Análisis de microRNA (miRNA) i proteínas de unión a RNA (RBPs) involucradas en la función de los circRNA en adenocarcinoma de pulmón

Joan Gibert Fernandez
Máster en bioinformatica y bioestadística

Nombre Consultor/a
Amadís Pagès Pinós
Nombre Profesor/a responsable de la asignatura
Carles Ventura Royo

02/01/2018

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Análisis de microRNA (miRNA) i proteínas de unión a RNA (RBPs) involucradas en la función de los circRNA en adenocarcinoma de pulmón*
Nombre del autor:	*Joan Gibert Fernandez*
Nombre del consultor/a:	*Amadís Pagès Pinós*
Nombre del PRA:	*Carles Ventura Royo*
Fecha de entrega (mm/aaaa):	01/2018
Titulación::	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Genómica Computacional*
Idioma del trabajo:	*Castellano*
Palabras clave	*Adenocarcinoma de pulmón, circRNA, miRNA, RNA binding proteins (RBPs).*

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

El desarrollo tumoral es un proceso de diversos pasos en el que la célula somática acumula alteraciones que le permiten, entre otras funciones, una proliferación descontrolada (Hanahan and Weinberg, 2011). Durante muchos años se ha intentado encontrar formas de detectar estas alteraciones y sus rasgos asociados con el fin de usarlos, también, como posibles dianas terapéuticas, aunque con resultados diferentes en función del origen tumoral (Kalia 2015). En este sentido, el adenocarcinoma de pulmón presenta uno de los índices más bajos de supervivencia a 5 años (18% - Siegel et al. 2016) por su difícil detección y tratamiento.

En este sentido, los RNA circulares (circRNA) se han planteado como moléculas candidatas para la detección precoz del adenocarcinoma de pulmón y se han propuesto diferentes mecanismos por los cuales estas moléculas podrían afectar al desarrollo tumoral (Memczak et al., 2013, Zhang et al., 2013).

En este TFM se observa que existen diferentes circRNAs que se podrían utilizar para discernir entre tejido pulmonar sano y tumoral mediante el análisis de la expresión génica en diferentes muestras de pacientes. Además, también se relacionan algunos de estos circRNA con la regulación de la expresión de diferentes proteínas en coordinación con la maquinaria de silenciamiento celular (miRNAs) o con proteínas asociadas al control traduccional (RBPs).

Abstract (in English, 250 words or less):

Tumor development is a multistep process where the somatic cell accumulates genetic alterations. These alterations give the cell the ability of uncontrolled cell division, between others (Hanahan and Weinberg, 2011). During the previous decades, different approaches in order to identify these alterations and its related traits and to use them as therapeutic targets with different success (Kalia 2015). In this sense, lung adenocarcinoma show one of the worst 5–year prognostic value (18% - Siegel et al., 2016) due to its difficult detection and treatment.

In this sense, circular RNA (circRNA) have been proposed as a candidate molecules for the detection of the lung adenocarcinoma and potential therapeutic markers due to its relation with different mechanisms related with the tumor development (Memczak et al., 2013, Zhang et al., 2013).

In this TFM, different circRNA, which could be used to discern between normal and tumor lung tissue, have been detected. Moreover, these sequences have been related with the protein expression in collaboration with the silencing machinery (miRNA) or with translational control proteins (RBPs).

Índice

Lista de figuras

1. Introducción

1.1 Contexto y justificación del Trabajo

Desde un punto de vista molecular, el cáncer es una enfermedad originada debido a alteraciones genéticas. Consiste en un proceso de diferentes pasos en que las células acumulan alteraciones que hace que las células puedan dividirse de forma aberrante (Hanahan and Weinberg, 2011). Un ejemplo de este desarrollo aberrante es el cáncer de pulmón, que es el tipo de cáncer que representa el mayor número de muertes al año y con uno de los índices más bajos de supervivencia a los 5 años (18%) (Siegel et al., 2016). El cáncer de pulmón se clasifica en cáncer de pulmón de célula pequeña (small cell lung cancer o SCLC, en inglés), que corresponde a un 15% de los casos, y el cáncer de pulmón que no es de célula pequeña (non-small cell lung cancer o NSCLC, en inglés) del cual se diagnostican alrededor del 85% de los casos. De este último, el adenocarcinoma de pulmón es el más común en la población no fumadora. Éste último presenta mutaciones en genes importantes que regulan la división celular como KRAS, EGFR o p53 (Roy et al., 2008).

La falta de biomarcadores para la detección precoz del adenocarcinoma de pulmón hace imprescindible la necesidad de utilizar nuevas aproximaciones. Estos biomarcadores también podrían ser utilizados como dianas terapéuticas en posteriores tratamientos. En este sentido, proponemos utilizar circRNA como biomarcadores del adenocarcinoma de pulmón. Los circRNA son cadenas de RNA circular que han aparecido con mucha fuerza en la literatura científica (Chen, 2016). Se generan, en parte, gracias a la maquinaria de empalmamiento alternativo de la célula y, algunas veces, con la ayuda de otras RBPs (Con et al., 2015 and Ashwal-Fluss et at., 2014).
La utilidad de los circRNA como biomarcadores en cáncer ha sido postulada anteriormente. Los circRNA presentan ciertas ventajas para actuar como tal: son relativamente abundantes debido a su elevada estabilidad por su estructura circular (Salzman et al., 2012), están conservados evolutivamente permitiendo su evaluación en diferentes modelos animales (Jeck et al., 2014) y se han encontrado en estructuras circulantes en la sangre, como los exosomas disminuyendo la invasividad para su detección (Li et al., 2011 y Memczak et al., 2015).
Aunque su potencial como biomarcador es claro, hay poca investigación hecha sobre este rol en adenocarcinoma de pulmón (Lui et al., 2016 y Zhu et al., 2017).
Por otro lado, aunque su función biológica no está del todo clara, se ha demostrado que pueden secuestrar diferentes miRNAs (Memczak et al., 2013) o regular la transcripción de sus RNA mensajeros (mRNA) asociados (Zhang et al., 2013). Este último punto también es importante ya que, además de poder ser utilizados para la detección precoz, discernir los mecanismos que regulan podría dar pistas sobre nuevos tratamientos para el adenocarcinoma de pulmón.

Por todo ello, y con el fin de ampliar el abanico de biomarcadores para el adenocarcinoma de pulmón, nos proponemos analizar la expresión de diferentes circRNA en ésta patología tumoral comparando la expresión de estas secuencias en muestras pareadas de tejido pulmonar tumoral y sano, y a partir de éstos, analizar cuál sería su mecanismo biológico en esta patología

1.2 Objetivos del Trabajo

Objetivos generales

1. Analizar la expresión de circRNA en tejido pulmonar sano y tumoral y identificar los que se encuentran diferencialmente expresados.

2. Explorar los dominios consenso de los circRNA diferencialmente expresados y identificar su capacidad de unión a otros factores implicados en el control traduccional del adenocarcinoma de pulmón (miRNA y RBPs) para inferir un rol biológico.

Objetivos específicos

1.1. Obtener las muestras sanas y con adenocarcinoma de pulmón a analizar a partir de bases de datos.

1.2. Identificar los circRNA diferencialmente expresados comparando tejido sano y tumoral usando programario específico para la detección de circRNA.

2.1. Identificar motivos (secuencias consenso) en la estructura de los circRNA diferencialmente expresados en el tejido tumoral.

2.2. Relacionar estos motivos con diferentes actores relacionados con el control traduccional mediante reconocimiento de secuencias (miRNA) o de motivos (RBPs).

2.3. Identificar en la bibliografía bases de datos donde se determine su relación con los circRNA en ese contexto.

2.4 Describir cual es el rol de estas proteínas en la biología del adenocarcinoma de pulmón para detectar posibles mecanismos de acción.

1.3 Enfoque y método seguido

Durante los últimos años, diferentes aproximaciones para la detección de circRNA han sido descritas por diversos autores (Jeck et al., 2014).

Estas técnicas están basadas, principalmente, en la detección de secuencias con eventos de “backsplicing”. Estas secuencias se caracterizan porque el orden de los exones en estas secuencias se encuentra en una orientación reversa comparándolas con la anotación genómica del organismo en cuestión.

El desarrollo de este TFM viene determinado, principalmente, por la metodología para detectar estos eventos de “backsplicing” y, por lo tanto, la detección de los circRNA. Esto es debido a que los diferentes analizadores de circRNA tienen diferentes requerimientos tanto en el método de secuenciación (*single end* versus *paired end*, por ejemplo) como en los alineadores de la anotación genómica usados o en la manera de enriquecer las secuencias candidatas a ser circRNA (Szabo y Salzman, 2016).

En un experimento ideal, se propondría que las muestras de RNA extraídas de los pacientes y sus respectivos controles fueran, primeramente, depletadas de otras especies de RNAs lineales mediante la degradación de los mismos usando por ejemplo, la RNAsa R (Suzuki et al., 2006). Este primer paso enriquecería las secuencias circulares y, por lo tanto, aumentaría la profundidad de secuenciación de éstos candidatos a circRNA ya de por sí poco abundantes (Salzman et al., 2013).

Dado que en este TFM no se puede realizar este primer paso, la base de datos ideal de adenocarcinoma de pulmón debe ser una con gran profundidad para poder conseguir el mayor número de reads asociados a posibles circRNA (Chen et al., 2015 y Zheng et al., 2017). Para la detección de circRNA en adenocarcinoma de pulmón se ha usado una parte de los datos de un estudio de secuenciación de tejido pulmonar sano y tumoral (15 casos, respectivamente) de diferentes pacientes (Seo J-S et al., 2012 - GSE40419) volcado en el repositorio GEO (https://www.ncbi.nlm.nih.gov/geo/) **(Objetivo 1.1)**.

Para este set de datos se utilizó un secuenciador “Illumina HiSeq 2000 RNA Sequencing” que permite unos 150M de *reads* por cada muestra.

Como ya se ha comentado anteriormente, existen diferentes tipos de analizadores de circRNA en la literatura (Szabo y Salzman 2016). Una revisión exhaustiva de la literatura referente al método para la detección de circRNA sugiere que el paquete MapSplice es el más adecuado (Wang et al., 2010). MapSplice puede trabajar con muestras de RNA-seq tanto *single end* como *paired end* y ha mostrado un alto porcentaje de circRNA validados después de su análisis (Szabo y Salzman 2016). Una vez obtenido el listado de circRNA en cada condición, se han utilizado estos datos para detectar posibles circRNA que puedan ser usados para diferenciar entre tejido sano y tumoral usando un algoritmo de selección (random.forest.importance, del paquete FSelector - https://CRAN.R-project.org/package=FSelector) que selecciona los atributos (en este caso, circRNAs) más relevantes para discernir entre diferentes

condiciones y descartar, a la vez, los atributos menos informativos **(Objetivo 1.2)**.

Una vez seleccionados los circRNA que son capaces de discernir entre tejido normal y tumoral, se ha procedido a analizar las secuencias de los mismos para buscar motivos enriquecidos teniendo en cuenta todas las secuencias identificadas a la vez usando el programario HOMER (Heinz S et al., 2010) **(Objetivo 2.1)**.

Por otro lado y, en este caso, analizando cada circRNA por separado, se han identificado los motivos conservados de unión a miRNA o RBPs en estas secuencias. Para el reconocimiento de interacciones entre miRNA y circRNA se ha usado miRanda (Enright et al., 2004), un analizador de las interacciones de miRNA con otras secuencias del genoma que ya ha sido utilizado para este propósito (Caimen et al., 2015). Por otro lado, para el reconocimiento de sitios de unión a RBPs, se ha utilizado ATTRACT en su versión *online* (v0.99b, Giudice et al., 2016) como analizador de los motivos enriquecidos relacionados con RBPs **(Objetivo 2.2)**.

Por último, se ha procedido a una validación parcial *in silico* de estos resultados. Para las interacciones circRNA-miRNA se ha propuesto estudiar datos de CLIP-seq de AGO2, la proteína encargada de la degradación y/o represión de las secuencias de RNA mediante la maquinaria de silenciamiento genético (Ye et al. 2015), usando la base de datos CLIPdb (Yang et al., 2015) **(Objetivo 2.3)**.

Por otro lado, para confirmar las interacciones circRNA-RBPs se ha usado, también, datos de CLIP-seq específicos de estas proteínas obtenidas de la base de datos ENCODE (The ENODE Project Consortium, 2012) y CLIPdb (Yang et al., 2015) **(Objetivo 2.3)**.

Esto dos últimos apartados permiten comprobar si realmente existe una unión directa entre el circRNA y la maquinaria de miRNA o la RBP candidata.

1.4 Planificación del Trabajo

1.1. Obtener las muestras sanas y con adenocarcinoma de pulmón a analizar a partir de bases de datos.

> Tarea 1. Descargar los datos de Seo J-S et al. (GSE40419) volcado en el repositorio GEO (https://www.ncbi.nlm.nih.gov/geo/). 2 días

1.2. Analizar los circRNA diferencialmente expresados comparando tejido sano y tumoral usando programario específico para la detección de circRNA.

Tarea 1. Identificar circRNAs expresados en tejido sano y tumoral. 1 día

Tarea 2. Proceder al análisis de circRNA diferencialmente expresados:

a. Creación del script de *machine learning* para discriminar circRNA diferencialmente expresados en adenocarcinoma de pulmón. 7 días
b. Aplicación del script a los circRNA identificados en los tejidos sano y tumoral. 7 días

2.1. Identificar motivos (secuencias consenso) en la estructura de los circRNA diferencialmente expresados en el tejido tumoral.

Tarea 1. Identificar, para todos los circRNA detectados, los motivos conservados en sus secuencias mediante HOMER. 8 días

2.2. Relacionar estos motivos con diferentes actores relacionados con el control traduccional mediante reconocimiento de secuencias (miRNA) o de motivos (RBPs).

