Detección de ARNs Circulares y Estudio de su Implicación en Adenocarcinoma Pulmonar

Alfonso Sánchez-Macián Pérez
Máster universitario en Bioinformática y bioestadística UOC-UB
Genómica Computacional

Amadís Pagès Pinós
Carles Ventura Royo

02/01/2018

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Detección de ARNs Circulares y Estudio de su Implicación en Adenocarcinoma Pulmonar*
Nombre del autor:	*Alfonso Sánchez-Macián Pérez*
Nombre del consultor/a:	*Amadís Pagès Pinós*
Nombre del PRA:	*Carles Ventura Royo*
Fecha de entrega (mm/aaaa):	01/2018
Titulación::	Máster universitario en Bioinformática y bioestadística UOC-UB
Área del Trabajo Final:	*Genómica computacional*
Idioma del trabajo:	*Español*
Palabras clave	*CircRNA, adenocarcinoma pulmonar, machine learning*

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

La finalidad del presente trabajo es la generación de un modelo bioinformático que permita establecer un conjunto de circRNAs que actúen como biomarcadores moleculares para el adenocarcinoma pulmonar, capaz de clasificar muestras como sanas o tumorales.

Aunque se descubrieron en los años 70, ha sido recientemente cuando se ha identificado la relación entre un incremento o reducción de la expresión de los circRNAs con la proliferación de algunos tumores. Dentro de los diferentes tipos de cáncer, el de pulmón tiene una alta incidencia. Entre los tipos de tumores pulmonares, el adenocarcinoma pulmonar es el más frecuente. Aunque los métodos diagnósticos han ido mejorando con los años, una eficaz y rápida detección puede reducir la mortalidad permitiendo detectar estos tumores en sus estadios iniciales.

Para conseguir la finalidad del proyecto, se ha generado un *pipeline* para identificar circRNAs a partir de dos herramientas existentes (CIRI2 y CircExplorer2), consolidando los resultados. Este pipeline se ha aplicado a un conjunto de muestras sanas y tumorales proporcionadas por un estudio previo y se ha generado un informe integrado.

Dicho informe se ha utilizado para construir y validar un modelo basado en árboles de decisión usando la métrica de *information gain*. Con el modelo se han identificado un conjunto de circRNAs susceptibles para ser usados como biomarcadores y se han contrastado contra información previa disponible en bases de datos.

El resultado muestra que el *pipeline* funciona y es capaz de dar como resultado

varios circRNAs prometedores, que deberían ser validados en estudios posteriores.

Abstract (in English, 250 words or less):

The goal of this project is the generation of a bioinformatics model capable of identifying circRNA biomarkers for the lung adenocarcinoma in order to classify samples as normal or tumoral.

Although circRNAs were discovered in the 1970s, their relation to the proliferation of tumor cells has only been revealed recently. Among the different types, lung cancer has a high incidence, with lung adenocarcinoma being the most frequent of lung cancers. Even though diagnostic methods have been refined during the years, a more efficient and quick detection procedure can reduce the mortality by spotting these tumors in their initial stages.

To achieve the expected project goals, a pipeline to identify circRNAs using two existing tools (CIRI2 and CircExplorer2) has been generated, consolidating the results produced by both applications. This pipeline has been applied to a set of normal and tumoral samples provided by a previous research study, and an integrated report has been produced.

This report has been used to build and validate a decision-tree based model using the information gain metric. A set of candidate circRNAs to be used as biomarkers has been identified with this model and they have been checked against existing databases.

The result shows that the complete pipeline works and it is able to produce, as a result, several potentially interesting circRNAs that should be validated in future studies.

Índice

Lista de figuras

Lista de tablas

1. Introducción

1.1 Contexto y justificación del Trabajo

La temática del proyecto consiste en la búsqueda de circRNAs que puedan ser utilizados como potenciales biomarcadores del adenocarcinoma pulmonar con el fin de proporcionar herramientas adicionales de diagnóstico.

Para ello se propone la utilización de un conjunto de métodos de identificación de circRNAs sobre muestras de tejido tumoral y sano de varios pacientes que estén disponibles en estudios publicados en bases de datos como The Cancer Genome Atlas [1]. A partir de la selección de unos valores coherentes obtenidos de dichos experimentos en cuanto a la información de circRNAs, se define la aplicación de métodos basados en machine learning, como los árboles de decisión, y medidas como Information Gain (IG) para resolver el problema de clasificación de muestras como sanas o tumorales en función de la diferencia en la expresión de dichos circRNAs (sobre/infrarregulación).

Adenocarcinoma pulmonar

Dentro de los diferentes tipos de cáncer, el de pulmón tiene una alta incidencia [2], siendo el tercero más diagnosticado en España en 2015 sin separar por sexo (tercero en varones y cuarto en mujeres) y el primero si se considera todo el mundo. La mortalidad de este tipo de tumores es alta, siendo el primero de los tumores con más fallecimientos en el mundo en 2012 y en España en 2014. Respecto a la prevalencia, debido a esta alta mortalidad, en 2012 se estimaba en sexto lugar entre los tumores en España. Aunque los métodos diagnósticos han ido mejorando con los años (lo que ha hecho subir la incidencia), una eficaz y rápida detección puede reducir la mortalidad permitiendo detectar estos tumores en sus estadios iniciales. Entre los tipos de tumores pulmonares, el adenocarcinoma pulmonar es el más frecuente.

Mediante el uso de la genómica computacional es posible identificar moléculas de ARN con diferente expresión génica en distintos transcriptomas. La disponibilidad de muestras de tejido no tumoral y sus equivalentes afectadas por un determinado tumor para el mismo individuo facilita el estudio de las diferencias en expresión en tejidos tumorales.

circRNAs

Los CircRNAs son un tipo de Ácido Ribonucléico (ARN) que se descubrió inicialmente en los años 70 en bacterias, virus y en células eucariotas, pero se supuso que eran un producto del procesado erróneo del material genético [3] [4]. En el año 2012 se descubrió que este tipo de ARN era la forma

predominante de expresión en un gran número de genes humanos [5]. La causa de pasar tanto tiempo desapercibido es su conformación especial en forma de ARN circular (con un bucle cerrado covalente) sin las típicas caperuza (CAP) y cola poly(A) que se incluyen en los ARN lineales, escapando por tanto, de las técnicas basadas en amplificación de ARN poliadenilado [4][6][7][8][9]. Se considera un ARN no codificante de cadena larga (lncRNA). Esto significa que las moléculas no se traducen en proteínas (aunque algunos experimentos plantean dudas sobre esta afirmación [10][11]) y están compuestas por un número grande de nucleótidos (por encima de 200).

Los circRNAs se forman debido a la creación de un enlace covalente especial cabeza a cola denominado “backsplice” [8]. Estos enlaces se producen entre sitios de empalme que son, generalmente, límites de exones o zonas con señales de este tipo reconocibles por el sistema encargado de realizar el corte y empalme (*spliceosome*). Se genera, por tanto, un ARN circular que puede ser detectado por la existencia de este tipo de uniones en las que el orden de los exones (o las secuencias) parece estar invertido. Un ejemplo de este tipo de uniones frente a las uniones canónicas lineales del ARN lineal se puede ver en la figura 1.

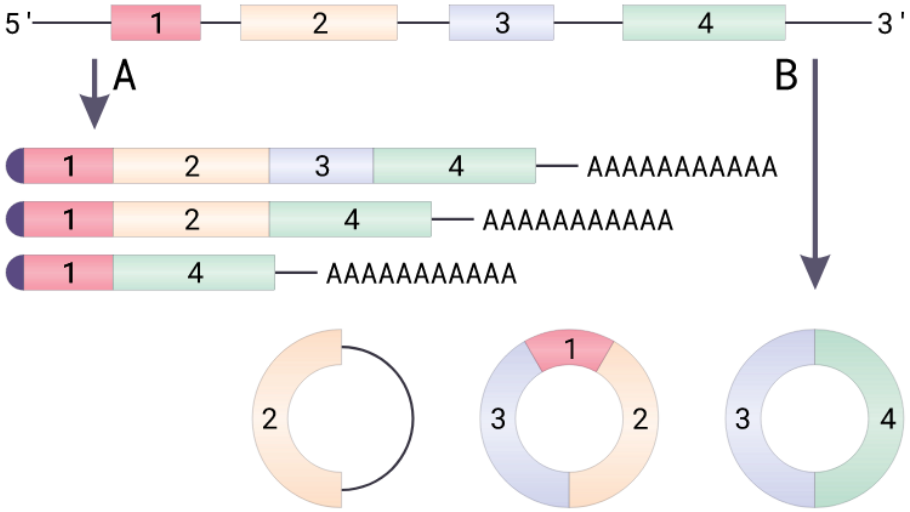


Figura 1. ARN lineal frente a ARN circular en función del tipo de empalme.[1]

circRNAs y cáncer

De entre las funciones identificadas para los circRNA [3][6][7][9] se encuentran las de actuar como estructuras para el ensamblaje de complejos proteicos, secuestro de proteínas de su localización nativa, modular la expresión de los genes de los que proceden, regular el empalmamiento (splicing) alternativo, interacción ARN-proteína y funcionar como esponjas de miRNA. Además, estudios recientes han mostrado que su incremento o reducción influye en la proliferación de algunos tumores, pudiendo actuar como biomarcadores

[1] Imagen extraída de https://upload.wikimedia.org/wikipedia/commons/4/42/CircRNA.svg con licencia Creative Commons Attribution-Share Alike 4.0 International creada por Helixitta.

tumorales gracias a su estabilidad, especificidad por tejido/órgano y etapa de desarrollo y diversidad [3][7].

La aplicación de algoritmos de machine learning y herramientas de identificación de circRNA a los transcriptomas relacionados con el adenocarcinoma pulmonar en bases de datos públicas puede permitir la selección de un conjunto de circRNAs susceptibles de poder ser utilizados como biomarcadores moleculares de este tumor tras una validación posterior.

1.2 Objetivos del Trabajo

Entre los objetivos generales de este proyecto destacan los siguientes:

1. Generación de un modelo bioinformático que permita establecer un conjunto de circRNAs que actúen como biomarcadores moleculares para el adenocarcinoma pulmonar, capaz de clasificar muestras de tejido pulmonar como sanas o tumorales.

Los objetivos específicos del proyecto se resumen a continuación:

1. Desarrollar un *pipeline* que permita identificar circRNAs.
2. Aplicar el *pipeline* a un conjunto de datos de secuenciación provinientes de muestras de tejido pulmonar de pacientes de adenocarcinoma pulmonar, para obtener información de perfiles de expresión de circRNAs en muestras pareadas de tejido tumoral y sano.
3. Procesar los datos obtenidos de expresión génica generando un modelo que facilite la clasificación de muestras como sanas y tumorales.
4. Presentar, a partir del modelo, un conjunto inicial de circRNAs que sean susceptibles de uso como biomarcadores tumorales para el adenocarcinoma pulmonar, con el fin de un futuro análisis y validación más detallado en posteriores proyectos.
5. Generar la documentación del proyecto y realizar la presentación del mismo.

El objetivo prioritario es el 1.4, por lo que se definió al inicio del proyecto que si era necesario acortar el alcance del mismo se intentaría dar prioridad a este sobre los anteriores, con acciones como utilizar una única herramienta/algoritmo en el *pipeline* de los objetivos 1.1 y 1.2 o realizar la aplicación sobre un número reducido de datos.

1.3 Enfoque y método seguido

Para obtener los objetivos presentados existen diferentes enfoques posibles.

En primer lugar, el pipeline para identificar circRNAs puede desarrollarse a partir de una o varias de las herramientas existentes, o generando una nueva herramienta. Esta última opción es más compleja, requiere un trabajo extenso, no asegura que el resultado sea mejor que las herramientas existentes y

distraería al proyecto del objetivo final de identificación de biomarcadores basados en ARNs circulares. Es importante destacar, por consiguiente, que la creación de una nueva herramienta no es el objetivo de este proyecto. Por otro lado, la divergencia en cuanto a resultados de las diferentes herramientas existentes sugiere que seleccionar una única opción disminuye la confianza en los resultados obtenidos. Por ello, se plantea la utilización de un número reducido de estas herramientas, buscando resultados de consenso y agregando resultados individuales. En este sentido, [8] compara 12 algoritmos diferentes, mientras que [12] hace una comparación exhaustiva de 11 algoritmos disponibles en 2017. Las conclusiones de este último estudio indican que CIRI [13], CircExplorer [14] y KNIFE [15] proporcionan un mayor equilibrio entre sensibilidad y precisión. Parte de la información ofrecida sobre estos algoritmos en cada uno de los dos estudios se muestra en las Tablas 1 y 2. Además, cada uno de ellos utiliza una estrategia diferente de identificación de circRNAs. En términos de recursos necesarios para su ejecución, puede haber problemas a la hora de ejecutar estos algoritmos debido a sus requisitos de uso de memoria. Además KNIFE parece requerir un mayor espacio, así como tiempo de ejecución mayor que los otros dos. En principio se seleccionarán estos 3 algoritmos en sus últimas versiones.

