC Symbol;Acc:HGNC:23212]
138	DCAKD	dephospho-CoA kinase domain containing [Source:HGNC Symbol;Acc:HGNC:26238]
139	NUP85	nucleoporin 85 [Source:HGNC Symbol;Acc:HGNC:8734]
140	SVEP1	sushi, von Willebrand factor type A, EGF and pentraxin domain containing 1 [Source:HGNC Symbol;Acc:HGNC:15985]
141	CEP290	centrosomal protein 290 [Source:HGNC

		Symbol;Acc:HGNC:29021]
142	FOSL1	FOS like 1, AP-1 transcription factor subunit [Source:HGNC Symbol;Acc:HGNC:13718]
143	PHF6	PHD finger protein 6 [Source:HGNC Symbol;Acc:HGNC:18145]
144	RAD54L	RAD54-like (S. cerevisiae) [Source:HGNC Symbol;Acc:HGNC:9826]
145	CUL1	cullin 1 [Source:HGNC Symbol;Acc:HGNC:2551]
146	RUNX1	runt related transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:10471]
147	CBFB	core-binding factor beta subunit [Source:HGNC Symbol;Acc:HGNC:1539]
148	FADD	Fas associated via death domain [Source:HGNC Symbol;Acc:HGNC:3573]
149	ALKBH1	alkB homolog 1, histone H2A dioxygenase [Source:HGNC Symbol;Acc:HGNC:17911]
150	KALRN	kalirin, RhoGEF kinase [Source:HGNC Symbol;Acc:HGNC:4814]
151	ACVR1B	activin A receptor type 1B [Source:HGNC Symbol;Acc:HGNC:172]
152	MAP3K13	mitogen-activated protein kinase kinase kinase 13 [Source:HGNC Symbol;Acc:HGNC:6852]
153	KLF4	Kruppel like factor 4 [Source:HGNC Symbol;Acc:HGNC:6348]
154	NCOR1	nuclear receptor corepressor 1 [Source:HGNC Symbol;Acc:HGNC:7672]
155	SETDB1	SET domain bifurcated 1 [Source:HGNC Symbol;Acc:HGNC:10761]
156	SEC24D	SEC24 homolog D, COPII coat complex component [Source:HGNC Symbol;Acc:HGNC:10706]
157	CDC25A	cell division cycle 25A [Source:HGNC Symbol;Acc:HGNC:1725]
158	CDC25B	cell division cycle 25B [Source:HGNC Symbol;Acc:HGNC:1726]
159	GOLGA5	golgin A5 [Source:HGNC Symbol;Acc:HGNC:4428]
160	CDC42	cell division cycle 42 [Source:HGNC Symbol;Acc:HGNC:1736]
161	CDH1	cadherin 1 [Source:HGNC Symbol;Acc:HGNC:1748]

ANEXO V: TABLA DE GENES DIFERENCIALMENTE EXPRESADOS OBTENIDOS CON EL PAQUETE DEseq

X	hgnc_symbol	description
1	PTPRU	protein tyrosine phosphatase, receptor type U [Source:HGNC Symbol;Acc:HGNC:9683]
2	RAD50	RAD50 double strand break repair protein [Source:HGNC Symbol;Acc:HGNC:9816]
3	AKAP9	A-kinase anchoring protein 9 [Source:HGNC Symbol;Acc:HGNC:379]
4	RBM5	RNA binding motif protein 5 [Source:HGNC Symbol;Acc:HGNC:9902]
5	CDK7	cyclin dependent kinase 7 [Source:HGNC Symbol;Acc:HGNC:1778]
6	CDKN1B	cyclin dependent kinase inhibitor 1B [Source:HGNC Symbol;Acc:HGNC:1785]
7	NDRG1	N-myc downstream regulated 1 [Source:HGNC Symbol;Acc:HGNC:7679]
8	CARM1	coactivator associated arginine methyltransferase 1 [Source:HGNC Symbol;Acc:HGNC:23393]
9	RPP38	ribonuclease P/MRP subunit p38 [Source:HGNC Symbol;Acc:HGNC:30329]
10	ARFGEF2	ADP ribosylation factor guanine nucleotide exchange factor 2 [Source:HGNC Symbol;Acc:HGNC:15853]
11	NOXA1	NADPH oxidase activator 1 [Source:HGNC Symbol;Acc:HGNC:10668]
12	EDAR	ectodysplasin A receptor [Source:HGNC Symbol;Acc:HGNC:2895]
13	STIP1	stress induced phosphoprotein 1 [Source:HGNC Symbol;Acc:HGNC:11387]
14	CHEK1	checkpoint kinase 1 [Source:HGNC Symbol;Acc:HGNC:1925]
15	CHEK2	checkpoint kinase 2 [Source:HGNC Symbol;Acc:HGNC:16627]
16	CLTC	clathrin heavy chain [Source:HGNC Symbol;Acc:HGNC:2092]
17	CSNK1G3	casein kinase 1 gamma 3 [Source:HGNC Symbol;Acc:HGNC:2456]
18	CTNNB1	catenin beta 1 [Source:HGNC Symbol;Acc:HGNC:2514]
19	E2F1	E2F transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:3113]
20	ARID2	AT-rich interaction domain 2 [Source:HGNC Symbol;Acc:HGNC:18037]

21	EIF4G1	eukaryotic translation initiation factor 4 gamma 1 [Source:HGNC Symbol;Acc:HGNC:3296]
22	ELF1	E74 like ETS transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:3316]
23	ESR1	estrogen receptor 1 [Source:HGNC Symbol;Acc:HGNC:3467]
24	RASGEF1A	RasGEF domain family member 1A [Source:HGNC Symbol;Acc:HGNC:24246]
25	CCT5	chaperonin containing TCP1 subunit 5 [Source:HGNC Symbol;Acc:HGNC:1618]
26	CLASP2	cytoplasmic linker associated protein 2 [Source:HGNC Symbol;Acc:HGNC:17078]
27	FLT3	fms related tyrosine kinase 3 [Source:HGNC Symbol;Acc:HGNC:3765]
28	ACSL6	acyl-CoA synthetase long-chain family member 6 [Source:HGNC Symbol;Acc:HGNC:16496]
29	FMR1	fragile X mental retardation 1 [Source:HGNC Symbol;Acc:HGNC:3775]
30	ZMYND8	zinc finger MYND-type containing 8 [Source:HGNC Symbol;Acc:HGNC:9397]
31	RALGAPA1	Ral GTPase activating protein catalytic alpha subunit 1 [Source:HGNC Symbol;Acc:HGNC:17770]
32	RASGRP3	RAS guanyl releasing protein 3 [Source:HGNC Symbol;Acc:HGNC:14545]
33	ASPM	abnormal spindle microtubule assembly [Source:HGNC Symbol;Acc:HGNC:19048]
34	GATA3	GATA binding protein 3 [Source:HGNC Symbol;Acc:HGNC:4172]
35	GDI1	GDP dissociation inhibitor 1 [Source:HGNC Symbol;Acc:HGNC:4226]
36	FOXP1	forkhead box P1 [Source:HGNC Symbol;Acc:HGNC:3823]
37	AFF4	AF4/FMR2 family member 4 [Source:HGNC Symbol;Acc:HGNC:17869]
38	ARHGAP35	Rho GTPase activating protein 35 [Source:HGNC Symbol;Acc:HGNC:4591]
39	MSH6	mutS homolog 6 [Source:HGNC Symbol;Acc:HGNC:7329]
40	ANXA1	annexin A1 [Source:HGNC Symbol;Acc:HGNC:533]
41	HCFC1	host cell factor C1 [Source:HGNC Symbol;Acc:HGNC:4839]
42	HLA-A	major histocompatibility complex, class I, A [Source:HGNC Symbol;Acc:HGNC:4931]
43	FOXA1	forkhead box A1 [Source:HGNC Symbol;Acc:HGNC:5021]
44	IRS1	insulin receptor substrate 1 [Source:HGNC Symbol;Acc:HGNC:6125]
45	AR	androgen receptor [Source:HGNC Symbol;Acc:HGNC:644]