Tarea 1. Relacionar, mediante programas de descubrimiento de motivos (miRanda), los circRNA con sus miRNA regulados. 5 días

Tarea 2. Relacionar, mediante programas de descubrimiento de dominios (ATTRACT), las RBPs que interaccionan con los circRNA diferencialmente expresados. 5 días

2.3. Identificar en la bibliografía bases de datos donde se determine su relación con los circRNA en ese contexto.

Tarea 1. Buscar en bases de datos de CLIP-seq la unión directa de AGO2 o las RBPs específicas asociados con los circRNA. 14 días

2.4 Describir cual es el rol de estas proteínas en la biología del adenocarcinoma de pulmón para detectar posibles mecanismos de acción.

Tarea 1. Buscar en la bibliografía el rol de los actores identificados en el adenocarcinoma de pulmón. 1 día

La planificación específica de este TFM se ha realizado con el programa ProjectLibre (https://www.projectlibre.com).

Name	Duration	Start	Finish
Desarrollo del trabajo - Fase 1	23 days	**19/10/17 08:00**	**20/11/17 17:00**
1.2 Análisis de circRNA	15 days	**19/10/17 08:00**	**08/11/17 17:00**
1.2.1. Identificación circRNA	1 day	19/10/17 08:00	19/10/17 17:00
1.2.2. Creación del script	10 days	20/10/17 08:00	02/11/17 17:00
1.2.3. Aplicación del script	4 days	03/11/17 08:00	08/11/17 17:00
2.1. Identificación de motivos	8 days	**09/11/17 08:00**	**20/11/17 17:00**
2.1.1. Creación del script	6 days	09/11/17 08:00	16/11/17 17:00
2.1.1. Creación del script	2 days	17/11/17 08:00	20/11/17 17:00
Desarrollo del trabajo - Fase 2	20 days	**21/11/17 09:00**	**19/12/17 09:00**
2.2. Relacionar circRNA con miRNA y RBPs	5 days	**21/11/17 09:00**	**28/11/17 09:00**
2.2.1. Relación circRNA-miRNA (miRanda)	5 days	21/11/17 09:00	28/11/17 09:00
2.2.2. Relación circRNA-RBPs (ATTRACT)	5 days	21/11/17 09:00	28/11/17 09:00
2.3. Buscar información y análisis de candidatos (miRNA/RBPs)	14 days	**28/11/17 09:00**	**18/12/17 09:00**
2.3.1. Busqueda bases de datos de CLIP-seq candidatos	2 days	28/11/17 09:00	30/11/17 09:00
2.3.2. Análsis CLIP-seq	12 days	30/11/17 09:00	18/12/17 09:00
2.4. Busqueda bibliografica del rol de los miRNA/RBPs en adenocarcinoma de pulmón	1 day	18/12/17 09:00	19/12/17 09:00
Redacción de la memoria	10 days	19/12/17 09:00	02/01/18 09:00
Elaboración de la presentación	5 days	03/01/18 09:00	10/01/18 09:00
Defensa pública	7 days	11/01/18 09:00	22/01/18 09:00

Figura 1. Planificación del proyecto

El diagrama de Gantt se adjunta en la figura suplementaria 1.

1.5 Breve sumario de productos obtenidos

El producto que se generará de este TFM será una relación de los circRNA diferencialmente expresados y su posible en rol biológico asociado con otros actores implicados en el control traduccional (miRNA/RBPs). Además, también se generarán una colección de scripts que llevarán a cabo tanto el análisis de los circRNA diferencialmente expresado como el descubrimiento de motivos de los mismos.

1.6 Breve descripción de los otros capítulos de la memoria

Contexto: En este apartado se pone en contexto del control transcripcional y traduccional, sus desregulaciones en cáncer y la presencia de los circRNA como parte de estos ambientes regulatorios. También se introduce la patología de estudio y la problemática a resolver.
Métodos: En este apartado se describen los diversos pasos efectuados para llevar a cabo el análisis de este TFM.
Resultados y Discusión: En este apartado se muestran los resultados obtenidos después del análisis así como la discusión de los mismos.
Conclusiones

2. Contexto

2.1. Del “Central Dogma” a la regulación génica

Francis Harry Compton Crick, junto con James Dewey Watson y Maurice Hugh Frederick Wilkins ganaron el Premio Nobel de Fisiología o Medicina en 1962 *“for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material"* (Nobelprize.org).

Usando datos de cristalografía de rayos X y construcción de modelos, fueron capaces de proponer un modelo que explica cómo se almacena la información dentro de una célula: la estructura de doble hélice del DNA (Watson y Crick, 1953).

Sin embargo, esta no era la única pregunta que Crick quería abordar en su carrera científica. Después de resolver la estructura de la información genética, Crick, centró su interés en una cuestión instrumental de la biología: ¿cómo codifica el DNA las proteínas? El misterio comienza a resolverse en 1958 cuando Crick publicó "On Protein Synthesis" donde Crick presenta el “Central Dogma” (Figura 2) de la biología que, hasta día de hoy es mayormente válido.

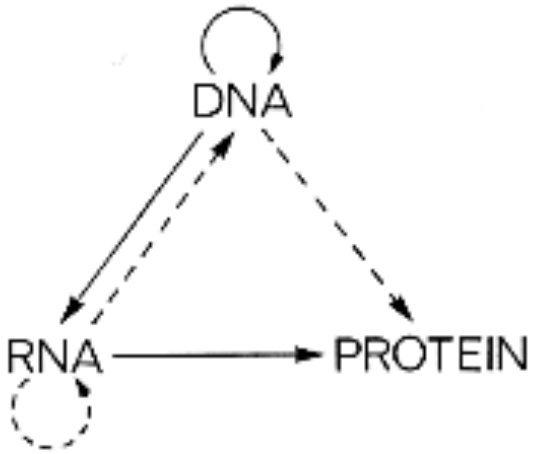


Figura 2. El “Central Dogma” de la expresión proteica propuesto por F. Crick en 1970.

Sin embargo, el “Central Dogma” le da un gran problema a la ciencia: ¿Cómo se controla esta transferencia de información? Esta pregunta se evaluó primero en bacterias. François Jacob y Jacques Lucien Monod declararon la presencia de "genes reguladores especializados", que controlan la transferencia de información del DNA a la proteína (Jacob y Monod, 1961).

En eucariotas, sin embargo, el proceso se vislumbra mucho más complejo. El primer paso para la producción de proteínas se lleva a cabo en el núcleo celular, donde el DNA es transcrito a RNA (Lee y Young, 2000), se lleva a cabo el empalmamiento alternativo (Papasaikas y Valcarcel, 2016) y, finalmente, el RNA es poliadenilado (Wahle, 1995) y translocado al citoplasma. Este proceso está altamente regulado en sus diferentes pasos mediante factores de transcripción (Gill, 2001) o remodeladores de la cromatina (Orphanides y Reinberg, 2002), per ejemplo.

Una vez el RNA se encuentra en el citoplasma, la maquinaria traduccional busca el codón de inicio y empieza a traducir el RNA maduro mediante un proceso complejo y altamente regulado (Figura 3 - Hinnebusch y Lorsch, 2012). Sin embargo, diferentes mecanismos relacionados con estos complejos de traducción pueden regular la vida media del RNA a traducir mediante el escaneo de mutaciones en la secuencia del RNA (van Hoof y Wagner, 2011), el almacenamiento de los RNA (Decker y Parker, 2012) o algunos mecanismos que regulan la expresión de RNA específicos utilizando secuencias específicas de reconocimiento (miRNAs - Lim et al., 2005) u otras proteínas (*RNA binding proteins* o RBPs - Castello et al., 2012).

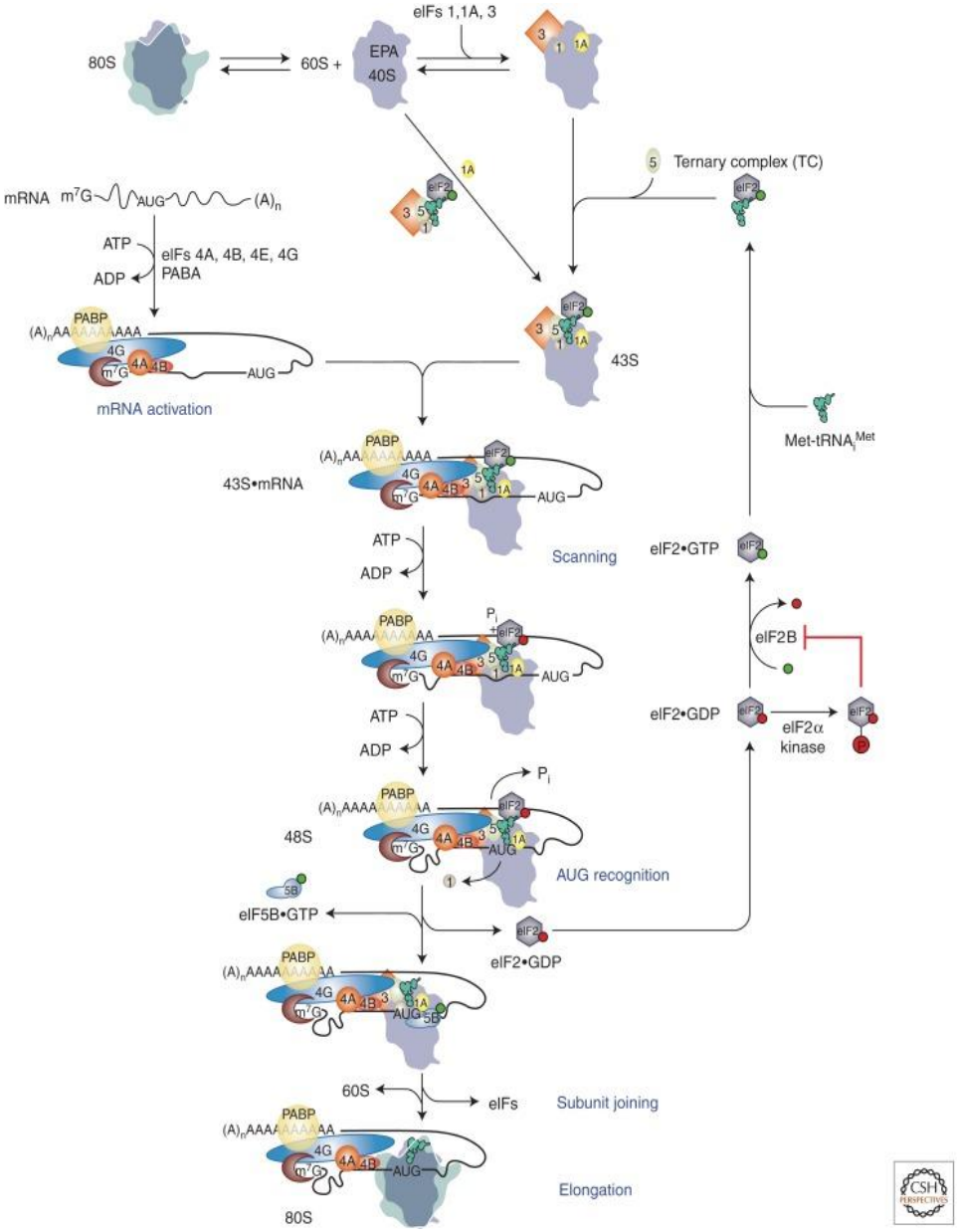


Figura 3. Esquema de la traducción dependiente de Cap. El complejo de preiniciación 43S (PIC) está formado por la adición de eIF1, 1A y 3 y el complejo ternario (TC) a la subunidad 40S. En paralelo, se forma el RNA circularizado y el PIC 43S escanea el 5'UTR hasta que encuentra un codón de inicio AUG adecuado. Estos desencadenan la hidrólisis de eIF2-GTP a eIF2-GDP y su liberación y reciclaje. Luego, la subunidad ribosómica 60S reconoce el 43S y forma el complejo de iniciación 80S con la hidrólisis de eIF5B-GTP a eIF5B-GDP y la liberación del resto de eIF (de (Hinnebusch y Lorsch, 2012).

Además de todos estos mecanismos de regulación, existen otras moléculas capaces de hacer de nexo entre la maquinaria transcripcional y la traduccional. En este conjunto encontramos los RNA circulares (circRNA). Los circRNA son cadenas de RNA circular que se generan gracias a la maquinaria de empalmamiento alternativo de la célula, en un fenómeno descrito como "backsplicing" que consiste en la unión inversa entre el final exón posterior con el inicio del exón anterior (Figura 4 - Huang et al., 2017). Algunas veces, este evento se lleva a cabo con la ayuda de otras RBPs (Con et al., 2015 y Ashwal-Fluss et at., 2014). Aunque su función biológica no está del todo clara, se ha demostrado que pueden regular la transcripción de sus RNA mensajeros (mRNA) asociados (Zhang et al., 2013) contribuyendo, así a la regulación transcripcional, y también, secuestrar diferentes miRNAs (Memczak et al., 2013) o incluso, RBPs (Ashwal-Fluss et at., 2014) regulando, de esta manera, la maquinaria de control traduccional. Además, los circRNA pueden modificar directamente el ratio de traducción de los RNA mensajeros cuando el primer exón se encuentra presente en el circRNA, este fenómeno se conoce como "mRNA traps" (Floris et al., 2017).