Tabla 1. Información sobre circExplorer, CIRI y KNIFE en [8]

Algoritmo	Capacidades	Puntos ciegos	Tasa de validación
circExplorer	Alta sensibilidad y bajo consumo de memoria comparado con similares	circRNAs que usen exones no anotados o múltiples genes	7/7 validados por combinación de varios métodos (incluido RT-PCR)
CIRI	Permite detectar circRNAs usando exones pequeños	circRNAs que usen señales de empalmamiento no canónicas	73% para circRNAs con más de 5 reads mediante PCR. 5/5 para exones de menos de 70 nucleótidos
KNIFE	Mejor validación estadística y permite enfoques dependientes e independientes de anotaciones	circRNAs en regiones de variación genómica o entre exones.	13/13 validados por resistencia a RNase R. 14/14 por PCR. 5/5 de novo por PCR y secuenciación Sanger.

Tabla 2. Información sobre circExplorer, CIRI y KNIFE en [12]

Algoritmo	Enfoque	Origen genómico
circExplorer	Identificación de uniones *backsplicing* desde la información de alineamiento.	Exones e intrones
CIRI	Identificación de uniones *backsplicing* desde la información de alineamiento.	Exones, intrones e intergénico
KNIFE	Construcción de todas los posibles candidatos exón-exón desordenadas y realiza el alineamiento frente a ellos.	Exones (intrones e intergénico en modo *de novo*)

Por otro lado, para la obtención de muestras de adenocarcinoma pulmonar, la única posibilidad en este momento y dadas las dimensiones del proyecto

consiste en tomar la información de algún repositorio público. The Cancer Genome Atlas es un claro ejemplo de este tipo de repositorios y dispone de datos de muestras emparejadas de pacientes con el tipo de tumor seleccionado. Existen también otros estudios [16] que proporcionan un número de muestras suficientemente amplio para los objetivos de este proyecto.

Para la construcción del modelo que facilite la clasificación de las muestras, se utilizarán algoritmos de Machine Learning basados en árboles de decisión y el uso de la métrica de Information Gain. Estos algoritmos mantienen un equilibrio entre simplicidad y eficiencia, están suficientemente probados en este campo de la genómica computacional y permiten identificar biomarcadores a partir de diferencias de expresión de circRNAs.

1.4 Planificación del Trabajo

Los objetivos del proyecto se dividen en las siguientes tareas:

1.1.1 Identificación de algoritmos y herramientas de identificación de circRNAs. Así como su disponibilidad y licencia de uso.

1.1.2 Análisis de la información disponible en The Cancer Genome Atlas. La finalidad es verificar la existencia, para los datos de adenocarcinoma pulmonar, del formato requerido por las herramientas existentes o su necesidad de conversión. También las limitaciones legales o normativas respecto al uso de dichos datos.

1.1.3 Generación del *pipeline* a partir de las herramientas seleccionadas. Que permita comparar y extraer datos coherentes.

1.2.1 División de la información en un grupo de entrenamiento y uno de test. Considerando la información disponible sobre adenocarcinoma pulmonar en The Cancer Genome Atlas.

1.2.2 Aplicación del *pipeline* definido al conjunto de datos de entrenamiento y al de test.

1.3.1 Construcción del modelo de Árboles de decisión e Information Gain aplicándolo al conjunto de entrenamiento.

1.3.2 Verificación del modelo con el conjunto de test.

1.4.1 Extracción de un conjunto de circRNAs con diferente expresión, a partir del modelo creado.

1.4.2 Comprobación de existencia de información sobre los circRNAs obtenidos.

1.5.1 Generación de Informe de seguimiento: Desarrollo del proyecto Fase I.

1.5.2 Generación de Informe de seguimiento: Desarrollo del proyecto Fase II.

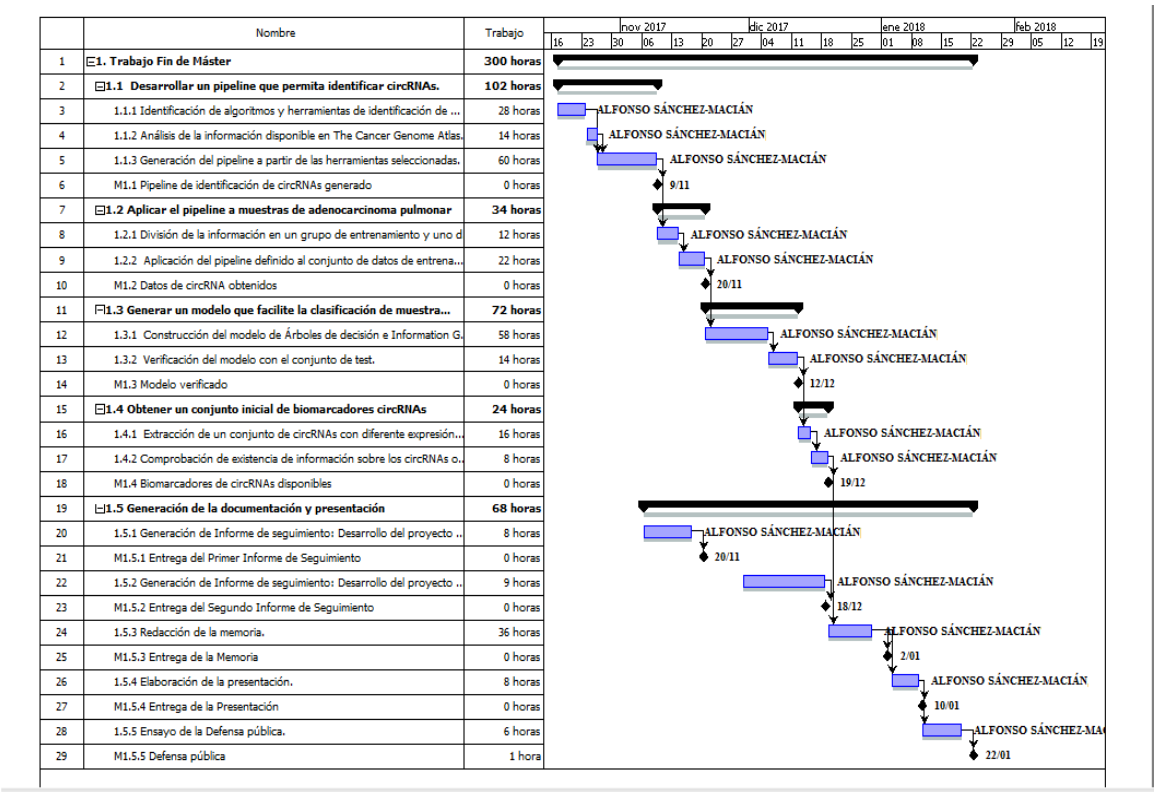
1.5.3 Redacción de la memoria.

1.5.4 Elaboración de la presentación.

1.5.5 Defensa pública.

Se realizó una planificación con la herramienta ProjectLibre. Para ello se definió el calendario de trabajo indicando los días festivos. Las tareas previamente comentadas junto con los hitos se presentan en un diagrama de Gantt asignando recursos y asignación de tiempos por tarea.

A continuación se muestra el trabajo en horas por cada objetivo y el diagrama de Gantt de las tareas (figura 2), que se han identificado en el apartado anterior. Como se observa, se han adaptado las tareas para cuadrarlas con las entregas del proyecto.



	Nombre	Trabajo
1	**1. Trabajo Fin de Máster**	**300 horas**
2	**1.1 Desarrollar un pipeline que permita identificar circRNAs.**	**102 horas**
3	1.1.1 Identificación de algoritmos y herramientas de identificación de ...	28 horas
4	1.1.2 Análisis de la información disponible en The Cancer Genome Atlas.	14 horas
5	1.1.3 Generación del pipeline a partir de las herramientas seleccionadas.	60 horas
6	M1.1 Pipeline de identificación de circRNAs generado	0 horas
7	**1.2 Aplicar el pipeline a muestras de adenocarcinoma pulmonar**	**34 horas**
8	1.2.1 División de la información en un grupo de entrenamiento y uno d	12 horas
9	1.2.2 Aplicación del pipeline definido al conjunto de datos de entrena...	22 horas
10	M1.2 Datos de circRNA obtenidos	0 horas
11	**1.3 Generar un modelo que facilite la clasificación de muestra...**	**72 horas**
12	1.3.1 Construcción del modelo de Árboles de decisión e Information G.	58 horas
13	1.3.2 Verificación del modelo con el conjunto de test.	14 horas
14	M1.3 Modelo verificado	0 horas
15	**1.4 Obtener un conjunto inicial de biomarcadores circRNAs**	**24 horas**
16	1.4.1 Extracción de un conjunto de circRNAs con diferente expresión...	16 horas
17	1.4.2 Comprobación de existencia de información sobre los circRNAs o..	8 horas
18	M1.4 Biomarcadores de circRNAs disponibles	0 horas
19	**1.5 Generación de la documentación y presentación**	**68 horas**
20	1.5.1 Generación de Informe de seguimiento: Desarrollo del proyecto ...	8 horas
21	M1.5.1 Entrega del Primer Informe de Seguimiento	0 horas
22	1.5.2 Generación de Informe de seguimiento: Desarrollo del proyecto ...	9 horas
23	M1.5.2 Entrega del Segundo Informe de Seguimiento	0 horas
24	1.5.3 Redacción de la memoria.	36 horas
25	M1.5.3 Entrega de la Memoria	0 horas
26	1.5.4 Elaboración de la presentación.	8 horas
27	M1.5.4 Entrega de la Presentación	0 horas
28	1.5.5 Ensayo de la Defensa pública.	6 horas
29	M1.5.5 Defensa pública	1 hora

Figura 2. Diagrama de Gantt de las tareas del proyecto

Se definieron tareas iniciales de identificación y análisis para asignar horas a tareas en las que el recurso que implementará el proyecto no dispone de experiencia.

1.5 Breve sumario de productos obtenidos

Se han definido los siguientes hitos en el proyecto:

> M1.1 Pipeline de identificación de circRNAs generado. Se dispone de un pipeline software generado para esa finalidad.
>
> M1.2 Datos de circRNA obtenidos. Se han procesado las muestras de adenocarcinoma pulmonar con el pipeline previamente construido,
>
> M1.3 Modelo verificado. El modelo de clasificación de muestras basado en Árboles de decisión ha sido generado y verificado.
>
> M1.4 Biomarcadores de circRNAs disponibles. Se dispone de una lista de biomarcadores de tipo circRNA.
>
> M1.5.1 Entrega del Primer Informe de Seguimiento.
>
> M1.5.2 Entrega del Segundo Informe de Seguimiento.
>
> M1.5.3 Entrega de la Memoria.
>
> M1.5.4 Entrega de la Presentación.
>
> M1.5.5 Defensa pública.

Como consecuencia de estos hitos, al final del proyecto se proporcionan, como resultado, los siguientes entregables:

- Memoria del proyecto. Este documento. Identifica la información acerca del desarrollo del proyecto y los resultados obtenidos.
- Producto. Es el conjunto de herramientas desarrolladas en el proyecto que incluyen tanto el pipeline de identificación de circRNAs, como los scripts para construir los modelos de árboles de decisión.
- Presentación virtual. Es la presentación que se realizará ante el Tribunal del Proyecto.
- Autoevaluación del proyecto. Realizada por parte del alumno.

Además, se identifican los siguientes entregables iniciales o intermedios:

- Plan de proyecto.
- Informes intermedios de seguimiento.

1.6 Breve descripción de los otros capítulos de la memoria

El resto de los capítulos de la memoria se encuentran organizados de la siguiente manera:

- Capítulo 2: Identificación de circRNAs. El segundo capítulo describe los métodos empleados y los resultados obtenidos en la identificación de circRNAs. En concreto describe la fuente de información utilizada, así como las herramientas empleadas en la identificación de circRNAs de muestras tumorales y sanas, y el método de integración de información de las mismas.
- Capítulo 3: Modelo de árbol de decisión. El tercer capítulo describe los métodos empleados y los resultados obtenidos en la generación del modelo de árbol de decisión. En concreto detalla el proceso utilizado para generar el modelo de árbol de decisión y su verificación, junto con información de los resultados obtenidos.
- Capítulo 4: Discusión. El cuarto capítulo hace un análisis inicial del conjunto de circRNAs identificados.
- Capítulo 5: Conclusiones. El quinto capítulo ofrece las conclusiones y sugiere líneas de trabajo futuro.
- Finalmente, se incluye el Glosario, la Bibliografía y los Anexos.

