



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

MAESTRÍA Y DOCTORADO EN CIENCIAS BIOQUÍMICAS

ANÁLISIS DE LA ESTRUCTURA DE LA RED DE REGULACIÓN POR NCRNA EN
PACIENTES CON CÁNCER DE MAMA

T E S I S

QUE PARA OPTAR POR EL GRADO DE:
MAESTRO EN CIENCIAS

PRESENTA:

DIANA DRAGO-GARCÍA

TUTOR PRINCIPAL

DR. ENRIQUE HERNÁNDEZ LEMUS
INSTITUTO NACIONAL DE MÉDICINA GENÓMICA

MIEMBROS DEL COMITÉ TUTOR

DR. ALFREDO HIDALGO MIRANDA
INSTITUTO NACIONAL DE MÉDICINA GENÓMICA

DR. ERNESTO PÉREZ RUEDA
INSTITUTO DE BIOTECNOLOGÍA

CIUDAD DE MÉXICO. Junio, 2017



Universidad Nacional
Autónoma de México

Dirección General de Bibliotecas de la UNAM

Biblioteca Central



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Formato: mgt_1
VoBo Tutor,
Revisión de tesis de Maestría.

02/03/2017
dd-mm-aaaa

Subcomité Académico

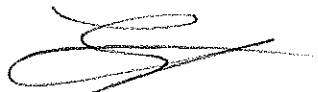
Yo Dr.: **Enrique Hernández Lemus** Tutor (asesor de tesis) del estudiante: **Diana Drago García.**

Manifiesto haber leído, revisado y corregido el manuscrito de tesis que lleva como título:

"Análisis de la estructura de la red de regulación por ncRNA en pacientes con cáncer de mama"

Por lo que autorizo a que sea entregado para su revisión a los sinodales que el Subcomité Académico asigne para la obtención del grado de Maestro en Ciencias por la Universidad Nacional Autónoma de México.

Atentamente



Dr. Enrique Hernández Lemus

Nombre completo y firma

Tutor

PMDCB/701/2017

Drago García Diana
Estudiante de Maestría en Ciencias Bioquímicas
P r e s e n t e

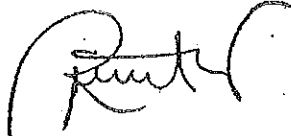
Los miembros del Subcomité Académico en reunión ordinaria del día 13 de marzo del presente año, conocieron su solicitud de asignación de **JURADO DE EXAMEN** para optar por el grado de **MAESTRÍA EN CIENCIAS**, con la réplica de la tesis "**Análisis de la estructura de la red de regulación por ncRNA en pacientes con cáncer de mama**", dirigida por el/la Dr(a) **Hernández Lemus Enrique**.

De su análisis se acordó nombrar el siguiente jurado integrado por los doctores:

PRESIDENTE	Morett Sánchez Juan Enrique
VOCAL	Flores Jasso Carlos Fabián
VOCAL	Espinal Enríquez Jesús
VOCAL	Jiménez Morales Silvia
SECRETARIO	Álvarez-Buylla Rocés María Elena

Sin otro particular por el momento, aprovecho la ocasión para enviarte un cordial saludo.

Atentamente
"Por mi raza hablará el espíritu"
Cd. Universitaria, Cd. Mx., a 13 de marzo de 2017.
COORDINADOR



Dr. ROBERTO CORIA ORTEGA

A mis padres

Agradecimientos

Quiero agradecer a todas personas que me apoyaron durante el desarrollo de este trabajo. Quiero agradecer de manera especial a mi tutor el Dr. Enrique Hernández Lemus por todo el apoyo que me ha brindado, y también al Dr. Jesús Espinal por sus valiosos consejos, ambos han formado parte importante de mi desarrollo académico y se han convertido también en amistades valiosas. También quiero agradecer a todos mis compañeros del laboratorio del CSB-IG que me ayudaron a introducirme en el mundo de la biología computacional, en especial a Rodrigo García y a Hugo Tovar que se tomaron el tiempo para ayudarme. Agradezco a los miembros de mi comité tutor, el Dr. Alfredo Hidalgo y el Dr. Ernesto Pérez Rueda por su guía y valiosos comentarios; y a los miembros de mi jurado por darse el tiempo de revisar este trabajo. Quiero agradecer a mi familia por su apoyo incondicional, en especial a mis padres Aída y Rolando. Y sobre todo quiero agradecer a Pedro Miramontes, por ofrecerme su sincera amistad y apoyo.

Agradezco al “Programa de Apoyo a los Estudios del Posgrado” (PAEP) y al Instituto Nacional de Medicina Genómica por los apoyos y facilidades brindados para el desarrollo de este proyecto.

“Un día, nos imaginamos que la biología y tratamiento del cáncer — que en el presente se trata de una combinación de retazos de biología celular, genética, histopatología, bioquímica, inmunología y farmacología— se convertirá en una ciencia con una estructura conceptual y coherencia lógica que rivalice con la química y la física”

— DOUGLAS HANAHAN & ROBER A WEINBERG, *The Hallmarks of Cancer* (2000)

Análisis de la Estructura de la red de Regulación por ncRNA en Pacientes con Cáncer de Mama

por

Diana Drago-García

Resumen

El cáncer de mama se define como un grupo de tumores epiteliales malignos que se originan en el tejido mamario, caracterizados por la invasión del tejido adyacente con una tendencia a formar metastasis. Como el cáncer más frecuente en mujeres a nivel mundial, el cáncer de mamá es un problema de salud pública. Grandes esfuerzos de investigación han demostrado que diferentes especies de RNAs pequeños, dentro de los que destacan los microRNAs (miRs), están relacionadas a las alteraciones transcripcionales características de esta enfermedad. Sin embargo la forma en la que los miRs participan en el establecimiento, progresión y diseminación del cáncer de mama aún no se entiende por completo. Para entender el papel de los miRs en la regulación del cáncer de mama construimos una red de regulación transcripcional miR-gen basada en Información Mutua (MI) con datos de secuenciación de tejido tumoral y tejido adyacente control de 86 pacientes del consorcio TCGA. Encontramos que la mayoría de los nodos y las interacciones forman parte de un 'componente gigante'. Aunque el número de conexiones por nodos (grado) es diferente entre las redes tumor y control, en ambos casos los nodos con los grados más altos resultaron ser miRs. Estudiando a los miRs como familias (debido a la relación estructural que mantienen). Encontramos que para la red de tumor la familia de miRs con grado más alto corresponde a miR-199, y miR-200 para los controles. Construimos subredes con los genes y miRs que tienen interacciones directas con los miRs de las familias miR-200 y miR-199, y encontramos que la mayoría de los miRs y los genes presentes se han asociado a la transición epitelio-mesenquima (EMT) y la transición mesenquima-epitelio (MET), como son: ZEB-1/2, TWIST-1/2, SNAI-2 y TGFBR2. Además encontramos que una gran cantidad de los miRs de la red inferida a partir de datos de tumor mapean a una región cromosómica específica en el cluster *DLK1-DIO3* (Chr14q32); y algunos de estos miRs también se han relacionado con la regulación de la EMT/MET. El análisis de las vías asociadas a EMT y TGF-Beta refuerza el papel de miR-200 y los miRs del cluster *DLK1-DIO3* en las redes de miR-genes en cáncer de mama. Dados los resultados obtenidos, incluir las relaciones de los genes y los miR que han sido asociados por medio de este enfoque basado en redes ha demostrado ser relevante para entender los mecanismos regulativos relacionados a la biología del cáncer de mama.

Análisis de la Estructura de la red de Regulación por ncRNA en Pacientes con Cáncer de Mama

by

Diana Drago-García

Abstract

Breast cancer is the most frequent cancer among women, and the second most common cancer in the world. Over the last years, microRNAs (miRs) have shown to be crucial for breast tumour establishment and progression. To understand the influence that miRs have over transcriptional regulation, here we constructed mutual information networks from 86 TCGA matched breast cancer and control tissue RNA-Seq and miRNA-Seq sequencing data. We show that miRs determine the structure of the networks inferred from tumour and control data. In tumour network, miR-200, miR-199 and neighbour miRs seem to cooperate in the epithelial-to-mesenchymal transition (EMT) and mesenchymal-to-epithelial transition (MET) regulation. Despite structural differences between tumour and control networks, a common core appears. The core expression signature, composed by miR-200 family members and genes such as VIM, ZEB-1/2 and TWIST-1/2 is particularly related to MET. Further, a large amount of miRs observed in tumour network mapped to a specific chromosomal location in *DLK1-DIO3* (Chr14q32); some of those miRs have also been associated with EMT/MET regulation. EMT and TGF-beta pathway analyses reinforce the relevance of miR-200 and *DLK1-DIO3* cluster in miR-mRNA networks. With this approach, we stress that miR inclusion in network construction would help to understand the regulatory mechanisms underlying breast cancer biology.

[Drago-García Diana, Espinal-Enríquez Jesús & Hernández-Lemus Enrique (submitted), Sci Rep]

Índice

I	Introducción	1
1.	Marco Teórico	2
1.1.	Cáncer de mama	2
1.2.	Alteraciones transcripcionales en cáncer de mama	4
1.3.	Los microRNAs y el cáncer de mama	6
1.4.	Tecnologías de secuenciación masiva	8
1.4.1.	RNA-Seq	9
1.4.2.	miR-Seq	9
1.5.	Redes de regulación genética como modelo para el estudio del cáncer	10
1.5.1.	Información Mutua	11
2.	Planteamiento del Problema	13
3.	Objetivos	15
3.1.	Objetivo general	15
3.2.	Objetivos particulares	15
II	Métodos	16
4.	Flujo de trabajo	17
4.1.	Obtención de los datos de expresión	19
4.2.	Preprocesamiento de los datos	20
4.2.1.	Datos de miR-Seq	20
4.2.2.	Datos de RNA-Seq	21
4.2.3.	Matrices de expresión	22
4.3.	Construcción de la Red	22

4.4. Análisis de la Red	25
4.5. Análisis de expresión diferencial	25
4.6. Análisis funcional	25
4.6.1. Análisis de enriquecimiento: Ontología de genes	25
4.6.2. Análisis de vías: Pathifier	26
4.6.3. Interacciones validadas y predichas: TargetScan y miRTarBase	28
III Resultados	29
5. Propiedades estructurales y funcionales de las redes	30
5.1. Los valores de MI de las redes poseen distintas distribuciones entre fenotipos	31
5.2. Propiedades topológicas de las redes	32
5.2.1. Los miRs mantienen la cohesión de las redes	37
5.2.2. Las redes muestran enriquecimiento diferencial entre fenotipos	37
5.3. miR-200 y miR-199 definen la estructura de la red	39
5.3.1. Los nodos de alto grado pertenecen a las familias miR-200 y miR-199	39
5.3.2. miR-200 es importante para la estructura de las redes independientemente de su fenotipo	40
5.3.3. Las redes de primeros vecinos de miR-200 muestran un centro común relacionado a EMT/MET	43
5.3.4. El comportamiento de miR-199 es determinante para la estructura de la red de datos de tumor	44
5.4. Los miRs de la red de datos de tumor muestran una tendencia a localizarse en el cluster <i>DLK1-DIO3</i>	47
5.5. Análisis de vías: miRs y la deregulación de las vías asociadas a EMT	50
5.6. La función de los miRs es consistente con su participación en la red de datos de tumor	56
5.7. Las asociaciones de los miRs y genes son consistentes con interacciones en TargetScan y miRTarBase	57

