Machine Learning para caracterizar ARNs circulares en exosomas de sangre periférica como biomarcadores

Carmen Gómez Valenzuela
Máster en Bioinformática y Bioestadística
Genómica computacional

Dr. Amadís Pagès Pinós
Dr. Carles Ventura Royo

Junio de 2018

Título del trabajo:	Machine Learning para caracterizar ARNs circulares en exosomas de sangre periférica como biomarcadores
Nombre del autor:	Carmen Gómez Valenzuela
Nombre del consultor:	Dr. Amadís Pagès Pinós
Nombre del PRA:	Dr. Carles Ventura Royo
Fecha de entrega:	06/2018
Titulación:	Máster en Bioinformática y Bioestadística
Área del Trabajo Final:	Genómica computacional
Idioma del trabajo:	Español
Palabras clave:	Machine Learning, circRNA, exosomes

Resumen

Los ARN circulares (circRNAs) han sido identificados recientemente como una clase de isoformas de ARN que forman una molécula cerrada de forma covalente y que se expresan de manera estable, generalizada y abundante en tejidos. Existen evidencias de que pueden actuar como esponjas de micro ARNs, y de su implicación en diferentes tipos de cáncer. Los exosomas son pequeñas vesículas derivadas de las células y de origen endocítico que participan en los procesos celulares actuando como una vía de comunicación entre ellas. Se ha demostrado que las células de un tumor suelen producir más exosomas que las sanas y que pueden detectarse en la sangre periférica humana. Adicionalmente, en la secuenciación de ARN, tanto en sangre completa como en exosomas, se han detectado miles de circRNAs de forma reproducible. En este trabajo se han cuantificado los circRNAs presentes en exosomas de sangre periférica de individuos sanos y con tres tipos de cáncer: colorrectal, hepatocelular y pancreático. Usando la expresión de estos circRNAs se ha conseguido discriminar, con una alta precisión, entre personas sanas y con cáncer, usando técnicas de Machine Learning, reforzando así la hipótesis de que los circRNAs presentes en exosomas de sangre periférica son biomarcadores prometedores que se expresan de forma diferente para distintos tipos de cáncer. Adicionalmente se hace una revisión sobre los micro ARNs sobre los que los circRNAs más expresados podrían estar actuando como esponjas, encontrando evidencias de la implicación de la desregulación de estos micro ARNs en los tres tipos de cáncer estudiados.

Abstract

Covalently closed circular RNA molecules (circRNAs) have recently emerged as a class of RNA isoforms with widespread and tissue-specific expression. circRNAs are remarkably stable and highly expressed molecules. Emerging evidence reveals that they might act as micro RNAs sponges, and also there is evidence about their involvement in different types of cancer. Exosomes are small membrane vesicles of endocytic origin secreted by most cell types and they are thought to play important roles in intercellular communications. It has been shown that tumor cells tend to produce more exosomes than healthy cells and that they can be detected in human peripheral blood. Additionally, in RNA sequencing, in whole blood as well as in exosomes, thousands of circRNAs have been consistently detected. Here we have quantified the circRNAs present in exosomes of peripheral blood of healthy people and with three types of cancer: colorectal, hepatocellular and pancreatic. Using the expression of these circRNAs, we have been able to discriminate, with high precision, between healthy individuals and patients of each cancer group using Machine Learning techniques, reinforcing the hypothesis that circRNAs present in peripheral blood exosomes are very promising biomarkers since they are expressed differently for different types of cancer. Additionally, a review has been made about the micro RNAs on which the most expressed circRNAs could be acting as sponges, finding evidence of the implication of the dysregulation of these micro RNAs in the three types of carcinomas.

Índice general

Lista de figuras

Lista de tablas

Capítulo 1

Introducción

1.1 Contexto y justificación del Trabajo

En este trabajo se ha llevado a cabo la **detección y caracterización**, mediante **técnicas bioinformáticas y de Machine Learning** (ML), de **ARNs circulares** (circRNAs) en exosomas de sangre periférica humana para ser usados como **biomarcadores oncológicos**.

La **sangre periférica** es un fluido orgánico fácilmente accesible mediante técnicas sencillas en la práctica clínica habitual. Encontrar **biomarcadores** en este tipo de **biopsias líquidas** puede permitir un diagnóstico precoz, un marcador de seguimiento de la evolución de la enfermedad una vez diagnosticada o incluso podría darnos información molecular sobre el tumor cuando no se pueden utilizar muestras de tejido sólido.

Los **circRNAs** han sido identificados recientemente como una clase de **isoformas** de ARN que se producen de **forma natural** que pueden regular la expresión génica en los mamíferos [1]. Se trata de un tipo de ARN que forma una **molécula cerrada de manera covalente** y que se expresa de manera generalizada y abundante en tejidos [2]. Además, los circRNAs pueden actuar como **esponjas de micro ARN** (miRNA), esponjas de determinados tipos de proteínas, o como reguladores de la transcripción [3]. La evidencia creciente de la **implicación** de los circRNAs en diferentes tipos de **cáncer**, y la alta **estabilidad** y **abundancia** de estas moléculas, apunta a que podrían usarse como **nuevos biomarcadores** en el diagnóstico oncológico [4].

Los **exosomas** son pequeñas vesículas derivadas de las células de 40 a 100 nm de tamaño y de origen endocítico [5]. Inicialmente se pensaba que eran los vehículos para la **eliminación** de las proteínas celulares innecesarias [6]. Pero recientemente se ha demostrado que su papel no es tan limitado ya que **parecen participar en varios procesos celulares** [7]. Por ejemplo, las células de defensa del sistema inmunitario mandan señales (antígenos) contenidas en exosomas al resto para que reconozcan las amenazas que han detectado y se activen contra ellas: orquestan la respuesta [8]. Incluso en el cerebro, las células de la glía (el soporte del sistema nervioso) son capaces de enviar proteínas al resto de las neuronas para que puedan resistir mejor diversos tipos de estrés [9]. Se trata de una **vía de comunicación entre las células** que, además, en el caso del **cáncer** puede ser especialmente **importante**. Se ha demostrado que **las células de un tumor suelen producir más exosomas** que las células sanas y que pueden **detectarse en la sangre** de los pacientes [10]. Por ejemplo, en el caso del glioma, un tipo de tumor cerebral, las células cancerígenas pueden mandar bolsas con proteínas encargadas de destruir tejido, lo que les permite avanzar y facilita la metástasis [11].

En la **secuenciación de ARN en sangre periférica humana**, tanto en sangre completa [12] como en exosomas [13], **se han detectado miles de circRNAs** de forma reproducible, encontrando que cientos de ellos están mucho más expresados que los correspondientes mRNAs lineales. Como consecuencia la expresión de circRNA en sangre humana revela y cuantifica la actividad de cientos de genes codificantes no accesibles mediante ensayos clásicos de mRNA específico. Estos hallazgos sugieren que los circRNAs podrían usarse como **biomarcadores** en muestras de **sangre** clínica estándar [12].

En la revisión *"CircRNA and Cancer"* [14] leemos que las **altas cantidades de circRNAs** en sangre derivan de su **estabilidad**. En plaquetas y otras fracciones encontramos circRNA intactos, pero solo mRNA y rRNA fragmentados. Debido a esto, se especuló temprano que los circRNAs podrían ser excelentes biomarcadores para la enfermedad. Dada su **estabilidad y especificidad** celular, incluso **pequeñas cantidades** de circRNAs que se originan a partir de células liberadas en el torrente sanguíneo a partir de un tumor localizado **serían detectables** con el tiempo. Sin embargo, este prometedor papel como biomarcadores de cáncer requiere que los circRNAs específicos se expresen de forma diferencial tras la malignización o en etapas tempranas de cánceres específicos. Alternativamente, **la desaparición** de un circRNA muy abundante también **podría utilizarse como un biomarcador** de enfermedad. Otros estudios recientes sugieren que el circRNA **podría tener un papel** en diferentes tipos de cáncer [15] [16] [17].

Por otra parte, e independientemente de su función putativa en el desarrollo y/o progresión del cáncer, los circRNA podrían ser **poderosos biomarcadores** del cáncer debido a su **larga vida media** y su **resistencia** a las vías de degradación comunes [18] [19]. En este sentido, varios estudios ya **han probado** la relevancia de los circRNAs como **biomarcadores de diferentes tipos de cáncer**, como en: *"Using circular RNA as a novel type of biomarker in the screening of gastric cancer"* [19] ó *"Characterization of hsa_circ_0004277 as a new biomarker for acute myeloid leukemia via circular rna profile and bioinformatics analysis"* [20].

Para finalizar, mencionar que existen estudios publicados que utilizan **circRNAs** detectados en **sangre completa periférica** como potenciales **biomarcadores** para la **enfermedad coronaria arterial** [21], la **tuberculosis** pulmonar activa [22] o **diabetes** tipo II [23]. Pero, por el contrario, aunque se han detectado miles de circRNAs en exosomas [13], en el momento de la elaboración del presente trabajo **aún no existen estudios publicados** que utilicen **circRNAs detectados únicamente en exosomas** procedentes de sangre periférica u otro fluido o tejido **para su posible caracterización como biomarcadores** de algún tipo. Es en este punto donde el presente trabajo **ha pretendido realizar alguna aportación** al campo de estudio, concretamente en la **detección y caracterización de circRNAs procedentes de exosomas de sangre periférica** humana para su utilización como **posibles biomarcadores**.

Para **lograr este objetivo**, en primer lugar, se realizó una **búsqueda** en bases de datos públicas **de muestras secuenciadas**, con el fin de obtener muestras secuenciadas de ARN (RNA-Seq) de **exosomas en sangre periférica** humana de individuos sanos y con cáncer, que además fueran **adecuadas para la detección** de circRNAs. Posteriormente se descargaron estas muestras, realizando un **control de calidad** sobre los datos brutos y un **preprocesado** para, a continuación, **alinear las muestras al genoma** de referencia y **usar herramientas** específicas para la **detección de circRNAs**.

Una vez **obtenidos los recuentos** de las lecturas **para cada circRNA** detectado en cada muestra, se utilizaron **técnicas estadísticas** de normalización y filtrado de manera que los datos **fueran adecuados** para ser utilizados **en algoritmos de ML**, ya que este tipo de métodos pueden verse afectados por la asimetría (skewness), la dependencia entre la media y la varianza (heterocedastidad) o la presencia de valores extremos.

Por último, tras implementar el modelo de **clasificación automática** mediante algoritmos de Machine Learning, con el fin de **discriminar** entre muestras de individuos sanos y con cáncer, se llevó a cabo una **selección** de los **circRNAs más relevantes** en cada tipo de cáncer, revisando a continuación en la literatura las **evidencias** disponibles sobre la **implicación** de estos circRNAs en el desarrollo o evolución de la enfermedad, por ejemplo, actuando como **esponjas de miRNAs** cuya desregulación esté implicada en los distintos tipos de cáncer estudiados.

1.2 Objetivos del Trabajo

Los **objetivos generales** del presente trabajo han sido:

1. Diseño y aplicación de un algoritmo que nos permitiera **identificar circRNAs** a partir de muestras secuenciadas de ARN de exosomas de sangre periférica humana y cuantificar su expresión.

2. Construcción de un **modelo de Machine Learning** para la clasificación de las muestras por grupos con el fin de discriminar entre individuos sanos y con cáncer, según los niveles de expresión de los circRNAs detectados y la caracterización de posibles biomarcadores. Es decir, caracterizar los circRNAs más relevantes en la oncogénesis del cáncer o cánceres de interés.

3. **Automatización de los procesos** de detección, expresión y clasificación a partir de los circRNAs detectados en cada muestra mediante el desarrollo una herramienta software.

Por otro lado, también se definieron una serie de **objetivos específicos**, que son:

1. Diseño y automatización del algoritmo que nos permitiera **identificar los circRNAs**.

2. Diseño y automatización del algoritmo para **anotar los circRNAs** detectados generando la tabla de conteos de lectura para cada muestra.

3. Diseño y automatización del algoritmo que, usando técnicas de Machine Learning, nos permitiera **clasificar las muestras en grupos** atendiendo al nivel de expresión de los circRNAs detectados.

4. **Caracterización** del rol del circRNA o **los circRNAs más relevantes en oncogénesis** del cáncer o cánceres de interés.

5. **Redacción de los documentos** intermedios, la memoria del TFM y elaboración de la presentación.

6. **Cumplir los objetivos** relacionados con los **criterios de evaluación** para alcanzar el mayor nivel de logro posible.

1.3 Enfoque y método seguido

Al abordar este trabajo, un **posible enfoque** podría haber sido el **secuencial**. Se trata de un **enfoque tradicional** en el que **se aborda cada objetivo de uno en uno** secuencialmente. En el caso del presente proyecto, en primer lugar, se diseñaría el algoritmo completando cada una de las fases: detección, expresión, clasificación y caracterización de biomarcadores. A continuación, se automatizaría cada una de las fases anteriores para obtener un script ejecutable. Por último, se redactaría la memoria y para finalizar se elaboraría la presentación.

Sin embargo, el **enfoque escogido** para la realización de este proyecto es más cercano a la **metodología ágil** [24]. Desde la primera iteración se ha obtenido un **mínimo producto viable** y el resto **se ha construido** en cada iteración posterior **de forma incremental**. De esta manera **se han minimizado los riesgos**: en cada una de las iteraciones operativas se ha obtenido **un producto entregable** que daba **cobertura** a cada uno de los **objetivos principales**, aunque fuera de forma parcial. Adicionalmente, al **repetir** en cada iteración del proyecto **cada una de las etapas**, incluyendo una etapa de seguimiento intermedio, **ha sido posible detectar errores de forma prematura** y corregirlos en iteraciones posteriores. Esto nos ha aportado una **experiencia adicional** que ha facilitado que fuera posible alcanzar un **mayor nivel de excelencia**.

El **desarrollo del TFM** se dividió en las **fases** detalladas en la Ilustración 1.

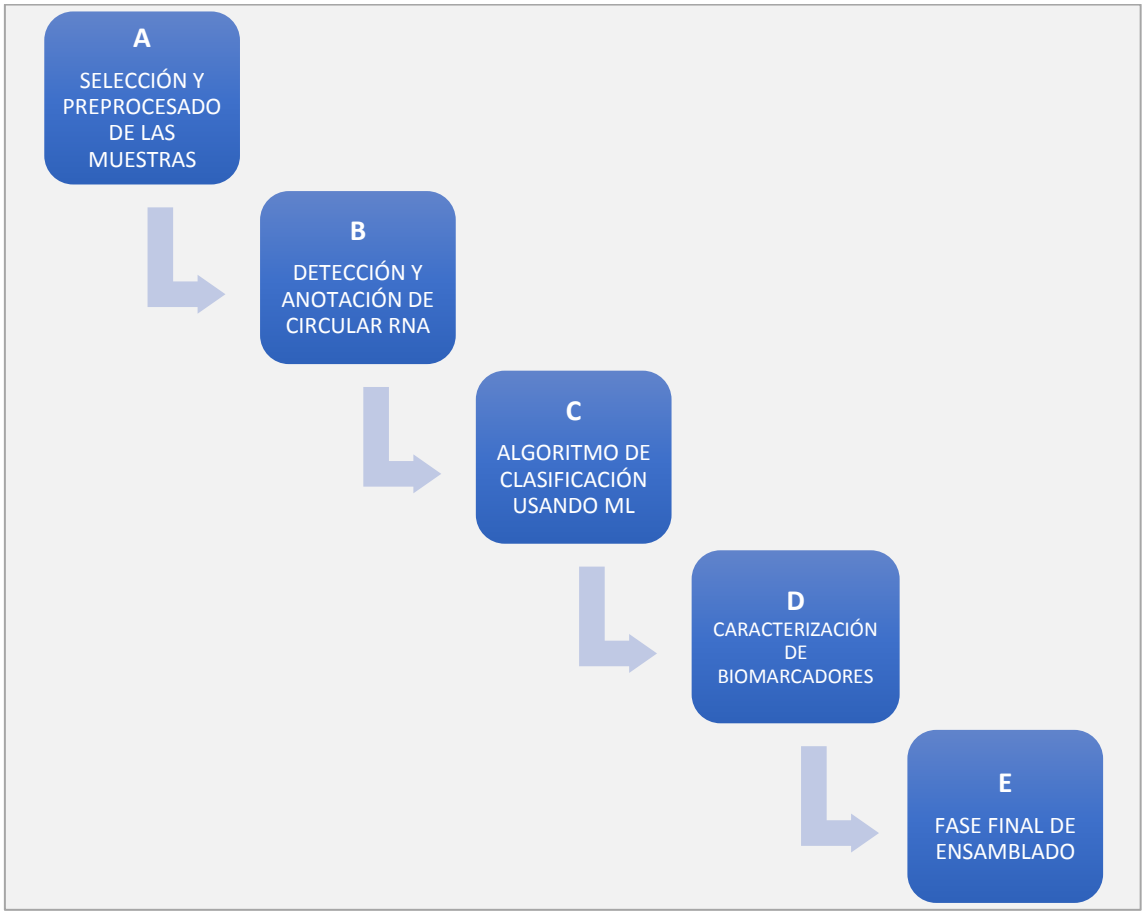

Ilustración 1. Fases del desarrollo del TFM

Los **objetivos generales de cada fase** han sido:

Fase A: Selección y preprocesado de las muestras secuenciadas.

En esta fase se realizó una **búsqueda** de muestras secuenciadas en bases de datos y en la bibliografía para localizar muestras de RNA-Seq en exosomas de sangre periférica humana con una calidad y contenido adecuados al problema que se estaba resolviendo. Asimismo, se realizó un **control de calidad** sobre las muestras y el **pre-procesado** necesario de acuerdo con los últimos estándares y buenas prácticas recomendadas.

Fase B: Detección y anotación de ARN Circular.

Se revisó el **estado del arte** en cuanto a **publicaciones y software específico** para la **detección y anotación de ARN circular** y **se escogió el más adecuado** en función de las **muestras** seleccionadas y a los **recursos** computacionales y de tiempo disponible. A continuación, se **procesaron** las muestras para **obtener** como salida **los circRNAs** detectados en cada una de las muestras secuenciadas.

Fase C: Algoritmo de Clasificación usando Machine Learning.

Se buscaron en la literatura los **algoritmos más prometedores** para el problema que se estaba resolviendo, así como las **mejores técnicas** de **normalización y filtrado** para preparar los datos. En la fase de diseño **se probaron distintos enfoques** y se **implementaron** aquellos que arrojaron los **mejores resultados**. Por último, **se aplicó el algoritmo** seleccionado al conjunto de datos a estudio **para obtener como retorno el grupo de cada muestra** en función del perfil de expresión de sus circRNAs.

Fase D: Caracterización de Biomarcadores

Se caracterizó **el rol** de los circRNAs más relevantes en oncogénesis de los cánceres de interés. Se **validaron en la literatura** y en bases de datos específicas de circRNAs (por ejemplo, en cirBase[25] y CSCD [26]), y se buscaron **posibles dianas** (miRNAs) para los que el circRNA pudiera estar haciendo de esponja.

Fase E: Fase final de ensamblado

Se **agruparon todos los productos** intermedios (documentación y software) con el fin de preparar el producto entregable finalizado.

Y, de forma general, todas las fases han incluido las **etapas** (Ilustración 2):

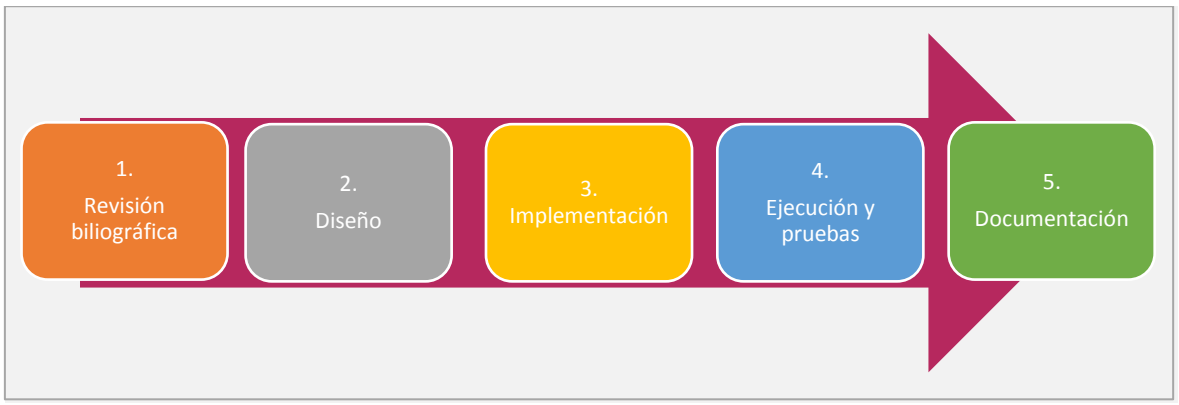

Ilustración 2. Etapas de cada fase del proyecto

1. Revisión bibliográfica. Antes de resolver el problema, era necesario identificar las características relevantes del mismo y el estado del arte de la fase que se estaba resolviendo. Se han utilizado recursos de libre acceso o disponibles a través de la biblioteca de la UOC y otros buscadores, y se han citado todas las fuentes utilizadas. Además, nos hemos asegurado de que el desarrollo de cada fase se realizara de acuerdo con las etapas y actividades previstas en la disciplina.

2. Diseño del algoritmo correspondiente a la fase que se estaba resolviendo. Antes de implementar una tarea se buscó el software disponible priorizando que se tratara de software libre. Se han documentado las dependencias y atribuciones en el manual de usuario y el documento final.

3. Implementación. La automatización de todo el proceso se ha conseguido automatizando cada fase incrementalmente. Se ha procurado que el producto final lograra los objetivos inicialmente definidos intentando alcanzar el mayor nivel de excelencia posible atendiendo a los criterios, normativas y buenas prácticas del área de trabajo.

4. Ejecución y pruebas. Se ha aplicado cada script implementado sobre los datos de trabajo, revisando los informes de salida y aplicando las medidas correctoras necesarias en caso de incidencias. Nos hemos asegurado de que el producto no contiene errores que afecten a su funcionamiento o viabilidad, produciendo el resultado esperado en las situaciones previstas en el trabajo. Asimismo, se ha comprobado que satisface los máximos criterios de calidad, en cuanto a eficiencia, usabilidad, modularidad, extensibilidad, facilidad de comprensión del código y existencia de casos de prueba.

5. Documentación.

5.1. Redacción de la memoria del TFM: en cada iteración se ha documentado el apartado correspondiente, revisando a continuación los criterios de autoevaluación para asegurar que los resultados se presentaban de una forma clara, estructurada y correcta.

5.2. Documentos intermedios de seguimiento y autoevaluación: al final de cada fase se redactó el informe correspondiente de seguimiento.

1.4 Planificación del Trabajo

1.4.1. Tareas

Las **tareas específicas** asociadas a cada fase son las enumeradas a continuación. En el apartado 1.4.3. Calendario se puede consultar su duración y temporización.

Fase A: Selección y preprocesado de las muestras secuenciadas.

A1. Búsqueda bibliográfica y en bases de datos de muestras secuenciadas

A2. Diseño de la estrategia de control de calidad: elección del software más adecuado y diseño de la estrategia de preprocesado para cada grupo

A3. Implementación del script de automatización

A4. Ejecución del script, revisión de los informes de salida y medidas correctoras en caso de incidencias.

A5. Documentación.

Fase B: Detección y anotación de ARN Circular.

B1. Búsqueda bibliográfica para acotar las posibles herramientas software a utilizar y las mejores estrategias.

B2. Diseño de la estrategia de detección: se seleccionarán las herramientas más adecuadas en función de los recursos (hardware y tiempo) disponibles.

B3. Implementación del script de automatización.

B4. Ejecución del script, revisión de los informes de salida y medidas correctoras en caso de incidencias.

B5. Documentación.

Fase C: Algoritmo de Clasificación usando Machine Learning.

C1. Búsqueda bibliográfica para determinar la mejor estrategia y el entorno de programación más adecuado.

C2. Diseño del algoritmo para realizar la clasificación

C3. Implementación

C4. Ejecución del script, revisión de los informes de salida y medidas correctoras en caso de incidencias.

C5. Interpretación de los resultados y documentación.

Fase D: Caracterización de Biomarcadores

D1. Búsqueda bibliográfica y en bases de datos

D2. Documentación

Fase E: Fase final de ensamblado

E1. Búsqueda bibliográfica y en bases de datos

E2. Diseño de la integración software y planificación.

E3. Implementación de la integración (software)

E4. Ejecución (fase de pruebas).

E5. Documentación: redacción de la memoria final agrupando todas las partes, elaboración de la presentación y elaboración del documento final de seguimiento (autoevaluación)

1.4.2. Hitos.

Los **hitos** marcan los **estados intermedios del proyecto** y permiten avanzar en sucesivas etapas de resultados prácticos. Si se hubiera producido un retraso en alguno de ellos, este hecho hubiera tenido una repercusión en el resto de actividades y tareas del proyecto, pudiendo retrasar los plazos marcados.

Se contemplaron **dos clases** de hitos: los **establecidos en el plan docente**, considerados **estrictos** y reflejados en el diagrama de Gantt, y los **establecidos en cada fase específica**, considerados **internos**. Estos últimos admitían cierta flexibilidad durante el desarrollo del proyecto en caso de que se hubieran producido incidencias.

Estos son los **hitos establecidos en el plan docente** (considerados estrictos):

19.03.2018: Entrega PEC1. Plan de trabajo.

23.04.2018: Entrega PEC2. Desarrollo del trabajo – Fase1

21.05.2018: Entrega PEC3. Desarrollo del trabajo – Fase2

06.06.2018: Entrega PEC4. Redacción de la memoria

14.06.2018: Entrega PEC5. Elaboración de la presentación

Y estos otros los **hitos específicos**, considerados internos:

26.03.2018: Finalización de Fase A. Selección de la muestra y preprocesado.

20.04.2018: Finalización de Fase B. Detección de circRNA.

14.05.2018: Finalización de Fase C. Algoritmo de clasificación usando ML.

21.05.2018: Finalización de Fase D. Caracterización de biomarcadores.

05.06.2018: Finalización de Fase E. Integración.

1.4.3. Calendario.

Se realizó un **diagrama de Gantt** (Ilustración 3) utilizando el entorno de gestión de proyectos Teamgantt, disponible online de forma gratuita, **identificando** las **tareas**, su **duración** y las **relaciones de dependencia** entre ellas. Este entorno ha sido utilizado para el seguimiento del proyecto en todas sus fases. El diagrama incluye las tareas esenciales para el proyecto y la metodología seleccionada. La planificación **se ajusta al marco temporal** del proyecto (fecha de inicio y final, calendario de hitos, 300 horas de duración total) y se **marcaron** en el diagrama **los hitos estrictos** detallados en el apartado anterior.

La duración estimada para cada tarea es **coherente** con su **complejidad**, las **características** del problema y las **competencias** del estudiante, y se tuvieron en cuenta el **nivel de conocimiento** y la **experiencia** del estudiante en cada ámbito del trabajo.

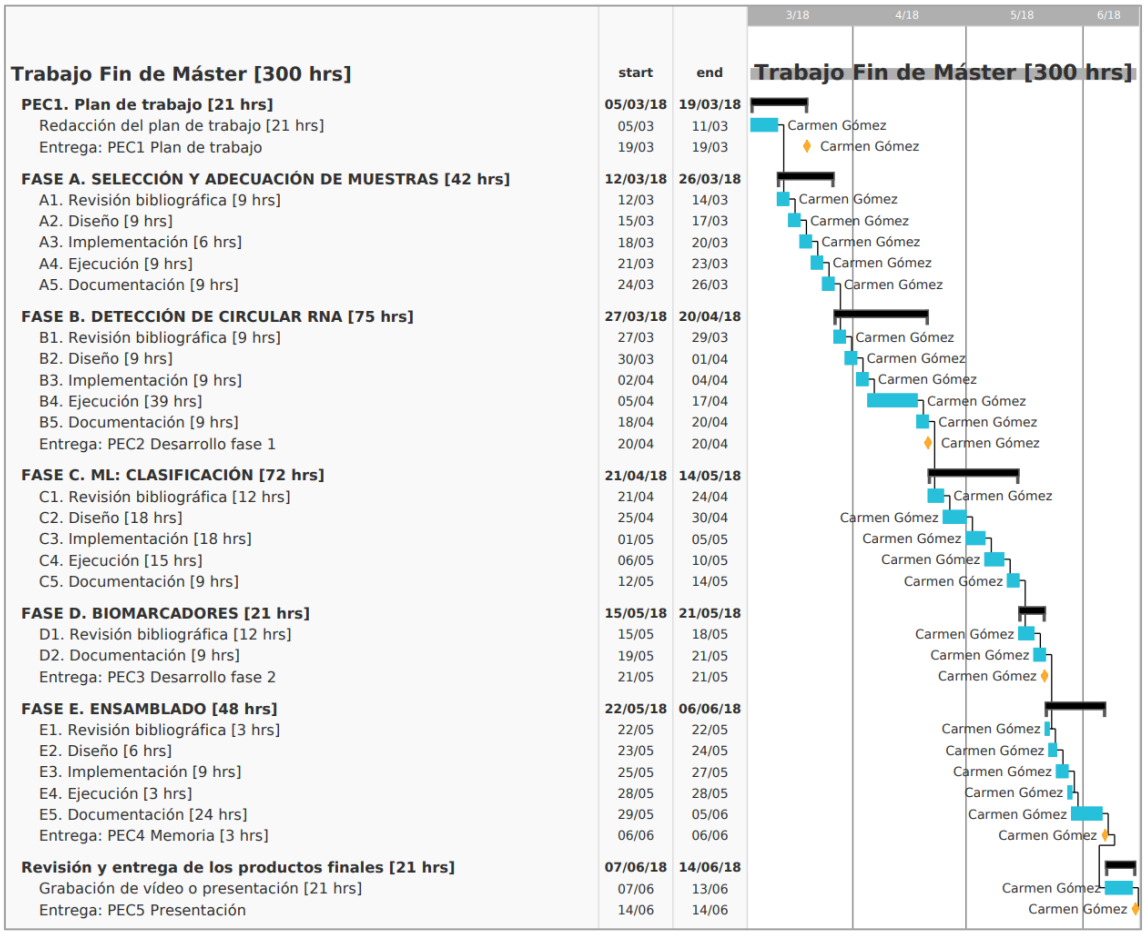

Ilustración 3. Diagrama de Gantt con la planificación del TFM

Con respecto a la disponibilidad de recursos, el único recurso sujeto a disponibilidad era el **clúster de cálculo** habilitado por la UOC, que **ha estado disponible** durante toda la duración del proyecto. Por último, se ha tenido en cuenta la **dedicación** del estudiante, asignando tres horas diarias de trabajo, en principio sin distinción de festivos o vacaciones. Esas fechas especiales han sido utilizadas para recuperar tiempo no dedicado en días laborables, en los que eran más probables las eventualidades.

1.5 Sumario de productos obtenidos

En la **realización** de este trabajo se han **generado** la **memoria** expuesta en este documento junto con un **conjunto de scripts** que, con sus correspondientes manuales de usuario y ejemplos de uso, pueden ser **descargados** del repositorio:

http://github.com/carmengmz/circRNA

Adicionalmente, en ese mismo repositorio, en la carpeta 'experiment', se encuentran todos los **resultados intermedios** obtenidos, que son: los informes de calidad de las librerías de RNA-Seq; la tabla circ_annotations.csv con las anotaciones, coordenadas y recuentos de lectura de los circRNAs detectados cada muestra; y el archivo phenodata.txt, con el identificador y grupo de las librerías de RNA-Seq utilizadas. También se han incluido los informes e implementación de cada uno de los modelos de clasificación.

1.6 Descripción del resto de capítulos de la memoria

El **capítulo 2** está dedicado a la detección de circRNAs en muestras de RNA-Seq de exosomas en sangre periférica humana. Comenzamos describiendo la estructura y biogénesis de exosomas y circRNAs. A continuación, hacemos un recorrido por los pasos más relevantes en la secuenciación del ARN, detallando los requisitos que debe cumplir una muestra secuenciada adecuada para la detección de circRNAs. Por último, se hace una revisión de las herramientas software disponibles para la limpieza, preprocesado y control de calidad de librerías de RNA-Seq; finalizando con una comparativa de las herramientas disponibles para la detección de circRNAs en muestras secuenciadas, argumentando, en cada caso, cuáles han sido los criterios utilizados para escoger las herramientas más adecuadas para el proyecto.

En la **segunda parte del capítulo 2** se detalla cómo se ha realizado la búsqueda, selección y descarga de muestras de RNA-Seq; y la implementación del preprocesado y control de calidad. Para terminar, describimos el proceso de detección y anotación de circRNAs con el fin de obtener la tabla con la identificación y recuentos lectura de los circRNA detectados en cada muestra.

El **capítulo 3** está dedicado a la clasificación de las muestras mediante Machine Learning, con una primera parte en la que se exponen los fundamentos matemáticos de la normalización mediante la transformación estabilizadora de la varianza, la selección de predictores utilizando un algoritmo de Random Forest, y los algoritmos de clasificación utilizados: Support Vector Machines y Neural Networks. Por último, se describe cómo se calcula y reporta la bondad de los modelos.

En la **segunda parte del capítulo 3** explicamos cómo han sido aplicados todos estos métodos con el fin de generar los modelos de Machine Learning para discriminar, de manera automática, entre muestras de RNA-Seq de exosomas de sangre periférica de individuos sanos y pacientes con cáncer colorrectal, hepatocelular y pancreático.

El **capítulo 4** está dedicado a la caracterización de los circRNAs más relevantes detectados en cada grupo, incluyendo una una revisión de las evidencias encontradas en la literatura sobre la posible implicación, en cada tipo de cáncer, de los miRNAs sobre los que podrían estar actúando como esponjas.

Por último, en el **capítulo 5**, se exponen los resultados y conclusiones del proyecto.

Adicionalmente, al final del documento, se ha incluido un **glosario** con la definición de los términos y acrónimos más relevantes utilizados en el este documento, y un apartado con las **referencias** bibliográficas.

Capítulo 2

Detección de ARNs circulares en exosomas de sangre periférica humana

2.1. Conceptos teóricos

2.1.1. Exosomas

La noción de **vesículas extracelulares** surgió por primera vez en 1983, cuando los investigadores en los grupos de Stahl [27] y Johnstone [28] describieron la observación de **endosomas tardíos multivesiculares** que liberaban **vesículas de reticulocitos** en el ambiente extracelular. Estas vesículas fueron posteriormente llamadas ***exosomas*** [29]

Recientemente, el interés en la biología de vesículas extracelulares ha crecido enormemente y **se han descrito diferentes tipos** de vesículas **dependiendo de sus propiedades** biofísicas y su biogénesis. Estas vesículas **se liberan desde la mayoría de las células** y pueden **aislarse fácilmente en los fluidos** corporales tales como suero, plasma, orina y fluido cerebroespinal.

Los **exosomas** son **partículas pequeñas y homogéneas** cuyo tamaño varía de 40 a 100 nm y de **origen endocítico**. En la endocitosis, las vesículas se forman en la membrana plasmática y se fusionan para formar endosomas tempranos. Estos maduran y se convierten en endosomas tardíos cuando las vesículas intraluminales se forman en el lumen intracitoplasmático. En lugar de fusionarse con el lisosoma, estos cuerpos multivesiculares se fusionan directamente con la membrana plasmática y liberan exosomas en el espacio extracelular [30]

Inicialmente se pensaba que su función era la de **excretar los desechos celulares** [29]. Posteriormente se ha sugerido que estas vesículas también **desempeñan un papel en la comunicación intercelular** (Ilustración 4) [31] y han sido asociados con **numerosas funciones fisiológicas y patológicas** [32]. Adicionalmente, los **exosomas de las células cancerosas** han demostrado **promover la angiogénesis**, ser **moduladores del sistema inmune** y **remodelar el tejido parenquimatoso circundante**, es decir, todos los **factores** que **favorecen** la **progresión tumoral** y la **metástasis** [33]. Y en particular se ha demostrado que los exosomas participan en la generación del nicho pre-metastásico [34][35][36].