2. Identificación de circRNAs

Este capítulo describe la fuente de información utilizada, así como las herramientas empleadas en la identificación de circRNAs de muestras tumorales y sanas, y el método de integración de información de las mismas.

2.1 Secuenciación, alineamiento e identificación de circRNAs

Las técnicas de secuenciación de ARN y ADN han evolucionado a lo largo de los últimos años de forma muy rápida, reduciendo la complejidad y el coste asociado a las mismas. Frente a la técnica original desarrollada por Frederick Sanger en la década de los 70 del siglo pasado, la secuenciación de nueva generación o Next-generation sequencing permite procesar de forma paralela y masiva una gran cantidad de información [17]. Esta secuenciación se basa en que los nucleótidos que se unen a una hebra determinada durante un proceso de síntesis controlado emiten una señal característica que puede ser debida a una molécula fluorescente, o a un cambio de PH, por ejemplo. En caso de secuenciar ADN, éste se divide inicialmente en pequeños fragmentos que se unen a un adaptador consistente en unos pocos nucleótidos. Cada uno de esos fragmentos se coloca en una superficie (*flow cell* por ejemplo) y se realizan reacciones simultáneas de forma paralela en todos ellos. Una máquina es capaz de detectar dichas señales de forma simultánea y generar una secuencia que, si existe un genoma de referencia, se mapeará al mismo y se detectarán variantes. Si no existe dicho genoma (análisis *de novo*), se ensamblarán para crear una propuesta de genoma.

Algunas técnicas permiten elegir entre secuenciación *single-read,* en la que la lectura se realiza en un solo sentido (a partir de uno de los extremos), o *pair-ended,* en la que se realiza una segunda lectura comenzando por el otro extremo del segmento a secuenciar. La secuenciación *pair-ended* tiene más posibilidades de alinearse con una referencia, mejora la calidad total del conjunto facilita el descubrimiento de fusiones entre genes y tránscritos nuevos [17].

En cuanto a la información generada en la secuenciación y el formato de los ficheros, los datos en “crudo” (*raw data*) contienen la información de las lecturas realizadas sobre los segmentos de secuencia junto con un valor de calidad asignado a dichas lecturas [18]. Dependiendo del fabricante del modelo de máquina utilizado para la secuenciación, existen diferentes formatos utilizados para distribuir la información de secuenciación como SFF (Standard Flowgram Format) de Roche, HDF5 (Hierarchical Data Format) utilizado por PacBio y FASTQ con sus variantes propietarias usadas por Illumina, pero también por otros fabricantes.

El Sequence Read Archive (SRA) [18] es una iniciativa y base de datos que permite acceder a información de datos de secuencias biológicas generadas por investigadores, con el fin de facilitar la reproducibilidad de los estudios y realizar nuevos hallazgos basados en datos agregados. Para compartir esa información, los datos originales enviados en los formatos comentados previamente, se transformas a un formato específico definido en el contexto de

dicho proyecto, con extensión ".sra", que incluye tanto los datos de secuenciación como la información de calidad de las bases secuenciadas. El kit de herramientas utilizadas para convertir los formatos originales a ".sra" puede utilizarse también para realizar la conversión inversa. Esto será necesario en el trabajo actual como se verá en la siguiente sección.

El alineamiento de las lecturas para asociarlas con el genoma de referencia puede realizarse con alguna de las múltiples herramientas disponibles para ello. En cuanto a las estrategias de dichas herramientas, originalmente se basaban en tablas de búsqueda *hash* (por ejemplo MAQ [19]), para posteriormente evolucionar al uso de otros algoritmos como la transformación Burrows-Wheeler (por ejemplo BWA [20]). Esa transformación reversible realiza una reordenación de las secuencias para facilitar su compresión y reducir los requisitos de memoria en el proceso de alineamiento.

Respecto a la identificación de ARNs circulares, existen diferentes herramientas, como se indicó en el capítulo 1, y siguen enfoques distintos. Algunas tratan de encontrar una coincidencia de los alineamientos con circRNAs ya identificados previamente y validados. Otras realizan un análisis *de novo*, detectando, por ejemplo, señales de alineamiento no lineal y señales de *splicing*. Finalmente, las últimas versiones de algunos paquetes realizan primero la identificación de circRNAs ya anotados, para después utilizar el resto de segmentos para completar un análisis *de novo*.

2.2 Fuente de información sobre muestras y transformación de formato

La propuesta inicial del proyecto consistía en obtener la información sobre los datos de secuencia de las muestras de adenocarcinoma pulmonar del servidor de The Cancer Genome Atlas. Sin embargo, en la fase 1 se materializó el riesgo inicialmente definido como "Fallo en la disponibilidad de los datos de muestras de adenocarcinoma pulmonar" correspondiente a dicho servidor. Se identificó que, aunque existen datos procesados disponibles para todo el público, los formatos requeridos por las herramientas a utilizar obligan a un proceso de registro de la institución y el investigador principal que puede dilatarse bastante en el tiempo. Este problema se descubrió gracias a la tarea definida 1.1.2: "Análisis de la información disponible en The Cancer Genome Atlas. La finalidad es verificar la existencia, para los datos de adenocarcinoma pulmonar, del formato requerido por las herramientas existentes o su necesidad de conversión. También las limitaciones legales o normativas respecto al uso de dichos datos."

Para resolver este inconveniente, se identificaron fuentes de datos alternativas junto al director del proyecto. Finalmente se utilizaron datos disponibles del estudio [16], que son ficheros *paired-end* (basados en lecturas pareadas, desde ambos extremos de cada fragmento) de un estudio sobre adenocarcinoma pulmonar. Dicho estudio consistía hasta el momento de su publicación en la secuenciación de RNA a mayor escala de este tipo de condición y se realizó para 200 tumores de pacientes de Corea del Sur, realizando una caracterización más profunda para 87 resecciones quirúrgicas en las que se realizó secuenciación del transcriptoma tanto del tejido tumoral como del tejido

normal adyacente (este último solo para 77). Estos 87 tejidos tumorales fueron seleccionados por no haber sido detectado mediante técnicas de *screening* previo. La finalidad de dicho estudio no está relacionada con circRNAs, sino con mutaciones en genes que se puedan relacionar con la condición mostrada por los pacientes, por lo que este proyecto proporciona un enfoque complementario a dicho artículo.

La información se encuentra en formato Sequence Read Archive (extensión ".sra"), que incluye datos de secuenciación. Las herramientas de alineamiento e identificación de circRNAs requieren su transformación a formato FASTQ.

Se han utilizado en este proyecto 40 ficheros correspondientes a 20 muestras de tejido tumoral y otras 20 muestras de tejido normal.

Por tanto, los primeros pasos del proceso del *pipeline* consisten, como se verá en las próximas secciones, en la descarga de cada uno de los ficheros que definen las muestras y su conversión al formato específico antes de su procesado.

2.3 Descripción general del *pipeline* de identificación de circRNAs

El *pipeline* de identificación de circRNAs está compuesto por dos partes. Una primera fase (Figura 3) realiza la descarga de los ficheros, su conversión y el procesado por las diferentes herramientas, finalizando con la integración de los resultados parciales en un único resultado consolidado. Una vez se dispone de la información integrada de los circRNAs para cada muestra, se realiza un procesado multimuestra (Figura 4) en el que se genera un único fichero con la información de expresión de todas las muestras, tabulado para su posterior análisis.

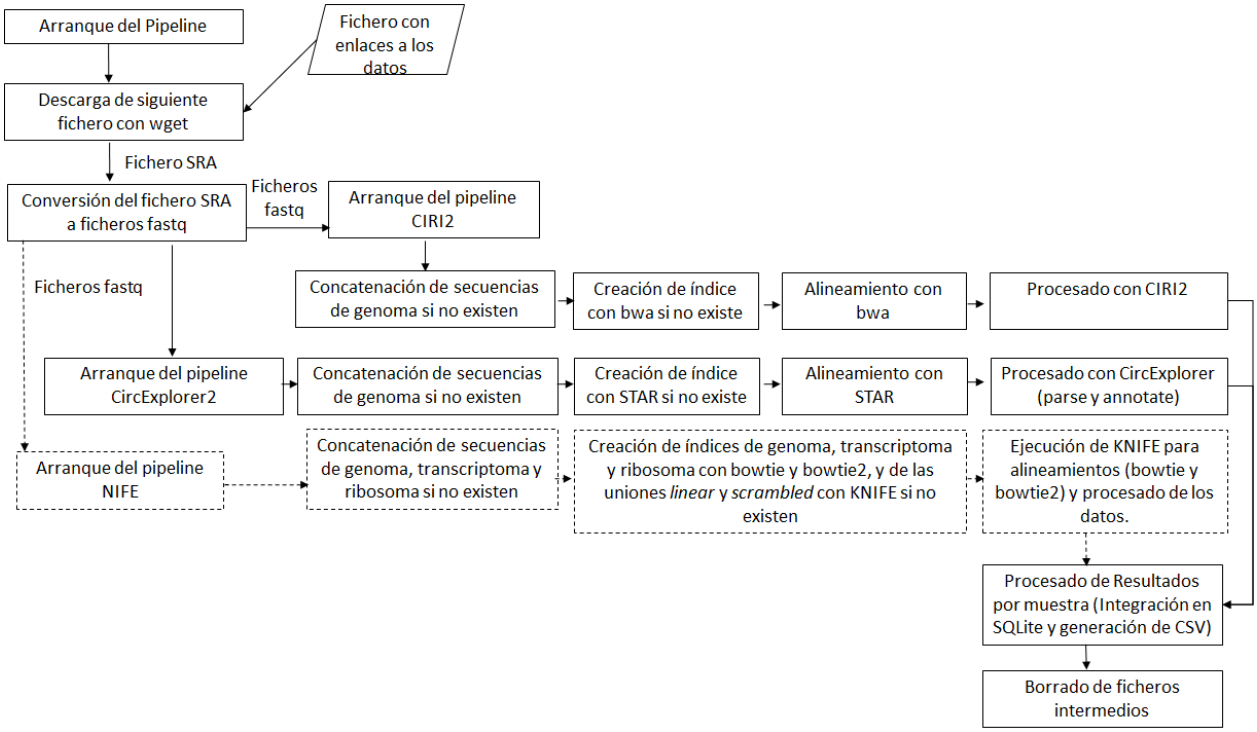


Figura 3. Parte del pipeline correspondiente a la ejecución por muestra.

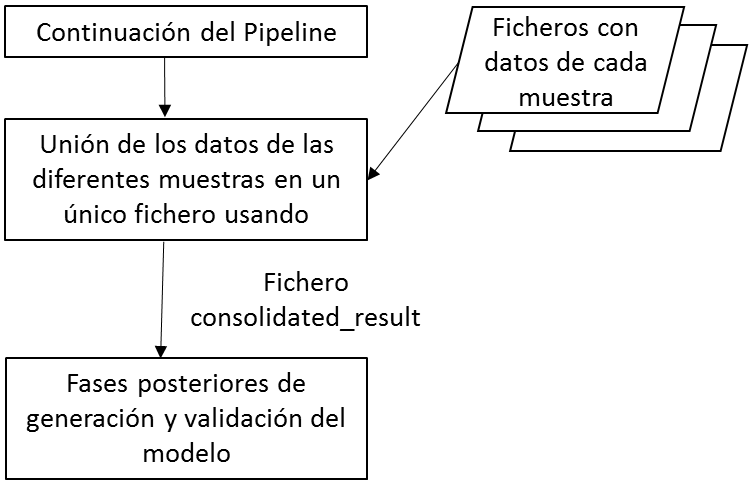


Figura 4. Parte del pipeline correspondiente a la integración de muestras.