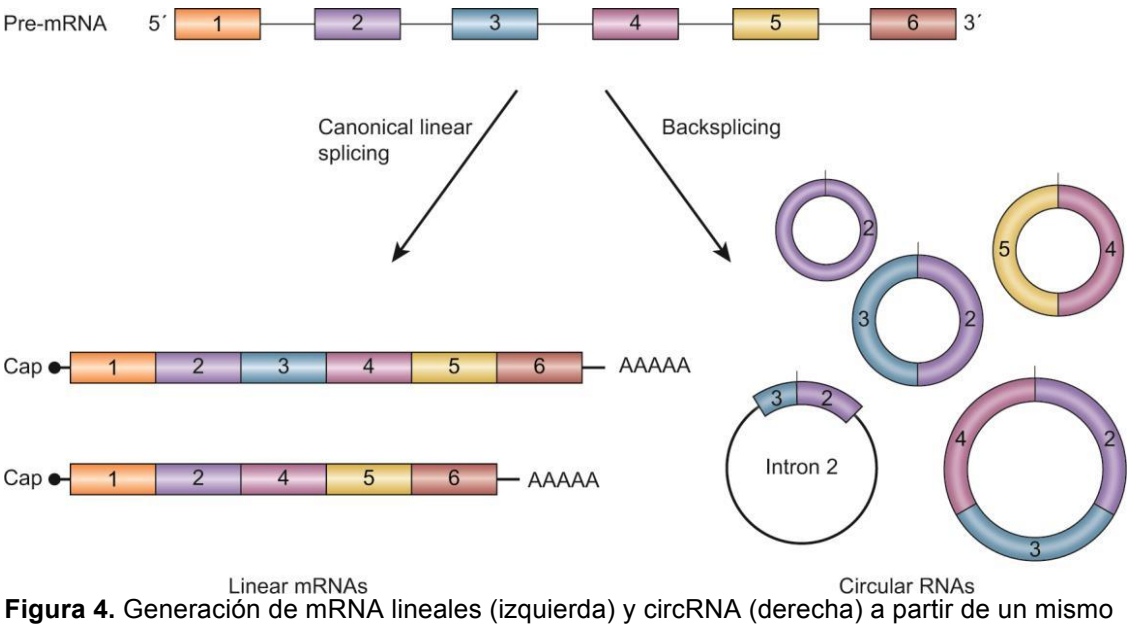


Figura 4. Generación de mRNA lineales (izquierda) y circRNA (derecha) a partir de un mismo pre-mRNA (de Huang et al., 2017).

2.2. Los "Hallmarks" del Cáncer y la regulación génica

Desde un punto de vista molecular, el cáncer es una enfermedad genética. Consiste en un proceso de pasos múltiples donde las células normales acumulan gradualmente alteraciones que les permiten superar su control homeostático y dividirse aberrantemente (Hanahan y Weinberg, 2011). Sin embargo, las células cancerosas no son un sistema aislado. La relación entre las células cancerosas y su microambiente circundante tiene un gran impacto en el desarrollo del cáncer. Estas interacciones, junto con sus vías desreguladas internas, determinan su desarrollo y regulan sus capacidades o "Hallmarks" (Figura 5).

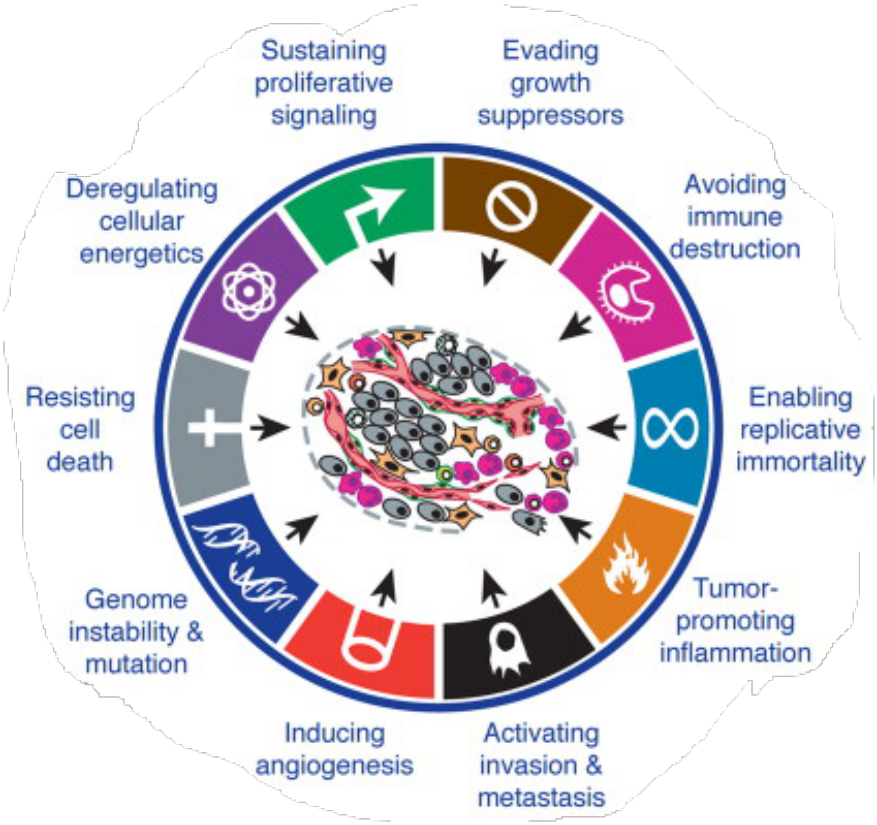


Figura 5. Los "Hallmarks" del cáncer. Las características distintivas de cáncer son vías que encierran los procesos más importantes en la biología de una célula cancerosa (de Hanahan y Weinberg, 2011).

Cada característica distintiva abarca alteraciones biológicas concretas en la fisiología celular que dictan el crecimiento maligno. Al observar la transferencia de información presentada al comienzo de esta Introducción, es lógico pensar que tanto la maquinaria de transcripción (factores de transcripción, remodeladores de cromatina o modificadores de histonas, entre otros - Lee y Young, 2013) juntamente con la maquinaria traduccional (miRNAs - Lin y Gregory, 2015; empalme alternativo - Chabot y Shkreta, 2016; estructuras secundarias en las zonas 5' no traducidas o 5'UTR como G-quadruplexes - Wolfe et al., 2014; otras secuencias 5'UTR o 3'UTR reconocidas por proteínas de unión a RNA - Xue et al., 2015 y Pereira et al., 2017 además de la maquinaria de traducción en sí - Ruggero, 2013) son usadas por las células cancerígenas en su propio beneficio. Estas vías desreguladas son, además, una gran fuente de información en la identificación de biomarcadores, donde muchas de estas moléculas implicadas han sido descritas para la diagnosis, estratificación predicción e, incluso, eficiencia de fármacos para distintos tipos de cáncer (Kamel y Al-Amodi, 2017).

En este sentido, los circRNA no son una excepción. Aunque su descubrimiento es relativamente reciente, muchos de ellos se han encontrado diferencialmente expresados en distintos tipos de cáncer y han sido descritos para diferentes funciones comentadas con anterioridad (Figura 6 – Kristensen et al., 2017).

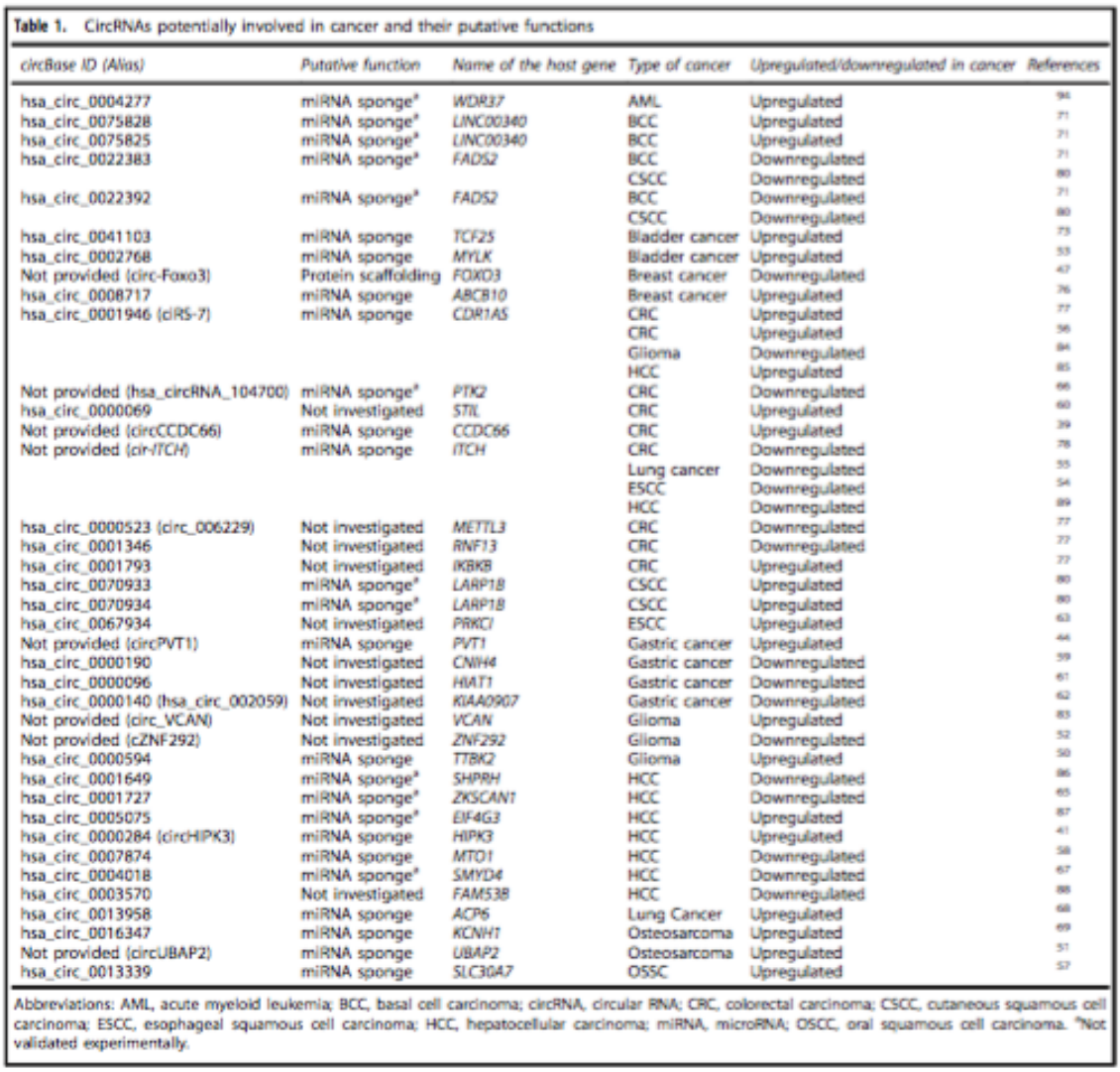


Table 1. CircRNAs potentially involved in cancer and their putative functions

circBase ID (Alias)	*Putative function*	*Name of the host gene*	*Type of cancer*	*Upregulated/downregulated in cancer*	*References*
hsa_circ_0004277	miRNA sponge[a]	*WDR37*	AML	Upregulated	94
hsa_circ_0075828	miRNA sponge[a]	*LINC00340*	BCC	Upregulated	71
hsa_circ_0075825	miRNA sponge[a]	*LINC00340*	BCC	Upregulated	71
hsa_circ_0022383	miRNA sponge[a]	*FADS2*	BCC	Downregulated	71
			CSCC	Downregulated	80
hsa_circ_0022392	miRNA sponge[a]	*FADS2*	BCC	Downregulated	71
			CSCC	Downregulated	80
hsa_circ_0041103	miRNA sponge	*TCF25*	Bladder cancer	Upregulated	73
hsa_circ_0002768	miRNA sponge	*MYLK*	Bladder cancer	Upregulated	53
Not provided (circ-Foxo3)	Protein scaffolding	*FOXO3*	Breast cancer	Downregulated	47
hsa_circ_0008717	miRNA sponge	*ABCB10*	Breast cancer	Upregulated	76
hsa_circ_0001946 (ciRS-7)	miRNA sponge	*CDR1AS*	CRC	Upregulated	77
			CRC	Upregulated	56
			Glioma	Downregulated	84
			HCC	Upregulated	83
Not provided (hsa_circRNA_104700)	miRNA sponge[a]	*PTK2*	CRC	Downregulated	66
hsa_circ_0000069	Not investigated	*STIL*	CRC	Upregulated	60
Not provided (circCCDC66)	miRNA sponge	*CCDC66*	CRC	Upregulated	39
Not provided (cir-*ITCH*)	miRNA sponge	*ITCH*	CRC	Downregulated	78
			Lung cancer	Downregulated	55
			ESCC	Downregulated	54
			HCC	Downregulated	89
hsa_circ_0000523 (circ_006229)	Not investigated	*METTL3*	CRC	Downregulated	77
hsa_circ_0001346	Not investigated	*RNF13*	CRC	Downregulated	77
hsa_circ_0001793	Not investigated	*IKBKB*	CRC	Upregulated	77
hsa_circ_0070933	miRNA sponge[a]	*LARP1B*	CSCC	Upregulated	80
hsa_circ_0070934	miRNA sponge[a]	*LARP1B*	CSCC	Upregulated	80
hsa_circ_0067934	Not investigated	*PRKCI*	ESCC	Upregulated	63
Not provided (circPVT1)	miRNA sponge	*PVT1*	Gastric cancer	Upregulated	44
hsa_circ_0000190	Not investigated	*CNIH4*	Gastric cancer	Downregulated	59
hsa_circ_0000096	Not investigated	*HIAT1*	Gastric cancer	Downregulated	61
hsa_circ_0000140 (hsa_circ_002059)	Not investigated	*KIAA0907*	Gastric cancer	Downregulated	62
Not provided (circ_VCAN)	Not investigated	*VCAN*	Glioma	Upregulated	83
Not provided (cZNF292)	Not investigated	*ZNF292*	Glioma	Downregulated	52
hsa_circ_0000594	miRNA sponge	*TTBK2*	Glioma	Upregulated	50
hsa_circ_0001649	miRNA sponge[a]	*SHPRH*	HCC	Downregulated	86
hsa_circ_0001727	miRNA sponge[a]	*ZKSCAN1*	HCC	Downregulated	65
hsa_circ_0005075	miRNA sponge[a]	*EIF4G3*	HCC	Upregulated	87
hsa_circ_0000284 (circHIPK3)	miRNA sponge	*HIPK3*	HCC	Upregulated	41
hsa_circ_0007874	miRNA sponge	*MTO1*	HCC	Downregulated	58
hsa_circ_0004018	miRNA sponge[a]	*SMYD4*	HCC	Downregulated	67
hsa_circ_0003570	Not investigated	*FAM53B*	HCC	Downregulated	88
hsa_circ_0013958	miRNA sponge	*ACP6*	Lung Cancer	Upregulated	68
hsa_circ_0016347	miRNA sponge	*KCNH1*	Osteosarcoma	Upregulated	69
Not provided (circUBAP2)	miRNA sponge	*UBAP2*	Osteosarcoma	Upregulated	51
hsa_circ_0013339	miRNA sponge	*SLC30A7*	OSSC	Upregulated	57

Abbreviations: AML, acute myeloid leukemia; BCC, basal cell carcinoma; circRNA, circular RNA; CRC, colorectal carcinoma; CSCC, cutaneous squamous cell carcinoma; ESCC, esophageal squamous cell carcinoma; HCC, hepatocellular carcinoma; miRNA, microRNA; OSCC, oral squamous cell carcinoma. [a]Not validated experimentally.