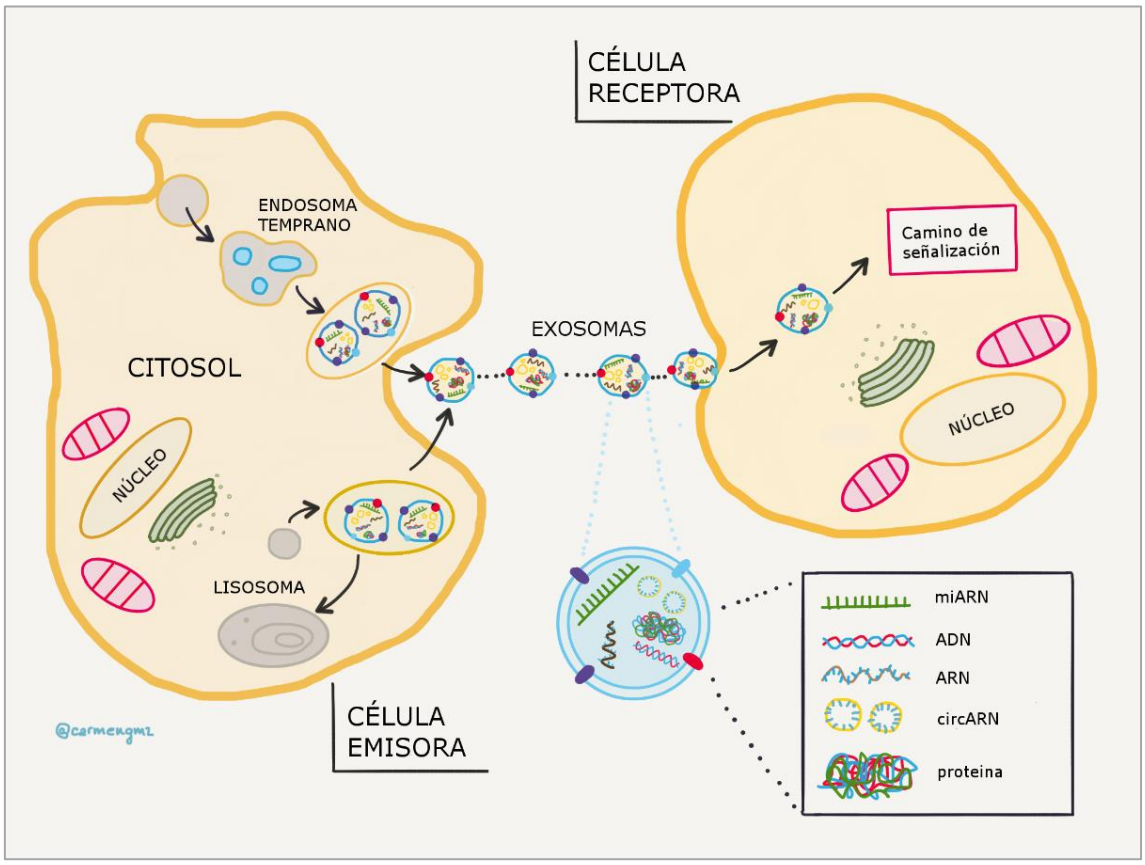

Ilustración 4. Comunicación intercelular, biogénesis y estructura de un exosoma[1]

[1] Todas las ilustraciones del documento son de elaboración propia, a no ser que se indique lo contrario.

2.1.2. ARN Circular

El **ARN circular** (circRNA) es un tipo de ARN que, a diferencia del ARN lineal, forma un **bucle continúo cerrado covalentemente**, es decir, en el ARN circular los extremos 3' y 5' normalmente presentes en una molécula de ARN están unidos. Esta característica les confiere numerosas propiedades.

Historia

Aunque **en un principio** los circRNAs se **categorizaron** como **ARN no codificante**, ha sido demostrado que a partir de ellos es posible la traducción de péptidos y proteínas [37][38][39]. Además, algunos circRNAs **han mostrado** recientemente **potencial** como **reguladores genéticos**. Y al igual que muchas otras isoformas alternativas de ARN, la función biológica de la mayoría de los circRNAs no está clara.

En 1991 **Janice Nigro** menciona en su publicación "Scrambled exons" en la revista Cell que habían encontrado evidencia de la presencia de **ARN que no cumplía las reglas canónicas del splicing**, sugiriendo una posible estructura circular [40].

Pero no fue hasta 2013 que **Norman Sharpless** y su equipo publican en la revista RNA [41] que **en fibroblastos** humanos existen alrededor de **25.000 circRNA diferentes** y que provienen de un 14,4% de los genes que estos expresan. Anteriormente los circRNA fueron considerados como simples errores en la extracción de ARN, pero Sharpless **demostró que están particularmente conservados** en una gran variedad de especies.

Biogénesis

El **dogma central** de la biología molecular describe el **flujo de información** que se almacena en los genes como ADN, se transcribe en ARN y finalmente se traduce en proteínas[42][43]. La **última expresión** de esta información genética modificada por factores ambientales caracteriza el **fenotipo** de un organismo. La **transcripción** de un subconjunto de genes en moléculas de ARN especifica la identidad de una célula y regula las actividades biológicas dentro de la célula. Colectivamente definidos como el **transcriptoma**, estas moléculas de ARN son **esenciales** para interpretar los **elementos funcionales del genoma** y comprender el **desarrollo** del individuo y las enfermedades.

La **transcripción del ADN** (Ilustración 5) es el proceso por el cual el ADN es copiado (transcrito) para convertirse en ARN conteniendo la información necesaria para la síntesis de proteínas. La transcripción se lleva a cabo mediante **dos grandes procesos**: en primer lugar, se forma el **ARN pre-mensajero** (pre-mRNA) con la participación de la encima **ARN-polimerasa** (ARNP). El proceso se basa en el emparejado de bases descrito por **Watson-Crick** y como resultado obtenemos una hebra simple de ARN que es la complementaria inversa de la secuencia de ADN original.

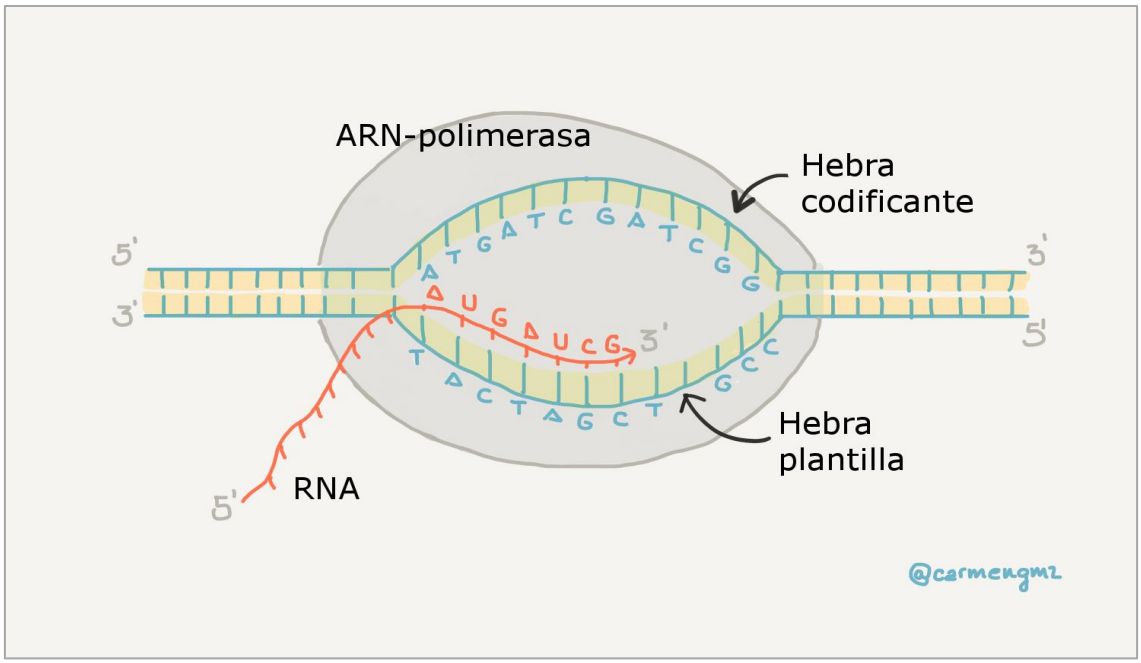

Ilustración 5. Transcripción del ADN

A continuación, **ambos extremos** de pre-mRNA son **modificados**. Al grupo del inicio (extremo 5') se le añade la **caperuza 5'** (5' cap). En este proceso, conocido como capping, se añade un nucleótido de guanina (G) modificado que confiere **estabilidad** al tránscrito y ayuda a que el ribosoma se una al ARN mensajero para leerlo y formar una proteína. Al grupo del final (extremo 3') se le añade una **cola de poli-(A)**. Cuando aparece una secuencia llamada señal de poliadenilación en una molécula de ARN durante la transcripción, una enzima corta el ARN en dos en ese sitio. Otra enzima agrega aproximadamente de 100 a 200 nucleótidos de adenina (A) al extremo cortado, formando una cola poli-(A). La cola hace que la transcripción sea más estable y ayuda a exportar el mRNA desde el núcleo al citosol.

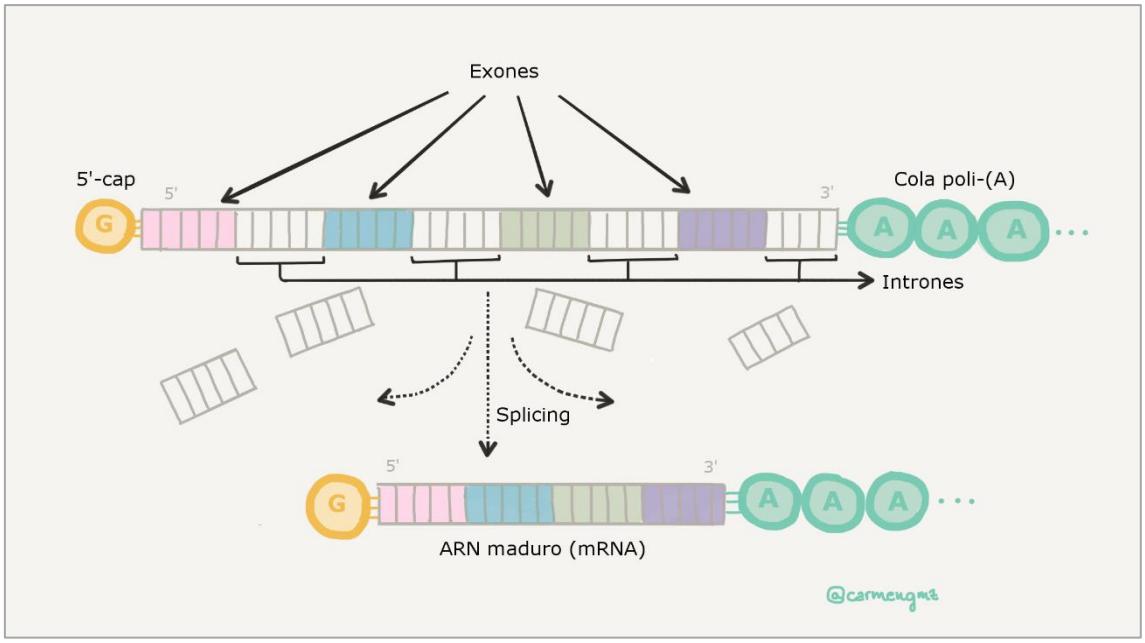

Ilustración 6. Maduración del ARN

El ARN pre-mensajero así formado contiene **intrones** que no son necesarios para la síntesis de proteínas. El ARN pre-mensajero se corta para eliminar los intrones y crear ARN mensajero (mRNA) en un proceso llamado corte y empalme de ARN (**splicing**) (Ilustración 6).

Pero, a diferencia de los mRNAs, en los **circRNAs** se produce **un enlace covalente y canónico** entre un sitio de corte y empalme 3' y un sitio de corte y empalme 5' cadena arriba en un proceso conocido como **backsplicing** (Ilustración 7) [44]. Los circRNA **carecen de colas de poli-(A)** y pueden contener un solo exón o exones múltiples, así como intrones. Debido a que los ARN circulares no tienen extremos 5' o 3', son **resistentes a la degradación** mediada por exonucleasa y son supuestamente **más estables** que la mayoría de los ARN lineales en las células[45].

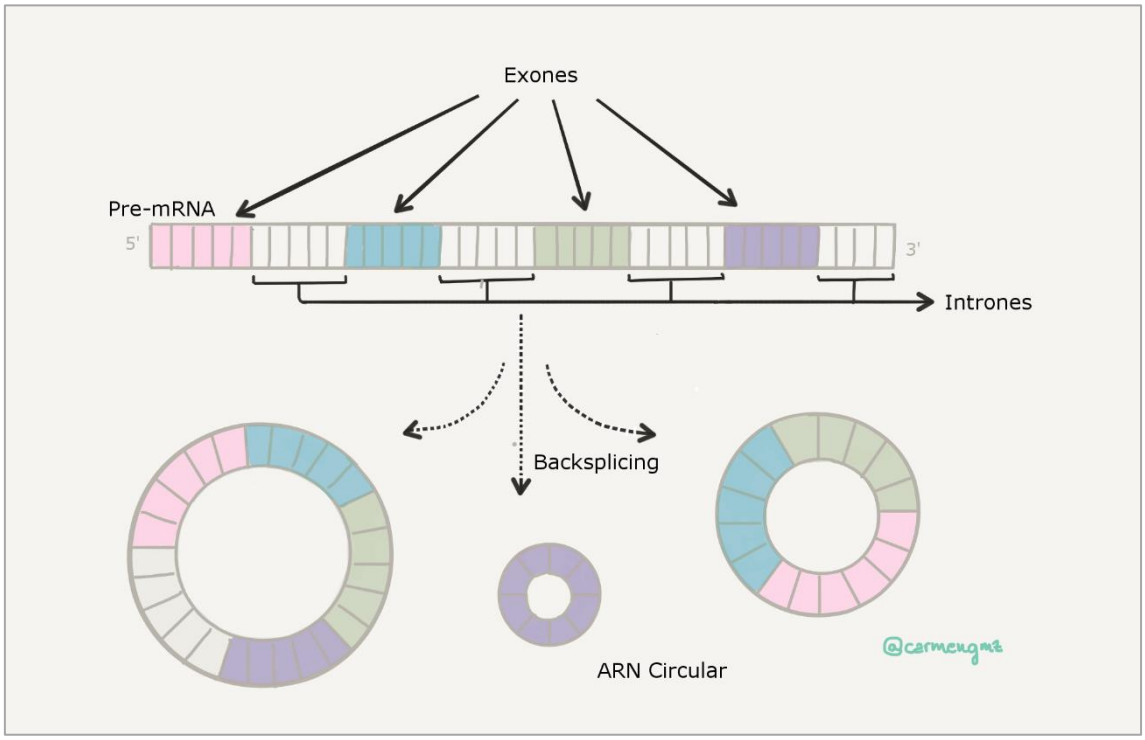

Ilustración 7. Biogénesis de los circRNAs

2.1.3. Detección de circRNA en RNA-Seq.

Las tecnologías de **secuenciación de última generación** (NGS, Next generation sequencing) han sido utilizadas para desarrollar la secuenciación del ARN (RNA-Seq: RNA Sequencing), un enfoque de secuenciación profunda y de alto rendimiento para transcriptomas. Los métodos previos de secuenciación de ARN se basaban en microarrays de alta densidad y eran métodos limitados ya que dependían del conocimiento previo de la secuencia [46].

Para realizar la **secuenciación**, el ARN total o fragmentado se convierte a ADN a través de la **transcripción inversa**. Como **resultado** se obtiene una **librería de ADN**

complementario (cDNA) que es secuenciada usando tecnologías NGS (habitualmente Illumina, SOLiD o Roche 454). Las secuencias resultantes pueden ser reensambladas desde cero o alineadas a una secuencia conocida para crear un mapa de la transcripción.

Hay **muchas maneras** en las que **el ARN es tratado** durante su conversión a **cDNA** y en la **preparación final** de las bibliotecas de secuenciación, pero, en general, el **flujo de trabajo experimental** incluye los siguientes **pasos**: **purificación** del ARN, **transcripción inversa** mediante Reverse Transcriptase (RT) para producir la primera hebra de cDNA, **síntesis de la segunda hebra** usando DNA polimerasa y la **preparación de la librería** para su secuenciación [47]

En esta enumeración se han omitido una gran cantidad de detalles, como, por ejemplo las estrategias de normalización y los procedimientos necesarios para tratar con ARN poco expresados o degradados [48]. Sin embargo, hay **dos consideraciones experimentales** importantes que **afectarían** las formas en que los **datos son analizados** y los **resultados interpretados**, como son, el **cebado** para la primera síntesis de la cadena de cADN, y las bibliotecas de lecturas **emparejadas** frente a las de lecturas de **una sola hebra**.

En primer lugar, la **transcripción inversa** mediante Reverse Transcriptase (RT) **necesita un cebador**. Se puede aprovechar el hecho de que **la mayoría de los ARNs están poliadentados** (tienen cola de poli-(A)) y se puede usar un **cebador oligo-dT** para (principalmente) restringir la síntesis de cADN a los ARNs maduros. En el caso de los **circRNAs**, este procedimiento **los descartaría** ya que, como se vio en el apartado anterior, carecen de cola de poli-(A). Alternativamente se puede usar una **mezcla de oligonucleótidos** (random hexamer priming) para cebar la RT en varios sitios internos independientemente del tipo de ARN y su estado de maduración. En función de esta elección, tendremos distintos tipos de RNAs incluidos en la secuencia final. Si lo que se intenta es estudiar ARN que no están poliadentados (como es nuestro caso) no es adecuado utilizar un cebador oligo-dT.

Por tanto, para **escoger una muestra secuenciada adecuada** para la detección de **circRNAs** debemos tener en cuenta que existen muchas **variaciones** en las preparaciones de **bibliotecas** RNA-Seq, que **afectan significativamente** la abundancia de circRNAs en los conjuntos de datos de RNA-Seq resultantes.

Los **pasos bioquímicos** en la preparación de la **biblioteca** que tienen la **influencia más importante** en la detección de circRNA son, primero, la **purificación** de ARN; segundo, las **selecciones de tamaño** a nivel de ARN o cDNA; y tercero, fragmentación del ARN, el **método de cebado** y/o la ligación del adaptador.

Actualmente las **bibliotecas** de RNA-Seq de RNA celular eucariótico son **típicamente poli(A)-selected o depleted de rRNA** antes de la preparación de la biblioteca. La única purificación que se predice que reducirá significativamente una muestra de circRNAs es una etapa de enriquecimiento de poli-(A), ya que los circRNA carecen de una cola poli-(A). Por el contrario, los circRNAs **se retienen en bibliotecas rRNA-depleted** y se **enriquecen** en bibliotecas **tratadas con RNasa-R** para digerir el ARN lineal. En la selección de tamaño normalmente se excluyen moléculas que están por debajo de 200 nm, esto no sería un gran

problema en la detección de circRNA ya que sólo excluiría aquellos de tamaño muy pequeño en caso de existir.

El **cebado aleatorio**, a diferencia del cebado oligo(dT), no requiere que esté presente una cola de **poli-(A)** y, por lo tanto, dará como resultado una biblioteca de RNA-Seq que no está sesgada contra circRNAs. Finalmente, las **bibliotecas de ARN pequeñas** estarán **sesgadas contra los circRNAs** solo **si el ARN no está fragmentado** antes de la ligadura o cebado del adaptador, ya que los circRNAs no tienen extremos libres a menos que estén cortados in vivo o in vitro [49]

Por último, en [50] se realiza un análisis comparativo para evaluar sistemáticamente el **sesgo de las predicciones de circRNA** a partir de los datos de RNA-Seq. El análisis mostró que era **factible** aprovechar al máximo la gran cantidad de datos RNA-Seq en las **bases de datos públicas** para las predicciones de circRNAs. Y, aunque dicha predicción adolece de falsos positivos, puede controlarse cuidadosamente mediante el uso de datos RNA-Seq con buena calidad y seleccionando preferentemente los circRNAs con alta relación entre lecturas de unión y recuento de lecturas.

2.1.4. Herramientas para pre-procesado de RNA-Seq y control de calidad

La **secuenciación directa del ADN complementario** (cDNA) usando tecnologías de secuenciación de alto rendimiento (RNA-Seq) permite una **comprensión muy precisa** del transcriptoma y sus isoformas. Se trata de un proceso que involucra **múltiples etapas**, como la transcripción inversa, amplificación, fragmentación, purificación, ligadura de adaptadores y secuenciación [47].

En teoría, mediante **RNA-Seq** deberíamos ser capaces de **identificar y cuantificar** con **precisión** todas las **especies de RNA**, pequeñas o grandes, en abundancia baja o alta. Sin embargo, las operaciones inadecuadas en cualquiera de los pasos necesarios pueden generar datos parciales o incluso inutilizables. Además, los **sesgos intrínsecos de RNA-Seq,** como el sesgo de composición de nucleótidos o el sesgo GC, afectan directamente a la precisión de muchas aplicaciones de RNA-Seq[51][52]. Por lo tanto, la **evaluación integral de la calidad** y el **preprocesado de los datos brutos** son los **primeros pasos y más críticos** para todos los análisis posteriores y la correcta interpretación de los resultados [53].

Existen **multitud de herramientas** para el control de calidad y preprocesado. Por ejemplo, **FASTQC** [54] es una herramienta basada en Java que proporciona perfiles de calidad por base y por lectura. **Cutadapt** [55] encuentra y elimina secuencias de adaptadores, cebadores, colas poli-(A) y otros tipos de secuencias no deseadas. **Trimmomatic** [56] es otra herramienta para eliminar adaptadores que además puede filtrar las lecturas en función de su calidad.

Una **combinación típica** es usar FASTQC para control de calidad, Cutadapt para eliminar los adaptadores y Trimmomatic para filtrar las lecturas. Pero la necesidad de **leer y cargar los datos en múltiples ocasiones** convierte el preprocesado en una **tarea lenta e ineficiente** con muchas lecturas y escrituras a disco. **AfterQC** [57] es una herramienta más

completa que puede realizar todas las operaciones necesarias en una sola pasada pero, al estar implementada en Python, es relativamente lenta.

Por todos estos **motivos**, en el presente trabajo **se ha optado** por el uso de la herramienta **fastp** [58] desarrollada en C++ y con suporte multi-hebra. Fastp incorpora **la mayoría de las características** de FASTQC, Cupdatadpt, Trimmomatic y AfterQC, pero **ejecutándose** de 2 a 5 veces **más rápido** que cualquiera de ellas por separado. Realiza el control de calidad, la eliminación de adaptadores, el filtrado de las lecturas y corrección de las bases **en una sola pasada**, soportando además lecturas emparejadas (paired-end reads).

2.1.5. Herramientas para la detección de ARN Circulares

En los últimos años se han desarrollado **múltiples herramientas y pipelines** para la **identificación** de los **circRNAs**. En paralelo podemos encontrar estudios publicados comparando las distintas herramientas de detección, por ejemplo: *"Detecting circular RNAs: bioinformatic and experimental challenges"* [49], *"Comparison of circular RNA prediction tools"* [59] ó *"A comprehensive overview and evaluation of circular RNA detection tools"* [60].

De este último estudio se ha extraido la comparativa entre 11 herramientas diferentes para la predicción de circRNA (Tabla 1)

Herramienta	Enfoque	Origen genómico	Ref.	Dependencias
CIRI	Lecturas basadas en segmentos	Exones, intrones, regiones intergénicas	[61]	BWA, perl
circRNA_finder	Lecturas basadas en segmentos	Exones, intrones, regiones intergénicas	[62]	STAR, samtools, perl
DCC	Lecturas basadas en segmentos	Exones, intrones, regiones intergénicas	[63]	STAR, python
Finc_circ	Lecturas basadas en segmentos	Exones, intrones, regiones intergénicas	[1]	Bowtie2, samtools, python
Segemehl	Lecturas basadas en segmentos	Exones, intrones, regiones intergénicas	[64]	Samtools
CIRCexplorer	Lecturas basadas en segmentos	Exones, intrones	[65]	STAR, bedtools, python
MapSplice	Lecturas basadas en segmentos	Exones, intrones	[66]	Bowtie, samtools, python
UROBORUS	Lecturas basadas en segmentos	Exones	[67]	Bowtie, Bowtie2, Tophat2, Samtools, perl
KNIFE	Lecturas basadas en segmentos	Exones	[68]	Bowtie, Bowtie2, samtools, python, R
PTESFinder	Basado en candidatos	Exones	[69]	Bowtie, Bowtie2, bedtools, Java
NCLScan	Basado en candidatos	Exones	[70]	BWA, BLAT, Novoalign, bedtools, samtools, python

Tabla 1. Herramientas para la detección de circRNA

Y de entre todas ellas concluye que CIRI, CIRCexplorer y KNIFE, logran un mejor rendimiento equilibrado entre su precisión y sensibilidad.

En el caso de KNIFE [68] , los requerimientos de memoria RAM y tiempo computacional son demasiado altos para los recursos disponibles. CIRI2 [61] y CIRCexplorer2 [65] son las opciones que se han tenido en cuenta. CIRI2 permite la detección de circRNAs en regiones intergénicas (no disponible en CIRCexplorer2) y, adicionalmente, al comparar la predicción forzosa de novo, CIRI2 es el predictor más confiable disponible [59].

Por todos estos motivos para la realización del presente trabajo se ha escogido la herramienta CIRI2 ya que, además, la documentación y casos de ejemplo para paired-end reads en el caso de CIRCexplorer2 son muy limitados en el momento de la realización del presente trabajo.

2.2. Implementación

2.2.1. Selección y descarga de las muestras secuenciadas

Teniendo en cuenta los requisitos expuestos en la primera parte de este capítulo, realizamos una búsqueda en el Gene Expression Omnibus repository (GEO) del NCBI (National Center for Biotecnology Information) [71] usando los siguientes criterios:

Organismo: Homo sapiens
Tipo de estudio: Expression profiling by high throughput sequencing
Cadena de búsqueda: "blood exosomes" AND cancer

De entre los resultados de la búsqueda, se seleccionaron tres proyectos de muestras de RNA-Seq de exosomas en sangre periférica adecuados para la detección de circRNA:

"RNA-seq reveals abundant circRNA, lncRNA and mRNA in ***blood exosomes*** *of patients with hepatocellular* ***carcinoma"*** [72]

"RNA-seq reveals abundant circRNA, lncRNA and mRNA in ***blood exosomes*** *of patients with colorectal* ***carcinoma"*** [73]

"RNA-seq reveals abundant circRNA, lncRNA and mRNA in ***blood exosomes*** *of patients with pancreatic* ***carcinoma"*** [74]

Estos proyectos fueron utilizados para la creación de exoRBase [75], un repositorio de ARNs circulares, ARN no codificante de cadena larga (lncRNA) y ARN mensajero, todos ellos procedentes del análisis de RNA-Seq de exosomas en sangre humana. A partir de esta última publicación seleccionamos un nuevo proyecto correspondiente a la secuenciación de exosomas de personas sanas:

"RNA-seq reveals abundant circRNA, lncRNA and mRNA in ***blood exosomes*** *of normal persons"* [76]

En los cuatro experimentos se parte de una muestra de 1 a 4 ml de plasma o suero. A continuación se extrae el ARN perteneciente a los exosomas usando el método distribuido por Quiagen "exoRNeasy Serum/Plasma Maxi kit"[77]. Después las muestras fueron tratadas con DNAse I y se eliminó el ARN ribosomal. Por último, las librerías fueron secuenciadas con la plataforma Illumina Hiseq.

Una vez seleccionados los proyectos **recuperamos los identificadores de las secuencias** correspondientes a cada muestra en el Sequence Read Archive (SRA) [78]. Para ello, en primer lugar, anotamos el identificador de cada proyecto:

- Pacientes con carcinoma hepatocelular: PRJNA390991
- Pacientes con carcinoma colorectal: PRJNA390615
- Pacientes con carcinoma pancreático: PRJNA391134
- Individuos sanos: PRJNA390988

Y utilizamos la herramienta "SRA Run Selector" del NCBI accesible en la url: https://www.ncbi.nlm.nih.gov/Traces/study/, para, a partir del identificador del proyecto, recuperar los **identificadores de las ejecuciones** (RUN) correspondientes a los SRAs que queremos descargar (Ilustración8).

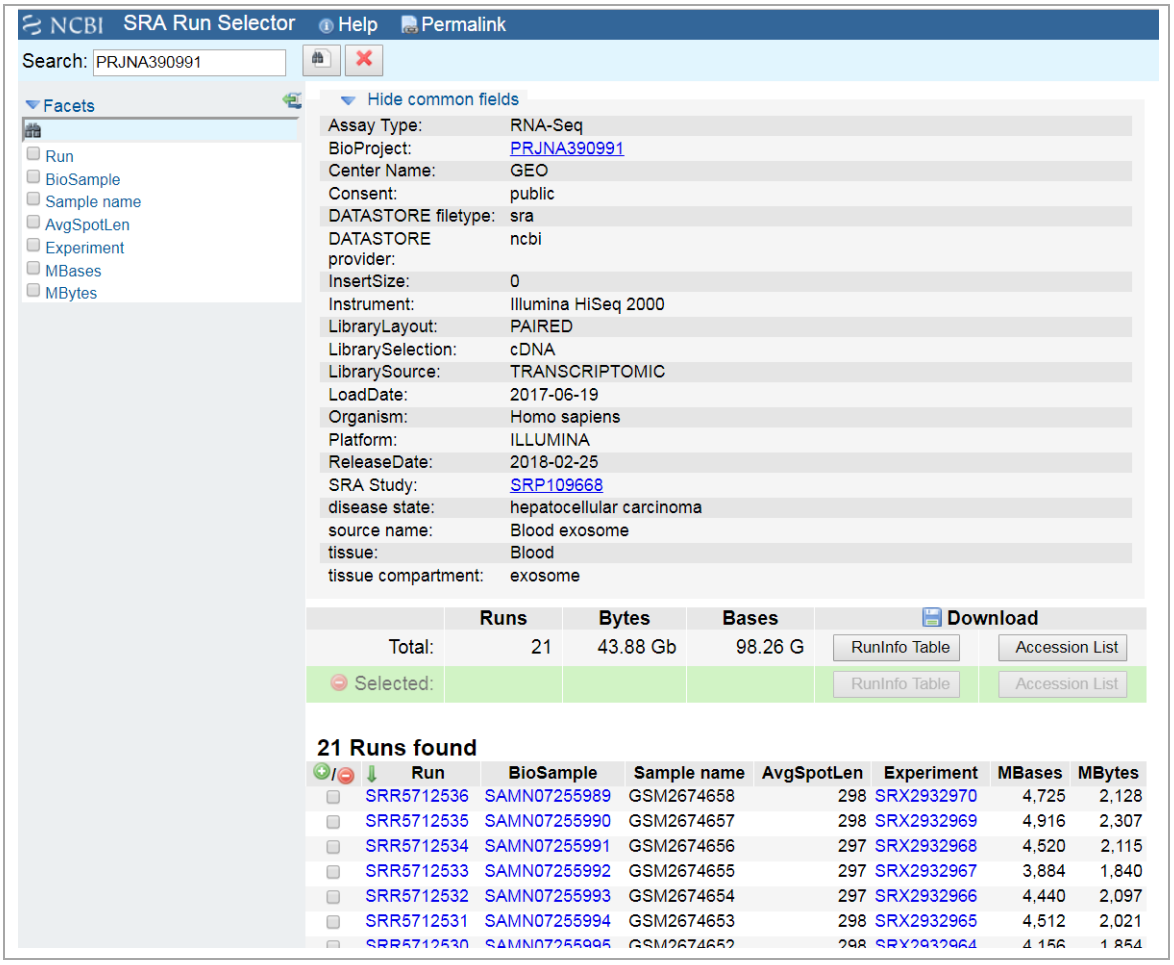

Ilustración 8. SRA Run Selector.

Como resultado obtuvimos el siguiente **número de muestras secuenciadas** para cada grupo (79 en total):

- 32 para el grupo de individuos sanos,
- 12 para cáncer colorectal,
- 21 para cáncer hepatocelular
- y 14 para cáncer pancreático.

Con esta información creamos un archivo de texto `phenodata.txt` delimitado por tabulaciones indicando en la primera columna el identificador de la ejecución y en una segunda columna el grupo al que pertenece (hepatocellular, colorectal, pancreatic, normal). Este archivo es necesario para el resto de las fases del presente trabajo. Podemos ver las primeras líneas para cada grupo en la Tabla 2.

File	Group
SRR5687235	colorectal
SRR5687236	colorectal
…	
SRR5712516	hepatocellular
SRR5712517	hepatocellular
…	
SRR5712482	normal
SRR5712483	normal
…	
SRR5714908	pancreatic
SRR5714909	pancreatic
…	

Tabla 2. Archivo phenodata.txt

Los **archivos SRA** que contienen las secuencias (listados en el archivo `phenodata.txt` en la columna `File`) **se pueden descargar** del repositorio: ftp://ftp-trace.ncbi.nih.gov, añadiendo el path al archivo SRA. El path se construye como:

```
/sra/sra-instant/reads/ByRun/sra/{SRR|ERR|DRR}/
<first_6_characters_of_accession>/<accession>/<accession>.sra
```

Por ejemplo, para el SRA con identificador SRR5687235 la ruta completa de descarga es:

ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR568/SRR5687235/SRR5687235.sra

El proceso de descarga se ha **automatizado** mediante la implementación del script en R llamado `Download.R`. Para la **ejecución** del script debemos tener instalado el entorno R (https://www.r-project.org/) en nuestro sistema. Este script se puede ejecutar en **cualquier sistema operativo** que soporte el entorno R. Por ejemplo, en un sistema Unix/Linux, desde línea de comandos, lanzaremos su ejecución con el comando:

```
> Rscript Download.R
```

usando como directorio de trabajo aquel en el que esté almacenado el archivo `phenodata.txt`. Como resultado, en nuestro directorio de trabajo, obtenemos los 79 archivos con extensión `.sra` cuyos identificadores estaban listados en la columna `File` del archivo `phenodata.txt`

Una vez **descargados** los archivos SRA con las seuencias de cada muestra, el siguiente paso que debemos dar es **convertir** los archivos al formato FASTQ [79], separando las lecturas en dos ficheros distintos si lo que tenemos son lecturas emparejadas (paired-end reads).

Para ello vamos a utilizar la herramienta `fastq-dump` incluida en el del "SRA Toolkit" del NCBI. Podemos descargar la herramienta en la url:

https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software

y está disponible para varios sistemas operativos.

El comando para convertir los archivos al formato FASTQ con `fastq-dump` es:

```
> fastq-dump --split-3 <fichero.sra>
```

La opción `--split-3` genera un solo fichero FASTQ si lo que tenemos son lecturas de una sóla hebra (single-end reads), o dos ficheros FASTQ separando las lecturas emparejadas (paired-end reads). En este último caso los nombres de los ficheros de salida llevarán los sufijos `_1.fastq` y `_2.fastq`. Además, podemos obtener un tercer archivo FASTQ sin sufijo y de un tamaño mucho menor, que contendrá lecturas "huérfanas". En caso de ocurrir, ese archivo lo descartaremos.

El proceso de conversión se ha **automatizado** mediante la implementación del script en R llamado `Trans.R`. Para la ejecución del script debemos tener instalado el entorno R (https://www.r-project.org/) y la herramienta `fastq-dump` del SRA-Toolkit que, además, debe poder ser invocada directamente desde el directorio de trabajo. En un sistema Unix/Linux, desde línea de comandos, lanzaremos su ejecución con el comando:

```
> Rscript Trans.R
```

usando como directorio de trabajo aquel en el que esté almacenado el archivo `phenodata.txt` junto con los archivos SRA. Este script se puede ejecutar en cualquier sistema operativo que soporte la ejecución del entorno R y la herramienta `fastq-dump`.

Como **resultado** obtenemos, en nuestro directorio de trabajo, 158 archivos con extensión `.fastq`, dos para cada identificador de la columna columna `File` del archivo `phenodata.txt`, ya que estamos trabajando con lecturas emparejadas. El archivo que contiene la primera lectura tiene el nombre `<identificador>_1.fastq` y el archivo con la segunda lectura emparejada, `<identificador>_2.fastq`.

2.2.2. Limpieza de los datos brutos y control de calidad

La herramienta seleccionada **fastp**[58] realiza el control de calidad, la eliminación de adaptadores, el filtrado de las lecturas de baja calidad y la corrección de las bases en una sola pasada, soportando además lecturas emparejadas (paired-end reads).

Pero, aunque fastp es una herramienta que produce informes de calidad muy completos, el **inconveniente** que encontramos, al igual que con el resto de herramientas para el control de calidad disponibles, es que generan **un informe por muestra**. En nuestro caso, un total de 79 informes con fastp o 158 con FASTQC (ya que realizamos un análisis de calidad previo al preprocesado y otro posterior). La herramienta **MultiQC** [80] puede **resumir** todos estos resultados permitiendo una visión más global del estado de la calidad de nuestros datos.

Y, aunque MultiQC soporta múltiples herramientas, a la fecha de realización de este estudio aún **no tiene implementada la integración con fastp**. Por este motivo, aunque sea redundante, **se ha utilizado también FASTQC** para el control de calidad previo y posterior, resumiendo a continuación los informes por grupo mediante MultiQC. Como resultado hemos obtenido 8 informes en vez de 158: 4 informes previos al preprocesado y 4 informes posteriores, agrupando los informes de las muestras del mismo grupo: normal, colorectal, hepatocellular y pancreatic.