Como configuración de entrada al proceso, se genera un fichero llamado *index_files.lnk* que incluye los datos de las URL de donde descargar los ficheros de secuencias de las muestras. A continuación se introducen unas líneas de ejemplo de dicho fichero:

```
#LC_S1(+nor)
ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-
instant/reads/ByStudy/sra/ERP/ERP001/ERP001058/ERR062334/ERR062334.sra
ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-
instant/reads/ByStudy/sra/ERP/ERP001/ERP001058/ERR164500/ERR164500.sra
#LC_S2(+nor)
ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-
instant/reads/ByStudy/sra/ERP/ERP001/ERP001058/ERR062335/ERR062335.sra
ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-
instant/reads/ByStudy/sra/ERP/ERP001/ERP001058/ERR164501/ERR164501.sra
```

El ejecutable principal se denomina *run_pipelines* e implementa un bucle encargado de leer cada uno de los enlaces (URLs) a los ficheros, verificar si ya se encuentra procesado y, en caso contrario, descargarlo, convertirlo y procesarlo. Para cada uno de esos pasos delega la tarea en otros *scripts*. Se usa la aplicación *wget* para las descargas y *fastq-dump,* parte del SRA Toolkit [18] del National Center for Biotechnology Information (NCBI), para la conversión de formato de los ficheros. Los ficheros convertidos se alimentan a *pipelines* específicos de cada aplicación de alineamiento e identificación de circRNAs. En las siguientes secciones se verán los pasos particulares de cada uno de estos diferentes caminos de identificación de circRNAs.

2.4 Descripción del *pipeline* basado en CIRI2.

CIRI2 [13] (CircRNA Identifier 2) es una herramienta de identificación de circRNAs de novo que utiliza un algoritmo basado en la identificación de señales junto con un filtrado sistemático posterior para eliminar falsos positivos.

Su funcionamiento se basa en la existencia de una secuencia genómica de referencia y un fichero de alineamiento local de los segmentos de la muestra con la secuencia de referencia utilizando el algoritmo BWA-MEM. A partir de estas entradas, CIRI2 es capaz de hacer una primera identificación de circRNAs candidatos mediante la detección de señales basadas en segmentos alineados de forma específica. Posteriormente, excluye los datos que no se ven validados por la información disponible de *paired-end* y por señales de *splicing* GT-AG.

El algoritmo BWA-MEM es uno de los implementados en la herramienta de alineamiento Burrows-Wheeler Aligner (BWA [20]) y se basa en la transformación de Burrows-Wheeler. La herramienta BWA requiere primero realizar una indexación del genoma de referencia para poder después completar el alineamiento de las muestras.

El *pipeline* comienza por la verificación de la existencia de la secuencia genómica de referencia. Si esta no está disponible, se genera concatenando de forma automática los datos del genoma que han de estar disponibles en una ubicación accesible desde el servidor y configurable en el programa.

Tras esto, como se ha comentado anteriormente, se utiliza la herramienta BWA para generar el índice del genoma de referencia, solo si este no ha sido ya creado en iteraciones previas sobre otras muestras.

Se ejecuta después el algoritmo BWA-MEM para alinear las lecturas obtenidas de las muestras con el genoma indexado.

Finalmente, el resultado es alimentado a CIRI2, que se encarga de identificar los circRNAs y volcar la información a un fichero.

2.5 Descripción del *pipeline* basado en CIRCExplorer2.

CIRCExplorer2 [14] es una herramienta de análisis de circRNAs con capacidad de anotación de los mismos. Las últimas versiones disponen también de un procesado alternativo con la posibilidad de identificar nuevos circRNAs realizando un ensamblado *de novo*. En este estudio se ha utilizado solamente el *pipeline* de anotación puesto que el análisis *de novo* requiere un mayor número de programas y extendería en exceso el alcance del proyecto.

El software soporta múltiples herramientas de alineamiento y una ejecución integrada con un solo comando. No obstante, para el uso de lecturas *paired-end*, el alineamiento ha de realizarse de forma manual y solo se pueden emplear STAR [21] (Spliced Transcripts Alignment to a Reference) y TopHat-Fusion[22] como herramientas para este fin.

En el presente proyecto se ha decidido utilizar STAR principalmente por una cuestión de simplicidad.

De forma similar a CIRI2, los primeros pasos consisten en generar la secuencia de genoma de referencia y el índice de dicho genoma utilizando la herramienta

de alineamiento (STAR), para después completar el procesado de las lecturas de las muestras en relación a este índice. Una vez se dispone del resultado de alineamiento, éste es utilizado por CIRCExplorer2 realizando la lectura (*parsing*) del mismo y generando un fichero en formato “BED”. Esta información, junto con el fichero de anotación del genoma, se utiliza para identificar los límites de los circRNAs y producir un fichero resultado.

2.6 Descripción del *pipeline* basado en KNIFE.

KNIFE [15] (Known and Novel IsoForm Explorer) es otra herramienta software para la identificación de circRNAs. Se basa en un algoritmo estadístico que cuantifica eventos de *splicing* lineales y circulares. Proporciona una versión ejecutable en una sola máquina y otra para poder hacer uso de un *cluster*.

KNIFE tiene unos requisitos en cuanto a memoria y procesador muy superiores a las otras herramientas utilizadas en el proyecto. En cuanto a software, necesita tener disponible *Bowtie* [23] y *Bowtie2* [24] (herramienta de alineamiento basada en la transformación de Burrows-Wheeler), *python* con bibliotecas científicas y R con el paquete *data.table*.

Como en los casos anteriores, el primer paso consiste en construir la secuencia de genoma de referencia si esta no existe.

A continuación será necesario descargar o generar un conjunto de índices de alineamiento. Existen índices pregenerados para algunos de los genomas de referencia, que son proporcionados por los creadores de la herramienta. En el caso humano, existe para la versión hg19. En el proyecto se ha utilizado la versión hg38, con lo que se optó por generar los índices de alineamiento para el genoma, el transcriptoma y el ribosoma de dicha versión del genoma humano. Además, se generaron los índices de las uniones lineales y desordenadas (*scrambled*) que solicita la herramienta.

Tras esto, KNIFE hace uso de Bowtie y Bowtie2 para realizar el alineamiento de las muestras y la posterior identificación de circRNAs basado en anotaciones y también mediante un análisis *de novo*. Para ello, en las muestras de tipo *paired-end* realiza la identificación de los eventos de splicing con la lectura en uno de los sentidos y valida con la lectura que comienza en el extremo opuesto y sentido contrario. En una segunda pasada, realiza el proceso en sentido inverso, para finalmente integrar el resultado en un fichero.

Como se comentó en el capítulo primero, en el proyecto se han encontrado varios inconvenientes a la hora de utilizar KNIFE. El primero está relacionado con los ya comentados amplios requisitos de memoria y proceso. Las limitaciones en cuanto a recursos físicos disponibles (máquina virtual 32 Gigabytes de RAM y 4 Terabytes de disco) han producido errores en la ejecución del programa, haciendo que el sistema operativo detuviera alguno de los procesos que se ejecutan en paralelo y produciendo, por tanto, resultados no válidos. Aunque se procedió a modificar los *scripts* del programa para secuenciar algunos de estos procesos, se ha observado que el comportamiento de la herramienta sigue sin ser estable en el entorno utilizado. Así, para

algunas de las muestras, se ha podido completar la identificación de circRNAs con la ejecución comentada en ambos sentidos, mientras que en otras muestras solo ha finalizado con éxito uno de los dos.

Por tanto, se ha decidido excluir KNIFE del análisis final, aunque se utilizará como referencia para observar la dificultad de añadir nuevas herramientas en este pipeline.

2.7 Consolidación de los resultados de los diferentes pipelines.

La información de circRNAs identificados por las distintas herramientas para una muestra concreta ha sido grabada en etapas previas en archivos específicos por herramienta. Para poder consolidar la información, se utiliza el software de gestión de base de datos SQLite [25] (creando la base de datos circRNA.db), junto con un programa en Python encargado de leer los ficheros y rellenar las tablas correspondientes.

Esos mismos programas se encargan de realizar consultas SQL para recuperar la información obtenida por las aplicaciones, consolidarla y validar si un mismo circRNA ha sido encontrado por diferentes herramientas. Para esto último, se establece una distancia máxima de 100 pares de bases al inicio y fin de secuencias identificadas por 2 programas para considerarlos como el mismo circRNA.

Como resultado, se añade la información a una tabla de la base de datos (circRNA_joined) y se genera un fichero CSV con los datos.

2.8 Integración de la información de las diferentes muestras.

La finalidad de este paso del pipeline consiste en integrar la información de las diferentes muestras y construir una tabla que pueda ser alimentada al modelo de árbol de decisión. El *script join_samples* se encarga de esta tarea.

Para ello, se selecciona la información de las muestras incluida en la tabla *circRNA_joined* y se comprueba si hay coincidencia en cuanto a los valores de inicio y fin de secuencia entre diferentes muestras. De nuevo se acepta una variación de hasta 100 pares de bases en ambos extremos. Los datos se integran en la tabla *circRNA_vals*.

Como resultado final del proceso, se genera un fichero en forma de tabla llamado *consolidated_result*. Las filas corresponden a cada muestra y las columnas a las lecturas (*reads*) sin normalizar para cada circRNA (identificado con su cromosoma, hebra y valores de inicio y fin de secuencia).

2.9 Resumen de resultados del *pipeline* para CIRI2 y CIRCExplorer2.

A continuación se describe la información principal resultado del *pipeline* de identificación de circRNAs.

De los resultados disponibles hasta el momento, hay un total de 7577 circRNAs detectados. De ellos, 118 se han detectado de forma simultánea por ambos programas en al menos 1 muestra. Curiosamente, un mismo circRNA ha podido ser detectado por ambos programas en una muestra concreta y solo por uno de ellos en otra (por ejemplo con un circRNA correspondiente al cromosoma 8 que se usará en el modelo).

Tabla 3. Listado de circRNAs detectados por CIRI2 y CIRCExplorer 2

Gene ID	Cromosoma	Hebra	Inicio	Fin
ABI3BP	chr3	-	100804791	100808235
ABI3BP	chr3	-	100823457	100824941
AHNAK	chr11	-	62516794	62517009
ANKRD36B	chr2	-	97549418	97549614
ANKRD36C	chr2	-	95884339	95886120
ANOS1	chrX	-	8585266	8587978
ANP32B	chr9	+	98011270	98012472
APP	chr21	-	26111978	26112146
ARID4B	chr1	-	235194012	235214026
ASPH	chr8	-	61680967	61684188
ATRX	chrX	-	77652113	77656653
AUTS2	chr7	+	70762869	70771644
B2M	chr15	+	44715422	44715701
B2M	chr15	+	44715545	44716356
CALD1	chr7	+	134933193	134933390
CCAR1	chr10	+	68761006	68773099
CCDC124	chr19	+	17933015	17936579
CD44	chr11	+	35186831	35190065
CD55	chr1	+	207331107	207331296
CDR1-AS	chrX	+	140783175	140784660
CDYL	chr6	+	4891712	4892379
CHD2	chr15	+	92996956	93002317
CHSY1	chr15	-	101235081	101235577
CLTC	chr17	+	59683386	59683536
COL1A1	chr17	-	50189167	50189538
COL1A1	chr17	-	50189858	50190108
COL1A1	chr17	-	50194129	50194447
COL1A1	chr17	-	50196313	50196670
COL1A2	chr7	+	94416404	94417831
COL1A2	chr7	+	94417723	94418552
COL1A2	chr7	+	94418498	94420286
COL3A1	chr2	+	189005349	189006259
COL6A2	chr21	+	46121060	46121618
COL9A3	chr20	+	62829454	62830413
CRIM1	chr2	+	36396613	36396787
CUX1	chr7	+	102070338	102104459
DCN	chr12	-	91157074	91164717
DEK	chr6	-	18236451	18258405
DMBT1	chr10	+	122579577	122598012
DNAJC2	chr7	-	103321931	103322794
DUSP1	chr5	-	172769156	172769574
DUSP1	chr5	-	172769574	172770306
EDEM1	chr3	+	5195208	5208263