Figura 6. circRNA implicados en cáncer y su posible función (Kristensen et al. 2017).

En este sentido, la función más estudiada de los circRNAs en cáncer es la regulación de la actividad de diferentes especies de miRNAs mediante su capacidad “esponja”, es decir, secuestrando diferentes miRNAs y evitando que ejerzan su función de silenciamiento génico (Meng et al., 2017).

Los circRNA presentan ciertas ventajas para actuar como biomarcadores: son relativamente abundantes debido a su elevada estabilidad por su estructura circular (Salzman et al., 2012), están conservados evolutivamente permitiendo su evaluación en diferentes modelos animales (Jeck et al., 2014), tienen una alta especificidad de tejido (Meng et al., 2017) y se han encontrado en estructuras circulantes en la sangre, como los exosomas, disminuyendo la invasividad para la detección tumoral (Li et al., 2011 y Memczak et al., 2015).

2.3. El adenocarcinoma de pulmón

El cáncer de pulmón es el tipo de cáncer que representa el mayor número de muertes al año y con uno de los índices más bajos de supervivencia a los 5 años (18%) (Siegel et al., 2016). El cáncer de pulmón se clasifica en cáncer de pulmón de célula pequeña (small cell lung cancer o SCLC, en inglés), que corresponde a un 15% de los casos, y el cáncer de pulmón que no es de célula pequeña (non-small cell lung cancer o NSCLC, en inglés) del cual se diagnostican alrededor del 85% de los casos. De este último, el adenocarcinoma de pulmón es el más común en la población no fumadora. Éste último presenta mutaciones en genes importantes que regulan la división celular como KRAS, EGFR o p53 (Roy et al., 2008).

La falta de herramientas para la detección precoz del adenocarcinoma de pulmón es la causa principal de la baja supervivencia de estos pacientes (Siegel et al., 2016), haciendo necesaria la investigación de nuevos biomarcadores para su detección. Los primeros biomarcadores con valor pronóstico en esta patología han estado muy relacionados con las mutaciones indicadas en el párrafo anterior aunque muchos de ellos presentan desventajas relativas a su baja detección y especificidad (Thunnissen et al., 2008). Con la aparición de la secuenciación masiva, diferentes tipos de miRNAs se han visto identificados en esta patología aunque con resultados, algunas veces, contradictorios (Wang et al., 2015).

Gracias a las ventajas en estabilidad y especificidad indicadas con anterioridad en este TFM se plantea la hipótesis que los circRNAs pueden ser unos potentes biomarcadores y, gracias a su función tamponadora de la maquinaria de control traduccional, pueden ser considerados unas dianas terapéuticas de gran interés. Por ello, y con el fin de ampliar el abanico de biomarcadores para el adenocarcinoma de pulmón, nos proponemos analizar la expresión de diferentes circRNA en ésta patología tumoral comparando la expresión de estas secuencias en tejido pulmonar y, a partir de estos, analizar cuál sería su mecanismo biológico en esta patología.

3. Métodos

3.1. Identificación de los circRNAs diferencialmente expresados en adenocarcinoma de pulmón

Para la detección de las diversas especies de circRNAs diferencialmente expresadas en adenocarcinoma de pulmón, se realizó una comparación entre muestras de tejido normal y tumoral de las cuales se había secuenciado todo su transcriptoma mediante RNA-seq.

En el análisis ideal, se hubiera requerido de una depleción inicial de los RNA lineales de estas muestras secuenciadas para que la cantidad de *reads* asociados a los circRNAs fuera la mayor posible. Sin embargo, la no disposición de estos datos nos hizo trabajar con otros sets sin enriquecimiento. Este hecho es importante de cara al análisis y la interpretación de los resultados.

Otra consideración a tener en cuenta es que quisimos utilizar las muestras volcadas en el repositorio del TCGA (https://cancergenome.nih.gov). En esta base de datos aparecen centralizados la mayoría de los estudios realizados usando secuenciación masiva, tanto a nivel de DNA, RNA e, incluso, proteína. Concretamente, se pretendía usar los datos TCGA LUAD, donde se han realizado secuenciación masiva de RNA (RNA-seq) de 515 pacientes. Sin embargo, la no disposición de los datos crudos nos hizo cambiar a otro set de datos, también público, volcado en el repositorio GEO (GSE40419). Aunque este hecho no implica ninguna diferencia *per se*, si que es importante decir que hubiera sido interesante obtener los datos del TCGA ya que hubiéramos podido contrastar los resultados de los circRNA con otros experimentos *high-throughput* disponibles en el repositorio como miRNA-seq o el proteoma de esos pacientes, aunque estos análisis no se encuentren especificados en los objetivos.

Una vez obtenidos las muestras de cada paciente, la identificación de los circRNAs de cada muestra se llevó a cabo mediante el software MapSplice (Wang et al., 2010) usando el clúster de cálculo cedido por la UOC para este TFM. Una vez obtenidos los datos de expresión de los circRNA en cada muestra, se decidió hacer una agrupación de los circRNAs identificados en regiones cercanas (distancia <100 pares de bases – bp). Este paso previo antes del análisis se llevó a cabo para aumentar la potencia de predicción del algoritmo de machine learning usado ya que, teniendo en cuenta la baja cantidad de *reads* asociados a circRNA en las muestras, se prevé que muchos de estos circRNA identificados por MapSplice en regiones cercanas sean, ciertamente, la misma especie de circRNA (Figura 7).

Una vez obtenidas las regiones determinadas para cada circRNA, hemos usado el algoritmo de selección Random Forest (random.forest.importance, del paquete FSelector - https://CRAN.R-project.org/package=FSelector). Random Forest consiste en el cálculo de los pesos de cada atributo (en este caso, los circRNAs) que se encuentran en cada nivel de expresión (en este caso, las

muestras secuenciadas) para discernir entre las muestras normales y tumorales. Es decir, Random Forest selecciona los diferentes atributos (circRNA) que mejor diferencian las muestras normal y tumoral. Random Forest hace este cálculo de predicciones a partir de una partición recursiva de las muestras (Lanz, 2015). Este hecho hace que este algoritmo pueda trabajar bien en condiciones de pocas muestras y muchas variables (atributos), problema conocido como “curse dimensionatility” (Bellman RE, 1957).

Este algoritmo de selección se encuentra entre los más precisos en la literatura y, además, es capaz de clasificar cada atributo identificado (en este caso, los circRNAs) de mayor a menor importancia (Touw et al., 2013). Además, Random Forest ya ha sido ya utilizado para la identificación de circRNAs en otros tipos de cáncer comparándolos, también, con tejido sano (Li et al., 2017).

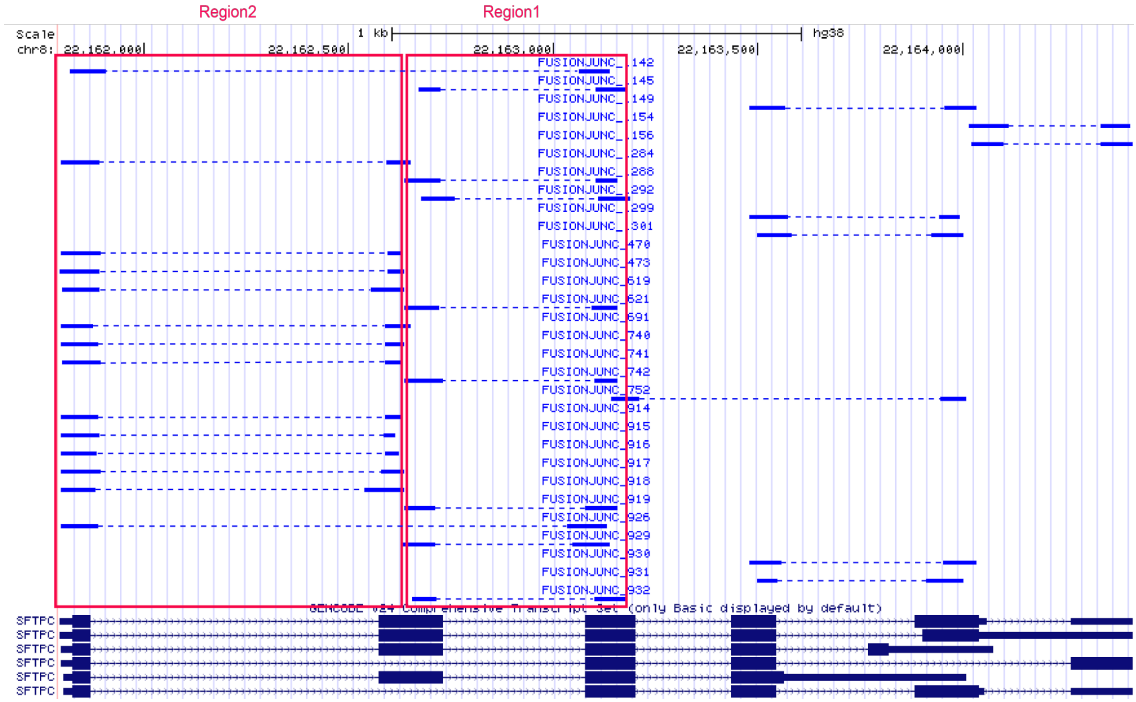


Figura 7. Presencia de circRNA en la región del cromosoma 8 asociada al gen SFTPC visualizado con Genome Browser (Kent et al., 2002). En azul oscuro, se aprecian las diferentes isoformas del gen SFTPC. En azul claro, se muestran las diferentes secuencias de circRNA identificadas por MapSplice. Las regiones en rojo corresponden a las dos regiones identificadas.

Un resumen de los pasos seguidos en este primer apartado se plasman en la figura resumen siguiente:

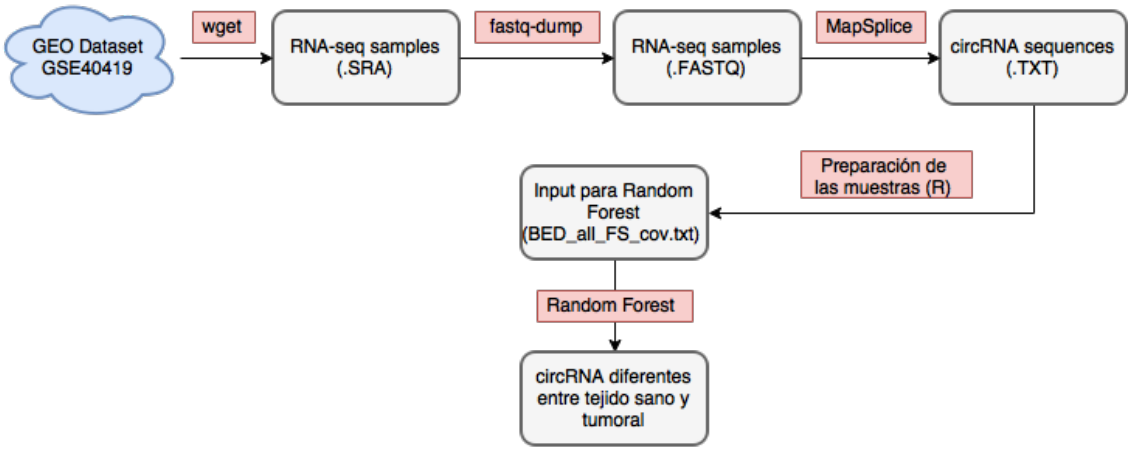


Figura 8. Diagrama de flujo de los pasos llevados a cabo en la primera fase del proyecto

3.2. Identificación de los motivos en la estructura de los circRNA diferencialmente expresados en el tejido tumoral de forma global y específica

Una vez identificados los diferentes circRNA diferencialmente expresados en adenocarcinoma de pulmón, se decidió hacer una primera exploración de los motivos consenso de todas las especies identificadas. Esta primera exploración se llevó a cabo con el programario HOMER (Heinz S et al., 2010) para obtener los motivos conservados en estas regiones. En esta primera exploración se pretendía discernir si había algún tipo de vía celular consenso entre los circRNAs identificados. Para ello, se cogieron las coordenadas genómicas de las regiones identificadas como circRNA diferencialmente expresados y se usaron para la identificación de los motivos. Es importante remarcar que en este análisis no se pudo plasmar la "circularidad" de las especies de los motivos identificados a causa de la necesidad de introducir coordenadas genómicas en el programa. Esto podría haber provocado la pérdida de algún motivo debido a la circularización del circRNA, cosa que si se ha tenido en cuenta en los análisis posteriores.

Con este último propósito, se construyó un archivo FASTA con las secuencias de los circRNAs detectados por MapSplice. En este caso, se han tenido en cuenta tanto las partes codificantes del mRNA maduro (exones), como las no traducidas (UTRs) en el caso de que alguno de los dos exones (donor o acceptor) presentaran secuencias UTR asociadas obtenidas de la base de datos GENCODE. La decisión de añadir las secuencias UTRs al circRNA ha sido basada en la bibliografía disponible. En este sentido, se han descrito artículos científicos explicando la presencia de regiones UTRs en la secuencia de algunos circRNA (Guo et al. 2014, Filippenkov et al. 2015). Por otro lado, y de cara al análisis de los motivos enriquecidos, se han usado las secuencias identificadas y se les ha añadido 50 bp de la primera región al final de la segunda con el fin de generar una secuencia semejante a la del circRNA de cara a la detección de motivos.

