Maestría en Ciencias

en Ciencias de la Vida con orientación en Microbiología

Análisis *in silico* de microRNAs secretados en diferentes sistemas de estudio: Explorando posibles determinantes para el empaquetamiento exosomal

Tesis

Para cubrir parcialmente los requisitos necesarios para obtener el grado de Maestro en Ciencias

Presenta:

Ricardo de Jesús Ehecatl Gómez Reyes

Ensenada, Baja California, México

2016

Tesis defendida por

Ricardo de Jesús Ehecatl Gómez Reyes

y aprobada por el Comité

_____

Dra. Kristina Marie Herbert

Director del Comité

Dr. Carlos Alberto Brizuela Rodríguez

Dr. Alejandro Huerta Saquero

_____

Dra. Clara Elizabeth Galindo Sanchez
Coordinador del Posgrado en Ciencias de la Vida

_____

Dra. Rufina Hernández Martínez
Directora de Estudios de Posgrado

Resumen de la tesis que presenta **Ricardo de Jesús Ehecatl Gómez Reyes** como requisito parcial para la obtención del grado de Maestro en Ciencias en Ciencias de la vida con orientación en Microbiología.

Análisis in silico de microRNAs secretados en diferentes sistemas de estudio: Explorando posibles determinantes para el empaquetamiento exosomal.

Resumen aprobado por:

_____

Dra. Kristina Marie Herbert
Directora de Tesis

El interés por estudiar microRNAs (miRNAs) en varios metazoarios se debe al potencial de estas biomoléculas para modular procesos que involucran el desarrollo celular, a partir de su función en la regulación de los niveles proteicos. Un tema emergente de gran interés es la secreción al medio extracelular de miRNAs y su posible función (en analogía con las citocinas y factores de crecimiento, ambas proteinas de señalización extracelular) en la comunicación entre células de un mismo tejido (señalización autócrina) o en la comunicación del sistema inmune con patógenos (e inversamente) o células infectadas (señalización parácrina). Se ha reportado que los miRNAs son selectivamente empaquetados y dirigidos vía exosomas para reprogramar la expresión genética de las células blanco a las cuales estas vesículas van dirigidas. Aunque muchos reportes sugieren que los miRNAs de carga exosomal pueden implementarse para el diagnóstico y/o terapia de diversas enfermedades incluyendo cáncer, aun se sabe poco sobre los factores que determinan el empaquetamiento de miRNAs dentro de exosomas. Existen sin embargo una gran cantidad de datos de proteómica y transcriptómica publicados en bases de datos de libre acceso, que pueden ser re-analizados para responder a hipótesis inspiradas en preguntas básicas respecto al empaquetamiento de proteínas y RNA en exosomas.

Palabras clave: microRNAs, ncRNAs, exosomas, vesículas extracelulares, comunicación intercelular, cáncer, infecciones microbianas, bioinformática

Abstract of the thesis presented by **Ricardo de Jesús Ehecatl Gómez Reyes** as a partial requirement to obtain the Master of Science degree in Life Science with orientation in Microbiology.

In silico analysis of microRNAs secreted in different study systems: Exploring determinants for exosomal packaging.

Abstract approved by:

_____

Dra. Kristina Marie Herbert
Thesis Director

MicroRNAs (miRNAs) play a great role in the regulation of cellular protein levels during cell development. Therefore there is large interest in studying them in several metazoa. An emerging topic of great interest is the secretion of miRNAs into the extracellular space and their possible function (in analogy with cytokines and growth factors, both extracellular signaling proteins) in cell-to-cell communication within the same tissue (autocrine signaling) or in the communication of the immune system with pathogens (and vice versa) or infected cells (paracrine signaling). MiRNAs have been reported to be selectively packaged into exosomes, which are then secreted and may reprogram the genetic expression of target cells which abosrb these vesicles. Although many reports suggest that the exosomal miRNAs can be utilized for the diagnosis and/or therapy of various diseases including cancer, little is known about the factors that determine the packaging of miRNAs within exosomes. There is, however, a large amount of proteomics, transcriptomics, and small RNA sequencing data published in open access databases, which can be re-analyzed to answer hypotheses based on fundamental questions regarding the packaging of proteins and RNAs into exosomes.

Key words: microRNAs, ncRNAs, exosomes, extracellular vesicles, cell communication, cancer, microbial infection, bioinformatics.

Dedicatorias

Gracias a la vida que me ha dado tanto
Me dio dos hermanos que cuando los abro
Perfecto distingo lo negro del blanco
Y en el alto cielo su fondo estrellado

Y en la multitud a los Padres que yo amo.

Gracias a la vida que me ha dado tanto
Me ha dado la marcha de mis pies motivados
Con ellos andaré ciudades y charcos
Playas y desiertos, montañas y llanos
Y la casa tuya, tu calle y tu patio

Gracias a la vida que me ha dado tanto
Me di el corazón que agita su marco
Cuando miro el fruto del cerebro humano
Cuando miro al bueno tan lejos del malo
Cuando miro el fondo de tus ojos claros

Así yo distingo dicha de quebranto
Los dos materiales que forman mi canto
Y el canto de ustedes que es el mismo canto
Y el canto de todos que es mi propio canto

Adaptación: Gracia a la vida, Parra Sandoval V., (1966)

Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología

Tabla de contenido

Resumen en español I

Resumen en inglés II

Dedicatorias III

Agradecimientos IV

Lista de figuras VIII

Lista de tablas X

Capítulo 1. Introducción 1

Capítulo 2. Fundamentos 3

2.1.1 Diversidad celular 3

2.1.2 Membranas biológicas 3

2.1.3 Genoma, DNA, mRNA y ncRNAs 4

2.1.4 MicroRNAs: origen, biogénesis y funcionamiento 4

2.1.5 Vesículas extracelulares: un mecanismo de comunicación intercelular. 5

2.1.6 Exosomas: biogénesis de vesículas intraluminales en el endosoma 6

2.2 Cargamento exosomas: RNAs circulantes para la expresión y reprogramación celular 7

2.2.1 Exo-miRNAs: cáncer, inmunosupresión e inmuno-evasión por parásitos. 8

2.3.1 Secuenciación de RNAs pequeños. 10

2.3.2 Formato fastq 10

2.3.3 Análisis “downstream” de bibliotecas de secuenciación 11

Capítulo 3. Antecedentes 12

Capítulo 4. Hipótesis 14

Capítulo 5. Objetivos 15

5.1 Objetivo general 15

5.2 Objetivos específicos 15

Capítulo 6. Metodología 16

6.1 Conexión al servicio computacional de cluster en CICESE 17

6.2 Entorno de trabajo 17

6.3 Obtención de datos NGS, genoma y microRNAs 18

6.4 Pre-procesamiento de los datos NGS 20

6.5 miRDeep2: Identificación de miRNAs expresados en los datos NGS. 20

6.5.1 Módulo Mapper 21

6.5.2 Módulo quantifier 22

6.6 Lenguaje R: Normalización de datos. 23

6.7 Lenguaje R: Análisis de la correlación 24

6.8 EdgeR: análisis diferencial de la expresión de miRNAs 24

6.8.1 Quasi-likelihood F-tests (Qlf) 25

6.9 Selección de los miRNAs enriquecidos en la fracción exosomal. 25

6.9.1 Preparación de los archivos de entrada para la búsqueda de elementos de acción en cis 26

6.9.1 MEME: búsqueda de elementos de acción cis 28

6.9.2 Localización de exo-miRNAs con un elemenento de acción en cis a lo largo de las bibliotecas 29

6.9.3 Tailor: análisis de la adición de nucleótidos pos-transcripcional 30

Capítulo 7. Resultados 32

7.1 Pre-procesamiento y Mapeo de los datos NGS 32

7.2 Anotación de microRNAs 39

7.3 Análisis de la correlación 43

7.4 Análisis del cociente de abundancia de microRNAs al comparar fracciones 49

7.5. Búsqueda de elementos de acción en cis 55

7.6 Tailor: análisis de la adición de nucleótidos pos-transcripcional 60

7.7 Tailor: análisis de la adición de nucleótidos pos-transcripcional en los miRNAs abundantes 67

Capítulo 8. Discusión 74

Literatura citada 79

Lista de figuras

Figura 1. Esquema de flujo de trabajo de la herramienta bioinformática mirDeep2, modulo mapper.pl. 21

Figura 2. Flujo del pipeline del programa Tailor 31

Figura 3. Comparación del mapeo de las lecturas después y antes del pre-procesamiento.....34

Figura 4. Distribución de lecturas por su tamaño en longitud de nucleótido a lo largo de las bibliotecas de secuenciación pre-procesadas; Fracción exosomal. 35

Figura 5. Distribución de lecturas por su tamaño en longitud de nucleótido a lo largo de las bibliotecas de secuenciación pre-procesadas; Fracción Celular. 36

Figura 6. Figura 5. Distribución del puntaje de calidad de las secuencias a lo largo de las bibliotecas analizadas. Fracción exosomal. 36

Figura 7. Distribución del puntaje de calidad de las secuencias a lo largo de las bibliotecas analizadas. Fracción celular 37

Figura 8. Distribución del puntaje de calidad a lo largo de los nucleótidos de las secuencias [Fracción Exosomal] 38

Figura 9. Comprobación de los adaptores removidos de las bibliotecas de secuenciación después del pre-procesamiento (Panel izquierdo). Se denota la presencia de adaptores illumina para secuenciación de RNAS pequeños (Panel derecho). 39

Figura 10. Numero de lecturas mapeadas contra microRNAs de referencia en comparativa con el número de lecturas totales de las bibliotecas y mapeadas contra el genoma de referencia. 41

Figura 11. La estructura secundaria de los miRNAs expresados se genera automáticamente con el paquete Vienna-RNAfold. El paquete Vienna-RNAfold también calcula la notación "()" y "." para denotar los alineamientos y no alineamientos de la secuencia precursora para calcular su estructura secundaria. 42

Figura 11. Correlación de los datos de la condición experimental: Linfoblastoides EBV+. Los conteos están normalizados por cpm (Counts Per Million) con una correlación Spearman de 0.75 [RN], 0.73 [IM] y 0.76 [IK]. 44

Figura 12. Correlación de los datos de la condición experimental: Leucemia de Burkitt EBV+. Los conteos están normalizados por cpm (Counts Per Million) con una correlación Spearman de 0.75 [M9] y 0.74 [M3]. 45

Figura 13. Correlación de los datos de la condición experimental: Linfoma. Los conteos están normalizados por cpm (Counts Per Million) con una correlación Spearman de 0.73. 45

Figura 14 Mapa de la matriz de relación de las bibliotecas. Las bibliotecas (Panel derecho) se agrupan a sus respectivas condiciones experimentales (Panel inferior) 47

Figura 15. Mapa de la matriz de relación de las bibliotecas. Las bibliotecas se agrupan a sus respectivas condiciones experimentales: 1) Linfoma (EBV+), 2) Leucemia de Burkit (EBV -) y Linfoblastoides (EBV+). 48

Figura 16. Distribución de microRNAs entre fracciones; Ambas fracciones (exosomal y celular) están expresadas como el valor LogFC ≤ -1 o LogFC ≥ 1, respectivamente. 51

Figura 17. Gráfico de volcano representativo de los microRNAs abundantes y significativos por fracción; Estrategia 1 (muestras: IK, IM, M3, M9 y RN). 53

Figura 18. Gráfico de volcano representativo de los microRNAs abundantes y significativos por fracción; Estrategia 2 (muestras: IK, IM y RN). El código de color es persistente con el de la figura 17 53

Figura 19. Gráfico de volcano representativo de los microRNAs abundantes y significativos por fracción; Estrategia 3 (muestras: BJ, M3 yM9). El código de color es persistente con el de la figura 17 54

Figura 20. Intersección de las listas de microRNAs enriquecidos en exosomas entre cada estrategia de análisis descrito en la tabla 15 55

Figura 21. Elemento de acción en cis 1 encontrado a lo largo de las tres estrategias experimentales para la búsqueda elementos de la fracción exosomal. El valor E para cada estrategia se añade debajo de cada una de las figuras. 59

Figura 22. Abundancia de los eventos de adición simple, adenilación, a lo largo de las bibliotecas correspondientes a la fracción celular (Panel derecho) y exosomal (Panel izquierdo). 63

Figura 23. Abundancia de los eventos de adición Múltiple, Adenilación, a lo largo de las bibliotecas correspondientes a la fracción celular (Panel derecho) y exosomal (Panel izquierdo). 64

Figura 24. Abundancia de los eventos de adición simple, Uridinilación, a lo largo de las bibliotecas correspondientes a la fracción celular (Panel derecho) y exosomal (Panel izquierdo). 65

Figura 25. Abundancia de los eventos de adición Múltiple, Uridinilación, a lo largo de las bibliotecas correspondientes a la fracción celular (Panel derecho) y exosomal (Panel izquierdo). 66

Figura 26. Conteos normalizados de microRNAs reportados con eventos de tailing en un gráfico de balón (Balloonplot) a lo largo de las estrategias experimentales: Fracción exosomal. 70

Figura 27. Heatmap comparativo del repertorio de miRNAs enriquecidos en la fracción exosoma con un elemento de acción en cis en común a lo largo de las bibliotecas. 73

Figura 28. Diagrama de venn de exo-microRNAs con el elemento de acción en cis 1 75

Lista de tablas

Tabla 1. Ejemplo de formato fastq.....11

Tabla 2. Herramientas implementadas durante esta tesis con una breve descripción de su funcionamiento 16

Tabla 3.Casos de estudio analizados durante la tesis 18

Tabla 4. Script en R implementado para descargar metadatos de bibliotecas que se analizaran 19

Tabla 6. Línea de comando de módulo quantifier de la herramienta mirdeep2 implementado durante esta tesis. 23

Tabla 7. Línea de comando para la normalización de los conteos implementando la herramienta edgeR desde su interfaz de lenguaje R 24

Tabla 8. Línea de comando para calcular la correlación de los conteos normalizados 24

Tabla 9. Código implementado para ejecutar la herramienta Tailor. 31

Tabla 12. Implementado el comando grep y wl del lenguaje Linux se cuantifico el número de lecturas que contienen las bibliotecas en bruto y después del pre-procesamiento. Ej. cat sample.fastq | grep "@" | wc –l >1 number_reads. 33

Tabla 13. Numero de lecturas mapeadas contra microRNAs de referencia en comparativa con el número de lecturas totales de las bibliotecas y mapeadas contra el genoma de referencia 40

Tabla 15. Diseños de las estrategias para el análisis binomial de la expresión diferenciada 48

Tabla 16. Tabla de los miRNAs enriquecidos por fracción y significativos.....52

Tabla 17. Tabla de los miRNAs enriquecidos significativamente (Valor P> 0.05) por fracción.....52

Tabla 18. Elemento de acción cis 1. Listas de exo-microRNAs que presentaran el elemento de acción cis 1 en común.....60

Tabla 19. Lista de exo-microRNAs que conservaron el elemento de acción cis 1 en común en adición al valor LogFC del enriquecimiento exosomal. 74

Tabla 20. Distribución de hebras a lo largo de la lista de microRNAs que conservan el elemento de acción cis 1 en común.....76

Tabla 21. Distribución de hebras 5p y 3p de las listas de exo-miRNAs significativos a lo largo de las estrategias de análisis. 77

Tabla 22. Lista de exo-miRNAs con un elemento de acción cis en común sin intersección aparente entre estrategias. Esta tabla se interpreta comparando las celdas de las estrategias 1 a 3. También se puede comparar el enriquecimiento relativo de los microRNAs entre las estrategias 79

Capítulo 1. Introducción

La secreción de exosomas no pasó de ser una observación con tintes de artefacto experimental cuando Raposo y colaboradores describieron este fenómeno en 1996 como un mecanismo de comunicación entre células dendríticas y linfocitos T (G Raposo et al., 1996). Décadas después la comunidad científica vuelve sus ojos a los exosomas, inspirada por evidencia experimental que demuestra la importancia funcional de estas vesículas de secreción. En animales, los exosomas son esferas lipídicas de origen endosomal secretadas a través de fluidos corporales y cuyo contenido molecular refleja, por un lado, el estado patológico/fisiológico de la célula donde se originan (Chahar, Bao, & Casola, 2015) mientras que por otro, altera el estado de la célula blanco que invagina el exosoma (Colombo, Raposo, & Théry, 2014).

Además de transferir proteínas y lípidos, se ha reportado también la presencia selectiva de una clase de ácidos ribonucleicos (RNAs) de interferencia (microRNAs) en adición a proteínas, RNAs mensajeros (mRNAs) y otros RNAs no codificantes como cargamento exosomal (Valadi et al., 2007). Los microRNAs (miRNAs) regulan la expresión génica a nivel pos-transcripcional (Appasani, 2009; Bartel, 2004) y su transferencia exosomal juega un papel activo durante cáncer o infecciones microbianas.

Aunque existen indicios sobre el empaquetamiento de microRNAs (y otros RNAs) en exosomas, los factores cis y trans con los cuales el empaquetamiento ocurre sigue siendo un paradigma por resolverse (Graça Raposo & Stoorvogel, 2013). Lo anterior ha derivado en múltiples estudios de exploración para comprender mejor este mecanismo de comunicación intercelular en microbios y humanos implementando métodos de secuenciación masiva. Los estudios de secuenciación masiva son en su mayoría de descubrimiento solamente, lo cual da pauta a que el re-análisis de los datos generados se pueda hacer desde múltiples puntos de vista, nutridos por hipótesis específicas. Recientemente, se ha implementado como carácter obligatorio en la mayoría de las revistas de publicación arbitrada por pares que al ser publicados, los datos generados en estudios de transcriptómica y proteómica sean depositados en bases de datos de acceso público. Esto ha derivado en la posibilidad del re-análisis de datos a partir de una hipótesis distinta a la que sustentó la publicación original. Lo anterior fundamenta el potencial innovador de esta tesis de

maestría – el re-análisis de datos depositados en repositorios, utilizando herramientas bioinformáticas e hipótesis que no fueron consideradas en la publicación en la que se reportaron los datos. El re-análisis de datos de secuenciación masiva es una tendencia importante (Rung & Brazma, 2012), la cual va en incremento. Varios ejemplos son los siguientes*: A reanalysis of mouse ENCODE comparative gene expression data* (Gilad & Mizrahi-Man, 2015), *Gene-expression analysis identifies global gene dosage sensitivity in cancer* (Fehrmann et al., 2015), *Meta-analysis of RNA-seq expression data across species, tissues and studies* (Sudmant, Alexis, & Burge, 2015), *Re-analysis of RNA-seq data transcriptoma data reveals new aspects of gene activity in Arabidopsis root hairs* (Li & Lan, 2015) y *General rules for functional microRNA targetings* (Kim et al., 2016).

Capítulo 2. Fundamentos

2.1.1 Diversidad celular

Los pilares moleculares de la vida suelen referirse a ácidos nucleicos, proteínas, lípidos y carbohidratos. Dadas sus propiedades, estas moléculas pueden interactuar física y químicamente para formar y mantener la vida celular. Por otro lado, las células han sido clasificadas en un sistema de tres dominios: Bacteria, Arquea y Eucariotas, todos ellos emergentes de la teoría del ancestro universal común (LUCA). Mientras que las bacterias y arqueas abarcan poblaciones de microorganismos unicelulares, el dominio eucariota suele enmarcar clados multicelulares y algunos organismos unicelulares como los son los parásitos protistas y levaduras. Además, en organismos multicelulares, las células se agregan para formar tejidos. Los tejidos a su vez componen un organismo macroscópico. Tanto hongos, animales y plantas son ejemplos de estos organismos (Craig *et al*., 2010).

2.1.2 Membranas biológicas

Las membranas biológicas son la frontera que delimita la interacción y organización celular así como la polaridad hacia solutos. Su función principal es mantener la composición y estado celular, además de asegurar la compartimentalización en su interior donde se establece la actividad de ciertas biomoléculas y el transporte de otras hacia el espacio citoplásmico a través de la membrana plasmática. Las membranas biológicas están constituidas principalmente por fosfolípidos y proteínas transmembrana que controlan el transporte de moléculas a través de la membrana y medían la interacción entre células vecinas (Lodish et al., 2008).

2.1.3 Genoma, DNA, mRNA y ncRNAs

Los seres vivientes comparten una característica inherente denominada genoma, el cual es el repositorio de la información necesaria para el inicio y propagación de la vida. El genoma es una larga cadena compuesta de bloques moleculares de DNA (Acido desoxiribonucleico). A lo largo del DNA se encuentran las instrucciones para la síntesis de productos proteicos necesarios durante el desarrollo celular; estas instrucciones suelen referirse a genes (Gerstein et al., 2007).

Durante la expresión génica (DNA → mRNA → proteína), el proceso de la transcripción es el primer paso para la síntesis de los productos funcionales a través de la polimerización de cadenas de ácidos ribonucleicos (RNA). La síntesis de RNAs puede producir transcritos codificantes de proteína (RNA mensajero – mRNA) y otros que no codifican (RNA no codificante – ncRNA) pero que tienen una función biológica (Cech, 2012). Una de las funciones de los ncRNAs puede derivar en la regulación de los niveles intracelulares de proteínas por un mecanismo específico mediado por RNAs de interferencia (RNAi), lo que corresponde a una variante del dogma central de la expresión de genes que resulta fascinante. Los RNAi fueron primero descubiertos en plantas como un mecanismo de silenciamiento de la expresión génica que funciona de manera autónoma o exógena (Okoye et al., 2014). Posteriormente, reportes en nematodos ampliaron la diversidad de este mecanismo pero esta vez, se definieron como microRNAs.

2.1.4 MicroRNAs: origen, biogénesis y funcionamiento

Los microRNAs (miRNA) han sido ampliamente estudiados en metazoarios que van desde *Caenorhabditis elegans* (Lee, Feinbaum, & Ambros, 1993; Wightman, Ha, & Ruvkun, 1993) a mamíferos (Friedman, Farh, Burge, & Bartel, 2009), como un mecanismo para la regulación pos-transcripcional de genes en procesos de desarrollo celular como: diferenciación, proliferación y muerte, segregación cromosómica y metabolismo (Appasani, 2009). Estos productos corresponden a RNAs endógenos de ~22 nucleótidos de longitud (nt) que pueden provenir de transcritos independientes (miRNA canónicos) o como derivados de intrones o exones (Bartel, 2004; Krol, Loedige, & Filipowicz, 2010). Los miRNA se unen por complementariedad a la región

no codificante (Untranslated Region, UTR) 3´- de mRNAs guiando al complejo de silenciamiento RISC (RNA *induced-silencing complex)* para interferir con la traducción de proteínas (Ameres & Zamore, 2013).

Durante la maduración canónica de miRNAs, una orquilla dúplex conocida como miRNA primario (pri-miRNA) es reconocido e hidrolizada por el microprocesador DROSHA/DGCR8 para formar un miRNA precursor (pre-miRNA) (Nguyen et al., 2015). Posteriormente los pre-miRNAs son transportados por alguna exportina hacia el citoplasma para finalizar su biogénesis y posterior funcionamiento. En el citoplasma los pre-miRNAs son hidrolizados por RNasas tipo III conocida como DICER en la zona apical del pre-miRNA dejando únicamente una doble hebra que es reconocida por una de las distintas proteínas Argonauta (AGO). Mientras que AGO sirve como efector, una sola hebra de miRNA es seleccionada para actuar como guía durante el silenciamiento en la expresión génica. Usualmente el complejo RISC se coloca en la región UTR 3´- para interrumpir la traducción ribosomal de RNAs mensajero; en otros casos, cuando la complementariedad del miRNA es perfecta con la región UTR 3´-, los RNAs son cortados para su degradación (Bartel, 2009; He & Hannon, 2004).

Estudios computacionales sugieren que el genoma humano contiene más de 1000 loci de miRNAs, los cuales probablemente modulen el 50% de los genes codificantes de proteínas (Gilbert, 2014). Además de su abundancia, se ha especulado que los miRNAs son altamente conservados entre especies (Friedman et al., 2009) y su expresión espacio-temporal modula la producción fisiológica de diversos productos proteicos durante el desarrollo animal (Ason *et al*., 2006).

2.1.5 Vesículas extracelulares: un mecanismo de comunicación intercelular.

El transporte de biomoléculas es altamente dinámico en el interior y exterior celular (Hurley, Boura, Carlson, & Różycki, 2010). Las biomoléculas que son dirigidas hacia el espacio extracelular suelen participar en el fenómeno de comunicación intra- e intercelular; llamadas también comunicación autócrina y parácrina, respectivamente. En la gran mayoría de organismos, las células se comunican entre sí a través de la secreción de biomoléculas (proteinas y segundos

mensajeros como lipidos y carbohidratos). Estas biomoléculas se asocian a otras células, induciendo una señal y respuesta que modifica el estado fisiológico de la célula blanco (Mathivanan, Ji, & Simpson, 2010).

Uno de los mecanismos para la comunicación intercelular esta mediado por vesículas secretadas al espacio extracelular (EVs, por sus siglas en inglés) (Colombo et al., 2014). El término transferencia de biomoléculas de carga vesicular, suele usarse para definir a productos que viajan de una célula origen a otra célula blanco a través de la compartimentalización, secreción y fusión de las EVs; estas biomoléculas incluyen ácidos nucléicos, proteínas y lípidos (Janas, Janas, Sapoń, & Janas, 2015). Aunque estudios pioneros sugirieron una función biológica de las EVs (Pan, Teng, Wu, Adam, & Johnstone, 1985; G Raposo et al., 1996), estas permanecieron en la creencia de que eran derivados de la degradación (basura) celular o artefactos experimentales (Théry, 2011). No obstante, la secreción de EVs es modulada en diferentes tipos de células del sistema inmune para adaptarse a cambios en el ambiente (Robbins & Morelli, 2014). La diversidad natural de las EVs ha sido tema de discusión para su nomenclatura en la literatura científica (Gould & Raposo, 2013), refiriéndose a ellas por diversos nombres en base a su tamaño, origen celular, función reportada o simplemente por su presencia extracelular (usando el prefijo exo o ecto) (Colombo et al., 2014).

En este documento usaremos el término acuñado en la literatura como “exosomas” para referirnos a nano-vesículas de entre 50 -100 nm de diámetro que se originan dentro del endosoma multivesicular (MVE) como vesículas intraluminales (ILVs, por sus siglas en inglés) hasta su liberación al espacio extracelular al fusionarse con la membrana plasmática.