El preprocesado y control de calidad se ha **automatizado** mediante la implementación del script en R llamado `Clean.R`. Las dependencias de este script son:

- el entorno R (https://www.r-project.org/)
- FASTQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
- fastp (https://github.com/OpenGene/fastp)
- y MultiQC (http://multiqc.info/)

que, además, deben poder ser invocadas directamente desde el directorio de trabajo. El sistema operativo debe ser un entorno Unix. Es necesario que el archivo `phenodata.txt` junto con los archivos FASTQ se encuentren en el directorio de trabajo. En un sistema Unix/Linux, desde línea de comandos, lanzaremos su ejecución con el comando:

```
> Rscript Clean.R
```

Como **resultado** obtenemos en nuestro directorio de trabajo:

- un conjunto de archivos `.fastq` resultado del preprocesado con fastp, uno por cada archivo de entrada, con nombre `<identificador>_clean_1.fastq` para la primera hebra e `<identificador>_clean_2.fastq` para la segunda.
- una carpeta `fastp` con los informes de calidad en formato html y json para cada identificador.
- una carpeta con los informes de calidad para los ficheros FASTQ sin procesar (datos brutos) para cada grupo listado en `phenodata.txt` llamada: `<grupo>_qc_raw`. Esta carpeta contiene los informes de la herramienta FASTQC y el informe en html unificado generado por la herramienta MultiQC, que reconocemos por llevar el sufjijo `MQC_`
- una carpeta con los informes de calidad para los ficheros FASTQ procesados para cada grupo listado en `phenodata.txt` llamada: `<grupo>_qc_clean`. Esta carpeta contiene los informes de la herramienta FASTQC y el informe en html unificado generado por la herramienta MultiQC, que reconocemos por llevar el sufjijo `MQC_`

En la Tabla 3 se muestra el resumen de las **estadísticas generales de control de calidad** generado por MultiQC a partir de los informes generados por FASTQC para el grupo colorectal tras el preprocesado.

General Statistics

Sample Name	% Dups	% GC	Length	M Seqs
SRR5687235_clean_1	58.3%	50%	132 bp	13.7
SRR5687235_clean_2	65.4%	49%	132 bp	13.7
SRR5687236_clean_1	62.5%	51%	132 bp	11.9
SRR5687236_clean_2	72.8%	50%	133 bp	11.9
SRR5687237_clean_1	61.2%	51%	133 bp	20.5
SRR5687237_clean_2	67.1%	50%	133 bp	20.5
SRR5687238_clean_1	70.8%	52%	131 bp	14.5
SRR5687238_clean_2	75.0%	51%	131 bp	14.5
SRR5687239_clean_1	54.7%	52%	136 bp	8.2
SRR5687239_clean_2	67.2%	50%	137 bp	8.2
SRR5687240_clean_1	72.3%	49%	142 bp	5.7
SRR5687240_clean_2	76.9%	47%	142 bp	5.7
SRR5687241_clean_1	67.5%	50%	139 bp	10.8
SRR5687241_clean_2	74.8%	49%	140 bp	10.8
SRR5687242_clean_1	53.2%	50%	133 bp	13.6
SRR5687242_clean_2	58.5%	49%	134 bp	13.6
SRR5687243_clean_1	66.7%	48%	132 bp	15.7
SRR5687243_clean_2	72.1%	47%	132 bp	15.7
SRR5687244_clean_1	55.4%	50%	140 bp	13.9
SRR5687244_clean_2	66.2%	48%	140 bp	13.9
SRR5687245_clean_1	71.1%	49%	134 bp	11.9
SRR5687245_clean_2	77.3%	48%	133 bp	11.9
SRR5687246_clean_1	75.9%	52%	139 bp	14.4
SRR5687246_clean_2	76.0%	50%	138 bp	14.4

Tabla 3. Colorectal: control de calidad de archivos fastq preprocesados.

Observamos un **porcentaje moderadamente elevado de duplicados**. Actualmente, los métodos cuantitativos de RNA-Seq se utilizan para trabajar con cantidades de muestras más pequeñas que requieren amplificación. Computacionalmente, los duplicados de lectura se definen por su posición de mapeo, que no distingue la PCR de los duplicados naturales y, por lo tanto, **no está claro cómo tratar las lecturas duplicadas**.

Se ha encontrado que una gran fracción de duplicados de lectura identificados computacionalmente no son duplicados de PCR y pueden explicarse por muestreo y sesgo de fragmentación. En consecuencia, **la eliminación computacional de duplicados no mejora la precisión** y, en realidad, **puede empeorar la potencia y la tasa de descubrimiento falso** (FDR) para la expresión génica diferencial [81].

Por tanto, **en base a esta evidencia**, aunque el porcentaje de duplicados se mantiene moderadamente elevado en todas nuestras librerías (no sólo en este grupo) **se ha optado por no eliminar computacionalmente los duplicados** en nuestras muestras.

Por otro lado, observamos también **un porcentaje medio de GC ligeramente elevado**, aunque en nuestras librerías no alcanza valores preocupantes. Es conocido que **la generación del cDNA** utizando un **cebado aleatorio induce sesgos en la composición** de nucleótidos al comienzo de lecturas de secuenciación de transcriptoma con Illumina Genome Analyzer. El sesgo **es independiente del organismo y del laboratorio** e impacta la uniformidad de las lecturas a lo largo del transcriptoma [52]. Este problema podemos observarlo con más detalle observando el gráfico de contenido de bases de una secuencia en la Ilustración 9.

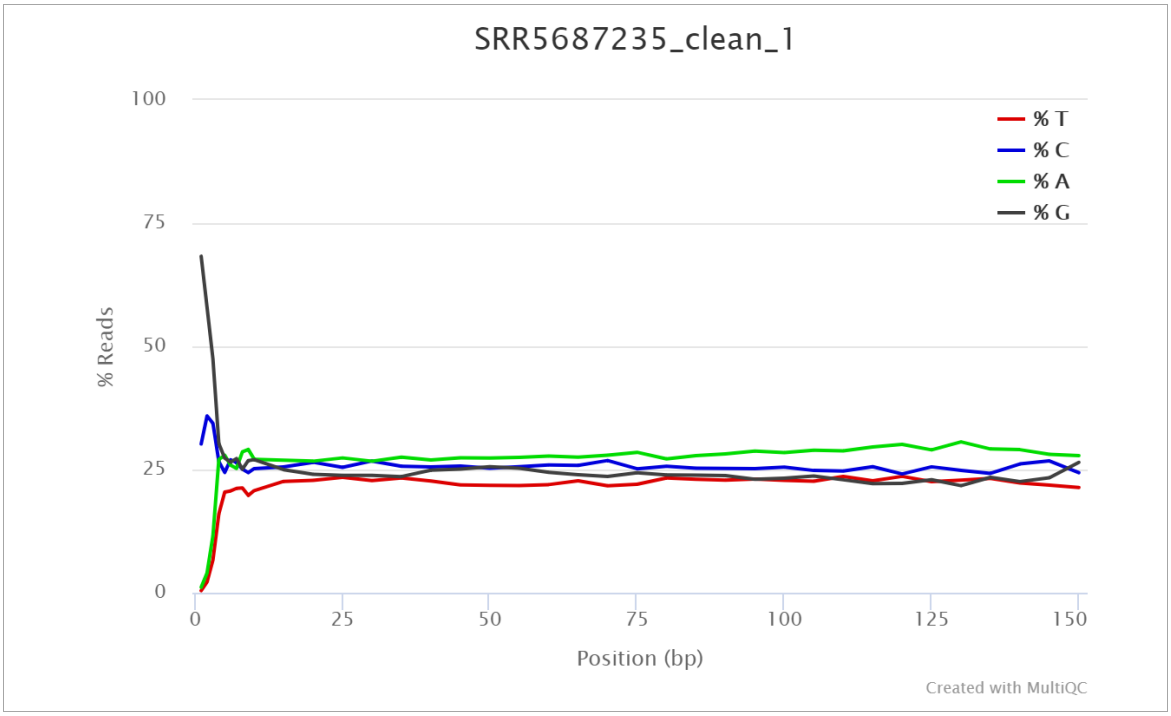

Ilustración 9. Contenido de bases de la secuencia SRR5687235 preprocesada.

Se observa claramente un sesgo en las primeras 10 bases, que después se disipa en el resto de la secuencia. A veces se opta por un recorte en el extremo 5' de las lecturas y así eliminar la parte sesgada, sin embargo, dado que la composición sesgada se crea mediante la selección de fragmentos y no por errores de secuenciación, el único efecto del recorte sería pasar de tener una biblioteca que comienza sobre posiciones sesgadas a tener una biblioteca que comience ligeramente más abajo de esas posiciones sesgadas.

También se podría pensar que es posible que el sesgo de selección introducido tenga un efecto significativo en la capacidad de la biblioteca para medir de manera justa el contenido de la población original de ARN debido a que ciertas secuencias se ven indebidamente favorecidas. Pero, en la práctica, parece que este hecho no supone un problema [52].

Por último, en las Ilustraciones 10 y 11 se muestra cómo la herramienta **fastp elimina** adecuadamente los **adaptadores**.

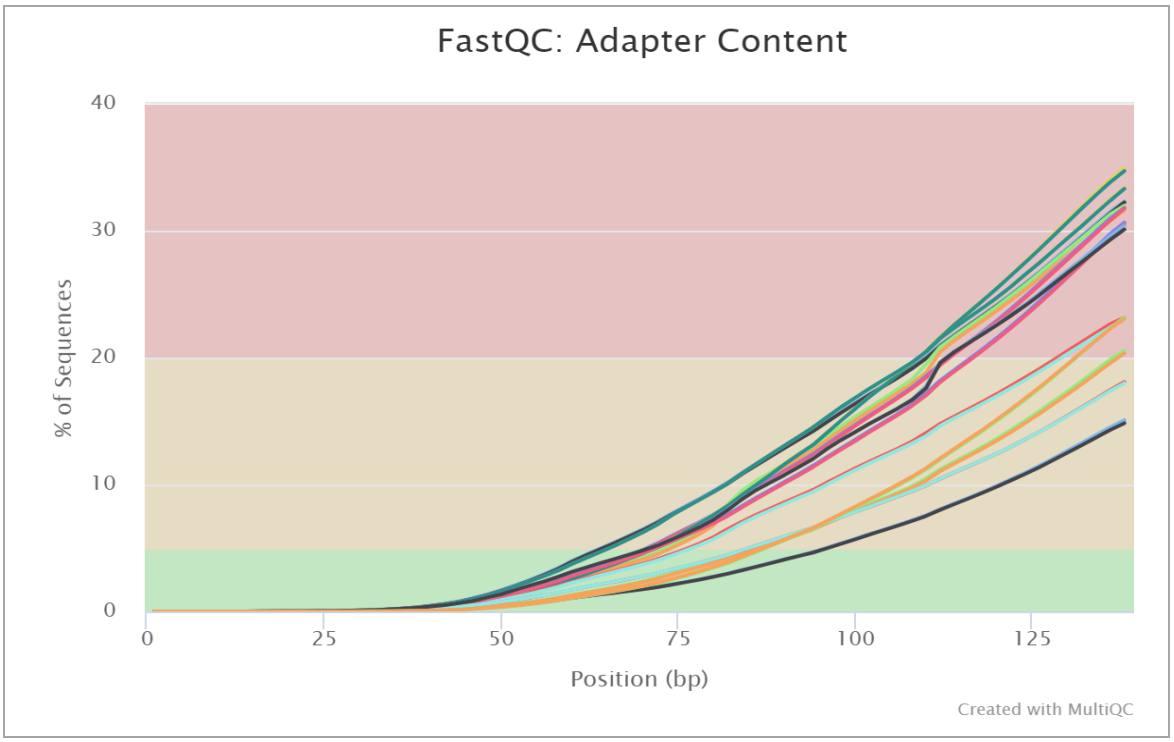

Ilustración 10. Contenido de adaptadores previo al preprocesado.

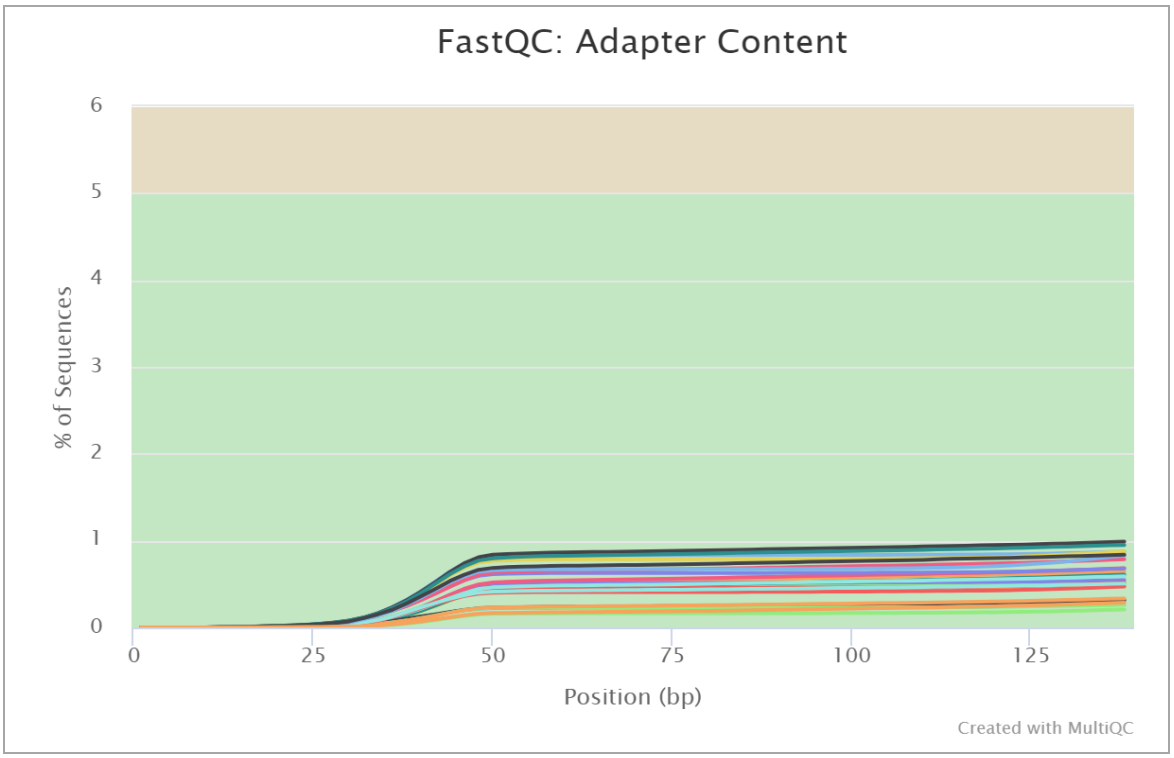

Ilustración 11. Contenido de adaptadores posterior al preprocesado.

Todos los informes de calidad se pueden descargar de:
https://github.com/carmengmz/circRNA/tree/master/experiment/quality_reports

2.2.3. Detección de circRNAs

Para poder utilizar **CIRI2** con el fin de **identificar los circRNAs** presentes en nuestras lecturas es necesario, en primer lugar, **alinear las lecturas a un genoma** de referencia usando el aligner **BWA** [82]. Se trata de una herramienta para mapear secuencias poco divergentes contra un gran genoma de referencia, como el genoma humano [83]. En este trabajo usaremos la última versión disponible del genoma de referencia **GRCh38**. Esta y otras versiones del genoma de referencia se pueden encontrar en la página web del Genome Reference Consortium http://genomereference.org

El **primer paso** para usar BWA es **construir un índice del genoma** de referencia en formato fasta. Para ello utilizaremos el siguiente comando:

```
> bwa index -a bwtsw hg38.fa
```

A continuación, **alineamos nuestras lecturas** al genoma de referencia usando el comando:

```
> bwa mem -t 10 -T 19 hg38.fa read_1.fastq read_2.fastq > aln-pe.sam
```

Donde la opción `-t 10` indica que se usarán 10 hebras y la opción `-T 19` indica que no se reporte como salida ninguna alineación con un score menor que 19. Este valor es el sugerido en el manual de CIRI2.

Por úlitmo pasamos el archivo `.sam` generado en el comando anterior a la herramienta **CIRI2:**

```
> perl CIRI2.pl -T 10 -I aln-pe.sam -O outfile -F hg38.fa -A hg38.gtf
```

La opción `-T 10` indica que se usarán 10 hebras y la salida **`outfile`** es un fichero de texto que contiene los **circRNAs identificados** en nuestras lecturas.

La identificación de circRNA se ha **automatizado** mediante la implementación del script en R llamado `Findcrna.R`. Las **dependencias** de este script son:

- el entorno R (https://www.r-project.org/)
- el lenguaje Perl (https://www.perl.org/)
- el aligner BWA (http://bio-bwa.sourceforge.net/)
- la herramienta CIRI2 (https://sourceforge.net/projects/ciri/)
- el genoma de referencia que almacenaremos en un fichero no comprimido llamado `hg38.fa`
- las anotaciones del genoma en un fichero llamado `hg38.gtf` (es muy importante descargar el genoma de referencia y sus anotaciones del mismo lugar)
- los ficheros fastq preprocesados (se espera `<nombre>_clean_1.fastq` y `<nombre>_clean_2.fastq`)
- el archivo `phenodata.txt`

El **sistema operativo debe ser un entorno Unix/Linux** y como **salida** obtendremos un fichero de texto por cada muetra llamado `<nombre>-outfile` que contiene una línea por cada uno de los circRNAs detectados, con sus coordenadas y el número de junction reads. Para **ejecutar** el script utilizaremos el siguiente comando:

```
> Rscript Findcrna.R
```

usando como directorio de trabajo aquel en el que estén almacenados los ficheros fastq preprocesados, el fichero fasta con el genoma de referencia, el fichero gtf con las anotaciones del genoma y el archivo `phenodata.txt`.

2.2.4. Anotación

Una vez obtenidos, mediante la herramienta CIRI2, los ficheros de salida con los circRNAs detectados en cada muestra, procederemos a la **creación de una única tabla** con los **recuentos de las lecturas** de los circRNAs de cada muestra.

Los circRNAs detectados mediante CIRI2 se encuentra en un archivo de texto con nombre `<identificador>-outfile` (un fichero por cada muestra) del que vamos a utilizar los campos listados en la Tabla 4.

Campo en la salida de CIRI2	Descripción
circRNA_ID	Identificador para el circRNA en el formato cromosoma:inicio|final. Por ejemplo, chr1:1223244|1223968
chr	Cromosoma
circRNA_start	Coordenada inicial
circRNA_end	Coordenada final
#junction_reads	Número de lecturas
strand	Hebra (+/-)

Tabla 4. Campos utlizados para la anotación de circRNA

Además del número de lecturas (`#junction_reads`) para cada circRNA detectado disponemos de su posición en un cromosoma y hebra con unas coordenadas de inicio y fin. Para construir la tabla de lecturas **se ha permitido un margen de 10 posiciones al inicio y al final** de las coordenadas para cosiderar que una detección pertenecía al mismo circRNA.

La **automatización** del proceso de anotación para construir la tabla con los recuentos de lectura se ha llevado a cabo en el script `Annotate.R`. El margen de posiciones de diferencia permitidas al inicio y final se puede configurar en la variable `range`. El script se puede ejecutar en cualquier sistema operativo que soporte el entorno R. En un entorno Unix/Linux lo haríamos con el siguiente comando:

```
> Rscript Annotate.R
```

En el directorio de trabajo deben hallarse los ficheros `<nombre>-outfile` con los circRNAs detectados por CIRI2 (un fichero por muestra) y el archivo `phenodata.txt`. Como **salida** obtendremos el fichero `circ_annotations.rds` que contiene, en las filas, un identificador para cada circRNA y, en las columnas, el identificador del circRNA, el cromosoma, hebra, posición inicial, posición final y el número de lecturas para ese circRNA en cada muestra.

En total, **el número de circRNA detectados** en todas las muestras **fue de 117.565**. Y por grupos, se detectaron 77.175 en el grupo de personas sanas, 53.767 en el grupo de cáncer hepatocelular, 39.420 en el grupo de cáncer pancreático y 39.931 en el grupo de cáncer colorectal.

Capítulo 3

Clasificación mediante Machine Learning

3.1. Conceptos teóricos

3.1.1. Adecuación de los datos: Variance Stabilizing Transformation

El **objetivo** que persigue el **presente análisis** es entrenar un algoritmo de **Machine Learning** que aprenda a **discriminar** entre muestras de circRNAs detectados en exosomas de sangre periférica de **individuos con cáncer** y de **individuos sanos**, para cada uno de los tres grupos de pacientes con cáncer estudiados: colorrectal, hepatocelular y pancreático.

Pero, para aplicar **métodos** de **Machine Learning**, primero es necesario **adecuar** los datos, ya que estos **algoritmos** pueden verse **afectados** por la **asimetría** (skewness), la **dependencia entre la media y la varianza** (heterocedastidad) o la **presencia de valores extremos**. Para evitar este tipo de problemas y facilitar posteriores análisis **es conveniente realizar transformaciones** sobre los datos de RNA-Seq [84]

Las **transformaciones** de las variables estadísticas se usan en estadística aplicada principalmente para **dos propósitos**: **normalización** y **estabilización de la varianza**. En la estabilización de la varianza se aplica una transformación para que sea posible el uso de técnicas estándar asociadas con variaciones continuas normales. Como su nombre indica, se pide a esta transformación que estabilice la varianza, esto es, que haga que la varianza de la variable transformada sea aproximadamente independiente, por ejemplo, de la media.

En los datos de conteos de RNA-Seq, con la **transformación de estabilización de la varianza** (VST) obtenemos una **matriz de recuentos de lectura** que es aproximadamente **homocedástica** (con varianza constante para todos los valores de la media). Esta transformación se usa generalmente para visualización, clustering u otras tareas de Machine Learning, no así para calcular la expresión diferencial.

En un principio se podría haber optado por una **transformación logarítmica**, **ajustando** además los conteos por el **tamaño de cada librería** para **reflejar** las **diferencias** entre la **profundidad** de la secuenciación. Y, como los valores de conteo para una clase (circRNAs en nuestro caso) pueden ser cero en algunos casos, hay quien aboga por el uso de **pseudo-conteos**. Son transformaciones de la forma $y = log_2(n + 1)$ o, de forma más general, $y = log_2(n + n_0)$, donde n representa los valores recuentos de lectura y n_0 es una constante positiva seleccionada. Esta transformación **soluciona problemas con los valores extremos,** pero aún **tendremos problemas de heterocedasticidad** (Ilustración 12).

Por otro lado, **el número total de lecturas** (tamaño de la librería) **no es una buena medida** de la profundidad de la secuenciación, pues **los circRNAs fuertemente expresados** pueden tener una **gran influencia** sobre el número total de lecturas. En su lugar, Anders y Huber proponen [85] ajustar los recuentos de lectura **estimando los factores de tamaño** $\hat{s}_j$ mediante la **media de los ratios observados** en los conteos k_{ij} como:

$$\hat{s}_j = median_i \frac{k_{ij}}{(\prod_{v=1}^{m} k_{iv})^{1/m}} \quad (1)$$

Una vez descartada la transformación logarítmica, si hiciésemos la suposición de que **las lecturas se muestrearon independientemente** en una población con fracciones fijadas de genes, los **recuentos de lectura** seguirían una **distribución multinomial**, que puede **aproximarse** mediante la distribución de **Poisson**. La distribución de Poisson tiene **un único parámetro**, que está determinado únicamente por su **media**. Su varianza y todas las demás propiedades se derivan de ella, en particular, la varianza es igual a la media. Sin embargo, se ha observado [86] que **la suposición** de la distribución de Poisson es **demasiado restrictiva**: predice variaciones más pequeñas que las que se ven en los datos. Por lo tanto, la prueba estadística resultante no controla el error de tipo I (la probabilidad de descubrimientos falsos) según lo anunciado. De hecho, las **distribuciones de probabilidad para los recuentos de lectura son heteroscedásticas** de forma natural, con **varianzas más grandes** para **conteos mayores** y se ha demostrado [87] que la **relación de media-varianza** para los **conteos de RNA-Seq** es aproximadamente **cuadrática**.

Para abordar este **problema de sobre-dispersión**, se ha propuesto **modelar datos de recuento** con **distribuciones binomiales negativas** (NB) [88]. Por ejemplo, este es el enfoque es utilizado en el paquete edgeR para el análisis de RNA-Seq [89]. La **distribución NB tiene parámetros** que están **determinados de forma única** por la **media** μ y la **varianza** σ^2. En edgeR, Robinson y Smyth supusieron [90] que la media y la varianza están **relacionadas por**:

$$\sigma^2 = \mu + \alpha \mu^2 \quad (2)$$

con una sola **constante de proporcionalidad** α que es la misma a lo largo del experimento y que **puede estimarse a partir de los datos**. Por lo tanto, solo se debe estimar un parámetro para cada clase, lo que permite la aplicación a experimentos con un pequeño número de réplicas.

El **modelo NB** es un **modelo mixto Gamma-Poission** que se puede interpretar como sigue: la distribución **Gamma** modela los **niveles no observados de la expresión** en cada muestra biológica y, condicionado al nivel de expresión, las **medidas de la máquina** secuenciadora siguen una distribución de **Poisson** [86]. Así, en la ecuación (2), el primer término representaría la varianza debida al error de muestreo de Poisson (la incertidumbre al medir la concentración de recuentos de lectura) y el segundo término la varizanza debida a la variación biológica entre réplicas.

El **parámetro α** es la **dispersión** y representa la variación de la expresión de un gen relativa a su media. La **relación** entre la **media y la dispersión** puede ser **parametrizada** con las constantes a_0 y a_1 como:

$$\alpha = a_0 + \frac{a_1}{\mu} \tag{3}$$

siendo $\boldsymbol{a_0}$ la **dispersión asintótica** y $\boldsymbol{a_1}$ el factor extra **de ruido de Poisson** (shot noise). El ruido de Poisson es el ruido **dominante para genes poco expresados** y la **dispersión** es el término **dominante para genes altamente expresados**.

Para **ajustar esta relación** (3) entre la dispersión y la media se utiliza un **modelo linear generalizado** (GLM, Generalized Linear Model) robusto [87]. Los GLMs son una generalización flexible de la regresión linear ordinaria para datos que no están normalmente distribuidos y **especifican las distribuciones** de probabilidad en función a sus **relaciones media-varianza**, por ejemplo, las relaciones cuadráticas entre la media y la varianza que siguen los recuentos de lectura de RNA-Seq. Los coeficientes del modelo ajustado son los parámetros a_0 y a_1.

Así, a partir de la ecuación (2) que modelaba la relación entre la media y la varianza en una distribución NB, la **varianza** queda como:

$$v = \mu + \alpha\,\mu^2 = \mu + \mu^2 a_0 + \mu\, a_1 \tag{4}$$

Una transformación estabilizadora de la varianza (**VST)** es una **transformación $\boldsymbol{u}$** de manera que, si X (nuestros recuentos de lectura) es una variable aleatoria con una relación entre la media y la varianza $Var(X) = v(E(X))$, entonces $u(X)$ tiene una **varianza estabilizada,** es decir, es homocedástica. Esta transformación u puede derivarse de una relación v entre la media y la varianza como [91]

$$u(x) = \int^x \frac{d\mu}{\sqrt{v(\mu)}} \tag{5}$$

Sustituyendo el valor de v de la ecuación (4), la transformación queda como:

$$u(x) = \int^x \frac{d\mu}{\sqrt{v(\mu)}} = \int^x \frac{d\mu}{\sqrt{\mu + \alpha\,\mu^2 = \mu + \mu^2 a_0 + \mu\, a_1}} \tag{6}$$

y resolvidendo la integral en la ecuación (6) tenemos una VST general con la siguiente transformación $u(x)$:

$$u(x) = \frac{\log\left(\frac{1+2x\,a_0+a_1+2\sqrt{x\,a_0\,(1+x\,a_0+a_1)}}{1+a_1}\right)}{\frac{4\,a_0}{\log(2)}} \qquad (7)$$

con: $a_0 > 0, a_1 > 0$ y $x > 0$ [92]

Antes de aplicar la transformación se normalizan los conteos de lectura (x) dividiendo por los factores de tamaño descritos en la ecuación (1).

Sobre **nuestro conjunto de recuentos de lectura** podemos observar **cómo se comporta la varianza** con respecto a la **media** (Ilustración 12) en los tres supuestos: los **datos brutos**, la transformación usando el **logaritmo** (más una constante igual a 1) y la **transformación VST**. Vemos cómo, en la transformación usando el logaritmo, se mantiene la dependencia de la media y la varianza para valores grandes y cómo, con la transformación VST, la varianza se mantiene mucho más estable para todos los valores de la media.

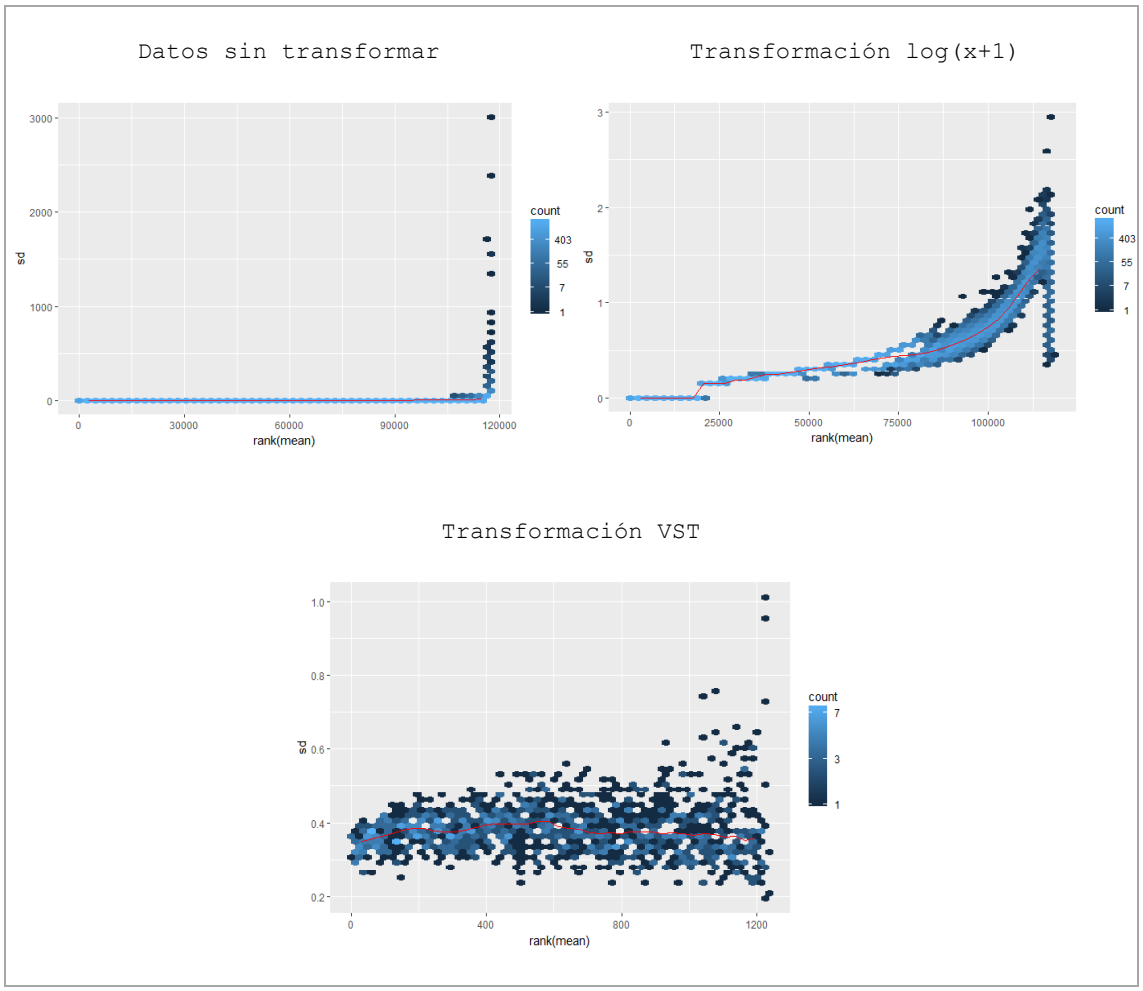

Ilustración 12.Relación entre la media (eje x) y varianza (eje y) para diferentes transformaciones.

3.1.2. Selección de predictores: Random Forest

En la selección de predictores **queremos conservar** aquellas **variables explicativas** de nuestro conjunto de datos que sean las **más relevantes para el modelado predictivo** del problema que estamos resolviendo. Cuando tenemos **datos con una alta dimensionalidad**, como es el caso de los recuentos de lectura de RNA-Seq, la **selección de predictores** es una **herramienta efectiva** para **reducir la dimensionalidad**, **incrementar la precisión** en el aprendizaje y, sobre todo, para **mejorar la comprensión** del modelo.

Destacar que, para generar un **modelo de Machine Learning válido**, es necesario **dividir el conjunto de datos** en dos subconjuntos: un primer conjunto de **entrenamiento**, sobre el que el algoritmo "aprende" y un segundo subconjunto de **pruebas** sobre el que probaremos la validez del modelo "aprendido", es decir, comprobaremos **si el modelo es capaz de generalizar** sus predicciones utilizando un conjunto de datos diferente al de entrenamiento. Y para **mantener la validez** del modelo la **selección de predictores** debe hacerse siempre **en el conjunto de entrenamiento**, en otro caso el modelo estaría sesgado ya que incorporaría información sobre la distribución del conjunto de pruebas.

Una de las técnicas usadas **para decidir la importancia de las variables** en el modelo son los **Random Forest**. Para ello, en primer lugar, **se ajusta un modelo de Random Forest** a los datos de entrenamiento para, a continuación, **recuperar la importancia de las variables** utilizadas en ese modelo de clasificación. Por último, **se selecciona un conjunto** adecuado y **relevante de predictores**, en función de su importancia.

Se trata de **métodos basados en árboles** que, en Machine Learning, se encuentran entre los más efectivos y útiles, capaces de producir resultados confiables y comprensibles, en la mayoría de los tipos de datos. Pero no es posible hablar de Random Forest sin introducir primero los árboles de decisión y clasificación y el algoritmo CART.

Árboles de decisión y clasificación

Los **árboles de decisión y clasificación** (CART, Classification and Regression Trees) fueron introducidos por primera vez en 1984 por un grupo liderado por Leo Breiman [93]. El **algoritmo CART** establece un **método** para realizar secuencialmente **particiones binarias** para las variables de entrada, lo que **resulta** en **una estructura de decisión semejante a un árbol**. Este algoritmo es capaz **de crear una respuesta** para **modelos** cuantitativos (numéricos) y **cualitativos** (categóricos). En este último caso, que es el que nos ocupa, **el espacio de decisión**, es decir, el grupo de las variables de entrada **se estratifica continuamente** utilizando un **error de clasificación**. En el caso de los modelos cuantitativos se usa la suma de cuadrados residual.

En una **tarea de clasificación**, si tenemos un conjunto de n observaciones de una variable de clase Y que toma valores $1,2,\dots,k$; y p variables predictoras, $X_1,\dots,X_p$, nuestra **meta** es **encontrar un modelo** para **predecir** los valores de Y para nuevos valores de X. Un **árbol de decisión** encuentra la **solución** al problema mediante una **partición** del espacio de X en k **conjuntos disjuntos**, $A_1,..,A_k$, de manera que el valor que se predice para Y es j si X

pertenece a A_j, para $j = 1, \dots k$. Los árboles de clasificación **se basan en conjuntos rectangulares** A_j generados **particionando recursivamente** el conjunto de datos X tomando las variables de una en una (Ilustración 13).

Para **implementar** este modelo, es necesario **establecer una metodología** que permita **crear las regiones** A_j, o lo que es lo mismo, decidir en **qué predictores** y en **qué valores** se **producirán las particiones**. Y como computacionalmente no es posible considerar todas las posibles particiones del espacio de los predictores, se utiliza **un enfoque voraz** (greedy) **de arriba hacia abajo** conocido como ***recursive binary splitting*** (división binaria recursiva).

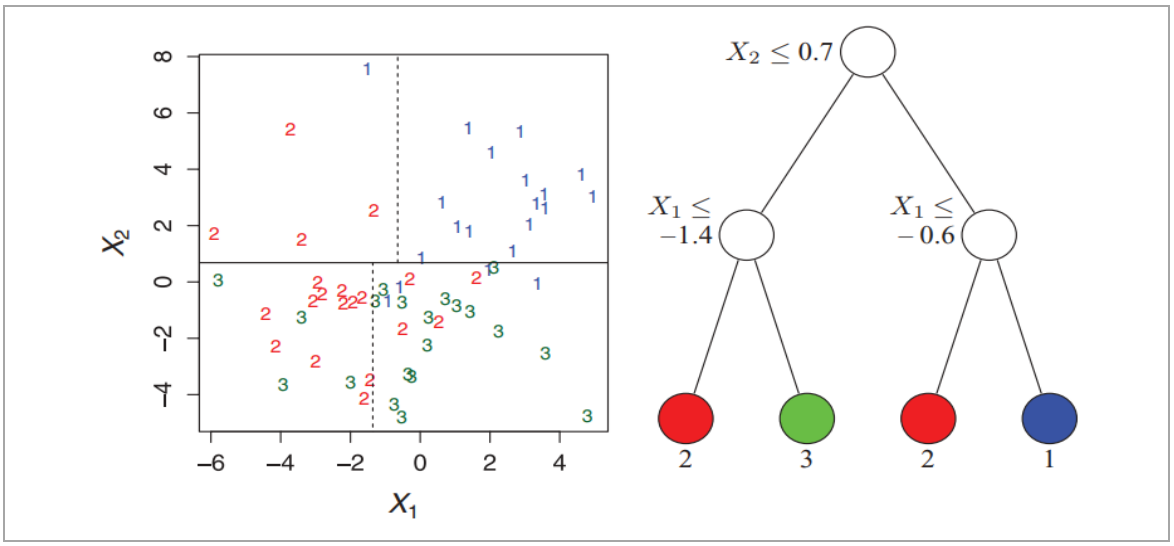

Ilustración 13. Particiones (izq.) y estructura de un árbol de clasificación.