Para la detección de las secuencias de unión de miRNAs, se usó el programario miRanda (miRanda-3.3a - Enright et al., 2004) instalado en la máquina virtual cedida por la UOC. Para identificar dichas regiones se utilizó un

archivo FASTA con los miRNA maduros en humano provenientes de la misma página (Figura 9). Por otro lado, para obtener las secuencias de unión a miRNA más significativas, se decidió mantener únicamente las uniones con un “Score” mayor de 150 y una energía libre menor de -20, características descritas con anterioridad en otros estudios (Richardson et al. 2011).

Además, con la finalidad de observar si los motivos identificados presentan un incremento respecto a una secuencia aleatoria, se generaron 200 versiones “shuffled” de la misma longitud que los circRNA identificados mediante el programa “fasta-shuffle-letters” del programario MEME (meme_4.12.0 - Timothy et al., 2009) y se identificaron los motivos de unión a los miRNA de la misma forma descrita anteriormente. Seguidamente se obtuvo un Z-score de cada circRNA comparándolo con las secuencias “shuffled” de la misma longitud. Este cálculo se hizo para plasmar el enriquecimiento de los lugares de reconocimiento de miRNA en los circRNA respecto a secuencias aleatorias. El valor de Z-score se obtuvo restando la media de los miRNA identificados en las secuencias “shuffled” (mS) al número de miRNA identificados en los circRNA (C) y dividido por la desviación estándar de los miRNA identificados en las secuencias “shuffled” (sdS), es decir: (C - mS)/sdS.

```
>hsa-miR-1183  INADL:1:62161618:62162663_50   142.00  -19.48  3 24    904 931 22      68.18%  72.73%
>hsa-miR-1184  DCN:12:91146252:91153096_50    140.00  -16.77  2 22    236 259 21      71.43%  80.95%
>hsa-miR-1184  RPS24:10:78037194:78040713_50  145.00  -18.04  3 22    113 135 19      68.42%  78.95%
>hsa-miR-1197  CDR1:X:140783176:140784660_50  140.00  -14.38  2 19    411 430 17      70.59%  82.35%
>hsa-miR-1197  CDR1:X:140783176:140784660_50  140.00  -14.38  2 19    450 469 17      70.59%  82.35%
>hsa-miR-1202  INADL:1:62161618:62162663_50   147.00  -19.55  2 16    574 594 14      78.57%  92.86%
>hsa-miR-1202  RPS24:10:78037194:78040713_50  151.00  -22.40  2 20    87 107  18      61.11%  77.78%
>hsa-miR-1202  SFTPC:8:22161655:22162732_50   144.00  -22.46  2 15    49 68   13      69.23%  92.31%
>hsa-miR-1203  SFTPC:8:22161655:22162732_50   144.00  -20.55  2 19    1 17    17      70.59%  88.24%
>hsa-miR-1203  SFTPC:8:22161655:22162732_50   149.00  -23.53  2 20    373 392 18      72.22%  88.89%
>hsa-miR-1205  SFTPC:8:22161655:22162732_50   159.00  -21.99  2 18    150 171 18      77.78%  88.89%
>hsa-miR-1206  SFTPC:8:22161655:22162732_50   142.00  -12.22  2 19    338 358 17      58.82%  88.24%
>hsa-miR-1206  SFTPC:8:22162574:22163202_50   142.00  -12.22  2 19    122 142 17      58.82%  88.24%
>hsa-miR-1208  FN1:2:215364985:215365505_50   145.00  -18.48  2 10    34 53   8       100.00% 100.00%
>hsa-miR-1208  INADL:1:62161618:62162663_50   146.00  -15.05  2 13    519 537 11      90.91%  90.91%
>hsa-miR-122-3pFN1:2:215364985:215365505_50   145.00  -16.90  2 19    106 130 20      70.00%  80.00%
```

Figura 9. Ejemplo del análisis reportado por miRanda. En el archivo se identifican por orden, el miRNA unido al circRNA, el circRNA, el Score asociado, la Energía libre asociada, la región del miRNA unida, la región del circRNA unida, la longitud de la unión y los porcentajes de nucleótidos unidos en esa región en el miRNA y el circRNA, respectivamente.

Para la detección de los sitios de unión de RBPs se usaron las mismas secuencias descritas en el apartado anterior. En este caso, se detectaron motivos de unión de RBPs mediante ATTRACT (versión 0.99b, Gaudice et al. 2016). El hecho del uso de ATTRACT se basa en que este programa detecta únicamente dominios de RBPs y no de otras proteínas de unión a DNA y que, además, genera directamente un “likehood ratio” respecto a la presencia de ese dominio de unión a RBPs del resto de secuencias del transcriptoma (Gaudice et al. 2016).

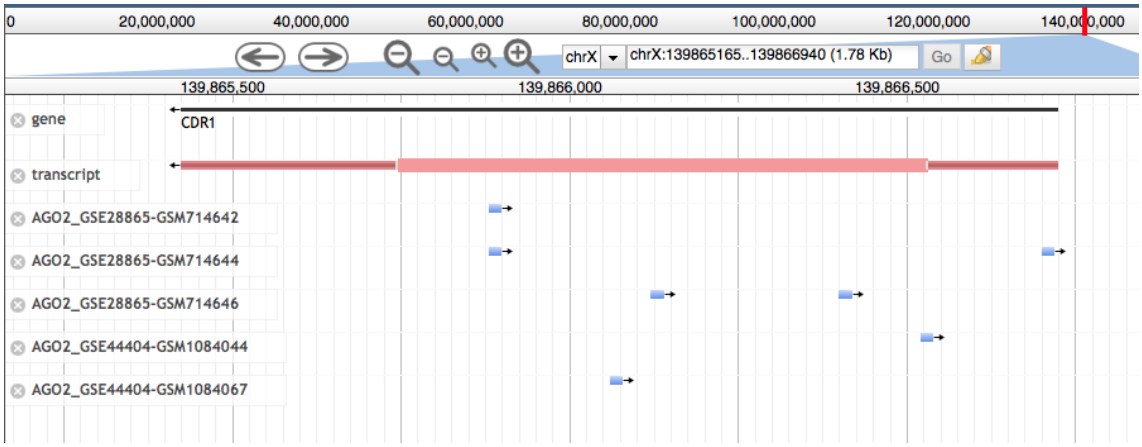
Para completar el análisis y con el fin de explorar cualitativamente los diferentes motivos de unión a RBPs detectados en los circRNAs, se generó un gráfico de intersección mediante UpSetR (https://cran.r-project.org/web/packages/UpSetR/index.html) y se destacaron aquellas RBPs que han sido descritas como proteínas de empalmamiento (GO:0008380) en la base de datos BioMart usando el paquete biomartR (https://CRAN.R-project.org/package=biomartr).

3.3. Validación *in silico* de motivos de la unión de miRNA y RBPs a los circRNA mediante datos de CLIP-seq y búsqueda bibliográfica de esta relación

Para la validación de la unión entre miRNA identificados y los circRNA diferencialmente expresados se utilizaron datos de CLIP-seq de AGO2, una de las proteínas más importantes en el mecanismo de silenciamiento mediado por miRNAs (Ye et al. 2015). Para ello se ha usado la base de datos CLIPdb (http://lulab.life.tsinghua.edu.cn/clipdb/). De esta manera, se pudo obtener evidencia directa de la interacción de la maquinaria de silenciamiento mediante miRNA con el circRNA en cuestión.

Para la validación de la unión entre RBPs identificadas y los circRNA diferencialmente expresados se usaron los datos de CLIP-seq de dos repositorios distintos. El primero, se analizaron los datos de CLIP-seq de las RBPs detectadas en el proyecto ENCODE (https://www.encodeproject.org) donde hay presentes un elevado número de CLIP-seq en dos líneas celulares (línea de carcinoma hepático HepG2 - https://www.atcc.org/products/all/HB-8065.aspx y línea celular de leucemia mieloide crónica K562 - https://www.atcc.org/products/all/CCL-243.aspx). En el segundo, se han buscado los datos de CLIP-seq de las RBPs detectadas en la base de datos CLIPdb (http://lulab.life.tsinghua.edu.cn/clipdb/).

Estos dos análisis se realizaron de forma cualitativa mediante la visualización de los *reads* de cada proteína en cada región de circRNA específica identificada mediante el uso de un visualizador genómico (Figura 10). Para la introducción de estas regiones en ENCODE, se ha usado la localización del genoma GRCh38/hg38 mientras que para CLIPdb dichas regiones han sido transformadas a GRCh37/hg19 por motivos de compatibilidad.



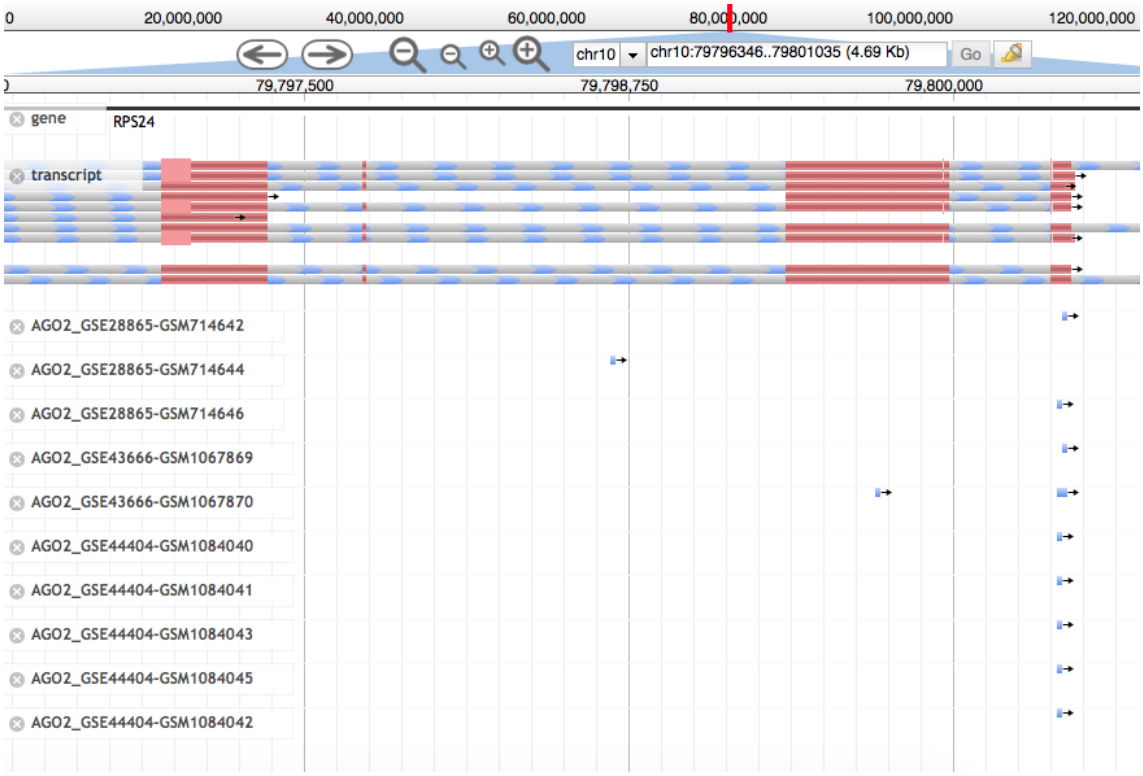


Figura 10. Capturas de pantalla de la identificación de las regiones de unión de AGO2 a dos circRNA diferencialmente expresados en adenocarcinoma de pulmón (CDR1, arriba y RPS24, abajo) mediante diferentes experimentos de CLIP-seq volcados en la base de datos CLIPdb (http://lulab.life.tsinghua.edu.cn/clipdb/).

Por último, para la búsqueda bibliográfica de la relación entre los diferentes miRNA y RBPs asociados a los circRNA identificados, se utilizó el buscador PubMed (https://www.ncbi.nlm.nih.gov/pubmed/).

Un resumen de los pasos seguidos en todo el segundo apartado se plasman en la figura resumen siguiente:

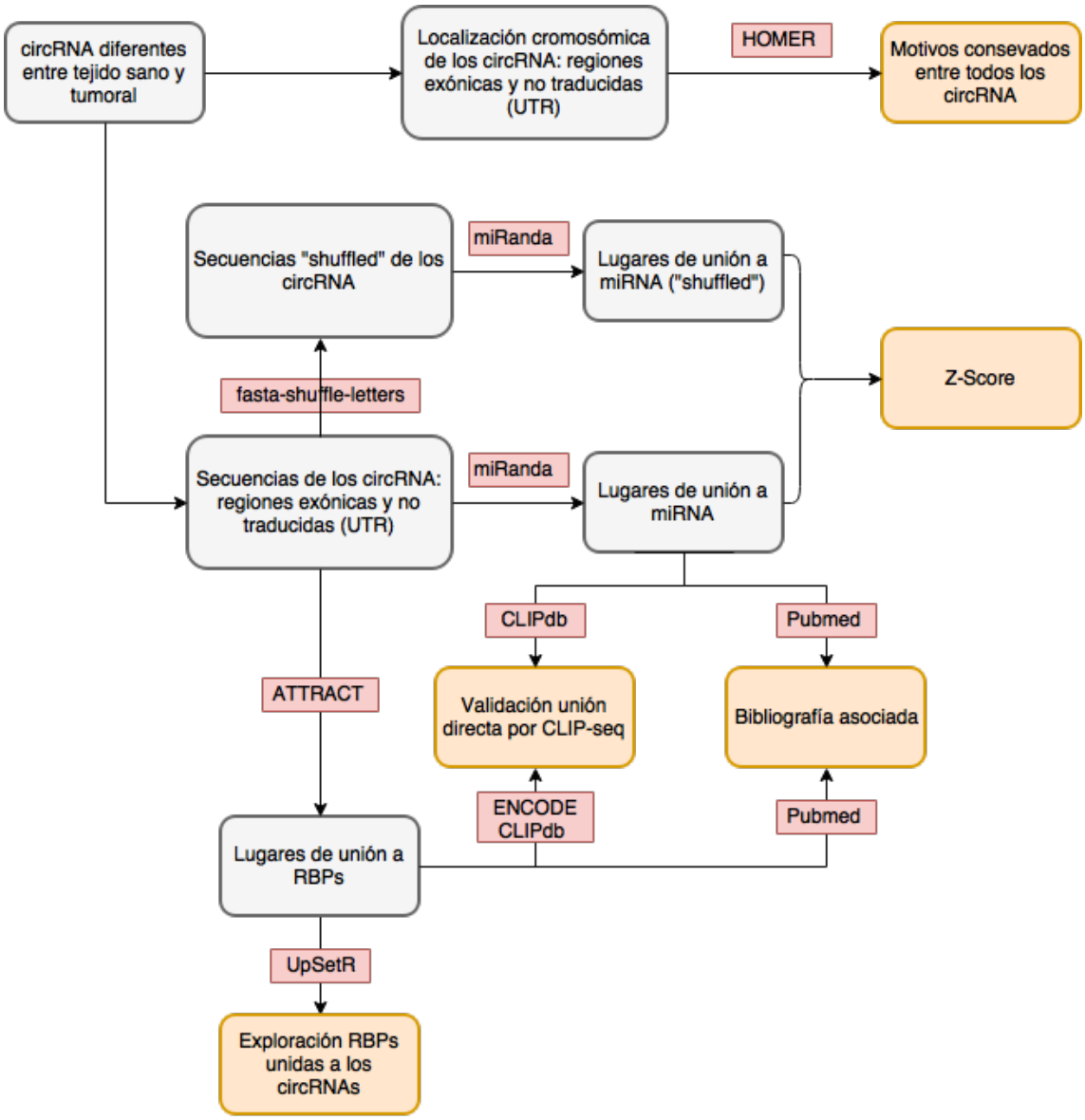


Figura 11. Diagrama de flujo de los pasos llevados a cabo en la segunda fase del proyecto

4. Resultados y Discusión

La producción inadecuada o incorrecta de proteínas es, finalmente, la principal causa de todas las enfermedades humanas. De la transferencia general de información descrita hace años (Crick, 1958) se hace evidente que cualquier error en la transferencia de información, ya sea a los factores estructurales (DNA, RNA o proteína) o sus componentes asociados, puede causar un amplio rango de desequilibrios en el comportamiento normal de la célula. La incapacidad para corregir estos desequilibrios pueden causar un mal funcionamiento y la enfermedad.