IV	Discusión	61
V	Conclusiones	68
VI	Apéndices	71
A.	Tablas con las propiedades de las redes con variaciones en su construcción	72
B.	Tablas con nodos con mayor grado para las redes con variaciones en su construcción	75
C.	Mapas de calor de los análisis de deregulación de vías para las vías de señalización asociadas a las redes de miR-200 y <i>DLK1-DIO3</i>	78

Parte I

Introducción

Capítulo 1

Marco Teórico

1.1. Cáncer de mama

De acuerdo a la Agencia Internacional para la Investigación del Cáncer (IARC) el cáncer de mama es la principal causa de muerte relacionada al cáncer en mujeres [Ferlay *et al.*, 2015], también en México el cáncer de mama representa la primera causa de muerte por tumores malignos en mujeres [Knaul *et al.*, 2009]. Debido a su elevada mortalidad e incidencia tanto a nivel nacional como internacional esta enfermedad se ha convertido en un problema de salud pública, por lo que se ha hecho evidente la necesidad de mejorar nuestra comprensión de los mecanismos moleculares que promueven la aparición de esta enfermedad. Un mejor entendimiento de la biología del cáncer de mama permitirá mejorar los tratamientos disponibles que se aplican actualmente en la clínica y desarrollar técnicas de diagnóstico que sean más sensibles y menos invasivas.

El cáncer de mama es una de las enfermedades que colectivamente se conocen como **cáncer**, y se caracterizan por compartir una serie de características comunes que promueven el desarrollo, mantenimiento y diseminación tumoral. Estas se conocen como Características del Cáncer o “Hallmarks of Cancer” [Hanahan y Weinberg, 2011], y representan aspectos clave de la biología del tumor. Debido a la gran heterogeneidad entre los distintos tipos de cáncer, el establecimiento de estas características generales revolucionó la manera en la que se percibe y se estudia esta enfermedad.



Figura 1-1: Características del cáncer. Entre las características representativas del cáncer se encuentran la resistencia a la muerte celular, la inducción de angiogénesis, la inmortalidad replicativa habilitada, la evasión de los supresores del crecimiento, la señalización proliferativa sostenida, la invasión activa y metástasis, la deregulación de la energética celular, la inestabilidad genómica y mutaciones, la evasión inmune y la inflamación promotora de tumores (Adaptado de Hanahan y Weinberg [2011]).

Las características descritas por Hanahan y Weinberg [2000] incluían originalmente una serie de capacidades comunes de las células cancerosas; las cuales incluyen a la resistencia a la muerte celular, la inducción de angiogénesis, la inmortalidad replicativa, la evasión de los supresores del crecimiento, la señalización proliferativa sostenida, así como la invasión activa y metástasis (Figura 1-1). Sin embargo, el trabajo de Hanahan y Weinberg [2000] también plantea que la estimulación entre el entorno celular o microambiente y las células cancerosas es vital para el establecimiento, desarrollo y diseminación del tumor; Aunque este microambiente se compone de células no cancerosas, su presencia y actividad favorecen a los tumores malignos.

Posteriormente se adicionaron a estas características la deregulación de la energética celular, la inestabilidad genómica y mutaciones, la evasión inmune y la inflamación promotora de tumores [Hanahan y Weinberg, 2011] (Figura 1-1). Es interesante que éstas últimas características se relacionan a respuestas sistémicas como la inmunidad e inflamación, exponiendo así que la biología del cáncer envuelve anormalidades celulares a diferentes escalas. Estas anormalidades incluyen desde propie-

dades celulares como la susceptibilidad a mutaciones e inestabilidad genómica y modificaciones en la señalización y energética celular por estimulación del microambiente, hasta respuestas sistémicas como la promoción de la inflamación y la reprogramación de la respuesta inmune que permiten el establecimiento de tolerancia inmunológica.

En el caso particular del cáncer de mama, éste se caracteriza por poseer una gran heterogeneidad morfológica y molecular que se traduce en diferencias importantes en cuanto a su curso clínico y respuesta a tratamiento. Se ha descrito que distintos grupos de tumores de mama presentan las características descritas por Hanahan y Weinberg [2011] de manera diferencial. Algunos grupos de tumores de mama que muestran una mayor tasa proliferativa y mayor invasividad tienden a poseer un peor pronóstico, sin embargo, todos ellos poseen características comunes a como su alta capacidad metastásica [Dai *et al.*, 2016]. La relevancia de entender estos mecanismos, poniendo por ejemplo las características moleculares que promueven la metastásis, radica en que el 90% de las muertes relacionadas con cáncer se deben a los crecimientos metastásicos [Weigelt *et al.*, 2005].

Dentro de la heterogeneidad que caracteriza al cáncer de mama; pruebas histológicas, moleculares y epidemiológicas han identificado principalmente cuatro subtipos intrínsecos: Luminal A, Luminal B, Her2-enriquecido y Basal. Debido a que estos subtipos presentan fenotipos, comportamientos moleculares y respuesta a tratamiento similares; han demostrado tener una gran aplicación clínica. Aunque estos subtipos se determinan de manera cuantitativa por la expresión de un grupo de 50 genes [Parker *et al.*, 2009], cada uno de ellos se asocia a un grupo de tumores con criterios histopatológicos comunes; como la presencia de receptores de hormonas (estrógeno y progesterona), el receptor HER2, y un marcador de proliferación (Ki-76) [Goldhirsch *et al.*, 2011]. Estos subtipos también se caracterizan por presentar la expresión particular de grupos de genes [Cancer Genome Atlas Network and others, 2012]; por lo que entender la relación entre la expresión de diferentes genes en el cáncer de mama es de vital importancia para mejorar su tratamiento y diagnóstico.

1.2. Alteraciones transcripcionales en cáncer de mama

El estudio del cáncer como enfermedad ha resultado ser complejo. No sólo debido a los factores económicos, políticos y sociales asociados a la enfermedad, sino a que a pesar de estar categorizada como una enfermedad genética el genotipo específico para cada tipo de cáncer no ha podido ser identificado.

Basándose en su papel en el cáncer de acuerdo a su genética, los genes asociados al cáncer se han clasificado principalmente en promotores de tumores (oncogenes) y supresores de tumor. También se ha establecido que es necesario que al menos dos de estos genes sufran mutaciones para permitir la carcinogenesis; pero dado a que se han identificado más de 1000 genes asociados a cáncer en humanos, las combinaciones de alteraciones necesarias para producir genotipos cancerosos sobrepasaría el millón de posibilidades [Wishart, 2015]. Sin embargo, de acuerdo a Forbes *et al.* [2015] este número realmente no alcanza a representar el número real de variantes genéticas presentes en los pacientes que sufren esta enfermedad, el cual de acuerdo a los datos contenidos en el catálogo de mutaciones somáticas (COSMIC) sobrepasan los 60 millones [Wishart, 2015].

En los últimos años se ha determinado que gran parte de estos oncogenes y genes supresores de tumores están relacionados al metabolismo celular. Esta percepción del cáncer como un desorden metabólico se considera en Hanahan y Weinberg [2011] como la alteración de la energética celular, ya que es necesario modificar la actividad de las principales vías metabólicas biosintéticas de forma que promuevan el crecimiento y proliferación del tumor [Wishart, 2015]. Además se han identificado un grupo de metabolitos llamados “oncometabolitos” que tienen la capacidad de inducir de manera directa o indirecta cambios en la expresión de genes que favorecen el desarrollo del cáncer. Debido a que la presencia de muchos de estos metabolitos es específica de un tipo de tumor y la posibilidad de su determinación por medios no invasivos hay expectativa de su aplicación como posibles biomarcadores [Yang *et al.*, 2013].

Los oncometabolitos como biomarcadores están principalmente limitados por la capacidad de detección de los equipos analíticos disponibles, así como por la estabilidad del metabolito a determinar [Yang *et al.*, 2013]. Debido a que las alteraciones metabólicas y genéticas suelen actuar mediante modificaciones en la expresión de los transcritos relacionados, las alteraciones transcripcionales reflejan el efecto de múltiples estímulos sobre las células tumorales, aunque ni los mismos estímulos ni su contribución exacta este completamente determinada. En este respecto las alteraciones transcripcionales han resultado especialmente útiles para caracterizar neoplasias las cuales aparentemente no tienen ninguna anomalía genética, resaltando el papel decisivo de las anomalías epigenéticas y su relación con la expresión anormal de transcritos en el cáncer [Feinberg *et al.*, 2016].

En el caso específico de los tumores de mama, los cambios en la expresión de los transcritos ha demostrado estar relacionada con la biología y las características clínicas de los tumores permitiendo su agrupamiento en subtipos [Cancer Genome Atlas Network and others, 2012]. Además, los niveles

de expresión de los transcritos de los genes que presentan anomalías genéticas [Chin *et al.*, 2006] y epigenéticas [Jones, 2005] también guardan relación. Por lo que se podría esperar que por medio de la comparación de los niveles de expresión de los transcritos en el tejido de mama tumoral con respecto a un control adecuado (tejido mamario no tumoral) se pueda obtener información relevante en cuanto a los procesos celulares que están siendo modulados para favorecer esta neoplasia. Debido a que la estimulación que promueve el desarrollo y progresión tumoral proviene de múltiples fuentes con distintas contribuciones, el estudio de las alteraciones transcripcionales y los análisis de expresión génica se han convertido en herramientas poderosas para el estudio del cáncer.

1.3. Los microRNAs y el cáncer de mama

En fechas recientes se ha observado el papel del RNA no codificante (ncRNA) como regulador de la transcripción en procesos normales y patológicos. Más aún, el constante descubrimiento de nuevos ncRNAs sugiere que su influencia en los procesos celulares es aún poco conocida. Dentro de los ncRNAs, los microRNAs (miRs) han cobrado gran importancia debido a que se han visto relacionados a procesos celulares tales como: diferenciación celular, respuesta inmune, señalización de estrés, infecciones virales, y múltiples enfermedades como el cáncer. De manera específica los miRs se han asociado al desarrollo y capacidad metastásica en el cáncer de mama [Lal y O'Day, 2010].