46	LCP1	lymphocyte cytosolic protein 1 [Source:HGNC Symbol;Acc:HGNC:6528]
47	LRP6	LDL receptor related protein 6 [Source:HGNC Symbol;Acc:HGNC:6698]
48	SMAD2	SMAD family member 2 [Source:HGNC Symbol;Acc:HGNC:6768]
49	MAX	MYC associated factor X [Source:HGNC Symbol;Acc:HGNC:6913]
50	MDM2	MDM2 proto-oncogene [Source:HGNC Symbol;Acc:HGNC:6973]
51	MAP3K1	mitogen-activated protein kinase kinase kinase 1 [Source:HGNC Symbol;Acc:HGNC:6848]
52	MMP1	matrix metallopeptidase 1 [Source:HGNC Symbol;Acc:HGNC:7155]
53	MSH2	mutS homolog 2 [Source:HGNC Symbol;Acc:HGNC:7325]
54	MYB	MYB proto-oncogene, transcription factor [Source:HGNC Symbol;Acc:HGNC:7545]
55	MYC	v-myc avian myelocytomatosis viral oncogene homolog [Source:HGNC Symbol;Acc:HGNC:7553]
56	MYH9	myosin heavy chain 9 [Source:HGNC Symbol;Acc:HGNC:7579]
57	MYT1	myelin transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:7622]
58	NAB1	NGFI-A binding protein 1 [Source:HGNC Symbol;Acc:HGNC:7626]
59	NF2	neurofibromin 2 [Source:HGNC Symbol;Acc:HGNC:7773]
60	NOTCH1	notch 1 [Source:HGNC Symbol;Acc:HGNC:7881]
61	NRAS	neuroblastoma RAS viral oncogene homolog [Source:HGNC Symbol;Acc:HGNC:7989]
62	NR4A2	nuclear receptor subfamily 4 group A member 2 [Source:HGNC Symbol;Acc:HGNC:7981]
63	ODC1	ornithine decarboxylase 1 [Source:HGNC Symbol;Acc:HGNC:8109]
64	PCSK6	proprotein convertase subtilisin/kexin type 6 [Source:HGNC Symbol;Acc:HGNC:8569]
65	PIK3R2	phosphoinositide-3-kinase regulatory subunit 2 [Source:HGNC Symbol;Acc:HGNC:8980]
66	PLK1	polo like kinase 1 [Source:HGNC Symbol;Acc:HGNC:9077]
67	PML	promyelocytic leukemia [Source:HGNC Symbol;Acc:HGNC:9113]
68	POLR2B	RNA polymerase II subunit B [Source:HGNC Symbol;Acc:HGNC:9188]
69	BNC2	basonuclin 2 [Source:HGNC Symbol;Acc:HGNC:30988]
70	BCOR	BCL6 corepressor [Source:HGNC Symbol;Acc:HGNC:20893]

71	FBXW7	F-box and WD repeat domain containing 7 [Source:HGNC Symbol;Acc:HGNC:16712]
72	DHTKD1	dehydrogenase E1 and transketolase domain containing 1 [Source:HGNC Symbol;Acc:HGNC:23537]
73	PRKAR1A	protein kinase cAMP-dependent type I regulatory subunit alpha [Source:HGNC Symbol;Acc:HGNC:9388]
74	BACH1	BTB domain and CNC homolog 1 [Source:HGNC Symbol;Acc:HGNC:935]
75	ZMIZ1	zinc finger MIZ-type containing 1 [Source:HGNC Symbol;Acc:HGNC:16493]
76	PPP4R3B	protein phosphatase 4 regulatory subunit 3B [Source:HGNC Symbol;Acc:HGNC:29267]
77	PTEN	phosphatase and tensin homolog [Source:HGNC Symbol;Acc:HGNC:9588]
78	MKL1	megakaryoblastic leukemia (translocation) 1 [Source:HGNC Symbol;Acc:HGNC:14334]
79	BAK1	BCL2 antagonist/killer 1 [Source:HGNC Symbol;Acc:HGNC:949]
80	CCNB1IP1	cyclin B1 interacting protein 1 [Source:HGNC Symbol;Acc:HGNC:19437]
81	RAD51	RAD51 recombinase [Source:HGNC Symbol;Acc:HGNC:9817]
82	RB1	RB transcriptional corepressor 1 [Source:HGNC Symbol;Acc:HGNC:9884]
83	CCND1	cyclin D1 [Source:HGNC Symbol;Acc:HGNC:1582]
84	BCL2	BCL2, apoptosis regulator [Source:HGNC Symbol;Acc:HGNC:990]
85	RFC4	replication factor C subunit 4 [Source:HGNC Symbol;Acc:HGNC:9972]
86	RHEB	Ras homolog enriched in brain [Source:HGNC Symbol;Acc:HGNC:10011]
87	APOBEC3G	apolipoprotein B mRNA editing enzyme catalytic subunit 3G [Source:HGNC Symbol;Acc:HGNC:17357]
88	RPL5	ribosomal protein L5 [Source:HGNC Symbol;Acc:HGNC:10360]
89	BID	BH3 interacting domain death agonist [Source:HGNC Symbol;Acc:HGNC:1050]
90	CLSPN	claspin [Source:HGNC Symbol;Acc:HGNC:19715]
91	BLM	Bloom syndrome RecQ like helicase [Source:HGNC Symbol;Acc:HGNC:1058]
92	MAP2K4	mitogen-activated protein kinase kinase 4 [Source:HGNC Symbol;Acc:HGNC:6844]
93	CERK	ceramide kinase [Source:HGNC Symbol;Acc:HGNC:19256]
94	SMARCA4	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4 [Source:HGNC Symbol;Acc:HGNC:11100]

95	SOS1	SOS Ras/Rac guanine nucleotide exchange factor 1 [Source:HGNC Symbol;Acc:HGNC:11187]
96	SP1	Sp1 transcription factor [Source:HGNC Symbol;Acc:HGNC:11205]
97	BRCA2	BRCA2, DNA repair associated [Source:HGNC Symbol;Acc:HGNC:1101]
98	STAT1	signal transducer and activator of transcription 1 [Source:HGNC Symbol;Acc:HGNC:11362]
99	AURKA	aurora kinase A [Source:HGNC Symbol;Acc:HGNC:11393]
100	TBX3	T-box 3 [Source:HGNC Symbol;Acc:HGNC:11602]
101	TCF7L2	transcription factor 7 like 2 [Source:HGNC Symbol;Acc:HGNC:11641]
102	TFDP1	transcription factor Dp-1 [Source:HGNC Symbol;Acc:HGNC:11749]
103	VEGFA	vascular endothelial growth factor A [Source:HGNC Symbol;Acc:HGNC:12680]
104	CSDE1	cold shock domain containing E1 [Source:HGNC Symbol;Acc:HGNC:29905]
105	CAD	carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase [Source:HGNC Symbol;Acc:HGNC:1424]
106	AHNAK	AHNAK nucleoprotein [Source:HGNC Symbol;Acc:HGNC:347]
107	ZNF655	zinc finger protein 655 [Source:HGNC Symbol;Acc:HGNC:30899]
108	DCAKD	dephospho-CoA kinase domain containing [Source:HGNC Symbol;Acc:HGNC:26238]
109	NUP85	nucleoporin 85 [Source:HGNC Symbol;Acc:HGNC:8734]
110	SVEP1	sushi, von Willebrand factor type A, EGF and pentraxin domain containing 1 [Source:HGNC Symbol;Acc:HGNC:15985]
111	CEP290	centrosomal protein 290 [Source:HGNC Symbol;Acc:HGNC:29021]
112	FOSL1	FOS like 1, AP-1 transcription factor subunit [Source:HGNC Symbol;Acc:HGNC:13718]
113	PHF6	PHD finger protein 6 [Source:HGNC Symbol;Acc:HGNC:18145]
114	RAD54L	RAD54-like (S. cerevisiae) [Source:HGNC Symbol;Acc:HGNC:9826]
115	CUL1	cullin 1 [Source:HGNC Symbol;Acc:HGNC:2551]
116	RUNX1	runt related transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:10471]
117	CBFB	core-binding factor beta subunit [Source:HGNC Symbol;Acc:HGNC:1539]
118	ALKBH1	alkB homolog 1, histone H2A dioxygenase [Source:HGNC Symbol;Acc:HGNC:17911]

119	KALRN	kalirin, RhoGEF kinase [Source:HGNC Symbol;Acc:HGNC:4814]
120	ACVR1B	activin A receptor type 1B [Source:HGNC Symbol;Acc:HGNC:172]
121	CDC25A	cell division cycle 25A [Source:HGNC Symbol;Acc:HGNC:1725]
122	CDC25B	cell division cycle 25B [Source:HGNC Symbol;Acc:HGNC:1726]
123	CDH1	cadherin 1 [Source:HGNC Symbol;Acc:HGNC:1748]

ANEXO VI: TABLA DE GENES DIFERENCIALMENTE EXPRESADOS OBTENIDOS CON EL PAQUETE tweeDEseq

X	hgnc_symbol	description
1	RAD50	RAD50 double strand break repair protein [Source:HGNC Symbol;Acc:HGNC:9816]
2	AKAP9	A-kinase anchoring protein 9 [Source:HGNC Symbol;Acc:HGNC:379]
3	RBM5	RNA binding motif protein 5 [Source:HGNC Symbol;Acc:HGNC:9902]
4	CDK7	cyclin dependent kinase 7 [Source:HGNC Symbol;Acc:HGNC:1778]
5	CDKN1B	cyclin dependent kinase inhibitor 1B [Source:HGNC Symbol;Acc:HGNC:1785]
6	NDRG1	N-myc downstream regulated 1 [Source:HGNC Symbol;Acc:HGNC:7679]
7	CARM1	coactivator associated arginine methyltransferase 1 [Source:HGNC Symbol;Acc:HGNC:23393]
8	RPP38	ribonuclease P/MRP subunit p38 [Source:HGNC Symbol;Acc:HGNC:30329]
9	ARFGEF2	ADP ribosylation factor guanine nucleotide exchange factor 2 [Source:HGNC Symbol;Acc:HGNC:15853]
10	NOXA1	NADPH oxidase activator 1 [Source:HGNC Symbol;Acc:HGNC:10668]
11	STIP1	stress induced phosphoprotein 1 [Source:HGNC Symbol;Acc:HGNC:11387]
12	CHEK1	checkpoint kinase 1 [Source:HGNC Symbol;Acc:HGNC:1925]
13	CHEK2	checkpoint kinase 2 [Source:HGNC Symbol;Acc:HGNC:16627]
14	CLTC	clathrin heavy chain [Source:HGNC Symbol;Acc:HGNC:2092]
15	CSNK1G3	casein kinase 1 gamma 3 [Source:HGNC Symbol;Acc:HGNC:2456]
16	E2F1	E2F transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:3113]
17	ARID2	AT-rich interaction domain 2 [Source:HGNC Symbol;Acc:HGNC:18037]
18	EIF4G1	eukaryotic translation initiation factor 4 gamma 1 [Source:HGNC Symbol;Acc:HGNC:3296]
19	ELF1	E74 like ETS transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:3316]
20	ESR1	estrogen receptor 1 [Source:HGNC Symbol;Acc:HGNC:3467]