Es un **enfoque de arriba abajo** ya que comienza en la raíz del árbol y sucesivamente va dividiendo el espacio de predictores: cada partición genera dos nuevas ramas hacia abajo en el árbol. Y es un **enfoque greedy** ya que, en cada paso de la construcción del árbol, se escoge la mejor partición para ese paso particular, sin tener en cuenta qué resultados podría producir esa partición en el futuro. Para ello, **seleccionamos el predictor** X_j y el **punto de corte** (umbral) s tal que, **al partir el espacio de los predictores** en regiones $\{X|X_j < s\}$ y $\{X|X_j > s\}$, se consigue la **mayor reducción posible en el error**. A continuación, **repetimos el proceso**, buscando el predictor que minimice el error en cada una de las regiones anteriores, hasta que se llega a un **punto de parada**, por ejemplo, hasta que todas las regiones contengan como máximo un número de observaciones establecido, o hasta que el árbol tenga un máximo de nodos terminales o que la incorporación del nodo reduzca el error en al menos un porcentaje mínimo.

Para cada **posible división** se calcula el **valor de la medida** en cada uno de los **dos nodos resultantes**. Se **suman** los dos valores **ponderando** cada uno por **la fracción de observaciones** que contiene cada nodo. Este paso es muy importante, ya que no es lo mismo dos nodos puros con 2 observaciones, que dos nodos puros con 100 observaciones. **La división con menor o mayor valor** (dependiendo de la medida empleada) se **selecciona** como división óptima. La **respuesta** para una **observación de test** es la **clase del mayor**

número de observaciones de entrenamiento **en la región** a la que la observación de test pertenezca.

El **error en un árbol de clasificación** se puede calcular como la fracción de las observaciones en esa región que no pertenecen a la clase más numerosa:

$$E = 1 - \max_k(\hat{p}_{mk}) \tag{8}$$

donde $\hat{p}_{mk}$ representa la proporción de observaciones de entrenamiento en la región m-ésima que pertenece a la clase k.

Por otro lado, también se puede usar el **Gini index**, definido como:

$$G = \sum_{k=1}^{k} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{9}$$

siendo una **medida de la varianza total** a lo largo de las K clases. Esta medida toma un valor pequeño si todos los $\hat{p}_{mk}$ son cercanos a cero. Por esta razón podemos pensar en el Gini index como una medida de la pureza: un valor pequeño indica que un nodo contiene predominantemente observaciones de una única clase [94].

Bagging

Los árboles no tienen el mismo nivel de precisión en la predicción que otros métodos de clasificación o regresión, pero **agregando** muchos árboles de decisión, se puede **mejorar su rendimiento** predictivo. La agregación **bootstrap** o bagging es un **procedimiento** general para **reducir la varianza** de un método de aprendizaje estadístico.

Dado un conjunto de n observaciones independientes $Z_1, \dots, Z_n$, cada una con varianza σ^2, la varianza de la media $\bar{Z}$ de las observaciones viene dada por σ^2/n. Es decir, **promediar** un conjunto de observaciones **reduce la varianza**. Pero como no tenemos acceso a múltiples conjuntos de entrenamiento, en su lugar hacemos un **muestreo** (bootstrap) tomando muestras repetidas del conjunto de entrenamiento. De esta manera generamos **B conjuntos** de entrenamiento diferentes. Entonces **entrenamos** el modelo en **cada conjunto** muestreado y, para **cada observación**, tomamos la **clase predicha** por cada uno de los B árboles y nos quedamos con el **voto mayoritario**: la predicción global será la más habitual de las ocurridas de entre las B predicciones.

La idea principal en el bagging es que los **árboles** son **ajustados** de **forma repetida** a los conjuntos muestreados de observaciones y, de media, **cada "bagged tree"** hace **uso** de alrededor de **2/3 de las observaciones**. Al resto de observaciones que **no se usaron** para ajustar ese modelo se les da el nombre de **observaciones "out-of-bag"** (OOB). Podemos predecir la respuesta para la i-ésima observación que quedó fuera del modelo (OOB-observation) usando cada uno de los árboles. A continuación, hacemos la media de esas B/3 predicciones para la i-ésima observación, y de esta manera se estima **el error de validación** cruzada para bagging (OOB-error).

Random Forest

Los **Random Forest** suponen una **mejora** sobre los "bagged trees" porque, de alguna manera, de correlacionan los árboles reduciendo la varianza cuando promediamos los árboles. Igual que en el bagging, **construimos un número de árboles** de decisión en **conjuntos de entrenamiento muestreados**. Pero, al construir esos árboles de decisión, **cada vez que se produce una partición** en un árbol, se considera **una selección aleatoria de m predictores** (de los p disponibles), y se **repite el proceso** para cada **partición**. Típicamente **el número de predictores considerados** en cada partición es la raíz cuadrada del número total de predictores ($m \cong \sqrt{p}$) [95]

Para conocer la **importancia** de las variables en **un único árbol** de clasificación, podemos utilizar la medida del **error** o el **Gini index** que es **único para cada predictor**, y ordenar los predictores en orden de error decreciente. En el caso de los **Random Forests** tenemos **múltiples árboles** de clasificación. Para calcular la **importancia** de un predictor, conocida como **MDI** (Mean Decrease Impurity), tomamos la **suma** sobre el **número de particiones** a lo largo de **todos los árboles** en los que se ha incluido este predictor, **proporcionalmente** al **número de muestras** que particiona [96].

3.1.4. Clasificación: Support Vector Machines.

Una vez que hemos seleccionado, en el conjunto de entrenamiento, los predictores más relevantes para nuestro problema, utilizaremos esos predictores para **entrenar un modelo de clasificación** que aprenda a **discriminar** entre muestras de **individuos con cáncer y sanos**. El primer enfoque que utilizaremos son las **Máquinas de Vectores de Soporte** (SVM, Support Vector Machines)

Las SVM fueron desarrolladas por Cortes y Vapnik en 1995 [97] para clasificación binaria. Se trata de un **modelo de aprendizaje supervisado** para clasificación cuya idea es buscar por un **hiperplano** que **separe de forma óptima** un conjunto de **muestras que pertenecen a clases diferentes**, maximizando el margen entre los puntos más cercanos entre ambas clases (Ilustración 14). Las líneas que pasan por los puntos que definen este margen se llaman **support vectors** y la línea en el medio es el **hiperplano** de separación óptima. Si hubiera puntos en el lado incorrecto del hiperplano, se les asigna un peso menor para reducir su influencia (soft margin).

Un hiperplano pude ser escrito como un conjunto de puntos $\boldsymbol{x}$ que satisfacen: $\boldsymbol{w}\,\boldsymbol{x} + b = 0$, donde $\boldsymbol{w}$ es el vector normal al hiperplano. El valor $\frac{b}{\|\boldsymbol{w}\|}$ es la distancia del hiperplano al origen usando el vector normal $\boldsymbol{w}$. Si los datos son linealmente separables entonces podemos seleccionar dos hiperplanos paralelos que separen los puntos de las dos clases. La región confinada entre estos dos hiperplanos es llamada margen [98], y el plano de máximo margen es aquel que se encuentra entre estos dos hiperplanos. La distancia entre los dos hiperplanos es $\frac{2}{\|\boldsymbol{w}\|}$, así que, para maximizar la distancia entre los planos, debemos minimizar $\|\boldsymbol{w}\|$.

La distancia se calcula usando la ecuación de la distancia de un punto a un plano añadiendo las siguientes restricciones, que impiden que un punto pueda quedar en el lado incorrecto del margen:

$$\begin{aligned} \boldsymbol{w} \cdot x_i + b \geq 1 \; si \; y_i = 1 \\ \boldsymbol{w} \cdot x_i + b \leq 1 \; si \; y_i = -1 \end{aligned} \tag{10}$$

es decir, que clasifique correctamente los vectores de ejemplos $\boldsymbol{x_i}$ en dos clases y_i. Estas restricciones pueden ser reescritas como: $y_i(w^T \cdot x_i + b) \geq 1 \;\; para \; 1 \leq i \leq n$

Por tanto, se trata de un problema de optimización en el que queremos minimizar $\|\boldsymbol{w}\|$ sujeto a $y_i(w^T \cdot x_i + b) \geq 1 \;\; para \; 1 \leq i \leq n$. El par (**w**,b) que resuelve esta ecuación determina nuestro clasificador [99]:

$$x \longmapsto signo(\boldsymbol{w} \cdot \boldsymbol{x} + b) \tag{11}$$

En las SVMs los datos de entrenamiento se usarán para "aprender" los pesos y luego se descartan. Sólo necesitamos $\boldsymbol{w}$ para clasificar nuevos datos.

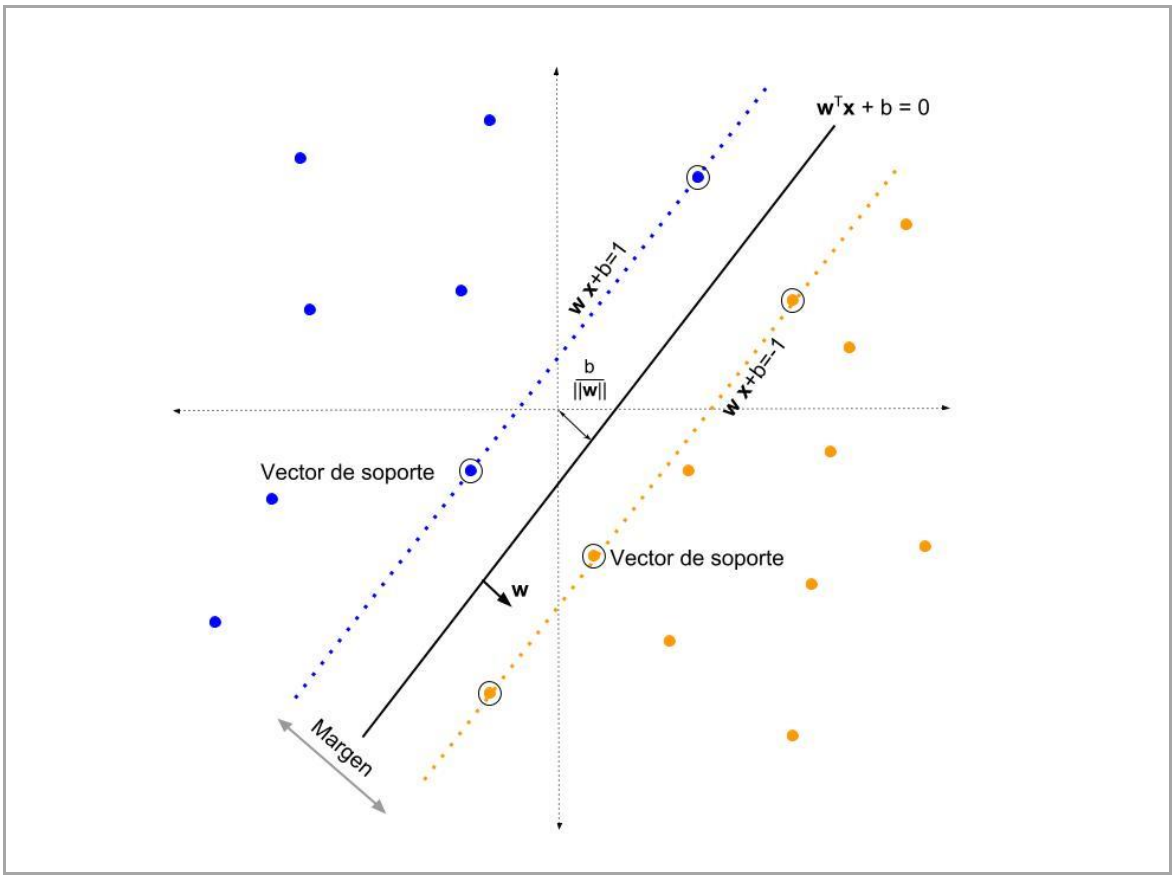

Ilustración 14. SVM lineal.

3.1.5. Clasificación: Redes neuronales artificiales.

El segundo enfoque para entrenar un modelo de clasificación que utilizaremos en este trabajo son las **Redes Neuronales Artificiales** (NN, Neural Networks). Se trata un conjunto de **algoritmos** diseñados para **reconocer patrones**. Una NN modela la **relación** entre un conjunto de **señales de entrada** y una señal de **salida** usando un modelo derivado de nuestro conocimiento de cómo el **cerebro biológico** responde a **estímulos** sensoriales. Igual que el cerebro utiliza una red de neuronas interconectadas para crear un procesador paralelo masivo, las NN utilizan una red de neuronas artificiales o nodos para resolver problemas de aprendizaje.

El modelo de una única neurona artificial puede ser entendido en términos muy similares al modelo biológico (Ilustración 15).

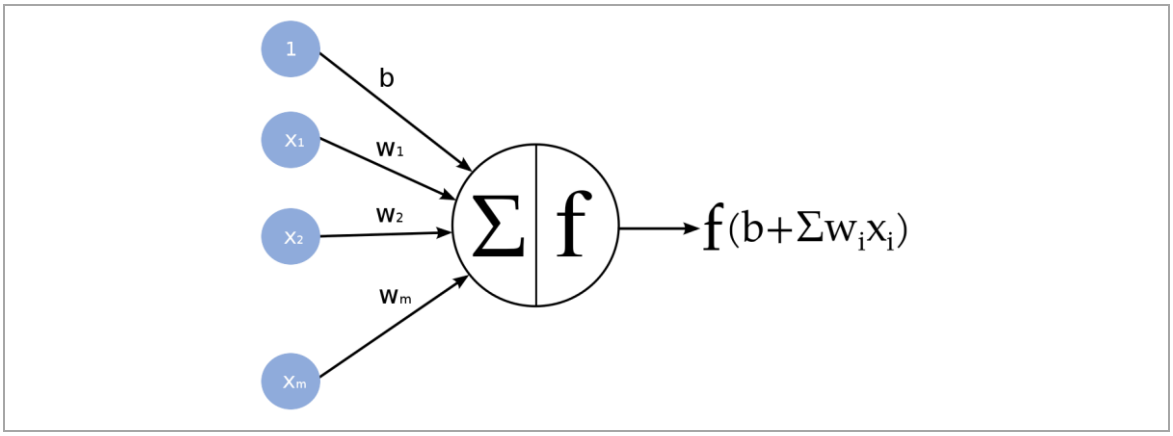

Ilustración 15. Neurona artificial

Una **red dirigida** de forma directa define una **relación** entre las **señales de entrada** recibidas por las dendritas (variables x) y la señal de **salida**. Tal como sucede en la neurona biológica, a cada **señal de entrada** de cada dendrita se le aplica un **peso** (valores w) de acuerdo con su **importancia**. Las señales de entrada se **adicionan** por el cuerpo celular y la señal se **retrasmite** de acuerdo con una **función de activación** (f).

Los **pesos** w permiten que cada una de las n entradas (x_i) **contribuyan** más o menos a la **suma total** de las señales de entrada. Esta suma es usada por la función de activación $f(x)$ y la señal resultante es la salida del axón.

Las NNs utilizan **neuronas** definidas de esta manera como los **bloques básicos** para construir modelos complejos. Pero, aunque hay **numerosas variantes**, todos pueden ser definidos en términos de las siguientes **características**:

- Una **función de activación**, que transforma la combinación de las señales de entrada de una neurona en una sola señal de salida que es retransmitida por la red.
- Una **arquitectura o topología** que describe el número de neuronas en el modelo, así como el número de capas y la forma en la que están interconectadas.

- El **algoritmo de aprendizaje** que especifica como se establecen los pesos de cada conexión de manera que se inhiban o activen las neuronas proporcionalmente a la señal de entrada.

La **función de activación** es el mecanismo por el cual la neurona artificial **procesa** la información entrante y la retransmite a través de la red. Para **calcular** la salida, lo que hace una neurona es calcular **una suma ponderada** de sus entradas, añade una **tendencia** (b) y decide si debe pasar la señal a la siguiente neurona o no (**disparo** o activación). El valor de Y puede ser cualquiera en el rango $(-inf, +inf)$.

$$Y = \sum (w_i * x_i) + b \tag{12}$$

Para decidir si la neurona se dispara o no se utilizan las **funciones de activación**. Estas funciones comprueban el valor de Y y deciden si la señal debe **continuar** (si la neurona debe activarse). Una de las más utilizadas es la **función sigmoidea** (más específicamente el sigmoide logístico, Ilustración 16), siendo e la base del logaritmo natural (aproximadamente 2,72). Los valores de salida están en el rango continuo (0,1).

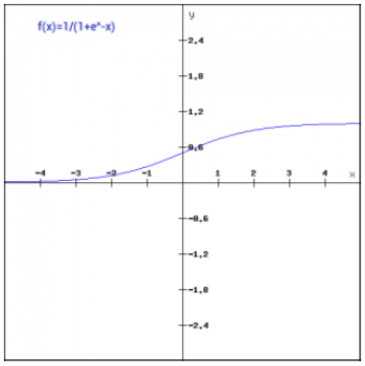

Ilustración 16. Función sigmoidea

$$f(x) = \frac{1}{1 + e^x} \tag{13}$$

Es una **función no lineal** lo que permite **apilar capas**, ya que las combinaciones de esta función también serán no lineales. También es una **función diferencial**, lo que significa que es posible calcular la derivada a través del rango entero de entradas. Esta característica es crucial para crear algoritmos de optimización eficientes.

Y aunque la sigmoidea es tal vez la más común y se usa a menudo por defecto, algunos algoritmos **permiten alternativas**, como funciones lineares o gaussianas. La principal **diferencia** entre estas funciones son los **rangos** de la señal de **salida**. La **elección** de la función de activación **condiciona** la red neuronal de manera que puede **ajustar** mejor ciertos tipos de información más apropiadamente, permitiendo la construcción de redes neuronales especializadas.

Por otro lado, la **habilidad** de una red neuronal de **aprender** está contenida en su **topología** o los patrones y estructuras que interconectan las neuronas. Aunque hay innumerables formas de arquitecturas de redes, se pueden **diferenciar** por tres **características** clave: el **número de capas**, si la información puede viajar hacia atrás (**backpropagation**) y el **número de nodos** en cada capa de la red.

La **topología** determina la **complejidad** de tareas que pueden ser aprendidas por la red. Generalmente, las redes más grandes y complejas son capaces de identificar patrones más sutiles y límites de decisión más complejos. Aún así la potencia de la red no es solo función del tamaño de la red, sino de cómo las unidades están interconectadas.

Para **definir la topología**, necesitamos una **terminología** que distinga a las **neuronas** artificiales en función de su **posición en la red**. Un conjunto de neuronas llamados **nodos de entrada** reciben las señales sin procesar directamente de los datos de entrada. Cada nodo de entrada es responsable de **procesar una sola** característica del conjunto de datos. El valor de la característica es **transformado** por la **función de activación** del nodo correspondiente. Las señales **enviadas** por los **nodos de entrada** son recibidas por los **nodos de salida**, que usan su propia **función de activación** para generar la **predicción final**.

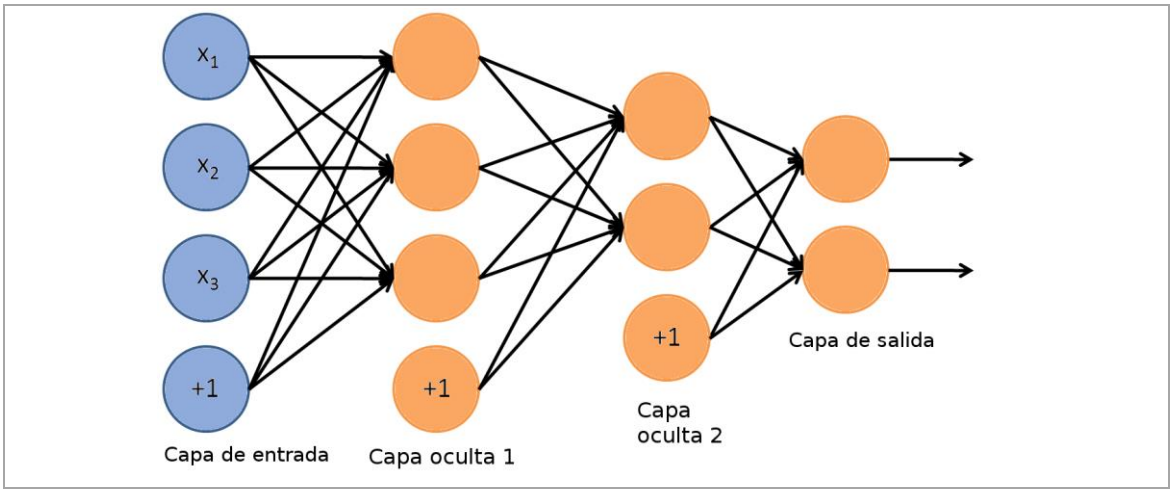

Ilustración 17. Red Neuronal multicapa

Los nodos de entrada y salida se disponen en grupos llamados **capas**. Podemos tener redes con una o más capas. Las redes con una sola capa son adecuadas para tareas de clasificación básicas, particularmente para patrones que son linealmente separables.

Una **red multicapa** (Ilustración 17) añade una o más capas ocultas que procesan las señales desde los nodos de entrada y los redirigen hacia el nodo de salida. La mayoría de las redes multicapa están **completamente conectadas**.

No hay una regla para determinar el número de neuronas en las capas ocultas. El número apropiado **dependerá** del número de nodos de **entrada**, la **cantidad de datos** de entrenamiento, la cantidad de **ruido** y la **complejidad** de la tarea de aprendizaje.

En general, una red más compleja permite aprender problemas más complejos, pero corremos el riesgo de sobreajuste (overfitting) generalizando peor datos futuros. La **mejor práctica** es utilizar el **menor número de nodos** que resulten en un **desempeño adecuado** en un test de validación.

Estas redes utilizan un algoritmo de aprendizaje denominado **backpropagation** cuya estrategia es **propagar los errores hacia atrás**. En su forma más general, el algoritmo **itera** a través de muchos ciclos de **dos procesos**. Cada ciclo se conoce como una **época** (epoch).

Como la red no contiene a priori ningún conocimiento, los pesos de salida son típicamente fijados aleatoriamente. Entonces el algoritmo itera a través de los procesos, hasta que se cumple el criterio de parada.

Cada **época** en el algoritmo de backpropagation incluye **una fase hacia adelante** (fordward phase) en la cual las neuronas son activadas de forma secuencial desde la capa de entrada hacia la capa de salida, aplicando los pesos de cada neurona y sus funciones de activación durante el camino, hasta que se llega a la capa final y se produce la señal de salida.

Y una fase hacia atrás (backward phase) en la que la señal de salida resultante de la fase anterior se compara con el valor de referencia verdadero en los datos de entrenamiento utilizando la función de coste, y se ajustan los pesos usando gradient-descent, que consiste en ir cambiando los pesos en pequeños incrementos después de cada iteración del conjunto de datos calculando la derivada (o gradiente) de la función de coste para determinar en qué dirección se encuentra el valor mínimo para el coste.

Una red neuronal con **múltiples capas ocultas** que utilice el algoritmo de aprendizaje de backpropagation se llama **Deep Neural Network (DNN)** y en la práctica a estos modelos se les suele denominar como de D**eep Learning** [100].

Otro enfoque de aprendizaje en redes neuronales, que es el que utilizaremos en el presente trabajo, es el algoritmo de **Extreme Learning Machine**. Se trata de una red neuronal **feedforward** con una **única capa oculta**, en la que no se utiliza la métrica de gradient-descent para en entrenamiento, como sucede en backpropagation.

En esta red se asignan de forma **aleatoria los pesos** de las capas ocultas y se calcula la salida mediante la función de activación, y su proceso de aprendizaje solo necesita una iteración. Se basan en la idea de **"random projection"** (una técnica para reducir la dimensionalidad en un espacio euclídeo) seguido de **linear regression** [101].

3.1.6. Cálculo de la bondad del modelo.

Para **evaluar** los resultados del modelo en la fase de **entrenamiento** y garantizar que son **independientes** de la **partición** entre datos de entrenamiento y prueba, utilizaremos **validación cruzada** (cross validation) en K iteraciones. Para ello, **dividimos** el conjunto de entrenamiento, en **K subconjuntos**. Uno de ellos lo utilizaremos como conjunto de prueba y el resto (K-1) como conjunto de entrenamiento, **repitiendo** el proceso durante **k iteraciones** con cada uno de los posibles subconjuntos de prueba. Se trata de un **método muy preciso** puesto que **evaluamos** a partir de **K combinaciones** de datos de entrenamiento y prueba. La **bondad** del modelo es la **media** de la bondad de los K modelos ajustados.

Para reportar la bondad del modelo utilizaremos el **AUC** (area under the curve). Se trata de una medida de la **sensibilidad** frente a la **especificidad** en un sistema de clasificación binario, es decir, representa la razón entre los **verdaderos positivos** y los **falsos positivos**. Un **AUC igual a 1** representa una **clasificación perfecta**, un **AUC de 0,5** representa a un **clasificador aleatorio**.

3.2. Implementación

El **objetivo** que se persigue es **entrenar** un **algoritmo de Machine Learning** que aprenda a **discriminar** entre muestras de circRNAs detectados en exosomas de sangre periférica de **personas con cáncer** y muestras de **personas sanas**, en cada uno de los tres grupos de pacientes con cáncer: hepatocelular, colorrectal y pancreático.

Para ello, en **primer lugar**, debemos **filtrar y preparar los datos** de manera que sean adecuados para la aplicación de técnicas de Machine Learning. Filtraremos aquellos circRNAs cuyos conteos de lectura estén por debajo de 10 en valor absoluto en al menos el 70% de las muestras, y utilizaremos la transformación estabilizadora de la varianza (VST) para eliminar la dependencia entre la media y la varianza. Después, **generaremos los conjuntos de entrenamiento y pruebas**, ya que el modelo debe ser **entrenado** en un subconjunto de muestras (entrenamiento) y después **validado** en un subconjunto de muestras diferente (pruebas).

Continuaremos **seleccionando**, mediante un algoritmo de Random Forest, **en el conjunto de entrenamiento,** los **circRNAs** considerados **más importantes** para, a continuación, **generar un modelo de clasificación** utilizando como predictores los circRNAs seleccionados, excepto en el caso del grupo de cáncer hepatocelular, en el que usaremos todos los circRNAs ya que al algoritmo de clasificación escogido (Neural Networks) se comporta mejor en ese caso. Aplicaremos **validación cruzada** para ajustar los **hiperparámetros** y reportar el **rendimiento medio del modelo** (AUC) en el conjunto de entrenamiento.

Para terminar, **aplicaremos el mejor modelo** de clasificación, de entre los generados en el proceso de entrenamiento, **sobre el conjunto de pruebas** para **comprobar** si los **resultados** de entrenamiento son **extrapolables** a un nuevo conjunto de datos, es decir, **si el algoritmo es capaz de discriminar automáticamente** entre las muestras de personas sanas y con cáncer asignando cada muestra su grupo de origen, esta vez en un conjunto de datos distinto al de entrenamiento. Reportaremos el **AUC** de cada uno de los tres modelos (uno por cada tipo de cáncer estudiado) como **medida de su rendimiento**.

La implementación se ha realizado en R y se ha automatizado en el script de R Markdown `Classify.Rmd` disponible en la carpeta "src" del repositorio: http://github.com/carmengmz/circRNA.

Este script se puede ejecutar en cualquier sistema operativo que disponga de un entorno gráfico en el que se pueda instalar Rstudio (https://www.rstudio.com/), un entorno de desarrollo que tiene una edición de libre distribución.

3.2.1. Preparación del conjunto de datos

En primer lugar, **recuperamos los recuentos de lectura** para cada circRNA en cada muestra, generados en el proceso de anotación, a partir del archivo `circ_annotations.rds`, y la **información sobre el grupo** al que pertenece cada muestra, a partir del fichero `phenodata.txt`

La tabla cargada contiene los **recuentos de lectura** para los **117.565 circRNAs detectados** en el total de las **79 muestras**. Adicionalmente, en las **cinco primeras columnas** de la tabla de recuentos de lectura tenemos el identificador interno asociado al circRNA, el cromosoma en el que está ubicado, la coordenada de origen, la coordenada final y la hebra.

A continuación, construimos **tres conjuntos de datos separando los grupos de dos en dos**: muestras de **personas sanas y con cada tipo de cáncer**; y **filtramos** los circRNAs poco expresados, descartando aquellos cuya expresión (en valor absoluto) sea menor a 10 en al menos el 70% de las muestras de ese conjunto.

De aquí en adelante nombraremos los conjuntos de datos como:

- **colorectal**: conjunto que incluye los recuentos de lectura de las muestras de personas sanas y con cáncer colorrectal.
- **hepatocellular**: conjunto que incluye los recuentos de lectura de las muestras de personas sanas y con cáncer hepatocelular.
- **pancreatic**: conjunto que incluye los recuentos de lectura de las muestras de personas sanas y con cáncer pancreático

Como **resultado**, tras el filtrado, obtenemos:

- **1194 circRNAs en el grupo colorectal** correspondientes a 44 muestras, 12 de ellas de cáncer colorrectal y las 32 de personas sanas.

- **1227 circRNAs en el grupo hepatocellular** correspondientes a 53 muestras, 21 de cáncer hepatocelular y las 32 de personas sanas,

- **1046 circRNAs en el grupo pancreatic** correspondientes a 46 muestras, 14 de ellas de cáncer pancreático y las 32 de personas sanas,

3.2.3. Variance Stabilizing Transoformation

Como se expuso en la parte teórica, la **transformación estabilizadora de la varianza** aplicada sobre los **recuentos de lectura** genera una matriz de recuentos para cada grupo que es aproximadamente **homocedástica**, es decir, con una varianza aproximadamente constante para todos los valores de la media. Realizaremos esta transformación usando la función `vst` de la librería DESeq2 [92] con el tipo de ajuste paramétrico.

```
library(DESeq2)
vst <- vst(as.matrix(counts), fitType="parametric")
```

Podemos visualizar como queda la **relación entre la media y la varianza para cada grupo** después de realizar la transformación: observamos como la varianza se mantiene aproximadamente constante para todos los valores de la media (Ilustraciones 18, 19 y 20).

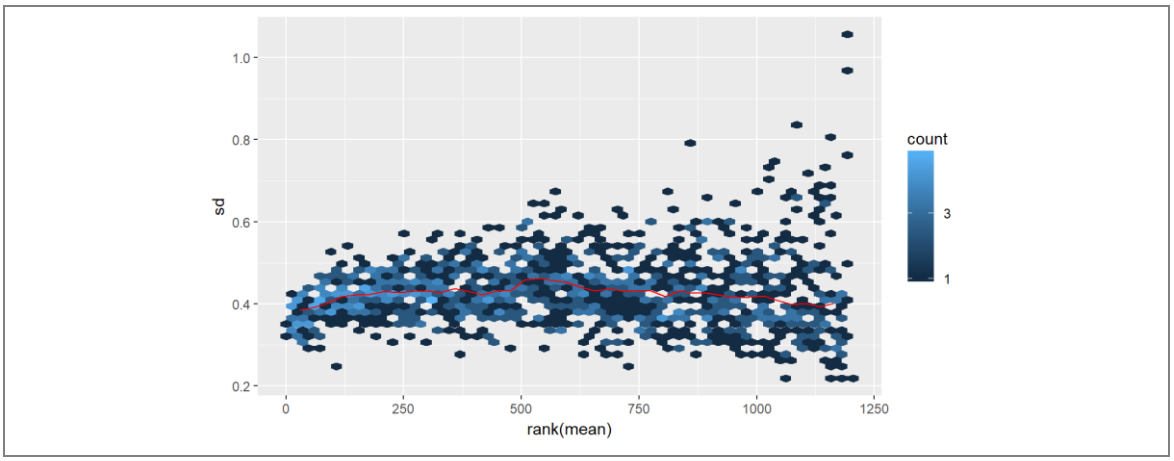

Ilustración 18. Colorectal. Relación entre la media y la varianza tras la transformación vst

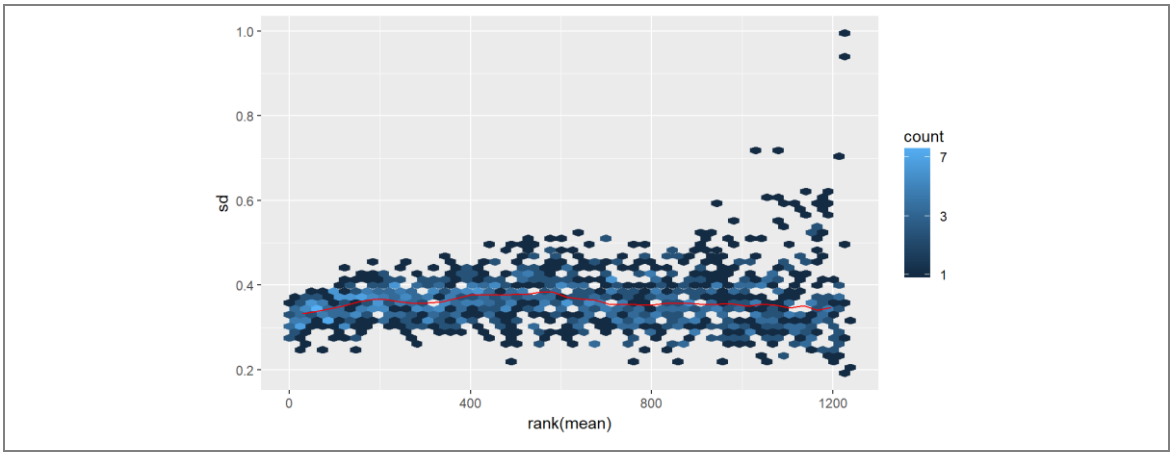

Ilustración 19. Hepatocellular. Relación entre la media y la varianza tras la transformación vst

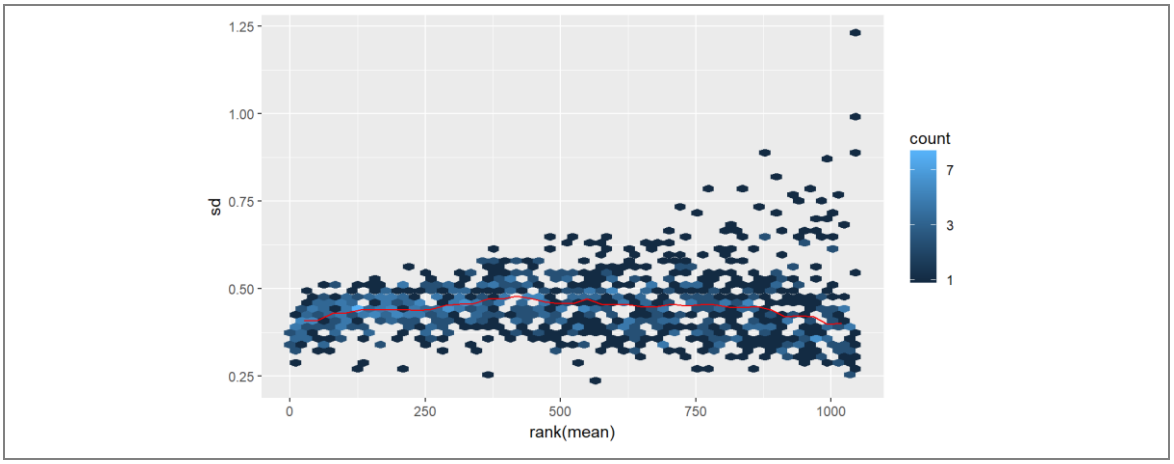

Ilustración 20. Pancreatic. Relación entre la media y la varianza tras la transformación vst

3.2.4. Conjuntos de entrenamiento y pruebas

Ahora **dividimos** cada uno de los tres grupos **en dos conjuntos** cada uno: un primer conjunto de **entrenamiento** (train) que contendrá aproximadamente el 80% de las muestras y un segundo grupo de **pruebas** (test) que contendrá las muestras restantes.

```
dat <- getPartition(as.factor(group), vst, ratio=0.2)
xtrain <- dat.hep$xtrain
xtest <- dat.hep$xtest
ytrain <- factor(dat.hep$ytrain$condition)
ytest <- factor(dat.hep$ytest$condition)
```

De esta manera, **entrenamos el modelo** en **el conjunto de entrenamiento** y **comprobamos su capacidad predictiva** en el de **pruebas**. Procuraremos que la partición esté **equilibrada**, calculando el porcentaje de muestras de cada tipo de individuos (sanos y con cáncer) por separado. Como resultado en la tabla 5 se muestra el **número de muestras por grupo** en cada conjunto de datos (entrenamiento y pruebas)

		Colorectal	Hepatocellular	Pancreatic
Entrenamiento	Normal	25	25	25
	Cancer	9	16	11
	Total	34	41	36
Pruebas	Normal	7	7	7
	Cancer	3	5	3
	Total	10	12	10

Tabla 5. Número de muestras de cada tipo de pacientes en los grupos de entrenamiento y pruebas

3.2.5. Cáncer colorrectal. Entrenamiento y evaluación del modelo de ML.

En primer lugar, realizamos una **selección de predictores** sobre el grupo de **entrenamiento**, para ello ajustaremos un **modelo de Random Forest** al conjunto de entrenamiento recuperando después la importancia de los predictores. Utilizamos la función randomForest de la librería en R con mismo nombre [102] , estableciendo el número de árboles a 1000 (`nTree = 1000`).

```
library(randomForest)
set.seed(12345)

forest.imp = randomForest(class ~. , data = data.frame(t(xtrain), class = ytrain),
                  ntree = 1000, importance = TRUE)

att.scores = as.data.frame(importance(forest.imp, type = 1))
```

Al graficar la **importancia de los predictores** (Ilustración 21) obtenidos al ajustar el modelo en función de su MDI y Gini index, vemos como hay un grupo de 7 u 8 circRNAs que destacan por su importancia sobre el resto, con un MDI por encima de 4 y un Gini index por encima de 0,2.