Los miRs (Figura 1-2) son ncRNAs cortos (~ 19-25 nucleótidos). Estos son principalmente transcritos por la polimerasa II y la polimerasa III, por lo que en su forma de microRNA primario (pri-microRNA) se caracterizan por ser transcritos largos poliadenilados con capuchón en su extremo 5' [Lee *et al.*, 2004]. El *pri-microRNA* es procesado en el núcleo por el complejo proteico formado por la proteína *DiGeorge Syndrome Critical Region 8* y *Drosha* (RNAasa de tipo III), el cual también se conoce como el *complejo microprocesador*. El producto de dicho procesamiento es una estructura de tallo-asa a la cual se le conoce como precursor de microRNA ó *pre-microRNA* [Han *et al.*, 2004]. El pre-microRNA nuclear es entonces transportado al citoplasma por medio de la *exportina 5 dependiente de Ran-GTP* [Bohnsack *et al.*, 2004].

Una vez en el citoplasma el pre-microRNA es procesado una vez más por el complejo formado por la RNAasa de tipo III *Dicer*, y las proteínas de unión a RNA de doble cadena *PACT/TRBP*, que en conjunto con *Argonauta* (*Ago*) forman el *Complejo de Silenciamiento Inducido por RNA (RISC) de carga* o *RLC* [MacRae *et al.*, 2008; Winter *et al.*, 2009; Kim, 2005; Ha y Kim, 2014]. El complejo RLC se encarga de procesar al pre-microRNA hasta convertirlo en un miR inmaduro de doble cadena.

Entonces, después de que una de las hebras es cargada en Ago, la otra puede ser degradada por un proceso dependiente de la actividad RNAasa de Ago o removida por medio de helicasas y otras proteínas. Cuando una de las *hebras de miR* ha sido cargada en Ago se forma el complejo *RISC*, ahora este complejo puede efectuar su actividad regulatoria sobre la expresión de los mensajeros [Winter *et al.*, 2009].

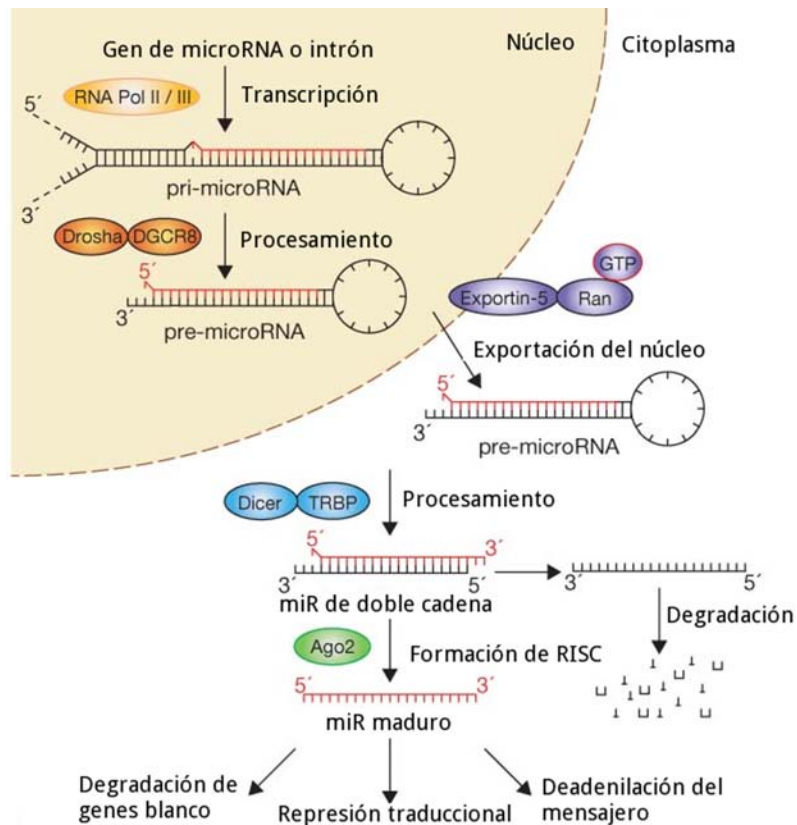


Figura 1-2: Vía de procesamiento de los miRs. La maduración canónica de los miRs incluye la producción de un transcrito primario (pri-microRNA) por la pol II/III y su procesamiento por el complejo microprocesador (Drosha-DGCR8) en el núcleo. La estructura resultante del procesamiento, llamado pre-microRNA es exportado desde el núcleo al citoplasma por la Exportina 5 dependiente de Ran-GTP. En el citoplasma el pre-microRNA es procesado de nuevo por la RNAsa Dicer en complejo con TRBP para producir un miR de doble de cadena con su longitud final. La hebra funcional de este miR es cargada en las proteínas Argonauta para formar el complejo de silenciamiento inducido por RNA (RISC). El complejo RISC entonces puede efectuar su actividad de regulación mediante la degradación de su RNA mensajero blanco (mRNA), su deadenilación o por medio de la represión traduccional (Adaptado de Winter *et al.* [2009]).

Existen diversos mecanismos por los que los miRs pueden efectuar su actividad regulatoria, estos van desde la degradación del blanco, el silenciamiento transcripcional, el silenciamiento traduccional ó incluso favorecer su estabilidad [Pasquinelli, 2012]. La evidencia apunta que la degradación del mRNA

por su desestabilización es la actividad de los miRs que impacta de manera más robusta los niveles de proteína de los genes blanco [Baek *et al.*, 2008], convirtiendo a las tecnologías que determinan los niveles de transcritos de manera global especialmente útiles para el análisis de la regulación por miRs.

Los microRNAs se han asociado principalmente con la represión de blancos por medio de su interacción directa, sin embargo los blancos indirectos han demostrado ser igualmente relevantes [Selbach *et al.*, 2008]. Además las interacciones de los miRs, directas o indirectas, pueden favorecer no sólo la regulación negativa sino también la regulación positiva de sus genes asociados [Vasudevan *et al.*, 2007]. El efecto que los miRs ejercen sobre la expresión de los genes es determinante para el fenotipo, debido a su interacción con otros elementos regulatorios como factores de transcripción y RNAs no codificantes [Chen *et al.*, 2011; Diaz *et al.*, 2015], jugando un papel central en mecanismos que aseguran la robustez biológica [Vidigal y Ventura, 2015].

La regulación transcripcional por parte de los miRs se ha asociado con diversos procesos celulares tanto fisiológicos como patológicos [Garofalo y Croce, 2011]. Incluso se ha observado que algunos miRs conocidos como “oncomiRs” juegan un papel decisivo en la promoción y progresión del cáncer, ya que tienen la capacidad de actuar como oncogenes o genes supresores de tumores [Cho, 2007; Esquela-Kerscher y Slack, 2006]. Algunos de estos miRs parecen tener actividades específicas en distintas neoplasias, como en el caso del cáncer de mama [Lal y O'Day, 2010]. De ahí la importancia de incluir estos elementos regulatorios en un análisis que pretenda comprender la biología de esta enfermedad.

1.4. Tecnologías de secuenciación masiva

La secuenciación del genoma humano significó en sus inicios un enorme esfuerzo económico y tecnológico para el sector público y privado [International Human Genome Sequencing Consortium and others, 2004]. Ahora gracias a los avances en las tecnologías actuales el costo y el tiempo requerido para realizar un análisis de secuenciación a nivel de genoma completo es significativamente menor, permitiendo su aplicación en múltiples protocolos de investigación.

La *secuenciación de alto rendimiento* ó *secuenciación de nueva generación* (NGS por sus siglas en inglés) es una técnica de secuenciación masiva que permite caracterizar el genoma completo de un organismo a un costo y tiempo menor al de todas las tecnologías precedentes [Reuter *et al.*, 2015]. La NGS se basa de manera general en la preparación de la secuencia genómica, amplificación clonal

y rondas de secuenciación masiva en paralelo [Reuter *et al.*, 2015].

Comparando la NGS con tecnologías similares, como los micro-arreglos, las plataformas de NGS han demostrado tener un mejor desempeño mostrando una mayor precisión, resolución y sensibilidad [Wang *et al.*, 2009]. Aunque ambas tecnologías están sujetas a distintos sesgos que han de ser considerados durante el análisis de los datos. Gracias a la robustez que ofrece, junto con la disminución de su costo y accesibilidad para las instituciones de investigación, se han podido desarrollar una gran cantidad de aplicaciones para la NGS [Reuter *et al.*, 2015]. Entre las más conocidas se encuentran el análisis transcriptómico RNA-Seq [Nagalakshmi *et al.*, 2008], el análisis de conformación de cromatina Hi-C [Lieberman-Aiden *et al.*, 2009], y la secuenciación de RNA por inmunoprecipitación (RIP-Seq) [Sephton *et al.*, 2010].

1.4.1. RNA-Seq

La identificación de ciertas regiones genómicas como las regiones no codificantes, los intrones y otros transcritos con actividades regulatorias siempre ha representado un desafío. Para abordar este problema Nagalakshmi *et al.* [2008] desarrollaron una derivación de la NGS conocida como **RNA-Seq** que permite cuantificar y secuenciar el transcriptoma completo de un tejido o tipo celular.

El RNA-Seq requiere en un principio la obtención del RNA a analizar, este puede ser una biblioteca de RNA total o alguna fracción enriquecida (RNAs poliadenilados); dicha biblioteca es entonces convertida a una biblioteca de cDNA con la ligación de adaptadores a ambos extremos [Wang *et al.*, 2009]. Estos adaptadores son requeridos para los consecuentes ciclos de amplificación y secuenciación típicos de las plataformas de secuenciación. Posteriormente las lecturas que se obtienen de la secuenciación (que generalmente poseen un tamaño de entre los 30-400 pares de bases) son alineadas con un genoma de referencia produciendo un mapa transcriptómico de la estructura y abundancia de los transcritos se que expresan en un tiempo y condiciones determinadas.

1.4.2. miR-Seq

Los RNAs pequeños han demostrado jugar un papel central en la regulación de una gran cantidad de procesos biológicos y enfermedades, incluyendo al cáncer [Lal y O'Day, 2010]. Para estudiarlos se utilizan protocolos modificados de RNA-Seq que tienen como finalidad la secuenciación de RNAs pequeños [Tam *et al.*, 2015]. Esta no es una tarea sencilla, ya que se deben de utilizar protocolos experimentales específicos para obtener una biblioteca enriquecida en la fracción de RNAs pequeños

que tienen una longitud menor a los 200 pares de bases y cuya abundancia relativa con otras especies de RNA normalmente es menor al 1 % en las células humanas [Palazzo y Lee, 2015]. Debido a su pequeño tamaño y baja abundancia, se requieren protocolos específicos para la preparación de las bibliotecas e incluso la interpretación de los datos requiere consideraciones especiales [Tam *et al.*, 2015].

1.5. Redes de regulación genética como modelo para el estudio del cáncer

En los últimos años se ha generado una gran cantidad de información biológica gracias a las tecnologías de alto rendimiento. El procesamiento de grandes cantidades de muestras mediante experimentos que determinan la expresión de los genes de manera masiva como el RNA-Seq, hace necesario utilizar métodos eficientes que sean capaces de integrar la enorme cantidad de datos obtenidos para hacer frente a los problemas de interés biológico.