2.1.6 Exosomas: biogénesis de vesículas intraluminales en el endosoma

Debido a su origen endosomal, los exosomas pueden distinguirse morfológica y bioquímicamente de las demás EVs que se originan en la membrana plasmática (ie. vesículas apoptóticas y micro-vesículas o micro-partículas) (György *et al.*, 2011). Los MVEs pueden formarse y coexistir dentro de una misma célula para cumplir con distintos destinos: degradación de carga y secreción de

exosomas o fábricas virales (*"viral Factories"*) (Diaz & Ahlquist, 2012; Gould, Booth, & Hildreth, 2003).

Como ya se mencionó en párrafos anteriores, antes de su liberación, los exosomas son generados como ILVs dentro del endosoma. Para ello se han reportado dos mecanismos: independientes o dependientes del complejo ESCRT (por sus siglas en inglés, *Endosomal Sorting Complex Responsible for Transport)* (Marsh & van Meer, 2008) que procesan la biogénesis de ILVs. Por un lado, el complejo ESCRT está dividido en cuatro sub-complejos proteicos que además, se asocian a proteínas accesorio durante la formación de ILVs (Piper & Katzmann, 2007); durante la formación dependiente de ESCRT, el sub-complejo ESCRT-0 reconoce una señal de ubiquitinación (Ub) de proteínas transmembrana en la cara citosólica del endosoma. Esto permite la unión de ESCRT-I quien recluta a ESCRT-II para iniciar la biogénesis de los ILVs nacientes. Por último, el sub-complejo ESCRT-III es reclutado para finalizar la formación de ILVs dentro del Endosoma hasta obtener el MVE. Aunque la señal de Ub es fundamental para colocar al complejo ESCRT, existen proteínas que no requieren de esta señal para formarse y liberarse como exosomas (Robbins & Morelli, 2014). Entre la composición lipídica enriquecida de los MVEs de secreción exosomal están el colesterol (Mobius et al., 2002) y los esfingolípidos-ceramidas. De hecho, las ceramidas pueden desencadenar la formación de ILVs dentro del endosoma en un mecanismo independiente del complejo ESCRT (Kosaka et al., 2010; Trajkovic et al., 2008).

2.2 Cargamento exosomas: RNAs circulantes para la expresión y reprogramación celular

El término de exosomas fue abordado por Johnstone y colaborarodess (1987) para describir nano-vesículas formadas dentro del endosoma multivesicular. Estos exosomas observados por microscopía electrónica transferían receptores acarreadores de hierro (TfR) entre reticulocitos durante su madurez (Johnstone, Adam, Hammond, Orr, & Turbide, 1987). Estudios *in vitro* e *in vivo* han reportado que los exosomas pueden modular procesos inmunológicos, tanto regulatorios (Février & Raposo, 2004; Mittelbrunn *et al*., 2011; Montecalvo *et al*., 2012; Zitvogel *et al*., 1998) como supresores del sistema inmune (Bergmann et al., 2009; Bobrie et al., 2012; Meckes et al., 2010; Peinado et al., 2012). Este espectro de funciones se debe a la diversidad de biomoléculas de carga que los exosomas pueden transferir.

Durante la formación de ILVs en el endosoma, diversas biomoléculas pueden ser empaquetadas para su liberación. Además de transferir proteínas y lípidos, estudios recientes reportaron la presencia de ácidos ribonucleicos de carga exosomal como mRNAs y microRNAs (Valadi et al., 2007). La relevancia de estos hallazgos es la posibilidad (confirmada en el caso del virus Epstein-Barr, EBV) de que los RNA de carga exosomal una vez introducidos a una célula blanco, puedan ser traducidos a proteínas, en el caso de los mRNAs, o puedan reprimir la expresión de genes en el caso de los miRNAs (Lotvall & Valadi, 2007). Aunque algunos autores rechazan la presencia de RNAs de carga vesicular (visto en: (Turchinovich, Tonevitsky, Cho, & Burwinkel, 2015)), numerosos trabajos reportan la presencia de diversas especies de RNAs exosomales.

Es interesante mencionar que algunos de los RNA reportados en exosomas presentan un coeficiente de abundancia relativamente superior al de los niveles intracelulares; acorde a esto, durante esta tesis reportamos que RNAs de entre 18-24 y 30-25 conservan un patrón en la abundancia en los RNAs exosomales respecto a la fracción celular; además, el análisis bayesiano demostró el enriquecimiento relativo en la fracción exosomal de datos NGS de entre 18-24 nucleótidos de longitud correspondientes de una línea celular en tres condiciones experimentales distintas lo cual apunta a una especificidad en el proceso de empaquetamiento activamente selectivo. Por otro lado, se ha descrito un mecanismo independiente de exosomas para secretar RNAs llevado a cabo por la fusión de complejos RNA/proteínas con la membrana plasmática (Ej. complejo circulante *miRNAs-AGO2*) (Arroyo et al., 2011). Lo anterior sugiere un mecanismo de especificidad en el empaquetamiento, muy posiblemente mediado por la unión de complejos ribonucleoproteicos (RNPs, por sus siglas en inglés) con la capacidad de dirigir RNAs hacia el MVE.

2.2.1 Exo-miRNAs: cáncer, inmunosupresión e inmuno-evasión por parásitos.

Los miRNAs de carga exosomal tienen un papel activo durante diversos procesos biológicos, incluyendo la sinapsis inmunológica (IS) (Mittelbrunn et al., 2011), metástasis (Zhou et al., 2014) y la inmuno-represión mediada por patógenos (Buck et al., 2014; Pegtel, van de Garde, & Middeldorp, 2011). A continuación se explican algunos ejemplos de ello. En el citoesqueleto, las

proteínas de unión estrecha "*tight junction protein*" ZO-1 proporcionan adhesión celular y permeabilidad selectiva en el tejido epitelial (Fanning, Jameson, Jesaitis, & Anderson, 1998). Durante el desarrollo de cáncer de mama la sobreexpresión intracelular y transferencia vía exosomas de miR-105 (normalmente expresado en bajas cantidades) reduce la síntesis de esta proteína, debilitando la barrera de la capa epitelial y promoviendo metástasis (Zhou et al. 2014). Por otro lado, parásitos infecciosos gastrointestinales pueden tener la capacidad de secretar diferentes miRNAs vía exosomas para tomar el control de la respuesta inmune innata en células del epitelio gastrointestinal. En modelos de ratón se ha reportado que *Heligmosomoides polygyrus* secretan exo-miRNAs para dirigir el silenciamiento de *Il33r* y *Dups1* (Buck et al. 2014). Normalmente, la expresión de estos genes celulares están involucrados en procesos inflamatorios y de respuesta inmune que media la muerte de *H. polygyrus.* Un tercer ejemplo aborda a la sinapsis unidireccional de células T hacia células presentadoras de antígeno (APCs) durante el reconocimiento de antígenos mediada por exo-miRNA (Mittelbrunn et al. 2011). Como último ejemplo, se ha reportado *in vitro* que el herpesvirus-4 EBV (por sus siglas en inglés, Epstein-Barr Virus) expresa un grupo de miRNAs de latencia (BART-miRNAs) que son empaquetados en exosomas para su transferencia hacia células B no infectadas, sirviendo como un mecanismo inmuno-evasivo (Pegtel et al., 2010). Se ha sugerido que los miRNAs de carga exosomal pueden implementarse como una estrategia terapéutica para el diagnóstico y/o terapia de diversas enfermedades.

2.3 Tecnologías de secuenciación

Con el reciente avance en tecnologías secuenciación de RNA (RNA-seq), los datos genómicos han incrementado cuantitativa y cualitativamente, complementando la biología tradicional (Nussinov, Bonhoeffer, Papin, & Sporns, 2015). Una de las ventajas de secuenciar parcial o totalmente el transcriptoma, es que la información obtenida puede revelar la cantidad, actividad y localización de mRNAs, isoformas y ncRNAs (Kukurba & Montgomery, 2015). Estas herramientas han permitido ampliar el conocimiento sobre el funcionamiento que los microRNAs pueden tener, incluyendo su capacidad para mediar comunicación intercelular**.**

2.3.1 Secuenciación de RNAs pequeños.

La tecnología RNA-seq ha suplementado los microarreglos como la herramienta por elección para conocer los perfiles de la expresión de la variedad de transcritos a lo largo de diversas condiciones experimentales. En un típico ensayo de RNA-seq, el RNA purificado de los experimentos biológicos es convertido a DNA complementario (cDNA) y secuenciado en una de las plataformas de siguiente generación (*high-throughput).* Los equipos *high-throughput* generan bibliotecas con millones de lecturas que pueden variar de 25 a 300 nucleótidos de longitud a partir de un solo extremo (single-end) o ambos (paired-end) de los fragmentos de cDNA (Chen. Yunshun, T. L. Aaron, 2014). *A posteriori*, estas bibliotecas requerirán de un procesamiento de cascada abajo *"downstream"* que valide la información obtenida en la secuenciación. Por esencia, las lecturas deberán alinearse a las características genómica del organismo cuyo RNA fue purificado. En el caso de esta tesis, el primer paso es implementar bibliotecas de datos correspondientes al genoma humano y viral para determinar la proporción de lecturas de las bibliotecas experimentales que mapean al genoma de referencia, de este modo se descifra la presencia de contaminantes biológicos hallados durante la secuenciación. Posteriormente, se implementan algunas herramientas de bioinformática para escudriñar las bibliotecas y generar un resumen de las lecturas (conteos) que se alinean a otras características genómicas (Ej. exones, mRNAs, microRNAs, etc). Ambos listados, de conteos y su respectiva características genómicas (Ej. miR-501) pueden ser asociadas con las condiciones biológicas de las que provienen las bibliotecas de secuenciación para conocer la expresión de genes (Liao, Smyth, & Shi, 2014).

2.3.2 Formato fastq

Usualmente, las técnicas de RNA-seq de cualquier tipo conllevan errores técnicos que disminuyen la veracidad de los nucleótidos secuenciados; las bibliotecas en su formato de archivo fastq contienen métricas que expresan la calidad de cada lectura secuenciada en unidades estándar conocidas como Puntaje Phred. Básicamente, el puntaje Phred consiste en un carácter ASCII que codifica un valor de calidad para cada nucleótido de la lectura secuenciada (Base-call position, BCP), El valor Phred está correlacionado con la probabilidad que el nucleótido secuenciado sea

incorrecto; mientras más bajo es el valor, mayor es la probabilidad de error. En principio, el formato fastq está estructurado por tres filas de información para cada lectura secuenciada (tabla 1).

Tabla 1. Ejemplo de formato fastq

@EAS54_6_R1_2_1_443_348	← Identificador de lectura secuenciada
GTTGCTTCTGGCGTGGGTGGGGGGG	← Lectura secuenciada
+EAS54_6_R1_2_1_443_348	
;;;;;;;;;;;9;7;;.7;393333	← Puntaje de calidad Phred para cada nucleótido en la secuencia

Como estándar, la primera línea inicia con un signo @ para denotar un identificador de la lectura seguido de una segunda fila que corresponde a la lectura secuenciada. La última fila corresponde al puntaje de calidad para cada nucleótido de la secuencia.

2.3.3 Análisis *"downstream"* de bibliotecas de secuenciación

El análisis de la información procedente de la secuenciación inicia pre-procesando las bibliotecas de interés. El pre-procesamiento permite eliminar los errores técnicos durante la secuenciación (eliminando lecturas de puntaje de calidad Phred bajo). Más tarde, las bibliotecas pre-procesadas son implementadas como entrada para programas que realizan el conteo de las características genómicas de interés. Antes de comparar otras mediciones, los datos de recuento tienen que ser normalizados. En un sentido práctico, la normalización permite comparar y hacer mediciones de los datos adimensionales.

Capítulo 3. Antecedentes

Se ha sugerido que existe una correlación entre los niveles de los microRNAs en exosomas e intracelulares. No obstante, esta correlación podría ser un producto de la difusión pasiva de microRNAs durante la formación de vesículas intraluminales en el Endosoma. Por otro lado, se ha reportado un enriquecimiento selectivo de ciertas especies de microRNAs sugiriendo un empaquetamiento selectivo (Cha et al., 2015; Koppers-lalic et al., 2014; Mittelbrunn et al., 2011; Villarroya-beltri, Gutie, Martin-cofreces, Martinez-herrera, & Pascual-montano, 2013). Aunque existen indicios sobre el empaquetamiento de mRNAs en exosomas, el mecanismo mediante el cual los miRNAs son incorporados a los exosomas es un paradigma por resolverse (Graça Raposo & Stoorvogel, 2013).

Una hipótesis es que al igual que los mRNAs (Batagov, Kuznetsov, & Kurochkin, 2011), los miRNAs son empaquetados mediante elementos de la secuencia de acción *cis*. Villarroya et. al. (2013) reportan cierto patrón en elementos de acción *cis* recurrentes en miRNAs enriquecidos en exosomas (exo-miRNAs) provenientes de un modelo de célula T humanas infectadas con retrovirus. Un análisis proteómico detectó la unión específica de la proteína hnRNPA2B1 a elementos de acción en *cis* del exo-miRNA miR-198*.* Además, esta proteína se encontró presente en dichos exosomas sugiriendo un empaquetamiento controlado por hnRNPA2B1 y la señal de empaquetamiento de miR-198 (Villarroya-beltri et al., 2013). En adición, Santagelo et al. (2016) reporta un repertorio de exo-miRNAS con elementos de acción en cis que interactúa con la proteína de unión a RNA SYNCRIP (Santangelo et al., 2016); ambos reportes sugieren que los elementos de accion en cis podrían representar un factor del empaquetamiento de microRNAs en exosomas.

Una segunda posibilidad es que las modificaciones pos-transcripcionales pueden influir en la trasferencia de miRNAs de carga exosomal (Koppers-lalic et al., 2014; Warf, Shepherd, Johnson, & Bass, 2012) Kopper y colaboradores (2014) reportan la adición de nucleótidos uracilo en la terminación 3´- de microRNAs exosomales mientras que los niveles de miRNAs adenilados en el extremo 3´- están enriquecidos en el medio intracelular de sistemas modelo de células B humanas infectadas por el virus EBV. Esa modificación pos-transcripcional podría tener un efecto en la

estabilidad de miRNAs dentro de exosomas, por lo que podría considerarse la Uridinilación como un rasgo del empaquetamiento exosomal. Además, se sugiere que estas modificaciones pueden promover la asociación o disociación de proteínas de unión a RNA (RNA-binding protein/RBPs, por sus siglas en inglés) para el empaquetamiento de miRNAs. Nos resulta importante mencionar que estas hipótesis corresponden al estudio con modelos de células humanas B, lo cual deja espacio para la investigación de exo-miRNA en otros sistemas modelo.

Finalmente, Cha et al., (2015) han reportado que los perfiles de exo-miRNAs son distintos de los perfiles miRNAs celulares durante cáncer colorrectal. Además, dichos autores sugieren que no existe una señal global en elementos de la secuencia de acción en *cis* que induzca el empaquetamiento exosomal, como es el caso de los miembros de la familia de microRNAs miR-320, que presentan elementos de la secuencia conservados (GCAG). Sin embargo, esta secuencia no fue observada en otros miRNAs (Ej. miR-218-5p) que también son secretados vía exosomas. Quisiéramos indicar que en ambos casos se trabajó con distintos sistemas experimentales, lo cual pudiera indicar un sesgo en los resultados y la necesidad de una estrategia experimental más adecuada.

Capítulo 4. Hipótesis

Al igual que en el caso de mRNA, el empaquetamiento de microRNAs parece ser selectivo y podría estar mediado por elementos de secuencia con acción regulatoria en cis así como modificaciones pos-transcripcionales. Estos elementos de especificidad van a variar en diferentes organismos (ej. nematodos vs humano vs virus), fenotipos celulares (linfocitos vs epitelio) o eventos biológicos (ej. cáncer vs normal).

Capítulo 5. Objetivos

5.1 Objetivo general

Mediante un análisis computacional de datos RNA-seq dirigido por la hipótesis en cuestión "*hypothesis-driven*", identificar y comparar entre sistemas biológicos qué microRNAs son selectivamente enriquecidos en exosomas y qué factores potenciales son los responsables para su enriquecimiento.

5.2 Objetivos específicos

1. Comparar los niveles de microRNAs exosomales y celulares a partir de conteos normalizados de bibliotecas de RNA-Seq.
2. Identificar y comparar elementos de secuencia "*sequence motifs*" de microRNAs que posiblemente regulen su empaquetamiento en exosomas en diferentes sistemas de estudio.
3. Identificar y comparar la adición pos-transcripcional de nucleótidos en la terminación 3′- de los microRNAs empaquetados en exosomas en diferentes sistemas de estudio.

Capítulo 6. Metodología

Para la metodología de este proyecto es importante considerar las siguientes características, la procedencia de los datos que se analizarán, los programas computacionales que se ajusten en tiempo, requerimientos y función al proyecto, los archivos de entrada que estos programas requieren y la capacidad de cómputo con la que se cuenta para implementar los programas seleccionados. Mientras que el análisis de elementos de acción *cis* puede proceder con cualquier dato de secuenciación, incluyendo arreglos de hibridación y NGS, las herramientas para el análisis de las ediciones pos-transcripcionales se basan exclusivamente en datos NGS de RNAs pequeños. Entonces, para el análisis propuesto en este documento nos restringiremos al uso de bibliotecas NGSs de RNAs pequeños basados en dos sistemas de estudio distintos; dichas bibliotecas se encuentra disponible en la base de datos de NCBI (http://www.ncbi.nlm.nih.gov/geo). La tabla 2 incluye las herramientas implementadas durante esta tesis.

Tabla 2. Herramientas implementadas durante esta tesis con una breve descripción de su funcionamiento

HERRAMIENTA	DESCRIPCIÓN
SCRIPT EN R	Extracción de datos NGS de la base de datos NCBI
FASTQC, FASTX TOOLKIT	Pre-procesamiento y visualización de calidad de datos NGS
MIRDEEP2 FRAMEWORK	
MAPPER.PL	Mapeo de lecturas al genoma de referencia
QUANRIFIER.PL	Anotación y cuantificación de microRNAs
R: BIOCONDUCTOR:	
EDGER	Análisis estadístico de expresión de microRNAs
DPLYR, GPLOT, GGPLOT2	Cribado y presentación de los resultados
MEME	Búsqueda de elementos de acción *cis*
TAILOR PIPELINE	Análisis de eventos de adición de nucleótidos

6.1 Conexión al servicio computacional de cluster en CICESE

El centro de Investigación Científica y Educación Superior de Ensenada, México cuenta con dos infraestructuras de cómputo de alto rendimiento con arquitectura cluster (cluster Omica e Ixachi, respectivamente) para la ejecución y desempeño óptimo de programas de alta demanda computacional. Para el acceso a cualquiera de estos dos clusters es necesario solicitar al departamento de telemática una cuenta usuario@cluster, la cual permitirá tener acceso a los programas informáticos que implementaremos desde la red de CICESE.

La conexión al servidor ixachi se realizó de la siguiente manera desde la línea de comandos en Ubuntu:

```
[user@ubuntu]$ ssh rgomez@cluster
> password: *******
[rgomez@cluster~]$: pwd
> /home/rgomez
```

De este breve código puede comprenderse que el signo “$” antepone la escritura de un comando ejecutable escrito por el usuario en la línea de comandos, mientras que el símbolo “>” nos indica la respuesta impresa del comando al ejecutarse. A través de este escrito, implementaremos el par de símbolos para ejemplificar la metodología realizada durante nuestros análisis hechos desde la “línea de comandos”**.**

6.2 Entorno de trabajo

Para conservar un entorno compatible con las infraestructuras antes mencionadas se recomienda el uso de la consola Linux a través de cualquiera de sus distribuciones comerciales. Durante el desarrollo de los análisis desarrollados durante esta tesis se utilizaron los equipos portátiles con las siguientes características: 1) RAM 4 GB, Intel i5 -4ta Generación, SSD 132 GB, distribución

Ubuntu 14 LTS x64; 2) RAM 8 GB, Intel Core i7 -6ta Generación, HD 1 TB, Sistema Operativo Windows 10. En ambos equipos se implementó la siguiente versión R, 3.1.1 y la siguiente sesión de trabajo, paquetes básicos adjuntos: stats, graphics, grDevices, utils, datasets, methods, base; loaded via a namespace , y no adjuntos: tools_3.3.0.

6.3 Obtención de datos NGS, genoma y microRNAs

Durante esta tesis se consideran dos fenotipos celulares: 1) Linfocitos B y 2) células epiteliales, ambos fenotipos de líneas celulares humanas. El primer caso de estudio corresponde a tres condiciones experimentales (Linfoma, Leucemia de Burkitt y células Linfoblastoides); dos de estos eventos biológicos se encuentran, además, en estadios de infección por un virus (EBV). El segundo caso de estudio corresponde a dos eventos biológicos: células epiteliales humanas sanas y células epiteliales humanas cancerígenas en estadio de metástasis. Esta información se resume en la tabla 3.

Tabla 3.Casos de estudio analizados durante la tesis

Fenotipo celular	**Evento biológico**	**Organismo**
Células B	Linfoblastoides	Línea celular humana, EBV +
	Leucemia de Burkitt	
	Linfoma	Línea celular humana, EBV -

Para cada sistema de estudio se generaron los respectivos archivos SraRunInfo.csv de metadatos desde el respectiva enlace web (Panel de opción *send to*, esquina superior-Derecha, Choose Destination: File, Download Format: Runinf, Create File). Utilizando un script en lenguaje R se descargó a un disco duro de la infraestructura ixachi la lista de datos SRR en el formato de archivo binario file.sra (tabla 4).

Tabla 4. Script en R implementado para descargar metadatos de bibliotecas que se analizaran

```
[rgomez@ixachi~]$: R
> R version 3.1.1 (2014-07-10) –
Copyright (C) 2014 The R Fundation for Statistical Computing
Plataform x86_64-redhat-linux-gnu (64-bit)

setwd(rgomez@cluster:/USB-5000/rgomez/data)

Srai <-read.csv("SraRunInfo.csv", stringsAsFactors=FALSE)
SRAfiles <-basename(Srai$download_path)
for (i in 1:length(files)) download.file (Srai$download_path[i], SRAfiles[i])
[rgomez@ixachi~]$: q()
```

Se utilizó la herramienta SRA Toolkit (Staff, 2011) previamente instalada en el cluster ixachi con el fin de convertir los archivos “file.sra” en el formato de entrada compatible “file.fastq” usando el siguiente comando desde el cluster: `fastq-dump /USB-5000/rgomez/data/file.sra`

Implementamos el código wget para descargar los genomas de referencia humano y del virus EBV como a continuación se indica:

```
[rgomez@ixachi~]$: wget ftp://hgdownload.cse.ucsc.edu/goldenPath/reference_genome/*.fa
```

Implementamos un script en bash para extraer la secuencia e ID de los miRNAs usados durante este trabajo como a continuación se muestra:

```
[rgomez@ixachi~]$: awk '{printf "@%s\t", substr($1,2); getline; l=length($1); printf "%s\n", $1;
}' mature.fa  > mature.csv
```

6.4 Pre-procesamiento de los datos NGS

El pre-procesamiento de las bibliotecas fastq implica básicamente los siguientes pasos: 1) Conservar solo lecturas de calidad obtenidas durante la secuenciación (ie. puntaje de calidad phred > 30 por nucleótido y > 80 por lectura), 2) remover lecturas que tengan un tamaño menor a 18 nucleótidos de longitud y 3) remover adaptadores añadidos durante la secuenciación. Se implementó el kit computacional *fastx* para esta labor, instalando los archivos pre-compilados en el siguiente enlace: http://hannonlab.cshl.edu/fastx_toolkit/download.html. Además, implementamos la herramienta computacional *fastqc* para conocer un resumen de calidad sobre las bibliotecas antes y después del pre-procesamiento, para detalles de esta herramienta visitar el sitio web de origen: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

6.5 miRDeep2: Identificación de miRNAs expresados en los datos NGS.

MiRDeep2 es una herramienta computacional eficaz para anotar las lecturas obtenidas en los datos NGS a un modelo biológico de miRNAs (Friedländer et al., 2008). miRDeep2 está dividido en 3 módulos: Mapper, Quantifier y mirDeep2 que pueden utilizarse en complemento o por separado. Para el funcionamiento correcto de esta herramienta se requiere instalar otros programas bioinformáticos como lo son bowtie, RNAfold-Vienna, SQUID y el paquete perl PDF:API2 (ver el archivo README de miRDeep2 para información detallada, https://www.mdc-berlin.de/8551903/en/).

6.5.1 Módulo Mapper

El script del módulo mapper.pl es útil para mapear las lecturas de los archivos fastq pre-procesados a un (o varios) genomas de referencias. La figura 1 presenta el flujo de trabajo de este módulo:

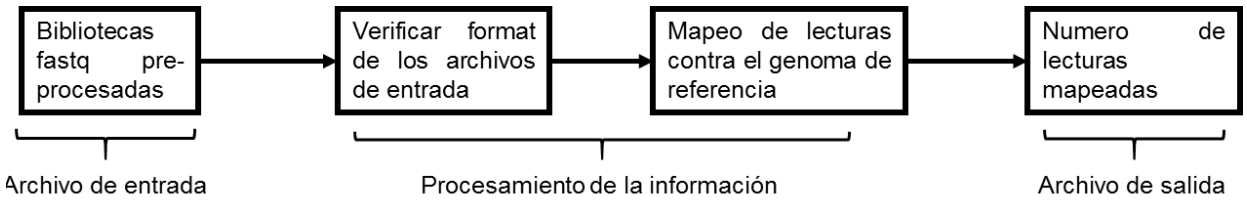


Figura 1. Esquema de flujo de trabajo de la herramienta bioinformática mirDeep2, modulo mapper.pl.

Este módulo toma como archivo de entrada los datos NGS crudos o pre-procesados de las bibliotecas NGS. Puesto que trabajamos con más de una biblioteca que se almacena en un formato seq.txt, file.fasta/file.fastq o un archivo config.txt que lleva una lista de diferentes archivos de entrada. El archivo config.txt de la tabla I contiene dos columnas de doce filas correspondientes a la columna “ID Nuevo” y “ID y se crea con el comando gedit `filename.txt`:

```
[rgomez@ixachi~]$: gedit config.txt
 BJC    SRR1563015.fastq
 BJE    SRR1563017.fastq
 RNC    SRR1563063.fastq
 RNE    SRR1563064.fastq
 ...    ...
```

Para optimizar el mapeo, los genomas de referencia implementados durante esta tesis (genoma humano y del virus EBV) fueron previamente indexados mediante un algoritmo que implementa la transformada de Burrows-Wheeler utilizando la herramienta bowtie-build. A continuación se

muestra la línea de comandos (Tabla 5) utilizados para las bibliotecas que fueron mapeadas contra los genomas de referencia obtenidos del portal bioinformático de la University of California Santa Cruz (UCSC). Lea el archivo *readme* del programa para comprender el significado de las literales.