EFNB2	chr13	-	106512528	106512812
EIF4G3	chr1	-	20849414	20904971
EIF5B	chr2	+	99360235	99371730
ESYT2	chr7	-	158759485	158764853
EXOC7	chr17	-	76101270	76103677
FN1	chr2	-	215373321	215391814
FN1	chr2	-	215384019	215384184
FOSB	chr19	+	45472550	45472706
FTO	chr16	+	53888831	53934109
GALK2	chr15	+	49235850	49239367
GALK2	chr15	+	49239220	49239367
GIGYF2	chr2	+	232819826	232819985
GPC6	chr13	+	93545262	93545421
H19	chr11	-	1997489	1997696
KIAA2026	chr9	-	5968018	5988545
LIMCH1	chr4	+	41494535	41606004
LINC01322	chr3	+	165491195	165511651
LINC01578	chr15	+	92885514	92892072
MAN1A2	chr1	+	117402185	117405645
MAN1A2	chr1	+	117402185	117420649
MAN1A2	chr1	+	117402185	117442325
MAP3K3	chr17	+	63692241	63692419
MAST3	chr19	+	18147442	18147624
METRNL	chr17	+	83084937	83085323
MTUS1	chr8	-	17675185	17684542
MUC16	chr19	-	8896810	8902222
MYRIP	chr3	+	40044049	40044271
NCOA4	chr10	-	46014984	46015125
NCOR1	chr17	-	16108785	16146548
NEDD8-MDP1	chr14	-	24217185	24218215
NEMF	chr14	-	49789145	49795944
PALM2-AKAP2	chr9	+	110136126	110138539
PARD6G	chr18	-	80202709	80202932
PDLIM5	chr4	+	94618003	94640450
PEA15	chr1	+	160213109	160213265
PHF14	chr7	+	10982371	10990847
PI4KA	chr22	-	20729312	20765693
PICALM	chr11	-	85974707	86003451
PID1	chr2	-	229155817	229155964
PLEKHH2	chr2	+	43742740	43745963
PLEKHO2	chr15	+	64848592	64848742
PRRC2C	chr1	+	171523220	171532961
RAB11FIP4	chr17	+	31525089	31525230
RABEPK	chr9	+	125200779	125203066
RBPMS	chr8	+	30474778	30504436
RILPL1	chr12	-	123498543	123499536
RIN3	chr14	+	92555750	92577477
RNF168	chr3	-	196487398	196488683
RTF1	chr15	+	41457671	41470392
RTKN2	chr10	-	62198282	62199861
SEC62	chr3	+	169975607	169988359
SEC62	chr3	+	169976945	169985865
SEC62	chr3	+	169976945	169988359

SFTPB	chr2	-	85665605	85665794
SFTPC	chr8	+	22161796	22162732
SFTPC	chr8	+	22163900	22164059
SLC39A8	chr4	-	102304316	102304481
SLC39A8	chr4	-	102304316	102315830
SMAD6	chr15	+	66711667	66716498
SMC4	chr3	+	160413472	160423840
SPARCL1	chr4	-	87482423	87491690
SPTBN1	chr2	+	54526371	54526566
STXBP6	chr14	-	24974664	24974850
TFPI	chr2	-	187484123	187503770
THRAP3	chr1	+	36259381	36282700
THYN1	chr11	-	134249827	134250343
TIPARP	chr3	+	156694019	156696025
TRPC4AP	chr20	-	35035122	35035308
TXNIP	chr1	-	145994928	145995476
UBXN4	chr2	+	135769768	135770735
UBXN4	chr2	+	135769768	135772547
WDR60	chr7	+	158869854	158876691
WFDC1	chr16	+	84324418	84326955
ZC3H6	chr2	+	112299848	112300029
ZNF384	chr12	-	6673215	6673395

De los 7577 circRNAs, 6217 solo aparecen en una muestra, 749 en dos muestras, 249 en 3 muestras y 112 en 4 muestras. El resto se muestra en la siguiente figura. En particular se observa que hay 2 circRNAs que han sido detectados en 37 de las 40 muestras.

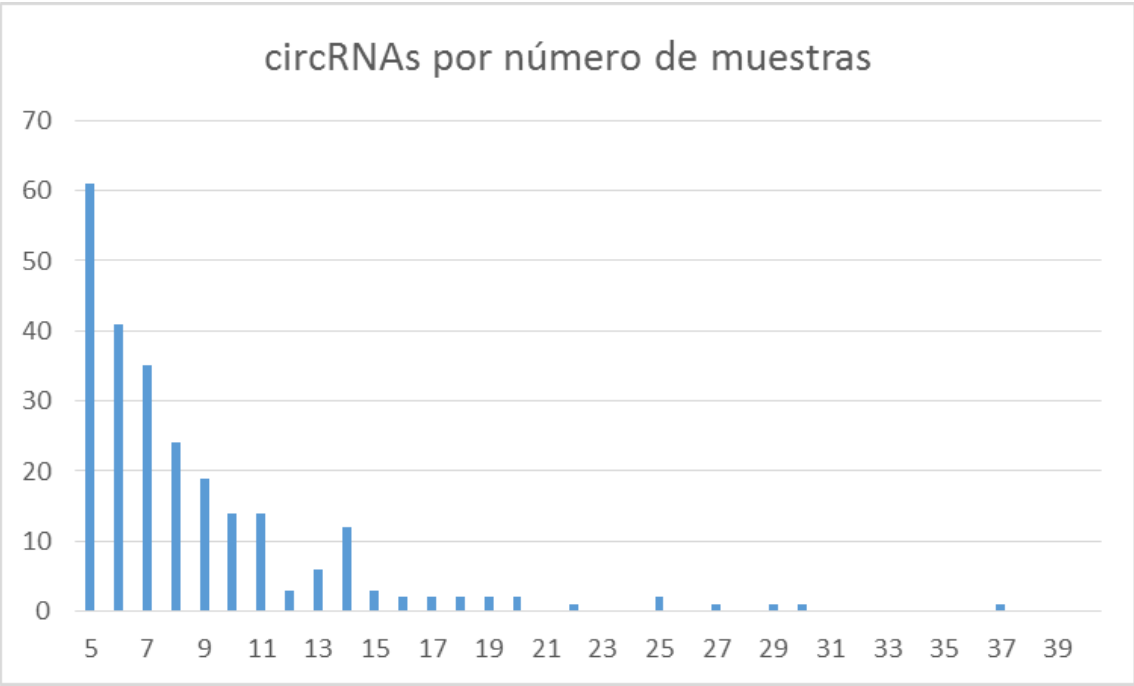


Figura 5. Número de circRNAs en función del número de muestras en las que se detectaron.

2.10 Incorporación de *pipeline* adicional

Como se ha comentado previamente, la herramienta KNIFE ha sido excluida del análisis anterior. Sin embargo, se ha realizado un procesado de 18 muestras (9 sanas y 9 tumorales) para las que se obtuvo información de KNIFE con el fin de validar la dificultad en introducir una herramienta adicional de identificación de circRNAs en el proceso.

Los elementos que se necesitan variar son los siguientes:

- Incorporar la línea correspondiente al nuevo pipeline en el fichero principal de ejecución.
- Generar el script del nuevo pipeline (en este caso ya estaba creado).
- Modificar las consultas de base de datos para integrar los resultados adicionales y consolidarlos con los demás para cada muestra.

El resto del proceso se mantiene igual.

Para estas 18 muestras se han identificado por los 3 *pipelines* un total de 8984 circRNAs. Las tres herramientas han coincidido en identificar 39 circRNAs en al menos una muestra. Destacan también 332 circRNAs detectados simultáneamente por CIRCExplorer2 y KNIFE, pero no por CIRI2 y 51 circRNAs coincidentes entre CIRI2 y KNIFE, pero no disponibles en CIRCExplorer2.

3. Modelo de árbol de decisión

3.1 Árboles de decisión e *information gain*

Dentro de los modelos de clasificación (y regresión) existentes en el campo de *Machine Learning*, los árboles de decisión son un método de aprendizaje supervisado basado en sentencias de control condicional que dividen las ramas del árbol y facilitan llegar a una conclusión (correspondiente a las hojas) respecto a un objeto especifico que se desea estudiar (una muestra en nuestro caso). Estos algoritmos equilibran simplicidad y eficiencia, lo que los hace apropiados para su uso en un proyecto como el aquí realizado, además de estar suficientemente probados en el campo de la genómica computacional

Para construir este modelo, es necesario entrenarlo proporcionándole un conjunto de datos que le permitan inferir las sentencias de control a crear y en qué orden dentro de la estructura del árbol. Dentro de las métricas posibles que facilitan al algoritmo identificar las condiciones de separación más útiles está *Information Gain*. La ganancia de información (*Information Gain)* selecciona primero aquellas características que reducen en mayor medida la entropía en la clasificación respecto al nivel anterior y, por tanto, mejoran la distribución de las muestras de entrenamiento en cada división (grupos divididos más homogéneos).

Crear software para la generación de este tipo de modelos está fuera de los objetivos de este trabajo. Existen numerosos paquetes de software perfectamente verificados, capaces de construir un modelo de árbol de decisión basado en la métrica de la ganancia de información. Entre ellos, dentro del software estadístico R existe el paquete *rpart* que se utilizará en este proyecto. Un *script* de R se encargará de ejecutar todos los pasos necesarios para construir y validar el modelo, así como para imprimir los resultados del mismo.

3.2 Normalización y preprocesado de la información

El fichero obtenido como resultado del *pipeline* de identificación de circRNAs descrito en el capítulo anterior contiene una tabla en el que las filas corresponden a las 40 muestras procesadas y las columnas a los circRNAs detectados. Por tanto la tabla incluye el número de *reads* de cada circRNA en cada muestra.

Al realizar el proceso experimental de secuenciación de cada muestra, el número de *spots* y el número de *reads* obtenidos (2 por *spot*) difieren entre muestras. Esto hace que sea posible tener una diferencia de expresión en cuanto a circRNAs derivada simplemente de un mayor o menor número de *reads* totales y no dependiente de la condición evaluada de cada muestra.

Por tanto, es necesario aplicar un proceso de normalización de las muestras, para lo que se ha decidido utilizar la métrica Reads Per Million Mapped Reads (RPM). Se calcula para cada circRNA utilizando la fórmula siguiente:

$$RPM = (reads\_{circRNA} * 10^6) / reads\_{sample}$$

donde $reads_{circRNA}$ es el número de *reads* para un circRNA concreto, y $reads_{sample}$ es el total de *reads* para la muestra.

Para incorporar a la tabla de circRNAs la información específica de cada muestra, que incluye el número de *spots* y la condición normal/tumoral de las mismas, se proporciona un fichero con formato de valores separados por coma denominado MapSampleReadsCondition. A continuación se muestran algunas entradas del mismo.

```
sample_id,spots,condition
ERR164473,22844380,0
ERR164474,42022150,0
ERR164475,36558196,0
ERR164476,32065580,0
ERR164477,35823521,0
ERR164478,42992272,0
```

Este fichero se carga en R y se realiza una fusión del mismo con la tabla de circRNAs. Tras esto se produce la normalización comentada sabiendo que existen 2 *reads* por cada *spot.*

3.3 Conjuntos de entrenamiento y prueba

Para realizar la división de las 40 muestras en conjuntos de entrenamiento y prueba se ha decidido, en primer lugar, asignar el 66% de las muestras para entrenar el modelo y el resto para la verificación posterior. Con el fin de que haya un mismo número de muestras de cada condición en ambos conjuntos, se ha empleado la función *createDataPartition* del paquete *caret* de R, estableciendo previamente una semilla (6333222) para hacer el proceso reproducible.

Las 28 muestras asignadas por este proceso al conjunto de entrenamiento se dividen a partes iguales en 14 muestras tumorales y 14 de tejido normal.

Las restantes 12 muestras (6 tumorales y 6 normales) se utilizan como conjunto de test para validar el modelo.

3.4 Creación y validación del árbol de decisión.

Como se ha indicado previamente, el árbol de decisión se ha generado utilizando la función *rpart* del paquete del mismo nombre de R. Se establece en ella como métrica de partición “information” (correspondiente a ganancia de información).

Una representación del modelo del árbol se muestra en la figura 6. En ella se ve que se centra en 3 circRNAs de los obtenidos para separar las 14 muestras tumorales de las de tejido normal. Teniendo en cuenta que se ha considerado un margen de error en cuanto al número de reads de +/- 100 pares de bases, se obtienen los siguientes elementos: chr5:151026992-151927531 (+), chr8:22161796-22162648 (+) y chr1:44810607-44810720 (+).

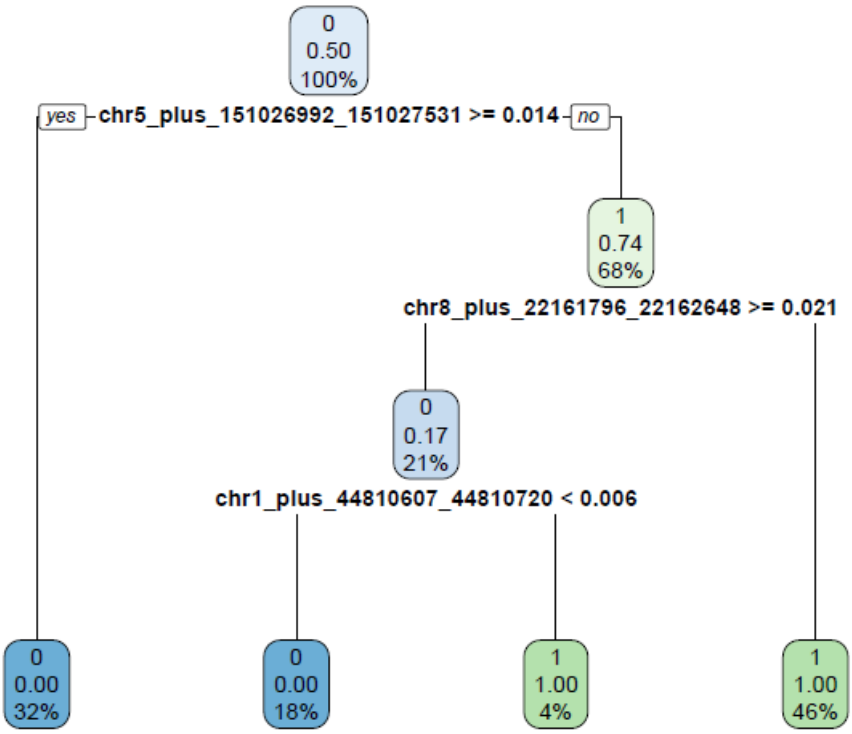


Figura 6. Árbol de decisión y condiciones de separación.