La comprensión de un proceso tan intrincado como la regulación de los mensajeros y la identificación del papel que juegan los miRs en ella, requiere de una gran cantidad de información, así como la capacidad de procesarla e integrarla de un modo coherente, por lo que la aplicación de herramientas matemáticas y computacionales para el análisis de dicha información se ha vuelto no únicamente útil, sino necesaria. Dentro de dichas herramientas, aquellos métodos que buscan inferir las interacciones regulatorias entre los genes a partir de datos experimentales por medio de herramientas computacionales se conocen como de algoritmos de ingeniería reversa [Bansal *et al.*, 2007]. Estos algoritmos han demostrado ser particularmente útiles para el modelado del paisaje transcripcional en las llamadas redes de regulación génica [Bansal *et al.*, 2007].

Una red de regulación génica se puede definir como una representación del estado celular basada en la teoría de redes, la cual permite definir las relaciones entre los genes de un grupo de células lo más homogéneo posible [Hernández-Lemus y Rangel-Escareño, 2011]. En las redes de regulación génica las interacciones entre los genes pueden representar tanto interacciones físicas (directas) o interacciones regulatorias (indirectas) por medio de proteínas, metabolitos o RNAs no codificantes [Bansal *et al.*, 2007]. Algunos algoritmos de inferencia de redes permiten determinar la dirección e incluso la fuerza de las interacciones produciendo redes dirigidas; cuando la dirección de la interacción no puede ser especificada por el algoritmo se producen redes no dirigidas [Bansal *et al.*, 2007].

Como se ha descrito antes, el desarrollo de una neoplasia depende de múltiples variables no lineales, incluyendo factores genéticos, metabólicos, epigenéticos e incluso ambientales. Modelar al cáncer como enfermedad representa un reto debido a la gran cantidad de variables involucradas y a la heterogeneidad propia de los tumores y sus manifestaciones clínicas. La capacidad de las redes para modelar fenómenos complejos permitiendo la integración de grandes cantidades de información las hace especialmente útiles para su aplicación en enfermedades como el cáncer.

Las redes como herramienta han permitido modelar distintos aspectos de la biología celular del cáncer, incluyendo la pérdida de regulación en las vías celulares, las consecuencias de las mutaciones oncogénicas, y el comportamiento del cáncer a nivel celular y tisular [Kreeger y Lauffenburger, 2010]. Entre las herramientas más comunes para inferir redes de regulación génica se encuentran los algoritmos de agrupamiento para inferir redes de coexpresión, los algoritmos basados en relaciones probabilísticas como las redes bayesianas y los abordajes basados en información mutua [Bansal *et al.*, 2007; Hernández-Lemus y Rangel-Escareño, 2011].

1.5.1. Información Mutua

La información mutua (MI) es una medida de la dependencia estadística basada en la teoría de la información, que permite seleccionar las variables representativas del fenómeno e inferir redes a partir de conjuntos de datos con relaciones no lineales y gran cantidad de ruido, como es el caso de los datos genómicos [Hernández-Lemus y Rangel-Escareño, 2011].

Como medida de la teoría de la información, la aplicación de la MI para relacionar los perfiles de expresión entre parejas de genes tiene ciertas implicaciones. La entropía de *Shannon-Weaver* es máxima para las variables uniformemente distribuidas, por lo tanto es una propiedad que se relaciona con la capacidad predictiva de una distribución. Esta medida se puede aplicar de forma condicional para medir la incertidumbre de un par de variables cuando una de ellas es conocida, por lo tanto la entropía condicional entre un par de variables es máxima cuando sus distribuciones son estadísticamente independientes [Hernández-Lemus y Rangel-Escareño, 2011]. Bajo esta definición, la entropía de un conjunto de genes aumentaría en la medida que su expresión se distribuyera de manera aleatoria en el experimento. En cambio, la reducción en la entropía condicional de la distribución conjunta de una variable con respecto a su entropía marginal es lo que se conoce como MI. Debido a la relación que existe entre MI y la entropía de Shannon como medida de la dependencia estadística, el MI entre los valores de expresión de una pareja de genes es igual a cero sólo si son estadísticamente independien-

tes, y diferente de cero cuando hay una asociación entre ambos genes; lo que significa que el conocer el perfil de expresión de uno de de ellos nos permite obtener información que no podríamos conocer a partir de sus perfiles individuales [Bansal *et al.*, 2007]. Debido a que la MI es simétrica, las redes inferidas por medio de ésta medida son no dirigidas.

Capítulo 2

Planteamiento del Problema

El cáncer de mama es el cáncer más común entre las mujeres, y el segundo cáncer más común en el mundo [Ferlay *et al.*, 2015]. La alta incidencia, mortalidad y heterogeneidad clínica muestra la importancia de comprender a mayor profundidad los mecanismos relacionados al desarrollo del cáncer de mama. A través de los años, con la introducción de las tecnologías de secuenciación de nueva generación, un grupo de RNAs pequeños no codificantes conocidos como miRs han demostrado ser cruciales para el establecimiento y progresión de los tumores cancerosos [Lal y O'Day, 2010]. La actividad de los miRs se ha asociado con la regulación transcripcional de la homeostasis celular y otros mecanismos celulares relevantes en el cáncer, entre los que se encuentran: la apoptosis, proliferación y migración [Garofalo y Croce, 2011].

Existe evidencia de que la actividad regulativa de los miRs impacta de manera más robusta sobre los niveles de proteína por medio de la desestabilización de los mensajeros [Baek *et al.*, 2008], convirtiendo a las tecnologías que permiten caracterizar y cuantificar el transcriptoma completo especialmente útiles para estudiar estas interacciones. Los mecanismos de regulación por miRs sean directos o indirectos pueden favorecer tanto la expresión como la represión de sus posibles blancos [Selbach *et al.*, 2008], demostrando ser importantes para asegurar la robustez biológica [Vasudevan *et al.*, 2007] y ser determinantes para el fenotipo [Vidigal y Ventura, 2015].

Grandes esfuerzos para entender la biología del cáncer han promovido la creación de consorcios internacionales que organizan y dirigen colaboraciones para generar información sobre la genómica del cáncer. Uno de estos consorcios, The Cancer Genome Atlas (TCGA) se ha dado a la tarea de crear un atlas de perfiles genómicos a gran escala del cáncer; para así mejorar el diagnóstico, tratamiento y

promover la prevención de esta enfermedad [Tomczak *et al.*, 2015].

Para entender la relación transcripcional entre los miRs y genes (RNAs mensajeros) en el presente trabajo planteamos la construcción de redes basadas en MI con muestras de cáncer de mama primario y controles pareados con datos de secuenciación de RNA-Seq y miR-Seq de TCGA. Utilizando análisis funcionales y de expresión diferencial para complementar la información obtenida en las redes, integrándola de manera que nos permita comprender los mecanismos de regulación entre los miR y los genes en cáncer de mama.

Capítulo 3

Objetivos

3.1. Objetivo general

- Entender la regulación transcripcional entre los microRNAs y los RNAs mensajeros en cáncer de mama.

3.2. Objetivos particulares

- Construir una red de regulación transcripcional de cáncer de mama a partir de datos de secuenciación de RNAs mensajeros (RNA-Seq).
- Construir una red de regulación transcripcional de cáncer de mama a partir de datos de secuenciación de miRs (microRNA-Seq).
- Integrar la información de ambas redes.
- Realizar contrastes de la expresión de RNAs mensajeros y microRNAs del tejido control y el tejido tumoral.
- Identificar las vías de señalización más relevantes en el contexto del cáncer de mama, y la participación de los microRNAs en su regulación.

Parte II

Métodos

Capítulo 4

Flujo de trabajo

Todo el código y la implementación de las herramientas utilizadas para desarrollar los resultados presentados en este trabajo se encuentran disponibles en línea: https://github.com/CSB-IG/miRseq_rnw. El análisis computacional se realizó principalmente con el lenguaje programación R (v.3.2.0).

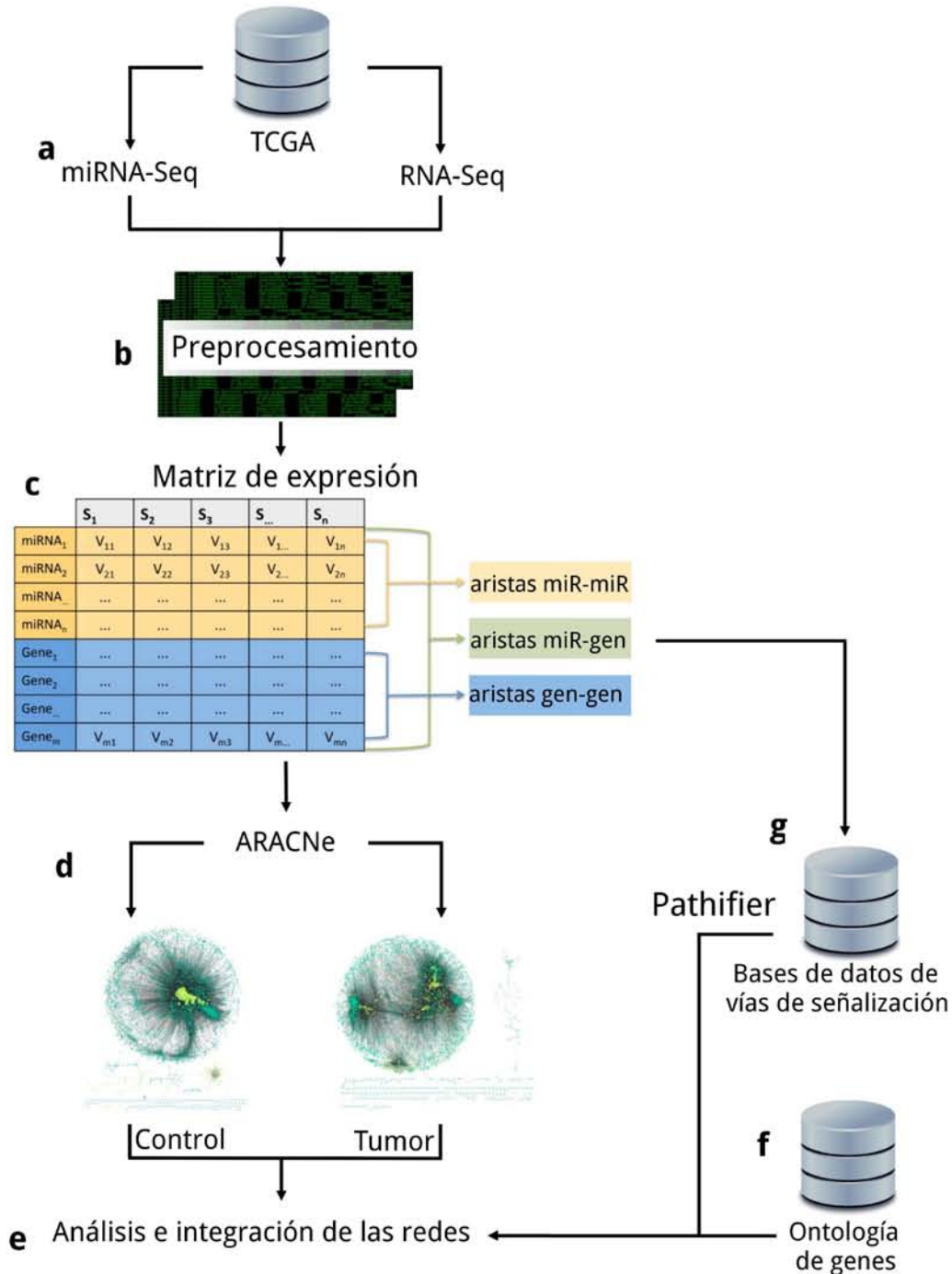


Figura 4-1: Diagrama metodológico. a) Obtención de datos de expresión de miRs y RNAs mensajeros por secuenciación de pacientes disponibles en TCGA. b) Preprocesamiento de los datos. c) Obtención de matrices de expresión normalizadas de miRs y RNAs mensajeros. d) Generación de redes basadas en Información Mutua a partir de las matrices de expresión por medio del algoritmo ARACNe. e) Análisis y visualización de las redes por medio de software especializado. f) Enriquecimiento de los genes presentes en las redes para casos y controles. g) Análisis de vías asociadas a las redes por medio de Pathifier.