21	RASGEF1A	RasGEF domain family member 1A [Source:HGNC Symbol;Acc:HGNC:24246]
22	DIS3	DIS3 homolog, exosome endoribonuclease and 3'-5' exoribonuclease [Source:HGNC Symbol;Acc:HGNC:20604]
23	CCT5	chaperonin containing TCP1 subunit 5 [Source:HGNC Symbol;Acc:HGNC:1618]
24	CLASP2	cytoplasmic linker associated protein 2 [Source:HGNC Symbol;Acc:HGNC:17078]
25	FLT3	fms related tyrosine kinase 3 [Source:HGNC Symbol;Acc:HGNC:3765]
26	MGA	MGA, MAX dimerization protein [Source:HGNC Symbol;Acc:HGNC:14010]
27	ACSL6	acyl-CoA synthetase long-chain family member 6 [Source:HGNC Symbol;Acc:HGNC:16496]
28	FMR1	fragile X mental retardation 1 [Source:HGNC Symbol;Acc:HGNC:3775]
29	SIRT1	sirtuin 1 [Source:HGNC Symbol;Acc:HGNC:14929]
30	ZMYND8	zinc finger MYND-type containing 8 [Source:HGNC Symbol;Acc:HGNC:9397]
31	MTOR	mechanistic target of rapamycin [Source:HGNC Symbol;Acc:HGNC:3942]
32	RALGAPA1	Ral GTPase activating protein catalytic alpha subunit 1 [Source:HGNC Symbol;Acc:HGNC:17770]
33	RASGRP3	RAS guanyl releasing protein 3 [Source:HGNC Symbol;Acc:HGNC:14545]
34	ASPM	abnormal spindle microtubule assembly [Source:HGNC Symbol;Acc:HGNC:19048]
35	GATA3	GATA binding protein 3 [Source:HGNC Symbol;Acc:HGNC:4172]
36	GDI1	GDP dissociation inhibitor 1 [Source:HGNC Symbol;Acc:HGNC:4226]
37	FOXP1	forkhead box P1 [Source:HGNC Symbol;Acc:HGNC:3823]
38	AFF4	AF4/FMR2 family member 4 [Source:HGNC Symbol;Acc:HGNC:17869]
39	ARHGAP35	Rho GTPase activating protein 35 [Source:HGNC Symbol;Acc:HGNC:4591]
40	MSH6	mutS homolog 6 [Source:HGNC Symbol;Acc:HGNC:7329]
41	ANXA1	annexin A1 [Source:HGNC Symbol;Acc:HGNC:533]
42	HCFC1	host cell factor C1 [Source:HGNC Symbol;Acc:HGNC:4839]
43	HLA-A	major histocompatibility complex, class I, A [Source:HGNC Symbol;Acc:HGNC:4931]
44	FOXA1	forkhead box A1 [Source:HGNC Symbol;Acc:HGNC:5021]
45	APC	APC, WNT signaling pathway regulator [Source:HGNC

		Symbol;Acc:HGNC:583]
46	IRS1	insulin receptor substrate 1 [Source:HGNC Symbol;Acc:HGNC:6125]
47	AR	androgen receptor [Source:HGNC Symbol;Acc:HGNC:644]
48	ARAF	A-Raf proto-oncogene, serine/threonine kinase [Source:HGNC Symbol;Acc:HGNC:646]
49	LCP1	lymphocyte cytosolic protein 1 [Source:HGNC Symbol;Acc:HGNC:6528]
50	LRP6	LDL receptor related protein 6 [Source:HGNC Symbol;Acc:HGNC:6698]
51	SMAD2	SMAD family member 2 [Source:HGNC Symbol;Acc:HGNC:6768]
52	MAX	MYC associated factor X [Source:HGNC Symbol;Acc:HGNC:6913]
53	MDM2	MDM2 proto-oncogene [Source:HGNC Symbol;Acc:HGNC:6973]
54	MAP3K1	mitogen-activated protein kinase kinase kinase 1 [Source:HGNC Symbol;Acc:HGNC:6848]
55	MLH1	mutL homolog 1 [Source:HGNC Symbol;Acc:HGNC:7127]
56	MMP1	matrix metallopeptidase 1 [Source:HGNC Symbol;Acc:HGNC:7155]
57	MSH2	mutS homolog 2 [Source:HGNC Symbol;Acc:HGNC:7325]
58	MYB	MYB proto-oncogene, transcription factor [Source:HGNC Symbol;Acc:HGNC:7545]
59	MYC	v-myc avian myelocytomatosis viral oncogene homolog [Source:HGNC Symbol;Acc:HGNC:7553]
60	MYH9	myosin heavy chain 9 [Source:HGNC Symbol;Acc:HGNC:7579]
61	NAB1	NGFI-A binding protein 1 [Source:HGNC Symbol;Acc:HGNC:7626]
62	NF1	neurofibromin 1 [Source:HGNC Symbol;Acc:HGNC:7765]
63	NF2	neurofibromin 2 [Source:HGNC Symbol;Acc:HGNC:7773]
64	NOTCH1	notch 1 [Source:HGNC Symbol;Acc:HGNC:7881]
65	NRAS	neuroblastoma RAS viral oncogene homolog [Source:HGNC Symbol;Acc:HGNC:7989]
66	NR4A2	nuclear receptor subfamily 4 group A member 2 [Source:HGNC Symbol;Acc:HGNC:7981]
67	ODC1	ornithine decarboxylase 1 [Source:HGNC Symbol;Acc:HGNC:8109]
68	PCSK6	proprotein convertase subtilisin/kexin type 6 [Source:HGNC Symbol;Acc:HGNC:8569]
69	PIK3R2	phosphoinositide-3-kinase regulatory subunit 2 [Source:HGNC Symbol;Acc:HGNC:8980]

70	PLK1	polo like kinase 1 [Source:HGNC Symbol;Acc:HGNC:9077]
71	PML	promyelocytic leukemia [Source:HGNC Symbol;Acc:HGNC:9113]
72	ATR	ATR serine/threonine kinase [Source:HGNC Symbol;Acc:HGNC:882]
73	BNC2	basonuclin 2 [Source:HGNC Symbol;Acc:HGNC:30988]
74	BCOR	BCL6 corepressor [Source:HGNC Symbol;Acc:HGNC:20893]
75	FBXW7	F-box and WD repeat domain containing 7 [Source:HGNC Symbol;Acc:HGNC:16712]
76	DHTKD1	dehydrogenase E1 and transketolase domain containing 1 [Source:HGNC Symbol;Acc:HGNC:23537]
77	PRKAR1A	protein kinase cAMP-dependent type I regulatory subunit alpha [Source:HGNC Symbol;Acc:HGNC:9388]
78	MAPK1	mitogen-activated protein kinase 1 [Source:HGNC Symbol;Acc:HGNC:6871]
79	BACH1	BTB domain and CNC homolog 1 [Source:HGNC Symbol;Acc:HGNC:935]
80	ZMIZ1	zinc finger MIZ-type containing 1 [Source:HGNC Symbol;Acc:HGNC:16493]
81	BAD	BCL2 associated agonist of cell death [Source:HGNC Symbol;Acc:HGNC:936]
82	PPP4R3B	protein phosphatase 4 regulatory subunit 3B [Source:HGNC Symbol;Acc:HGNC:29267]
83	PTEN	phosphatase and tensin homolog [Source:HGNC Symbol;Acc:HGNC:9588]
84	MKL1	megakaryoblastic leukemia (translocation) 1 [Source:HGNC Symbol;Acc:HGNC:14334]
85	BAK1	BCL2 antagonist/killer 1 [Source:HGNC Symbol;Acc:HGNC:949]
86	CCNB1IP1	cyclin B1 interacting protein 1 [Source:HGNC Symbol;Acc:HGNC:19437]
87	BAX	BCL2 associated X, apoptosis regulator [Source:HGNC Symbol;Acc:HGNC:959]
88	RAD51	RAD51 recombinase [Source:HGNC Symbol;Acc:HGNC:9817]
89	RAP1A	RAP1A, member of RAS oncogene family [Source:HGNC Symbol;Acc:HGNC:9855]
90	RB1	RB transcriptional corepressor 1 [Source:HGNC Symbol;Acc:HGNC:9884]
91	RBBP7	RB binding protein 7, chromatin remodeling factor [Source:HGNC Symbol;Acc:HGNC:9890]
92	CCND1	cyclin D1 [Source:HGNC Symbol;Acc:HGNC:1582]
93	BCL2	BCL2, apoptosis regulator [Source:HGNC Symbol;Acc:HGNC:990]
94	RFC4	replication factor C subunit 4 [Source:HGNC