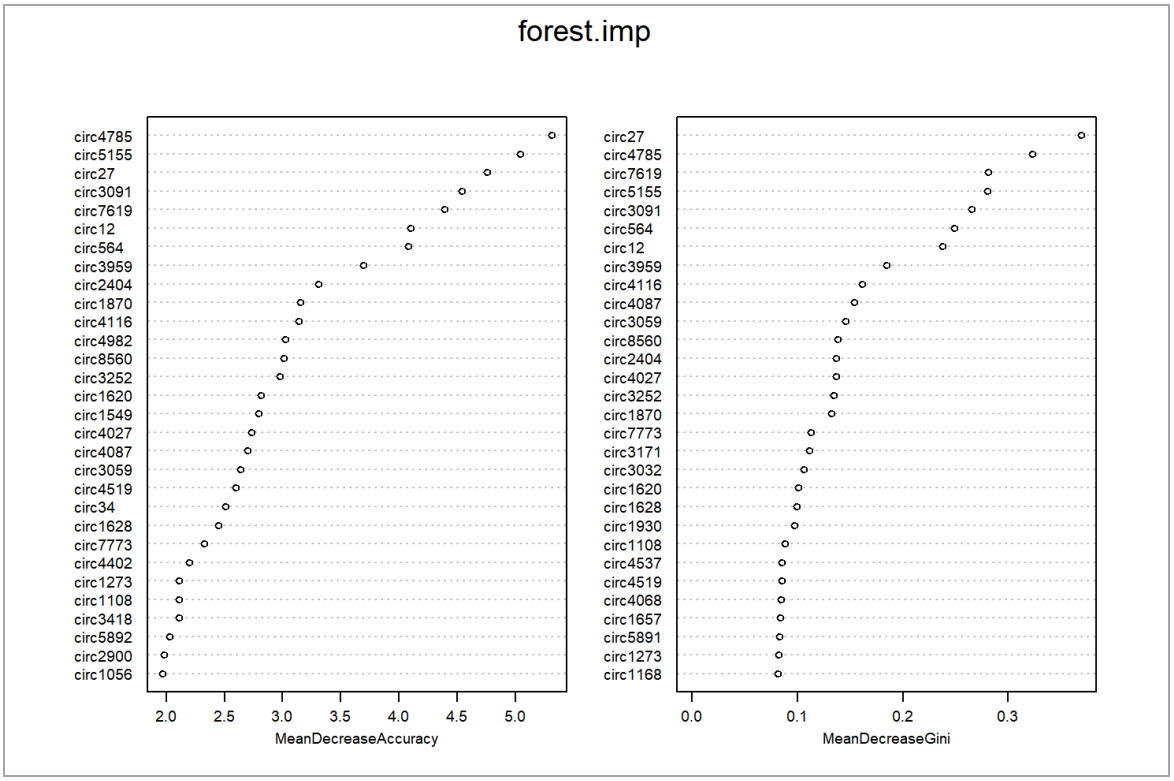

Ilustración 21. Colorectal. Importancia de los predictores de un modelo RF en el conjunto de entrenamiento.

Y en un gráfico de escalado multidimensional (MDS) utilizando la distancia euclídea observamos, en la Ilustración 22, cómo se comportan las muestras si tomamos solamente los ocho primeros circRNAs, separándose naturalmente en grupos.

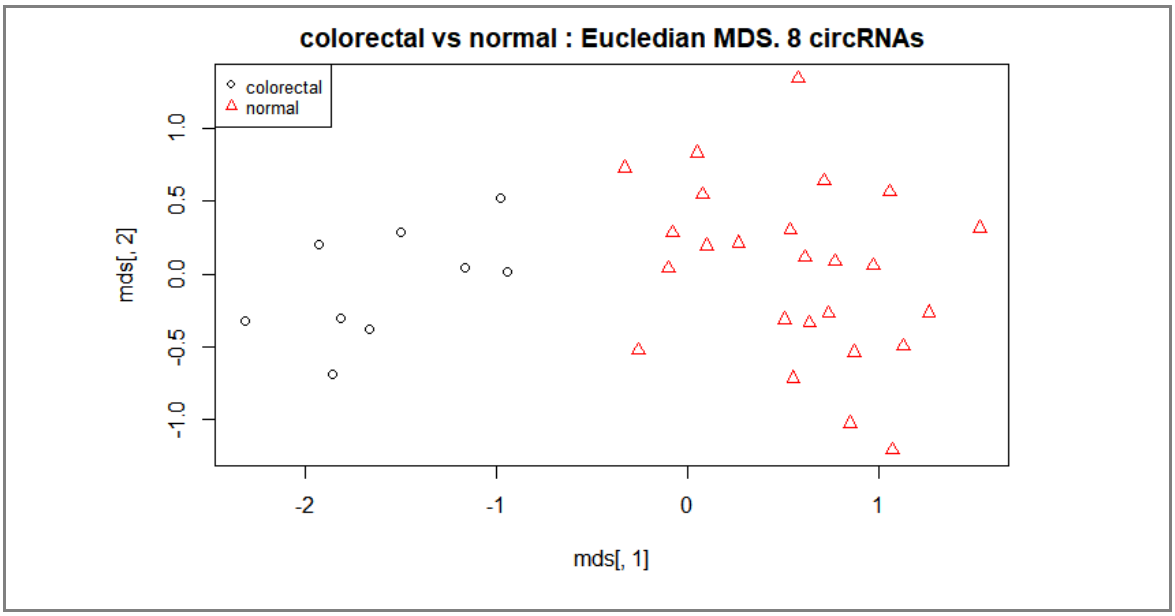

Ilustración 22. Colorectal. Gráfico MDS para conjunto de entrenamiento y los 8 circRNAs más importantes

En base a esta información, vamos a **entrenar un modelo** con un algoritmo de **Support Vector Machines** (SVM) utilizando los **ocho primeros circRNAs** en orden de importancia: circ4785, circ5155, circ27, circ3091, circ7619, circ12, circ564, y circ3959[2]. Si las muestras del grupo de pruebas se comportan de la misma manera, es decir, si la separación que observamos entre muestras en el grupo de entrenamiento es extrapolable al grupo de pruebas, un modelo de SVM con un kernel lineal debería clasificar las muestras con una alta precisión, ya que observamos como las muestras son separables linealmente.

Vamos a utilizar la función `tune.svm` de la librería e1071 [103] que permite ajustar los hiperparámetros y entrenar un modelo SVM usando validación cruzada. Concretamente entrenaremos el modelo con **5-fold cross-validation** y un **kernel lineal.** El hiperparámetro que estamos entrenando (cost) es la penalización que se aplica cuando un punto de datos no está bien clasificado.

```
library(e1071)
set.seed(12345)

svm_tune <- tune.svm(class ~ ., data=data.frame(class=ytrain, t(xtrain.cut)),
                     cost = 2^(0:9), tunecontrol=tune.control(cross=5), kernel="linear")
```

```
Parameter tuning of 'svm':
- sampling method: 5-fold cross validation
- best parameters:
 cost
    1
- best performance: 0
```

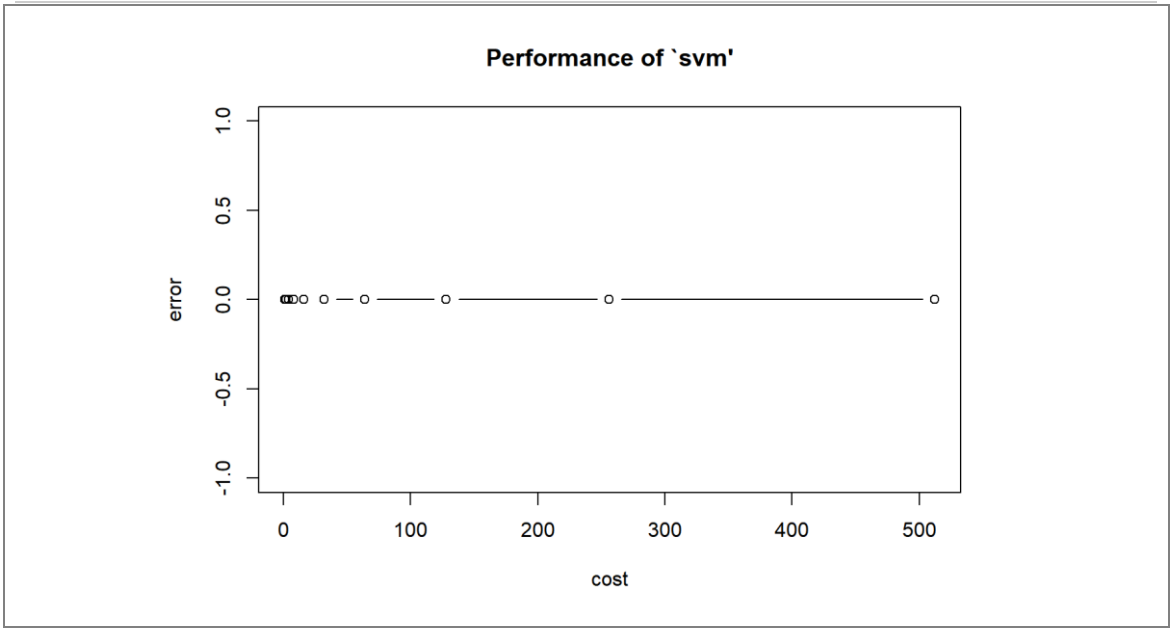

Ilustración 23. Colorectal. Error de entrenamiento del modelo SVM

Obtenemos un **error igual a 0** para las **cinco validaciones** y todos los valores que se han probado para ajustar el coste (Ilustración 23). Era un resultado previsible, que ahora se confirma, debido a la amplia separación entre grupos que observamos en el gráfico MDS.

[2] La información sobre los circRNAs referenciados en este trabajo se puede encontrar en la tabla circ_annotatios.csv del repositorio http://github.com/carmengmz/circRNA/experiment/

A continuación, vamos a ver si el resultado es extrapolable al conjunto de pruebas. Para ello **aplicamos el modelo al conjunto de pruebas** y evaluamos su rendimiento, obteniendo un AUC igual a 1. Todas **las muestras han sido clasificadas correctamente** según el grupo de pacientes del que proceden: sanos y con cáncer colorrectal.

```
library(caret)

pred.svm <- predict(svm_tune$best.model,t(xtest))
confusionMatrix(pred.svm, ytest)
```

```
 Confusion Matrix and Statistics

             Reference
 Prediction   colorectal normal
   colorectal          3      0
   normal              0      7

                Accuracy : 1
                  95% CI : (0.6915, 1)
     No Information Rate : 0.7
     P-Value [Acc > NIR] : 0.02825

                   Kappa : 1
  Mcnemar's Test P-Value : NA

             Sensitivity : 1.0
             Specificity : 1.0
          Pos Pred Value : 1.0
          Neg Pred Value : 1.0
              Prevalence : 0.3
          Detection Rate : 0.3
    Detection Prevalence : 0.3
       Balanced Accuracy : 1.0

        'Positive' Class : colorectal
```

Este resultado nos indica que los circRNAs seleccionados se comportan de la misma manera en el grupo de pruebas y, por tanto, **es posible clasificar muestras** de pacientes con cáncer colorrectal y de pacientes sanos **con una precisión del 100%** utilizando solamente esos 8 circRNAs en nuestra población a estudio.

Por último, dejar constancia de que sobre este grupo se han entrenado también los modelos de Random Forest (sobre los 8 circRNAs) y Neural Networks (sobre todos los circRNAS), obteniendo en todos los casos un AUC igual a 1 en la validación cruzada en el conjunto de entrenamiento y un AUC de 1 en el conjunto de pruebas.

La implementación completa del experimento, Colorectal-Classify.Rmd junto con el informe de salida, Colorectal-Classify.hmtl[3] se encuentra disponible en la carpeta "experiment" del repositorio: http://github.com/carmengmz/circRNA

[3] El informe html se puede visualizar on-line en la url:
https://carmengmz.github.io/circRNA/experiment/Colorectal-Classify.html

3.2.6. Cáncer hepatocelular. Entrenamiento y evaluación del modelo de ML.

De nuevo realizamos una **selección de predictores** sobre el grupo de **entrenamiento** ajustando un **modelo de Random Forest** al conjunto de entrenamiento y recuperando después la importancia de los predictores graficándolos en función de su importancia (Ilustración 24).

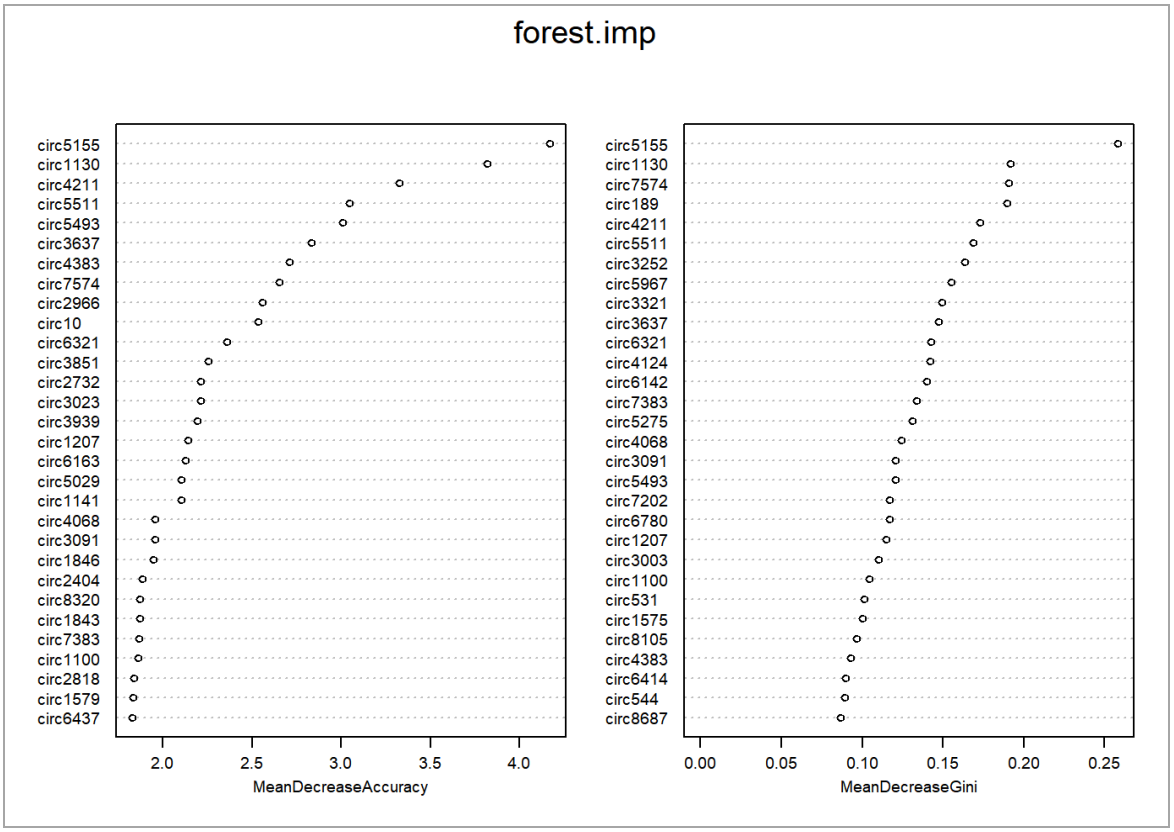

Ilustración 24. Hepatocellular. Importancia de los predictores. Conjunto de entrenamiento.

En el grupo colorectal, los circRNAs utilizados tenían un MDI mayor que 4 y un Gini index mayor que 0,2. Pero en este grupo sólo hay un predictor que cumpla con esos requisitos: el circ5155. Por tanto, en este caso **no tenemos un conjunto tan definido de circRNAs** en función de su importancia. De hecho, debemos tomar los 95 primeros circRNAs (MDI >1.4) para observar cierta separación entre grupos en el conjunto de entrenamiento (Ilustración 25).

Adicionalmente, al ajustar al conjunto de entrenamiento un modelo de Support Vector Machines(SVM) probando dos kernels, uno lineal y otro radial, y otro modelo de Random Forest (RF), obtenemos un AUC de 0.7 utilizando 95 predictores con el modelo RF y de 0.9 (especificidad = 100% y sensibilidad = 80%) utilizando 50 predictores en el modelo SVM, ya que, en el mejor de los casos, tenemos siempre una muestra de cáncer hepatocelular que no es clasificada correctamente en su grupo.

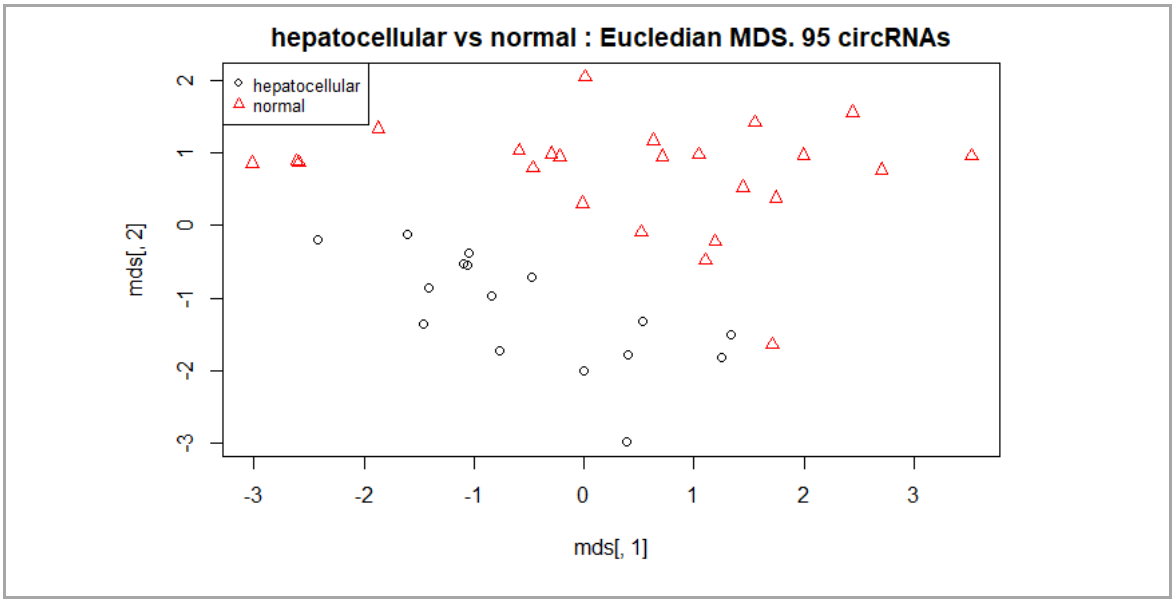

Ilustración 25. Hepatocellular. Gráfico MDS con 95 circRNAs. Conjunto de entrenamiento.

Por otro lado, los modelos también sufren de "overfitting" (sobreajuste) ya que obtenemos, con ambos algoritmos (RF y SVM) un error de entrenamiento igual a cero que no se extrapola a los datos de pruebas.

Todos estos datos nos llevan a pensar que la **complejidad** en el grupo hepatocellular **es mayor** y, por tanto, vamos a entrenar un modelo usando un algoritmo de **Neural Networks**. Este tipo de algoritmos son capaces de **capturar** mejor los **patrones** en conjuntos más **complejos**. Para entrenar el modelo utilizaremos **todos los circRNAs** como predictores y no únicamente un subconjunto como hicimos en el grupo colorectal, ya que este tipo de modelos funcionan mejor con más predictores.

Se ha usado el paquete **elmNN** para entrenar una red neural mediante el algoritmo de Exterme Learning Machine [101]. Los **hiperparámetros** que podemos ajustar en este modelo son la **función de activación** y el número de **nodos ocultos**.

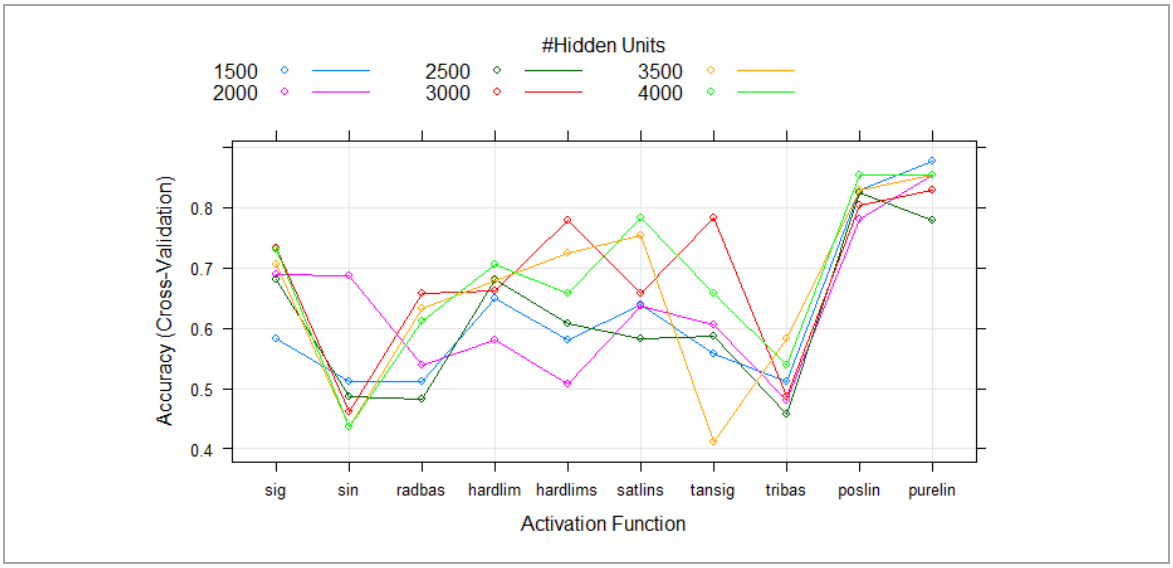

Ilustración 26. Hepatocellular. Búsqueda de los hiperparámetros de la Red Neuronal.

Realizamos una búsqueda en todo el espacio de funciones de activación disponibles y varios valores para hidden units (nodos ocultos), encontrando que la **mejor función** de activación para nuestros datos es "**purelin**" (Ilustración 26). Entrenaremos la red utilizando esta función de activación y restringiendo el espacio de búsqueda de las hidden units al rango 1000-3500. Los valores habituales que suelen clasificar mejor están en el rango del número de predictores. En nuestro caso estamos usando 1.227 predictores (circRNAs).

El **AUC** del mejor modelo obtenido por validación cruzada en el **conjunto de entrenamiento** es de **0,913** usando 2.000 nodos en la capa oculta y la función de activación "poslin" (lustración 27)

```
library(caret)
set.seed(12345)

numFolds <- trainControl(method = 'cv', number = 5, search="grid")
tunegrid <- expand.grid(nhid=c(1000,1500,2000,2500,3000,3500), actfun=c("purelin"))

model.nn <- train(x = t(xtrain), y = as.factor(ytrain), method = "elm",
                  trControl = numFolds, tuneGrid=tunegrid, metric="ROC")
```

```
Extreme Learning Machine
  41 samples
1227 predictors
   2 classes: 'hepatocellular', 'normal'
No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 33, 33, 33, 33, 32
Resampling results across tuning parameters:
  nhid  ROC        Sens       Spec
  1000  0.8200000  0.6166667  0.80
  1500  0.8933333  0.8166667  0.84
  2000  0.9133333  0.8166667  0.88
  2500  0.8800000  0.7500000  0.88
  3000  0.8933333  0.8166667  0.88
  3500  0.8933333  0.8666667  0.84

Tuning parameter 'actfun' was held constant at a value of purelin
ROC was used to select the optimal model using the largest value.
The final values used for the model were nhid = 2000 and actfun = purelin.
```

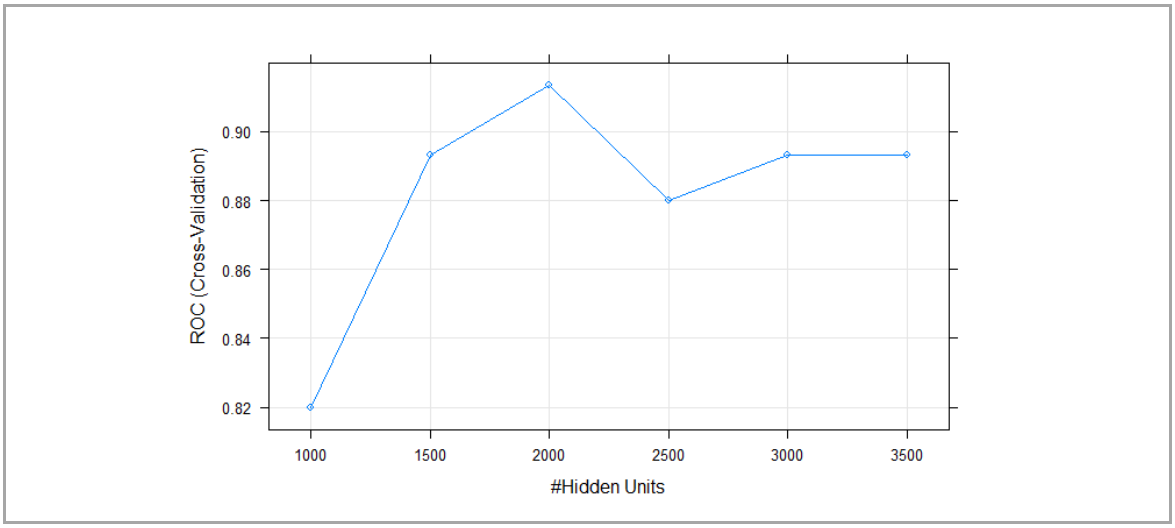

Ilustración 27. Hepatocellular. AUC de entrenamiento de la Red Neuronal.

Y al **evaluar** el modelo en el conjunto de datos de **prueba** obtenemos un **AUC igual a 1**. Este resultado nos dice que es posible, con una **precisión del 100%** en nuestra población, **discriminar** entre muestras de **personas sanas** y con **cáncer hepatocelular** utilizando los circRNAs detectados en exosomas de sangre periférica. Igual que para el cáncer colorrectal, es un resultado prometedor, aunque, en este caso, la complejidad del conjunto de circRNAs es mayor.

```
pred.nn <- predict(model.nn,t(xtest))
confusionMatrix(pred.nn, ytest)
```

```
Confusion Matrix and Statistics

                Reference
 Prediction       hepatocellular normal
   hepatocellular              5      0
   normal                      0      7
                Accuracy : 1
                  95% CI : (0.7354, 1)
     No Information Rate : 0.5833
     P-Value [Acc > NIR] : 0.001552

                   Kappa : 1
  Mcnemar's Test P-Value : NA
             Sensitivity : 1.0000
             Specificity : 1.0000
          Pos Pred Value : 1.0000
          Neg Pred Value : 1.0000
              Prevalence : 0.4167
          Detection Rate : 0.4167
    Detection Prevalence : 0.4167
       Balanced Accuracy : 1.0000

        'Positive' Class : hepatocellular
```

La implementación completa del experimento, Hepatocellular-Classify.Rmd junto con el informe de salida, Hepatocellular-Classify.hmtl[4] se encuentra disponible en la carpeta "experiment" del repositorio: http://github.com/carmengmz/circRNA

[4] El informe html se puede visualizar on-line en la url: https://carmengmz.github.io/circRNA/experiment/Hepatocellular-Classify.html

3.2.7. Cáncer pancreático. Entrenamiento y evaluación del modelo de ML.

Volvemos a realizar una **selección de predictores** sobre el grupo de **entrenamiento** ajustando un **modelo de Random Forest** al conjunto de entrenamiento y recuperando después la importancia de los predictores graficándolos en función de su importancia (Ilustración 28).

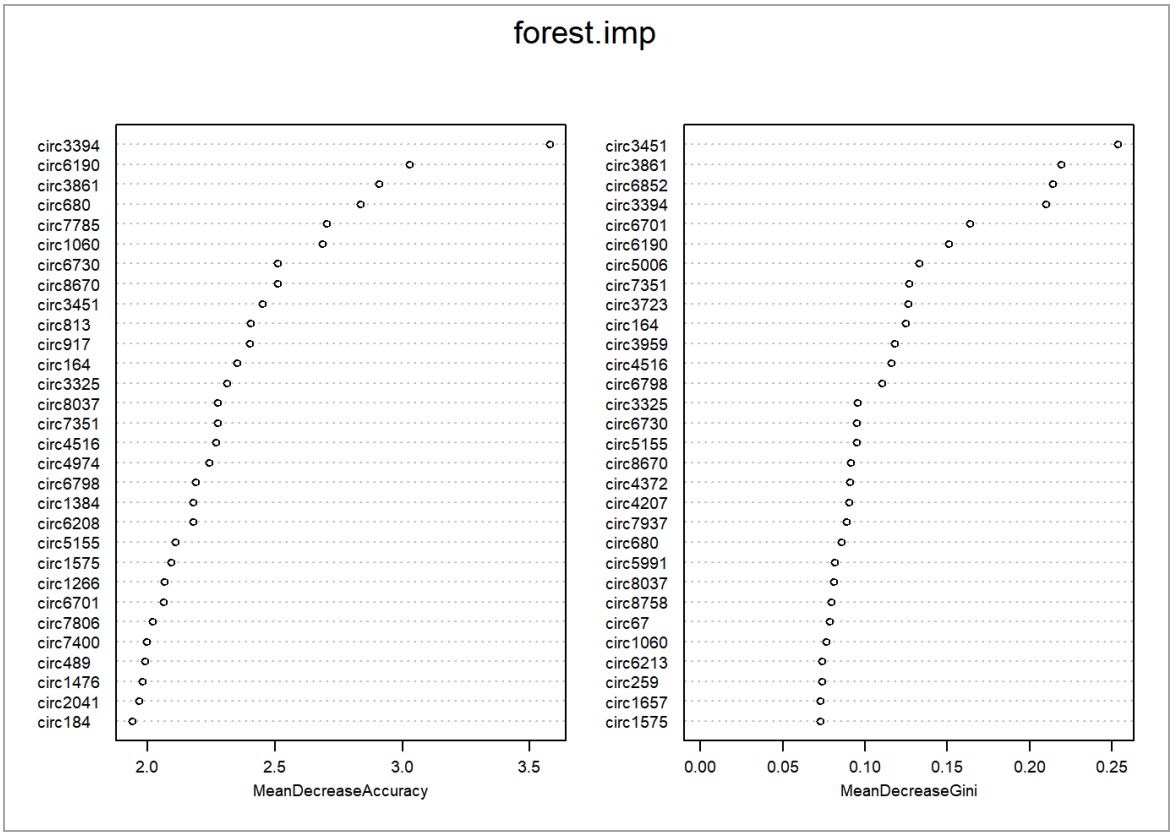

Ilustración 28. Pancreatic. Importancia de los predictores.

En este grupo vemos como el valor mayor del MDI es de 3,5 para un único predictor, el circ3394, el resto de los predictores quedan por debajo de 3,0. Como ocurría en el grupo hepatocellular tenemos una mayor complejidad y debemos tomar unos **50 predictores** para empezar a observar alguna **separación** en grupos el el gráfico de escalado multidimensional (Ilustración 29).

Para este grupo se entrenaron tres modelos usando los algoritmos de Random Forest, Redes Neuronales y Support Vector Machines, en este caso probando un kernel lineal y otro radial. La Red Neuronal y el modelo de Random Forest no fueron capaces de clasificar correctamente ninguna muestra de cáncer pancreático aunque, en entrenamiento, reportaban un AUC igual a 1, es decir, todos los modelos sufrían de sobreajuste.

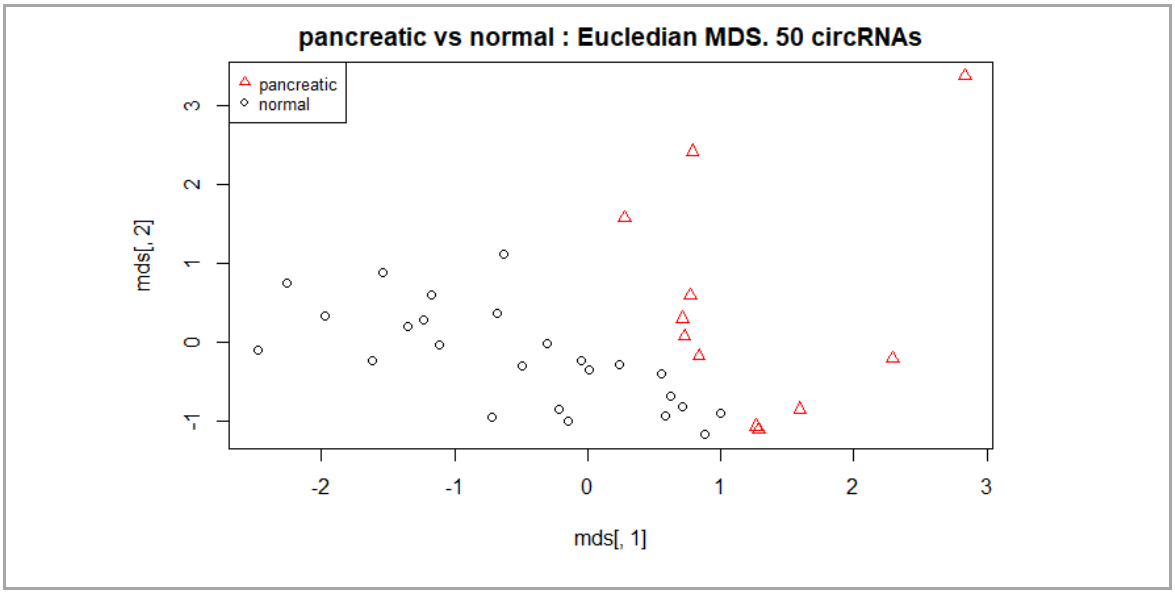

Ilustración 29. Pancreatic. Gráfico MDS con 50 circRNAs. Conjunto de entrenamiento.