Tabla 5. Línea de comandos para ejecutar la herramienta bowtie (para indexar los genomas) y mapper (para mapear las lecturas secuenciadas).

```
[rgomez@ixachi~]$: bowtie-build hg38_ebvg.fa genomes.index && mapper.pl config.txt -d -e -j -m -p genomes_index -s reads_collapsed.fa –t reads_collapsed_vs_hgenome.arf -v 2>reportmapper.log
```

Este módulo colapsa las lecturas obtenidas en el secuenciador en grupos de número de incidencias de las lecturas del archivo en un archivo llamado reads_collapsed.fa, de este modo se crean índices para el mapeo con el genoma de referencia. Por otro lado, el archivo reads_collapsed_vs_hgenome.arf contiene la información detallada de los alineamientos de las lecturas colapsadas con el genoma de referencia, incluyendo la ubicación cromosomal del alineamiento, el tamaño de la lectura alineada, la secuencia alineada, la secuencia cromosomal y la dirección cromosomal del alineamiento.

6.5.2 Módulo quantifier

El objetivo de este módulo es re-mapear las lecturas colapsadas en el archivo reads_collapsed.fa contra los miRNAs conocidos de la especie en estudio anotados en la base de datos de miRNAs miRBase (Griffiths-Jones, Grocock, van Dongen, Bateman, & Enright, 2006; Kozomara & Griffiths-Jones, 2014) a modo de conocer el perfil de expresión de miRNAs de las lecturas de la secuenciación, obteniendo un primer acercamiento hacia aquellos miRNAs presentes en la fracción exosomal y celular respectivamente sin un análisis estadístico correctivo. El script extract.pl extrae de la base de datos mirbase (mirbase.org) la lista de miRNAs maduros, precursores y stars procedentes de diversos organismos modelo de interés. Para esta actividad implmentamos la versión de mirbase 21 para las bases de datos de miRNAs humanos y del virus

EBV. El siguiente fragmento de código (tabla 6) expresa el modo en cómo se configuraron los comandos de este módulo:

Tabla 6.Línea de comando de módulo quantifier de la herramienta mirdeep2 implementado durante esta tesis.

```
[rgomez@ixachi~]$: extract_miRNAs.pl mature.fa ref_specie mature >mature_ref_specie.fa
&& extract_miRNAs.pl hairpin.fa ref_specie >hairpin_ref_specie.fa
[rgomez@ixachi~]$: quantifier.pl -p precursors_ref_this_species.fa -m
mature_ref_this_species.fa -p precursos_ref_this_species.fa -r reads_collapsed.fa –t hsa -y
now
```

6.6 Lenguaje R: Normalización de datos.

R es un lenguaje y entorno de programación para el análisis estadístico y gráfico de datos cuantitativos y cualitativos. Durante esta metodología implementamos diversas paqueterías y funciones de R para manipular y evaluar la relación estadística de los datos procedentes del programa miRDeep2.

Antes de realizar cualquier análisis estadístico las tablas de lecturas tienen que ser normalizadas. Desde un sentido biológico la distribución de los transcritos anotados en las bibliotecas de RNA-seq presenta un comportamiento binomial. Por ello es importante que se consideren parámetros que determinen el tamaño, distribución y procedencia de las bibliotecas de RNA-seq (Parametric analysis of RNA-seq, Konishi 2016).

Debido a que el módulo quantifier.pl de la paquetería miRDeep2 implementa un cálculo básico pasa normalizar los conteos de lecturas procedentes de las bibliotecas analizadas fue necesario implementar la paquetería edgeR del lenguaje R para normalizar los conteos de lecturas de una manera adecuada para los análisis posteriores implementando el código de la tabla 7.

Tabla 7. Línea de comando para la normalización de los conteos implementando la herramienta edgeR desde su interfaz de lenguaje R

```
x <- tbl_df(read.csv('miRNAs_expressed_all_samples_counts.csv', header = T,sep = ","))
y <-DGEList(counts= x)


lognormal <- as.data.frame(cpm(y, normalized.lib.sizes=TRUE, log=T, prior.count=2)
```

6.7 Lenguaje R: Análisis de la correlación

Basándose en la premisa de que la correlación entre las fracciones celular y exosomal de las bibliotecas analizadas refleja aquellas especies de microRNA que se empaquetan en exosomas por difusión pasiva se realizó un breve cálculo de la correlación lineal Spearman para cada par de muestras (fracción celular y exosomal) procedentes de los diferentes eventos experimentales (n=6) así como por fracciones (Figuras 12,13 y 14)**.** Se configuró además el comando heatmap para graficar la distancia de la matriz de correlación *n* que hay entre las bibliotecas de lecturas (Tabla 8) (Figura 15 y 16).

Tabla 8. Línea de comando para calcular la correlación de los conteos normalizados

```

y <- as.matrix(dist(t(y)))
y <- y /max(y)
heatmap(y)
```

6.8 EdgeR: análisis diferencial de la expresión de miRNAs

Las bibliotecas de lecturas de secuenciación de RNA (RNA-seq) son variables discretas que tienen una fuerte relación estadística a pesar de la complejidad de los diseños experimentales (ie. Múltiples tratamientos que pueden inducir la expresión de distintos genes entre cada muestra) (Chen. Yunshun, T. L. Aaron, 2014). En esta vía, el paquete EdgeR fue diseñado para distinguir relación y variación de genes entre bibliotecas con variación biológica usando métodos empíricos Bayesianos, lineal generalizado (glms) y clásico. Este paquete fue escrito en lenguaje R y está

disponible como un componente “open source” de bioconductor (Robinson, McCarthy, & Smyth, 2010).

6.8.1 Quasi-likelihood F-tests (Qlf)

Los datos no normalizados de los archivos de salida read.csv provenientes del módulo quantifier.pl son utilizados para calcular la dispersión de los datos con la función qlf del paquete edgeR. El modelo estadístico qlf es utilizado para estimar la dispersión de un grupo de datos enteros no negativos {0, 1, 2, 3,…} con una distribución binomial. Los datos de Koppers-lalic et al. (2014) están basados en tres condiciones experimentales distintas (Tabla 3).

6.9 Selección de los miRNAs enriquecidos en la fracción exosomal.

Para almacenar los datos del análisis qlf implementamos el siguiente código. Esto es importante solo si se desea conocer los resultados en otras plataformas distintas a R.

```
#Sava and filter DE data
Dataframe <- as.data.frame(TopTags(qlf, n=length(qlf))
Dataframe <-$ID <- x$miRNAs
write.table(Dataframe,"C:/Users/Windows_user/test/qlf_test.csv", sep="\t")
```

La tabla de datos qlf contiene columnas que describen la significancia (columna del valor P) y enriquecimiento (columna LogFC) de los microRNAs anotados.

6.9.1 Preparación de los archivos de entrada para la búsqueda de elementos de acción en cis

Una mejor estrategia consistió en anotar los miRNAs expresados diferencialmente por fracción y significancia.

```
#Logical Filter
fraction <- mutate(dataframe,
                   frac=ifelse(logFC<=-1, "EXO Fr",

       ifelse(logFC>=1, "CELL Fr", "non-active")))

mutate(fraction, sig=ifelse(negl10pv>1.3, "sig", "NA"))
```

También podemos verificar el número de miRNAs significativos y cuántos de ellos pertenecen a cada fracción:

```
#Logical Filter
table (fraction$frac)
table(sig$sig)
```

Implementamos el siguiente código para obtener la lista de miRNAs de la fracción exosomal que fuesen significativamente enriquecidos.

```
#Logical Filter
DEseq_psig_enriched <- fraction %>%
 select(ID, frac,sig) %>%
 filter(frac==" EXO Fr") %>%
 filter(sig=="sig")
```

Aquellos miRNAs que cruzaron el umbral significativo (valor P < 0.05) serán observados en un gráfico de dispersión, tomando como eje X los valores Log_2FC (Fold Change) y eje Y =-log_{10}(PValue). Los datos con un Log_2FC <= -1 pertenecen a los miRNAs hallados en la fracción

exosomal, mientras que los datos con un $Log_2FC >= 1$ corresponden a los miRNAs expresados en la fracción celular.

Finalmente se extrajeron las secuencias de los miRNAs maduros anotados en la base de datos de miRNAs miRBase (ftp://mirbase.org/pub/mirbase/CURRENT/) publicada el 23 de Junio del 2014, versión 21. Se interceptan las secuencias que pertenecen a cada miRNA enriquecido en exosomas implementando la herramienta Venn de la paquetería gplots en r. El siguiente código fue desarrollado para extraer la lista de secuencias que pertenecen a los miRNAs enriquecidos significativamente en exosomas y celulares:

```
miRbase <- tbl_df(read.csv('mirbase20V_rstudio.csv', header = T,sep = ","))
rownames(miRbase)=make.names(miRbase$ID, unique=T)
miRbase$ID_format <-rownames(miRbase)

Meme_input<- miRbase  %>%
 select(ID_format,sequence) %>%
 inner_join(seq_psig_enriched,
          by=c("ID_format"="ID"))

write.table(Meme_input,
"C:/Users/Windows_user/Desktop/ESCRITO_TESIS_OCTUBRE/RESULTADOS/RESULTADOS_EBV
/0.2)DE_TEST/input_meme.csv", sep="\t",col.names = F, row.names = F)
```

6.9.1 MEME: búsqueda de elementos de acción *cis*

Las herramientas de la suite MEME son útiles para descubrir elementos de acción cis de secuencias de DNA/RNA y aminoácidos. Un elemento de acción cis es un patrón de nucleótidos o amino ácidos que aparecen repetidamente en un grupo de secuencias relacionadas de DNA/RNA o proteínas. La suite MEME representa estos elementos de acción cis como matrices de puntuación dependientes de la posición (Bailey et al., 2009). La herramienta MEME puede ser instalada como un programa o puede usarse como un servicio en línea en suite-meme.org/tolos/meme. Durante la actual tesis se tomaron en cuenta los siguientes criterios para configurar la herramienta:

1. El modo discriminativo, tomando en cuenta todos los miRNAs procedentes del archivo de salida de EdgeR que no pertenecían a las fracciones celular o exosomal respectivamente.

2. Modelo de Zoops: MEME supone que cada secuencia puede contener como máximo una ocurrencia de cada motivo. Esta opción es útil cuando se sospecha que algunos motivos puede que no se encuentre en algunas de las secuencias. En ese caso, los motivos encontrados serán más precisos que el uso de la primera opción.

3. La búsqueda se limitó a encontrar una lista de cuatro probables motivos de entre 3 y 10 nucleótidos de longitud.

6.9.2 Localización de exo-miRNAs con un elemento de acción en cis a lo largo de las bibliotecas

Este apartado describe la búsqueda de los exo-miRNAs con un elemento de acción en cis en común a lo largo de los conteos normalizados del archivo 'miRNAs_expressed_all_samples_counts.csv' obtenido del módulo quantifier. Esta vez se tomaron conteos normalizados adicionando el cálculo correctivo de la varianza implementando el método descrito por Anderson et. Al. (2010) y la herramienta DESeq (Anders & Huber, 2010), el siguiente fragmento de código describe el cálculo:

```
cdsFull = estimateSizeFactors( cdsFull )
cdsFull = estimateDispersions( cdsFull )
cdsFullBlind = estimateDispersions(  cdsFull , method = "blind" )
vsdFull = varianceStabilizingTransformation( cdsFullBlind )
```

El siguiente fragmento de código describe la extracción de los valores normalizados de los conteos correspondientes a los exo-microRNAs que presentaron un elemento de acción en cis en común:

```
exo_motif  <- tbl_df(read.csv("exo_motif.txt", header = T,sep = "\t"))
filter_libs_vs_exomotif <-   vsdFull %>%
 select(BJC,IKC,IMC,M3C,M9C,RNC,
       BJE,IKE,IME,M3E,M9E,RNE,ID) %>%
 inner_join(exo_motif,
          by=c("ID"=" exo_motif"), copy=T)
```

Finalmente se graficó de manera descendente los valores obtenidos en una figura *heatmap*; este gráfico puede visualizarse en la sección de resultados de esta tesis.

```
select = order(rowMeans(libs_vs_exomotif), decreasing=TRUE)[1:24] #select the most expresed
hmcol = colorRampPalette(brewer.pal(9, "GnBu"))(100)

heatmap.2(as.matrix(libs_vs_exomotif[select,]),
        main = "exo-miRNAs with a common motif: \
        normalized data with \
        (variance Stabilization) ",
        col = hmcol, trace="none",
        density.info=c("none"),
        dendrogram = c("none"),
        margin=c(5, 15))
```

6.9.3 Tailor: análisis de la adición de nucleótidos pos-transcripcional

Estudios recientes han sugerido que la adición de nucleótidos no templados (*tailing*) en la terminación 3’ de RNAs puede jugar un papel fundamental en la biogénesis de RNAs pequeños de silenciamiento (sRNAs). La identificación *in silico* ha facilitado de gran manera el entendimiento de los eventos de *tailing* a través de implementar datos de secuenciación de RNAs pequeños. En esta vía, la herramienta computacional Tailor integra herramientas bioinformáticas (Ej. Bowtie) para el análisis de datos NGS distinguiendo entre el mapeo y *tailing* de las lecturas pre-procesadas de las bibliotecas NGS de interés (Chou et al., 2015). El principio de detección de esta herramienta consiste en dada una lectura R (de M tamaño de bases) y todos los sufijos (Si) de una secuencia de referencia G (de N tamaño de bases), se puede encontrar el mayor prefijo en común (Longest common prefix, LCP) entre R y Si reportando el alineamiento desde la primer base hasta la última de la lectura R. La figura 2 de esta tesis resume el flujo de la información de entrada y salida de la herramienta Tailor; por otro lado, la tabla 9 representa la configuración del código implementado para ejecutar la herramienta Tailor, utilizando como archivo de entrada las

bibliotecas pre-procesadas de datos de RNAs pequeños de interés, el genoma de referencia humano indexado implementado durante esta tesis, la base de datos de secuencias maduras y hairpin de microRNAs humanos a sí como las características genómicas anotadas al genoma humano hasta la fecha.

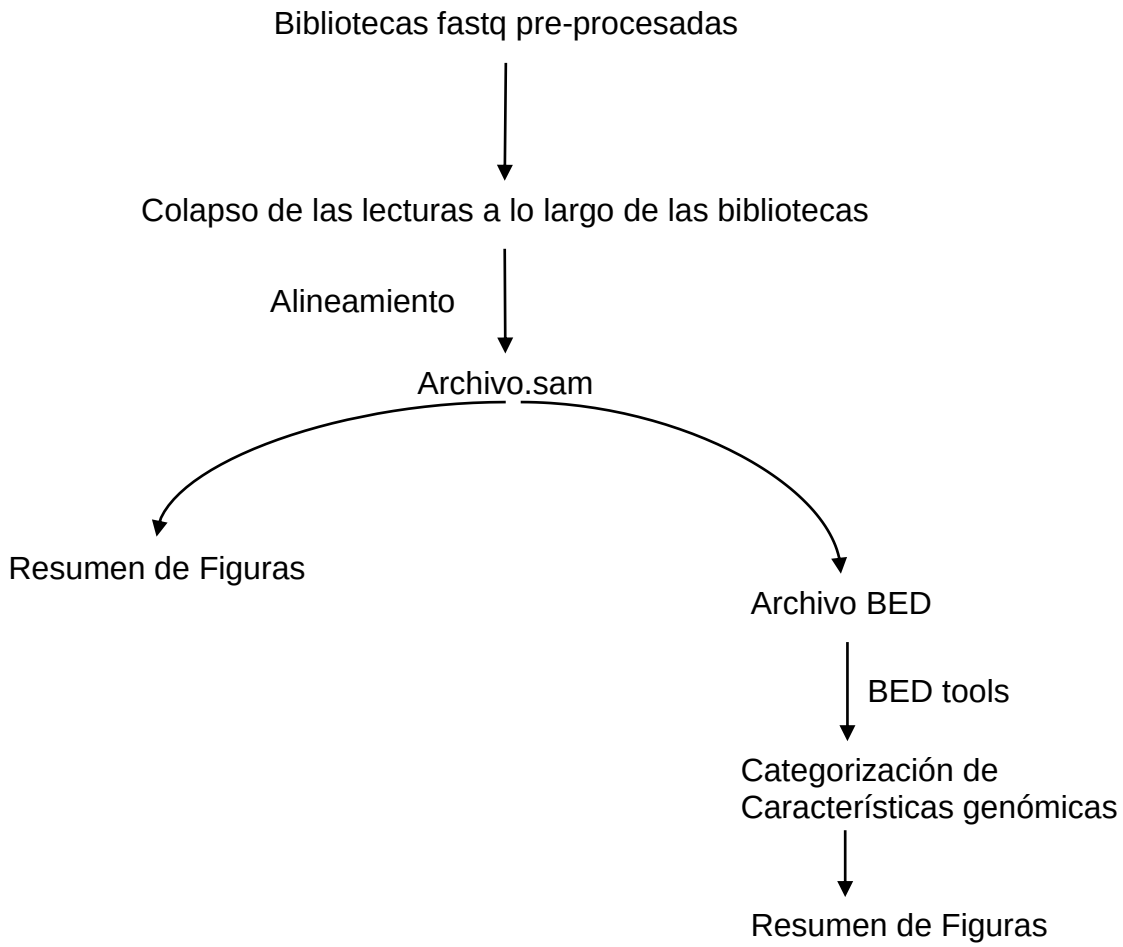


Figura 2. Flujo del pipeline del programa Tailor

Tabla 9. Código implementado para ejecutar la herramienta Tailor.

```
run_tailing_pipeline.sh -i archivo.fastq -g hg38.fa -t  /hg38.genomic_features –H hsa.hairpin.fa
-M hsa.mature.fa -c 22
```

Capítulo 7. Resultados

7.1 Pre-procesamiento y Mapeo de los datos NGS

El sistema de estudio referido en la tabla 3 corresponde a dos grupos de bibliotecas (fracción células y fracción exosomal). Antes de la secuenciación, Koppers-lalic et al. (2014) aislaron y purificaron las fracciones exosomales por protocolos descritos por Verweij et al., 2013. Las bibliotecas fueron preparadas usando el kit de secuenciación de RNAs pequeños TruSeq (Adaptor: ATCTCGTATGCCGTCTTCTGCTTG Layout: Single-end) y el equipo Illumina Hiseq 2000. Más detalles se indican en la tabla 10 de esta tesis.

Tabla 10. Metadatos del archivo SraInfoRun para el sistema de Células B infectadas por el virus Epstein Barr (EBV). ID del proyecto: SRP046046. Koppers-lalic et al. (2014) realizaron tres condiciones experimentales y n replicas dando como resultado 12 bibliotec

Biblioteca	fracción	ID de la muestra	Línea celular	Condición experimental
SRR1563015	celular	[BJ] C	BJAB sRNA	Linfoma (No infectivo) n=1
SRR1563017	exosomal	[BJ] E		
SRR1563018	celular	[M3] C	MUTU I clone 3 sRNA	Leucemia de Burkitt, EBV+ n=2
SRR1563056	exosomal	[M3] E		
SRR1563057	celular	[M9] C	MUTU I clone 9 sRNA	
SRR1563058	exosomal	[M9] E		
SRR1563059	celular	[IK] C	IK140508 sRNA	Linfoblastoides EBV+ n=3
SRR1563060	exosomal	[IK] E		
SRR1563061	celular	[IM] C	IM-1 sRNA	
SRR1563062	exosomal	[IM] E		
SRR1563063	celular	[RN] C	RN sRNA	
SRR1563064	exosomal	[RN] E		

Tras obtener los archivos fastq por el método descrito en el capítulo 7.2 de esta tesis las bibliotecas fueron pre-procesadas utilizando los módulos fastx_clipper y fastq_quality_filter del

kit computacional fastx, en la siguiente tabla se representa la sintaxis del código implementado para pre-procesar las bibliotecas (tabla 11).

Tabla 11. Línea de comando para el pre-procesamiento de las bibliotecas

```
[rgomez@ixachi~]$: $ fastx_clipper -i sample.fastq -l 18 -a [SECUENCIA DE ADAPTOR] -o clipped.fastq & fastq_quality_filter -q 30 -p 80 -v -i clipped.fastq -o preprocessed.fastq &
```

La tabla 12 muestra el tamaño de las bibliotecas antes y después de pre-procesarles así como el porcentaje de mapeo global e individual de las bibliotecas contra el genoma humano.

Tabla 12. Implementado el comando grep y wl del lenguaje Linux se cuantifico el número de lecturas que contienen las bibliotecas en bruto y después del pre-procesamiento. Ej. cat sample.fastq | grep "@" | wc –l >1 number_reads.

Biblioteca	**Número de lecturas en bruto**	**Número de lecturas después del pre-procesamiento**	**Número de lecturas pre-procesadas mapeadas contra el genoma de referencia**
SRR1563015	2.56E+07	1.48E+07	5.47E+06
SRR1563017	8.72E+06	2.35E+06	5.42E+05
SRR1563018	1.08E+07	5.97E+06	3.40E+06
SRR1563056	2.01E+07	1.08E+07	1.04E+06
SRR1563057	7.13E+06	4.23E+06	2.84E+06
SRR1563058	1.27E+07	6.37E+06	4.41E+05
SRR1563059	1.13E+07	6.89E+06	4.20E+06
SRR1563060	1.18E+07	6.16E+06	1.53E+06
SRR1563061	1.27E+07	8.32E+06	3.17E+06
SRR1563062	3.00E+06	1.77E+06	1.65E+06
SRR1563063	1.88E+07	1.22E+07	5.99E+06
SRR1563064	1.52E+07	8.18E+06	1.50E+06

Además, se implementó el módulo mapper.pl de la herramienta computacional mirDeep2 para comparar el número de lecturas mapeadas contra el genoma de referencia antes y después del pre-procesamiento (Figura 3).

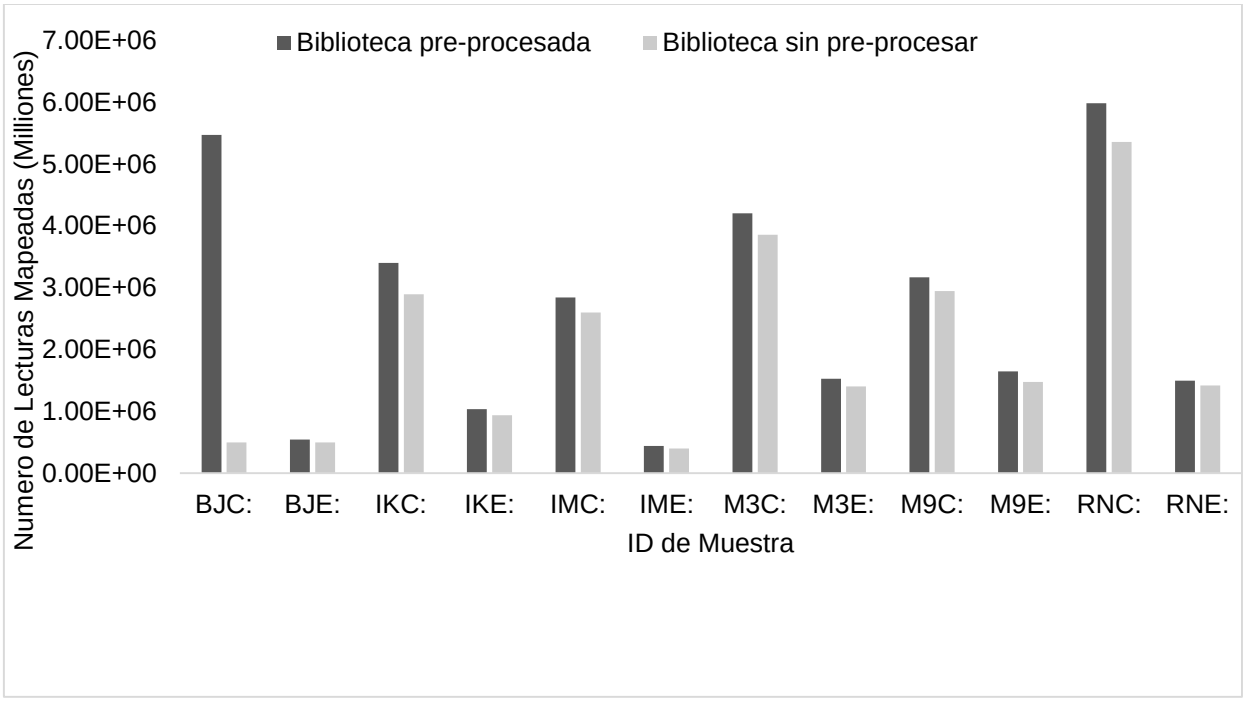


Figura 3. Comparación del mapeo de las lecturas después y antes del pre-procesamiento

Como valor global, el 50% de las lecturas procedentes a las 12 bibliotecas del primer caso de estudio mapearon contra los genomas de referencia (ie. humano y de virus EBV). El porcentaje de mapeo de las lecturas puede reflejar la información genética de organismos contaminantes encontrados durante los procedimientos experimentales previos a la secuenciación. Obviándose de este modo la calidad de la secuenciación. Para fines de esta tesis, estos valores fueron aceptables.

Implementando el análisis con la herramienta fastqc se corroboró la efectividad del pre-procesamiento. El puntaje por base de lectura y por lectura global nos señala de manera más detallada la calidad de las lecturas después del pre-procesamiento realizado para esta tesis. La figura 4 y 5 indica la distribución de lecturas por su tamaño en longitud de nucleótido a lo largo de las bibliotecas. Se agrupan en dos gráficos estos resultados por las respectivas fracciones experimentales (ie. celular y exosomal). El tamaño relativo de los picos entre fracciones indica

que RNAs de entre 32-33 nucleótidos de longitud son enriquecidos en exosomas (Figura 4). Además, todas las bibliotecas mostraron una distribución de calidad phred por lectura mayor a 30 (figura 6 y 7).

A pesar del hecho de que se analizaron 12 bibliotecas, la figura 8 indica el resultado relativo de la distribución del puntaje de calidad a lo largo de los nucleótidos de las secuencias en todas las bibliotecas. Se denota que el puntaje de calidad phred por nucleótido en la lectura es mayor a 30 (Figura 8). Ambos resultados, la distribución de calidad por lectura y nucleótido contribuyen a reconocer la calidad de la secuenciación almacenada en las bibliotecas que se analizaron durante esta tesis. Finalmente, se comprobó por biblioteca que los adaptores añadidos durante la secuenciación hayan sido removidos satisfactoriamente después del pre-procesamiento, la figura 9 representa el esquema general de cómo se visualiza esta información para las 12 bibliotecas analizadas.