		Symbol;Acc:HGNC:9972]
95	ACTB	actin beta [Source:HGNC Symbol;Acc:HGNC:132]
96	RHEB	Ras homolog enriched in brain [Source:HGNC Symbol;Acc:HGNC:10011]
97	RPGR	retinitis pigmentosa GTPase regulator [Source:HGNC Symbol;Acc:HGNC:10295]
98	RPL5	ribosomal protein L5 [Source:HGNC Symbol;Acc:HGNC:10360]
99	BID	BH3 interacting domain death agonist [Source:HGNC Symbol;Acc:HGNC:1050]
100	CLSPN	claspin [Source:HGNC Symbol;Acc:HGNC:19715]
101	BLM	Bloom syndrome RecQ like helicase [Source:HGNC Symbol;Acc:HGNC:1058]
102	MAP2K4	mitogen-activated protein kinase kinase 4 [Source:HGNC Symbol;Acc:HGNC:6844]
103	SFPQ	splicing factor proline and glutamine rich [Source:HGNC Symbol;Acc:HGNC:10774]
104	CERK	ceramide kinase [Source:HGNC Symbol;Acc:HGNC:19256]
105	SMARCA4	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4 [Source:HGNC Symbol;Acc:HGNC:11100]
106	SOS1	SOS Ras/Rac guanine nucleotide exchange factor 1 [Source:HGNC Symbol;Acc:HGNC:11187]
107	SP1	Sp1 transcription factor [Source:HGNC Symbol;Acc:HGNC:11205]
108	BRCA2	BRCA2, DNA repair associated [Source:HGNC Symbol;Acc:HGNC:1101]
109	STAT1	signal transducer and activator of transcription 1 [Source:HGNC Symbol;Acc:HGNC:11362]
110	AURKA	aurora kinase A [Source:HGNC Symbol;Acc:HGNC:11393]
111	TBX3	T-box 3 [Source:HGNC Symbol;Acc:HGNC:11602]
112	TCF7L2	transcription factor 7 like 2 [Source:HGNC Symbol;Acc:HGNC:11641]
113	TCF12	transcription factor 12 [Source:HGNC Symbol;Acc:HGNC:11623]
114	TFDP1	transcription factor Dp-1 [Source:HGNC Symbol;Acc:HGNC:11749]
115	TSC2	tuberous sclerosis 2 [Source:HGNC Symbol;Acc:HGNC:12363]
116	VEGFA	vascular endothelial growth factor A [Source:HGNC Symbol;Acc:HGNC:12680]
117	XRCC3	X-ray repair cross complementing 3 [Source:HGNC Symbol;Acc:HGNC:12830]
118	CSDE1	cold shock domain containing E1 [Source:HGNC Symbol;Acc:HGNC:29905]

119	CAD	carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase [Source:HGNC Symbol;Acc:HGNC:1424]
120	AHNAK	AHNAK nucleoprotein [Source:HGNC Symbol;Acc:HGNC:347]
121	ZNF655	zinc finger protein 655 [Source:HGNC Symbol;Acc:HGNC:30899]
122	DCAKD	dephospho-CoA kinase domain containing [Source:HGNC Symbol;Acc:HGNC:26238]
123	NUP85	nucleoporin 85 [Source:HGNC Symbol;Acc:HGNC:8734]
124	SVEP1	sushi, von Willebrand factor type A, EGF and pentraxin domain containing 1 [Source:HGNC Symbol;Acc:HGNC:15985]
125	CEP290	centrosomal protein 290 [Source:HGNC Symbol;Acc:HGNC:29021]
126	FOSL1	FOS like 1, AP-1 transcription factor subunit [Source:HGNC Symbol;Acc:HGNC:13718]
127	PHF6	PHD finger protein 6 [Source:HGNC Symbol;Acc:HGNC:18145]
128	RAD54L	RAD54-like (S. cerevisiae) [Source:HGNC Symbol;Acc:HGNC:9826]
129	CUL1	cullin 1 [Source:HGNC Symbol;Acc:HGNC:2551]
130	RUNX1	runt related transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:10471]
131	CBFB	core-binding factor beta subunit [Source:HGNC Symbol;Acc:HGNC:1539]
132	FADD	Fas associated via death domain [Source:HGNC Symbol;Acc:HGNC:3573]
133	ALKBH1	alkB homolog 1, histone H2A dioxygenase [Source:HGNC Symbol;Acc:HGNC:17911]
134	KALRN	kalirin, RhoGEF kinase [Source:HGNC Symbol;Acc:HGNC:4814]
135	ACVR1B	activin A receptor type 1B [Source:HGNC Symbol;Acc:HGNC:172]
136	MAP3K13	mitogen-activated protein kinase kinase kinase 13 [Source:HGNC Symbol;Acc:HGNC:6852]
137	NCOR1	nuclear receptor corepressor 1 [Source:HGNC Symbol;Acc:HGNC:7672]
138	AQR	aquarius intron-binding spliceosomal factor [Source:HGNC Symbol;Acc:HGNC:29513]
139	CDC25A	cell division cycle 25A [Source:HGNC Symbol;Acc:HGNC:1725]
140	CDC25B	cell division cycle 25B [Source:HGNC Symbol;Acc:HGNC:1726]
141	GOLGA5	golgin A5 [Source:HGNC Symbol;Acc:HGNC:4428]
142	CDC42	cell division cycle 42 [Source:HGNC Symbol;Acc:HGNC:1736]
143	CDH1	cadherin 1 [Source:HGNC Symbol;Acc:HGNC:1748]

ANEXO VII: TABLA DEL SUBGRUPO DE GENES QUE MEJOR CLASIFICÓ LOS DATOS EN EL EXPERIMENTO 2 DE ML

X	hgnc_symbol	description
1	COQ7	coenzyme Q7, hydroxylase [Source:HGNC Symbol;Acc:HGNC:2244]
2	HEXIM1	hexamethylene bisacetamide inducible 1 [Source:HGNC Symbol;Acc:HGNC:24953]
3	FAM114A2	family with sequence similarity 114 member A2 [Source:HGNC Symbol;Acc:HGNC:1333]
4	MAGED2	MAGE family member D2 [Source:HGNC Symbol;Acc:HGNC:16353]
5	ADCY9	adenylate cyclase 9 [Source:HGNC Symbol;Acc:HGNC:240]
6	CIRBP	cold inducible RNA binding protein [Source:HGNC Symbol;Acc:HGNC:1982]
7	LRRC56	leucine rich repeat containing 56 [Source:HGNC Symbol;Acc:HGNC:25430]
8	BORCS7	BLOC-1 related complex subunit 7 [Source:HGNC Symbol;Acc:HGNC:23516]
9	TTC8	tetratricopeptide repeat domain 8 [Source:HGNC Symbol;Acc:HGNC:20087]
10	DEGS2	delta 4-desaturase, sphingolipid 2 [Source:HGNC Symbol;Acc:HGNC:20113]
11	UBXN10	UBX domain protein 10 [Source:HGNC Symbol;Acc:HGNC:26354]
12	CPEB2	cytoplasmic polyadenylation element binding protein 2 [Source:HGNC Symbol;Acc:HGNC:21745]
13	C9orf116	chromosome 9 open reading frame 116 [Source:HGNC Symbol;Acc:HGNC:28435]
14	RUNDC1	RUN domain containing 1 [Source:HGNC Symbol;Acc:HGNC:25418]
15	C1orf64	chromosome 1 open reading frame 64 [Source:HGNC Symbol;Acc:HGNC:28339]
16	AGR3	anterior gradient 3, protein disulphide isomerase family member [Source:HGNC Symbol;Acc:HGNC:24167]
17	AK8	adenylate kinase 8 [Source:HGNC Symbol;Acc:HGNC:26526]
18	DACH1	dachshund family transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:2663]
19	SAMD15	sterile alpha motif domain containing 15 [Source:HGNC Symbol;Acc:HGNC:18631]
20	ADAMTS15	ADAM metallopeptidase with thrombospondin type 1 motif 15

		[Source:HGNC Symbol;Acc:HGNC:16305]
21	ABAT	4-aminobutyrate aminotransferase [Source:HGNC Symbol;Acc:HGNC:23]
22	SMIM14	small integral membrane protein 14 [Source:HGNC Symbol;Acc:HGNC:27321]
23	CCDC125	coiled-coil domain containing 125 [Source:HGNC Symbol;Acc:HGNC:28924]
24	SUSD3	sushi domain containing 3 [Source:HGNC Symbol;Acc:HGNC:28391]
25	ERBB4	erb-b2 receptor tyrosine kinase 4 [Source:HGNC Symbol;Acc:HGNC:3432]
26	ESR1	estrogen receptor 1 [Source:HGNC Symbol;Acc:HGNC:3467]
27	FBP1	fructose-bisphosphatase 1 [Source:HGNC Symbol;Acc:HGNC:3606]
28	ARSG	arylsulfatase G [Source:HGNC Symbol;Acc:HGNC:24102]
29	KDM4B	lysine demethylase 4B [Source:HGNC Symbol;Acc:HGNC:29136]
30	TBC1D9	TBC1 domain family member 9 [Source:HGNC Symbol;Acc:HGNC:21710]
31	MED13L	mediator complex subunit 13 like [Source:HGNC Symbol;Acc:HGNC:22962]
32	SLC7A8	solute carrier family 7 member 8 [Source:HGNC Symbol;Acc:HGNC:11066]
33	KLHDC2	kelch domain containing 2 [Source:HGNC Symbol;Acc:HGNC:20231]
34	FUT8	fucosyltransferase 8 [Source:HGNC Symbol;Acc:HGNC:4019]
35	GAMT	guanidinoacetate N-methyltransferase [Source:HGNC Symbol;Acc:HGNC:4136]
36	FBXL5	F-box and leucine rich repeat protein 5 [Source:HGNC Symbol;Acc:HGNC:13602]
37	GATA3	GATA binding protein 3 [Source:HGNC Symbol;Acc:HGNC:4172]
38	GFRA1	GDNF family receptor alpha 1 [Source:HGNC Symbol;Acc:HGNC:4243]
39	C5AR2	complement component 5a receptor 2 [Source:HGNC Symbol;Acc:HGNC:4527]
40	FOXA1	forkhead box A1 [Source:HGNC Symbol;Acc:HGNC:5021]
41	APBB2	amyloid beta precursor protein binding family B member 2 [Source:HGNC Symbol;Acc:HGNC:582]
42	ACADSB	acyl-CoA dehydrogenase, short/branched chain [Source:HGNC Symbol;Acc:HGNC:91]
43	TRIM23	tripartite motif containing 23 [Source:HGNC Symbol;Acc:HGNC:660]