El mejor resultado lo obtuvimos entrenando un modelo usando el algoritmo de **Support Vector Machines** con un **kernel radial** que reportó un **error en entrenamiento** de **0,078** usando **5-fold cross-validation** (Ilustración 30).

```
set.seed(12345)
library(e1071)

svm_tune <- tune.svm(class ~ ., data=data.frame(class=ytrain, t(xtrain.cut)),
        cost = c(1,2,3,4,5,10,30), kernel = "radial", tunecontrol=tune.control(cross=5))
```

```
Parameter tuning of 'svm':

 - sampling method: 5-fold cross validation
 - best parameters:
       cost 3
 - best performance: 0.07857143
```

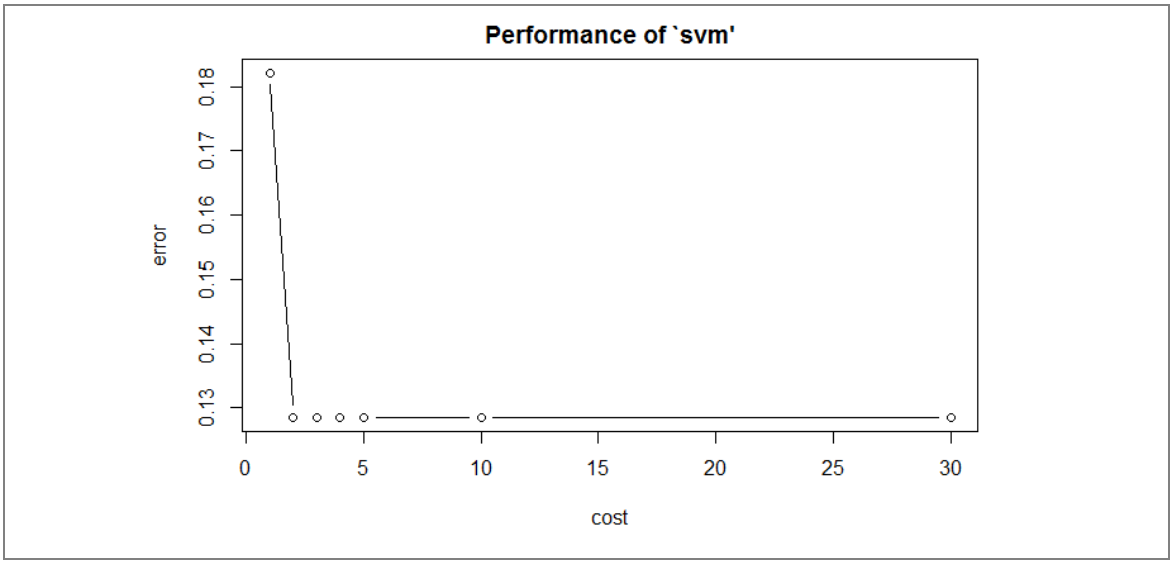

Ilustración 30. Pancreatic. Error de entrenamiento del modelo SVM

El **AUC** obtenido en el **conjunto de pruebas** es igual a **0,833**. Para todos los ajustes de hiperparámetros el mejor modelo siempre clasifica de forma incorrecta una muestra de cáncer pancreático. Como solamente contamos con 3 muestras en el conjunto de pruebas, una sola muestra mal clasificada reduce mucho el rendimiento del modelo.

```
library(caret)
pred.svm <- predict(svm_tune$best.model,t(xtest))
confusionMatrix(pred.svm, ytest)
```

```
Confusion Matrix and Statistics

             Reference
 Prediction   normal pancreatic
   normal          7          1
   pancreatic      0          2

               Accuracy : 0.9
                 95% CI : (0.555, 0.9975)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 0.1493

                  Kappa : 0.7368
 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 1.0000
            Specificity : 0.6667
         Pos Pred Value : 0.8750
         Neg Pred Value : 1.0000
             Prevalence : 0.7000
         Detection Rate : 0.7000
   Detection Prevalence : 0.8000
      Balanced Accuracy : 0.8333

       'Positive' Class : normal
```

```
library(pROC)

roc_obj <- roc(as.numeric(as.factor(ytest)), as.numeric(as.factor(pred.svm)))
auc(roc_obj)
```

```
Area under the curve: 0.8333
```

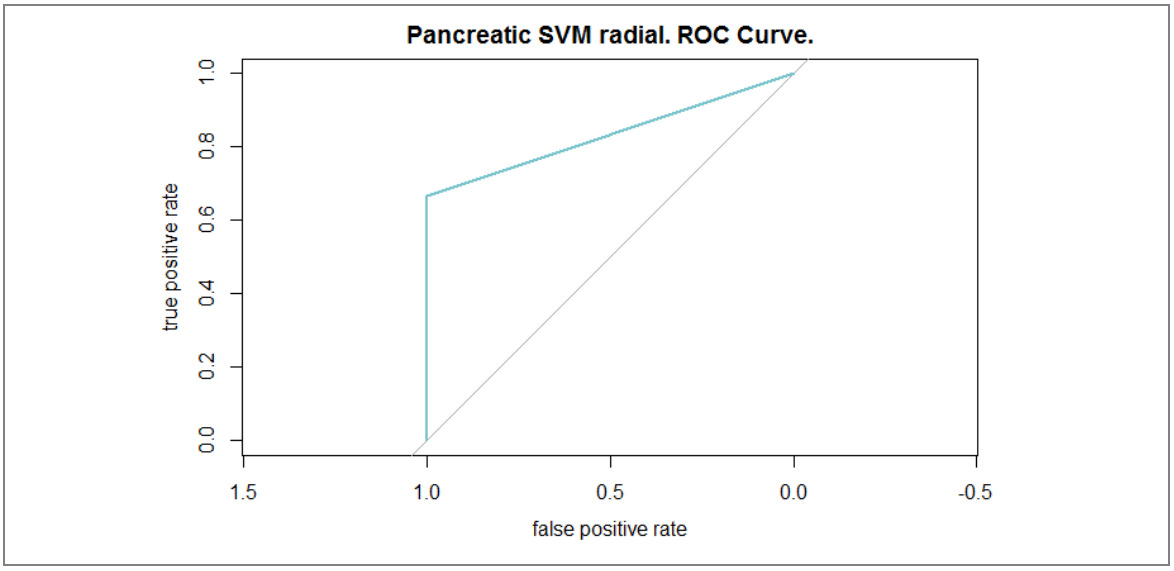

Ilustración 31. Pacnreatic. Curva ROC para el modelo SVM.

Los **resultados** obtenidos nos dan una **sensibilidad del 100%,** pero una **especificidad del 66,67%**, es decir, si un individuo está sano el modelo lo clasifica correctamente, pero si tiene cáncer pancreático, en un 33% de las ocasiones (en nuestro caso para una muestra) no lo detecta y lo clasifica como sano. Por tanto, hemos sido capaces de discriminar entre pacientes sanos y con cáncer usando los circRNAs contenidos en exosomas de sangre periférica para cáncer pancreático, pero con una precisión balanceada del 83,33%.

La implementación completa del experimento, Pancreatic-Classify.Rmd junto con el informe de salida, Pancreatic-Classify.hmtl[5] se encuentra disponible en la carpeta "experiment" del repositorio: http://github.com/carmengmz/circRNA

[5] El informe html se puede visualizar on-line en la url: https://carmengmz.github.io/circRNA/experiment/Pancreatic-Classify.html

Capítulo 4

Identificación de biomarcadores

En el capítulo anterior **entrenamos** un **modelo de Machine Learning**, para cada tipo de cáncer, **capaz de discriminar**, con una **alta precisión**, entre **muestras de individuos sanos y con cáncer**. Este hecho **confirma la hipótesis** de que, en nuestra población a estudio, los **circRNAs** detectados en **exosomas** de **sangre periférica se expresan de forma diferente** en individuos **sanos** y con **cáncer**, lo que los convierte en **biomarcadores prometedores**.

Para **caracterizar el rol** que pueden tener los **circRNAs más relevantes** de cada grupo en cada tipo de cáncer a estudio, en primer lugar, vamos a **determinar**, utilizando un **algoritmo de Random Forest**, cuáles son los circRNAs **más importantes**, esta vez utilizando el **conjunto completo** de recuentos de lectura.

Anteriormente, para entrenar el modelo de Machine Learning de clasificación, dividíamos cada grupo en un conjunto de entrenamiento y otro de pruebas y, para recuperar la importancia de los circRNAs, ajustábamos un modelo de Random Forest sólo en el conjunto de entrenamiento.

En esta ocasión, con el fin de seleccionar los circRNAs más relevantes en cada cáncer, se han ajustado **10^5 modelos de clasificación**, pero utilizando el **conjunto completo** de recuentos de lectura de cada grupo, recuperando, a partir de cada uno de los 10^5 modelos, el

circRNA considerado más importante en cada ocasión. Por tanto, los **circRNAs que vamos a caracterizar** son los que han sido **considerados el más importante** por el conjunto de modelos **un mayor número de veces**.

Ahora todos los circRNAs pasan a formar parte del conjunto de entrenamiento, generando hipótesis sobre los circRNAs más relevantes en nuestra población a estudio, que deberían ser validadas en trabajos posteriores sobre poblaciones diferentes.

Para **cada circRNA,** de los considerados **relevantes** en cada grupo, hemos realizado una **revisión de las evidencias** disponibles en la literatura sobre su **implicación** en cada tipo de **cáncer** estudiado, ya que sabemos que los circRNAs pueden estar implicados en el desarrollo y evolución del cáncer [104].

Además, recientemente se ha descubierto que los **circRNAs pueden unirse a los micro RNAs** (miRNAs) actuando como **esponjas** y, consecuentemente, **desregular o suprimir su función** [3]. Los **miRNAs** son pequeños ARNs (de 21 a 23 nucleótidos) que actúan como **reguladores post-transcripcionales** de la expresión génica emparejándose directamente a sitios diana en regiones no traducidas de ARN mensajeros [105]. Debido a que los miRNAs **controlan** un amplio conjunto de **procesos biológicos**, la actividad de los **circRNAs** como **esponjas** de miRNAs puede **afectar a estos procesos**.

Para los circRNAs considerados más importantes **hemos recopilado** de la "Cancer-Specific CircRNA Database" [26] **los miRNAs diana** sobre los que podrían estar haciendo de esponja y se **ha revisado la literatura** para buscar **evidencias** de si la **desregulación** de esos miRNAs pudiera estar **implicada** en los cánceres de interés.

Por último, también se ha llevado a cabo un **análisis de expresión diferencial**[6] usando la librería **DESeq2** [92] para determinar si la **diferencia de expresión** observada en las gráficas de recuentos de lectura normalizados de los distintos circRNAs es **estadísticamente significativa** (con un nivel de significación de al menos el 10%, $\alpha = 0{,}1$). Reportaremos el p-valor ajustado para contrastes múltiples (p-adj).

Y, aunque hemos comprobado que la **expresión** de los **circRNAs** en **cáncer pancreático** es **suficientemente diferente** a la de personas sanas como para **clasificar** las muestras de **manera automática** con una precisión balanceada del 83,33%, para este grupo **el test estadístico no reportó circRNAs diferencialmente expresados** de manera significativa (todos los p-valores ajustados eran mayores a 0,8), por lo que no se reportarán los p-valores, y sólo se describirá la expresión relativa entre grupos observada en las gráficas de recuentos de lectura normalizados.

[6] El informe del análisis de expresión diferencial puede ser consultado on-line en: https://carmengmz.github.io/circRNA/experiment/DifferentialExpression.html y el código fuente se encuentra en la carpeta "experiment" del repositorio: https://github.com/carmengmz/circRNA

4.1. ARNs circulares más relevantes en cáncer colorrectal y hepatocelular

En **cáncer colorectal** encontramos que el circRNA **circ5155** fue considerado el más importante por 6.985 modelos (70%) de los 10.000 entrenados, seguido por el **circ4785** que fue considerado el más importante en 2.439 modelos (24%). El número de circRNAs considerados más importantes por algún modelo es de 6 (Ilustración 32), confirmando el hecho de la **baja complejidad** que ya observábamos en este conjunto. Además, los circRNAs considerados aquí los más relevantes se encontraban entre los predictores considerados más importantes en el conjunto de entrenamiento en el capítulo anterior.

Adicionalmente, excepto el circ1870, que además fue considerado importante sólo en una ocasión, el resto de los circRNAs se encuentran **diferencialmente expresados** (sobre expresados) de manera **estadísticamente significativa** (p-adj < 0.05) con respecto al grupo de personas sin enfermedad.

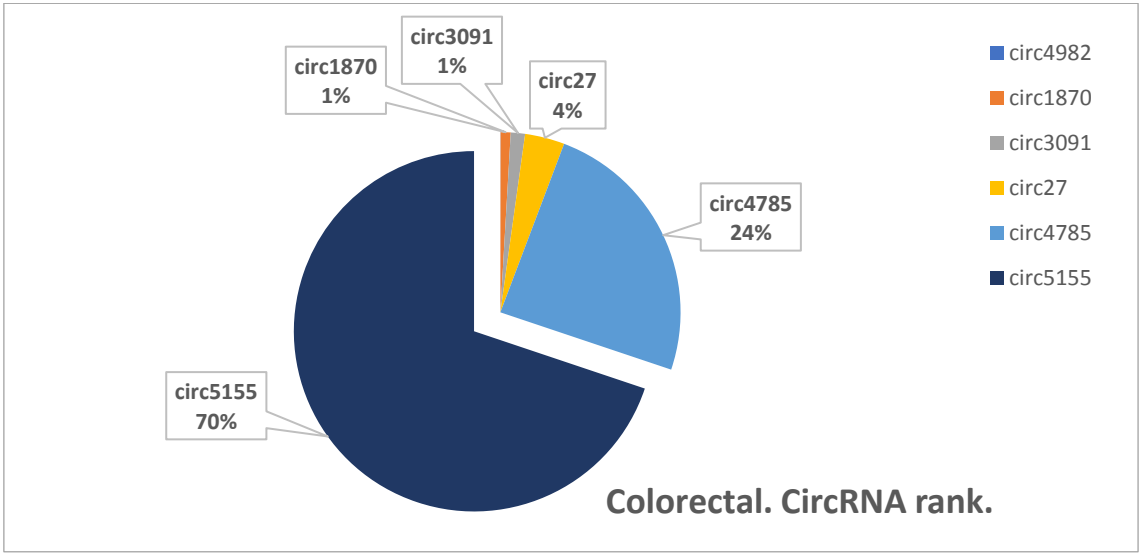

Ilustración 32. CircRNAs más relevantes para cáncer colorrectal.

En **cáncer hepatocelular encontramos** que el mismo circRNA **circ5155** destaca claramente sobre el resto, considerado el más importante en 6.067 modelos (61%). El siguiente en importancia es el **circ3637**, considerado el más importante por 1.076 modelos (11%). Un menor número de veces fueron considerados como los más importantes 29 circRNAs adicionales (Ilustración 33), confirmando así la **dispersión y complejidad** que observábamos en este grupo al ajustar el modelo de clasificación.

Adicionalmente, el circ5155 fue considerado el más importante en el conjunto de entrenamiento al seleccionar los predictores más relevantes para entrenar el modelo de clasificación. Destacar también que en el grupo de **cáncer hepatocelular** la **expresión diferencial** de los circRNAs revisados **no suele estadísticamente significativa**, excepto en el caso del circ5155.

Vamos a revisar la **implicación del circ5155** en **cáncer colorrectal y hepatocelular** conjuntamente, ya que ha sido considerado el más relevante en ambos grupos. Además, el hecho de que un mismo circRNA haya sido considerado el más importante para los dos tipos de cáncer sugiere que, de alguna manera, los mecanismos celulares y moleculares de ambos cánceres podrían estar relacionados.

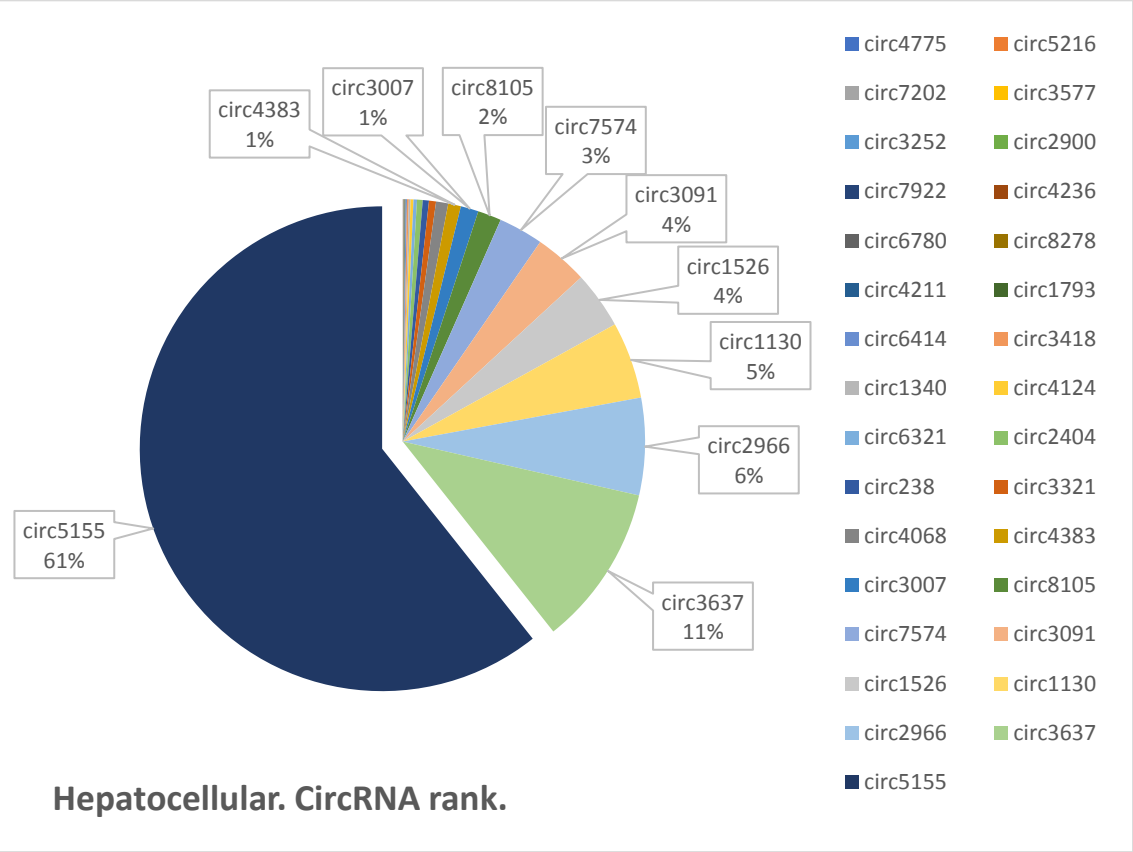

Ilustración 33. CircRNAs más relevantes para cáncer hepatocelular

4.1.1. circ5155 (hsa_circ_0001190)

El **circ5155** es un circRNA que se transcribe a partir de los exones 3, 4, 5 y 6 del gen **DYRK1A** (Ilustración 34), en el **cromosoma 21**. Sus **coordenadas** (GRCh38) son chr21:37420298-37421866. En circBase recibe el nombre **hsa_circ_0001190** y en los ensayos con microarrays, hsa_circRNA_001826.

El **gen DYRK1A** pertenece a la **familia** de **quinasas reguladas por fosforilación de tirosina de doble especificidad** (DYRK, dual-specificity tyrosine phosphorylation-regulated kinase) que ha surgido recientemente como una **nueva diana terapéutica** para diferentes tipos de **cáncer** y **enfermedades neurodegenerativas** [106]. En estudios recientes se ha demostrado que los DYRK juegan un **papel clave** en las enfermedades **neurodegenerativas**, la **supervivencia** de las células cancerosas, y su **inhibición** induce la

apoptosis de las células cancerosas. En concreto el **DYRK1A**, dependiendo del **contexto celular**, se sabe que funciona tanto como un **supresor tumoral** o como un **oncogen** [107]. Este gen **promueve la supervivencia celular** y se encuentra **sobre expresado** en muchos tipos de cáncer, sin embargo, en otros, como en la leucemia mieloide aguda, actuaría como supresor tumoral y se encuentra sub expresado [108].

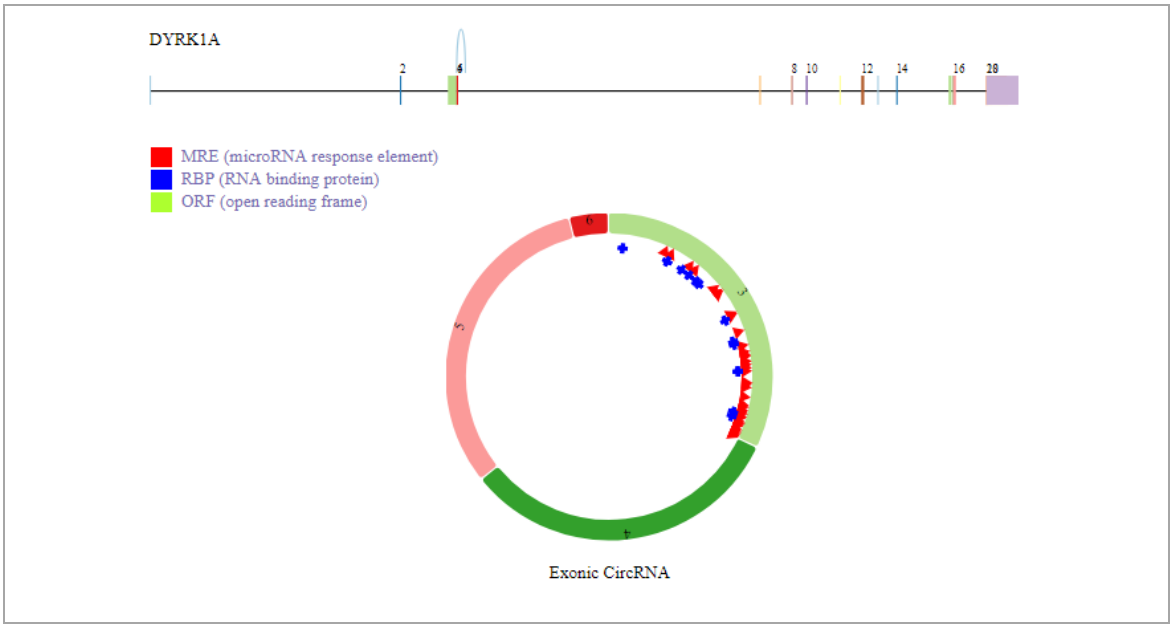

Ilustración 34. Estructura del circRNA hsa_circ_0001190 (circ5155)

En nuestro conjunto de muestras a estudio, el **circ5155** aparece **sobre expresado** en pacientes con **cáncer colorrectal** (p-adj≈0) **y hepatocelular** (p-adj = 0.0939) con respecto a la expresión en personas sanas, y **menos expresado** en **cáncer pancreático** (Ilustración 35). El papel que pueda jugar este circRNA en el desarrollo y progresión de estos tipos de cáncer aún no está descrito, encontrando una única referencia en la literatura sobre su **baja expresión en cáncer gástrico** [109]**.**

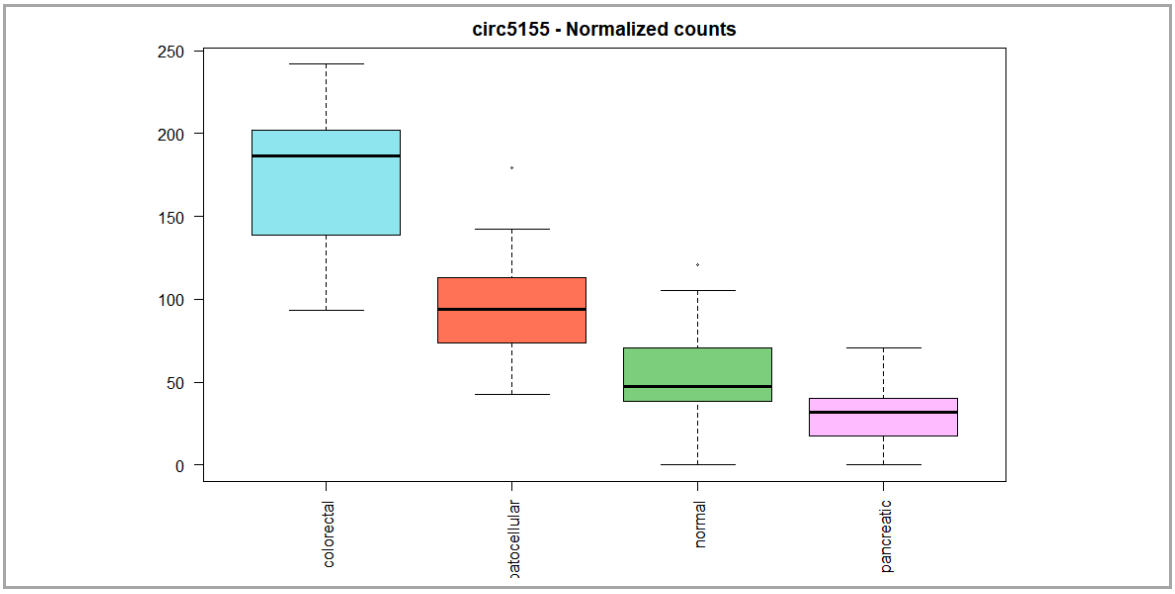

Ilustración 35. Box-plot con los recuentos de lectura normalizados del circ5155 en cada grupo.

Adicionalmente sabemos que los **circRNAs** pueden actuar como **esponjas de miRNA**. Se Recuperamos de la CSCD [26] los **miRNA diana** de este circRNA y realizamos una búsqueda en la literatura encontrando abundantes **evidencias** de que la **desregulación** de muchos de ellos está **implicada** en la **progresión del cáncer** colorrectal y hepatocelular. En concreto, encontramos que existen **115 miRNAs** diana para el **circ5155**, encontrando **evidencias** de la **relación** de **18** de estos miRNAs con **cáncer colorrectal** y **15** con **cáncer hepatocelular**.

Llama la atención que casi todos los **circRNAs más importantes** detectados para **cáncer hepatocelular y colorrectal** (excepto uno en cáncer hepatocelular), están **sobre expresados** (Ilustraciones 35, 36 y 37), y todos los **miRNAs** con los que podría **interaccionar** el circ5155, para los que hemos encontrado **estudios publicados**, son **oncoprotectores**. Esto significaría que, si los **circRNAs sobre expresados** que estamos estudiando **capturan** a estos miRNAs, harían **desparecer su efecto oncoprotector**, favoreciendo el desarrollo y progresión de la enfermedad. Se trata de una hipótesis prometedora que requeriría más trabajo para ser confirmada, pero cuya confirmación quizás abriría la puerta a nuevas terapias.

En la siguiente tabla (Tabla 6) detallamos la **revisión de la literatura** sobre la implicación en **cáncer colorrectal** de los **miRNA diana del circ5155**.

miRNA	**Evidencia para cáncer colorrectal**	**Ref.**
hsa-miR-138	"Down-regulation of miR-138 promotes colorectal cancer metastasis via directly targeting TWIST2."	[110]
	"The tumor suppressor miR-138-5p targets PD-L1 in colorectal cancer"	[111]
hsa-miR-149	"miR-149 in Human Cancer: A Systemic Review"	[112]
	"MicroRNA-149 Increases the Sensitivity of Colorectal Cancer Cells to 5-Fluorouracil by Targeting Forkhead Box Transcription Factor FOXM1."	[113]
hsa-miR-4728	"miR-4728-3p Functions as a Tumor Suppressor in Ulcerative Colitis-associated Colorectal Neoplasia Through Regulation of Focal Adhesion Signaling."	[114]
hsa-miR-6883 hsa-miR-6785	"miR-6883 Family miRNAs Target CDK4/6 to Induce G1 Phase Cell-Cycle Arrest in Colon Cancer Cells."	[115]
hsa-miR-17-5p	"Elevated oncofoetal miR-17-5p expression regulates colorectal cancer progression by repressing its target gene P130 "	[116]
hsa-miR-93	"MicroRNA-93 suppress colorectal cancer development via Wnt/β-catenin pathway downregulating."	[117]
	"miR-93 suppresses proliferation and colony formation of human colon cancer stem cells"	[118]
	"MicroRNA-93 inhibits tumor growth and early relapse of human colorectal cancer by affecting genes involved in the cell cycle"	[119]

hsa-miR-519	"miR-519 reduces cell proliferation by lowering RNA-binding protein HuR levels."	[120]
	"miR-519 suppresses tumor growth by reducing HuR levels"	[121]
hsa-mir-198	"MiR-198 represses tumor growth and metastasis in colorectal cancer by targeting fucosyl transferase "	[122]
hsa-miR-363	"The miR-363-GATA6-Lgr5 pathway is critical for colorectal tumourigenesis."	[123]
	"MiR-363-3p inhibits the epithelial-to-mesenchymal transition and suppresses "metastasis in colorectal cancer by targeting Sox4	[124]
hsa-miR-4510	"Identification and Validation of Potential Biomarkers for the Detection of Dysregulated microRNA by qPCR in Patients with Colorectal Adenocarcinoma"	[125]
hsa-miR-4533	"Infrequently expressed miRNAs in colorectal cancer tissue and tumor molecular phenotype"	[126]
hsa-miR-491	"Functional screening identifies a microRNA, miR-491 that induces apoptosis by targeting Bcl-X(L) in colorectal cancer cells."	[127]
hsa-miR-5582	"Novel miR-5582-5p functions as a tumor suppressor by inducing apoptosis and cell cycle arrest in cancer cells through direct targeting of GAB1, SHC1, and CDK2."	[128]
hsa-miR-625	"Decreased expression of microRNA-625 is associated with tumor metastasis and poor prognosis in patients with colorectal cancer."	[129]
hsa-miR-6803	"Exosomal miR-6803-5p as potential diagnostic and prognostic marker in colorectal cancer"	[130]
hsa-miR-7	"miR-7 inhibits colorectal cancer cell proliferation and induces apoptosis by targeting XRCC2"	[131]
	"microRNA-7 is a novel inhibitor of YY1 contributing to colorectal tumorigenesis."	[132]
hsa-miR-92	"miR-92 is a key oncogenic component of the miR-17-92 cluster in colon cancer.	[133]
	MicroRNA regulation network in colorectal cancer metastasis"	[134]

Tabla 6. Mi-RNAs diana del circ5155 cuya desregulación está implicada en cáncer colorrectal.

Y esta otra tabla (Tabla 7) se detalla la revisión sobre la **implicación de los miRNAs diana** del **circ5155** en el **cáncer hepatocelular**.

miRNA	**Evidencia para cáncer hepatocelular**	**Ref.**
hsa-miR-1249	"Identification of pathogenesis-related microRNAs in hepatocellular carcinoma by expression profiling"	[135]
	"Induced MiR-1249 expression by aberrant activation of Hedegehog signaling pathway in hepatocellular carcinoma."	[136]
hsa-miR-1293	"Identification of Recurrence-Related microRNAs from Bone Marrow in Hepatocellular Carcinoma Patients"	[137]
hsa-miR-138	"MicroRNA-138 targets SP1 to inhibit the proliferation, migration and invasion of hepatocellular carcinoma cells"	[138]
	"MicroRNA-138 inhibits cell proliferation in hepatocellular carcinoma by targeting Sirt1."	[139]
hsa-miR-149	"miR-149 represses metastasis of hepatocellular carcinoma by targeting actin-regulatory proteins PPM1F."	[140]
hsa-miR-17-5p	"miR-17-5p as a novel prognostic marker for hepatocellular carcinoma"	[141]
hsa-miR-519	"miR-519 suppresses tumor growth by reducing HuR levels"	[121]
hsa-miR-198	"miR-198 inhibits migration and invasion of hepatocellular carcinoma cells by targeting the HGF/c-MET pathway"	[142]
hsa-miR-21	"miR-21 expression predicts prognosis in hepatocellular carcinoma"	[143]
hsa-miR-342	"miR-342-3p affects hepatocellular carcinoma cell proliferation via regulating nF-ϰB pathway"	[144]
hsa-miR-363	"miR-363 inhibits the growth, migration and invasion of hepatocellular carcinoma cells by regulating E2F3."	[145]
hsa-miR-5582	"Novel miR-5582-5p functions as a tumor suppressor by inducing apoptosis and cell cycle arrest in cancer cells through direct targeting of GAB1, SHC1, and CDK2"	[128]
hsa-miR-608	"MiR-608 rs4919510 is associated with prognosis of hepatocellular carcinoma."	[146]
hsa-miR-625	"miR-625 suppresses tumour migration and invasion by targeting IGF2BP1 in hepatocellular carcinoma."	[147]
hsa-miR-7	"MicroRNA-7 inhibits tumor growth and metastasis by targeting the phosphoinositide 3-kinase/Akt pathway in hepatocellular carcinoma."	[148]
	"Quantitative Proteomics Reveals the Regulatory Networks of Circular RNA CDR1as in Hepatocellular Carcinoma Cells."	[149]
hsa-miR-92	"Deregulation of miR-92a expression is implicated in hepatocellular carcinoma development."	[150]

Tabla 7. Mi-RNAs diana del circ5155 cuya desregulación está implicada en cáncer hepatocelular.

4.1.2. Otros circRNAs considerados relevantes en cáncer colorrectal

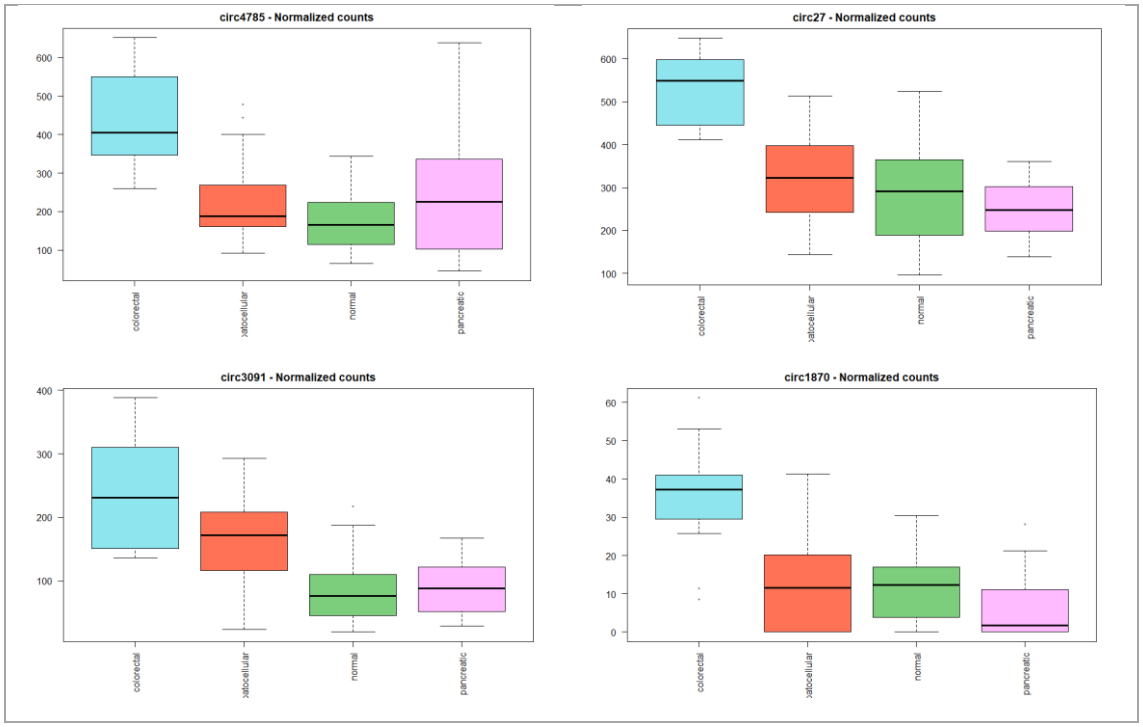

Ilustración 36. Otros circRNAs relevantes para cancer colorrectal.

Otros circRNAs que el modelo de Random Forest consideró **relevantes para cáncer colorrectal** (Ilustración 36) se detallan a continuación:

Circ4785: en circBase recibe el nombre de **hsa_circ_0001092** y hsa_circRNA_102894 en los ensayos con microarrays. Se transcribe a partir del **gen CFLAR** en el **cromosoma 2**, y sus **coordenadas** GRCh38g son chr2:201145378-201149835. En el gráfico podemos ver que está muy **sobre expresado** (p-adj≈0) en **cáncer colorrectal**. Encontramos una única referencia a este circRNA en la literatura: en este estudio [151] se reporta que el hsa_circ_0001092 está sobre expresado en cáncer gástrico.

Circ27: en circBase recibe el nombre de **hsa_circ_0006354**. Se transcribe a partir del **gen VAMP3** en el **cromosoma 1**. Sus **coordenadas** GRCh38g son chr1:7777160-7778169. En el gráfico observamos que está muy **sobre expresado** en cáncer colorrectal (p-adj = 0.0008). No encontramos referencias en la literatura sobre este circRNA.

Circ3091: no está catalogado en circBase. Es un circRNA que se transcribe a partir de una **región intergénica** en el **cromosoma 15**. Sus **coordenadas** GRCh38g son chr15:100564688-100565265. En el gráfico observamos que está **sobre expresado** para cáncer colorrectal (p-adj≈0).

Circ1870: no está catalogado en circBase. Se transcribe a partir del **gen BIN2** en el **cromosoma 12**. Sus **coordenadas** GRCh38g son chr12:51299203-51299714. En el gráfico observamos que tiene una expresión muy baja, estando **más expresado para cáncer colorrectal**, aunque no de forma significativa (p-adj = 0.33)

4.1.3. Otros circRNAs considerados relevantes en cáncer hepatocelular

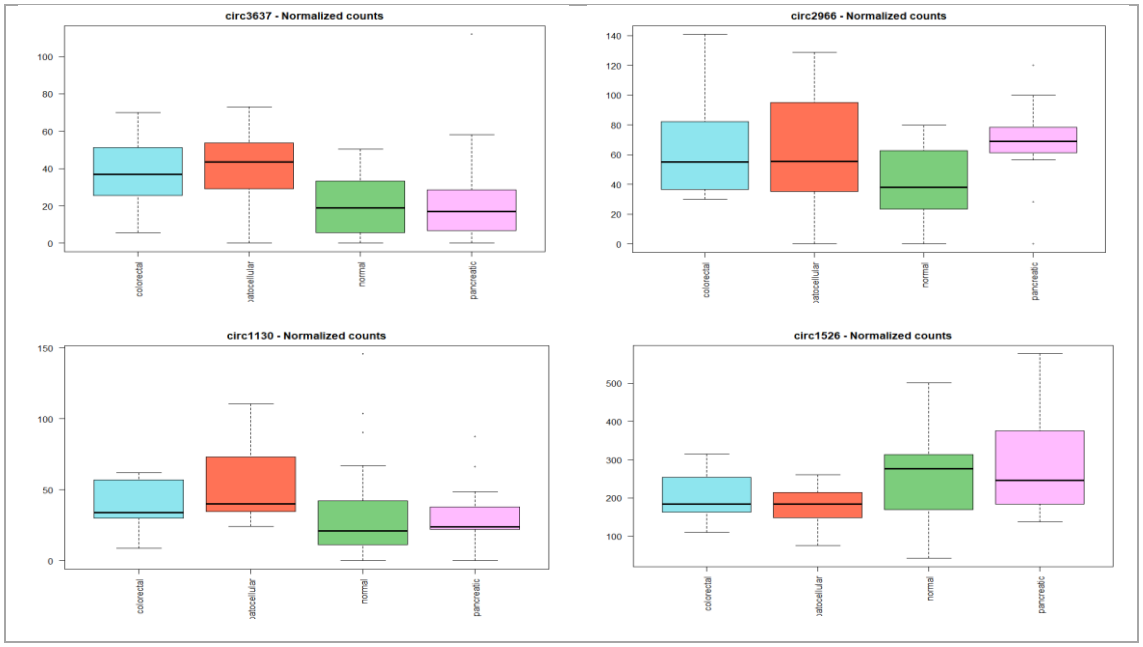

Ilustración 37. Otros circRNAs relevantes para cáncer hepatocelular.