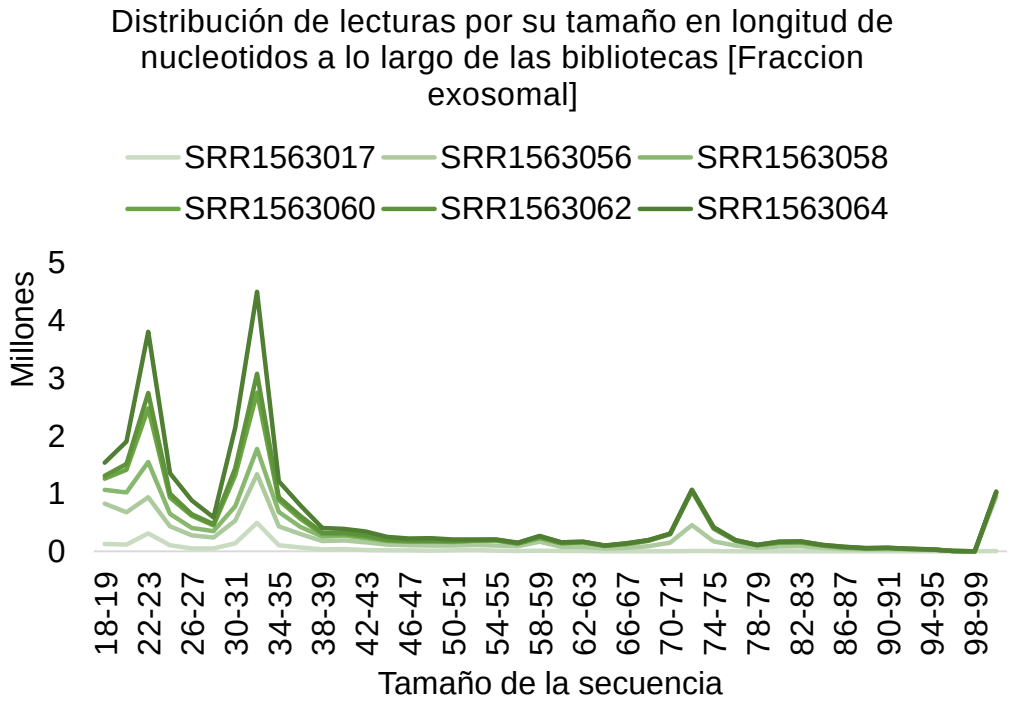


Figura 4. Distribución de lecturas por su tamaño en longitud de nucleótidos a lo largo de las bibliotecas de secuenciación pre-procesadas; (Fracción exosomal).

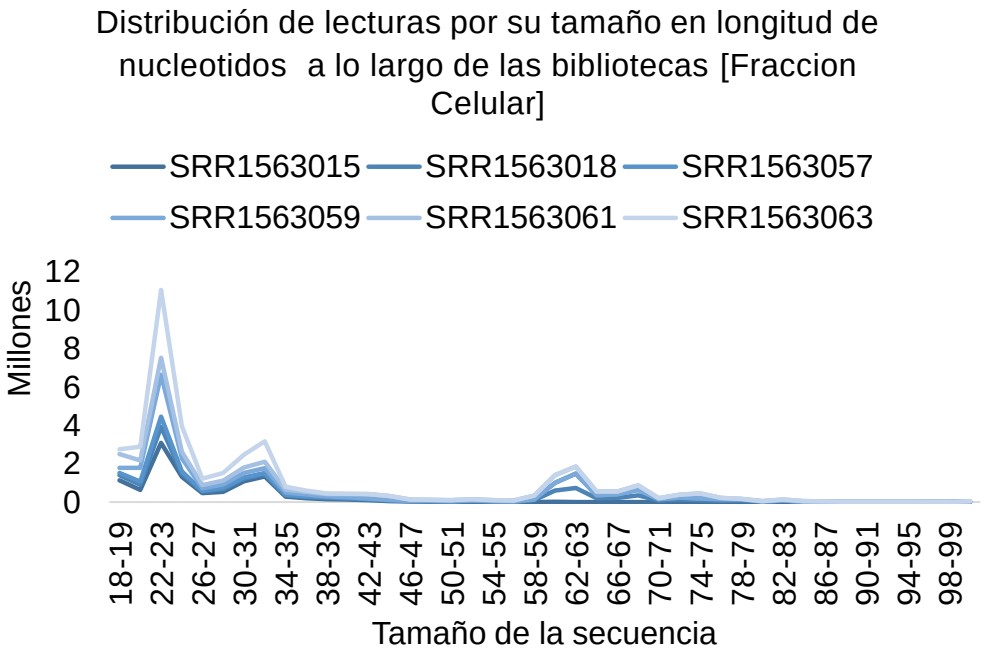


Figura 5. Distribución de lecturas por su tamaño en longitud de nucleótidos a lo largo de las bibliotecas de secuenciación pre-procesadas; (Fracción Celular).

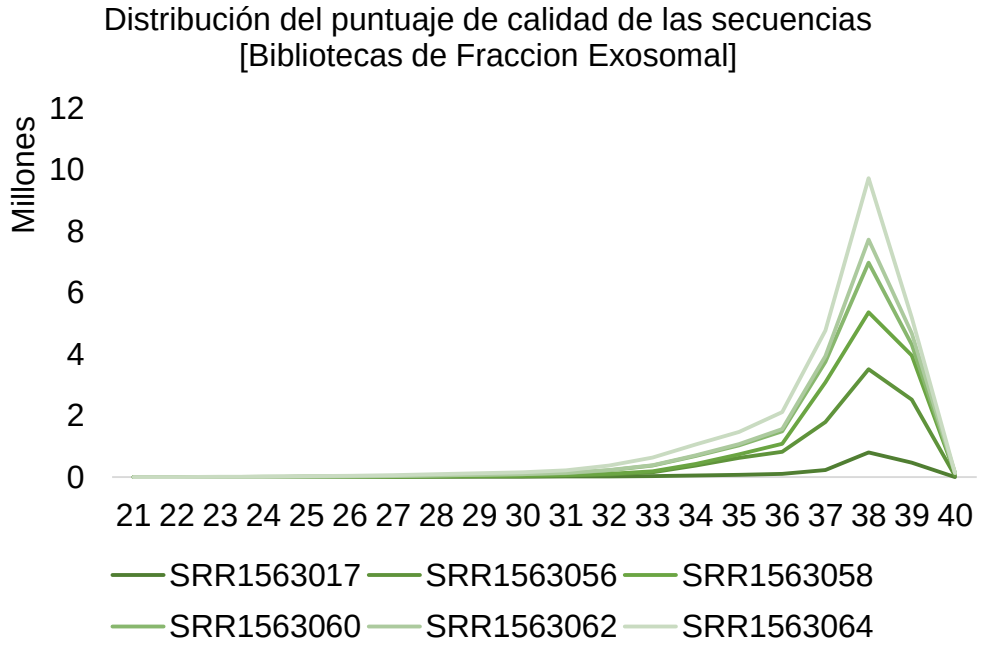


Figura 6. Figura 5. Distribución del puntaje de calidad de las secuencias a lo largo de las bibliotecas analizadas. (Fracción exosomal).

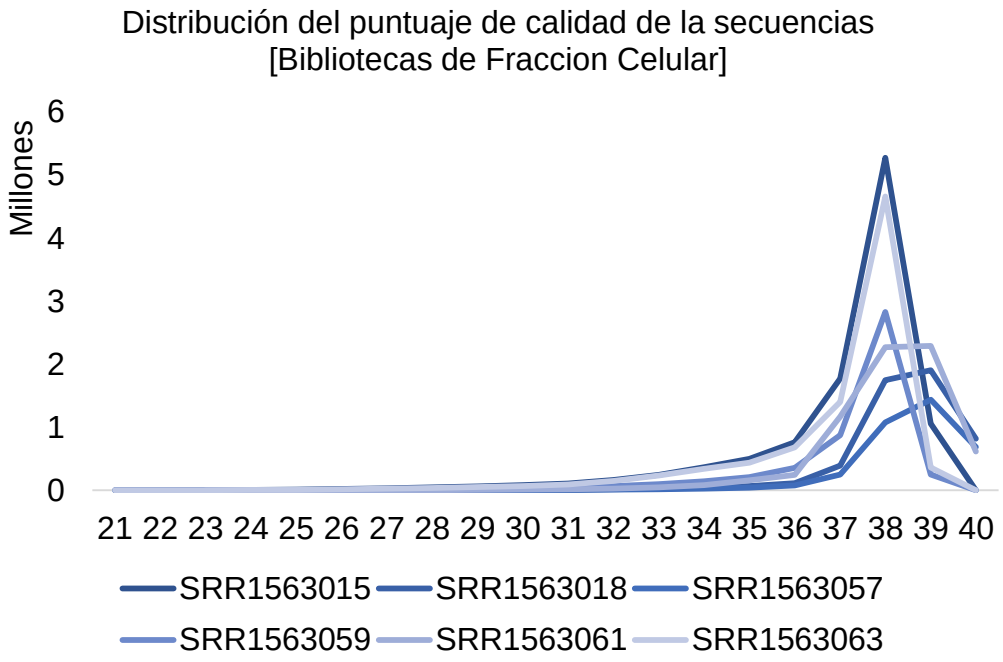


Figura 7. Distribución del puntaje de calidad de las secuencias a lo largo de las bibliotecas analizadas. (Fracción celular).

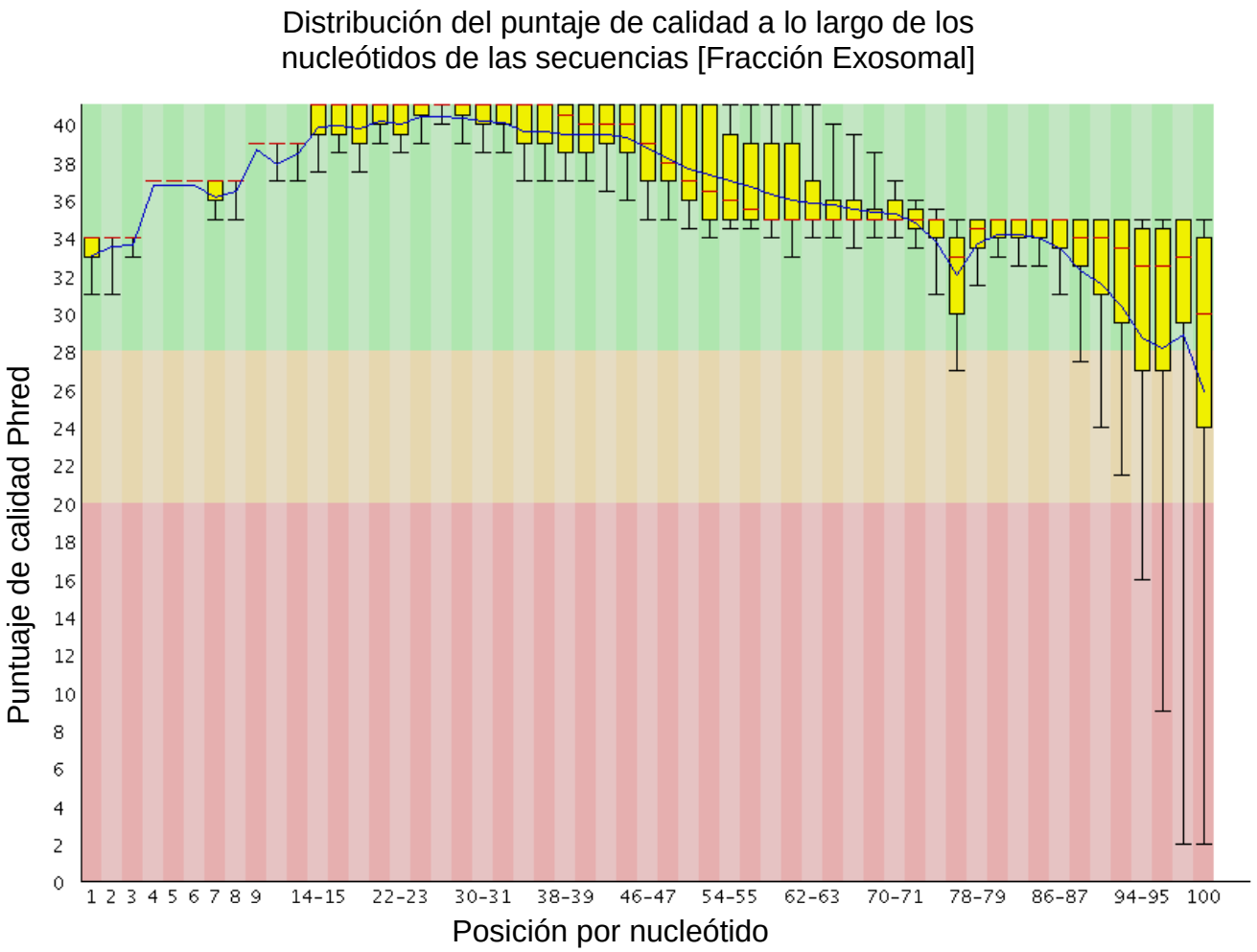


Figura 8. Distribución del puntaje de calidad a lo largo de los nucleótidos de las secuencias [Fracción Exosomal]

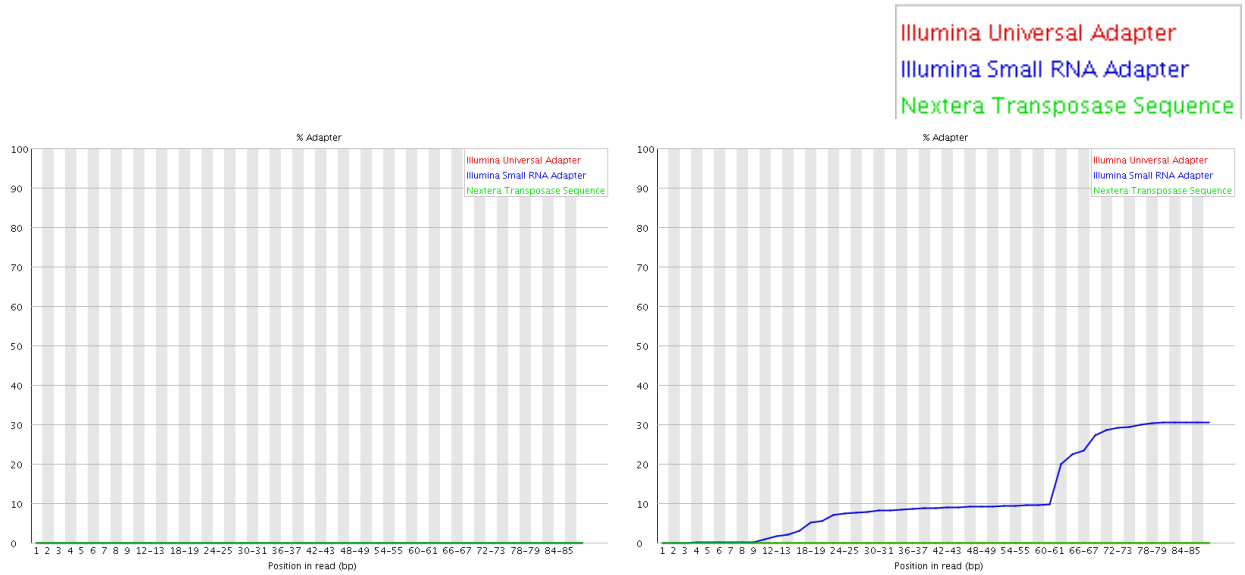


Figura 9. Comprobación de los adaptores removidos de las bibliotecas de secuenciación después del pre-procesamiento (Panel izquierdo). Se denota la presencia de adaptores illumina para secuenciación de RNAS pequeños (Panel derecho).

7.2 Anotación de microRNAs

Implementando la paquetería mirdeep2 se cuantificó el perfil de expresión de microRNAs que fueron secuenciados y almacenados en las bibliotecas que pre-procesamos con anterioridad. La tabla 13 y figura 10 señalan el conteo global de las lecturas mapeadas contra microRNAs y genoma de referencia en comparativa con el número de lecturas totales de las bibliotecas pre-procesadas.

Tabla 13. Número de lecturas mapeadas contra microRNAs de referencia en comparativa con el número de lecturas totales de las bibliotecas pre-procesadas y mapeadas contra el genoma de referencia

Biblioteca	**ID de la muestra**	**Número de lecturas después del pre-procesamiento**	**Número de lecturas pre-procesadas mapeadas contra el genoma de referencia**	**Número de lecturas pre-procesadas mapeadas contra microRNAs de referencia**
SRR1563015	BJC	1.48E+07	5.47E+06	5.44E+06
SRR1563017	BJE	2.35E+06	5.42E+05	4.76E+05
SRR1563018	M3C	5.97E+06	3.40E+06	1.47E+06
SRR1563056	M3E	1.08E+07	1.04E+06	7.01E+05
SRR1563057	M9C	4.23E+06	2.84E+06	9.55E+05
SRR1563058	M9E	6.37E+06	4.41E+05	7.62E+05
SRR1563059	IKC	6.89E+06	4.20E+06	4.39E+06
SRR1563060	IKE	6.16E+06	1.53E+06	1.03E+06
SRR1563061	IMC	8.32E+06	3.17E+06	1.14E+06
SRR1563062	IME	1.77E+06	1.65E+06	3.74E+05
SRR1563063	RNC	1.22E+07	5.99E+06	6.23E+06
SRR1563064	RNE	8.18E+06	1.50E+06	1.20E+06

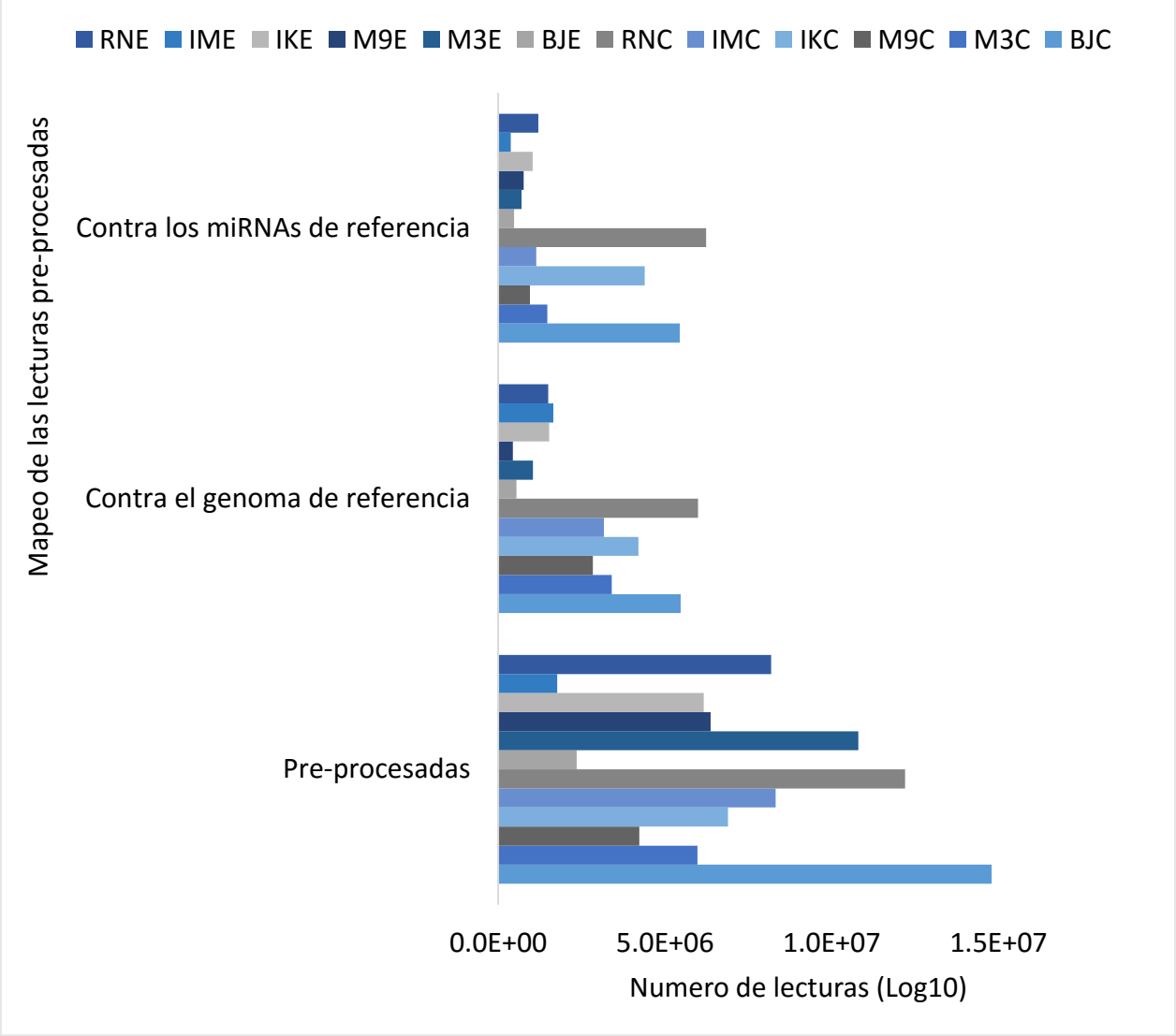


Figura 10. Número de lecturas mapeadas contra microRNAs de referencia en comparativa con el número de lecturas totales de las bibliotecas y mapeadas contra el genoma de referencia.

También, la paquetería miRDeep2 genera archivos pdf que muestran dos importantes gráficos: 1) La estructura secundaria y secuencia de los miRNAs encontrados para cada biblioteca que se analizó y 2) La frecuencia de los alineamientos obtenidos para cada biblioteca contra la secuencia de la estructura secundaria de los miRNAs anotados en miRBase (indicando si el alineamiento fue en la región 5’ o 3’ de la secuencia precursora) (Figura 11).

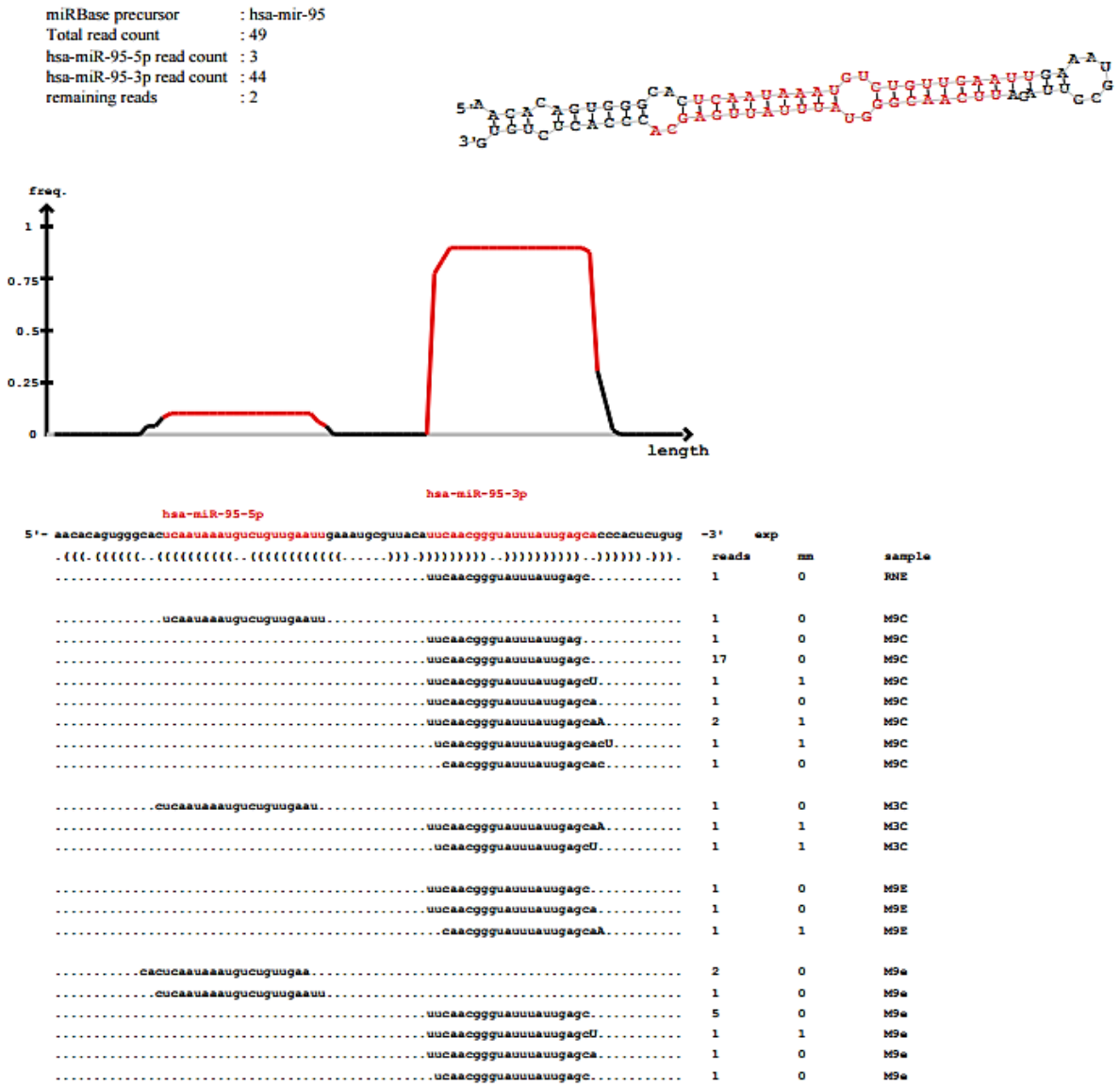


Figura 11. La estructura secundaria de los miRNAs expresados se genera automáticamente con el paquete Vienna-RNAfold. El paquete Vienna-RNAfold también calcula la notación "()" y "." para denotar los alineamientos y no alineamientos de la secuencia precursora.

Finalmente, se obtuvieron el archivo de salida file.csv que muestra dos meta-columnas: "miRNA ID" y "conteos por muestra", los cuales incluyen los IDs de los microRNAs anotados y los conteos (ie. Número de veces que las lecturas mapearon a un particular microRNA) hallados en cada muestra, respectivamente. Como resultado se obtuvo una matriz de 12 columnas y 2866 filas (correspondientes a 2866 miRNAs cuantificados a lo largo de las 12 bibliotecas). La tabla 14 muestra la cabecera del contenido del archivo file.csv implementado para los posteriores análisis.