La verificación del modelo se realiza con el conjunto de test. Aplicado este árbol de decisión a dicho conjunto se obtiene que clasifica correctamente 11 de 12 muestras (hay 1 muestra tumoral que clasifica como normal) como se observa en la tabla siguiente. En concreto, la muestra 29 (ERR164554) es la que se clasifica erróneamente.

	pred_class	condition
2	1	1
6	0	0
11	0	0
15	0	0
17	0	0
20	0	0
22	0	0
26	1	1
29	0	1
31	1	1
38	1	1
40	1	1

Esa misma información más detallada, generada por la función *CrossTable* del paquete *gmodel* se observa en la siguiente tabla.

Tabla 4. Predicción (pred_class) obtenida vs. condición real (condition)

	test[, "pred_class"]		
test[, "condition"]	0	1	Row Total
0	6	0	6
	1.000	0.000	0.500
	0.857	0.000	
	0.500	0.000	
1	1	5	6
	0.167	0.833	0.500
	0.143	1.000	
	0.083	0.417	
Column Total	7	5	12
	0.583	0.417	

En estas condiciones, la sensibilidad correspondería a un 83,33% (5/6) y la especificidad sería de un 100%. Es necesario considerar que se ha utilizado un número limitado de muestras (40), por lo que sería necesario incorporar muestras adicionales para tener una mayor seguridad en el modelo. No obstante, sí que muestra claramente que el *pipeline* construido funciona.

3.5 Resultados del modelo

Observando en detalle el modelo creado y la importancia de las variables consideradas en el mismo, se observa la información presentada en la figura 7.

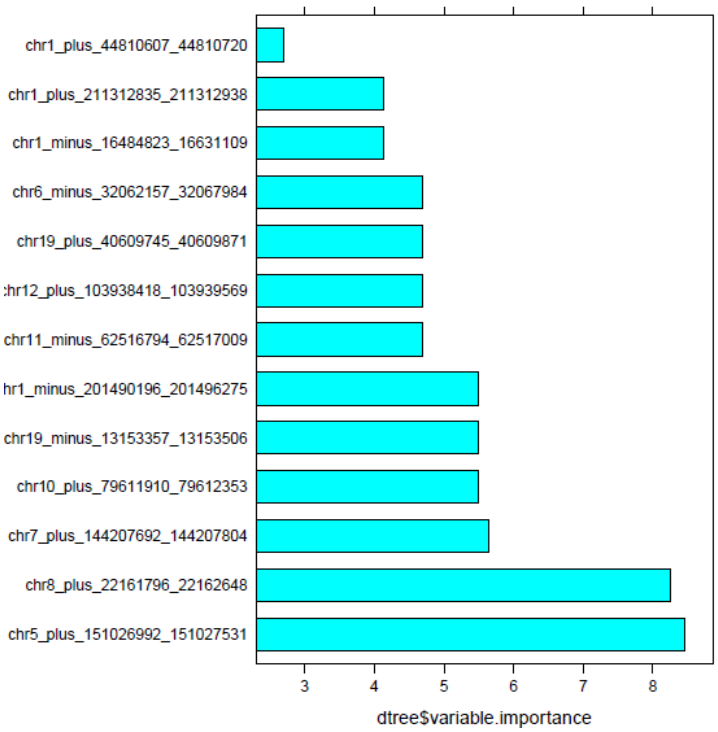


Figura 7. Importancia de las variables en el modelo del árbol de decisión.

En ella vemos que se han seleccionado las 2 variables con más importancia en las primeras divisiones y la menos representativa en el nivel más bajo del árbol. Al observar la diferencia de expresión de las variables correspondientes a los circRNAs en cuanto a aparición en muestras normales y tumorales, se visualiza el gráfico de la figura 8.

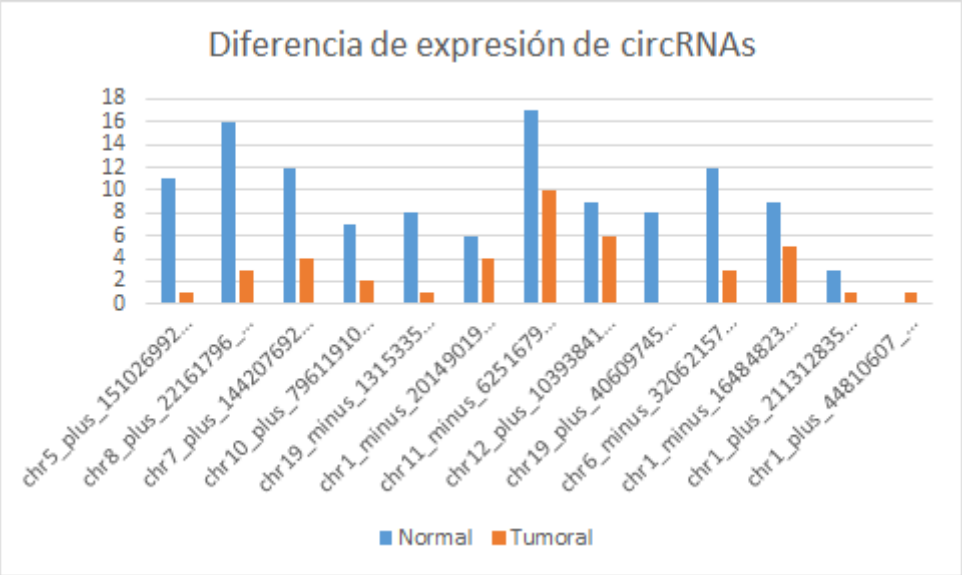


Figura 8. Diferencia de expresión de los circRNAs en muestras sanas y tumorales.

Además de los 2 circRNAs con mayor importancia, usados en la clasificación, se ven también otros posible circRNAs que difieren en expresión entre ambos tipos de muestras (por ejemplo, los correspondientes al cromosoma 19 o al cromosoma 7).

Se observa también que el circRNA chr1:44810607-44810720 (+) solo ha afectado a la construcción del modelo y no ayuda (en este caso) a la clasificación posterior del conjunto de test pues corresponde a una única muestra de entre todas las de entrenamiento y test.

4. Discusión

4.1 Bases de datos de circRNAs.

Para validar la información relativa a los circRNAs obtenidos en el modelo de clasificación del capítulo anterior, se han revisado las bases de datos de circRNAs con el fin de contrastar si se han identificado previamente en otros estudios. En algunos casos ha sido necesario emplear la utilidad liftover [26] al tener algunas de esas bases de datos solo identificada la información para el ensamblado hg19, diferente del usado. En caso de no encontrar el circRNA, se ha validado si existe relación con algún exón de alguna proteína o RNA no codificante existente.

Las bases de datos utilizadas son las siguientes:

- circBase [27] (http://www.circbase.org/). Agrupa información de circRNAs identificados en diferentes estudios. Es posible realizar consultas online o descargar un Excel con todos los datos.
- circRNADb [28] (http://202.195.183.4:8000/circrnadb/circRNADb.php). Otra base de datos de este tipo de secuencias, disponible de forma gratuita para uso no comercial.
- CSCD [29] - Cancer Specific CircRNA DB (http://gb.whu.edu.cn/CSCD/). En este caso, presenta también información específica de circRNAs que han sido asociados en algún momento con expresión de células tumorales.

Se han consultado también las siguientes páginas web:

- Protein Data Bank [30] del Research Collaboratory for Structural Bioinformatics, con el fin de identificar el encaje o no de los circRNAs con exones de diferentes proteínas.
- The Human Protein Atlas [31] para verificar si las proteínas a las que corresponden los circRNAs se corresponden con biomarcadores pulmonares.
- UCSC (University of California, Santa Cruz) Genome Browser [32] para visualizar los circRNA situados en un gen frente a los exones/intrones del mismo.

4.2 Información sobre los circRNAs identificados.

Se han verificado los circRNAs detectados en el capítulo anterior, tanto los usados en la clasificación del modelo de árbol de decisión, como el resto de variables que se analizaron en el mismo.

Los resultados de este análisis se muestran a continuación:

- chr5:151026992-151027531 (+) (chr5:150406553-150407092 en hg19). Se superpone sobre un circRNA existente en la base de datos de circBase para el gen GPX3. Además, se corresponde con 1 o 2 exones y 1 intrón de GPX3. Se encuentra dentro de un circRNA (chr5:151025375|151028130) de GPX3 que ha sido identificado en CSCD como relacionado con cáncer. La

expresión de este gen se encuentra infrarregulada en varios tumores, incluido el de pulmón, debido a hipermetilación del promotor [33].

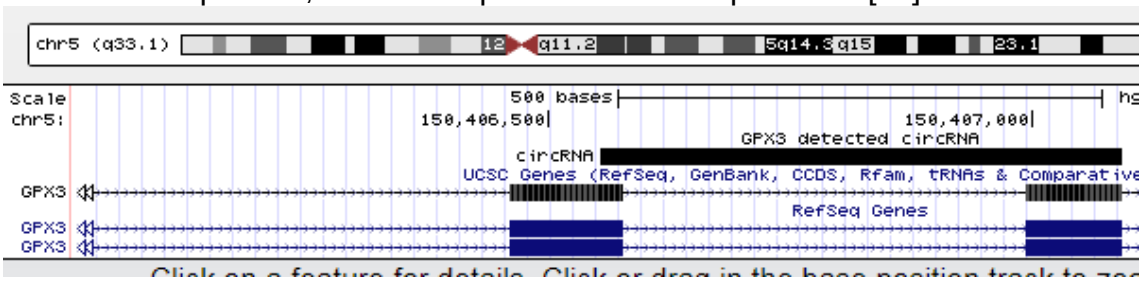


Figura 9. CircRNA chr5:151026992-151027531 vs. exones de GPX3.

- chr8:22161796-22162648 (+) (chr8:22019309-22020161 en hg19). Se encuentra dentro de un circRNA identificado en circRNAdb (chr8:22019183-22020245). Corresponde aproximadamente a 2 exones de SFTPC (considerando un margen de error en la detección de los circRNAs). Es importante destacar que las células que expresan SFTPC suelen estar en el origen de los adenocarcinomas pulmonares [34].

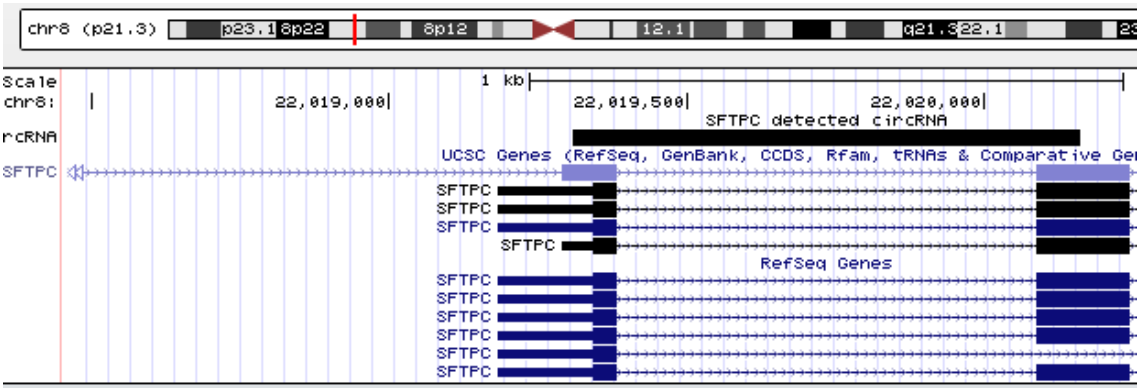


Figura 10. CircRNA chr8:22161796-22162648 vs. exones de SFTPC.

- chr7:144207692-144207804 (+) (chr7:143904785-143904897 en hg19). Clasificado por CircExplorer como parte de RP4-545C24.1. No existe un circRNA identificado justo en esta posición según las bases de datos consultadas. Existe un circRNA que corresponde a RP4-545C24.1 (chr7:144209437|144210726) relacionado con cáncer según CSCD, aunque bastante alejado del inicio (casi 2000 pares de bases).