En el caso de **cáncer hepatocelular** otros circRNAs que el modelo de Random Forest consideró **relevantes** (Ilustración 37) son:

Circ3637: recibe el nombre de **hsa_circ_0045272** en circBase y hsa_circRNA_102165 en ensayos con microarrays. Se transcribe a partir del **gen ERN1** en el **cromosoma 17**. Sus **coordenadas** GRCh38g son chr17:64052780-64053371. En la literatura encontramos un estudio en el que se describe su potencial función en **lupus eritematoso** [152]. En la gráfica vemos qué tiene una expresión muy baja para personas sanas y con cáncer pancreático, estando más expresado para cáncer colorrectal y hepatocelular, aunque no de forma significativa (p-adj > 0.1).

Circ2966: recibe el nombre **hsa_circ_0035957** en circBase y hsa_circRNA_101566 en ensayos con microarrays. Se transcribe a partir del **gen DENND4A** en el **cromosoma 15**. Sus **coordenadas** GRCh38g son chr15:65752379-65761438. No encontramos referencias en la literatura sobre este circRNA y su expresión es aparentemente mayor para los tres tipos de cáncer que las personas sanas, pero no es significativa (p-adj > 0.1).

Circ1130: recibe el nombre de **hsa_circ_0000239** en circBase. Se transcribe a partir del **gen RUFY2** en el cromosoma 10. Sus **coordenadas** GRCh38g son chr10:68393138-68394451. En la gráfica vemos que está **algo más expresado** para cáncer hepatocelular, pero no de forma significativa (p-adj > 0.1).

Circ1526: recibe el nombre de **hsa_circ_0000339** en circBase y hsa_circRNA_100884 en los ensayos con microarrays. Se transcribe a partir del **gen RAB6A** en el **cromosoma 11**. Sus **coordenadas** GRCh38g son chr11:73707420-73718718. Es el primero, de entre los considerados más importantes, que **aparece menos expresado** para cáncer colorectal y hepatocelular, con respecto a las personas sanas, aunque esta diferencia no es significativa estadísticamente en ninguno de los dos grupos (p-adj > 0.1).

4.2. ARNs circulares más relevantes en cáncer pancreático

En este grupo, al igual que ocurría en cáncer hepatocelular, encontramos un **conjunto amplio de circRNAs** (Ilustración 38) que han sido **considerados los más importantes** en alguna de las 10^5 ejecuciones del algortimo Random Forest. De entre todos estos circRNAs **destacan** el **circ3602**, que se consideró el más importante en 4.034 ocasiones (40%), el **circ3861**, considerado el más importantes en 3.467 ocasiones (35%) y el **circ164** considerado el más importante en 1.935 ocasiones (19%).

Como ya se mencionó al inicio, no obtuvimos ningún circRNA cuyo p-valor ajustado fuera menor a 0.1 en la expresión diferencial, por lo que las diferencias de expresión que se comentan en este apartado responden solamente a la observación en el gráfico de los recuentos de lectura normalizados de cada circRNA en cada grupo.

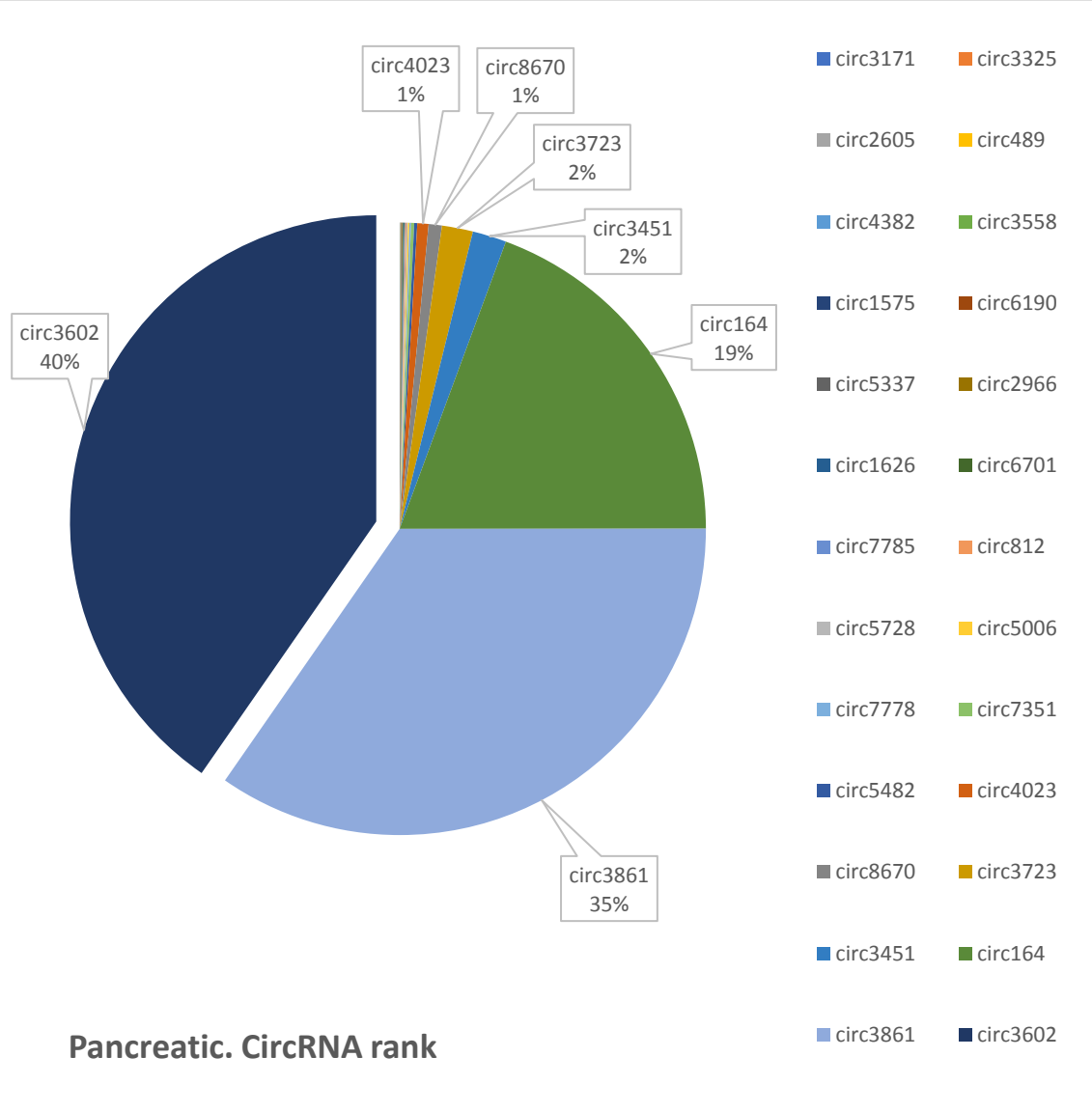

Ilustración 38. CircRNAs más relevantes para cáncer pancreático.

4.2.1. Circ3602

Es un circRNA **no catalogado** en circBase. Se transcribe a partir de los exones 14, 15, 16 y 17 del **gen RPS6KB1** (Ilustración 39). Sus **coordenadas** GRCh38g son chr17:59935193-59940943.

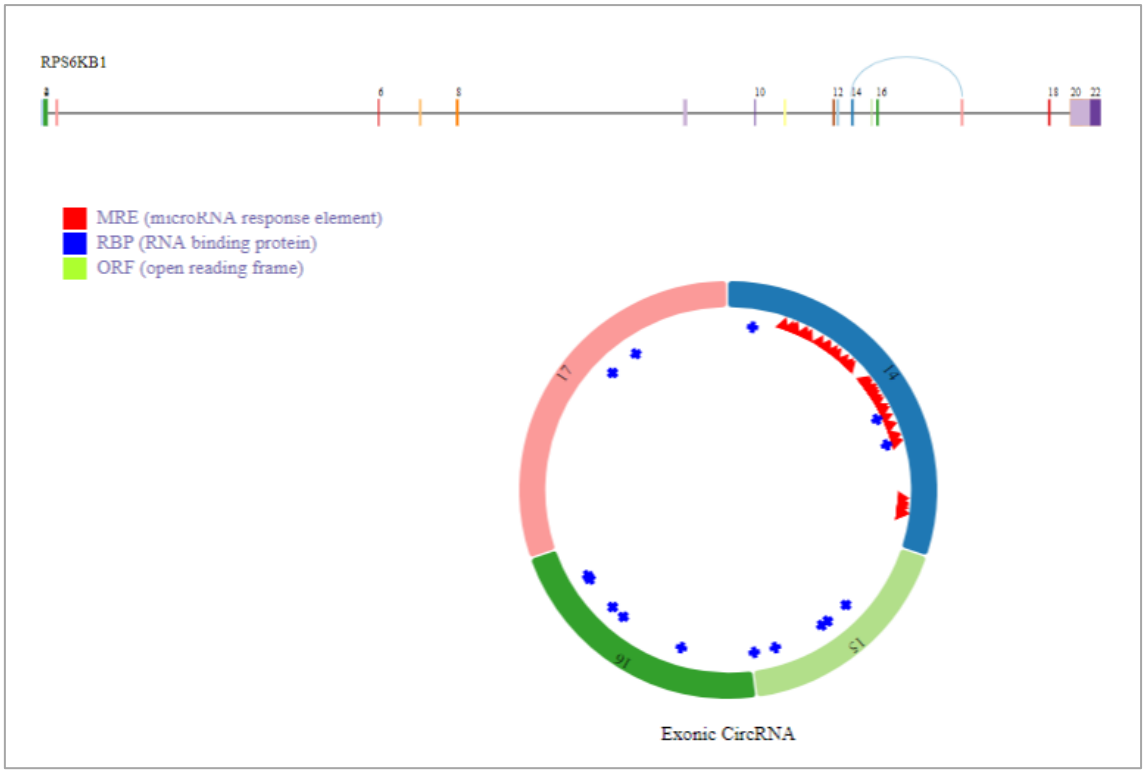

Ilustración 39. Estructura de circ3602.

La **proteína ribosómica S6 quinasa, polipéptido 1** (RPS6KB1) pertenece a la familia ribosómica S6 quinasa serina/treonina quinasa. La proteína funciona en la **señalización de mTOR**, la **síntesis de proteínas**, el **crecimiento celular** y la **proliferación celular**. Se observan mutaciones sin sentido, mutaciones silenciosas e inserciones y eliminaciones del marco de lectura en cánceres tales como cáncer de endometrio, cáncer intestinal y cáncer de estómago [153]. Adicionalmente encontramos **evidencias** de la **sobre expresión** de este gen, junto con su proteína fosforilada p-RPS6KB1, en **tumores neuroendocrinos** [154] , **cáncer de pulmón** [155], **cáncer de mama** [156] y **cáncer pancreático** [157]

En la Ilustración 40 observamos que **la media** de los **recuentos de lectura** normalizados para el circ3602 **es menor** para cáncer pancreático que para el resto de los grupos, pero el **valor máximo** está **igualado** al valor del resto de grupos. Es decir, observamos una **amplia dispersión** en los valores de recuentos de lectura normalizados para este circRNA en el grupo de cáncer pancreático.

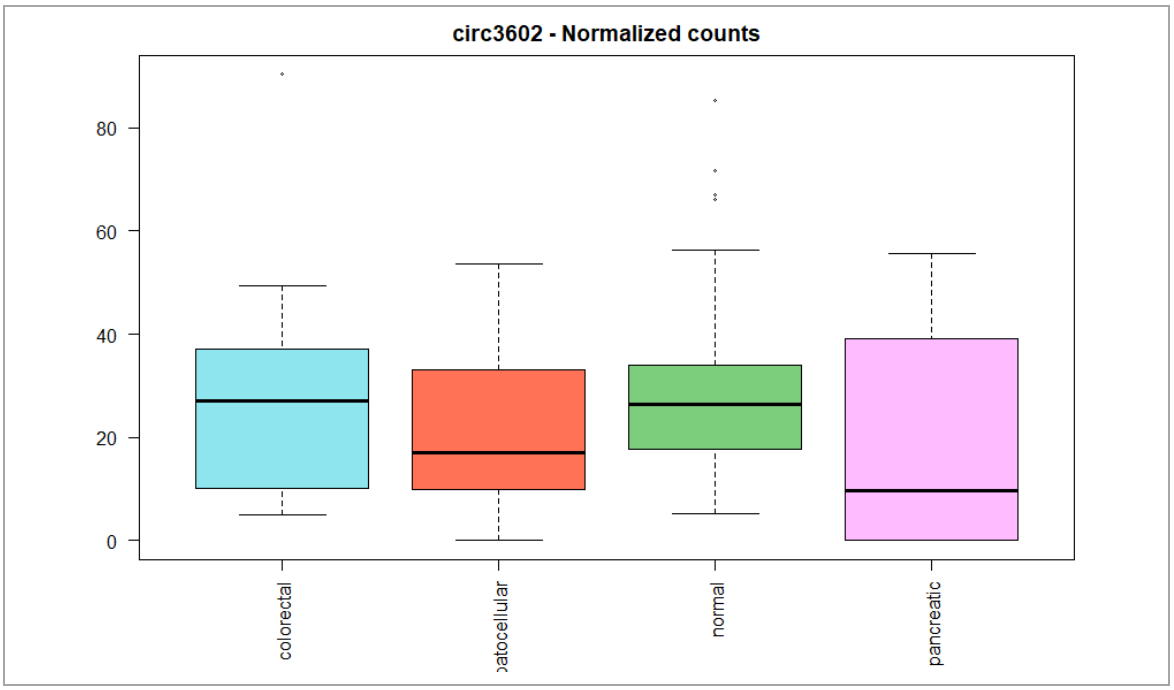

Ilustración 40. Box-plot con los recuentos de lectura normalizados del circ3602 en cada grupo.

En la "Cáncer-Specific CircRNA Database" [26] encontramos **68 miRNAs diana** para el circ3602. Se ha revisado la literatura encontrando evidencias de la **implicación de 19** de ellos en **cáncer pancreático**. Pero, al contrario de lo que ocurría en cáncer hepatocelular y colorrectal, se trata de **miRNAs que promueven el cáncer** cuando están sobre expresados.

Este hecho es consistente con un valor medio menor para los recuentos de lectura en cáncer pancreático para el circ3602 que los estaría capturando. En los casos en los que se encuentre menos expresado dejaría de hacer de esponja para estos miRNAs, lo que podría estar favoreciendo el desarrollo y progresión de la enfermedad.

En la siguiente tabla (Tabla 8) se detallan las **evidencias** encontradas sobre la **implicación** de los **miRNAs** diana del circ3602 en **cáncer pancreático**:

mi-RNA	Evidencias de su participación en cáncer pancreático	Ref.
hsa-miR-1185	Plasma MicroRNAs as Novel Biomarkers for Patients with Intraductal Papillary Mucinous Neoplasms of the Pancreas	[158]
hsa-miR-3679	Salivary microRNAs show potential as a noninvasive biomarker for detecting resectable pancreatic cancer	[159]
	miRNAs As Diagnostic and Prognostic Biomarkers in Pancreatic Ductal Adenocarcinoma and Its Precursor Lesions: A Review	[160]
hsa-miR-1224	The microRNA expression signature of pancreatic ductal adenocarcinoma by RNA sequencing: anti-tumour functions of the microRNA-216 cluster	[161]
hsa-miR-1237	Aberrant expression of microRNAs in serum may identify individuals with pancreatic cancer	[162]
hsa-miR-1488	Contribution of microRNAs in understanding the pancreatic tumor microenvironment involving cancer associated stellate and fibroblast cells	[163]

hsa-miR-1488	The pancreatic tumor microenvironment drives changes in miRNA expression that promote cytokine production and inhibit migration by the tumor associated stroma	[164]
hsa-miR-4697	The microRNA expression signature of pancreatic ductal adenocarcinoma by RNA sequencing: anti-tumour functions of the microRNA-216 cluster	[161]
hsa-miR-1275	Noninvasive urinary miRNA biomarkers for early detection of pancreatic adenocarcinoma	[165]
hsa-miR-1290	MicroRNA array analysis finds elevated serum miR-1290 accurately distinguishes patients with low-stage pancreatic cancer from healthy and disease controls.	[166]
hsa-miR-182	Circulating microRNA-182 in plasma and its potential diagnostic and prognostic value for pancreatic cancer.	[167]
hsa-miR-128	Serum miR-128-2 Serves as a Prognostic Marker for Patients with Hepatocellular Carcinoma	[168]
	miR-128 and its target genes in tumorigenesis and metastasis	[169]
hsa-miR-1290	Serum miR-1290 as a Marker of Pancreatic Cancer-Response	[170]
hsa-miR-182	MicroRNA-182 promotes pancreatic cancer cell proliferation and migration by targeting β-TrCP2.	[171]
hsa-miR-199 hsa-miR-214	MicroRNA-199a and -214 as potential therapeutic targets in pancreatic stellate cells in pancreatic tumor	[172]
	Dysregulation of miR-15a and miR-214 in human pancreatic cancer.	[173]
hsa-miR-2355	Next-generation sequencing reveals novel differentially regulated mRNAs, lncRNAs, miRNAs, sdRNAs and a piRNA in pancreatic cancer	[174]
hsa-miR-4713	The microRNA expression signature of pancreatic ductal adenocarcinoma by RNA sequencing: anti-tumour functions of the microRNA-216 cluster	[161]
hsa-miR-548	MicroRNA-548a-5p promotes proliferation and inhibits apoptosis in hepatocellular carcinoma cells by targeting Tg737	[175]
hsa-miR-629	MiR-629 promotes human pancreatic cancer progression by targeting FOXO3	[176]
hsa-miR-938	Plasma microRNA panels to diagnose pancreatic cancer: Results from a multicenter study	[177]
	Plasma miRNAs Effectively Distinguish Patients With Pancreatic Cancer From Controls: A Multicenter Study.	[178]

Tabla 8. Mi-RNAs diana del circ3602 cuya desregulación está implicada en cáncer pancreático

4.1.3. Otros circRNAs relevantes en cáncer pancreático.

En este grupo hay que destacar que el **valor medio** de recuentos de lectura normalizados para los **circRNAs** considerados relevantes **suele ser menor** que el del grupo de **personas sanas** (Ilustración 41). Sucede **lo contrario** que en **cáncer colorrectal y hepatocelular**, gupos en los que los circRNAs considerados relevantes se encontraban generalmente sobre expresados, de forma estadísticamente significativa en muchos casos, principalmente en cáncer colorectal.

Destacar también que el **comportamiento** observado de los recuentos normalizados de lectura de los circRNAs revisados en este capítulo ser el **opuesto** en cáncer pancreático al resto de cánceres: si los valores medios de un circRNA son menores en cáncer pancreático que en las personas sanas, en cáncer colorectal y hepatocelular están por encima; dándose también la situación contraria: para valores mayores en cáncer pancreático encontramos que ese mismo circRNA se encuentra menos expresado en cáncer colorrectal y hepatocelular.

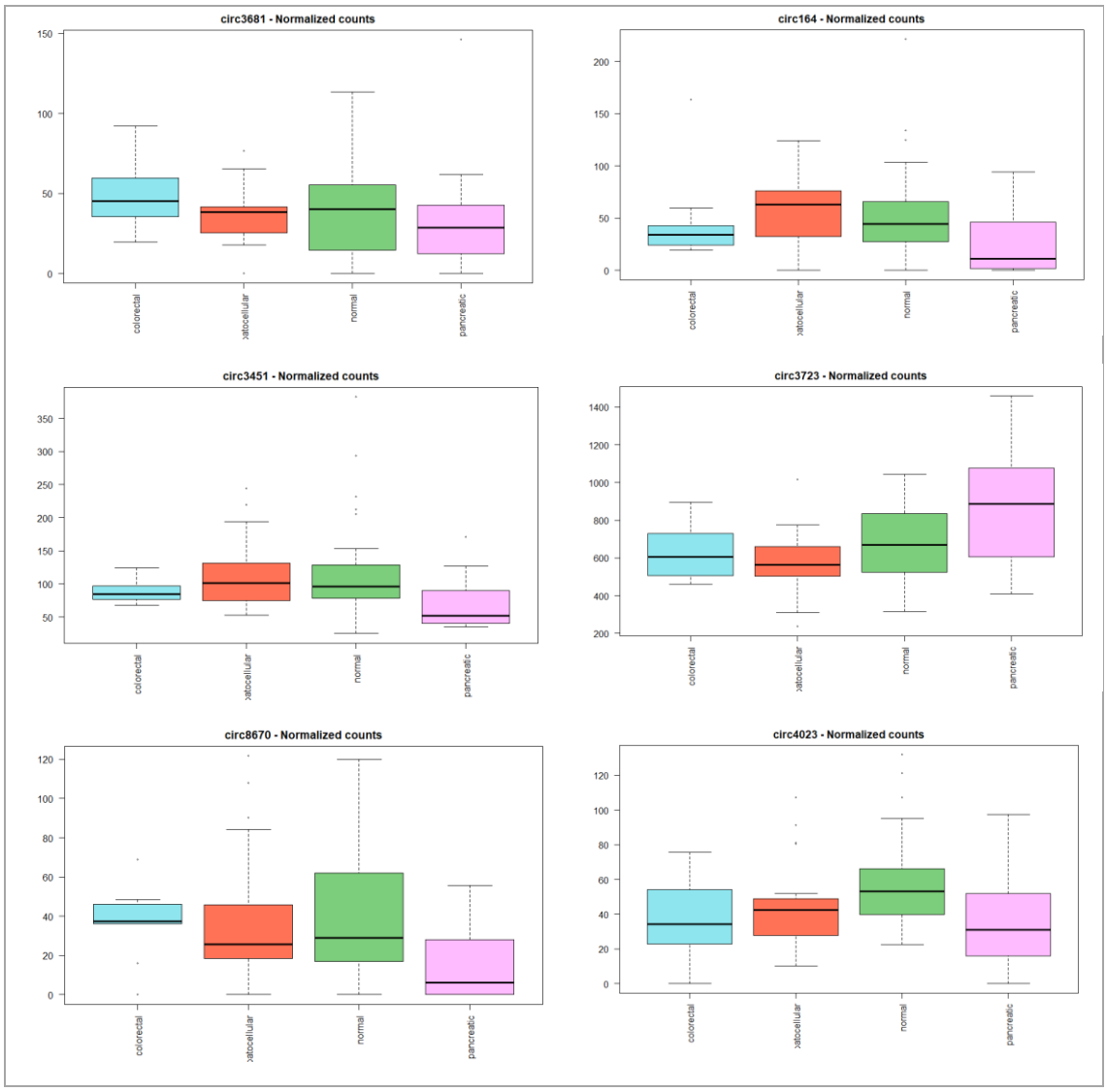

Ilustración 41. Otros circRNAs relevantes para cáncer pancreático

Otros **circRNAs** considerados **relevantes** en **cáncer pancreático** (Ilustración 41) se describen a continuación, no habiendo encontrado referencias en la literatura para ninguno de ellos.

Circ3861: recibe el nombre en circBase de **hsa_circ_0003428**. Se transcribe a partir del **gen POLI** en el **cromosoma 18** y sus **coordenadas** GRCh38g son chr18:51804072-51813781. Su **valor medio** de recuentos de lectura normalizados es **menor** en cáncer pancreático que para individuos sanos y resto de cánceres.

Circ164: recibe el nombre en circBase de **hsa_circ_0011167**. Se transcribe a partir del **gen EPB41** en el **cromosoma 1**. Sus **coordenadas** GRCh38g son chr1:28987431-28987905. En el gráfico se observa un **valor medio** de recuentos de lectura normalizados es **menor** que el del resto de grupos.

Circ3451: recibe el nombre en circBase **hsa_circ_0000754**. Se transcribe a partir del **gen SSH2** en el **cromosoma 1**. Sus **coordenadas** GRCh38 g son chr17:29684563-29703062. De nuevo el **valor medio** de recuentos de lectura normalizados es **menor** al del resto de grupos.

Circ3723: su nombre en circBase es **hsa_circ_0000826**. Se transcribe a partir del **gen ANKRD12** en el **cromosoma 1**. Sus **coordenadas** GRCh38g son chr18:9182382-9221999. Tiene un **rango de expresión amplio**, siendo el **único** circRNA de los considerados relevantes cuyo **valor medio** para recuentos de lectura normalizados es **superior** al resto de grupos en cáncer pancreático.

Circ8670: su nombre en circBase es **hsa_circ_0001934**. Se transcribe a partir del **gen ATP7A** en el **cromosoma X**. Sus **coordenadas** GRCh38g son chrX:78014662-78020398. Su **valor medio** de recuentos de lectura normalizados **es menor** al del resto de grupos.

Circ4023: recibe el nombre en circBase de **hsa_circ_0026311**. Se transcribe a partir del **gen TFCP2** en el cromosoma 12. Sus **coordenadas** GRCh38g son chr12:51110877-51117747. Observamos un **valor medio** de recuentos de lectura normalizados **menor** al del resto de grupos.

Capítulo 5

Conclusiones

El **objetivo** del presente trabajo ha sido **detectar y cuantificar** los **circRNAs** presentes en muestras de RNA-Seq de **exosomas** de **sangre periférica** humana de individuos **sanos** y con **tres tipos de cáncer**: colorrectal, hepatocelular y pancreático; y, utilizando los recuentos de lectura de los circRNAs detectados en cada muestra, entrenar un **modelo de Machine Learning** que aprendiera a **discriminar** entre muestras de **pacientes con cáncer** e individuos **sanos** para, por último, **caracterizar el rol** de los **circRNAs** más relevantes detectados en cada grupo en el **desarrollo y progresión** de cada tipo de **cáncer**. Además, se debía implementar un script ejecutable que automatizara todas las fases.

En primer lugar, destacar que se **han logrado los objetivos** propuestos inicialmente llevándose a cabo **según la planificación** prevista, ajustándose a las **tareas y tiempos** establecidos en el plan de trabajo. Adicionalmente, **se han automatizado** todas las fases del proyecto, generando un **conjunto de scripts** ejecutables que, junto los casos de ejemplo, manual de usuario y **resultados intermedios**, se pueden descargar en el **repositorio**: http://github.com/carmengmz/circRNA

Para llevar a cabo el presente estudio, en primer lugar, conseguimos **encontrar** en el Gene Expression Omnibus repository (GEO) del NCBI (National Center for Biotecnology Information) [71] muestras secuencias de **RNA-Seq** de **exosomas de sangre periférica** humana **adecuadas** para la **detección de circRNA**. A continuación, se **descargaron** y se **pre procesaron** las muestras con la herramienta fastp[58] para eliminar los adaptadores, filtrar las lecturas de baja calidad y corregir las bases; realizando un **control de calidad** posterior [54] que nos aseguró que las muestras **tenían la calidad suficiente** para su posterior análisis.

A continuación, se **detectaron y cuantificaron** los **circRNAs** presentes en las muestras de RNA-Seq con la herramienta CIRI2 [61] y se **anotaron** los circRNAs detectados en cada muestra para obtener la tabla de **recuentos de lectura**. En total, **el número de circRNA**

detectados en todas las muestras **fue de 117.565**. Y por grupos, se detectaron 77.175 en el grupo de personas sanas, 53.767 en el grupo de cáncer hepatocelular, 39.420 en el grupo de cáncer pancreático y 39.931 en el grupo de cáncer colorrectal.

A partir de la tabla de recuentos de lectura, se generaron **tres conjuntos de datos** conteniendo los **recuentos** de lectura de circRNAs de individuos **sanos** y con cada uno de los **tres tipos de cáncer** con el fin de **entrenar** un modelo **de Machine Learning** en cada grupo que fuera capaz de **discriminar** entre individuos **sanos** y pacientes con **cáncer**. Para ello, en primer lugar, se **filtraron** los circRNAs poco expresados, descartando aquellos cuya expresión (en valor absoluto) fuera menor a 10 en al menos el 70% de las muestras en cada uno de los tres conjuntos. Como **resultado** obtuvimos **1194 circRNAs** en el grupo **de cáncer colorrectal**, **1227 circRNAs** en el grupo de **cáncer hepatocelular** y **1046 circRNAs** en el grupo de **cáncer pancreático**.

Una vez preparados los conjuntos de datos, se aplicó una **transformación estabilizadora de la varianza** (VST) [92] para eliminar la dependencia entre la media y la varianza, de manera que los **datos fueran adecuados** para la aplicación de técnicas de **Machine Learning**. Después **dividimos** cada grupo en dos subconjuntos: un primer conjunto de datos para **entrenar** el modelo y un segundo conjunto de **pruebas** para validar el modelo entrenado.

En el **conjunto de entrenamiento**, en primer lugar, ajustamos un **modelo de Random Forest** [95] para determinar la **importancia** de los circRNAs presentes en cada grupo encontrando que el grupo de **cáncer colorrectal** permitía visualizar la **separación** de los datos en grupos mediante un gráfico de escalado multidimensional (MDS), utilizando **solamente los 8 circRNAs** considerados más importantes. Sin embargo, los grupos de **cáncer hepatocelular y pancreático** mostraban una **mayor complejidad**, siendo necesario el uso de 95 y 50 circRNAs (respectivamente) para visualizar la separación de los datos en grupos en el gráfico MDS.

Para **cáncer colorrectal** se entrenó, mediante validación cruzada, un modelo de **Support Vector Machines** [97] logrando un **AUC igual a 1**, tanto en **entrenamiento** como en **pruebas**. Este resultado nos indica que **los circRNAs seleccionados** permiten **clasificar muestras** de pacientes con **cáncer colorrectal** y de pacientes **sanos** con una **precisión del 100%** utilizando solamente esos **8 circRNAs** en nuestra población a estudio. Se trata de un **resultado prometedor** que requeriría ser **validado** en un conjunto más amplio, pero que **confirma**, en nuestra población a estudio, que los **circRNAs de exosomas** de sangre periférica humana **se expresan de manera diferente** en pacientes con **cáncer colorectal** y personas **sanas**.

Adicionalmente, el hecho de que la **detección** de los circRNAs se hiciera a partir de una **muestra de sangre periférica** podría abrir la puerta, si esta hipótesis se confirmara en una población más amplia, al posible desarrollo de **pruebas diagnósticas poco invasivas**, con una alta especificidad y sensibilidad, usando la expresión **de unos pocos circRNAs**, para este tipo de cáncer.

Para **cáncer hepatocelular** contábamos con un **conjunto** de datos **más complejo**, así que se entrenó un **modelo** con **Redes Neuronales** [100] usando el algoritmo de **Extreme Machine Learning** [101]. Este tipo de modelos son capaces de reconocer **patrones más complejos**, pero necesitan un **mayor número de predictores**, así que se utilizaron los recuentos de lectura de **todos** los **circRNAs** del conjunto de entrenamiento.

Como **resultado** fuimos capaces de **entrenar** un modelo, mediante validación cruzada, que reportó un **AUC de 0,913 en entrenamiento** y un **AUC igual a 1 en el conjunto de pruebas**. De nuevo fuimos capaces de **discriminar**, con una **precisión del 100%,** entre pacientes con **cáncer hepatocelular** e individuos **sanos**; en esta ocasión, utilizando un conjunto de circRNAs mucho más amplio.

Este resultado nos dice que **la expresión**, en nuestra población a estudio, de los **circRNAs** en **exosomas** de sangre periférica de pacientes con cáncer hepatocelular **es más compleja**, pero que también es **suficientemente diferente** a la de las personas sanas como para poder **clasificar automáticamente** los individuos en grupos con una **precisión del 100%.**

Por último, para **cáncer pancreático**, se entrenó un modelo de **Support Vector Machines** [97] con un kernel radial, obteniendo un **error en entrenamiento de 0,078** y un **AUC de 0,833** en el conjunto de **pruebas**. Con cualquiera de los modelos entrenados, en el mejor de los casos, siempre teníamos **una muestra** de cáncer pancreático que era **clasificada en el grupo** de individuos **sanos**. Como solo disponíamos de **tres muestras** de cáncer pancreático en el conjunto de pruebas, el modelo arroja una **sensibilidad** del 100%, pero una **especificidad** del 66,67%. Así que, **es posible discriminar** entre pacientes sanos y con cáncer usando los circRNAs contenidos en exosomas de sangre periférica para cáncer pancreático, pero con una **precisión balanceada del 83,33%.**

Para terminar, se realizó una **caracterización** de los **circRNAs** más relevantes detectados en cada grupo, encontrando que el mismo circRNA, el **hsa_circ_0001190** que se transcribe a partir de los exones del **gen DYRK1A**, en el **cromosoma 21** era, con diferencia, el **más relevante** para **cáncer colorrectal y hepatocelular**, encontrándose **sobre expresado** (padj<0.1) con respecto a los individuos sanos en los dos casos. Este hecho sugiere que, de alguna manera, los mecanismos celulares y moleculares de ambos cánceres podrían estar relacionados.

Adicionalmente sabemos que los **circRNAs** pueden actuar como **esponjas de miRNAs** [18], así que realizamos una **revisión** de las **evidencias** disponibles en la literatura sobre la **implicación**, en cada tipo de cáncer, de los **miRNAs diana** para este circRNA en cada tipo de cáncer, encontrando que **todos ellos**, cuando se encuentran normalmente expresados, tienen un **efecto oncoprotector**. Además, llama la atención que todos los **circRNAs más importantes** detectados para **cáncer hepatocelular y colorrectal**, suelen estar **sobre expresados**, y como ya hemos dicho, todos los **miRNAs** con los que podría **interaccionar** el **hsa_circ_0001190**, para los que hemos encontrado **estudios publicados**, son **oncoprotectores** en cáncer hepatocelular y colorrectal. Esto significaría que, si los **circRNAs sobre expresados** que estamos estudiando **capturan** estos miRNAs, **estaría desapareciendo el efecto oncoprotector**, favoreciendo el **desarrollo** y **progresión** de la

enfermedad. Se trata de una hipótesis **prometedora** que requeriría **más trabajo** para ser confirmada, pero cuya confirmación quizás abriría la puerta a nuevas terapias.

En el caso de **cáncer pancreático** encontramos que los **circRNAs** más **relevantes** se encontraban **menos expresados** que en en el grupo de personas sanas y resto de cánceres, no encontrando, en el análisis de expresión diferencial, ningún circRNA que estuviera diferencialmente expresado de manera significativa.

En nuestro modelo, el circRNA más importante para cáncer pancreático, fue el **circ3602** (que aún no tiene nombre en circBase [25]). Se trata de un circRNA que se transcribe a partir de los exones del **gen RPS6KB1,** en el **cromosoma 17** y cuyo valor medio para los recuentos de lectura normalizados es menor en cáncer pancreático que en el resto de grupos.

Por último, realizamos una **revisión** de las **evidencias** disponibles sobre la implicación de los **miRNAs diana** para este circRNA en cáncer pancreático y, al contrario de lo que ocurría en cáncer hepatocelular y colorrectal, se trataría de **miRNAs que promueven el desarrollo y progresión del cáncer** cuando se encuentran **sobre expresados**, lo cual es **consistente** con un **valor menos expresado** en cáncer pancreático para el **circ3602** que los estaría capturando. Al estar sub expresado **dejaría de hacer de esponja** para estos miRNAs, lo que **podría estar favoreciendo** el desarrollo y progresión de la enfermedad.

De hecho, al contrario de lo que ocurría en cáncer colorrectal y hepatocelular, cuyos circRNAs solían estar sobre expresados, todos los **circRNAs** considerados **más relevantes** en **cáncer pancreático** se suelen encontrar **menos expresados** con respecto al resto de grupos. Esto nos dice que los **mecanismos** celulares y moleculares del cáncer pancreático podrían ser **muy diferentes** a los del cáncer colorrectal y hepatocelular. Además, el conjunto de circRNAs detectado en exosomas **es suficientemente diferente** al de las personas sanas como para poder **clasificar** muestras **con cierta precisión**, pero en este caso no del 100%, al menos en nuestra población a estudio.