Las células cancerosas son uno de los ejemplos más letales de este mal funcionamiento. Aunque se presentó por primera vez como una masa homogénea aislada de células proliferativas incontroladas, hoy en día es bien sabido que las células malignas tienen una compleja red de señalización que las rodea (Hanahan y Weinberg, 2011). En este contexto, los "Hallmarks" distintivos de cáncer se resuelven en varias vías desreguladas que son responsables de la capacidad de este comportamiento incontrolado.

En este TFM se intenta discernir el valor pronóstico (y, en un futuro, terapéutico) de los RNA circulares (circRNA) en uno de los tipos de cáncer más letales de la actualidad: el adenocarcinoma de pulmón.

4.1. Presencia de circRNA diferencialmente expresados en adenocarcinoma de pulmón

Para explorar, primeramente, la capacidad de predecir en evento tumoral en muestras humanas, en este TFM se decidió realizar un análisis de circRNA mediante MapSplice, utilizando muestras de RNA-seq de tejido pulmonar humano normal (15 pacientes) y tumoral (15 pacientes).

Una vez obtenidos los circRNAs expresados en cada muestra, se utilizó un algoritmo de selección *Random Forest* para identificar los circRNA que diferenciaban mejor el tejido pulmonar sano del tumoral (Figura 12).

Región identificada	Gen asociado	Importancia por Random Forest	↑↓	Enlace circBase	
chr8.22161796.22162638	SFTPC	0.96249521	Down	No	
chr8.22162628.22163139	SFTPC	0.93850772	Down	No	
chr10.78037241.78040667	RPS24	0.62343293	Up	hsa_circ_0018961	
chr1.62161618.62162663	INADL	0.46044129	Down	No	
chr2.215364898.215365545	FN1	0.45976596	Down	Contienen la región:	hsa_circ_0058080 a hsa_circ_0058084
chr12.91146236.91153152	DCN	0.42047922	Down	Contienen la región:	hsa_circ_0027711
chr2.188996124.188996468	COL3A1	0.30464720	Up	Contienen la región:	hsa_circ_0057342 a hsa_circ_0057370
chrX.140783175.140784659	CDR1	0.24943918	Up	hsa_circ_0001946	

Figura 12. circRNAs diferencialmente expresados en adenocarcinoma de pulmón comparados con tejido sano. Identificación por Random Forest de los circRNA en adenocarcinoma de pulmón, por orden: región identificada, nombre del gen asociado a la

región, importancia determinada por Random Forest, expresión del circRNA comparada con tejido normal y enlace (si existe) a la base de datos circBase (http://www.circbase.org).

Estos resultados muestran que existen diferentes especies de circRNA que se encuentran expresadas de forma diferente en tejido sano y tumoral. Algunos de ellos (o alguna parte de la región identificada) han sido descritos en estudios anteriores volcados en la base de datos circBase (Glažar et al., 2014). El hecho de que algunos circRNA estén identificados de forma parcial respecto a circBase puede ser debido a la baja cantidad de *reads* asociados a circRNA en las muestras analizadas. Por otro lado, la no identificación de otras especies de circRNA puede ser debido a la especificidad de tejido comentada anteriormente (Meng et al., 2017). Esto podría provocar que algunos de estos circRNA no estén identificados todavía, debido a la falta de análisis de circRNA en tejido pulmonar en la base de datos circBase. En este sentido, los circRNA asociados al gen SFTPC secundan esta idea, ya que este gen se encuentra expresado de forma específica en tejido pulmonar (Figura 13).

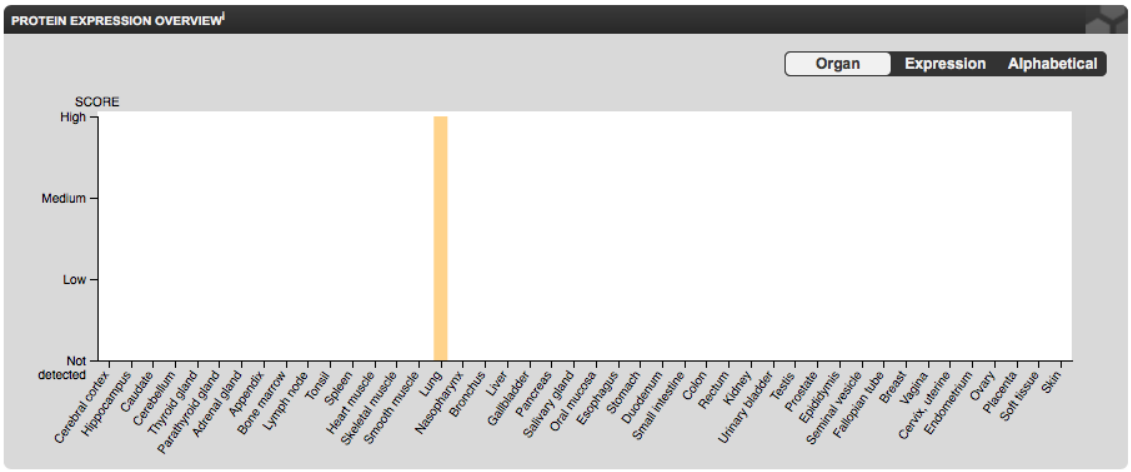


Figura 13. Captura de pantalla de la expresión de proteína del gen SFTPC obtenida de Human Proteome Atlas (https://www.proteinatlas.org/ENSG00000168484-SFTPC/tissue)

4.2. Análisis de los motivos consenso en los circRNAs detectados

La identificación de diferentes especies de circRNA en adenocarcinoma de pulmón podría indicar algún rol biológico de estas secuencias en la patología del tumor. Con el fin de analizar este posible mecanismo, se decidió analizar, primeramente, la existencia de motivos consenso en los circRNAs identificados mediante HOMER (Wang et al., 2010). Este análisis únicamente pretendía discernir la hipótesis de que todos (o la gran mayoría) los circRNAs identificados pudieran tener una vía de acción consenso (Figura 14).

De entre los motivos identificados, únicamente la primera entrada corresponde a un dominio consenso entre algunos circRNAs identificados. Este motivo aparece en los circRNAs asociados a SFTPC y a FN1.

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1		1e-12	-2.774e+01	27.27%	0.00%	83.9bp (0.0bp)	hsa-miR-1228 MIMAT0005583 Homo sapiens miR-1228 Targets (miRBase)(0.703) More Information | Similar Motifs Found	motif file (matrix)
2 *		1e-10	-2.383e+01	54.55%	0.68%	56.6bp (78.0bp)	hsa-miR-4277 MIMAT0016908 Homo sapiens miR-4277 Targets (miRBase)(0.555) More Information | Similar Motifs Found	motif file (matrix)
3 *		1e-9	-2.112e+01	36.36%	0.12%	60.1bp (71.4bp)	hsa-miR-34a* MIMAT0004557 Homo sapiens miR-34a* Targets (miRBase)(0.745) More Information | Similar Motifs Found	motif file (matrix)
4 *		1e-8	-2.001e+01	45.45%	0.54%	33.2bp (80.7bp)	hsa-miR-1206 MIMAT0005870 Homo sapiens miR-1206 Targets (miRBase)(0.683) More Information | Similar Motifs Found	motif file (matrix)
5 *		1e-8	-1.982e+01	27.27%	0.03%	1.4bp (76.4bp)	hsa-miR-3142 MIMAT0015011 Homo sapiens miR-3142 Targets (miRBase)(0.674) More Information | Similar Motifs Found	motif file (matrix)

Figura 14. Motivos identificados por el programa HOMER. Se identifican, por orden, el motivo identificado, el P-valor y su transformación logarítmica, el porcentaje de secuencias con el motivo y su comparación con el background. El asterisco rojo indica que los motivos identificados podrían ser falsos positivos.

4.3. Análisis y validación de los sitios de unión a miRNA en los circRNAs detectados

La capacidad más importante de los circRNA descrita hasta la fecha consiste en la función de "esponja" de diferentes tipos de miRNAs (Figura 6 – Kristensen et al., 2017). Para determinar si los circRNAs detectados podrían tener una función tamponadora de algunos miRNAs se efectuó un análisis de sus secuencias mediante miRanda (Enright et al., 2004). Una vez detectadas las uniones de miRNAs en las secuencias de circRNAs y, antes de analizar la función de algunos de los miRNAs detectados en adenocarcinoma de pulmón, se realizó una primera aproximación para sugerir que, efectivamente, las secuencias de circRNAs identificadas podrían tener un rol biológico de "esponja". Para ello, se generaron versiones "shuffled" de las secuencias de circRNAs detectadas y se detectaron el número de lugares de unión en estas secuencias donde, teóricamente, no existe ninguna presión evolutiva para conservar esos lugares de unión a miRNA. Seguidamente, se calculó un Z-score en representación del incremento de secuencias de unión a miRNA de los circRNAs candidatos respecto a la secuencias "shuffled" (Figura 15).

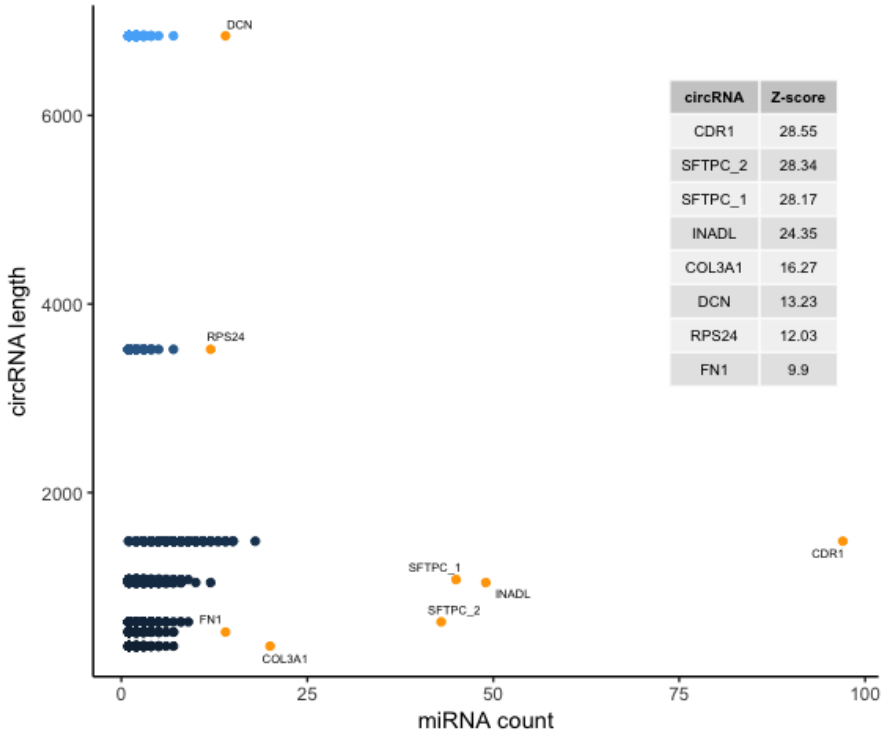


Figura 15. Diagrama de puntos del número de los sitios de unión a miRNA identificados en cada circRNA (puntos amarillos) comparados con las versiones "shuffled" generadas mediante "fasta-shuffle-letters" (puntos azules). El valor de Z-score ((C - mS)/sdS) se encuentra especificado en la tabla superior derecha.

Por los resultados obtenidos, se hace evidente que existe una función biológica relacionada con los lugares de unión a miRNAs detectados en los circRNAs candidatos. Para validar que estos secuencias de reconocimiento por miRNA son, efectivamente, funcionales, se decidió explorar si existía unión directa entre estos exones y la maquinaria de silenciamiento por miRNA. En este caso, se buscaron datos de CLIP-seq de la proteína AGO2. En este sentido, únicamente se encontraron uniones directas a 2 de los 8 circRNAs estudiados (CDR1 y RPS24 – Figura 10). Sin embargo, la no aparición de unión directa de AGO2 a estas secuencias no implica directamente que estos circRNA no puedan ejercer esa función esponja. Este hecho también puede ser debido a que el mRNA en cuestión no se encuentra expresado en las células donde se ha realizado el experimento. Un ejemplo claro (y ya comentado) es el caso del gen SFPTC.