4.1. Obtención de los datos de expresión

Uno de los principales problemas de trabajar con datos disponibles públicamente es el integrar información de múltiples fuentes, además de las diferencias en las metodologías experimentales y el análisis de los datos. Para evitar estos problemas, TCGA se ha dedicado a crear flujos de trabajo estandarizados para coleccionar, seleccionar y analizar muestras de tejidos humanos a gran escala [Tomczak *et al.*, 2015]. Además del procesamiento estandarizado, TCGA tiene la ventaja de que en su base de datos tiene pacientes con muestras pareadas tumor-control; lo que quiere decir es que para algunos pacientes además del análisis de su tejido tumoral también se encuentra disponible el análisis del tejido adyacente del mismo paciente.

De la base de datos de TCGA se obtuvieron los datos de expresión nivel 3 de los experimentos de secuenciación de mRNA (RNA-Seq, plataforma RNAseq Illumina Hiseq V2) y de secuenciación de miR (miR-Seq Illumina HiSeq 2000) de 86 pacientes (Figura 4-1 a).

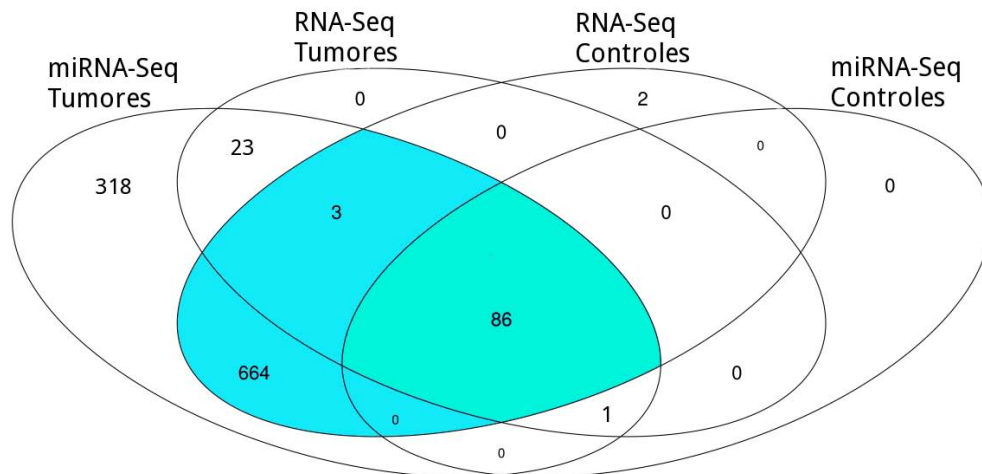


Figura 4-2: Datos de expresión de pacientes en TCGA. En este diagrama de Venn se representa el número de pacientes para los cuales se encuentran disponibles datos de RNA-Seq y miR-Seq para sus muestras de tumor y control contenidas en TCGA (Septiembre 2015), la intersección turquesa muestra el número de pacientes para los cuales se encuentran disponibles sus datos de expresión de miRs y genes de forma pareada para su tejido control y tumoral.

Seleccionamos sólo aquellos pacientes para los cuales estuvieran disponibles de manera pareada los datos de RNA-Seq y miR-Seq del tejido tumoral y tejido control, para así controlar la variabilidad entre las plataformas para las comparaciones entre tumores y controles (Figura 4-2).

4.2. Preprocesamiento de los datos

4.2.1. Datos de miR-Seq

Los datos de expresión de miRs se obtuvieron los conteos correspondientes a los miRs maduros de acuerdo a la metodología recomendada por el centro que realizó el análisis de los datos de secuenciación para TCGA (Canada's Michael Smith Genome Sciences Centre) en su "Documentación del Flujo de Trabajo para el Perfilado de miRs" (<https://github.com/bcgsc/mirna>). Para este análisis se utilizó la versión v.21 de la miRBase ([Kozomara y Griffiths-Jones, 2014; Griffiths-Jones *et al.*, 2008, 2006]) y la información de isoformas que provee TCGA.

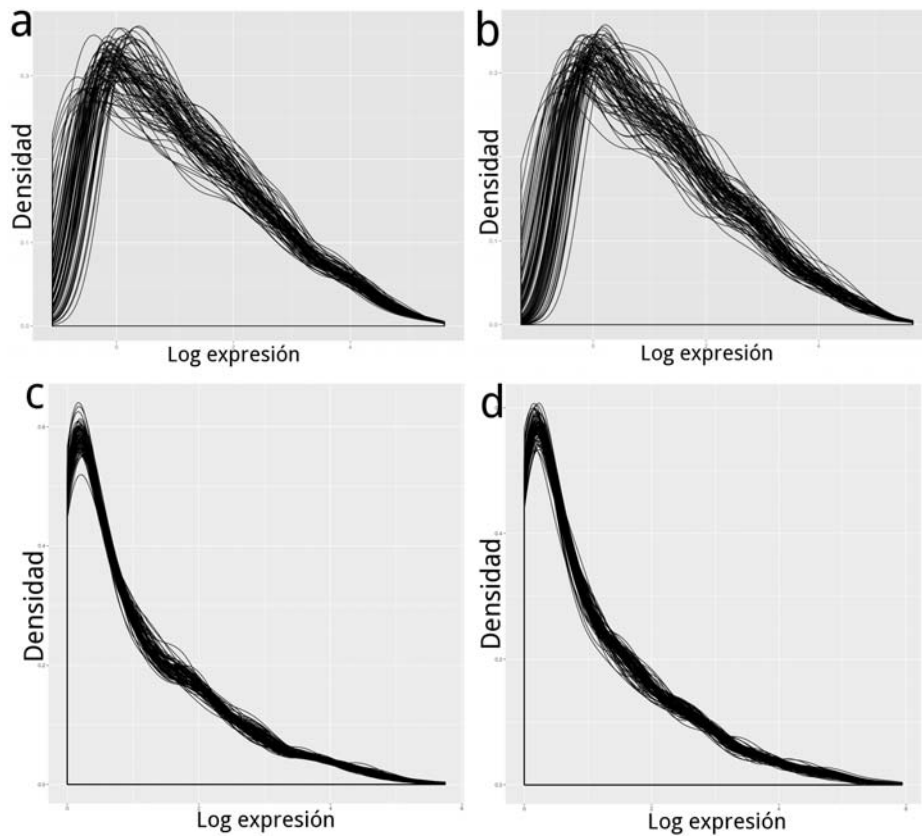


Figura 4-3: Diagrama de densidad de los datos de expresión de las 86 muestras de miR-Seq; sin normalizar para a) controles y b) tumores, y normalizando mediante el método de TMM para c) controles y d) tumores (Un diagrama de densidad nos permite visualizar la distribución de los datos sobre un intervalo continuo).

Una vez que los conteos de los miR maduros fueron calculados (Figura 4-1 b), todos los miRs con menos de 5 conteos en el 25% de las muestras fueron removidos del análisis [Tam *et al.*, 2015]. El conjunto resultante se normalizó por el método TMM ('Trimmed mean of M normalization') ([Robinson *et al.*, 2010b]) por medio del paquete EdgeR (v.3.12.0) ([Robinson *et al.*, 2010a]) de acuerdo a lo recomendado por Tam *et al.* [2015] (Figura 4-3). También se realizaron comparaciones con el método de 'Upper Quartile' (UQ), y su combinación con TMM pero no mejoraron la normalización de los datos por lo que fueron descartadas.

4.2.2. Datos de RNA-Seq

El consorcio de expertos de TCGA decidió incluir en su análisis de datos de RNA-Seq un par de herramientas computacionales diseñadas para optimizar el alineamiento y la cuantificación de las lecturas ambiguas. Esto debido a que las lecturas ambiguas son un problema común para los algoritmos de alineamiento debido a la gran cantidad de lecturas que mapean a posiciones múltiples en un genoma de referencia o conjunto de transcritos.

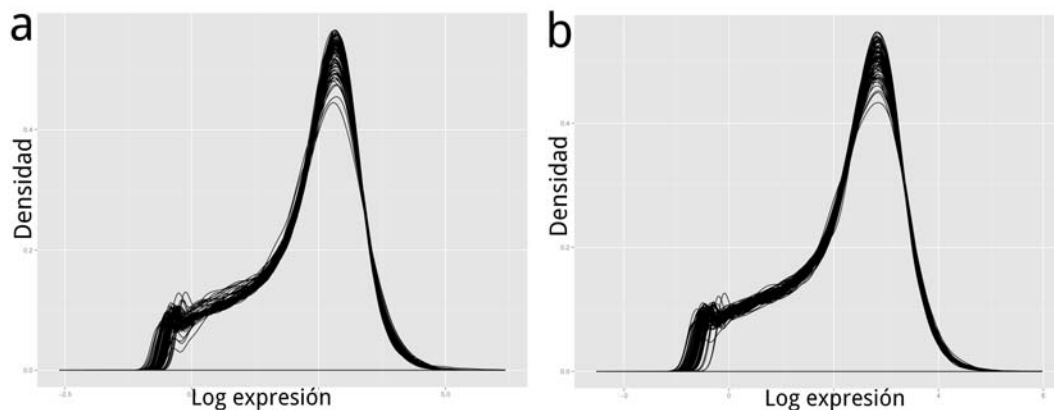


Figura 4-4: Diagrama de densidad de los datos de expresión de las 86 muestras de RNA-Seq; normalizadas por el método de Upper Quartile a) controles y b) tumores.