- chr10:79611910-79612353 (+) (chr10:81371666-81372109). No se encuentran circRNAs relacionados en las bases de datos consultadas. Puede corresponder con 2 exones de SFTPA1 o con 1 intrón en función del empalmamiento alternativo. Esta proteína, aunque no usada como biomarcador, aparece sobreexpresada en algunos tipos de cáncer de pulmón al utilizar algunos de los anticuerpos específicos, según The Human Protein Atlas.

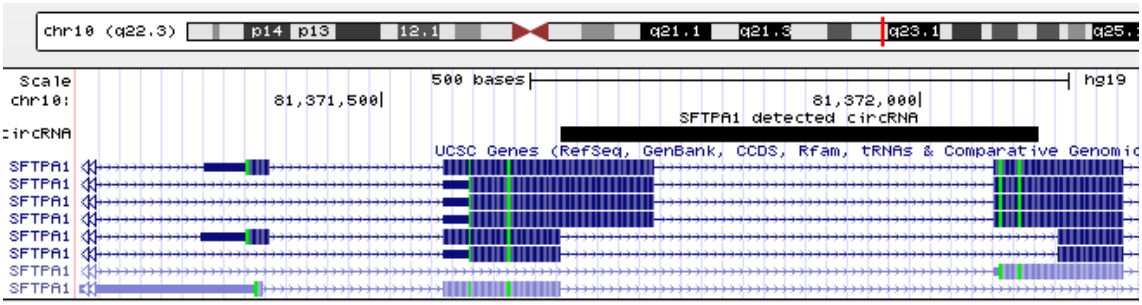


Figura 11. CircRNA chr10:79611910-79612353 vs. exones de SFTPA1.

- chr19:13153357-13153506 (-) (chr19:13264171-13264320 en hg19). No se encuentran circRNAs que coincidan con el encontrado, sin embargo, en CSCD existe el circRNA chr19:13264051|13264179 cercano al identificado. Por otro lado, la secuencia está situada dentro del gen CTC-250I14.6 (ENSG00000267598), en el intrón situado entre los dos exones incluidos en la misma.

- chr1:201490196-201496275 (-) (chr1:201459324-201465403 en hg19). Corresponde a 2 exones del gen CSRP1 (201459304-201459472 y 201465320-201465431). Existe un circRNA en la base de datos CSCD (chr1:201459327|201465386) que se corresponde en gran medida con el circRNA encontrado. CSRP1 ha sido identificado como biomarcador para cáncer renal y de cabeza y cuello según The Human Protein Atlas.

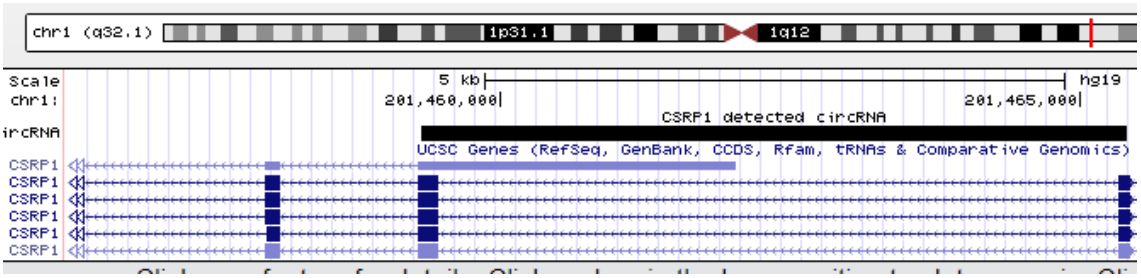


Figura 12. CircRNA chr1:201490196-201496275 vs. exones de CSRP1.

- chr11:62516794-62517009 (-) (chr11:62284266-62284481 en hg19). Existe un circRNA en CSCD (chr11:62284210|62284417) localizado en el gen AHNAK, que ha sido identificado como biomarcador en otros tipos de cáncer (pancreático y urotelial según The Human Protein Atlas).

- chr12:103938418-103939569 (+) (chr12:104332196-104333347 en hg19). Se corresponde a parte del gen HSP90B1. Existen dos circRNAs que se solapan parcialmente con este en CSCD, pero no coinciden completamente. Esta proteína se utiliza como biomarcador de cáncer renal según The Human Protein Atlas, con una media/alta expresión en cáncer de pulmón.

- chr19:40609745-40609871 (+) (chr19:41115651-41115777 en hg19). Se corresponde exactamente a un exon de la proteína LTBP4. No se encuentran circRNAs relacionados en las diferentes bases de datos. El gen es biomarcador de cáncer de ovario según The Human Protein Atlas.

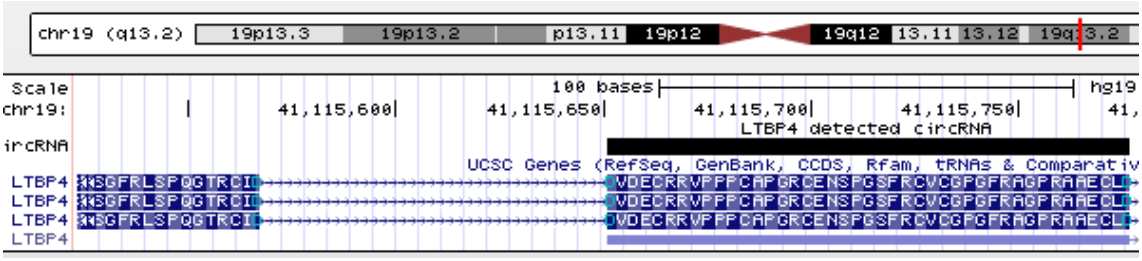


Figura 13. CircRNA chr19:40609745-40609871 vs. exones de LTBP4.

- chr6:32062157-32067984 (-) (chr6:32029934-32035761 en hg19). Parece un circRNA poco probable, pues no se corresponde a ningún circRNA y su larga longitud hace que se solape parcialmente con el gen TNXB, pero que la mayor parte del mismo corresponda a regiones entre genes. Además, ha sido reconocido solo por CIRI2 en el análisis *de novo*.

- chr1:16484823-16631109 (-) (chr1:16811318-16957604 en hg19). La longitud de este elemento, cubriendo varios genes, hace probable que sea un falso circRNA generado por el análisis *de novo* realizado por CIRI2 aunque al verificar con los datos reducidos obtenidos con KNIFE, se ve que también es identificado por esta herramienta.

- chr1:211312835-211312938 (+) (chr1:211486177-211486280 en hg19). Ubicado dentro de un exon de RCOR3. Existe en CSCD un circRNA (chr1:211485784|211486360) dentro del cual se encuentra contenido el identificado por las herramientas. RCOR3 está considerado como biomarcador favorable en cáncer de pulmón (también en glioma).

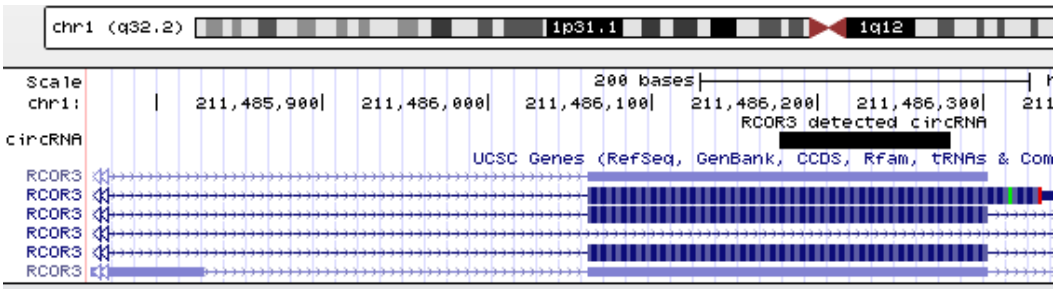


Figura 14. CircRNA chr1:211312835-211312938 vs. exones de RCOR3.

- En cuanto a chr1:44810607-44810720 (+) (chr1:45276279-45276392 en hg19) se corresponde a una parte muy corta de BTBD19, de una región de un intrón y, como comentamos, solo aplica a una muestra. Esta proteína se usa como biomarcador de cáncer renal.

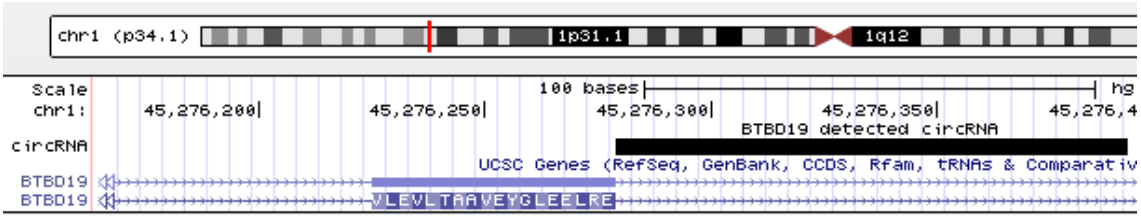


Figura 15. CircRNA chr1:44810607-44810720 vs. exones de BTBD19.

A continuación se muestra una tabla que relaciona cada uno de los circRNAs comentados con la herramienta que lo ha detectado, proporcionando una medida de confianza en su validez. Se ha incluido la columna relativa a KNIFE de forma meramente informativa ya que no ha podido procesar todas las muestras ni toda la información para ellas. No obstante, aquellos circRNAs detectados por KNIFE pueden considerarse con una cierta confianza adicional. La no detección por parte de KNIFE no implica una reducción de confianza pues puede deberse a que no haya procesado las secuencias de las muestras que los incluían.

Tabla 5. Confianza en los circRNA en función de las herramientas.

circRNA	CIRI2	CircExplorer2	KNIFE[2]	Total
chr5:151026992-151027531 (+)	SÍ	NO	NO	1
chr8:22161796-22162648 (+)	SI	SÍ	SÍ	3
chr7:144207692-144207804 (+)	NO	SI	SI	2
chr10:79611910-79612353 (+)	NO	SI	NO	1
chr19:13153357-13153506 (-)	NO	SI	NO	1
chr1:201490196-201496275 (-)	NO	SI	NO	1
chr11:62516794-62517009 (-)	SI	SI	NO	2
chr12:103938418-103939569 (+)	NO	SI	NO	1
chr19:40609745-40609871 (+)	NO	SI	NO	1
chr6:32062157-32067984 (-)	SI	NO	SI	2
chr1:16484823-16631109 (-)	SI	NO	SI	2
chr1:211312835-211312938 (+)	NO	SI	NO	1
chr1:44810607-44810720 (+)	NO	SI	NO	1

[2] Solo para aquellos datos para los que fue procesado.

5. Conclusiones

El objetivo principal del proyecto se ha cumplido pues se ha generado un modelo bioinformático que permite establecer biomarcadores a partir de circRNAs para el adenocarcinoma pulmonar con el fin de clasificar muestras como sanas o tumorales. Este modelo se basa en un árbol de decisión que hace uso de la métrica de *information gain*.

Gracias al pipeline implementado se han detectado 7577 circRNAs. De ellos, 118 se han detectado de forma simultánea por CIRI2 y CIRCExplorer2 en al menos 1 muestra. Tras la normalización de los datos, se ha podido construir un modelo de árbol de decisión que ha sido capaz de predecir con un 91% de precisión (11 de 12 muestras del conjunto de test) la condición normal o tumoral de las muestras. La sensibilidad es de un 83,33% (5/6 muestras tumorales correctamente clasificadas) y la especificidad de un 100% (6/6 muestras normales correctamente clasificadas). Además, se han consultado tres bases de datos de circRNAs y se ha comprobado que los circRNAs más relevantes resultado del trabajo han sido identificados y validados experimentalmente.