Tabla 14. Se muestra la cabecera del contenido del archivo file.scv de los datos reads_collapsed.fa que fueron mapeados contra el archivo mature.fa de referencia (humano y virus EBV); la primer columna indica las muestras analizadas, para cada fila de la primer columna se extienden los primeros tres microRNAs contados por muestra.

Dimensiones		**miRNA ID**		
12 x 2866		ebv-miR-BART1-5p	ebv-miR-BART1-3p	ebv-miR-BART10-5p
Conteos por muestra	[BJ]C	0	2	0
	[BJ]E	0	0	0
	[IK]C	5630	1261	626
	[IK]E	1700	514	95
	[IM]C	785	290	69
	[IM]E	736	279	34
	[M3]C	3	1	0
	[M3]E	3	2	5
	[M9]C	0	1	0
	[M9]E	1	0	0
	[RN]C	17934	11120	0
	[RN]E	3029	669	1

7.3 Análisis de la correlación

Se ha sugerido que el enriquecimiento de microRNAs en exosomas ocurre por un fenómeno no selectivo durante la formación de exosomas (visto en: (Turchinovich, Tonevitsky, & Burwinkel, 2016)). Los siguientes resultados reportan que existe una correlación lineal ascendente entre los conteos de microRNAs del par de bibliotecas celulares y exosomales de cada caso de estudio (figura 12, 13 y 14). Esto puede reflejar un proceso de empaquetamiento no selectivo de microRNAs intracelulares durante la formación de exosomas. De estas correlaciones concluimos que una larga proporción de microRNAs pueden ser empaquetados pasivamente en exosomas puesto que las concentraciones exosomales reflejan relativamente las concentraciones celulares.

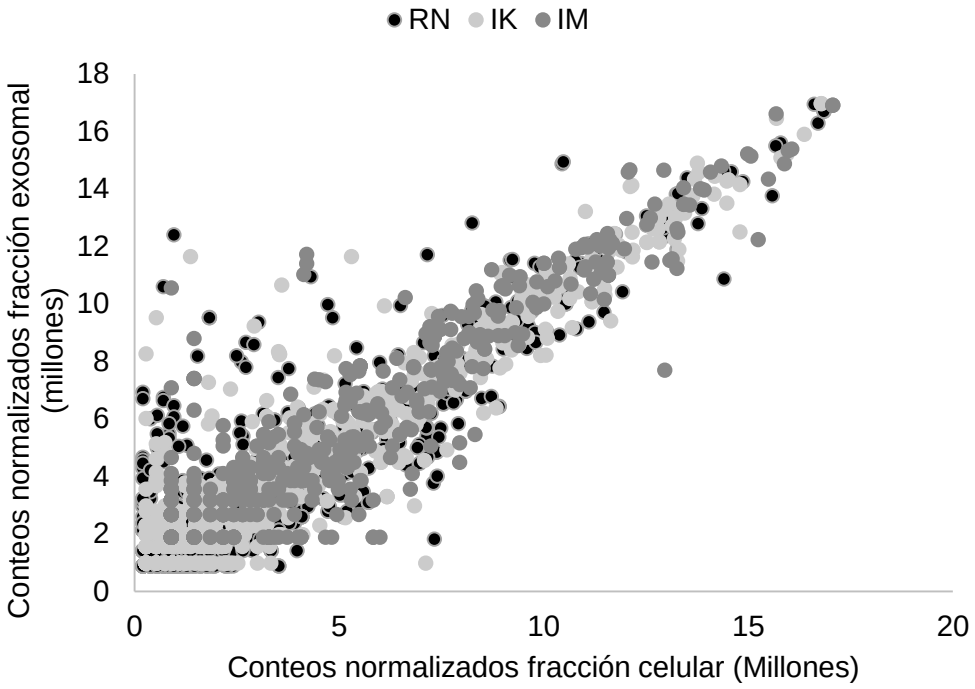


Figura 12. Correlación de los datos de la condición experimental: Linfoblastoides EBV+. Los conteos están normalizados por cpm (Counts Per Million) con una correlación Spearman de 0.75 [RN], 0.73 [IM] y 0.76 [IK].

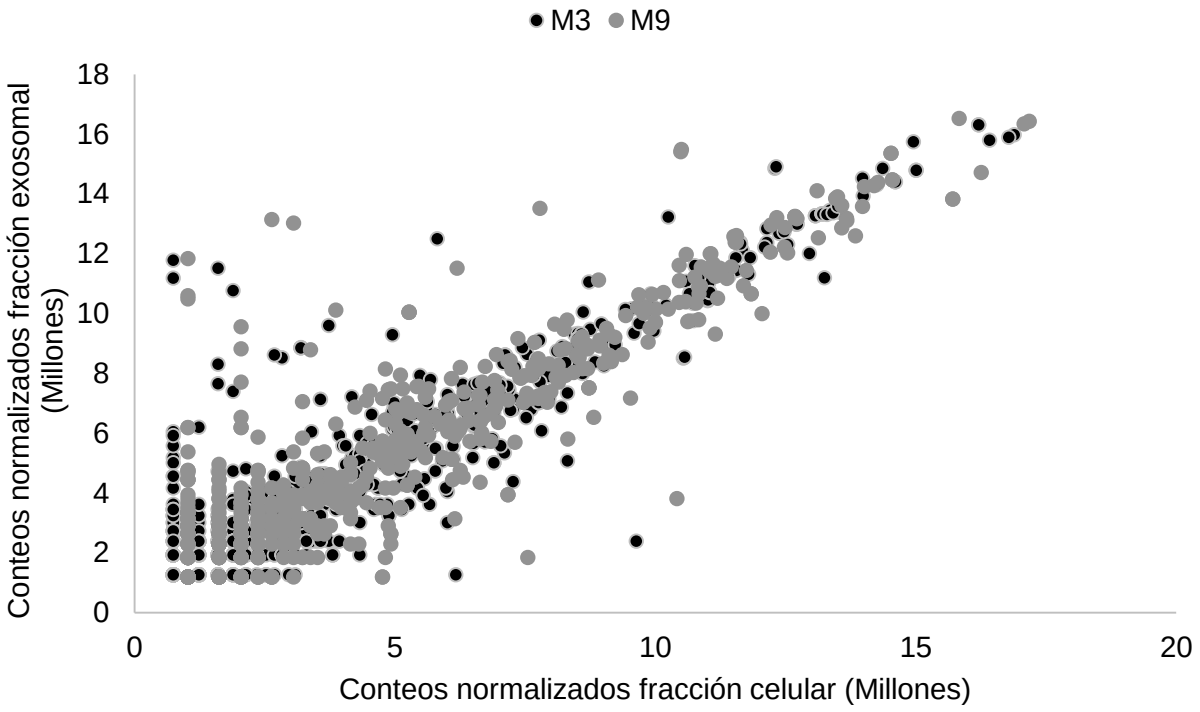


Figura 13. Correlación de los datos de la condición experimental: Leucemia de Burkitt EBV+. Los conteos están normalizados por cpm (Counts Per Million) con una correlación Spearman de 0.75 [M9] y 0.74 [M3].

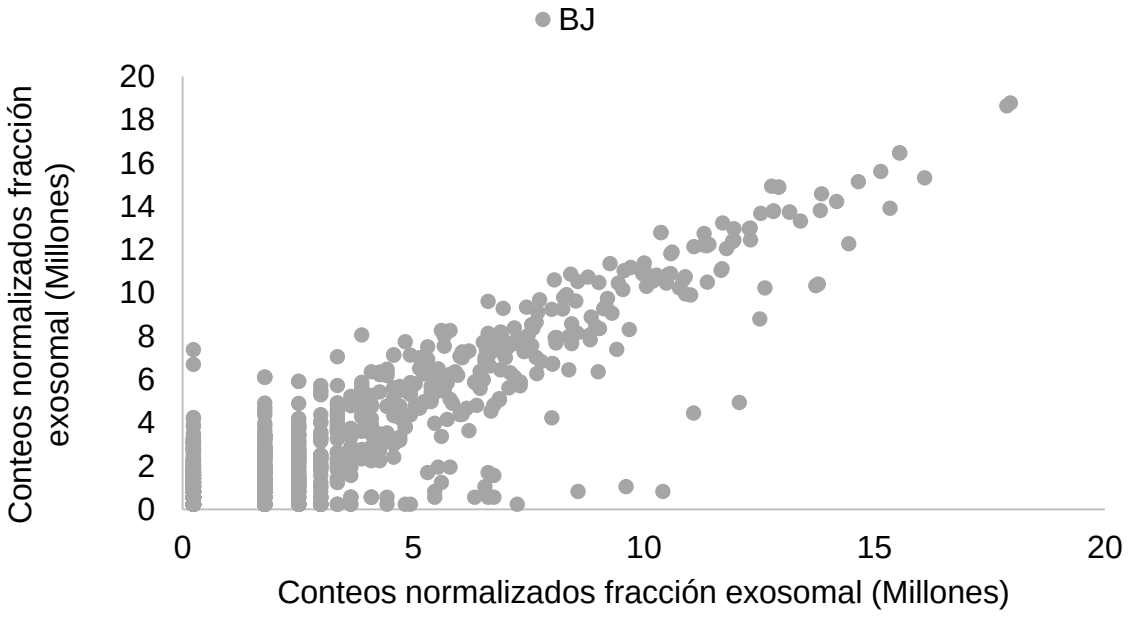


Figura 14. Correlación de los datos de la condición experimental: Linfoma. Los conteos están normalizados por cpm (Counts Per Million) con una correlación Spearman de 0.73.

Cuando comparamos la matriz de correlaciones de todas las bibliotecas se encontró un agrupamiento que coincide con el estado experimental de donde provienen las bibliotecas de secuenciación. Las muestras [IM], [IK] y [IM] reportan una relación mayor entre sí que con las demás bibliotecas agrupándose en la parte superior derecha de la figura 15. Por otra parte, las muestras [M3], [M9] en adición a [BJ] tiene una relación más íntima que con el primer grupo de datos pertenecientes a otra procedencia experimental (ie. Linfoma EBV+) (Figura 15, extremo izquierdo inferior). También comparamos la matriz de correlación separando las bibliotecas por fracción dando los mismos resultados que los primeros señalados (Figura 16 y Figura 17); excluyendo las muestras negativas del virus EBV (ie. [BJ]), el agrupamiento de la matriz de correlación para las muestras Linfoma de burkit y linfoma corresponde a la latencia 1 y 3 del estado infectivo del virus EBV, respectivamente. Además, se incluyó una estrategia en la cual solo se consideraron las muestras positivas del virus EBV; basándose en estos criterios se establecieron tres estrategias para el análisis binomial de la expresión diferenciada (DE) (tabla 15).

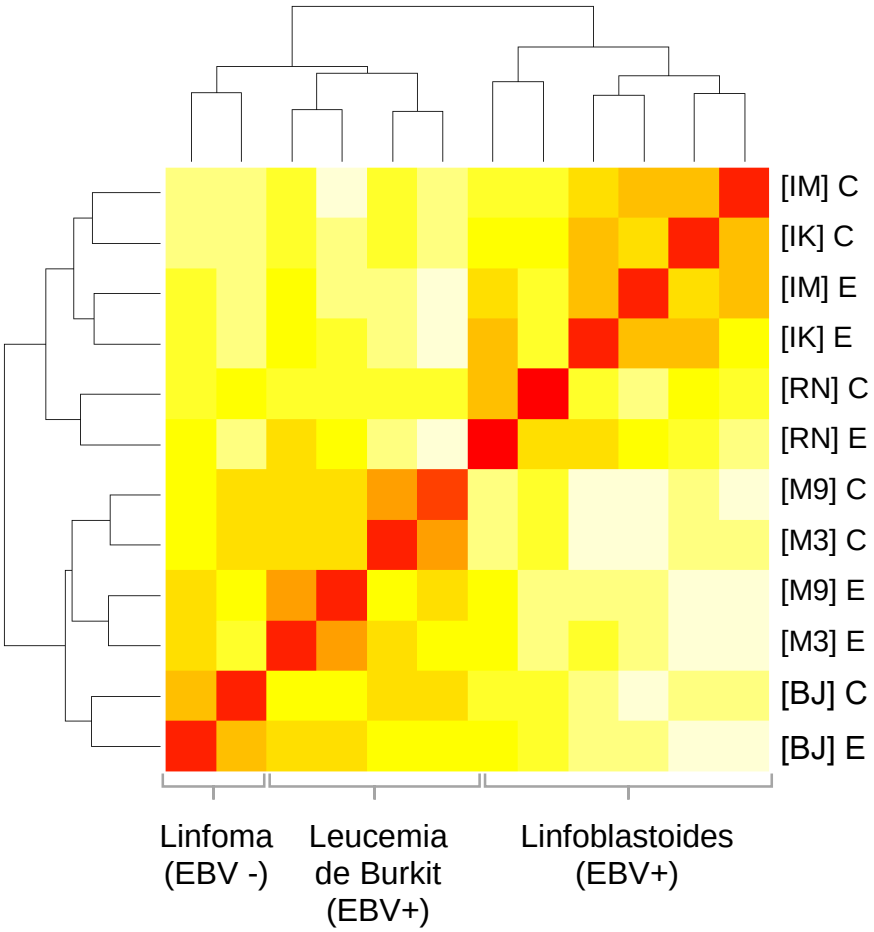


Figura 15 Mapa de la matriz de relación de las bibliotecas. Las bibliotecas (Panel derecho) se agrupan a sus respectivas condiciones experimentales (Panel inferior)

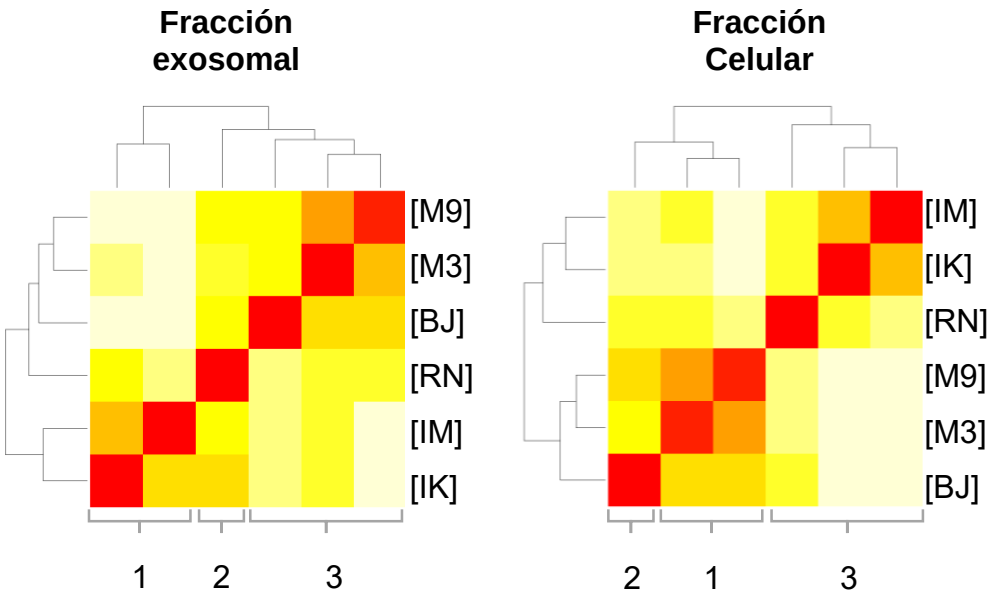


Figura 16. Mapa de la matriz de relación de las bibliotecas. Las bibliotecas se agrupan a sus respectivas condiciones experimentales: 1) Linfoma (EBV+), 2) Leucemia de Burkit (EBV -) y Linfoblastoides (EBV+).

Tabla 15. Diseños de las estrategias para el análisis binomial de la expresión diferenciada

ESTRATEGIA	MUESTRA	CONDICIÓN EXPERIMENTAL
1	IK, IM, RN, M3 y M9	EBV+
2	IK, IM y RN	Linfoblastoides (EBV+)
3	M3, M9 y BJ	Leucemia de Burkit (EBV+) y de Linfoma (EBV-)

Finalmente, se almacenaron las réplicas de las columnas deseadas en una tabla clase DGE-List la cual está diseñada para almacenar información asociada a nuestras tablas tipo file.csv, posteriormente construimos un modelo de matriz basado en las estrategias experimental y separando los datos en dos grupos correspondientes a las fracciones exosomales y celulares.

7.4 Análisis del cociente de abundancia de microRNAs al comparar fracciones

Cargamos la tabla file.csv para desarrollar el análisis Quasi-likelihood F-tests (Qlf)

```
tb <- tbl_df(read.csv('miRNAs_expressed_all_samples_counts.csv', header = T,sep = ","))
```

Basado en los resultados de la correlación se agruparon las siguientes estrategias del análisis

```
#Grouping samples
Desing_1<-as.data.frame(select(tb, IKE,IME,M3E,M9E,RNE,IKC,IMC,M3C,M9C,      RNC))
esing_2<- as.data.frame(select(tb, IKE, IME,RNE,IKC, IMC, RNC))
esing_3<-as.data.frame(select(tb, BJE,M3E,M9E,BJC,M3C,M9C))
```

Se diseñaron dos matrices en los cuales se agrupan ambas fracciones para las estrategias experimentales.

```
group <- factor(c(0,0,0,1,1,1))
group1 <- factor(c(0,0,0,0,0,1,1,1,1,1))
design <-model.matrix(~group)
design1 <-model.matrix(~group1)
```

Almacenamos estas características en una lista de clase DGE:

```
y_a<-DGEList(counts= Desing_2 , group =group, genes =rownames(tb))
y_b <-DGEList(counts= Desing_3, group =group, genes =rownames(tb))
y1 <-DGEList(counts= Desing_1, group =group1, genes =rownames(tb))
```

Calculamos factor de normalización para las muestras y visualizamos.

```
y_a <-calcnormfactors(y_a)
mfactors(y_b)
y1 <-calcNormFactors(y1)
```

El análisis qlf se realizó de la siguiente manera:

```
fit_a <-glmQLFit(y_a,design)
Fit(y_b,design)
 fit1 <-glmQLFit(y1,design1)
 qlf_a <-glmQLFTest(fit_a,coef=2)
 qlf_b <-glmQLFTest(fit_b,coef=2)
 qlf1 <-glmQLFTest(fit1,coef=2)
```

Como resultado se obtiene una tabla denotada en la variable qlf. Dicha tabla contiene una lista de datos que describen la probabilidad estadística (Valor P) y enriquecimiento (cociente expresado en base Logarítmica, LogFC) de las lecturas de miRNAs basado en las estrategias de análisis hechos (design <-model.matrix (~group)). Este análisis incluye además una lista de valores de 0 a 1 que describen la probabilidad de falsibilidad del análisis (ie. Análisis de la diferenciación en la expresión entre fracciones Exosomal vs Celular). Dichos valores denominados como FDR (False Discovery Rate) son asignados en función del valor P. A continuación se muestra un ejemplo de 3 filas de datos correspondientes al análisis qlf, con su respectiva información:

```
> topTags(qlf, n=3)
```

miRNA ID	logFC	logCPM	F	PValue	FDR
hsa-miR-136-3p	8.60	5.55	60.23	0.0000	0.0014
hsa-miR-144-3p	9.50	8.30	56.54	0.0000	0.0014
hsa-miR-432-5p	8.27	5.23	53.80	0.0000	0.0014

Para el interés del trabajo actual se consideraron las columnas LogFC, $-\text{Log}_{10}$ (PValue), FDR y la columna inicial de la lista de miRNAs expresados durante los casos de estudio analizados en esta tesis. El grupo de figuras 17 muestran la distribución de microRNAs entre fracciones; Ambas fracciones (exosomal y celular) están expresadas como el coeficiente de abundancia LogFC, valor LogFC ≤ -1 o LogFC ≥ para las fracciones respectivas; as figuras de cada estrategia experimental revelan que gran parte de los microRNAs biológicamente expresados tienen un coeficiente de abundancia cercano a cero (Línea recta que se prolonga por encima del valor 1500 en el eje x); Esto último refleja una gran proporción de los conteos de microRNAs sin enriquecimiento preferencial en cualquiera de las fracciones, celular o exosomal. Mientras que por debajo del valor

1000, en el eje x, el coeficiente de abundancia (LogFC) empieza a distribuirse hacia cualquiera de las fracciones.

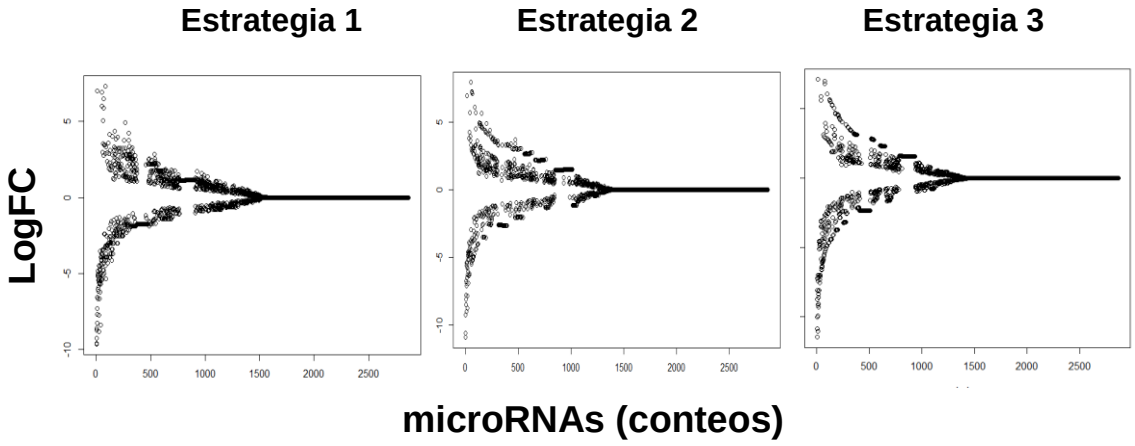


Figura 17. Distribución de microRNAs entre fracciones; Ambas fracciones (exosomal y celular) están expresadas como el valor LogFC ≤ -1 o LogFC ≥ 1, respectivamente.

Tomando la hipótesis de que existen elementos de acción en *cis* como factor del empaquetamiento selectivamente activo se tomaron solo las lista de transcritos enriquecidos en la fracción exosomal (LogFC ≤ -1) para el análisis bioinformático propuesto en el objetivo 2 de esta tesis. Como primera instancia se consideró tomar aquella lista de microRNAs exosomales que crucen el umbral del valor P ≤ 0.05 (ie. $-\log_{10}$ (PValue) ≥ 1.3) (tabla 16); no obstante, un análisis más riguroso señaló proporciones de microRNAs por encima del umbral del valor P con una probabilidad FDR menor a 0.05 (ie. $\log_{10}$ (FDR) ≤ -1.5); Este dato pudiera fungir como corroborativo de los resultados de los análisis posteriores. Siguiendo el raciocinio de este análisis implementamos el siguiente código para extraer la lista de miRNAs enriquecidos para ambas fracciones así como aquellos estadísticamente significativos (valor P ≤ 0.05). La tabla 17 indica la proporción de miRNAs enriquecidos significativamente para cada fracción.

```
fraction <- mutate(dataframe,
            frac=ifelse(logFC<=-1, "EXO Fr",
                        ifelse(logFC>=1,  "CELL  Fr",  "Passive"))fraction  <-  mutate(fraction,
 sig=ifelse(negl10pv>1.3, "sig", "NA"))
```

Tabla 16. Tabla de los miRNAs enriquecidos por fracción y significativos.

ESTRATEGIA	[CEL-MIRNAS]	[EXO-MIRNAS]	SIN ABUNDANCIA	SIGNIFICATIVOS VALOR P > 0.05
1 [IK, IM,M3,M9, RN]	511	366	1989	202
2 [IK, IM, RN]	586	299	1981	127
3 [BJ, M3, M9]	446	402	2018	172

Tabla 17. Tabla de los miRNAs enriquecidos significativamente (Valor P> 0.05) por fracción

ESTRATEGIA	[CEL-MIRNAS]	[EXO-MIRNAS]
1 [IK, IM,M3,M9, RN]	58	144
2 [IK, IM, RN]	50	77
3 [BJ, M3, M9]	41	131

El grupo de figuras 18,19 Y 20 señala dos elementos importantes: 1) aquellos microRNAs enriquecidos diferencialmente entre la fracción exosomal y celular y 2) Aquellos microRNAs significativamente enriquecidos que cruzan el umbral de la probabilidad P <0.05 deseada.

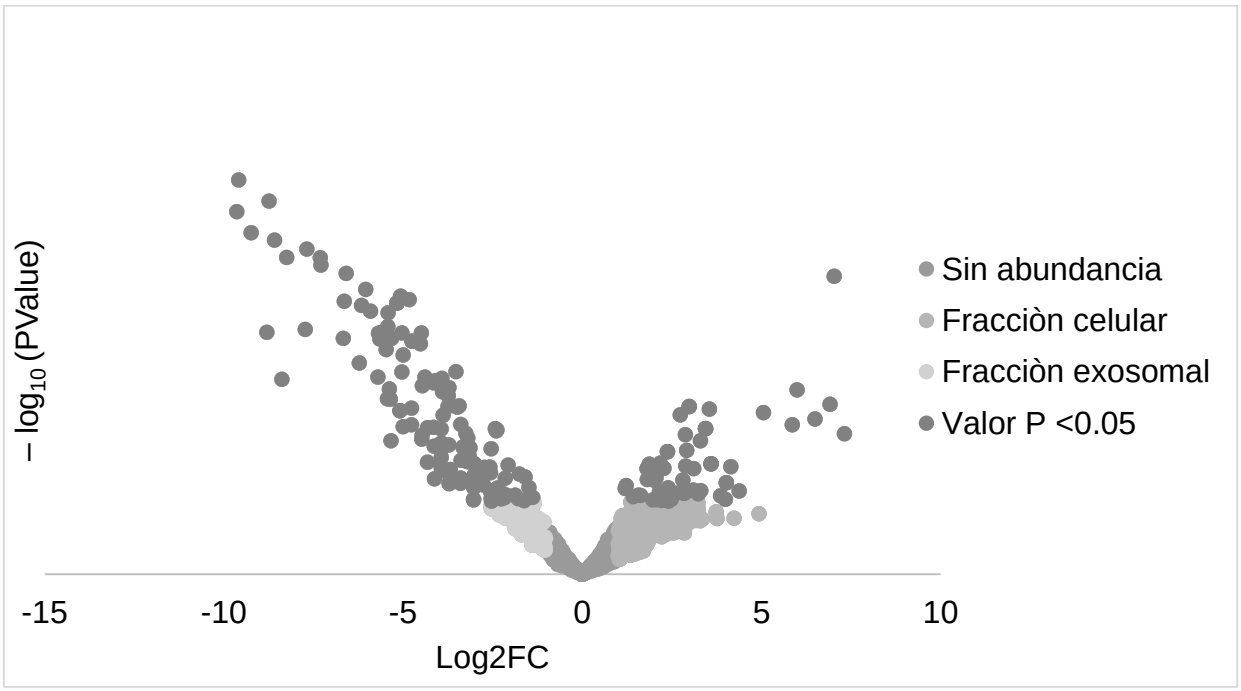


Figura 18. Gráfico de volcán representativo de los microRNAs abundantes y significativos por fracción; Estrategia 1 (muestras: IK, IM, M3, M9 y RN).