Por tanto, podemos concluir que los **circRNAs** detectados en **exosomas** de **sangre periférica** humana permiten **discriminar**, en la población estudiada, entre pacientes con **cáncer** e individuos **sanos** con una **alta precisión**, del 100% en cáncer colorrectal y hepatocelular. Además, se han encontrado **evidencias** que los **miRNAs** con los que podrían estar **interactuando** los **circRNAs** más relevantes en cada grupo podrían estar **implicados** en el **desarrollo** y **progresión** de cada uno de los tres tipos de **cáncer** cuando se encuentran desregulados.

Se trata de resultados **prometedores** que abrirían la puerta a **nuevos trabajos**, por ejemplo, **validar** los resultados en una **población mayor**, **cuantificar los miRNAs** en las mismas muestras de RNA-Seq de exosomas de sangre periférica usadas para cuantificar los circRNAs, para **validar** si se encuentran **desregulados** para los tres tipos de cáncer, o **investigar** si la **desregulación** de los circRNAs **interfiere** de alguna manera en la **expresión** de sus **homólogos lineales**. Así mimo sería muy **interesante** determinar cuál es el **papel** del **hsa_circ_0001190** en **cáncer colorrectal y hepatocelular**, y si podría ser utilizado como un **biomarcador fiable.** Y si todas las **hipótesis se confirman**, tal vez se podría intentar **diseñar** una **prueba diagnóstica poco invasiva** a partir de muestras de sangre periférica.

Glosario

ADN: ácido desoxirribonucleico (DNA en inglés)

ARN: ácido ribonucleico (RNA en inglés)

circRNA: ARN circular (del inglés, circular RNA).

miRNA: micro ARN (del inglés, micro RNA)

mRNA: ARN mensajero (del inglés, messenger RNA)

RNA-Seq: Secuenciación del ARN (del inglés, RNA sequencing)

ML: aprendizaje automático (del inglés, Machine Learning)

VST: transformación estabilizadora de la varianza (del inglés, Variance Stabilizing Transformation)

SVM: Máquinas de vectores de Soporte (del inglés, Support Vector Machines)

RF: Random Forest.

NN: Redes Neuronales Artificiales (del inglés, Neural Networks)

AUC: Área Bajo la Curva (del inglés, Area Under the [ROC] curve)

Bibliografía

[1] S. Memczak *et al.*, "Circular RNAs are a large class of animal RNAs with regulatory potency," *Nature*, vol. 495, no. 7441, pp. 333–338, 2013.

[2] J. Salzman, C. Gawad, P. L. Wang, N. Lacayo, and P. O. Brown, "Circular RNAs Are the Predominant Transcript Isoform from Hundreds of Human Genes in Diverse Cell Types," *PLoS One*, vol. 7, no. 2, p. e30733, 2012.

[3] T. B. Hansen *et al.*, "Natural RNA circles function as efficient microRNA sponges," *Nature*, vol. 495, no. 7441, pp. 384–388, 2013.

[4] S. Meng *et al.*, *CircRNA: Functions and properties of a novel potential biomarker for cancer*, vol. 16, no. 1. 2017.

[5] J. Conde-Vancells *et al.*, "Characterization and comprehensive proteome profiling of exosomes secreted by hepatocytes," *J. Proteome Res.*, vol. 7, no. 12, pp. 5157–5166, 2008.

[6] R. M. Johnstone, A. Mathew, A. B. Mason, and K. Teng, "Exosome formation during maturation of mammalian and avian reticulocytes: Evidence that exosome release is a major route for externalization of obsolete membrane proteins," *J. Cell. Physiol.*, vol. 147, no. 1, pp. 27–36, 1991.

[7] G. Van Niel, I. Porto-Carreiro, S. Simoes, and G. Raposo, "Exosomes: A common pathway for a specialized function," *Journal of Biochemistry*, vol. 140, no. 1. pp. 13–21, 2006.

[8] A. K. Ludwig and B. Giebel, "Exosomes: Small vesicles participating in intercellular communication," *Int. J. Biochem. Cell Biol.*, vol. 44, no. 1, pp. 11–15, 2012.

[9] C. Frühbeis *et al.*, "Neurotransmitter-Triggered Transfer of Exosomes Mediates Oligodendrocyte-Neuron Communication," *PLoS Biol.*, vol. 11, no. 7, 2013.

[10] K. O'Brien *et al.*, "Exosomes from triple-negative breast cancer cells can transfer phenotypic traits representing their cells of origin to secondary cells," *Eur. J. Cancer*, vol. 49, no. 8, pp. 1845–1859, 2013.

[11] P. Kucharzewska *et al.*, "Exosomes reflect the hypoxic status of glioma cells and mediate hypoxia-dependent activation of vascular cells during tumor development," *Proc. Natl. Acad. Sci.*, vol. 110, no. 18, pp. 7312–7317, 2013.

[12] S. Memczak *et al.*, "Identification and Characterization of Circular RNAs As a New Class of Putative Biomarkers in Human Blood.," *PLoS One*, vol. 10, no. 10, p. e0141214, Oct. 2015.

[13] Y. Li *et al.*, "Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis," *Cell Res.*, vol. 25, no. 8, p. 981, 2015.

[14] I. L. Patop and S. Kadener, "circRNAs in Cancer," *Curr. Opin. Genet. Dev.*, vol. 48, pp. 121–127, Feb. 2018.

[15] Q. Yang *et al.*, "A circular RNA promotes tumorigenesis by inducing c-myc nuclear translocation," *Cell Death Differ.*, vol. 24, p. 1609, 2017.

[16] J. Guarnerio and M. J. J. C. P. S. V. . B. K. N. M. M. . L.-C. F. T. Y. B. A. H. . P. P. P. Bezzi, "Oncogenic Role of Fusion-circRNAs Derived from Cancer-Associated Chromosomal Translocations.," *Cell*, vol. 166, no. 4, 2016.

[17] F. Li *et al.*, "Circular RNA ITCH has inhibitory effect on ESCC by suppressing the Wnt/beta-catenin pathway," *Oncotarget*, vol. 6, no. 8, pp. 6001–6013, 2015.

[18] F. R. Kulcheski, A. P. Christoff, and R. Margis, "Circular RNAs are miRNA sponges and can be used as a new class of biomarker," *J. Biotechnol.*, vol. 238, pp. 42–51, Nov. 2016.

[19] P. Li *et al.*, "Using circular RNA as a novel type of biomarker in the screening of gastric cancer," *Clin. Chim. Acta*, vol. 444, pp. 132–136, 2015.

[20] W. Li *et al.*, "Characterization of hsa_circ_0004277 as a new biomarker for acute myeloid leukemia via circular RNA profile and bioinformatics analysis," *Int. J. Mol. Sci.*, vol. 18, no. 3, p. 597, Mar. 2017.

[21] Z. Zhao *et al.*, "Peripheral blood circular RNA hsa_circ_0124644 can be used as a diagnostic biomarker of coronary artery disease," *Sci. Rep.*, vol. 7, p. 39918, Jan. 2017.

[22] Z. Qian *et al.*, "Potential Diagnostic Power of Blood Circular RNA Expression in Active Pulmonary Tuberculosis," *EBioMedicine*, 2017.

[23] Z. Zhao, X. Li, D. Jian, P. Hao, L. Rao, and M. Li, "Hsa_circ_0054633 in peripheral blood can be used as a diagnostic biomarker of pre-diabetes and type 2 diabetes mellitus," *Acta Diabetol.*, vol. 54, no. 3, pp. 237–245, 2017.

[24] D. W. Kane, M. M. Hohman, E. G. Cerami, M. W. McCormick, K. F. Kuhlmman, and J. A. Byrd, "Agile methods in biomedical software development: A multi-site experience report," *BMC Bioinformatics*, vol. 7, 2006.

[25] P. Glažar, P. Papavasileiou, and N. Rajewsky, "circBase: a database for circular RNAs," *RNA*, vol. 20, no. 11, pp. 1666–1670, 2014.

[26] S. Xia *et al.*, "CSCD: a database for cancer-specific circular RNAs," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D925–D929, Jan. 2017.

[27] C. Harding, J. Heuser, and P. Stahl, "Endocytosis and intracellular processing of transferrin and colloidal gold-transferrin in rat reticulocytes: demonstration of a pathway for receptor shedding.," *Eur. J. Cell Biol.*, vol. 35, no. 2, pp. 256–63, 1984.

[28] B. T. Pan and R. M. Johnstone, "Fate of the transferrin receptor during maturation of sheep reticulocytes in vitro: Selective externalization of the receptor," *Cell*, vol. 33, no. 3, pp. 967–978, 1983.

[29] R. M. Johnstone, M. Adam, J. R. Hammond, L. Orr, and C. Turbide, "Vesicle formation during reticulocyte maturation. Association of plasma membrane activities with released vesicles (exosomes).," *J. Biol. Chem.*, vol. 262, no. 19, pp. 9412–9420, 1987.

[30] N. P. Hessvik and A. Llorente, "Current knowledge on exosome biogenesis and release," *Cellular and Molecular Life Sciences*, vol. 75, no. 2. pp. 193–208, 2018.

[31] S. Mathivanan, H. Ji, and R. J. Simpson, "Exosomes: Extracellular organelles important in intercellular communication," *Journal of Proteomics*, vol. 73, no. 10. pp. 1907–1920, 2010.

[32] M. Record, K. Carayon, M. Poirot, and S. Silvente-Poirot, "Exosomes as new vesicular lipid transporters involved in cell-cell communication and various pathophysiologies," *Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids*, vol. 1841, no. 1. pp. 108–120, 2014.

[33] A. Becker, B. K. Thakur, J. M. Weiss, H. S. Kim, H. Peinado, and D. Lyden, "Extracellular Vesicles in Cancer: Cell-to-Cell Mediators of Metastasis," *Cancer Cell*, vol. 30, no. 6. pp. 836–848, 2016.

[34] H. Peinado *et al.*, "Melanoma exosomes educate bone marrow progenitor cells toward a pro-metastatic phenotype through MET," *Nat. Med.*, vol. 18, no. 6, pp. 883–891, 2012.

[35] B. Costa-Silva *et al.*, "Pancreatic cancer exosomes initiate pre-metastatic niche formation in the liver," *Nat. Cell Biol.*, vol. 17, no. 6, pp. 816–826, 2015.

[36] A. Hoshino *et al.*, "Tumour exosome integrins determine organotropic metastasis," *Nature*, vol. 527, no. 7578, pp. 329–335, 2015.

[37] N. R. Pamudurti *et al.*, "Translation of CircRNAs," *Mol. Cell*, vol. 66, no. 1, p. 9–21.e7, Jan. 2017.

[38] J. T. Granados-Riveron and G. Aquino-Jarquin, "The complexity of the translation ability of

circRNAs," *Biochim. Biophys. Acta - Gene Regul. Mech.*, vol. 1859, no. 10, pp. 1245–1251, Oct. 2016.

[39] Y. Y. Yang *et al.*, "Extensive translation of circular RNAs driven by N6-methyladenosine," *Cell Res*, vol. 27, no. 5, pp. 626–641, 2017.

[40] J. M. Nigro *et al.*, "Scrambled exons," *Cell*, vol. 64, no. 3, pp. 607–613, 1991.

[41] L. S. Kristensen, T. L. H. Okholm, M. T. Venø, and J. Kjems, "Circular RNAs are abundantly expressed and upregulated during human epidermal stem cell differentiation," *RNA Biol*, pp. 1–12, 2017.

[42] F. Crick, "On Protein Synthesis," *Symp. Soc. Exp. Biol.*, pp. 138–166, 1958.

[43] F. H. C. Crick, "Central Dogma of Molecular Biology," *Nature*, vol. 227, no. 5258. pp. 561–563, 1970.

[44] L.-L. L. Chen, "The biogenesis and emerging roles of circular RNAs," *Nat. Rev. Mol. Cell Biol.*, vol. 17, no. 4, p. 205, 2016.

[45] K. Andreeva and N. G. F. Cooper, "Circular RNAs: New Players in Gene Regulation," *Adv. Biosci. Biotechnol.*, vol. Vol.06No.0, p. 9, 2015.

[46] M. L. Metzker, "Sequencing technologies the next generation," *Nature Reviews Genetics*, vol. 11, no. 1. pp. 31–46, 2010.

[47] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: A revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1. pp. 57–63, 2009.

[48] X. Adiconis *et al.*, "Comparative analysis of RNA sequencing methods for degraded or low-input samples," *Nat. Methods*, vol. 10, no. 7, pp. 623–629, 2013.

[49] L. Szabo and J. Salzman, "Detecting circular RNAs: bioinformatic and experimental challenges," *Nat. Rev. Genet.*, vol. 17, no. 11, pp. 679–692, 2016.

[50] J. Wang, K. Liu, Y. Liu, Q. Lv, F. Zhang, and H. Wang, "Evaluating the bias of circRNA predictions from total RNA-Seq data," *Oncotarget*, vol. 8, no. 67, pp. 110914–110921, 2017.

[51] Y. Benjamini and T. P. Speed, "Summarizing and correcting the GC content bias in high-throughput sequencing," *Nucleic Acids Res.*, vol. 40, no. 10, 2012.

[52] K. D. Hansen, S. E. Brenner, and S. Dudoit, "Biases in Illumina transcriptome sequencing caused by random hexamer priming," *Nucleic Acids Res.*, vol. 38, no. 12, 2010.

[53] X. Li, A. Nair, S. Wang, and L. Wang, "Quality control of RNA-seq experiments," in *RNA Bioinformatics*, 2015, pp. 137–146.

[54] S. Andrews, "FastQC: A quality control tool for high throughput sequence data.," *Http://Www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/*, 2010. .

[55] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet.journal*, vol. 17, no. 1, p. 10, 2011.

[56] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: A flexible trimmer for Illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, 2014.

[57] S. Chen, T. Huang, Y. Zhou, Y. Han, M. Xu, and J. Gu, "AfterQC: Automatic filtering, trimming, error removing and quality control for fastq data," *BMC Bioinformatics*, vol. 18, 2017.

[58] S. Chen, Y. Zhou, Y. Chen, and J. Gu, "fastp: an ultra-fast all-in-one FASTQ preprocessor," *bioRxiv*, p. 274100, 2018.

[59] T. B. Hansen, M. T. Venø, C. K. Damgaard, and J. Kjems, "Comparison of circular RNA prediction tools," *Nucleic Acids Res.*, vol. 44, no. 6, pp. e58–e58, 2016.

[60] X. Zeng, W. Lin, M. Guo, and Q. Zou, *A comprehensive overview and evaluation of circular RNA detection tools*, vol. 13, no. 6. 2017, p. e1005420.

[61] Y. Gao, J. Wang, and F. Zhao, "CIRI: An efficient and unbiased algorithm for de novo circular RNA identification," *Genome Biol.*, vol. 16, no. 1, p. 4, Jan. 2015.

[62] J. O. Westholm *et al.*, "Genome-wide Analysis of Drosophila Circular RNAs Reveals Their Structural and Sequence Properties and Age-Dependent Neural Accumulation," *Cell Rep.*, vol. 9, no. 5, pp. 1966–1981, 2014.

[63] J. Cheng *et al.*, "Specific identification and quantification of circular RNAs from sequencing data.," *Bioinformatics*, vol. 32, no. 7, pp. 1094–1096, Apr. 2015.

[64] S. Hoffmann *et al.*, "A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection," *Genome Biol*, vol. 15, no. 2, p. R34, 2014.

[65] X. O. Zhang, H. Bin Wang, Y. Zhang, X. Lu, L. L. Chen, and L. Yang, "Complementary sequence-mediated exon circularization," *Cell*, vol. 159, no. 1, pp. 134–147, 2014.

[66] K. Wang *et al.*, "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery," *Nucleic Acids Res*, vol. 38, no. 18, p. e178, 2010.

[67] X. Song *et al.*, "Circular RNA profile in gliomas revealed by identification tool UROBORUS," *Nucleic Acids Res.*, vol. 44, no. 9, pp. e87–e87, May 2016.

[68] L. Szabo *et al.*, "Erratum to: Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development [Genome Biology, 16, (2016) (126)], DOI: 10.1186/s13059-015-0690-5," *Genome Biology*, vol. 17, no. 1. 2016.

[69] O. G. Izuogu, A. A. Alhasan, H. M. Alafghani, M. Santibanez-Koref, D. J. Elliot, and M. S. Jackson, "PTESFinder: A computational method to identify post-transcriptional exon shuffling (PTES) events," *BMC Bioinformatics*, vol. 17, no. 1, 2016.

[70] T. J. Chuang, C. S. Wu, C. Y. Chen, L. Y. Hung, T. W. Chiang, and M. Y. Yang, "NCLscan: Accurate identification of non-co-linear transcripts (fusion, trans-splicing and circular RNA) with a good balance between sensitivity and precision," *Nucleic Acids Res.*, vol. 44, no. 3, 2016.

[71] R. Edgar, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Res.*, vol. 30, no. 1, pp. 207–210, 2002.

[72] S. Li, "RNA-seq reveals abundant circRNA, lncRNA and mRNA in blood exosomes of patients with hepatocellular carcinoma," *Gene Expression Omnibus*, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE100207. [Accessed: 14-Mar-2018].

[73] S. Li, "RNA-seq reveals abundant circRNA, lncRNA and mRNA in blood exosomes of patients with colorectal carcinoma," *Gene Expression Omnibus*, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE100063. [Accessed: 14-Mar-2018].

[74] S. Li, "RNA-seq reveals abundant circRNA, lncRNA and mRNA in blood exosomes of patients with pancreatic carcinoma," *Gene Expression Omnibus*, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE100232. [Accessed: 14-Mar-2018].

[75] S. Li *et al.*, "exoRBase: a database of circRNA, lncRNA and mRNA in human blood exosomes," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D106–D112, Jan. 2018.

[76] Y. Li *et al.*, "RNA-seq reveals abundant circRNA, lncRNA and mRNA in blood exosomes of normal persons," *Gene Expression Omnibus*, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE100206. [Accessed: 14-Mar-2018].

[77] D. Enderle *et al.*, "Characterization of RNA from exosomes and other extracellular vesicles isolated by a novel spin column-based method," *PLoS One*, vol. 10, no. 8, 2015.

[78] R. Leinonen, H. Sugawara, and M. Shumway, "The sequence read archive," *Nucleic Acids Res.*,

vol. 39, no. SUPPL. 1, 2011.

[79] P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants," *Nucleic Acids Res.*, vol. 38, no. 6, pp. 1767–1771, 2009.

[80] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, "MultiQC: Summarize analysis results for multiple tools and samples in a single report," *Bioinformatics*, vol. 32, no. 19, pp. 3047–3048, 2016.

[81] S. Parekh, C. Ziegenhain, B. Vieth, W. Enard, and I. Hellmann, "The impact of amplification on differential expression analyses by RNA-seq," *Sci. Rep.*, vol. 6, 2016.

[82] H. Li, "Aligning new-sequencing reads by BWA BWA : Burrows-Wheeler Aligner," *PPT*, 2010.

[83] E. S. Lander *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[84] I. Zwiener, B. Frisch, and H. Binder, "Transforming RNA-Seq data to improve the performance of prognostic gene signatures," *PLoS One*, vol. 9, no. 1, 2014.

[85] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biol.*, vol. 15, no. 12, 2014.

[86] M. D. Robinson and G. K. Smyth, "Moderated statistical tests for assessing differences in tag abundance," *Bioinformatics*, vol. 23, no. 21, pp. 2881–2887, 2007.

[87] D. J. McCarthy, Y. Chen, and G. K. Smyth, "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation," *Nucleic Acids Res.*, vol. 40, no. 10, pp. 4288–4297, 2012.

[88] L. WHITAKER, "ON THE POISSON LAW OF SMALL NUMBERS," *Biometrika*, vol. 10, no. 1, pp. 36–71, 1914.

[89] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.," *Bioinformatics*, vol. 26, no. 1, pp. 139–40, 2010.

[90] M. D. Robinson and G. K. Smyth, "Small-sample estimation of negative binomial dispersion, with applications to SAGE data," *Biostatistics*, vol. 9, no. 2, pp. 321–332, 2008.

[91] R. Tibshirani, "Estimating transformations for regression via additivity and variance stabilization," *J. Am. Stat. Assoc.*, vol. 83, no. 402, pp. 394–405, 1988.

[92] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biol.*, vol. 11, no. 10, 2010.

[93] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, vol. 19. 1984.

[94] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. 2013.

[95] L. Breiman, "Random forests," *Mach. Learn.*, 2001.

[96] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," *Adv. Neural Inf. Process. Syst. 26*, pp. 431–439, 2013.

[97] C. Cortes and V. Vapnik, "Support vector machine," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[98] V. N. Vapnik, *Statistical Learning Theory*. 1998.

[99] E. Osuna, R. Freund, and F. Girosi, "Support Vector Machines: Training and Applications," Mar. 1997.

[100] S. Haykin, *Neural Networks and Learning Machines*. 2008.

[101] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, 2006.

[102] a Liaw and M. Wiener, "Classification and Regression by randomForest," *R news*, vol. 2, no. December, pp. 18–22, 2002.

[103] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, "Misc functions of the Department of Statistics (e1071), TU Wien," *R package version 1.6-2*. 2014.

[104] S. Qu *et al.*, *The emerging functions and roles of circular RNAs in cancer*, vol. 414. 2018, pp. 301–309.

[105] D. P. D. P. Bartel *et al.*, "MicroRNAs: target recognition and regulatory functions.," *Cell*, vol. 136, no. 2, pp. 215–33, 2009.

[106] W. Becker, "Emerging role of DYRK family protein kinases as regulators of protein stability in cell cycle control.," *Cell Cycle*, vol. 11, no. 18, pp. 3389–94, 2012.

[107] P. Fernández-Martínez, C. Zahonero, and P. Sánchez-Gómez, "DYRK1A: the double-edged kinase as a protagonist in cell growth and tumorigenesis," *Mol. Cell. Oncol.*, 2015.

[108] Q. Liu *et al.*, "Tumor suppressor DYRK1A effects on proliferation and chemoresistance of AML cells by downregulating c-Myc," *PLoS One*, 2014.

[109] T. Li *et al.*, "Plasma circular RNA profiling of patients with gastric cancer and their droplet digital RT-PCR detection," *J Mol Med*, 2018.

[110] L. Long, G. Huang, H. Zhu, Y. Guo, Y. Liu, and J. Huo, "Down-regulation of miR-138 promotes colorectal cancer metastasis via directly targeting TWIST2," *J. Transl. Med.*, 2013.

[111] L. Zhao *et al.*, "The tumor suppressor miR-138-5p targets PD-L1 in colorectal cancer," *Oncotarget*, 2016.

[112] Y. He, D. Yu, L. Zhu, S. Zhong, J. Zhao, and J. Tang, "miR-149 in human cancer: A systemic review," *J. Cancer*, 2017.

[113] X. Liu, T. Xie, X. Mao, L. Xue, X. Chu, and L. Chen, "MicroRNA-149 increases the sensitivity of colorectal cancer cells to 5-fluorouracil by targeting forkhead box transcription factor FOXM1," *Cell. Physiol. Biochem.*, 2016.

[114] J. Pekow *et al.*, "MIR-4728-3p Functions as a Tumor Suppressor in Ulcerative Colitis-associated Colorectal Neoplasia Through Regulation of Focal Adhesion Signaling," *Inflamm. Bowel Dis.*, 2017.

[115] A. R. Lulla *et al.*, "miR-6883 family miRNAs target CDK4/6 to induce G1phase cell-cycle arrest in colon cancer cells," *Cancer Res.*, 2017.

[116] Y. Ma *et al.*, "Elevated oncofoetal miR-17-5p expression regulates colorectal cancer progression by repressing its target gene P130," *Nat. Commun.*, 2012.

[117] Q. Tang *et al.*, "MicroRNA-93 suppress colorectal cancer development via Wnt/β-catenin pathway downregulating," *Tumor Biol.*, 2015.

[118] X. F. Yu, J. Zou, Z. J. Bao, and J. Dong, "miR-93 suppresses proliferation and colony formation of human colon cancer stem cells," *World J Gastroenterol*, 2011.

[119] I.-P. Yang *et al.*, "MicroRNA-93 inhibits tumor growth and early relapse of human colorectal cancer by affecting genes involved in the cell cycle," *Carcinogenesis*, 2012.

[120] K. Abdelmohsen, S. Srikantan, Y. Kuwano, and M. Gorospe, "miR-519 reduces cell proliferation by lowering RNA-binding protein HuR levels," *Proc. Natl. Acad. Sci.*, 2008.

[121] K. Abdelmohsen *et al.*, "miR-519 suppresses tumor growth by reducing HuR levels," *Cell Cycle*, 2010.

[122] M. Wang *et al.*, "MiR-198 represses tumor growth and metastasis in colorectal cancer by targeting fucosyl transferase 8," *Sci. Rep.*, 2014.

[123] S. Tsuji *et al.*, "The miR-363-GATA6-Lgr5 pathway is critical for colorectal tumourigenesis," *Nat. Commun.*, 2014.

[124] F. Hu *et al.*, "MiR-363-3p inhibits the epithelial-to-mesenchymal transition and suppresses metastasis in colorectal cancer by targeting Sox4," *Biochem. Biophys. Res. Commun.*, 2016.

[125] X. Wu *et al.*, "Identification and validation of potential biomarkers for the detection of dysregulated microrna by QPCR in patients with colorectal adenocarcinoma," *PLoS One*, 2015.

[126] M. L. Slattery *et al.*, "Infrequently expressed miRNAs in colorectal cancer tissue and tumor molecular phenotype," *Mod. Pathol.*, 2017.

[127] H. Nakano, T. Miyazawa, K. Kinoshita, Y. Yamada, and T. Yoshida, "Functional screening identifies a microRNA, miR-491 that induces apoptosis by targeting Bcl-X(L) in colorectal cancer cells," *Int J Cancer*, 2010.

[128] H. J. An *et al.*, "Novel miR-5582-5p functions as a tumor suppressor by inducing apoptosis and cell cycle arrest in cancer cells through direct targeting of GAB1, SHC1, and CDK2," *Biochim. Biophys. Acta - Mol. Basis Dis.*, 2016.

[129] X. Lou, X. Qi, Y. Zhang, H. Long, and J. Yang, "Decreased expression of microRNA-625 is associated with tumor metastasis and poor prognosis in patients with colorectal cancer," *J. Surg. Oncol.*, 2013.

[130] S. Yan *et al.*, "Exosomal miR-6803-5p as potential diagnostic and prognostic marker in colorectal cancer," *J. Cell. Biochem.*, 2018.

[131] K. Xu, Z. Chen, C. Qin, and X. Song, "mir-7 inhibits colorectal cancer cell proliferation and induces apoptosis by targeting xrcc2," *Onco. Targets. Ther.*, 2014.

[132] N. Zhang *et al.*, "MicroRNA-7 is a novel inhibitor of YY1 contributing to colorectal tumorigenesis," *Oncogene*, 2013.

[133] A. Tsuchida *et al.*, "miR-92 is a key oncogenic component of the miR-17-92 cluster in colon cancer," *Cancer Sci.*, 2011.

[134] J.-J. Zhou, S. Zheng, L.-F. Sun, and L. Zheng, "MicroRNA regulation network in colorectal cancer metastasis.," *World J. Biol. Chem.*, 2014.

[135] Y. Katayama *et al.*, "Identification of pathogenesis-related microRNAs in hepatocellular carcinoma by expression profiling," *Oncol. Lett.*, 2012.

[136] Y. Ye *et al.*, "Induced MiR-1249 expression by aberrant activation of Hedegehog signaling pathway in hepatocellular carcinoma," *Exp. Cell Res.*, 2017.

[137] K. Sugimachi *et al.*, "Identification of Recurrence-Related microRNAs from Bone Marrow in Hepatocellular Carcinoma Patients.," *J. Clin. Med.*, 2015.

[138] X. Chen, L. Bo, W. Lu, G. Zhou, and Q. Chen, "MicroRNA-148b targets Rho-associated protein kinase 1 to inhibit cell proliferation, migration and invasion in hepatocellular carcinoma," *Mol. Med. Rep.*, 2016.

[139] J. Luo, P. Chen, W. Xie, and F. Wu, "MicroRNA-138 inhibits cell proliferation in hepatocellular carcinoma by targeting Sirt1," *Oncol. Rep.*, 2017.

[140] G. Luo *et al.*, "miR-149 represses metastasis of hepatocellular carcinoma by targeting actin-regulatory proteins PPM1F," *Oncotarget*, 2015.

[141] L. Chen, M. Jiang, W. Yuan, and H. Tang, "miR-17-5p as a novel prognostic marker for hepatocellular carcinoma.," *J. Invest. Surg.*, 2012.

[142] S. Tan, R. Li, K. Ding, P. E. Lobie, and T. Zhu, "MiR-198 inhibits migration and invasion of hepatocellular carcinoma cells by targeting the HGF/c-MET pathway," *FEBS Lett.*, 2011.

[143] W.-Y. Wang *et al.*, "miR-21 expression predicts prognosis in hepatocellular carcinoma.," *Clin.*

Res. Hepatol. Gastroenterol., 2014.

[144] L. Zhao and Y. Zhang, "miR-342-3p affects hepatocellular carcinoma cell proliferation via regulating nF-κB pathway," *Biochem. Biophys. Res. Commun.*, 2015.

[145] J. Ye, W. Zhang, S. Liu, Y. Liu, and K. Liu, "MiR-363 inhibits the growth, migration and invasion of hepatocellular carcinoma cells by regulating E2F3," *Oncol. Rep.*, 2017.

[146] M. X.-P. *et al.*, "MiR-608 rs4919510 is associated with prognosis of hepatocellular carcinoma," *Tumor Biol.*, 2016.

[147] X. Zhou *et al.*, "MiR-625 suppresses tumour migration and invasion by targeting IGF2BP1 in hepatocellular carcinoma," *Oncogene*, 2015.

[148] Y. Fang, J. L. Xue, Q. Shen, J. Chen, and L. Tian, "MicroRNA-7 inhibits tumor growth and metastasis by targeting the phosphoinositide 3-kinase/Akt pathway in hepatocellular carcinoma," *Hepatology*, 2012.

[149] X. Yang *et al.*, "Quantitative Proteomics Reveals the Regulatory Networks of Circular RNA CDR1as in Hepatocellular Carcinoma Cells," *J. Proteome Res.*, vol. 16, no. 10, p. acs.jproteome.7b00519, 2017.

[150] M. Shigoka *et al.*, "Deregulation of miR-92a expression is implicated in hepatocellular carcinoma development," *Pathol. Int.*, 2010.

[151] A. F. Vidal *et al.*, "The comprehensive expression analysis of circular RNAs in gastric cancer and its association with field cancerization," *Sci. Rep.*, vol. 7, no. 1, p. 14551, 2017.

[152] L.-J. Li *et al.*, "Circular RNA expression profile and potential function of hsa_circ_0045272 in systemic lupus erythematosus," *Immunology*, May 2018.

[153] C. Swanton, "My Cancer Genome: a unified genomics and clinical trial portal," *Lancet Oncol.*, vol. 13, no. 7, pp. 668–669, 2012.

[154] Z. R. Qian *et al.*, "Prognostic significance of MTOR pathway component expression in neuroendocrine tumors," *J. Clin. Oncol.*, vol. 31, no. 27, pp. 3418–3425, 2013.

[155] B. Chen *et al.*, "Hyperphosphorylation of RPS6KB1, rather than overexpression, predicts worse prognosis in non-small cell lung cancer patients.," *PLoS One*, vol. 12, no. 8, p. e0182891, 2017.

[156] B. Davidson, K. Valborg Reinertsen, D. Trinh, W. Reed, and P. J. Bøhler, "BAG-1/SODD, HSP70, and HSP90 are potential prognostic markers of poor survival in node-negative breast carcinoma," *Hum. Pathol.*, 2016.

[157] C.-C. Shen *et al.*, "High Phosphorylation Status of AKT/mTOR Signal in DESI2-Reduced Pancreatic Ductal Adenocarcinoma," *Pathol. Oncol. Res.*, vol. 21, no. 2, pp. 267–272, Apr. 2015.

[158] J. Permuth-Wey *et al.*, "Plasma MicroRNAs as Novel Biomarkers for Patients with Intraductal Papillary Mucinous Neoplasms of the Pancreas.," *Cancer Prev. Res. (Phila).*, vol. 8, no. 9, pp. 826–34, 2015.

[159] Z. Xie *et al.*, "Salivary microRNAs show potential as a noninvasive biomarker for detecting resectable pancreatic cancer," *Cancer Prev. Res.*, vol. 8, no. 2, pp. 165–173, 2015.

[160] B. Alemar, C. Gregório, and P. Ashton-Prolla, "miRNAs As Diagnostic and Prognostic Biomarkers in Pancreatic Ductal Adenocarcinoma and Its Precursor Lesions: A Review.," *Biomark. Insights*, vol. 10, pp. 113–24, 2015.

[161] K. Yonemori *et al.*, "The microRNA expression signature of pancreatic ductal adenocarcinoma by RNA sequencing: anti-tumour functions of the microRNA-216 cluster.," *Oncotarget*, vol. 8, no. 41, pp. 70097–70115, 2017.

[162] M. S. Lin, W. C. Chen, J. X. Huang, H. J. Gao, and H. H. Sheng, "Aberrant expression of microRNAs in serum may identify individuals with pancreatic cancer," *Int. J. Clin. Exp. Med.*, vol. 7, no. 12, pp. 5226–5234, 2014.

[163] S. Ali *et al.*, "Contribution of microRNAs in understanding the pancreatic tumor microenvironment involving cancer associated stellate and fibroblast cells.," *Am. J. Cancer Res.*, vol. 5, no. 3, pp. 1251–64, 2015.

[164] S. Han *et al.*, "The pancreatic tumor microenvironment drives changes in miRNA expression that promote cytokine production and inhibit migration by the tumor associated stroma," *Oncotarget*, vol. 8, no. 33, 2017.

[165] S. Debernardi *et al.*, "Noninvasive urinary miRNA biomarkers for early detection of pancreatic adenocarcinoma," *Am. J. Cancer Res.*, vol. 5, no. 11, pp. 3455–3466, 2015.

[166] A. Li *et al.*, "MicroRNA array analysis finds elevated serum miR-1290 accurately distinguishes patients with low-stage pancreatic cancer from healthy and disease controls," *Clin. Cancer Res.*, vol. 19, no. 13, pp. 3600–3610, 2013.

[167] Q. Chen, L. Yang, Y. Xiao, J. Zhu, and Z. Li, "Circulating microRNA-182 in plasma and its potential diagnostic and prognostic value for pancreatic cancer," *Med. Oncol.*, vol. 31, no. 11, p. 225, 2014.

[168] L. Zhuang, L. Xu, P. Wang, and Z. Meng, "Serum miR-128-2 serves as a prognostic marker for patients with hepatocellular carcinoma," *PLoS One*, vol. 10, no. 2, 2015.

[169] M. Li, W. Fu, L. Wo, X. Shu, F. Liu, and C. Li, "MiR-128 and its target genes in tumorigenesis and metastasis," *Experimental Cell Research*, vol. 319, no. 20. pp. 3059–3064, 2013.

[170] A. E. Frampton, J. Krell, G. Kazemier, and E. Giovannetti, "Serum miR-1290 as a marker of pancreatic cancer-letter," *Clinical Cancer Research*, vol. 19, no. 18. pp. 5250–5251, 2013.

[171] S. Wang *et al.*, "MicroRNA-182 promotes pancreatic cancer cell proliferation and migration by targeting β-TrCP2," *Acta Biochim. Biophys. Sin. (Shanghai).*, vol. 48, no. 12, pp. 1085–1093, 2016.

[172] P. R. Kuninty *et al.*, "MicroRNA-199a and -214 as potential therapeutic targets in pancreatic stellate cells in pancreatic tumor," *Oncotarget*, vol. 7, no. 13, 2016.

[173] X. J. Zhang, H. Ye, C. W. Zeng, B. He, H. Zhang, and Y. Q. Chen, "Dysregulation of miR-15a and miR-214 in human pancreatic cancer," *J. Hematol. Oncol.*, vol. 3, 2010.

[174] S. Müller *et al.*, "Next-generation sequencing reveals novel differentially regulated mRNAs, lncRNAs, miRNAs, sdRNAs and a piRNA in pancreatic cancer," *Mol. Cancer*, 2015.

[175] G. Zhao *et al.*, "MicroRNA-548a-5p promotes proliferation and inhibits apoptosis in hepatocellular carcinoma cells by targeting Tg737," *World J. Gastroenterol.*, 2016.

[176] H. Yan *et al.*, "MiR-629 promotes human pancreatic cancer progression by targeting FOXO3.," *Cell Death Dis.*, 2017.

[177] J. Kleeff *et al.*, "Plasma microRNA panels to diagnose pancreatic cancer: Results from a multicenter study," *Oncotarget*, 2016.

[178] J. Xu *et al.*, "Plasma miRNAs effectively distinguish patients with pancreatic cancer from controls a multicenter study," *Ann. Surg.*, 2016.