Una búsqueda exhaustiva de los dos candidatos validados (CDR1 y RPS24) nos llevó a resultados interesantes. De entre la literatura estudiada, cabe destacar el rol de miR-7, un miRNA que se ha descrito como promotor tumoral en adenocarcinoma de pulmón (Zhao et al. 2015) y que ha sido fuertemente relacionado con CDR1 o ciRS-7, seguramente el circRNA mejor descrito hasta la fecha y con una funcionalidad "esponja" contrastada (Hansen et al. 2013). En el análisis efectuado por miRanda se identificaron 29 motivos de unión únicos de miR-7 a la secuencia del circRNA de CDR1.

Por otro lado, también se encontraron lugares de unión a miRNA en la secuencia del circRNA asociado a RPS24. En este sentido, el miR-27, identificado el la secuencia del circRNA de RPS24 ha sido descrito como un

regulador de la maquinaria de empalmamiento alternativo que reduce el crecimiento tumoral (Jiang et al., 2014).

4.4. Análisis y validación de los sitios de unión a RBPs en los circRNAs detectados

Siguiendo la hipótesis de que los circRNA pueden tamponar la función de otras proteínas reguladoras, en este TFM se propuso también la exploración de esta función mediante la unión de RBPs a los circRNAs identificados. Esta hipótesis también había sido planteada con anterioridad en la literatura con la proteína MBL y su circRNA asociado (Ashwal-Fluss et at., 2014). Para identificar los diferentes lugares de unión a RBPs en los circRNA candidatos se utilizó el software ATTRACT (Giudice et al. 2016). Para comprobar la distribución de las RBPs obtenidas, se decidió hacer una comparativa entre las diferentes RBPs obtenidas para cada circRNA y identificar aquellas pertenecientes a la maquinaria de empalmamiento alternativo de la que MBL forma parte (Figura 16). En este análisis también se tiene en cuenta si el circRNA identificado presenta regiones no traducidas (UTRs) ya que estas presentan un incremento de unión de RBPs no involucradas en empalmamiento alternativo (Szostak y Gebauer, 2013).

Observando la gráfica, se puede observar como los circRNA que contienen UTRs (RPS24, CDR1, DCN, INADL, SFTPC_1) presentan un porcentaje mayor de RBPs que no están asociadas con el empalmamiento alternativo. Sin embargo, las que presentan únicamente zonas exónicas (SFTPC_2, FN1), presentan unos porcentajes invertidos.

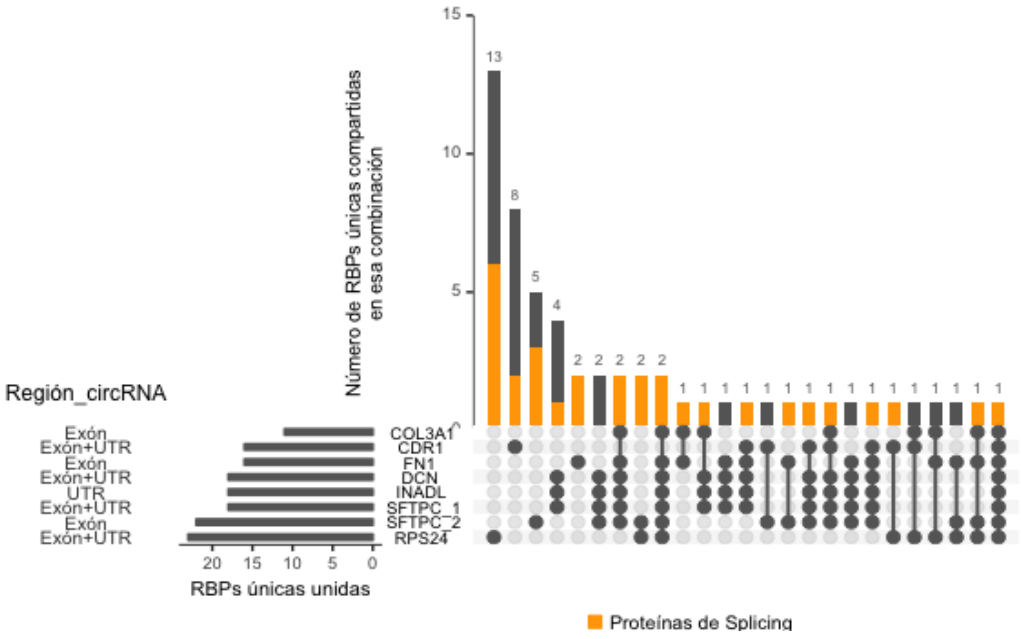


Figura 16. Gráfico de intersección entre los diferentes dominios de unión a RBPs de los distintos circRNAs. De izquierda a derecha, se describe la región comprendida del circRNA, el número de dominios de unión únicos detectados en cada circRNA, el nombre del circRNA y la interacción y cuantificación entre los distintos circRNA única de cada combinación.

Para validar si estas RBPs se encuentran, efectivamente, unidas a las secuencias de los circRNAs, se han usado los datos de CLIP-seq de dos repositorios distintos. El primero, se han buscado los datos de CLIP-seq de las RBPs detectadas en el proyecto ENCODE (https://www.encodeproject.org) donde hay presentes un elevado número de CLIP-seq en dos líneas celulares (línea de carcinoma hepático HepG2 y línea celular de leucemia mieloide crónica K562). En este caso, donde los datos estaban generados en dos líneas celulares concretas, los niveles de expresión del ARN asociado a cada circRNA mediante RNA-seq han sido también añadidos usando los datos del Human Protein Atlas (https://www.proteinatlas.org) (Tabla 1 - Anexo). En el segundo, se han buscado los datos de CLIP-seq de las RBPs detectadas en la base de datos CLIPdb (http://lulab.life.tsinghua.edu.cn/clipdb/) (Tabla 2 - Anexo). Como en el apartado anterior, el hecho de la no aparición de unión por parte de la RBP al circRNA en los datos de CLIP-seq no implica que no exista la unión ya que, como se puede observar en la expresión de RNA en la líneas usadas en el repositorio ENCODE, hay algunos RNAs que no se encuentran expresados.

Por último, se ha buscado en diversos artículos la importancia de algunas de las RBPs detectadas en la progresión tumoral. En este sentido, está ampliamente reportado el rol del empalmamiento alternativo en la progresión tumoral, donde diferentes isoformas de un mismo gen pueden provocar fenotipos totalmente contrarios. En este sentido, proteínas implicadas en el splicing alternativo como SRSF1 y SRSF2, entre otras (Gout et al., 2012 y Jiang et al., 2016) han sido descritas como importantes factores en la progresión tumoral del cáncer de pulmón e incluso estando, algunas de ellas, desplazadas de sus funciones habituales por la formación de uniones ARN:proteína (Ji et al. 2014).

3. Conclusiones

Durante el transcurso de este TFM se pretendía analizar la presencia de especies de circRNAs diferencialmente expresadas en adenocarcinoma de pulmón respecto al tejido tumoral e inferir el rol biológico de los circRNAs mediante el análisis de la unión de miRNAs y RBPs.

Repasando los resultados expuestos en esta memoria, se puede concluir que los objetivos planteados han sido llevados a cabo en este trabajo. Además, se ha podido abordar diferentes tópicos desarrollados durante el transcurso del máster de Bioinformática y Bioestadística de la UOC (Análisis de datos ómicos, análisis de bases de datos, machine learning, etc). También se han podido dar respuestas bioinformáticas a preguntas estadísticas, y se ha podido manipular datos *high-throughput* mediante diferentes ambientes de programación (R, perl y bash).

El seguimiento de la planificación ha sido adecuado aunque se han presentado algunos problemas durante el transcurso del TFM. En este sentido, se han tenido que introducir algunos cambios respecto a la base de datos (TCGA vs GSE40419), de datos para el análisis (RNA-seq vs CLIP-seq) o de analizadores (MEME-suite vs ATTRACT). Estos cambios han sido justificados por diversos motivos como la disponibilidad de los datos o la capacidad para obtener resultados conclusivos. También se han podido completar algunos apartados con nueva información para dar más robustez al trabajo.

Con todos estos resultados obtenidos, se puede concluir que, en el trascurso de este TFM, se han podido identificar 8 secuencias de circRNAs que son capaces de discernir entre tejido pulmonar sano y tumoral y pudiendo, por tanto, clasificarlos como biomarcadores diagnóstico. Además, se ha podido hipotetizar parte del mecanismo de acción de los mismos mediante el análisis de sus regiones de reconocimiento para miRNA y RBPs y su validación mediante datos de CLIP-seq.

En el caso de plantear experimentos futuros relacionados con los datos obtenidos, y centrándonos en su capacidad “esponja” para miRNAs, se podría sugerir la exploración de datos de proteómica e discernir las vías de señalización que los circRNA regulan cuando evitan la capacidad de silenciamiento de sus miRNA diana. Otro punto interesante podría ir relacionado con la capacidad de estos circRNA de regular la maquinaria de empalmamiento, un atributo interesante ya que esta maquinaria es la encargada de generar la versión lineal o circular de un RNA dado (Ashwal-Fluss, et al., 2014) generando así un bucle de retroalimentación circRNA/maquinaria de empalmamiento.

4. Glosario

circRNA: RNA circular

miRNA: micro RNA

RBP: RNA binding protein

GEO: Gene Expression Onmybunus

Biomarcador: Molécula que sirve como indicador de una situación o condición

RNA-seq: Secuenciación masiva de las secuencias de RNA de una muestra

5. Bibliografía

Ashwal-Fluss R, Meyer M, Pamudurti NR, Ivanov A, Bartok O, Hanan M, Evantal N, Memczak S, Rajewsky N, Kadener S. circRNA biogenesis competes with pre-mRNA splicing. Mol Cell. 2014 Oct 2;56(1):55-66.

Bellman RE RandCorporation. Dynamic Programming. Princeton: Princeton University Press; 1957. p. 342.

Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, Davey NE, Humphreys DT, Preiss T, Steinmetz LM, Krijgsveld J, Hentze MW. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell. 2012;149(6):1393-406.

Chabot B, Shkreta L. Defective control of pre–messenger RNA splicing in human disease. The Journal of Cell Biology. 2016;212(1):13-27.

Chen I, Chen C, Chuang T. Biogenesis, identification, and function of exonic circular RNAs . Wiley Interdisciplinary Reviews RNA. 2015;6(5):563-579.

Chen LL. The biogenesis and emerging roles of circular RNAs. Nat Rev Mol Cell Biol. 2016 Apr;17(4):205-11.

Crick FH. On protein synthesis.Symp Soc Exp Biol. 1958;12:138-63.

Decker CJ, Parker R. P-Bodies and Stress Granules: Possible Roles in the Control of Translation and mRNA Degradation. Cold Spring Harbor Perspectives in Biology. 2012;4(9):a012286.

Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. MicroRNA targets in Drosophila. Genome Biology. 2004;5(1):R1.

Filippenkov IB, Sudarkina OY, Limborska SA, Dergunova LV. Circular RNA of the human sphingomyelin synthase 1 gene: Multiple splice variants, evolutionary conservatism and expression in different tissues. RNA Biology. 2015;12(9):1030-1042. doi:10.1080/15476286.2015.1076611.

Floris G, Zhang L, Follesa P, Sun T. Regulatory Role of Circular RNAs and Neurological Disorders. Mol Neurobiol. 2017 Sep;54(7):5156-5165. doi: 10.1007/s12035-016-0055-4.

Gill G. Regulation of the initiation of eukaryotic transcription. Essays Biochem. 2001;37:33-43.

Giudice G, Sánchez-Cabo F, Torroja C, Lara-Pezzi E. ATtRACT—a database of RNA-binding proteins and associated motifs. Database: The Journal of Biological Databases and Curation. 2016;2016:baw035. doi:10.1093/database/baw035.

Glažar P, Papavasileiou P, Rajewsky N. circBase: a database for circular RNAs. RNA. 2014 Sep 18.

Gout S, Brambilla E, Boudria A, et al. Abnormal Expression of the Pre-mRNA Splicing Regulators SRSF1, SRSF2, SRPK1 and SRPK2 in Non Small Cell Lung Carcinoma. Medeiros R, ed. PLoS ONE. 2012;7(10):e46539. doi:10.1371/journal.pone.0046539.

Guo JU, Agarwal V, Guo H, Bartel DP. Expanded identification and characterization of mammalian circular RNAs. Genome Biology. 2014;15(7):409. doi:10.1186/s13059-014-0409-z.

Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011 Mar 4;144(5):646-74.

Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, Damgaard CK, Kjems J. Natural RNA circles function as efficient microRNA sponges. Nature. 2013 Mar 21;495(7441):384-8.

Heinz S, Benner C, Spann N, Bertolino E et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell 2010 May 28;38(4):576-589.

Hinnebusch AG, Lorsch JR. The Mechanism of Eukaryotic Translation Initiation: New Insights and Challenges. Cold Spring Harbor Perspectives in Biology. 2012;4(10):a011544.

Huang S, Yang B, Chen BJ, Bliim N, Ueberham U, Arendt T, Janitz M. The emerging role of circular RNAs in transcriptome regulation. Genomics. 2017 Oct;109(5-6):401-407. doi: 10.1016/j.ygeno.2017.06.005

Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol. 1961;3:318-56.

Jeck WR, Sharpless NE. Detecting and characterizing circular RNAs. Nature biotechnology. 2014;32(5):453-461.

Ji Q, Zhang L, Liu X, et al. Long non-coding RNA MALAT1 promotes tumour growth and metastasis in colorectal cancer through binding to SFPQ and releasing oncogene PTBP2 from SFPQ/PTBP2 complex. British Journal of Cancer. 2014;111(4):736-748. doi:10.1038/bjc.2014.383.

Jiang J, Lv X, Fan L, Huang G, Zhan Y, Wang M, Lu H. MicroRNA-27b suppresses growth and invasion of NSCLC cells by targeting Sp1. Tumour Biol. 2014 Oct;35(10):10019-23. doi: 10.1007/s13277-014-2294-1.