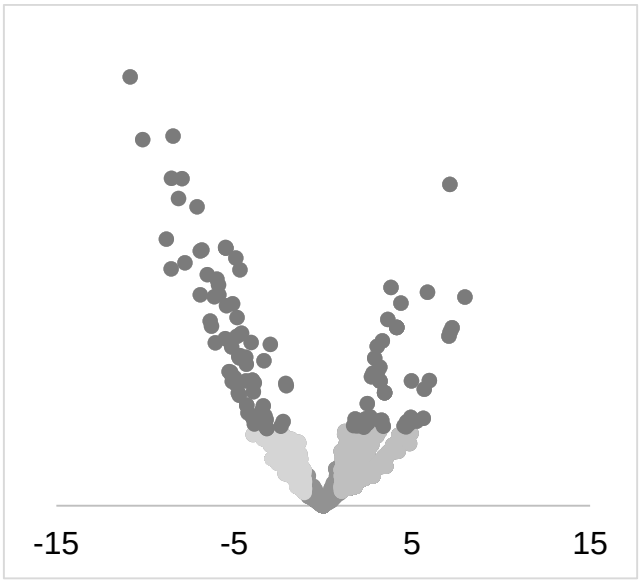


Figura 19. Gráfico de volcán representativo de los microRNAs abundantes y significativos por fracción; Estrategia 2 (muestras: IK, IM y RN). El código de color es persistente con el de la figura 18

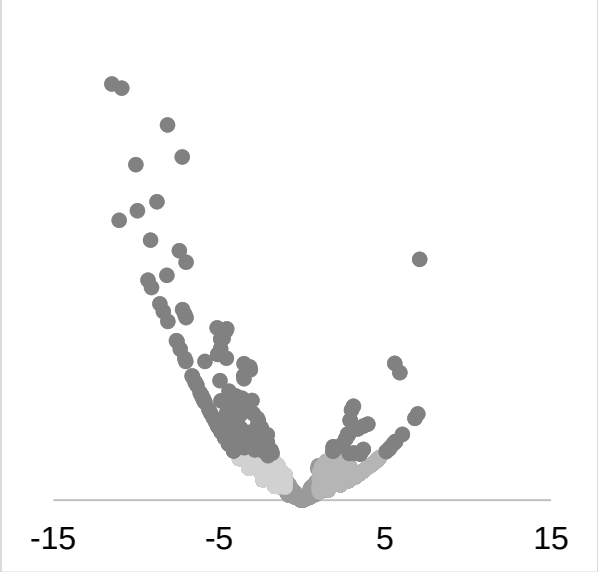


Figura 20. Gráfico de volcán representativo de los microRNAs abundantes y significativos por fracción; Estrategia 3 (muestras: BJ, M3 yM9). El código de color es persistente con el de la figura 18

Como último ensayo se analizó si existía intersección de las listas de microRNAs enriquecidos en exosomas entre cada estrategia de análisis descrita en la tabla 15**.** El siguiente diagrama de Venn (Figura 21) señala una intersección de 55 exo-microRNAs en común entre las tres estrategias de análisis. Esta relación puede interpretarse como una lista de transcritos que se expresan en común a pesar de las diversas condiciones experimentales de la cuales proceden las bibliotecas de secuenciación. En esta vía, la estrategia 1 (Bibliotecas de Línea celulares infectadas (EBV+): Leucemia de Burkit y Linfoma) demostró tener relación con 13 transcritos en la estrategia 2 (Línea celulares infectadas (EBV+): Linfoma) mientras que se hallaron 31 transcritos relacionados entre la estrategia 3 (Línea celulares parciales: Linfoma y linfoblastocitos) y estrategia 1. Las estrategias 2 y 3 no presentaron exo-microRNAs relacionados. Es importante resaltar que se descifraron microRNAs que pueden estar relativamente enriquecidos en exosomas sin relación alguna entre estas estrategias de análisis. De modo ascendente, 55, 4 y 38 son los valores de transcritos relativamente abundantes en exosomas a lo largo de cada estrategia (1, 2 y 3).

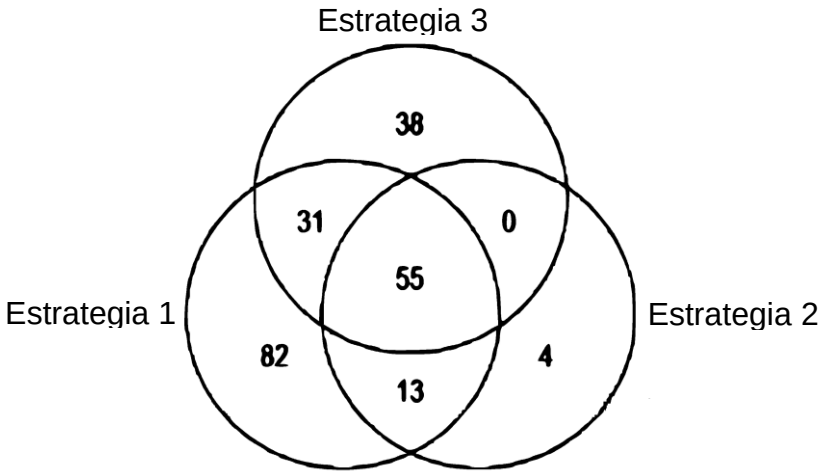


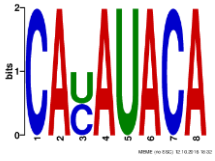
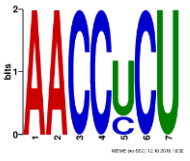
Figura 21. Intersección de las listas de microRNAs enriquecidos en exosomas entre cada estrategia de análisis descrito en la tabla 15

7.5. Búsqueda de elementos de acción en *cis*

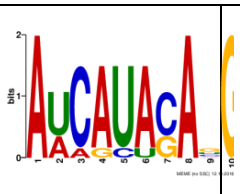
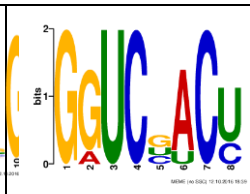
Durante esta tesis se llevó acabo la búsqueda de elementos de acción *cis* implementando la herramienta MEME en un método denominado modo discriminativo (DM, *Discriminative Mode*). Este método permite guiar la búsqueda de elementos de acción *cis* relativamente enriquecidos en la lista de secuencias de interés. Dicho método requirió dos archivos de entrada: 1) La lista de secuencias de microRNAs en las que se desea hallar el elemento de acción en cis (p.ej. Aquellos microRNAs enriquecidos en la fracción exosomal [exo-miRNAs] de cada estrategia de análisis) y 2) lista de control, es decir, aquellos microRNAs que no están enriquecidos en la fracción exosomal pero si en la celular [cel-miRNAs].

Desarrollamos el análisis de elementos de acción en *cis* como se describe en los métodos de esta tesis. Tomando como archivos de entrada las listas descritas en la tabla 17 donde se establece la lista de microRNAs que fueron significativos (Valor P < 0.05) para cada fracción (celular y exosomal).

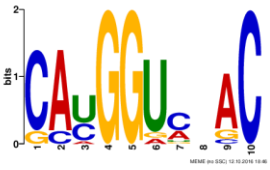
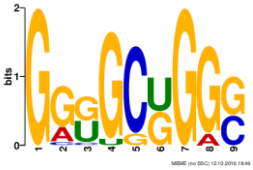
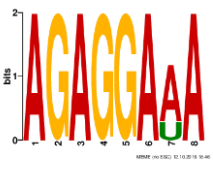
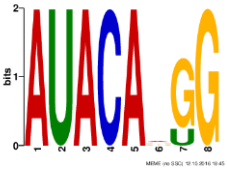
Como resultado, la lista de microRNAs enriquecidos en exosomas con elementos de acción *cis* en común de la estrategia 1 fue la siguiente:

	Estrategia 1_ fracción exosomal [IK, IM,M3,M9, RN]			
Elemento de acción				
Valor E	1.9e-006	6.2e+003	7.9e+003	4.9e+003

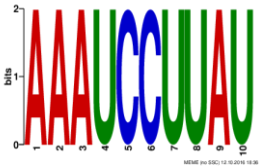
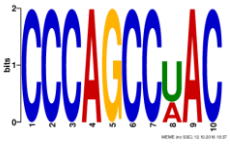
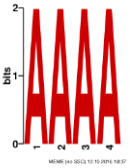
Por otro lado, la lista de microRNAs enriquecidos en exosomas con elementos de acción *cis* en común de la estrategia 2 corresponde a:

	Estrategia 2_ fracción exosomal [IK, IM, RN]			
Elemento de acción				
Valor E	7.6e-004	2.6e+002	8.5e+002	5.6e+002

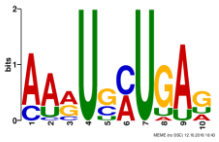
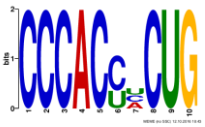
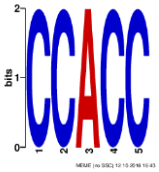
Finalmente, la lista de microRNAs enriquecidos en exosomas con elementos de acción *cis* en común de la estrategia 3 resultó en:

	Estrategia 3_ fracción exosomal [BJ, M3, M9]			
Elemento de acción				
Valor E	7.4e-003	7.5+001	1.5+002	2.4+003

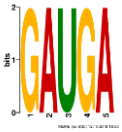
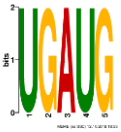
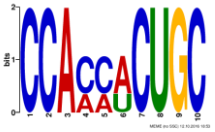
Mientras que la lista de microRNAs enriquecidos significativamente en la fracción celular con elementos de acción cis en común para la estrategia 1 resultó en:

	Estrategia 1_ fracción celular [IK, IM,M3,M9, RN]			
Elemento de acción				
Valor E	1.0e+001	3.3e+002	5.4e+002	2.2e+003

Mientras que la lista de microRNAs enriquecidos significativamente en la fracción celular con elementos de acción cis en común para la estrategia 2 resultó en:

	Estrategia 2_ fracción_celular [IK, IM,M3,M9, RN]			
Elemento de acción				
Valor E	1.6e+000	7.1e+001	6.4e+002	1.5e+003

Mientras que la lista de microRNAs enriquecidos significativamente en la fracción celular con elementos de acción cis en común para la estrategia 3 resultó en:

	Estrategia 3_ fracción_celular [IK, IM,M3,M9, RN]			
Elemento de acción				
Valor E	5.3e+000	5.0e+002	4.6e+002	2.3e+002

Como primera impresión, los elementos de acción cis para la fracción celular de cada estrategia presentaron un valor E positivo, lo cual se traduce como un umbral con poco significado estadístico. El parámetro E nos permite definir la probabilidad de que dicho elemento de acción *cis* exista de acuerdo a la lista de transcritos que introducimos en el análisis MEME; Cuanto menor sea el valor de E, más significado probabilístico tendrán los elementos de acción cis. En esta vía, solo el primer elemento de acción cis hallado a lo largo de las tres estrategias de la fracción exosomal presentaron un valor E con mayor significado estadístico que algún otro (Figura 22). Curiosamente, estos elementos de acción cis (exo-motivo) presentan un patrón similar (CAUGG). Por ello se consideró que existe un elemento de acción cis relacionado entre las estrategias 1, 2 y 3 que denominaremos elemento de acción en cis 1. La tabla 18 describe la lista de exo-miRNAs en los que se halló el elemento de acción en cis 1 a lo largo de los ensayos experimentales.

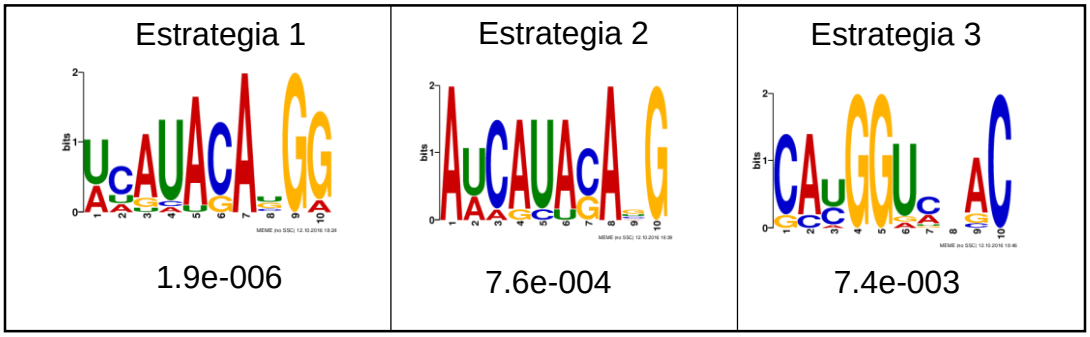


Figura 22. Elemento de acción en cis 1 encontrado a lo largo de las tres estrategias experimentales para la búsqueda elementos de la fracción exosomal. El valor E para cada estrategia se añade debajo de cada una de las figuras.

Tabla 18. Elemento de acción cis 1. Listas de exo-microRNAs que presentaran el elemento de acción cis 1 en común.

Estrategia 1	**Estrategia 2**	**Estrategia 3**
hsa.miR.487a.3p	hsa.miR.487a.3p	hsa.miR.379.3p
hsa.miR.485.3p	hsa.miR.376b.3p	hsa.miR.1197
hsa.miR.494.3p	hsa.miR.494.3p	hsa.miR.411.3p
hsa.miR.655.3p	hsa.miR.376c.3p	hsa.miR.323a.3p
hsa.miR.369.3p	hsa.miR.539.3p	hsa.miR.758.3p
hsa.miR.539.3p	hsa.miR.487b.3p	hsa.miR.655.3p
hsa.miR.1185.2.3p	hsa.miR.382.3p	hsa.miR.495.3p
hsa.miR.1185.1.3p	hsa.miR.369.3p	hsa.miR.382.3p
hsa.miR.487b.3p	hsa.miR.410.3p	hsa.miR.299.3p
hsa.miR.495.3p	hsa.miR.487a.3p	hsa.miR.494.3p
hsa.miR.376b.3p	hsa.miR.376b.3p	hsa.miR.493.5p
hsa.miR.382.3p	hsa.miR.494.3p	hsa.miR.363.3p
hsa.miR.493.5p	hsa.miR.376c.3p	hsa.miR.504.5p
hsa.miR.376c.3p		hsa.miR.493.3p
hsa.miR.496 (3p)		hsa.miR.379.3p
hsa.miR.410.3p		hsa.miR.1197 (3p)
		hsa.miR.411.3p
		hsa.miR.323a.3p
		hsa.miR.758.3p

7.6 Tailor: análisis de la adición de nucleótidos pos-transcripcional

Con el objetivo de visualizar los resultados de la paquetería Tailor, los archivos de salida correspondientes al alineamiento (ie. Las secuencias hairpin de los microRNAs versión 22 en las que mapearon las lecturas a lo largo de las bibliotecas pre-procesadas) fueron manipulados implementado el siguiente código en R; tomando como información de entrada los archivos con extensión .hairpin.single_nt_sum:

```
tables <- dir(pattern = ".*p20.hairpin.single_nt_sum")

table<-list()

n_count <- list()
```

Y ejecutando el siguiente pipeline para filtrar y almacenar la información de interés a lo largo de los archivos con extensión .hairpin.single_nt_sum:

```
for (i in 1:length(tables)) {

  table[[i]] <- read.delim(tables[i], sep = "")

  colnames(table[[i]]) = c("Length", "Tail", "Count")

table[[i]] = transform (table[[i]],

                        order = factor (Tail,

                                        levels=c('A',

                                                 'AA',

                                                 'U',

                                                 'UU'),

                                        ordered=TRUE))  table[[i]]= na.omit(table[[i]])

  table[[i]] <- table[[i]] %>% filter(Length<40)

    n_count[[i]] <- table[[i]] %>%

    select(Count)

   table[[i]]$Count<- cpm(n_count[[i]], normalized.lib.sizes=TRUE, log=T, prior.count=2)

write.table(table[[i]], sep = "\t")

}
```

El siguiente grupo de figuras resumen los eventos de adición de nucleótidos (Adenilación y Uridinilación) encontrados a lo largo de las bibliotecas (fracción celular y exosomal). Dichas figuras reportan mayor abundancia de eventos de adición de nucleótidos a lo largo de lecturas de 22 nt cuyo tamaño es consistente con las dimensiones de los miRNAs maduros. Este resultado conserva relación con las figuras 4 y 5 cuya información describe abundancia en secuencias de tamaño entre 22 a 23 nt en lecturas de las bibliotecas de las fracciones exosomales (figura 4) así como abundancia en secuencias del mismo tamaño en lecturas de las bibliotecas de las fracciones celulares (figura 5).

Aunque la distribución de eventos de adición de adenina simple (figura 23) y múltiple (figura 24) aparenta encontrarse en proporciones similares a lo largo de las bibliotecas de ambas fracciones se reporta que los eventos de adición simple y múltiple de uracilos son prominente en microRNAs de 22 nt de las bibliotecas de fracción exosomal (figura 25, panel izquierdo) en comparación con microRNAs de 22 nt que se encuentran en las bibliotecas de la fracción célula (figura 25, panel derecho), el valor acumulativo de los eventos de Uridinilación superan los 100 millones en las bibliotecas de la fracción celular y por debajo de los 85 millones en la fracción celular. Un comportamiento similar se observa en los eventos de múltiple Uridinilación (Figura 26).

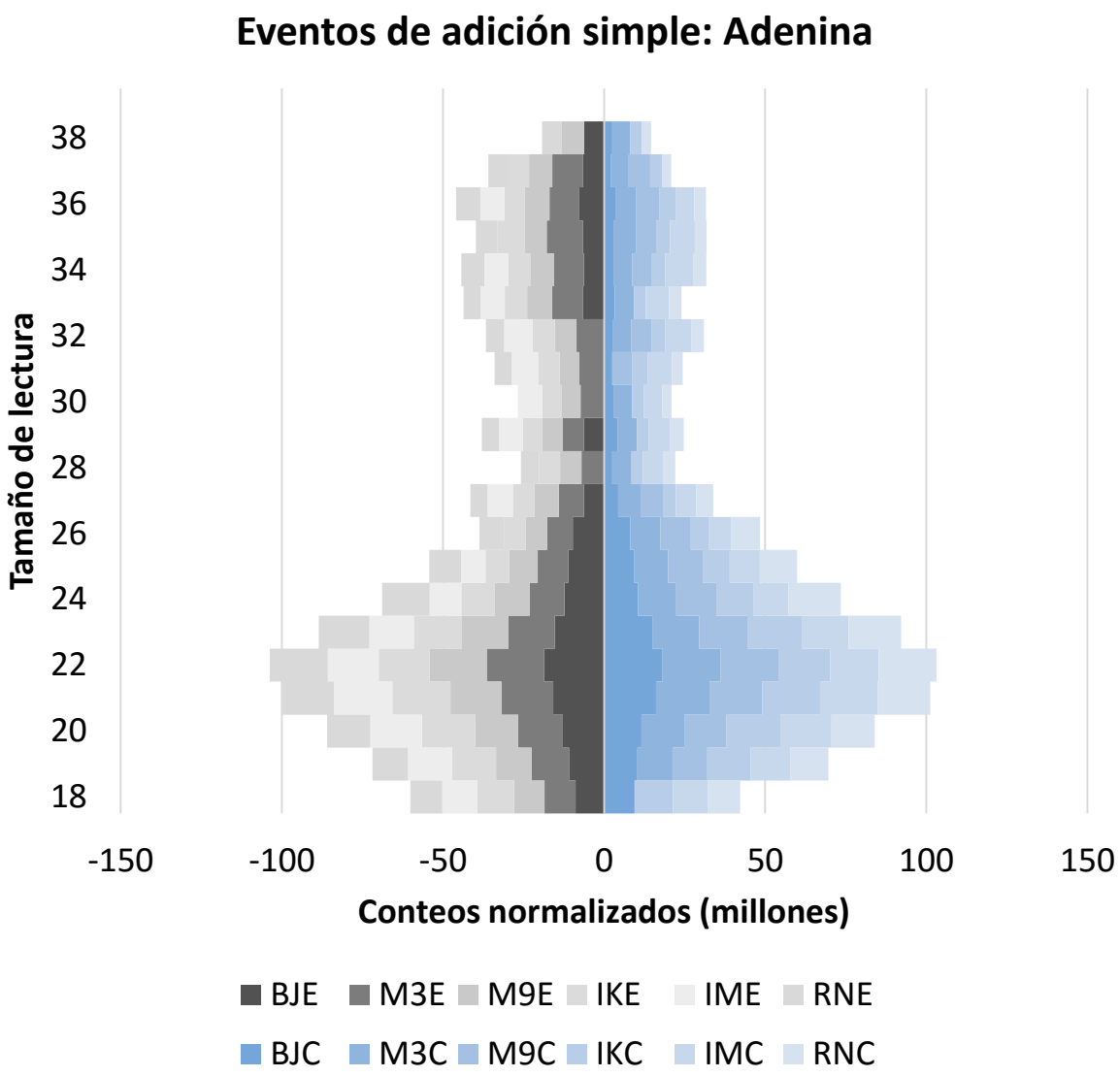


Figura 23. Abundancia de los eventos de adición simple, adenilación, a lo largo de las bibliotecas correspondientes a la fracción celular (Panel derecho) y exosomal (Panel izquierdo).

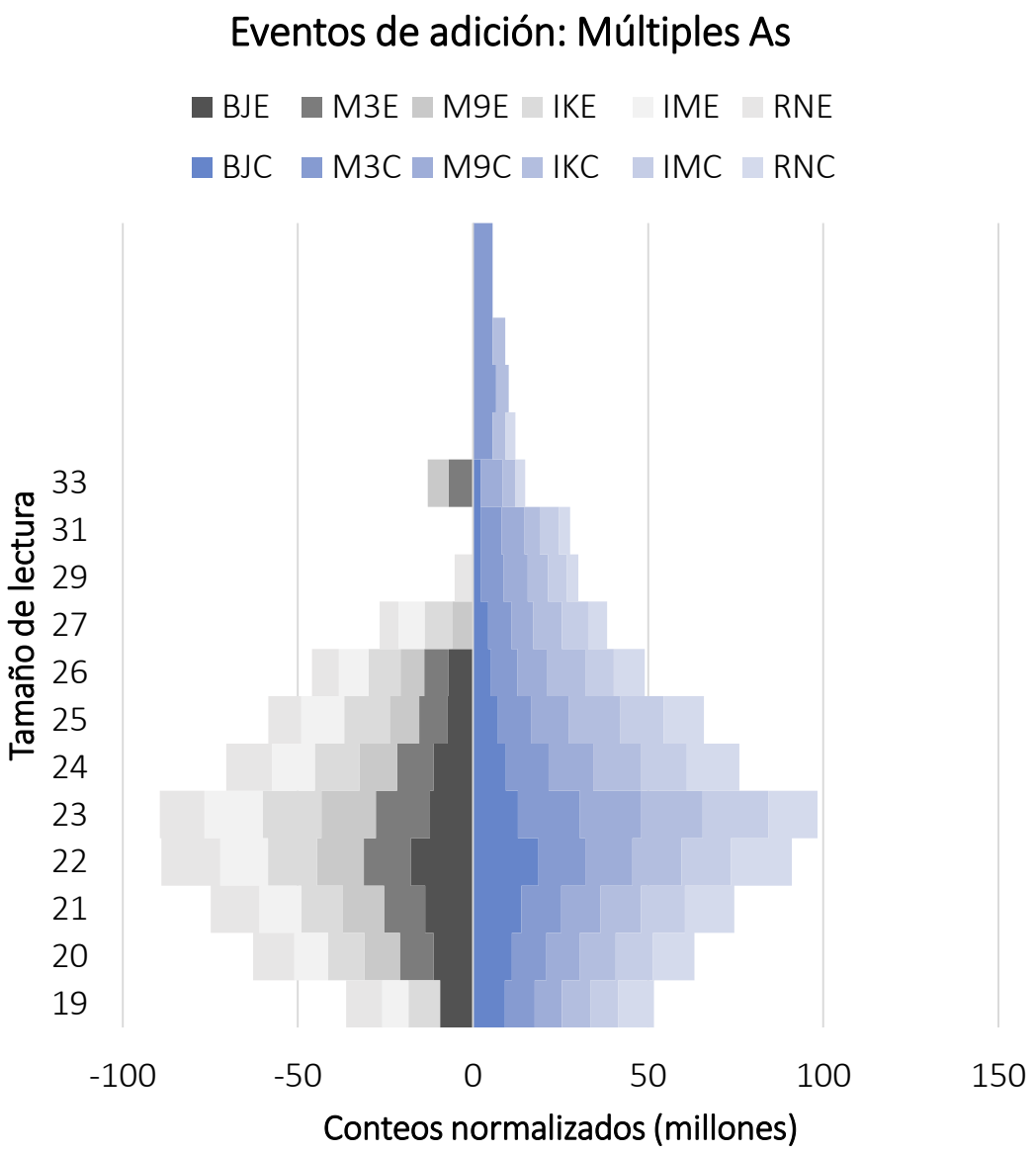


Figura 24. Abundancia de los eventos de adición Múltiple, Adenilación, a lo largo de las bibliotecas correspondientes a la fracción celular (Panel derecho) y exosomal (Panel izquierdo).