44	CCDC103	coiled-coil domain containing 103 [Source:HGNC Symbol;Acc:HGNC:32700]
45	FAM187A	family with sequence similarity 187 member A [Source:HGNC Symbol;Acc:HGNC:35153]
46	AFF3	AF4/FMR2 family member 3 [Source:HGNC Symbol;Acc:HGNC:6473]
47	MAPT	microtubule associated protein tau [Source:HGNC Symbol;Acc:HGNC:6893]
48	PHOSPHO2	phosphatase, orphan 2 [Source:HGNC Symbol;Acc:HGNC:28316]
49	DCTN4	dynactin subunit 4 [Source:HGNC Symbol;Acc:HGNC:15518]
50	CDK17	cyclin dependent kinase 17 [Source:HGNC Symbol;Acc:HGNC:8750]
51	CSAD	cysteine sulfinic acid decarboxylase [Source:HGNC Symbol;Acc:HGNC:18966]
52	EVL	Enah/Vasp-like [Source:HGNC Symbol;Acc:HGNC:20234]
53	CXXC5	CXXC finger protein 5 [Source:HGNC Symbol;Acc:HGNC:26943]
54	P4HTM	prolyl 4-hydroxylase, transmembrane [Source:HGNC Symbol;Acc:HGNC:28858]
55	GIN1	gypsy retrotransposon integrase 1 [Source:HGNC Symbol;Acc:HGNC:25959]
56	RALGPS2	Ral GEF with PH domain and SH3 binding motif 2 [Source:HGNC Symbol;Acc:HGNC:30279]
57	SYBU	syntabulin [Source:HGNC Symbol;Acc:HGNC:26011]
58	MKL2	MKL1/myocardin like 2 [Source:HGNC Symbol;Acc:HGNC:29819]
59	EPB41L5	erythrocyte membrane protein band 4.1 like 5 [Source:HGNC Symbol;Acc:HGNC:19819]
60	WDR19	WD repeat domain 19 [Source:HGNC Symbol;Acc:HGNC:18340]
61	ANKRA2	ankyrin repeat family A member 2 [Source:HGNC Symbol;Acc:HGNC:13208]
62	BBS1	Bardet-Biedl syndrome 1 [Source:HGNC Symbol;Acc:HGNC:966]
63	BBS4	Bardet-Biedl syndrome 4 [Source:HGNC Symbol;Acc:HGNC:969]
64	RAD17	RAD17 checkpoint clamp loader component [Source:HGNC Symbol;Acc:HGNC:9807]
65	SIAH2	siah E3 ubiquitin protein ligase 2 [Source:HGNC Symbol;Acc:HGNC:10858]
66	SKP1	S-phase kinase associated protein 1 [Source:HGNC Symbol;Acc:HGNC:10899]

67	SLC22A5	solute carrier family 22 member 5 [Source:HGNC Symbol;Acc:HGNC:10969]
68	BTF3	basic transcription factor 3 [Source:HGNC Symbol;Acc:HGNC:1125]
69	TLE3	transducin like enhancer of split 3 [Source:HGNC Symbol;Acc:HGNC:11839]
70	XBP1	X-box binding protein 1 [Source:HGNC Symbol;Acc:HGNC:12801]
71	CA12	carbonic anhydrase 12 [Source:HGNC Symbol;Acc:HGNC:1371]
72	MAP3K12	mitogen-activated protein kinase kinase kinase 12 [Source:HGNC Symbol;Acc:HGNC:6851]
73	DNALI1	dynein axonemal light intermediate chain 1 [Source:HGNC Symbol;Acc:HGNC:14353]
74	MLPH	melanophilin [Source:HGNC Symbol;Acc:HGNC:29643]
75	ARMT1	acidic residue methyltransferase 1 [Source:HGNC Symbol;Acc:HGNC:17872]
76	ZNF552	zinc finger protein 552 [Source:HGNC Symbol;Acc:HGNC:26135]
77	THSD4	thrombospondin type 1 domain containing 4 [Source:HGNC Symbol;Acc:HGNC:25835]
78	SPEF2	sperm flagellar 2 [Source:HGNC Symbol;Acc:HGNC:26293]
79	BBOF1	basal body orientation factor 1 [Source:HGNC Symbol;Acc:HGNC:19855]
80	CCDC170	coiled-coil domain containing 170 [Source:HGNC Symbol;Acc:HGNC:21177]
81	IFT88	intraflagellar transport 88 [Source:HGNC Symbol;Acc:HGNC:20606]
82	TIGD6	tigger transposable element derived 6 [Source:HGNC Symbol;Acc:HGNC:18332]
83	APH1B	aph-1 homolog B, gamma-secretase subunit [Source:HGNC Symbol;Acc:HGNC:24080]
84	DNAL1	dynein axonemal light chain 1 [Source:HGNC Symbol;Acc:HGNC:23247]
85	NUDT12	nudix hydrolase 12 [Source:HGNC Symbol;Acc:HGNC:18826]
86	DYNLRB2	dynein light chain roadblock-type 2 [Source:HGNC Symbol;Acc:HGNC:15467]
87	PCBD2	pterin-4 alpha-carbinolamine dehydratase 2 [Source:HGNC Symbol;Acc:HGNC:24474]
88	ATRIP	ATR interacting protein [Source:HGNC Symbol;Acc:HGNC:33499]
89	TMEM25	transmembrane protein 25 [Source:HGNC Symbol;Acc:HGNC:25890]

90	SPRYD3	SPRY domain containing 3 [Source:HGNC Symbol;Acc:HGNC:25920]
91	BTRC	beta-transducin repeat containing E3 ubiquitin protein ligase [Source:HGNC Symbol;Acc:HGNC:1144]
92	EFCAB12	EF-hand calcium binding domain 12 [Source:HGNC Symbol;Acc:HGNC:28061]
93	RABEP1	rabaptin, RAB GTPase binding effector protein 1 [Source:HGNC Symbol;Acc:HGNC:17677]
94	NEK9	NIMA related kinase 9 [Source:HGNC Symbol;Acc:HGNC:18591]
95	ATP6AP1L	ATPase H+ transporting accessory protein 1 like [Source:HGNC Symbol;Acc:HGNC:28091]
96	SYTL4	synaptotagmin like 4 [Source:HGNC Symbol;Acc:HGNC:15588]
97	CELSR1	cadherin EGF LAG seven-pass G-type receptor 1 [Source:HGNC Symbol;Acc:HGNC:1850]
98	FNIP1	folliculin interacting protein 1 [Source:HGNC Symbol;Acc:HGNC:29418]
99	GREB1	growth regulation by estrogen in breast cancer 1 [Source:HGNC Symbol;Acc:HGNC:24885]
100	DAZAP2	DAZ associated protein 2 [Source:HGNC Symbol;Acc:HGNC:2684]
101	KIAA0141	KIAA0141 [Source:HGNC Symbol;Acc:HGNC:28969]