Jiang L, Huang J, Higgs BW, et al. Genomic Landscape Survey Identifies SRSF1 as a Key Oncodriver in Small Cell Lung Cancer. Hammerman P, ed. PLoS Genetics. 2016;12(4):e1005895. doi:10.1371/journal.pgen.1005895.

Kalia M. Biomarkers for personalized oncology: recent advances and future challenges. Metabolism. 2015 Mar;64(3 Suppl 1):S16-21.

Kamel HFM, Al-Amodi HSAB. Exploitation of Gene Expression and Cancer

Biomarkers in Paving the Path to Era of Personalized Medicine. Genomics, Proteomics & Bioinformatics. 2017;15(4):220-235. doi:10.1016/j.gpb.2016.11.005.

Kent WJ, Sugnet Charles W., Furey Terrence S., et al. The Human Genome Browser at UCSC. Genome Research. 2002;12(6):996-1006.

Kristensen LS, Hansen TB, Venø MT, Kjems J. Circular RNAs in cancer: opportunities and challenges in the field. Oncogene. 2017 Oct 9. doi: 10.1038/onc.2017.361.

Lanz, B. Machine Learning with R - Second Edition: Expert techniques for predictive modeling to solve all your data analysis problems. 2015 ISBN-13: 978-1784393908

Lee TI, Young RA. Transcription of eukaryotic protein-coding genes. Annu Rev Genet. 2000;34:77-137.

Lee TI, Young RA. Transcriptional Regulation and its Misregulation in Disease. Cell. 2013;152(6):1237-1251.

Li Y, Dong Y, Huang Z, et al. Computational identifying and characterizing circular RNAs and their associated genes in hepatocellular carcinoma. Coleman WB, ed. PLoS ONE. 2017;12(3):e0174436. doi:10.1371/journal.pone.0174436.

Li Y, Zheng Q, Bao C, et al. Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis. Cell Research. 2015;25(8):981-984.

Li Z, Huang C, Bao C, Chen L, Lin M, Wang X, Zhong G, Yu B, Hu W, Dai L, Zhu P, Chang Z, Wu Q, Zhao Y, Jia Y, Xu P, Liu H, Shan G. Exon-intron circular RNAs regulate transcription in the nucleus. Nat Struct Mol Biol. 2015 Mar;22(3):256-64. doi: 10.1038/nsmb.2959.

Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature. 2005 17;433(7027):769-73.

Lin S, Gregory RI. MicroRNA biogenesis pathways in cancer. Nature reviews Cancer. 2015;15(6):321-333.

Memczak S, Papavasileiou P, Peters O, Rajewsky N. Identification and Characterization of Circular RNAs As a New Class of Putative Biomarkers in Human Blood. Pfeffer S, ed. PLoS ONE. 2015;10(10):e0141214.

Memczak, S. et al. Circular RNAs are a large class of animal RNAs with regulatory potency. Nature 2013;495, 333–338.

Meng S, Zhou H, Feng Z, et al. CircRNA: functions and properties of a novel potential biomarker for cancer. Molecular Cancer. 2017;16:94. doi:10.1186/s12943-017-0663-2.

Orphanides G, Reinberg D. A unified theory of gene expression. Cell. 2002;108(4):439-51.

Papasaikas P, Valcárcel J. The Spliceosome: The Ultimate RNA Chaperone and Sculptor.Trends Biochem Sci. 2016;41(1):33-45.

Pereira B, Billaud M, Almeida R. RNA-Binding Proteins in Cancer: Old Players and New Actors. Trends Cancer. 2017;3(7):506-528.

Richardson K, Lai C-Q, Parnell LD, Lee Y-C, Ordovas JM. A genome-wide survey for SNPs altering microRNA seed sites identifies functional candidates in GWAS. BMC Genomics. 2011;12:504. doi:10.1186/1471-2164-12-504.

Ruggero D. Translational Control in Cancer Etiology. Cold Spring Harbor Perspectives in Biology. 2013;5(2):a012336.

Salzman J, Chen RE, Olsen MN, Wang PL, Brown PO. Cell-Type Specific Features of Circular RNA Expression. Moran JV, ed. PLoS Genetics. 2013;9(9):e1003777.

Salzman J, Gawad C, Wang PL, Lacayo N, Brown PO. Circular RNAs Are the Predominant Transcript Isoform from Hundreds of Human Genes in Diverse Cell Types. Preiss T, ed. PLoS ONE. 2012;7(2):e30733.

Seo J-S, Ju YS, Lee W-C, et al. The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome Research. 2012;22(11):2109-2119. doi:10.1101/gr.145144.112.

Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA Cancer J Clin. 2016 Jan-Feb;66(1):7-30.

Suzuki H, Zuo Y, Wang J, Zhang MQ, Malhotra A, Mayeda A. Characterization of RNase R-digested cellular RNA source that consists of lariat and circular RNAs from pre-mRNA splicing. Nucleic Acids Research. 2006;34(8):e63.

Szabo L, Morey R, Palpant NJ, et al. Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. Genome Biology. 2015;16(1):126.

Szabo L, Salzman J. Detecting circular RNAs: bioinformatic and experimental challenges. Nature reviews Genetics. 2016;17(11):679-692.

Szostak E, Gebauer F. Translational control by 3′-UTR-binding proteins. Briefings in Functional Genomics. 2013;12(1):58-65. doi:10.1093/bfgp/els056.

The ENCODE Project Consortium. An Integrated Encyclopedia of DNA Elements in the Human Genome. Nature. 2012;489(7414):57-74. doi:10.1038/nature11247.

Thunnissen E, van der Oord K, den Bakker M. Prognostic and predictive biomarkers in lung cancer. A review. Virchows Arch. 2014 Mar;464(3):347-58. doi: 10.1007/s00428-014-1535-4.

Timothy L. Bailey, Mikael Bodén, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, William S. Noble, "MEME SUITE: tools for motif discovery and searching", Nucleic Acids Research, 37:W202-W208, 2009.

Touw WG, Bayjanov JR, Overmars L, et al. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? Briefings in Bioinformatics. 2013;14(3):315-326. doi:10.1093/bib/bbs034.

van Hoof A, Wagner EJ. A brief survey of mRNA surveillance. Trends in biochemical sciences. 2011;36(11):585-592.

Wahle E. Poly(A) tail length control is caused by termination of processive synthesis. J Biol Chem. 1995;270(6):2800-8.

Wang H, Wu S, Zhao L, Zhao J, Liu J, Wang Z. Clinical use of microRNAs as potential non-invasive biomarkers for detecting non-small cell lung cancer: a meta-analysis. Respirology. 2015 Jan;20(1):56-65. doi: 10.1111/resp.12444.

Wang K, Singh D, Zeng Z, et al. MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Research. 2010;38(18):e178.

Watson J.D. and Crick F.H.C. A Structure for Deoxyribose Nucleic Acid. Nature. 1953;171:737-738

Wolfe AL, Singh K, Zhong Y, et al. RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer. Nature. 2014;513(7516):65-70.

Xue S, Tian S, Fujii K, Kladwang W, Das R, Barna M. RNA regulons in Hox 5′UTRs confer ribosome specificity to gene regulation. Nature. 2015;517(7532):33-38.

Yang Y-CT, Di C, Hu B, et al. CLIPdb: a CLIP-seq database for protein-RNA interactions. BMC Genomics. 2015;16(1):51.

Ye Z, Jin H, Qian Q. Argonaute 2: A Novel Rising Star in Cancer Research. Journal of Cancer. 2015;6(9):877-882. doi:10.7150/jca.11735.

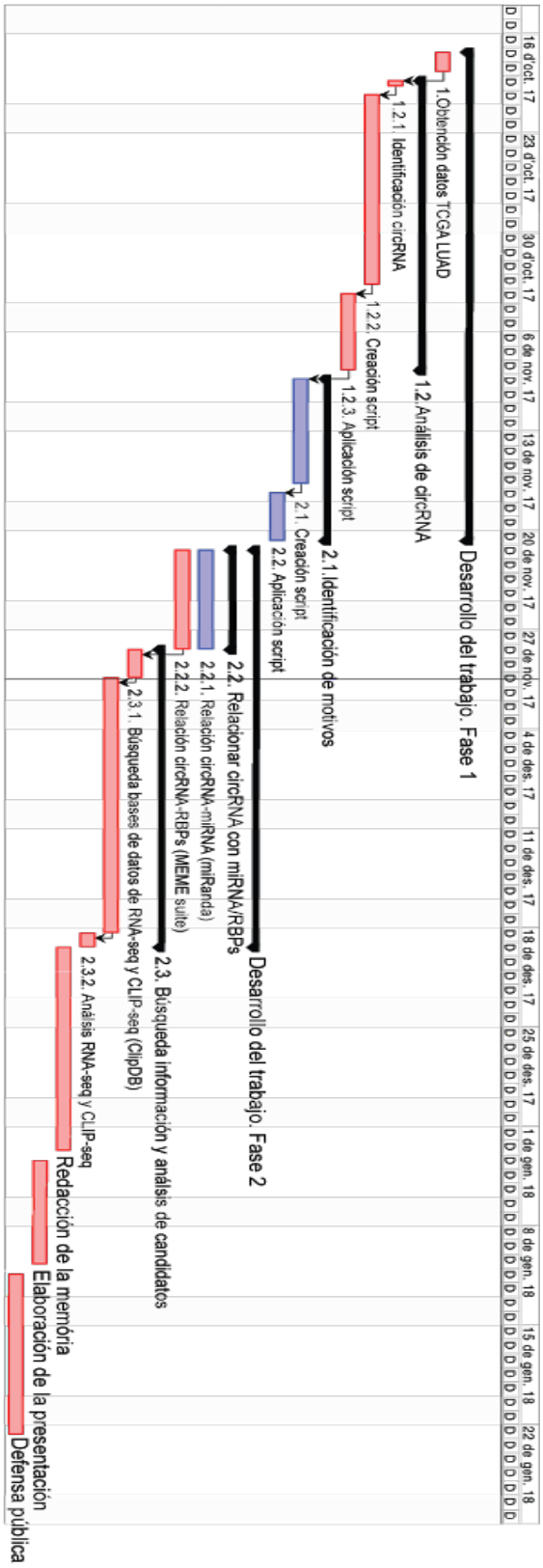
Zhang, Y. et al. Circular intronic long noncoding RNAs. Mol. Cell 2013;51, 792–806 (2013).

Zhao J, Tao Y, Zhou Y, et al. MicroRNA-7: a promising new target in cancer therapy. Cancer Cell International. 2015;15:103. doi:10.1186/s12935-015-0259-0.

Zhu X, Wang X, Wei S, Chen Y, Chen Y, Fan X, Han S, Wu G. hsa_circ_0013958: a circular RNA and potential novel biomarker for lung adenocarcinoma. FEBS J. 2017 Jul;284(14):2170-2182.

6. Anexos

Figura 1



Archivos obtenidos y comandos utilizados

Todos los archivos obtenidos durante el análisis, así como los comandos ejecutados se desglosan a continuación:

circRNA_ML_pipeline.zip: Archivo comprimido con: (1) circRNA detectados en las secuenciaciones de tejido normal (BED_normal.bed), (2) circRNA detectados en las secuenciaciones de tejido tumoral (BED_tumor.bed), (3) agrupación de los circRNA a una distancia menor de 100bp con el valor de coverage (BED_all_format_cov.bed), (4) agrupación de los circRNA a una distancia menor de 100bp con el codigo del paciente (BED_all_format_name.bed), (5) tabla para el uso en el algoritmo de selección (BED_all_FS_cov.txt), (6) reads totales de cada muestra secuenciada (total_reads.csv), (7) coordenadas de las secuencias de los circRNA detectados usados por HOMER (hg19_coord_circRNA.bed) (8) Carpetas con los archivos de los circRNA.txt para cada paciente (normal y cancer) (9) archivo con los comandos usados para los diferentes pasos de obtención, preparación y selección de los circRNA diferencialmente expresados en adenocarcinoma de pulmón y la identificación de sus motivos. También aparecen listados las regiones de circRNA obtenidas después del análisis con *Random Forest.* Por último se identifican las regiones concretas de cada circRNA para el análisis de motivos con HOMER (circRNA_ML_pipeline.Rmd).

FINAL_circRNAs_plus50bp_GENCODE.txt: Secuencias y localización cromosómica de los circRNA detectados por MapSplice en formato FASTA.

miRNA_pipeline.zip: Archivo comprimido con (1) los dominios de unión obtenidos por miRanda de los circRNA detectados por MapSplice (miRanda_FINAL_circRNA_plus50bp_GENCODE_format.txt), (2) los archivos generados para cada circRNA (miRNA_targets_circX.txt, donde X representa cada uno de los 8 circRNA estudiados), (3) los dominios de unión detectados por miRanda para las secuencias "shuffled" generadas por "fasta-shuffle-letters" (miRanda_out_shuffle200_format.txt) y (4) el archivo de R donde se especifica como se han generado los archivos anteriores en el clúster de calculo de la UOC y como se ha realizado el proceso de enriquecimiento de las secuencias de miRNA de los circRNA detectados comparándolos con las secuencias "shuffled" (miRNA_pipeline.R).

ATTRACT_output.zip: Archivo comprimido con (1) las secuencias de unión a cada RBP (ATTRACT_output.tsv), (2) su adaptación a formato Excel (ATTRACT_output.xlsx), (3) los archivos generados para cada circRNA (X.txt, donde X representa cada uno de los 8 circRNA estudiados), (4) el archivo para el formatado inicial de los resultados de ATTRACT y el análisis posterior de los resultados para la comparación de las distintas RBPs detectadas para cada circRNA (RPBs_pipeline.R), (5) los resultados de esta comparativa (Intersections_RBPs.txt), (6) el archivo con la presencia o ausencia de unión de cada RBP a la región de cada circRNA mediante las bases de datos ENCODE y CLIPdb (ATTRACT_vs_CLIP_data.xlsx) y (7) la expresión de las diferentes líneas celulares estudiadas en el Human Protein Atlas (HPA_rna_celline.tsv).