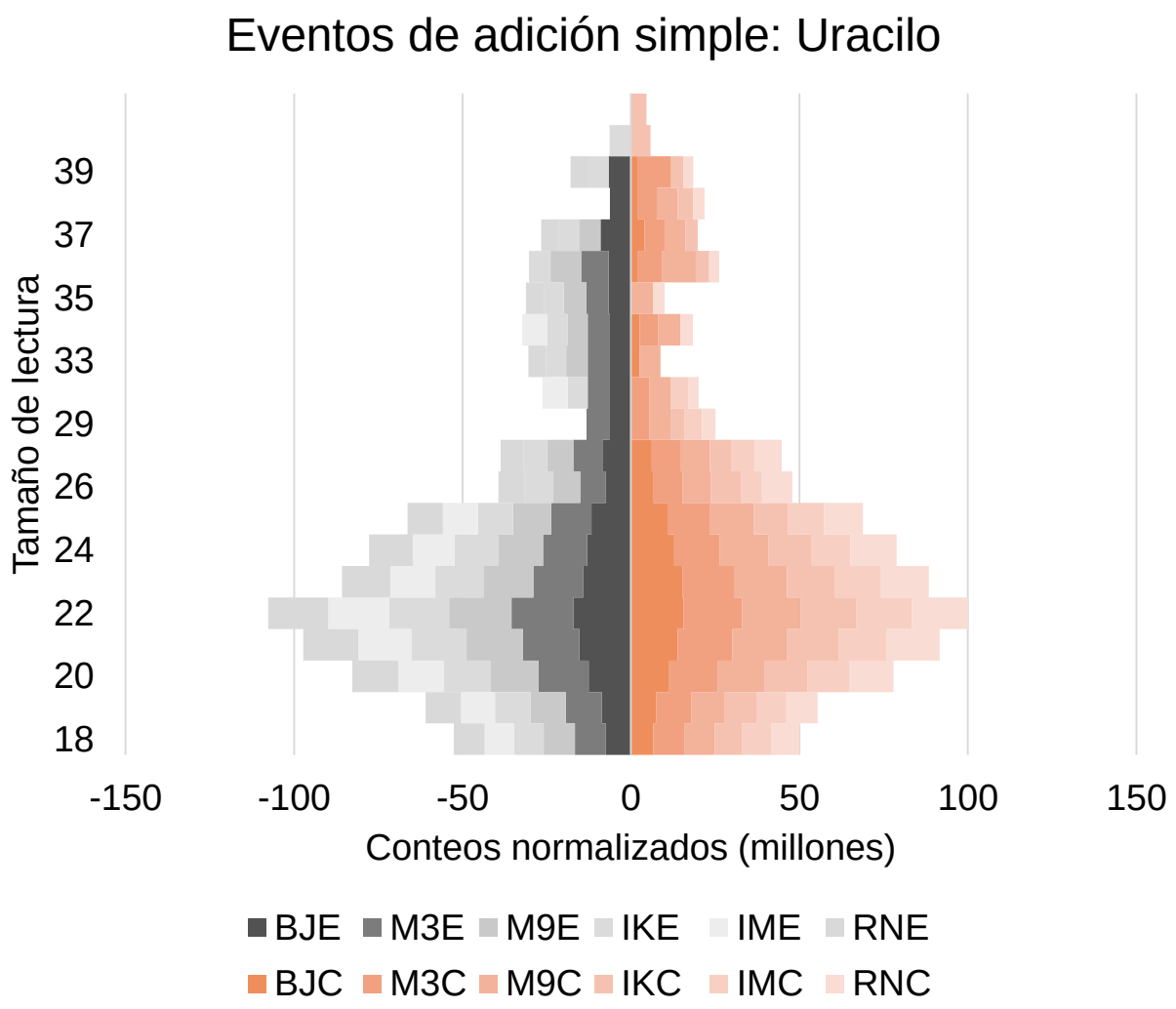


Figura 25. Abundancia de los eventos de adición simple, Uridinilación, a lo largo de las bibliotecas correspondientes a la fracción celular (Panel derecho) y exosomal (Panel izquierdo).

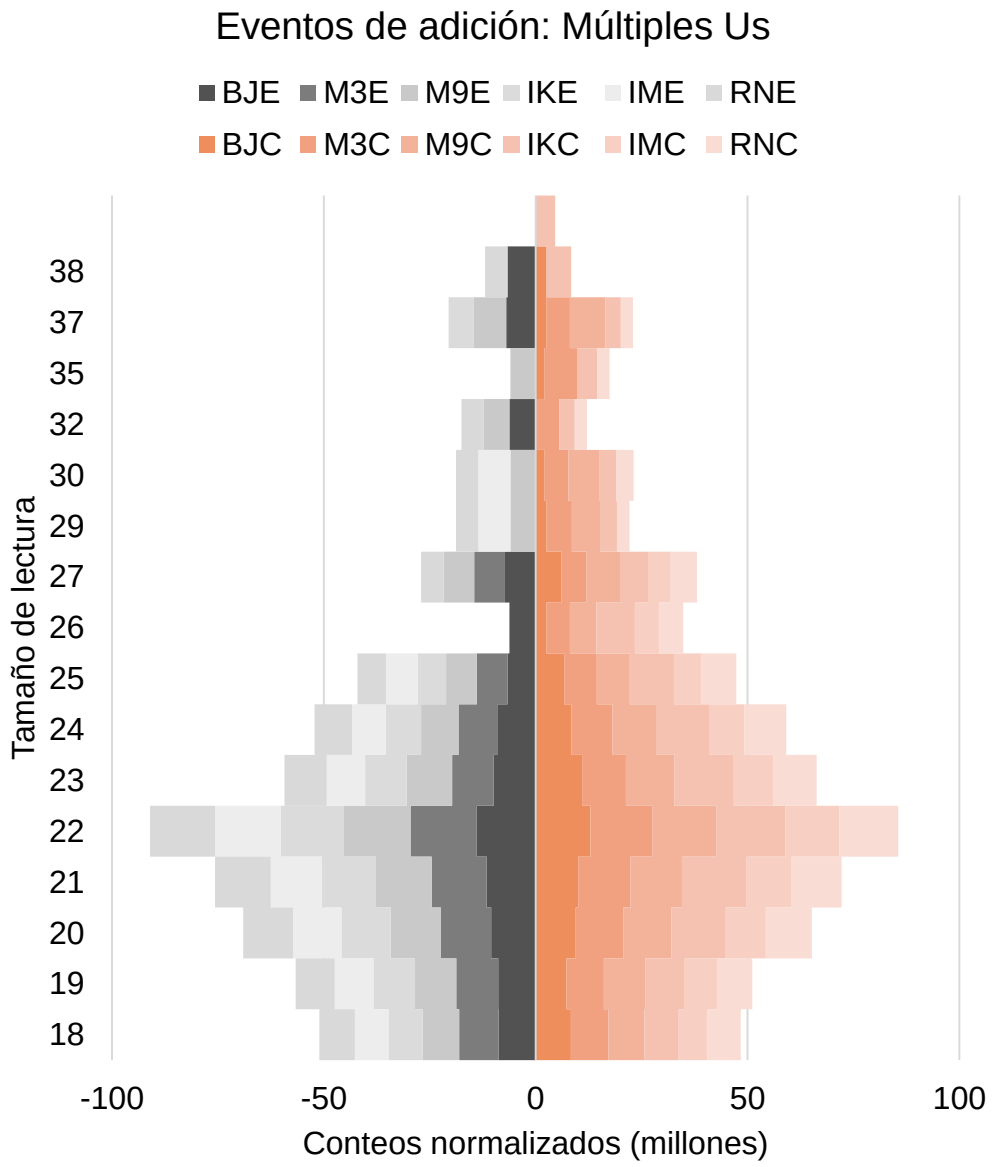


Figura 26. Abundancia de los eventos de adición Múltiple, Uridinilación, a lo largo de las bibliotecas correspondientes a la fracción celular (Panel derecho) y exosomal (Panel izquierdo).

7.7 Tailor: análisis de la adición de nucleótidos pos-transcripcional en los miRNAs abundantes

Así mismo, comparamos a detalle los eventos de *tailing* a lo largo de la lista de microRNAs abundantes significativamente entre ambas fracciones (celular y exosomal) (tabla 17); tomando como archivos todos con extensión relative.bed obtenidos por la herramienta Tailor; dichos archivos contienen información de interés correspondientes a los microRNAs anotados con su respectivo conteo en adición al potencial evento de *tailing* identificado en ellos. El siguiente código fue implementado para normalizar por cpm los conteos para los microRNAs hallados a lo largo de las bibliotecas de estudio durante esta tesis:

```
library(dplyr,warn= F)
library(edgeR)
setwd("C:/Users/Windows_user/Desktop/ESCRITO_TESIS_OCTUBRE/RESULTADOS/RESULTADOS_EBV/3)TAILING_MODIFICATION_TEST/balloon_input/")
tables <- dir(pattern = ".*relative.bed")
mirRelativePos<-list()
n_count <-list()

for (i in 1:length(tables)){ mirRelativePos[[i]] <- read.delim(tables[i], sep="\t",stringsAsFactors=F, header = F)
 n_count[[i]] <- mirRelativePos[[i]] %>%
  select(V4)
 mirRelativePos[[i]]$V4 <-as.integer(cpm((n_count[[i]]), normalized.lib.sizes=TRUE, log=T, prior.count=2))
 colnames(mirRelativePos[[i]]) = c("ID", "Start", "end",
                                   "count", "Loci", "strand",
                                   "sequence","tailing", "length")
}
```

Recordemos el análisis de la abundancia descrito en capítulos anteriores que sugiere el enriquecimiento de un repertorio de microRNAs en la fracción exosomal (exo-miRNAs);

podríamos hipotetizar que las modificaciones pos-trancripcionales son una característica de los exo-miRNAs; por ello, fue racional cuantificar los niveles de adición de uracilos, Adeninas, Guaninas y Citosinas reportados por la herramienta Tailor a lo largo de los exo-microRNAs.

La tabla 17 presenta valores distintos en el repertorio de exo-miRNAs a lo largo de las estrategias experimentales, es por ello que evaluamos los eventos de *tailing* de la lista de exo-miRNAs y ubicuos de la fracción celular (cel-miRNAs) correspondiente a cada una de las bibliotecas presentes en su respectiva estrategia experimental. Los gráficos de balón fueron desarrollados para esta etapa, implementando el siguiente seudo-código:

```
sample <-  mirRelativePos[[i]] %>%
  select(ID, tailing,count) %>%
  inner_join(d3, by=c("ID"="enriched_list"), copy=T)

# grouping tailing events
T_EVENTS <-  sample[i] [grep("T", sample[i]$tailing), ]
A_EVENTS <-  sample[i] [grep("A", sample[i]$tailing), ]
G_EVENTS <-  sample[i] [grep("GG", sample[i]$tailing), ]
C_EVENTS <-  sample[i] [grep("CC", sample[i]$tailing), ]

sum(T_EVENTS$count)
sum(A_EVENTS$count)
sum(C_EVENTS$count)
sum(G_EVENTS$count)

library(gplots)
balloonplot(x= sample[i]$tailing,
          y=list(sample[i] …),
          z=count.,
          sort=F,
          show.zeros=TRUE,
          cum.margins=FALSE,
          ylab=list("Desing","Samples"),
          xlab = "Event",
          text.size = 1,
          main=
            "BalloonPlot : tailing events"
)
```

La figura 27 y figura 28 reportan abundancia en adición de Timinas (Uracilos) y Adeninas en contraste a los eventos de *tailing* de Guaninas y Citosinas; curiosamente se observa un ligero aumento en los eventos de Uridinilación en los exo-miRNAs a lo largo de las bibliotecas (fracción exosomal) en cada una de las estrategias experimentales (Figura 27); inversamente, los eventos de Adenilación son mayoritarios en los cel-microRNAs de las bibliotecas de la fracción celular (Figura 28).

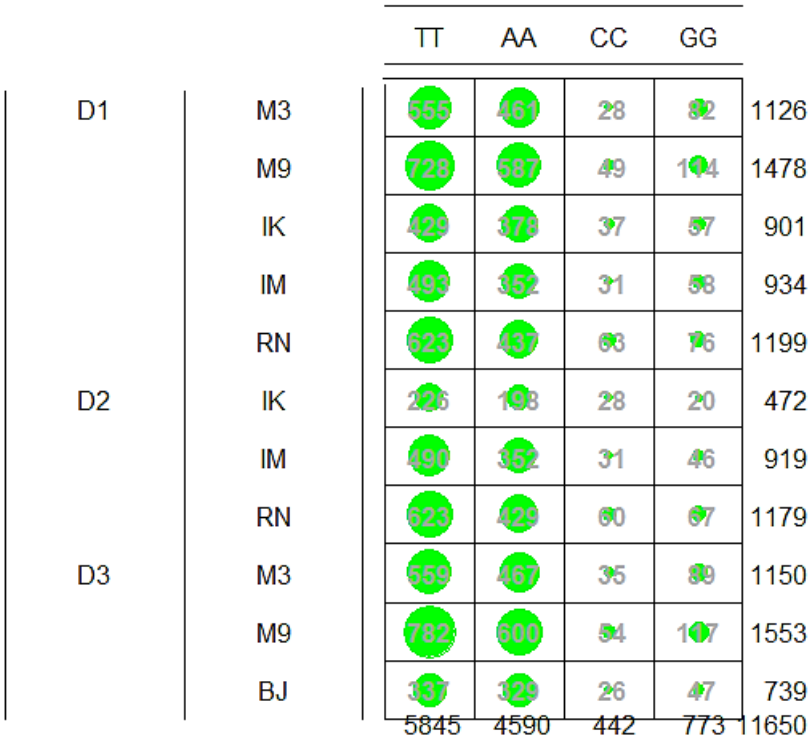


Figura 27. Conteos normalizados de microRNAs reportados con eventos de tailing en un gráfico de balón (*Balloonplot*) a lo largo de las estrategias experimentales; (Fracción exosomal).

		TT	AA	CC	GG	
D1	M3	113	462	21	53	649
	M9	163	509	5	41	718
	IK	58	399	4	36	497
	IM	163	771	12	67	1013
	RN	63	386	2	43	494
D2	IK	11	43	3	2	59
	IM	43	154	3	11	211
	RN	22	46	0	7	75
D3	M3	4	10	5	2	21
	M9	26	12	0	1	39
	BJ	-9	-26	-13	-9	-57
		657	2766	42	254	3719

Figura 28. Conteos normalizados de microRNAs reportados con eventos de tailing en un gráfico de balón (*Ballonplot*) a lo largo de las estrategias experimentales; Fracción celular.

Finalmente, sensibilizamos la prueba comparativa enfocándose entre aquellos exo-microRNAs que presentaron el elemento de acción en *cis* en común (tabla 18) y los archivos relative bed de la fracción exosomal discriminando los exo-miRNAs repetidos entre las estrategias experimentales y utilizando una sola lista de los exo-miRNAs descritos en la tabla 18; en esta vía nos enfocamos en los eventos de Adenilación y Uridinilación a lo largo de los exo-miRNAs, desafortunadamente no se encontró algún evento de *tailing* en Uracilos y Adeninas en algún exo-miRNA reportado con un elemento de acción en *cis* en común.

Debido a estos resultados, enfocamos un nueva prueba comparativa nuevamente entre los exo-miRNAs que presentaron un elemento de acción en *cis* en común (tabla 18) y los archivos relative bed de la fracción celular. Esta prueba fue guiada en el sentido de identificar el repertorio de miRNAs enriquecidos en la fracción exosomal en concentraciones relativas en su fracción citoplasmática; curiosamente, ningún conteo representativo o evento de *tailing* fue hallado a lo largo de los archivos relative bed de las respectivas fracciones celulares.

Esto último pudiera deberse a la eficiencia de la herramienta Tailor para anotar mapear lecturas a las miRNAs de referencia en comparación a la herramienta miRDeep2; es por ello que implementamos una búsqueda definitiva de los exo-miRNAs con un elemento e acción en cis en común a lo largo de los conteos normalizados del archivo 'miRNAs_expressed_all_samples_counts.csv' obtenido del módulo quantifier; esta vez tomamos conteos normalizados adicionando el cálculo correctivo de la varianza implementado el método descrito por Anderson et. Al. (2010) y la herramienta DESeq (Anders & Huber, 2010) descrito en los métodos de esta tesis. Como resultado, la figura 28 reporta el enriquecimiento de los microRNAs con elemento de acción cis en común únicamente en las bibliotecas de la fracción exosomal. En el apartado de discusiones nos referiremos a más detalle a este resultado.

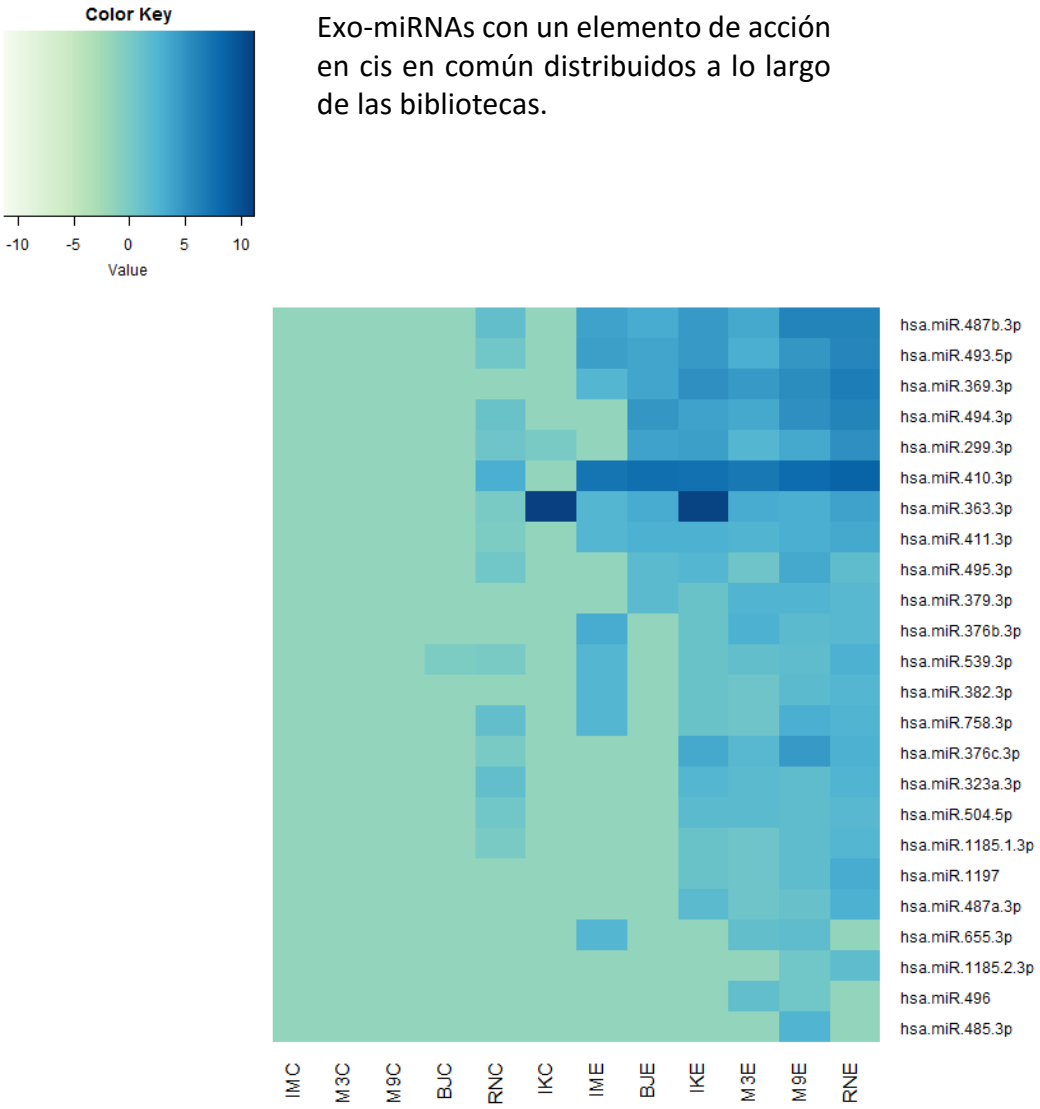


Figura 28. *Heatmap* comparativo del repertorio de miRNAs enriquecidos en la fracción exosomal con un elemento de acción en *cis* en común a lo largo de las bibliotecas.

Capítulo 8. Discusión

Respecto a la figura 22, se observa que el primer elemento de acción *cis* para la estrategia 1 conserva una relación prominente con el primer elemento de acción cis de la estrategia 2 en la secuencia CAUACA. También, se describe que la secuencia CAUGG del elemento de acción cis de la estrategia 3 presenta un patrón similar a lo largo de las estrategias restantes. Para cada respectiva estrategia, la tabla 18 señala la lista de exo-microRNAs que demostraron tener dicho patrón en común. De esta lista, reportamos la intersección de 2 exo-microRNAs (hsa.miR.494.3p y hsa.miR.382.3p) que presentaron el elemento de acción en cis 1 a lo largo de las tres estrategias (tabla19); En la tabla 19 se añade el enriquecimiento (valor LogFC) del par de exo-miRNAs mencionados a lo largo de las estrategias; en todos los casos, los valores de enriquecimiento LogFC se encuentran en los parámetros significativos del valor $P < 0.05$ (dato no mostrado). El diagrama de ven en la figura 29 resume la prueba comparativa descrita, como resultados adicionales, se identifica a los extremos del diagrama que las estrategias 1,2 y 3 tienen respectivamente cuatro, cero y nueve exo-microRNAs con el elemento de acción en cis 1 únicamente en esas estrategias experimentales. Esto último puede obviarse por la variación de la procedencia experimental de las bibliotecas de secuenciación analizadas.

Tabla 19. Lista de exo-microRNAs que conservaron el elemento de acción cis 1 en común en adición al valor LogFC del enriquecimiento exosomal.

	Exo-microRNA con elemento de accion cis 1	LogFC		
		Estrategia 1	Estrategia 2	Estrategia 3
intersección (Grupo 1, 2 y 3)	**hsa.miR.494.3p**	-5.47	-5.14	-8.09
	hsa.miR.382.3p	-4.77	-4.98	-4.41

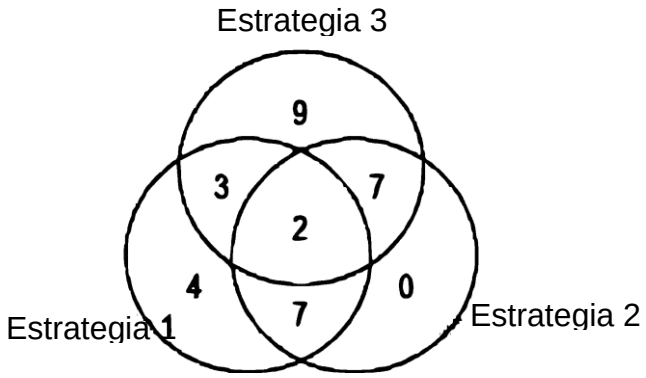


Figura 29. Diagrama de venn de exo-microRNAs con el elemento de acción en cis 1

Por otro lado, basada en la figura 30, elaboramos la tabla 22 que describe aquellos exo-miRNAs con un elemento de acción cis en común únicos por cada estrategia del análisis descrito en la tabla 15; al comparar los exo-miRNAs únicos de la estrategia 1, miR-496 y miR-485-3p, se reporta un valor logFC igual a cero en la estrategia 2 (tabla 22); es importante denotar que a diferencia de la estrategia 2, la estrategia 1 incluye las bibliotecas M3 y M9; en ambas bibliotecas de la fracción exosomal se reportan la expresión de miR-496, mientras que miR-485-3p solo se ve expresado en la biblioteca M9E; ambos miRNAs con un elemento de acción en cis en común no reportan expresión en alguna de las bibliotecas de la fracción celular y el resto de las bibliotecas de la fracción celular de la estrategia 1. Por otro lado, la comparación de exo-miRNA miR-363-3p de la estrategia 3 reporta valores logFC positivos a lo largo de las estrategias restantes; es decir, para las estrategias 1 y 2 este microRNA podría ser abundante en la fracción celular en vez de la exosomal aunque sin relevancia significativa (tabla 19).

En la figura 28, se observa que la expresión del miR-363-3p es consistente a lo largo de las bibliotecas de la fracción exosomal y curiosamente se ve expresado en mayor grado en ambas fracciones de la muestra IK; nuevamente nos referimos al diseño de las estrategias, a diferencia de las estrategias 3, únicamente las estrategias 1 y 2 incluyen a las bibliotecas IK; esto de algún modo podría dar razón a los valores logFC positivos de dicho miRNA a lo largo de las estrategias 1 y 2 visto en la tabla 22 pero sin relevancia significativa alguna para el análisis de búsqueda elementos de acción en cis.

Tabla 20. Lista de exo-miRNAs con un elemento de acción cis en común sin intersección aparente entre estrategias. Esta tabla se interpreta comparando las celdas de las estrategias 1 a 3. También se puede comparar el enriquecimiento relativo de los microRNAs entre las estrategias

	Exo-miRNA		Estrategia 1	Estrategia 2	Estrategia 3
		LogFC	LogFC		
Estrategia 1	hsa.miR.485.3p	-4.12		**0.00**	-4.86
	hsa.miR.1185.2.3p	-3.03		-3.36	-2.35
	hsa.miR.1185.1.3p	-2.73		-2.51	-4.11
	hsa.miR.496	-3.02		**0.00**	-3.73
Estrategia 3	hsa.miR.379.3p	-5.85	-5.35	-4.29	
	hsa.miR.1197	-4.11	-4.99	-5.28	
	hsa.miR.411.3p	-6.14	-5.03	-4.75	
	hsa.miR.323a.3p	-4.66	-1.75	-1.65	
	hsa.miR.758.3p	-5.35	-2.00	-1.56	
	hsa.miR.299.3p	-6.60	-4.39	-4.31	
	hsa.miR.363.3p	-6.45	**0.59**	**0.65**	
	hsa.miR.504.5p	-4.66	-2.58	-2.15	
	hsa.miR.493.3p	-5.85	-7.73	-7.75	

Por otro lado, encontramos que la mayoría de los microRNAs que conservaban el elemento de acción *cis* 1 a lo largo de las estrategias de análisis forman parte de la hebra guía -3p. La tabla 20 señala la distribución de hebras a lo largo de la lista de microRNAs que conservan el elemento de acción *cis* 1 en común.