Los objetivos parciales consistían en:

- *Desarrollar un pipeline que permita identificar circRNAs.*
  El capítulo 2 describe dicho *pipeline*. Si bien se ha tenido que prescindir de la herramienta KNIFE, los resultados de las otras dos herramientas han satisfecho el objetivo inicial. Además, se ha utilizado KNIFE para analizar el trabajo necesario para añadir otras aplicaciones adicionales de identificación de circRNAs al *pipeline*.
- *Aplicar el pipeline a muestras de adenocarcinoma pulmonar para obtener información de perfiles de expresión de circRNAs en muestras pareadas de tejido tumoral y sano.*
  También descrito en el mismo capítulo, se han utilizado datos de 40 muestras, 20 de tejido con adenocarcinoma pulmonar y 20 de tejido normal. Se ha aplicado el *pipeline* y se ha obtenido un fichero consolidado con una tabla que indica la expresión de circRNAs hallada.
- *Procesar los datos obtenidos de expresión génica generando un modelo que facilite la clasificación de muestras como sanas y tumorales.*
  El capítulo 3 explica cómo se ha normalizado la información obtenida del *pipeline*, divididos los datos en conjuntos de entrenamiento y test, construido el modelo y verificado. Los resultados son especialmente positivos al clasificar solo una de las doce muestras del conjunto de test como errónea.
- *Presentar, a partir del modelo, un conjunto inicial de circRNAs que sean susceptibles de uso como biomarcadores tumorales para el adenocarcinoma pulmonar, con el fin de un futuro análisis y validación más detallado en posteriores proyectos.*
  La parte final del capítulo 3 junto con el capítulo 4 describen este conjunto de circRNAs candidados. En concreto, hay varios de estos circRNAs que habían sido identificados por estudios previos. Otros se corresponden con

exones de genes que se traducen en proteínas que se usan como biomarcadores de algunos tipos de cáncer.

- *Generar la documentación del proyecto y realizar la presentación del mismo.*
  La documentación correspondiente al proyecto incluye la propuesta, el plan de proyecto, los informes de seguimiento y se completa con esta memoria.
  Respecto a la presentación del proyecto, aunque no se encuentra aún completada, se estima que se finalizará en el plazo previsto.

En relación a las lecciones aprendidas en el transcurso del proyecto, la materialización de varios de los riesgos previstos ha hecho que se haya tenido que tomar acciones correctoras. En concreto:

- "Fallo en la disponibilidad de los datos de muestras de adenocarcinoma pulmonar" correspondiente al servidor de The Cancer Genome Atlas". La información en bruto no siempre está disponible de forma directa y el acceso a la misma puede ser complejo.
- "Fallo de los sistemas informáticos utilizados para el desarrollo del TFM o falta de recursos hardware para la ejecución de las herramientas necesarias". La cantidad de recursos requeridos por las herramientas bioinformáticas no siempre se encuentran disponibles. La máquina virtual inicialmente disponible carecía de recursos de memoria, procesador y espacio en disco necesarios para el trabajo de los dos proyectos que albergaba y fue necesario solicitar una más potente.
- "Posibilidad de no conseguir hacer funcionar herramientas disponibles, con licencia abierta o capaces de realizar la identificación de circRNAs". Aunque las herramientas hayan sido utilizadas en estudios previos, su nivel de madurez puede no ser lo suficientemente avanzado, o los recursos que requiere pueden ser demasiado altos para los medios disponibles.
- "Resultados no coherentes de la aplicación del pipeline sobre los datos". El número de circRNAs coincidentes entre las herramientas es más limitado de lo esperado. Se decidió cambiar el enfoque y generar el resultado como unión de los circRNAs detectados por los diferentes programas, pero incluyendo información extra sobre si el mismo ha sido identificado en 1 o 2 de los programas.

Respecto a las líneas de trabajo futuro, se plantean las siguientes:

- Aplicación del *pipeline* a un conjunto mayor de muestras de adenocarcinoma pulmonar, si fuera posible a aquellos sugeridos inicialmente.
- Aplicación del *pipeline* a otros tipos de muestras. Los *scripts* generados no tienen especificidades relativas al adenocarcinoma pulmonar, por lo que se podría utilizar para cualquier otro tipo de cáncer o condición siempre que se disponga de información al respecto.
- Verificación experimental de los circRNAs identificados como posibles biomarcadores de adenocarcinoma pulmonar.
- Incorporación de otras herramientas de identificación del circRNA al análisis para poder obtener un mayor número de circRNAs. Sería posible en estos

casos seleccionar solo aquellos circRNAs que sean identificados por más de una herramienta.

6. Glosario

ADN/DNA: ácido desoxirribonucléico/ deoxyribonucleic acid.
ARN/RNA: ácido ribonucléico/ ribonucleic acid.
BWA: Burrows-Wheeler aligner.
circRNA: circular RNA.
CIRI: Circular RNA Identier
CSCD: Cancer Specific CircRNA Database.
CSV: Comma-Separated Values
HDF5: Hierarchical Data Format 5
KNIFE: Known and Novel IsoForm Explorer
lncRNA: long non-coding RNA
MAQ: Mapping and Assembly with Quality
miRNA: microRNA
NCBI: National Center for Biotechnology Information
PCR: Polymerase chain reaction
RAM: Random Access Memory
RPM: Reads per million mapped reads.
SFF: Standard Flowgram Format
SQL: Structured Query Language
SRA: Sequence Read Archive
STAR: Spliced Transcripts Alignment to a Reference

7. Bibliografía

[1] Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K. et al. Cancer Genome Atlas Research Network. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, *45*(10), 1113-1120.

[2] Las Cifras del Cáncer en España 2017. Sociedad Española de Oncología Médica, 2017 https://www.seom.org/es/prensa/el-cancer-en-espanyacom/105941-las-cifras-del-cancer-en-espana-2017?showall=1 (29/12/2017)

[3] Meng, S., Zhou, H., Feng, Z., Xu, Z., Tang, Y., Li, P., & Wu, M. (2017). CircRNA: functions and properties of a novel potential biomarker for cancer. *Molecular cancer*, *16*(1), 94.

[4] Chen, L. L., & Yang, L. (2015). Regulation of circRNA biogenesis. *RNA biology*, *12*(4), 381-388.

[5] Salzman, J., Gawad, C., Wang, P. L., Lacayo, N., & Brown, P. O. (2012). Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PloS one*, *7*(2), e30733.

[6] Chen, L. L. (2016). The biogenesis and emerging roles of circular RNAs. *Nature reviews Molecular cell biology*, *17*(4), 205-212.

[7] Wang, F., Nazarali, A. J., & Ji, S. (2016). Circular RNAs as potential biomarkers for cancer diagnosis and therapy. *American journal of cancer research*, *6*(6), 1167.

[8] Szabo, L., & Salzman, J. (2016). Detecting circular RNAs: bioinformatic and experimental challenges. *Nature reviews. Genetics*, *17*(11), 679.

[9] Qu, S., Zhong, Y., Shang, R., Zhang, X., Song, W., Kjems, J., & Li, H. (2016). The emerging landscape of circular RNA in life processes. *RNA biology*, 1-8.

[10] Chen, C. Y., & Sarnow, P. (1995). Initiation of protein synthesis by the eukaryotic translational apparatus on circular RNAs. *Science*, *268*(5209), 415.

[11] Wang, Y., & Wang, Z. (2015). Efficient backsplicing produces translatable circular mRNAs. *Rna*, *21*(2), 172-179.

[12] Zeng, X., Lin, W., Guo, M., & Zou, Q. (2017). A comprehensive overview and evaluation of circular RNA detection tools. *PLoS computational biology*, *13*(6), e1005420.

[13] Gao, Y., Zhang, J., & Zhao, F. (2017). Circular RNA identification based on multiple seed matching. *Briefings in Bioinformatics*, 16(4).

[14] Zhang, X. O., Dong, R., Zhang, Y., Zhang, J. L., Luo, Z., Zhang, et al. (2016). Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome research*, *26*(9), 1277-1287.

[15] Szabo, L., Morey, R., Palpant, N. J., Wang, et al. (2015). Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. *Genome biology*, *16*(1), 126.

[16] Seo, J. S., Ju, Y. S., Lee, W. C., Shin, J. Y., Lee, J. K., Bleazard, T. et al. (2012). The transcriptional landscape and mutational profile of lung adenocarcinoma. *Genome research*.

[17] An introduction to Next-Generation Sequencing Technology (2017). Illumina. https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf (29/12/2017)

[18] SRA Handbook (2010-), National Center for Biotechnology Information, https://www.ncbi.nlm.nih.gov/books/NBK242622/ (29/12/2017).

[19] Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research*, *18*(11), 1851-1858.

[20] Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754-1760
[21] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J. et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15-21.
[22] Kim, D., & Salzberg, S. L. (2011). TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome biology*, *12*(8), R72.
[23] Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, *10*(3), R25.
[24] Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), 357-359.
[25] Hipp, R, et. al. (2015). SQLite (Version 3.7.17) [Computer software]. SQLite Development Team. Retrieved November 18, 2017. Available from <https://www.sqlite.org/download.html>
[26] Liftover Genome Annotation Application. Online [https://genome.ucsc.edu/cgi-bin/hgLiftOver].
[27] Glažar, P., Papavasileiou, P., & Rajewsky, N. (2014). circBase: a database for circular RNAs. *Rna*, *20*(11), 1666-1670.
[28] Chen, X., Han, P., Zhou, T., Guo, X., et al (2016). circRNADb: a comprehensive database for human circular RNAs with protein-coding annotations. *Scientific reports*, *6*.
[29] Xia, S., Feng, J., Chen, K., Ma, Y., et al (2017). CSCD: a database for cancer-specific circular RNAs. *Nucleic Acids Research*.
[30] Berman, H, Westbrook, J, Feng, Z, Gilliland, G. et al. (2000) The Protein Data Bank, *Nucleic Acids Research*, 28 (1), 235–242
[31] Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., et al (2017). A pathology atlas of the human cancer transcriptome. *Science*, *357*(6352), eaan2507
[32] Kent, WJ, Sugnet, CW, Furey, TS, Roskin, KM et al (2002). The human genome browser at UCSC. *Genome Res.,* 12(6), 996-1006.
[33] Oh, I. J., Kim, H. E., Song, S. Y., Na, K. J., Kim, K. S., Kim, Y. C., & Lee, S. W. (2014). Diagnostic value of serum glutathione peroxidase 3 levels in patients with lung cancer. *Thoracic cancer*, *5*(5), 425-430.
[34] Cicchini, M., Buza, E. L., Sagal, K. M., Gudiel, A. A., Durham, A. C., & Feldser, D. M. (2017). Context-Dependent Effects of Amplified MAPK Signaling during Lung Adenocarcinoma Initiation and Progression. *Cell reports*, *18*(8), 1958-1969.

8. Anexos

8.1 Listado de programas desarrollados

Los *scripts* utilizados y los ficheros de configuración de apoyo se muestran a continuación:

- *run_pipeline*: script principal del proyecto que realiza llamadas a programas del sistema y al resto de scripts.
- *sra_to_fq*: script que convierte todos los ficheros en formato SRA de un directorio en formato fastq.
- *merge_fa*: script que concatena ficheros FASTA de secuencia de un genoma disponibles en un directorio (todos o solo aquellos que corresponden a cromosomas 1-22, X y M) guardándolos en un fichero.
- *merge_ribo_fa y merge_trans_fa*: scripts que combinan ficheros FASTA de ribosoma o transcriptoma en un único fichero.
- *ciri_pipeline, circexplorer_pipeline (y knife_pipeline)*: scripts que realizan los diferentes pasos requeridos por cada una de las herramientas (generación de índices y alineamiento con bwa/STAR/bowtie) para después llamar a la propia herramienta.
- *process_ciri_results, process_circexplorer_results (y process_knife_results)* que leen los ficheros generados por los *pipelines* e integran la información en la base de datos de SQLite.
- *process_combined*: script en python que procesa el resultado de las dos herramientas (CIRI2 y CIRCExplorer2) y usa la base de datos SQLite para extraer información común.
- *join_samples*: script en python que procesa el resultado consolidado de cada muestra y lo integra en un único fichero común.
- *R_normalize_dectree.Rscript*: script R que se encarga de preparar la información para su procesado, construir el modelo de árbol de decisión, validarlo y obtener la información sobre circRNAs
- *index_files.lnk*: ejemplo de fichero para la descarga de datos.
- *MapSampleReadsCondition*: fichero de configuración que asocia cada muestra con el número de spots y su condición (normal/tumoral).
- *process_combined_3*: script en python que procesa el resultado de las tres herramientas (CIRI2, CIRCExplorer2 y KNIFE) y usa la base de datos SQLite para extraer información común.

8.2 Listado de ficheros de resultado

Se han generado un conjunto de ficheros resultado. En concreto:

- Ficheros con extensión ".ciri" para cada muestra que son el resultado de la herramienta CIRI2.
- Ficheros con sufijo "_known.txt" para cada muestra que son el resultado de la herramienta CircExplorer2.
- Ficheros con sufijo "-consolidated.csv" que son el resultado de procesar las salidas de los distintos programas y dar un resultado integrado.
- Fichero "*consolidated_result*" con los resultados de todas las muestras consolidados en un solo fichero.

- Ficheros “*final_result*” y “*final_result.pdf*” con las salidas del script de R de texto y gráficas respectivamente.
- Ficheros con sufijo “*__circJuncProbs.txt*” para cada muestra que son el resultado de la herramienta KNIFE [3]. Solo se utilizan en un análisis secundario.