ANEXO VIII: TABLA DEL SUBGRUPO DE GENES QUE MEJOR CLASIFICÓ LOS DATOS EN EL EXPERIMENTO 3 DE ML

hgnc_symbol	description
TOM1	target of myb1 membrane trafficking protein [Source:HGNC Symbol;Acc:HGNC:11982]
RAD50	RAD50 double strand break repair protein [Source:HGNC Symbol;Acc:HGNC:9816]
AKAP9	A-kinase anchoring protein 9 [Source:HGNC Symbol;Acc:HGNC:379]
CDK4	cyclin dependent kinase 4 [Source:HGNC Symbol;Acc:HGNC:1773]
CDK7	cyclin dependent kinase 7 [Source:HGNC Symbol;Acc:HGNC:1778]
CDKN1B	cyclin dependent kinase inhibitor 1B [Source:HGNC Symbol;Acc:HGNC:1785]
STAG1	stromal antigen 1 [Source:HGNC Symbol;Acc:HGNC:11354]
CARM1	coactivator associated arginine methyltransferase 1 [Source:HGNC Symbol;Acc:HGNC:23393]
RPP38	ribonuclease P/MRP subunit p38 [Source:HGNC Symbol;Acc:HGNC:30329]
USP16	ubiquitin specific peptidase 16 [Source:HGNC Symbol;Acc:HGNC:12614]
CTCF	CCCTC-binding factor [Source:HGNC Symbol;Acc:HGNC:13723]
NOXA1	NADPH oxidase activator 1 [Source:HGNC Symbol;Acc:HGNC:10668]
EDAR	ectodysplasin A receptor [Source:HGNC Symbol;Acc:HGNC:2895]
CHD4	chromodomain helicase DNA binding protein 4 [Source:HGNC Symbol;Acc:HGNC:1919]
CHEK1	checkpoint kinase 1 [Source:HGNC Symbol;Acc:HGNC:1925]
CSNK1G3	casein kinase 1 gamma 3 [Source:HGNC Symbol;Acc:HGNC:2456]
JAKMIP1	janus kinase and microtubule interacting protein 1 [Source:HGNC Symbol;Acc:HGNC:26460]
DAG1	dystroglycan 1 [Source:HGNC Symbol;Acc:HGNC:2666]
DHX15	DEAH-box helicase 15 [Source:HGNC Symbol;Acc:HGNC:2738]
EGFR	epidermal growth factor receptor [Source:HGNC Symbol;Acc:HGNC:3236]
ARID2	AT-rich interaction domain 2 [Source:HGNC Symbol;Acc:HGNC:18037]
EIF4G1	eukaryotic translation initiation factor 4 gamma 1 [Source:HGNC Symbol;Acc:HGNC:3296]
ELF1	E74 like ETS transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:3316]
ERBB2	erb-b2 receptor tyrosine kinase 2 [Source:HGNC Symbol;Acc:HGNC:3430]
ERCC2	ERCC excision repair 2, TFIIH core complex helicase subunit [Source:HGNC Symbol;Acc:HGNC:3434]

AKT1	AKT serine/threonine kinase 1 [Source:HGNC Symbol;Acc:HGNC:391]
ESR1	estrogen receptor 1 [Source:HGNC Symbol;Acc:HGNC:3467]
MECOM	MDS1 and EVI1 complex locus [Source:HGNC Symbol;Acc:HGNC:3498]
FAU	FAU, ubiquitin like and ribosomal protein S30 fusion [Source:HGNC Symbol;Acc:HGNC:3597]
RASGEF1A	RasGEF domain family member 1A [Source:HGNC Symbol;Acc:HGNC:24246]
FER	FER tyrosine kinase [Source:HGNC Symbol;Acc:HGNC:3655]
CCT5	chaperonin containing TCP1 subunit 5 [Source:HGNC Symbol;Acc:HGNC:1618]
MYCBP2	MYC binding protein 2, E3 ubiquitin protein ligase [Source:HGNC Symbol;Acc:HGNC:23386]
CLASP2	cytoplasmic linker associated protein 2 [Source:HGNC Symbol;Acc:HGNC:17078]
FLT3	fms related tyrosine kinase 3 [Source:HGNC Symbol;Acc:HGNC:3765]
ACSL6	acyl-CoA synthetase long-chain family member 6 [Source:HGNC Symbol;Acc:HGNC:16496]
FMR1	fragile X mental retardation 1 [Source:HGNC Symbol;Acc:HGNC:3775]
ZMYND8	zinc finger MYND-type containing 8 [Source:HGNC Symbol;Acc:HGNC:9397]
RALGAPA1	Ral GTPase activating protein catalytic alpha subunit 1 [Source:HGNC Symbol;Acc:HGNC:17770]
ASPM	abnormal spindle microtubule assembly [Source:HGNC Symbol;Acc:HGNC:19048]
ERAL1	Era like 12S mitochondrial rRNA chaperone 1 [Source:HGNC Symbol;Acc:HGNC:3424]
GDI1	GDP dissociation inhibitor 1 [Source:HGNC Symbol;Acc:HGNC:4226]
FOXP1	forkhead box P1 [Source:HGNC Symbol;Acc:HGNC:3823]
FILIP1	filamin A interacting protein 1 [Source:HGNC Symbol;Acc:HGNC:21015]
ARHGAP35	Rho GTPase activating protein 35 [Source:HGNC Symbol;Acc:HGNC:4591]
MSH6	mutS homolog 6 [Source:HGNC Symbol;Acc:HGNC:7329]
ANXA1	annexin A1 [Source:HGNC Symbol;Acc:HGNC:533]
FOXA1	forkhead box A1 [Source:HGNC Symbol;Acc:HGNC:5021]
APC	APC, WNT signaling pathway regulator [Source:HGNC Symbol;Acc:HGNC:583]
IRS1	insulin receptor substrate 1 [Source:HGNC Symbol;Acc:HGNC:6125]
AR	androgen receptor [Source:HGNC Symbol;Acc:HGNC:644]
ARAF	A-Raf proto-oncogene, serine/threonine kinase [Source:HGNC Symbol;Acc:HGNC:646]
JAK1	Janus kinase 1 [Source:HGNC Symbol;Acc:HGNC:6190]
LRP6	LDL receptor related protein 6 [Source:HGNC Symbol;Acc:HGNC:6698]
ARNTL	aryl hydrocarbon receptor nuclear translocator like [Source:HGNC

	Symbol;Acc:HGNC:701]
SMAD1	SMAD family member 1 [Source:HGNC Symbol;Acc:HGNC:6767]
MAX	MYC associated factor X [Source:HGNC Symbol;Acc:HGNC:6913]
MAP3K1	mitogen-activated protein kinase kinase kinase 1 [Source:HGNC Symbol;Acc:HGNC:6848]
MMP1	matrix metallopeptidase 1 [Source:HGNC Symbol;Acc:HGNC:7155]
MSH2	mutS homolog 2 [Source:HGNC Symbol;Acc:HGNC:7325]
MYB	MYB proto-oncogene, transcription factor [Source:HGNC Symbol;Acc:HGNC:7545]
MYH9	myosin heavy chain 9 [Source:HGNC Symbol;Acc:HGNC:7579]
MYH11	myosin heavy chain 11 [Source:HGNC Symbol;Acc:HGNC:7569]
ATF1	activating transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:783]
MYT1	myelin transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:7622]
ATIC	5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase [Source:HGNC Symbol;Acc:HGNC:794]
ATM	ATM serine/threonine kinase [Source:HGNC Symbol;Acc:HGNC:795]
NF1	neurofibromin 1 [Source:HGNC Symbol;Acc:HGNC:7765]
NF2	neurofibromin 2 [Source:HGNC Symbol;Acc:HGNC:7773]
CNOT3	CCR4-NOT transcription complex subunit 3 [Source:HGNC Symbol;Acc:HGNC:7879]
NRAS	neuroblastoma RAS viral oncogene homolog [Source:HGNC Symbol;Acc:HGNC:7989]
NR4A2	nuclear receptor subfamily 4 group A member 2 [Source:HGNC Symbol;Acc:HGNC:7981]
ODC1	ornithine decarboxylase 1 [Source:HGNC Symbol;Acc:HGNC:8109]
PCSK6	proprotein convertase subtilisin/kexin type 6 [Source:HGNC Symbol;Acc:HGNC:8569]
PAX5	paired box 5 [Source:HGNC Symbol;Acc:HGNC:8619]
CDK12	cyclin dependent kinase 12 [Source:HGNC Symbol;Acc:HGNC:24224]
PIGR	polymeric immunoglobulin receptor [Source:HGNC Symbol;Acc:HGNC:8968]
PIK3CB	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit beta [Source:HGNC Symbol;Acc:HGNC:8976]
PIK3R2	phosphoinositide-3-kinase regulatory subunit 2 [Source:HGNC Symbol;Acc:HGNC:8980]
PML	promyelocytic leukemia [Source:HGNC Symbol;Acc:HGNC:9113]
BNC2	basonuclin 2 [Source:HGNC Symbol;Acc:HGNC:30988]
BCOR	BCL6 corepressor [Source:HGNC Symbol;Acc:HGNC:20893]
DHTKD1	dehydrogenase E1 and transketolase domain containing 1 [Source:HGNC Symbol;Acc:HGNC:23537]