Los datos de RNA-Seq empleados en este trabajo son la segunda versión de su análisis y pertenecen al tercer nivel de TCGA lo que quiere decir que sólo se encuentran públicamente disponibles los datos de expresión terminalmente procesados obtenidos por la metodología que se describe a continuación. Estos datos fueron alineados por medio de la herramienta MapSplice [Wang *et al.*, 2010], y la cuantificación de las lecturas ambiguas se realizó por medio del algoritmo RSEM [Li y Dewey, 2011] (<https://wiki.nci.nih.gov/display/tcga/rnaseq+version+2>). MapSplice [Wang *et al.*,

2010] está diseñado para trabajar con base en las lecturas con la mejor calidad de alineamiento basado en las uniones de splicing. Por su parte, RSEM [Li y Dewey, 2011] usa un algoritmo de esperanza-maximización para realizar una estimación estadísticamente rigurosa por máxima verosimilitud de los niveles de expresión de los genes. Los estimados de los conteos por transcrito obtenidos por RSEM fueron normalizados por el método de UQ, tras determinar que era posible trabajar con ellos (Figura 4-4) todos aquellos genes con menos de 10 conteos por transcrito en promedio fueron filtrados del análisis (Figura 4-1 b).

4.2.3. Matrices de expresión

Una matriz de expresión es una matriz de datos que contiene la información de la expresión de los genes que se utilizaran el análisis para cada uno de los sujetos de estudio o muestras. En una matriz de expresión generalmente las columnas representan a las muestras (estas pueden ser muestras de tejido de diferentes pacientes, modelos animales o líneas celulares en cultivo), y las filas corresponden a los genes seleccionados.

Continuando con el análisis de los datos, una vez que nos aseguramos de que los datos se encontraban correctamente normalizados (Figura 4-1 b), se obtuvieron dos matrices de expresión por cada conjunto de datos (RNA-Seq y miR-Seq): una correspondiente a los datos del tejido tumoral de 86 pacientes, y otra correspondiente a los datos del tejido control perteneciente a los mismos 86 pacientes (Figura 4-1 c).

4.3. Construcción de la Red

Las redes obtenidas por ingeniería reversa se conocen como Redes de Regulación Génica representan a los transcritos y a sus asociaciones regulatorias como un grafo. Existen distintos métodos que buscan integrar la información de expresión de miRs y genes [Peng *et al.*, 2009; Huang *et al.*, 2011; Sales *et al.*, 2010; Hua *et al.*, 2013; Jung *et al.*, 2015; Andrews *et al.*, 2016; Le *et al.*, 2015] en este tipo de modelos, la mayoría utilizando medidas de correlación lineal [Peng *et al.*, 2009; Huang *et al.*, 2011] o modelos bayesianos [Huang *et al.*, 2007]. Sin embargo, considerando que la mayoría de las relaciones biológicas tienen una naturaleza no lineal, se han hecho esfuerzos para implementar el uso de algoritmos que utilicen medidas de la correlación no lineal para capturar las asociaciones regulatorias entre miRs y genes [Sales *et al.*, 2010; Hua *et al.*, 2013; Jung *et al.*, 2015].

ARACNe (Algorithm for the Reconstruction of Accurate Cellular Networks) [Margolin *et al.*, 2006a] es un algoritmo basado en la teoría de la información que al considerar a los RNAs (genes y miRNAs) como nodos y sus asociaciones como interacciones; calcula dichas asociaciones como la dependencia estadística o MI entre el perfil de expresión de una pareja de RNAs.

Es importante mencionar que las diferencias en rango dinámico entre los miR y los genes (mRNA) no son un impedimento para el análisis, ya que ARACNe utiliza un Kernel Gaussiano para la estimación de la densidad de probabilidad. Utilizar este tipo de estimador disminuye la influencia de las transformaciones arbitrarias de los datos y elimina la necesidad de determinar anchos de banda asociados con los datos uniformemente distribuidos [Margolin *et al.*, 2006a]. Sin embargo, aunque las diferencias en rango dinámico no son un problema, el efecto del ancho de kernel sobre la estimación de la información mutua podría serlo. Esto debido a que para un número de muestras finito no existe un ancho de kernel universal, por lo que Margolin *et al.* [2006a] diseñaron ARACNe para trabajar con los rangos de los valores de MI más que con la estimación exacta de los valores individuales de MI.

Por medio de una versión paralelizada del algoritmo ARACNe 2 [Margolin *et al.*, 2006a; Tovar *et al.*, 2015] y el paquete de R Minet (v.3.28.0) [Meyer *et al.*, 2008], a partir de las matrices de expresión de RNAseq y miRseq se calcularon las redes no dirigidas correspondientes a los datos de tejido tumoral y a los datos pareados de tejido control (Figura 4-1 d). Encontramos que los valores de MI de las interacciones entre miR y genes (miR-miR y miR-gen) tienen una distribución distinta a las asociaciones entre genes (gen-gen), y que las interacciones miR-miR y miR-gen tienden a poseer valores más bajos. Por lo que decidimos un valor de corte ligeramente menos restrictivo a las interacciones miR-miR y miR-gen en comparación con las interacciones gen-gen.

Una de las mayores limitaciones de inferir redes por medio de métodos estadísticos es el gran número de falsos positivos, por lo tanto es necesario aplicar un filtro que nos permita discernir entre las asociaciones relevantes de los artefactos de la metodología. En las redes basadas en MI, existe una relación entre el valor de MI, el número de muestras utilizado para la inferencia de la red y la significancia estadística de la red [Margolin *et al.*, 2006b]. En este caso este filtro principalmente consiste en filtrar las interacciones de la red que no sean significativas de acuerdo al p-valor corregido por el método de Bonferroni adecuado. Utilizamos un valor de corte de MI correspondiente a un p-valor corregido de 7.118841×10^{-26} para las interacciones gen-gen de la red inferida de los datos control, y un valor de MI correspondiente a un p-valor corregido de 1.565302×10^{-19} para las interacciones gen-gen de la red inferida de los datos de tumor. Para las interacciones miR-miR y miR-gen utilizamos un valor de

corte de MI correspondiente a un p-valor corregido de 1.120557×10^{-11} para las interacciones de la red inferida de los datos control, y un valor de MI correspondiente a un p-valor corregido de 0.008220069 para las interacciones de la red inferida de los datos de tumor. Estos p-valores junto con la aplicación de la Desigualdad del Procesamiento de Datos (DPI) con una tolerancia (τ) del 10% son considerados como apropiados para obtener una relación razonable entre el número de falsos positivos y falsos negativos, manteniendo la arquitectura de la red [Margolin *et al.*, 2006a]. El DPI es un teorema de la teoría de la información que permite eliminar asociaciones espurias, éste establece que la información que se transfiere de manera directa siempre es mayor que la que se transfiere de manera indirecta [Jang *et al.*, 2013]. El DPI identifica el menor valor de cada triplete de interacciones y lo descarta si dicho valor es menor a un valor de tolerancia (τ), el valor τ permite mantener estructuras de tripletes en la red en el caso de que las tres interacciones sean similarmente fuertes, y es independiente para cada triplete. Validaciones experimentales han demostrado que el DPI tiene la capacidad de eliminar una gran cantidad de falsos positivos, produciendo modelos de regulación altamente precisos [Margolin *et al.*, 2006b,a; Jang *et al.*, 2013].

Para probar la susceptibilidad de las redes inferidas a la elección del p-valor para filtrar sus interacciones, construimos redes adicionales utilizando variaciones de los valores originales. Los valores de corte originales asociados a los p-valores antes mencionados corresponden a 25,334 asociaciones miR-miR y miR-gen, y a 14,892 asociaciones gen-gen. Para construir las redes adicionales, mantuvimos uno de estos p-valores constante y ajustamos el otro para obtener el mismo número de interacciones. De manera que obtuvimos una red con 25,334 interacciones miR-miR y miR-gen, y 25,334 interacciones gen-gen; y una red con 14,892 interacciones miR-miR y miR-gen, y 14,892 interacciones gen-gen. También construimos redes aumentando y disminuyendo en un orden de magnitud la cantidad de interacciones original para analizar el efecto del corte del número de interacciones en las redes; obteniendo una red más pequeña con 2,533 asociaciones miR-miR y miR-gen, y 1,489 interacciones gen-gen, y una red más grande con 253,340 interacciones miR-miR y miR-gen, y 148,920 interacciones gen-gen.

La capacidad de determinar la dependencia estadística entre el perfil de expresión de una pareja de genes es especialmente útil para reconstruir las relaciones canónicas y no-canónicas entre los miR y los genes. Además tiene como ventaja que no es necesario asumir un comportamiento lineal entre los miR y los genes, ni utilizar criterios biológicos *a priori* ya que mucha de la información aún no se conoce. De ahí que la metodología propuesta en este trabajo cobra gran importancia para estudiar el papel de los miRs en el cáncer de mama.

4.4. Análisis de la Red

Para estudiar las propiedades topológicas y visualizar las redes utilizamos Cytoscape [Shannon *et al.*, 2003]. Al tratarse de redes no pesadas y no dirigidas, nuestro análisis se centró en la medida de centralidad del grado, enfatizando en las propiedades biológicas de los nodos más conectados y sus primeros vecinos (Nodos directamente conectados) (Figura 4-1 e). La mayoría de las visualizaciones se obtuvieron mediante el algoritmo de visualización “Spring Embedded”. La visualización “Hiveplot” de las redes fue preparada de acuerdo a lo recomendado por Krzywinski *et al.* [2012], para obtener una visualización optimizada de las redes para realizar comparaciones. El diagrama aluvial presentado se creó por medio de la herramienta web RAWGraphs (<http://rawgraphs.io/>).

4.5. Análisis de expresión diferencial

La expresión diferencial es una forma de representar las diferencias en la abundancia de un transcrito entre fenotipos. Existen distintas metodologías que se utilizan para cuantificar estas diferencias considerando el ruido que se introduce por la dispersión de los valores de los datos experimentales disponibles permitiendo su interpretación [Love *et al.*, 2014]. El análisis de expresión diferencial de los datos de RNA-Seq y miR-Seq se llevó a cabo usando el paquete de R DESeq2 (v.1.10.1) [Love *et al.*, 2014]. Los genes diferencialmente expresados fueron usados para identificar los nodos con un posible rol biológico importante, un miR o un gen se consideraban como diferencialmente expresados cuando su expresión cambiaba al menos dos veces en comparación al control y poseían un p-valor ajustado < 0.01 (FDR: Benjamini-Hochberg False Discovery Rate) .

4.6. Análisis funcional

4.6.1. Análisis de enriquecimiento: Ontología de genes

Para estudiar las implicaciones biológicas de la estructura y las propiedades de las redes inferidas, realizamos un análisis de enriquecimiento por sobre-representación. Utilizamos el plug-in de Cytoscape BiNGO [Maere *et al.*, 2005] (Figura 4-1 f) para analizar los genes que corresponden a mRNA en los nodos de nuestras redes; esta herramienta provee información relevante en cuanto a las ontologías de las categorías de proceso biológico, función molecular y componente celular. La categoría de proceso biológico se relaciona a las vías de señalización y los procesos celulares; función molecular se refiere a la actividad de los productos de los genes; y componente celular se refiere a la localización subcelular

donde los productos de los genes son activos. El análisis de sobre-representación utiliza la información de los genes de interés y las bases de datos que contienen a las ontologías para calcular, mediante una prueba hipergeométrica, la probabilidad de que ciertos genes pertenezcan a una determinada categoría. Decidimos centrarnos en aquellos procesos con mayor significancia estadística ($FDR < 0.01$) y que están constituidos por menos de 1,000 genes.