Tabla 21. Distribución de hebras a lo largo de la lista de microRNAs que conservan el elemento de acción cis 1 en común

Distribución	**Estrategia 1** [IK, IM,M3,M9, RN]	**Estrategia 2** [IK, IM, RN]	**Estrategia 3** [BJ, M3, M9]
-3 p	15	9	12
-5 p	1	0	2

Este hallazgo nos llevó a evaluar la distribución de hebras dentro de las listas de exo-miRNAs a lo largo de las estrategias para descifrar si existía una distribución mayoritaria de las hebras -3p. La tabla 21 demuestra que la distribución de hebras 5p y 3p de las listas de exo-miRNAs es aproximadamente proporcional.

Tabla 22. Distribución de hebras 5p y 3p de las listas de exo-miRNAs significativos a lo largo de las estrategias de análisis.

Distribución	**Estrategia 1** [IK, IM,M3,M9, RN]	**Estrategia 2** [IK, IM, RN]	**Estrategia 3** [BJ, M3, M9]
-3 p	34	54	53
-5 p	33	56	51

De este resultado, se puede inferir: 1) que el empaquetamiento podría estar dirigido por alguna proteína de unión a RNA (RBP) que reconoce el elemento de acción cis encontrado en estos exo-miRNAs y 2) en el transcurso previo al empaquetamiento la selección de los exo-miRNAs puede estar favorecida por la accesibilidad estructural del brazo -3p o -5p de los pre-miRNAs.

Cabe mencionar que, al igual que los RNAs mensajeros (mRNAs) se ha reportado que los miRNAs son empaquetados mediante proteínas RBP que reconocen elementos de la secuencia de acción cis en microRNAs del citoplasma (Santangelo et al., 2016; Villarroya-beltri et al., 2013) . Aunque en esta tesis se reporta un elemento de acción distinto al reportado por otros autores, nuestros resultados sugieren que elementos en secuencia de acción cis además de la adición de nucleótidos uracilo en el sitio -3p son un rasgo de exo-miRNAs y podrían ser factores asociados a la interacción RNA/Proteína para el empaquetamiento de microRNAs en exosomas. Esta idea puede ser puesta a prueba con la biogénesis conocida de los microRNAs para regular su actividad en el citoplasma. Se ha estudiado a fondo la interacción y estructura del complejo lin28/TUT4 para unirse al miRNA precursor (pre-mir) let-7 regulando su actividad y decaimiento (Heo et al., 2009; Nam, Chen, Gregory, Chou, & Sliz, 2011; Thornton, Chang, Piskounova, & Gregory, 2012). Lin28 conserva dos dominios de unión a RNA, CSD (Cold Shock Domain) y CCHCx2 (repeticiones Cys-Cys-His-Cys) que reconocen el lazo formado en la estructura dsRNA de los pre-miRNA y el elemento de acción cis GGAG de pre-miRNA let-7 respectivamente. Usualmente, la enzima TUT4/zcchc11 es reclutada por lin28 para añadir un “tallo” de uracilos en la región 3’ de la secuencia precursora del miRNA evitando así su maduración; este mecanismo dependiente de lin28 ha sido reportado en otras poblaciones de miRNAs en mamíferos (Thornton et al., 2014); aunque este mecanismo induce la degradación de miRNAs hay que mencionar que la poli-

Uridinilación funge como señal para dirigir RNAs hacia el Endosoma donde posteriormente se forman Vesículas Intraluminales (más tarde exosomas).

Basado en esto, realizamos una búsqueda en la base de datos RNPDB.com con el objetivo de hallar alguna coincidencia entre el elemento de acción en cis encontrado en esta tesis y alguna proteína reportada de unión. Datos no mostrados en esta tesis dieron como resultado la proteína YB-1 (o YBX-1) con potencial reconocimiento al elemento de acción en cis CAUGG que a su vez reportamos tienen homología en secuencia con la proteína Lin28 a lo largo de un dominios CDS. Aunque no se ha reportado la interacción entre YB-1 y Lin28, la literatura ha reportado la asociación entre YB-1 y la proteína MBNL1 como respuesta al estrés en células humanas (Onishi et al., 2008); curiosamente MBNL1 está directamente asociada a Lin28 durante la regulando y decaimiento de microRNAs (Rau et al., 2011).

Literatura citada

Ameres, S. L., & Zamore, P. D. (2013). Diversifying microRNA sequence and function. *Nature Reviews. Molecular Cell Biology*, *14*(8), 475–88. Recuperado de: http://doi.org/10.1038/nrm3611

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106. Recuperado de: http://doi.org/10.1186/gb-2010-11-10-r106

Appasani, K. (2009). *MicroRNAs: From Basic Science to Disease Biology*. (K. Appasan, V. R. Ambros, & S. Altman, Eds.) (1st ed.). New York: Cambridge University Press. Recuperado de: http://www.cambridge.org/us/academic/subjects/life-sciences/genomics-bioinformatics-and-systems-biology/microrna s-basic-science-disease-biology?format=HB

Arroyo, J. D., Chevillet, J. R., Kroh, E. M., Ruf, I. K., Pritchard, C. C., Gibson, D. F., … Tewari, M. (2011). Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(12), 5003–8. Recuperado de:http://doi.org/10.1073/pnas.1019055108

Ason, B., Darnell, D. K., Wittbrodt, B., Berezikov, E., Kloosterman, W. P., Wittbrodt, J., … Plasterk, R. H. A. (2006). Differences in vertebrate microRNA expression. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(39), 14385–9. Recuperado de: http://doi.org/10.1073/pnas.0603529103

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., … Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, *37*(Web Server issue), W202-8. Recuperado de: http://doi.org/10.1093/nar/gkp335

Bartel, D. P. (2004). MicroRNAsGenomics, Biogenesis, Mechanism, and Function. *Cell*, *116*(2), 281–297. Recuperado de:http://doi.org/10.1016/S0092-8674(04)00045-5

Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell*, *136*(2), 215–33. Recuperado de: http://doi.org/10.1016/j.cell.2009.01.002

Batagov, A. O., Kuznetsov, V. A., & Kurochkin, I. V. (2011). Identification of nucleotide patterns enriched in secreted RNAs as putative cis-acting elements targeting them to exosome nano-vesicles. *BMC Genomics*, *12 Suppl 3*(Suppl 3), S18. Recuperado de: http://doi.org/10.1186/1471-2164-12-S3-S18

Bergmann, C., Strauss, L., Wieckowski, E., Czystowska, M., Albers, A., Wang, Y., … Whiteside, T. L. (2009). Tumor-derived microvesicles in sera of patients with head and neck cancer and their role in tumor progression. *Head & Neck*, *31*(3), 371–80. Recuperado de: http://doi.org/10.1002/hed.20968

Bobrie, A., Krumeich, S., Reyal, F., Recchi, C., Moita, L. F., Seabra, M. C., … Théry, C. (2012). Rab27a supports exosome-dependent and -independent mechanisms that modify the tumor microenvironment and can promote tumor progression. *Cancer Research*, *72*(19), 4920–30. Recuperado de: http://doi.org/10.1158/0008-5472.CAN-12-0925

Buck, A. H., Coakley, G., Simbari, F., McSorley, H. J., Quintana, J. F., Le Bihan, T., … Maizels, R. M. (2014). Exosomes secreted by nematode parasites transfer small RNAs to mammalian cells and modulate innate immunity. *Nature Communications*, *5*, 5488. Recuperado de: http://doi.org/10.1038/ncomms6488

Cech, T. R. (2012). The RNA worlds in context. *Cold Spring Harbor Perspectives in Biology*, *4*(7), 1–5. Recuperado de: http://doi.org/10.1101/cshperspect.a006742

Cha, D. J., Franklin, J. L., Dou, Y., Liu, Q., Higginbotham, J. N., Demory Beckler, M., … Patton, J. G. (2015). KRAS -dependent sorting of miRNA to exosomes. *eLife*, *4*, e07197. Recuperado de: http://doi.org/10.7554/eLife.07197

Chahar, H. S., Bao, X., & Casola, A. (2015). Exosomes and Their Role in the Life Cycle and Pathogenesis of RNA Viruses. *Viruses*, *7*(6), 3204–25. Recuperado de: http://doi.org/10.3390/v7062770

Chen. Yunshun, T. L. Aaron, L. and G. K. S. (2014). *Differential Expression Analysis of Complex RNA-seq Experiments Using edgeR*.

Chou, M.-T., Han, B. W., Hsiao, C.-P., Zamore, P. D., Weng, Z., & Hung, J.-H. (2015). Tailor: a computational framework for detecting non-templated tailing of small silencing RNAs. *Nucleic Acids Research*, *43*(17), e109. Recuperado de: http://doi.org/10.1093/nar/gkv537

Colombo, M., Raposo, G., & Théry, C. (2014). Biogenesis, Secretion, and Intercellular Interactions of Exosomes and Other Extracellular Vesicles. *Annual Review of Cell and Developmental Biology*, *30*(1), 255–289. Recuperado de: http://doi.org/10.1146/annurev-cellbio-101512-122326

Craig, N. L., Cohen-Fix, O., Green, R., Greider, C. W., Storz, G., & Wolberger Cynthia. (2010). Genomes and the flow of biological information. In *Molecular Biology: Principles of Genome Function* (p. 839). OUP Oxford. Recuperado de: https://books.google.com/books?id=iX6sAQAAQBAJ&pgis=1

Diaz, A., & Ahlquist, P. (2012). Role of host reticulon proteins in rearranging membranes for positive-strand RNA virus replication. *Current Opinion in Microbiology*, *15*(4), 519–24. Recuperado de: http://doi.org/10.1016/j.mib.2012.04.007

Fanning, A. S., Jameson, B. J., Jesaitis, L. A., & Anderson, J. M. (1998). The tight junction protein ZO-1 establishes a link between the transmembrane protein occludin and the actin cytoskeleton. *The Journal of Biological Chemistry*, *273*(45), 29745–53. Recuperado de: http://www.ncbi.nlm.nih.gov/pubmed/9792688

Fehrmann, R. S. N., Karjalainen, J. M., Krajewska, M., Westra, H.-J., Maloney, D., Simeonov, A., … Franke, L. (2015). Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nature Genetics*, *47*(2), 115–25. Recuperado de: http://doi.org/10.1038/ng.3173

Février, B., & Raposo, G. (2004). Exosomes: endosomal-derived vesicles shipping extracellular messages. *Current Opinion in Cell Biology*, *16*(4), 415–21. Recuperado de: http://doi.org/10.1016/j.ceb.2004.06.003

Friedman, R. C., Farh, K. K.-H., Burge, C. B., & Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, *19*(1), 92–105. Recuperado de: http://doi.org/10.1101/gr.082701.108

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., … Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Research*, *17*(6), 669–81. Recuperado de: http://doi.org/10.1101/gr.6339607

Gilad, Y., & Mizrahi-Man, O. (2015). A reanalysis of mouse ENCODE comparative gene expression data [v1; ref status: indexed, http://f1000r.es/5ez]. *F1000Research*, *121*(4), 1–32. Recuperado de: http://doi.org/10.12688/f1000research.6536.1

Gilbert, S. F. (2014). Differential gene expresión in development. In *Developmental Biology* (p. 719). Sinauer. Recuperado de: https://books.google.com/books?id=_52xmgEACAAJ&pgis=1

Gould, S. J., Booth, A. M., & Hildreth, J. E. K. (2003). The Trojan exosome hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(19), 10592–7. Recuperado de: http://doi.org/10.1073/pnas.1831413100

Gould, S. J., & Raposo, G. (2013, February 15). As we wait: coping with an imperfect nomenclature for extracellular vesicles. *Journal of Extracellular Vesicles*. Recuperado de: http://www.journalofextracellularvesicles.net/index.php/jev/article/view/20389/26067

György, B., Szabó, T. G., Pásztói, M., Pál, Z., Misják, P., Aradi, B., … Buzás, E. I. (2011). Membrane vesicles, current state-of-the-art: emerging role of extracellular vesicles. *Cellular and Molecular Life Sciences : CMLS*, *68*(16), 2667–88. Recuperado de: http://doi.org/10.1007/s00018-011-0689-3

He, L., & Hannon, G. J. (2004). MicroRNAs: small RNAs with a big role in gene regulation. *Nature Reviews. Genetics*, *5*(7), 522–31. Recuperado de: http://doi.org/10.1038/nrg1379

Heo, I., Joo, C., Kim, Y.-K., Ha, M., Yoon, M.-J., Cho, J., … Kim, V. N. (2009). TUT4 in concert with Lin28 suppresses microRNA biogenesis through pre-microRNA uridylation. *Cell*, *138*(4), 696–708. Recuperado de: http://doi.org/10.1016/j.cell.2009.08.002

Hurley, J. H., Boura, E., Carlson, L.-A., & Różycki, B. (2010). Membrane budding. *Cell*, *143*(6), 875–87. Recuperado de: http://doi.org/10.1016/j.cell.2010.11.030

Janas, T., Janas, M. M., Sapoń, K., & Janas, T. (2015). Mechanisms of RNA loading into exosomes. *FEBS Letters*, *589*(13), 1391–1398. Recuperado de: http://doi.org/10.1016/j.febslet.2015.04.036

Johnstone, R. M., Adam, M., Hammond, J. R., Orr, L., & Turbide, C. (1987). Vesicle formation during reticulocyte maturation. Association of plasma membrane activities with released vesicles (exosomes). *The Journal of Biological Chemistry*, *262*(19), 9412–20. Recuperado de: http://www.ncbi.nlm.nih.gov/pubmed/3597417

Kim, D., Sung, Y. M., Park, J., Kim, S., Kim, J., Park, J., … Baek, D. (2016). General rules for functional microRNA targeting. *Nature Genetics*. Recuperado de: http://doi.org/10.1038/ng.3694

Koppers-lalic, D., Hacken-, M., Meijer, G. A., Pegtel, D. M., Koppers-lalic, D., Hackenberg, M., … Sadek, P. (2014). Nontemplated Nucleotide Additions Distinguish the Small RNA Composition in Cells from Exosomes Report Nontemplated Nucleotide Additions Distinguish the Small RNA Composition in Cells from Exosomes. *Cell Reports*, 1649–1658. Recuperado de: http://doi.org/10.1016/j.celrep.2014.08.027

Kosaka, N., Iguchi, H., Yoshioka, Y., Takeshita, F., Matsuki, Y., & Ochiya, T. (2010). Secretory mechanisms and intercellular transfer of microRNAs in living cells. *The Journal of Biological Chemistry*, *285*(23), 17442–52. Recuperado de: http://doi.org/10.1074/jbc.M110.107821

Krol, J., Loedige, I., & Filipowicz, W. (2010). The widespread regulation of microRNA biogenesis, function and decay. *Nature Reviews. Genetics*, *11*(9), 597–610. Recuperado de: http://doi.org/10.1038/nrg2843

Kukurba, K. R., & Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor Protocols*, pdb.top084970-. Recuperado de: http://doi.org/10.1101/pdb.top084970

Lee, R. C., Feinbaum, R. L., & Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, *75*(5), 843–854. Recuperado de: http://doi.org/10.1016/0092-8674(93)90529-Y

Li, W., & Lan, P. (2015). Re-analysis of RNA-seq transcriptome data reveals new aspects of gene activity in Arabidopsis root hairs. *Front Plant Sci*, *6*(June), 421. Recuperado de: http://doi.org/10.3389/fpls.2015.00421

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, *30*(7), 923–30. Recuperado de: http://doi.org/10.1093/bioinformatics/btt656

Lodish, H., Berk, A., Kaiser, C. A., Monty Krieger, Bretscher, A., Hidd, e P., … Scott. (2008). Biomembrane Structure. In *Molecular Cell Biology* (p. 1150). W. H. Freeman. Recuperado de: https://books.google.com/books?id=K3JbjG1JiUMC&pgis=1

Lotvall, J., & Valadi, H. (2007). Cell to cell signalling via exosomes through esRNA. *Cell Adhesion & Migration*, *1*(3), 156–8. Recuperado de: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2634021&tool=pmcentrez&rendertype=abstract

Marsh, M., & van Meer, G. (2008). Cell biology. No ESCRTs for exosomes. *Science (New York, N.Y.)*, *319*(5867), 1191–2. Recuperado de: http://doi.org/10.1126/science.1155750

Mathivanan, S., Ji, H., & Simpson, R. J. (2010). Exosomes: extracellular organelles important in intercellular communication. *Journal of Proteomics*, *73*(10), 1907–20. Recuperado de: http://doi.org/10.1016/j.jprot.2010.06.006

Meckes, D. G., Shair, K. H. Y., Marquitz, A. R., Kung, C.-P., Edwards, R. H., & Raab-Traub, N. (2010). Human tumor virus utilizes exosomes for intercellular communication. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(47), 20370–5. Recuperado de: http://doi.org/10.1073/pnas.1014194107

Mittelbrunn, M., Gutiérrez-Vázquez, C., Villarroya-Beltri, C., González, S., Sánchez-Cabo, F., González, M. Á., … Sánchez-Madrid, F. (2011). Unidirectional transfer of microRNA-loaded exosomes from T cells to antigen-presenting cells. *Nature Communications*, *2*, 282. Recuperado de: http://doi.org/10.1038/ncomms1285

Mobius, W., Ohno-Iwashita, Y., Donselaar, E. G. v., Oorschot, V. M. J., Shimada, Y., Fujimoto, T., … Slot, J. W. (2002). Immunoelectron Microscopic Localization of Cholesterol Using Biotinylated and Non-cytolytic Perfringolysin O. *Journal of Histochemistry & Cytochemistry*, *50*(1), 43–55. Recuperado de: http://doi.org/10.1177/002215540205000105

Montecalvo, A., Larregina, A. T., Shufesky, W. J., Stolz, D. B., Sullivan, M. L. G., Karlsson, J. M., … Morelli, A. E. (2012). Mechanism of transfer of functional microRNAs between mouse dendritic cells via exosomes. *Blood*, *119*(3), 756–66. Recuperado de: http://doi.org/10.1182/blood-2011-02-338004

Nam, Y., Chen, C., Gregory, R. I., Chou, J. J., & Sliz, P. (2011). Molecular basis for interaction of let-7 microRNAs with Lin28. *Cell*, *147*(5), 1080–91. Recuperado de: http://doi.org/10.1016/j.cell.2011.10.020

Nguyen, T. A., Jo, M. H., Choi, Y.-G., Park, J., Kwon, S. C., Hohng, S., … Woo, J.-S. (2015). Functional Anatomy of the Human Microprocessor. *Cell*, *161*(6), 1374–1387. Recuperado de: http://doi.org/10.1016/j.cell.2015.05.010

Nussinov, R., Bonhoeffer, S., Papin, J. A., & Sporns, O. (2015). From "What Is?" to "What Isn't?" Computational Biology. *PLoS Computational Biology*, *11*(7), e1004318. Recuperado de: http://doi.org/10.1371/journal.pcbi.1004318

Okoye, I. S., Coomes, S. M., Pelly, V. S., Czieso, S., Papayannopoulos, V., Tolmachova, T., … Wilson, M. S. (2014). MicroRNA-Containing T-Regulatory-Cell-Derived Exosomes Suppress Pathogenic T Helper 1 Cells. *Immunity*, *41*(1), 89–103. Recuperado de: http://doi.org/10.1016/j.immuni.2014.05.019

Onishi, H., Kino, Y., Morita, T., Futai, E., Sasagawa, N., & Ishiura, S. (2008). MBNL1 associates with YB-1 in cytoplasmic stress granules. *Journal of Neuroscience Research*, *86*(9), 1994–2002. Recuperado de: http://doi.org/10.1002/jnr.21655

Pan, B. T., Teng, K., Wu, C., Adam, M., & Johnstone, R. M. (1985). Electron microscopic evidence for externalization of the transferrin receptor in vesicular form in sheep reticulocytes. *The Journal of Cell Biology*, *101*(3), 942–8. Recuperado de: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2113705&tool=pmcentrez&rendertype=abstract

Pegtel, D. M., Cosmopoulos, K., Thorley-Lawson, D. A., van Eijndhoven, M. A. J., Hopmans, E. S., Lindenberg, J. L., … Middeldorp, J. M. (2010). Functional delivery of viral miRNAs via exosomes. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(14), 6328–33. Recuperado de: http://doi.org/10.1073/pnas.0914843107

Pegtel, D. M., van de Garde, M. D. B., & Middeldorp, J. M. (2011). Viral miRNAs exploiting the endosomal-exosomal pathway for intercellular cross-talk and immune evasion. *Biochimica et Biophysica Acta*, *1809*(11–12), 715–21. Recuperado de: http://doi.org/10.1016/j.bbagrm.2011.08.002

Peinado, H., Alečković, M., Lavotshkin, S., Matei, I., Costa-Silva, B., Moreno-Bueno, G., … Lyden, D. (2012). Melanoma exosomes educate bone marrow progenitor cells toward a pro-metastatic phenotype through MET. *Nature Medicine*, *18*(6), 883–91. Recuperado de: http://doi.org/10.1038/nm.2753

Piper, R. C., & Katzmann, D. J. (2007). Biogenesis and function of multivesicular bodies. *Annual Review of Cell and Developmental Biology*, *23*, 519–47. Recuperado de: http://doi.org/10.1146/annurev.cellbio.23.090506.123319

Raposo, G., Nijman, H. W., Stoorvogel, W., Liejendekker, R., Harding, C. V, Melief, C. J., & Geuze, H. J. (1996). B lymphocytes secrete antigen-presenting vesicles. *The Journal of Experimental Medicine*, *183*(3), 1161–72. Recuperado de: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2192324&tool=pmcentrez&rendertype=abstract

Raposo, G., & Stoorvogel, W. (2013). Extracellular vesicles: exosomes, microvesicles, and friends. *The Journal of Cell Biology*, *200*(4), 373–83. Recuperado de: http://doi.org/10.1083/jcb.201211138

Rau, F., Freyermuth, F., Fugier, C., Villemin, J.-P., Fischer, M.-C., Jost, B., … Charlet-Berguerand, N. (2011). Misregulation of miR-1 processing is associated with heart defects in myotonic dystrophy. *Nature Structural & Molecular Biology*, *18*(7), 840–5. Recuperado de: http://doi.org/10.1038/nsmb.2067

Robbins, P. D., & Morelli, A. E. (2014). Regulation of immune responses by extracellular vesicles. *Nature Reviews. Immunology*, *14*(3), 195–208. Recuperado de: http://doi.org/10.1038/nri3622

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, *26*(1), 139–40. Recuperado de: http://doi.org/10.1093/bioinformatics/btp616

Rung, J., & Brazma, A. (2012). Reuse of public genome-wide gene expression data. *Nature Reviews Genetics*, *14*(2), 89–99. Recuperado de: http://doi.org/10.1038/nrg3394

Santangelo, L., Giurato, G., Cicchini, C., Montaldo, C., Mancone, C., Tarallo, R., … Tripodi, M. (2016). The RNA-Binding Protein SYNCRIP Is a Component of the Hepatocyte Exosomal Machinery Controlling MicroRNA Sorting. *Cell Reports*, *17*(3), 799–808. Recuperado de: http://doi.org/10.1016/j.celrep.2016.09.031

Staff, S. R. A. S. (2011). Using the SRA Toolkit to convert .sra files into other formats. National Center for Biotechnology Information (US). Recuperado de: http://www.ncbi.nlm.nih.gov/books/NBK158900/

Sudmant, P. H., Alexis, M. S., & Burge, C. B. (2015). Meta-analysis of RNA-seq expression data across species, tissues and studies. *Genome Biology*, *16*(1), 287. Recuperado de: http://doi.org/10.1186/s13059-015-0853-4

Théry, C. (2011). Exosomes: secreted vesicles and intercellular communications. *F1000 Biology Reports*, *3*(15), 15. Recuperado de: http://doi.org/10.3410/B3-15

Thornton, J. E., Chang, H.-M., Piskounova, E., & Gregory, R. I. (2012). Lin28-mediated control of let-7 microRNA expression by alternative TUTases Zcchc11 (TUT4) and Zcchc6 (TUT7). *RNA (New York, N.Y.)*, *18*(10), 1875–85. Recuperado de: http://doi.org/10.1261/rna.034538.112

Thornton, J. E., Du, P., Jing, L., Sjekloca, L., Lin, S., Grossi, E., … Gregory, R. I. (2014). Selective microRNA uridylation by Zcchc6 (TUT7) and Zcchc11 (TUT4). *Nucleic Acids Research*, *42*(18), 11777–91. Recuperado de: http://doi.org/10.1093/nar/gku805

Trajkovic, K., Hsu, C., Chiantia, S., Rajendran, L., Wenzel, D., Wieland, F., … Simons, M. (2008). Ceramide triggers budding of exosome vesicles into multivesicular endosomes. *Science (New York, N.Y.)*, *319*(5867), 1244–7. Recuperado de: http://doi.org/10.1126/science.1153124

Turchinovich, A., Tonevitsky, A. G., & Burwinkel, B. (2016). Extracellular miRNA: A Collision of Two Paradigms. *Trends in Biochemical Sciences*, *41*(10), 883–892. Recuperado de: http://doi.org/10.1016/j.tibs.2016.08.004

Turchinovich, A., Tonevitsky, A. G., Cho, W. C., & Burwinkel, B. (2015). Check and mate to exosomal extracellular miRNA: new lesson from a new approach. *Frontiers in Molecular Biosciences*, *2*, 11. Recuperado de: http://doi.org/10.3389/fmolb.2015.00011

Valadi, H., Ekström, K., Bossios, A., Sjöstrand, M., Lee, J. J., & Lötvall, J. O. (2007). Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. *Nature Cell Biology*, *9*(6), 654–9. Recuperado de: http://doi.org/10.1038/ncb1596

Villarroya-beltri, C., Gutie, C., Martin-cofreces, N., Martinez-herrera, D. J., & Pascual-montano, A. (2013). Sumoylated hnRNPA2B1 controls the sorting of miRNAs into exosomes through binding to specific motifs. *Nature Communications*, 1–10. Recuperado de: http://doi.org/10.1038/ncomms3980

Wightman, B., Ha, I., & Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, *75*(5), 855–862. Recuperado de: http://doi.org/10.1016/0092-8674(93)90530-4

Zhou, W., Fong, M. Y., Min, Y., Somlo, G., Liu, L., Palomares, M. R., … Wang, S. E. (2014). Cancer-Secreted miR-105 Destroys Vascular Endothelial Barriers to Promote Metastasis. *Cancer Cell*, *25*(4), 501–515. Recuperado de: http://doi.org/10.1016/j.ccr.2014.03.007

Zitvogel, L., Regnault, A., Lozier, A., Wolfers, J., Flament, C., Tenza, D., … Amigorena, S. (1998). Eradication of established murine tumors using a novel cell-free vaccine: dendritic cell derived exosomes. *Nature Medicine*, *4*(5), 594–600. Recuperado de: http://doi.org/10.1038/nm0598-594