MAPK1	mitogen-activated protein kinase 1 [Source:HGNC Symbol;Acc:HGNC:6871]
NUP107	nucleoporin 107 [Source:HGNC Symbol;Acc:HGNC:29914]
ZMIZ1	zinc finger MIZ-type containing 1 [Source:HGNC Symbol;Acc:HGNC:16493]
PPP4R3B	protein phosphatase 4 regulatory subunit 3B [Source:HGNC Symbol;Acc:HGNC:29267]
PTGS1	prostaglandin-endoperoxide synthase 1 [Source:HGNC Symbol;Acc:HGNC:9604]
SRGAP1	SLIT-ROBO Rho GTPase activating protein 1 [Source:HGNC Symbol;Acc:HGNC:17382]
BAK1	BCL2 antagonist/killer 1 [Source:HGNC Symbol;Acc:HGNC:949]
CCNB1IP1	cyclin B1 interacting protein 1 [Source:HGNC Symbol;Acc:HGNC:19437]
BAX	BCL2 associated X, apoptosis regulator [Source:HGNC Symbol;Acc:HGNC:959]
RAC1	ras-related C3 botulinum toxin substrate 1 (rho family, small GTP binding protein Rac1) [Source:HGNC Symbol;Acc:HGNC:9801]
RAD51	RAD51 recombinase [Source:HGNC Symbol;Acc:HGNC:9817]
RALA	RAS like proto-oncogene A [Source:HGNC Symbol;Acc:HGNC:9839]
RB1	RB transcriptional corepressor 1 [Source:HGNC Symbol;Acc:HGNC:9884]
BCL2	BCL2, apoptosis regulator [Source:HGNC Symbol;Acc:HGNC:990]
RFC4	replication factor C subunit 4 [Source:HGNC Symbol;Acc:HGNC:9972]
ACTB	actin beta [Source:HGNC Symbol;Acc:HGNC:132]
RHEB	Ras homolog enriched in brain [Source:HGNC Symbol;Acc:HGNC:10011]
RPGR	retinitis pigmentosa GTPase regulator [Source:HGNC Symbol;Acc:HGNC:10295]
RPL5	ribosomal protein L5 [Source:HGNC Symbol;Acc:HGNC:10360]
BID	BH3 interacting domain death agonist [Source:HGNC Symbol;Acc:HGNC:1050]
CLSPN	claspin [Source:HGNC Symbol;Acc:HGNC:19715]
BLM	Bloom syndrome RecQ like helicase [Source:HGNC Symbol;Acc:HGNC:1058]
MAP2K4	mitogen-activated protein kinase kinase 4 [Source:HGNC Symbol;Acc:HGNC:6844]
ITSN1	intersectin 1 [Source:HGNC Symbol;Acc:HGNC:6183]
CERK	ceramide kinase [Source:HGNC Symbol;Acc:HGNC:19256]
SMARCA4	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4 [Source:HGNC Symbol;Acc:HGNC:11100]
SOS1	SOS Ras/Rac guanine nucleotide exchange factor 1 [Source:HGNC Symbol;Acc:HGNC:11187]
SP1	Sp1 transcription factor [Source:HGNC Symbol;Acc:HGNC:11205]
SPTAN1	spectrin alpha, non-erythrocytic 1 [Source:HGNC Symbol;Acc:HGNC:11273]
BRCA2	BRCA2, DNA repair associated [Source:HGNC Symbol;Acc:HGNC:1101]
ZFP36L1	ZFP36 ring finger protein like 1 [Source:HGNC Symbol;Acc:HGNC:1107]

ZFP36L2	ZFP36 ring finger protein like 2 [Source:HGNC Symbol;Acc:HGNC:1108]
AURKA	aurora kinase A [Source:HGNC Symbol;Acc:HGNC:11393]
TFDP1	transcription factor Dp-1 [Source:HGNC Symbol;Acc:HGNC:11749]
TP53	tumor protein p53 [Source:HGNC Symbol;Acc:HGNC:11998]
TSC1	tuberous sclerosis 1 [Source:HGNC Symbol;Acc:HGNC:12362]
VEGFA	vascular endothelial growth factor A [Source:HGNC Symbol;Acc:HGNC:12680]
CSDE1	cold shock domain containing E1 [Source:HGNC Symbol;Acc:HGNC:29905]
CAD	carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase [Source:HGNC Symbol;Acc:HGNC:1424]
AHNAK	AHNAK nucleoprotein [Source:HGNC Symbol;Acc:HGNC:347]
NUP85	nucleoporin 85 [Source:HGNC Symbol;Acc:HGNC:8734]
SVEP1	sushi, von Willebrand factor type A, EGF and pentraxin domain containing 1 [Source:HGNC Symbol;Acc:HGNC:15985]
CEP290	centrosomal protein 290 [Source:HGNC Symbol;Acc:HGNC:29021]
CHD9	chromodomain helicase DNA binding protein 9 [Source:HGNC Symbol;Acc:HGNC:25701]
FOSL1	FOS like 1, AP-1 transcription factor subunit [Source:HGNC Symbol;Acc:HGNC:13718]
ARID1A	AT-rich interaction domain 1A [Source:HGNC Symbol;Acc:HGNC:11110]
CAST	calpastatin [Source:HGNC Symbol;Acc:HGNC:1515]
BAP1	BRCA1 associated protein 1 [Source:HGNC Symbol;Acc:HGNC:950]
CASP3	caspase 3 [Source:HGNC Symbol;Acc:HGNC:1504]
CASP9	caspase 9 [Source:HGNC Symbol;Acc:HGNC:1511]
RAD54L	RAD54-like (S. cerevisiae) [Source:HGNC Symbol;Acc:HGNC:9826]
CUL1	cullin 1 [Source:HGNC Symbol;Acc:HGNC:2551]
USP38	ubiquitin specific peptidase 38 [Source:HGNC Symbol;Acc:HGNC:20067]
RUNX1	runt related transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:10471]
CBFB	core-binding factor beta subunit [Source:HGNC Symbol;Acc:HGNC:1539]
ALKBH1	alkB homolog 1, histone H2A dioxygenase [Source:HGNC Symbol;Acc:HGNC:17911]
KALRN	kalirin, RhoGEF kinase [Source:HGNC Symbol;Acc:HGNC:4814]
ACVR1B	activin A receptor type 1B [Source:HGNC Symbol;Acc:HGNC:172]
AQR	aquarius intron-binding spliceosomal factor [Source:HGNC Symbol;Acc:HGNC:29513]
SETDB1	SET domain bifurcated 1 [Source:HGNC Symbol;Acc:HGNC:10761]
G3BP2	G3BP stress granule assembly factor 2 [Source:HGNC Symbol;Acc:HGNC:30291]
CDC25A	cell division cycle 25A [Source:HGNC Symbol;Acc:HGNC:1725]

GOLGA5	golgin A5 [Source:HGNC Symbol;Acc:HGNC:4428]
THRAP3	thyroid hormone receptor associated protein 3 [Source:HGNC Symbol;Acc:HGNC:22964]
MED12	mediator complex subunit 12 [Source:HGNC Symbol;Acc:HGNC:11957]
CDC42	cell division cycle 42 [Source:HGNC Symbol;Acc:HGNC:1736]
CDH1	cadherin 1 [Source:HGNC Symbol;Acc:HGNC:1748]

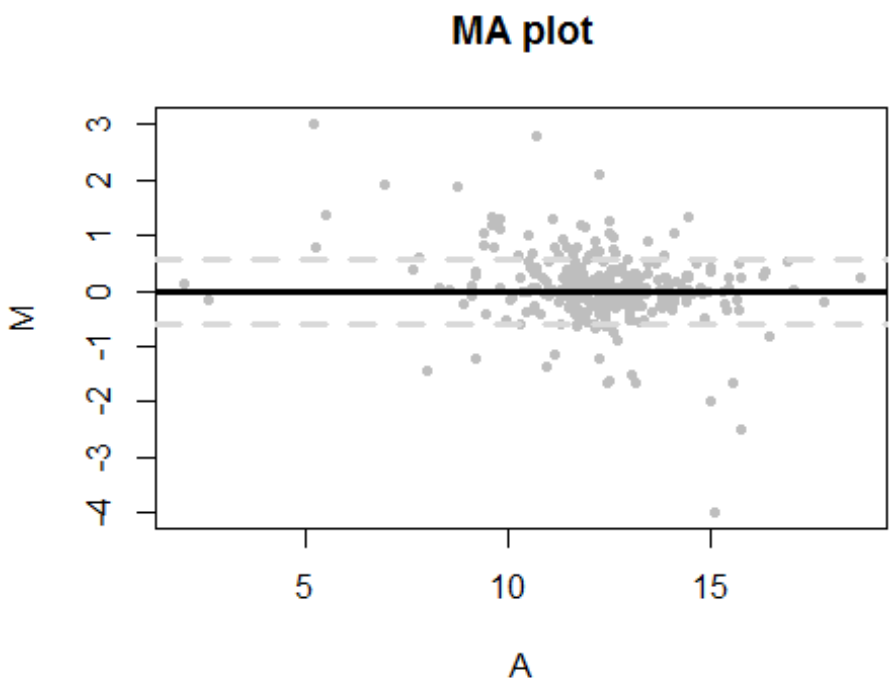
ANEXO IX: CÓDIGO UTILIZADO EN EL PROTOCOLO DE BIOCONDUCTOR

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

library(tweeDEseqCountData)

library(Biobase)
library(tweeDEseq)


library(edgeR)

library(DESeq)

library(S4Vectors)

library("curl")

library("plyr")

library("jsonlite")

library("devtools")

library("RCurl")

library("bitops")
library("urltools")

library("rWikiPathways")
library(biomaRt)


#datos de expresión
load("C:/Users/JOSE/Desktop/Máster
Jose/TFM/PRUEBAS/EJEMPLO/data/breastTCGA.RData")
breastTCGA

## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 20532 features, 757 samples
##  element names: exprs
## protocolData: none
## phenoData
##  sampleNames: TCGA-A1-A0SB TCGA-A1-A0SD ... TCGA-GI-A2C8 (757
##   total)
##  varLabels: bcr_patient_barcode additional_pharmaceutical_therapy
##   ... year_of_initial_pathologic_diagnosis (85 total)
##  varMetadata: labelDescription
## featureData
##  featureNames: 1 2 ... 20532 (20532 total)
##  fvarLabels: entrezIDs
##  fvarMetadata: labelDescription
## experimentData: use 'experimentData(object)'
## Annotation:
```