4.6.2. Análisis de vías: Pathifier

Pathifier [Drier *et al.*, 2013] es un algoritmo semi supervisado, que utiliza toda la información disponible de las mediciones experimentales de los niveles de expresión de los genes para evaluar la deregulación de una vía para una condición con respecto a un control [García-Campos *et al.*, 2015]. El algoritmo Pathifier realiza un análisis de componentes principales, evaluando la expresión de los genes pertenecientes a una determinada vía de señalización en un sistema de coordenadas creando una nube de puntos. Posteriormente, mediante el algoritmo de Hastie y Stuetzle [1989]) los puntos son usados para calcular una curva principal, usando a las muestras control para determinar el punto inicial del centroide. Finalmente, los puntos que corresponden a las muestras son proyectados a su punto más cercano en la curva principal asignando un valor de “Score de Deregulación de Vías” (PDS) entre 0 y 1, que corresponde a la distancia relativa desde la proyección de la muestra al punto inicial en el centroide [Drier *et al.*, 2013]. Lo que quiere decir que los valores de PDS cercanos a 0 representan valores de expresión similares al grupo control y los valores cercanos a 1 representan a las muestras cuya expresión muestra las mayores diferencias respecto al control. Utilizamos el algoritmo Pathifier (Figura 4-1 g) sobre las vías que cumplían con nuestros criterios.

Seleccionamos todas las vías de señalización contenidas en las bases de datos: WikiPathways, Reactome y KEGG; para entonces seleccionar aquellas vías que contuvieran al menos un gen presente en las redes de primeros vecinos de los miR de interés en la red de los datos de tumor. De las vías elegidas, filtramos todas aquellas que tuvieran menos de 4 genes y que tuvieran más genes que nuestro número de muestras.

Utilizamos los plug-ins de Cytoscape CytoKEGG, Reactome FI [Wu *et al.*, 2014] y WikiPathways [Kutmon *et al.*, 2014] para crear las visualizaciones de las vías elegidas. A estas visualizaciones les adicionamos las interacciones inferidas por medio de nuestras redes para estudiar la participación de los miR en las vías.

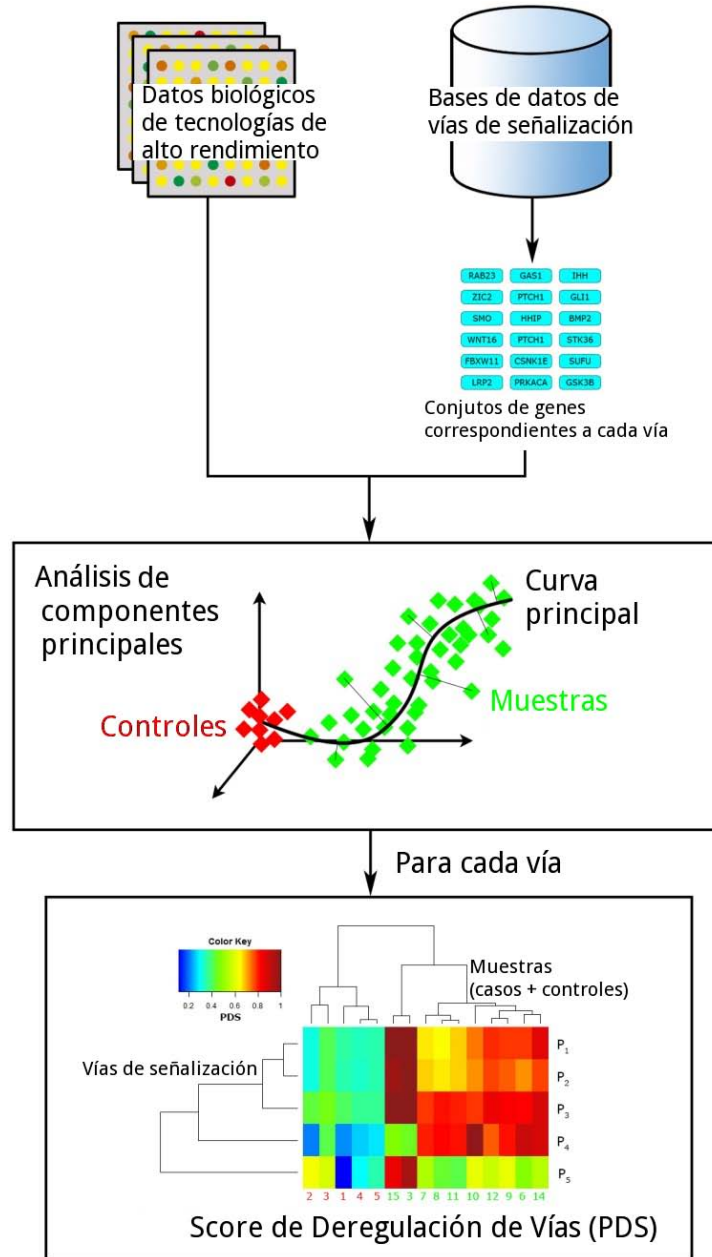


Figura 4-5: Pathifier [Drier *et al.*, 2013] utiliza información biológica de experimentos masivos e información de vías de señalización de interés para calcular un “Score de Deregulación de Vías” (PDS). Usando los datos biológicos y la información de las vías crea un espacio multidimensional donde para cada muestra los perfiles de expresión de los genes que pertenecen a una vía se representan como un punto, estos puntos son utilizados para calcular una curva principal. Posteriormente a cada muestra se le asigna un valor de PDS de acuerdo a la distancia relativa de los puntos hacia la curva. Esta información se representa en mapas de calor, ya que estas representaciones nos permiten identificar de manera visual las vías con mayor diferencia en sus PDS (Adaptado de García-Campos *et al.* [2015]).

4.6.3. Interacciones validadas y predichas: TargetScan y miRTarBase

Usamos la información disponible de las interacciones experimentalmente validadas presentes en la base de datos miRTarBase y las predicciones de asociación entre miR-blanco de la base de datos de targetScan para explorar las interacciones comunes presentes en estas bases de datos y nuestras redes inferidas a partir de datos experimentales usando Cytoscape [Shannon *et al.*, 2003].

Parte III

Resultados

Capítulo 5

Propiedades estructurales y funcionales de las redes

Definimos los nodos de las redes como genes y miRs (miR maduro); para la mayoría de los análisis los miR están agrupados como familias de miR. Las interacciones de las redes inferidas se definen como la MI entre el perfil de expresión de un par de nodos, resultando en una red no dirigida con 3 tipos de interacciones: entre miRs (miR-miR), entre miRs y genes (miR-gen), y entre genes (gen-gen).

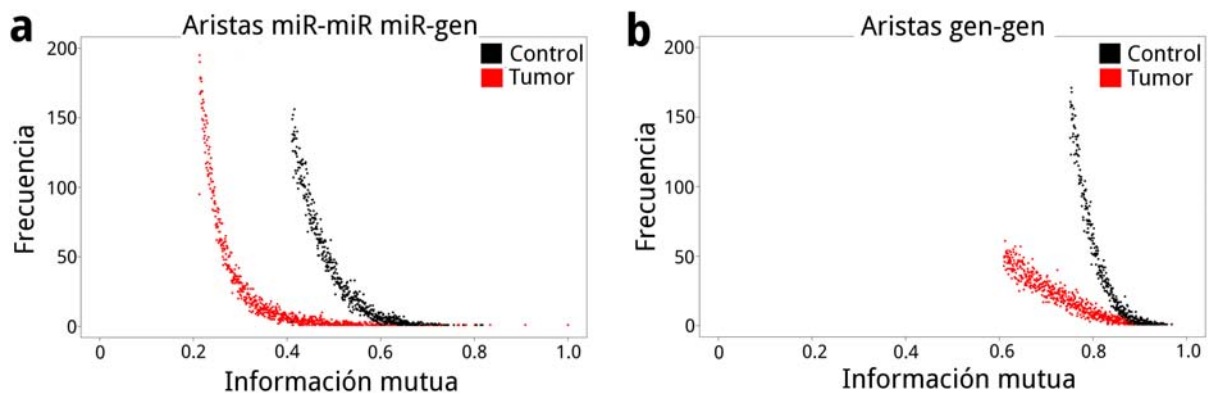


Figura 5-1: Histograma de MI de las interacciones de las redes; a) interacciones miR-miR y miR-gen del componente gigante de la red de datos de tumor (rojo), y de la red de datos control (negro) (representando el 0.259% de las interacciones más fuertes); b) interacciones gen-gen del componente gigante de la red de datos de tumor (rojo), y de la red de datos control (negro) (representando el 0.013% de las interacciones más fuertes).

5.1. Los valores de MI de las redes poseen distintas distribuciones entre fenotipos

Por medio del algoritmo ARACNe [Margolin *et al.*, 2006a] a partir de los datos de expresión pareados de tejido tumoral de mama y tejido adyacente control, construimos redes basadas en información mutua. Decidimos usar distintos puntos de corte dependiendo de la naturaleza de la interacción (miR-miR, miR-gen o gen-gen) (Ver Métodos), debido a que el tipo de interacción se refleja de manera cuantitativa en la distribución de los valores de MI de sus interacciones (Figura 5-1).

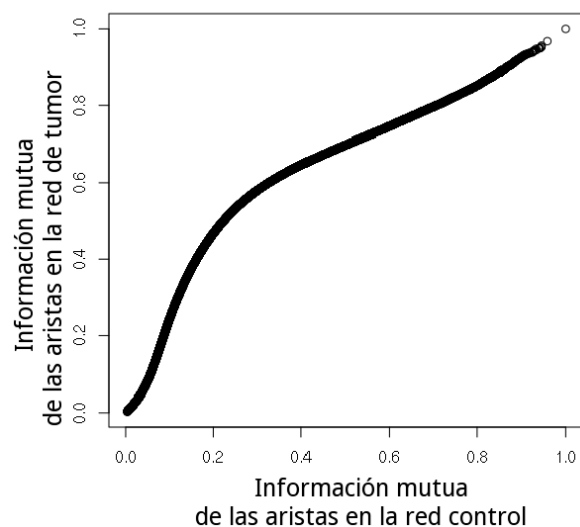


Figura 5-2: Representación q-q plot, esta figura representa una comparación entre las distribuciones de las interacciones entre las redes de datos control y de tumor. En esta figura podemos observar que los valores de MI de las interacciones de los datos de tumor son menores a valores bajos de MI, pero casi no presentan diferencias con respecto a la distribución de MI de los controles a valores altos de MI.