```
counts<-exprs(breastTCGA)
pData<-pData(breastTCGA)
#datos de anotación
data("annotEnsembl63")


dim(counts)

## [1] 20532  757

dim(annotEnsembl63)

## [1] 50451   8

#eliminamos de la base de datos fenotípicos los pacientes que no son ni negativos ni positivos en el status del ER
elim_indet<-which(pData$breast_carcinoma_estrogen_receptor_status == "Indeterminate"|pData$breast_carcinoma_estrogen_receptor_status == "Not Performed"|pData$breast_carcinoma_estrogen_receptor_status == "Performed but Not Available")
pData_new<-pData[-elim_indet, ]
dim(pData_new)

## [1] 721 85

#ahora eliminamos los mismos pacientes pero en la base de datos de expresión (counts)
#trasponemos la base de datos
counts_t<-t(counts)
counts  new<-counts  t[-elim_indet,]
counts_new<-t(counts_new)
dim(counts_new)

## [1] 20532  721

####CHECKEAMOOS!
stopifnot(all(rownames(pData_new) == colnames(counts_new)))



#intersecto para obtener únicamente los genes humanos del RNAseq

#eliminamos todas las filas que no tengan un valor de EntrezID
annotEnsembl63_na.omit<-annotEnsembl63[!(is.na(annotEnsembl63$EntrezID)),]
dim(annotEnsembl63_na.omit)

## [1] 19866   8

genes.ok<-intersect(rownames(counts_new), annotEnsembl63_na.omit[,5])
length(genes.ok)

## [1] 18257

counts.ok<-counts_new[genes.ok, ]
annot.ok<-annotEnsembl63_na.omit[match(genes.ok,
annotEnsembl63_na.omit$EntrezID),]
```

```
####### NORMALIZACIÓN #######

##RPKM normalization
width<-annot.ok$Length
counts.rpkm<-t(t(counts.ok/width*1000)/colSums(counts.ok)*1e6)
#TMM normalization
counts.tmm<-normalizeCounts(counts.ok, method = "TMM")


row.names(counts.ok) <- row.names(annot.ok)
annotation<-annot.ok[,c("Length", "GCcontent")]
counts.cqn<-normalizeCounts(counts.ok, method = "cqn", annot = annotation)

row.names(counts.cqn)<-row.names(counts.rpkm)


###BREAST CANCER GENES####
intogen<-read.table("C:\\Users\\JOSE\\Desktop\\Máster Jose\\TFM\\intogen-BRCA-
drivers-data.tsv", sep = "\t", header = T)
genes_intogen<-as.vector(intogen$SYMBOL) #184 intogenes
wikigene<-getXrefList(pathway = "WP1984", systemCode = "H")
genes_breast_cancer<-unique(c(genes_intogen, wikigene)) #313 genes en hgnc
symbol que no están repetidos
mart<-useMart(biomart="ensembl", dataset = "hsapiens_gene_ensembl")
ID_BC_genes<-getBM(attributes = c("ensembl_gene_id",
"entrezgene","hgnc_symbol"), filters = "hgnc_symbol", values =genes_breast_cancer,
mart = mart)
BC_genes_entrezid<-unique(as.vector(ID_BC_genes$entrezgene)) #308 BREAST
CANCER genes in entrezID
genes.brcan<-intersect(rownames(counts.cqn), BC_genes_entrezid)
genes.brcan<-as.character(genes.brcan)
counts_brcan<-counts.cqn[genes.brcan, ]
annot_brcan<-annot.ok[match(genes.brcan, annot.ok$EntrezID),]



################################   ANÁLISIS DE EXPRESIÓN
#########################

group<-as.factor(pData_new$breast_carcinoma_estrogen_receptor_status)
table(group)

## group
## Negative Positive
##   166   555

#edgeR
d<-DGEList(counts = counts_brcan, group = group)
d<-estimateCommonDisp(d)
d<-estimateTagwiseDisp(d)
res.edgeR.common<-exactTest(d,pair=c("Positive", "Negative"), dispersion =
"common")
res.edgeR.tagwise<-exactTest(d, pair=c("Positive", "Negative"), dispersion ="tagwise")

#DESeq
cds <- newCountDataSet(counts_brcan, group)
```

```
cds <- estimateSizeFactors(cds)
cds <- estimateDispersions(cds)

## Warning: glm.fit: algorithm did not converge

res.DEseq <- nbinomTest(cds, "Positive", "Negative")

#tweeDEseq
res.tweeDE <- tweeDE(counts_brcan, group,pair =c("Positive", "Negative"))

#####################################  RESULTADOS
##########################################
topTags(res.edgeR.common)

## Comparison of groups: Negative-Positive
##      logFC  logCPM    PValue      FDR
## 2099 -4.036173 13.201620 0.000000e+00 0.000000e+00
## 1956  2.842718 8.937368 0.000000e+00 0.000000e+00
## 10913 2.798331 3.365696 0.000000e+00 0.000000e+00
## 4312  2.081870 10.292939 7.891863e-234 5.938627e-232
## 2625 -2.534320 13.873091 1.378186e-201 8.296681e-200
## 23305 1.944089 5.101864 2.412014e-197 1.210027e-195
## 8061  1.885408 6.917181 1.277821e-188 5.494632e-187
## 3169 -2.039803 13.123211 1.476634e-140 5.555835e-139
## 4661 -1.818782 6.164080 4.559353e-115 1.524850e-113
## 5046 -1.701918 10.520449 4.268332e-103 1.284768e-101

sum(topTags(res.edgeR.common, n=Inf)$table[,4]<0.05)

## [1] 202

sum(res.DEseq$padj<0.05)

## [1] 181

sum(res.tweeDE$pval.adjust<0.05)

## [1] 188

MAplot(res.tweeDE,
   cex=0.7,
   log2fc.cutoff=log2(1.5),
   main="MA plot")
```




```
#################### COMPARACIÓN ENTRE LOS GENES DE LOS TRES TIPOS DE MÉTODOS ####################################3

edger<-res.edgeR.common$table
edger_filt<-edger[which(edger$PValue<0.001),]
dim(edger_filt)

## [1] 161  3

head(edger_filt)

##       logFC  logCPM    PValue
## 23305 1.9440885 5.101864 2.412014e-197
## 91  -0.5864205 10.288551 1.056227e-15
## 27125 -0.5896827 11.355598 7.337494e-16
## 79026 -0.8495744 14.582484 3.273425e-30
## 10142 -0.5050453 11.107589 3.786049e-12
## 8846 -0.4030196 8.550692 2.513847e-08

write.csv(edger_filt, file = "edgeR_genes.csv")



dim(res.DEseq)

## [1] 301  8

deseq_filt<-res.DEseq[which(res.DEseq$padj<0.001),]
dim(deseq_filt)

## [1] 123  8

head(deseq_filt)
```

```
##    id  baseMean  baseMeanA baseMeanB foldChange log2FoldChange
## 3 23305  119.3054  71.84697  277.9769 3.8690136   1.9519658
## 5   91 4415.6819 4780.58937 3195.6601 0.6684657   -0.5810745
## 6 27125 9286.5131 10037.96422 6774.1315 0.6748511   -0.5673588
## 7 79026 87305.8151 96996.36363 54906.6920 0.5660696   -0.8209487
## 9 10142 7796.5969 8352.28285 5938.7312 0.7110309   -0.4920158
## 11 8846 1325.1392 1401.25892 1070.6427 0.7640577   -0.3882464
##      pval     padj
## 3 1.229223e-11 6.379242e-11
## 5 7.093775e-13 4.448388e-12
## 6 9.735540e-15 7.711572e-14
## 7 6.233553e-33 1.876299e-31
## 9 4.470789e-11 2.206078e-10
## 11 2.377459e-04 6.169095e-04

write.csv(deseq_filt, file = "DEseq_genes.csv")


dim(res.tweeDE)

## [1] 301  7

tweede_filt<-res.tweeDE[which(res.tweeDE$pval.adjust<0.001),]
dim(tweede_filt)

## [1] 143  7

head(tweede_filt)

## Comparison of groups: Negative - Positive
## Showing genes ranked by P-value
## Minimum absolute log2 fold-change of 0
## Maximum adjusted P-value of 1
##    overallMean  Negative  Positive   log2fc   stat     pval
## 79026  91035.632 56834.205 101265.249 -0.8333078 12.904878 4.224864e-38
## 27125  9675.069  7023.277 10468.218 -0.5757996 11.183167 4.929876e-29
## 91    4595.745  3321.608  4976.838 -0.5833474 7.968795 1.602294e-15
## 10142  8164.580  6210.163  8749.144 -0.4945108 6.593795 4.287236e-11
## 23305   123.877  286.771   75.155 1.9319594 4.729099 2.255180e-06
## 60   425119.766 478835.187 409053.532 0.2272395 4.296645 1.734024e-05
##    pval.adjust
## 79026 3.179211e-36
## 27125 1.483893e-27
## 91  1.303488e-14
## 10142 2.048346e-10
## 23305 6.590382e-06
## 60  4.499494e-05

write.csv(tweede_filt, file = "tweeDE_genes.csv")


intersect_1<-intersect(rownames(edger_filt),deseq_filt[,1])
intersect_2<-intersect(intersect_1, rownames(tweede_filt))
write.csv(intersect_2, file = "common_genes.csv")
length(intersect_2)

## [1] 116
```