Encontramos que la distribución de MI de las interacciones es diferente entre los fenotipos tumoral y control. En la Figura 5-1 se pueden observar las distribuciones de las aristas de las asociaciones miR-miR y miR-gen (Figura 5-1 a), y las asociaciones gen-gen (Figura 5-1 b) para los datos control (negro) y de tumor (rojo). La distribución de las interacciones con valores de MI más bajos está sesgada hacia la izquierda en el caso de la red de tumor. En cambio, al mismo corte, los valores de MI de las interacciones de la red control tienden a ser más altos. Estas diferencias se mantienen en el conjunto completo de interacciones en las redes, como puede observarse en la Figura 5-2.

5.2. Propiedades topológicas de las redes

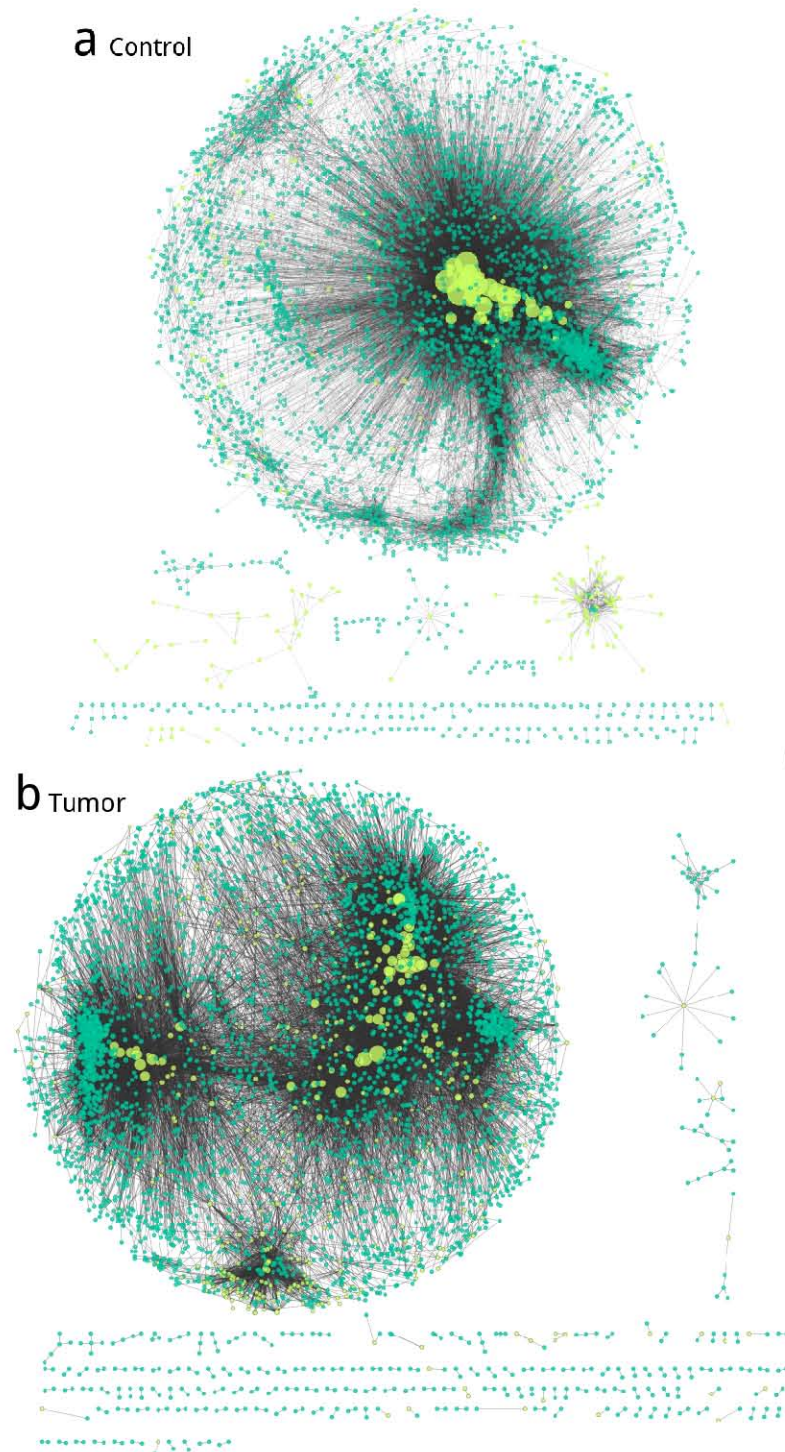


Figura 5-3: Representación de la red completa inferida por medio de los datos de expresión de las muestras a) control y de b) tumor. Los nodos verde claro representan a los miR, y los nodos turquesa representan a los genes, el tamaño del nodo es proporcional a su número de conexiones; las interacciones se representan de color gris.

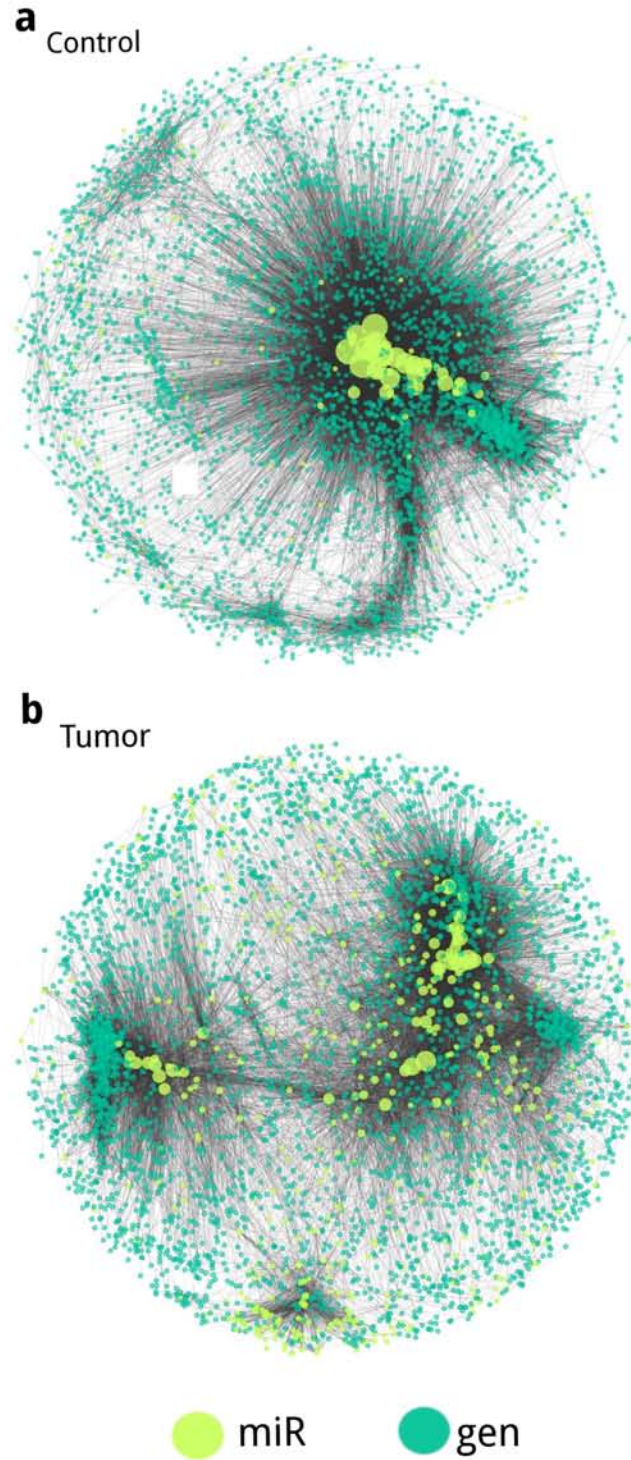


Figura 5-4: Representación de los Componentes Gigantes de las redes inferidas a partir de los datos control a) y de tumor b). Para estas redes el color de los nodos representa su tipo, los nodos turquesa representan a los genes y los nodos verde claro a los miR, además el tamaño de los nodos es proporcional a su número de conexiones; las interacciones se representan de color gris.

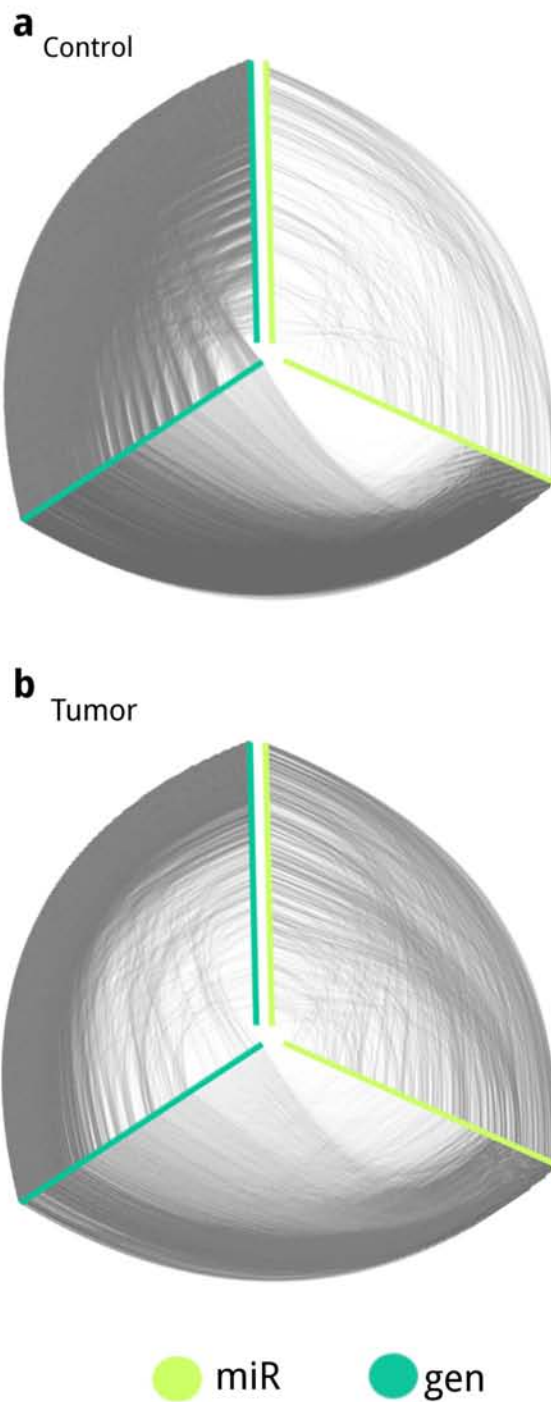


Figura 5-5: Representación Hiveplot de las redes de datos control a) y de tumor b). En esta representación bidimensional de la red tenemos a los nodos acomodados de forma creciente de acuerdo a grado en los ejes, y las interacciones entre los nodos se representan como aristas entre los ejes. Los ejes turquesa contienen a los nodos que corresponden a genes, y los ejes verde claro contienen a los nodos que corresponden a miRs; las interacciones se representan de color gris. Como puede observarse en la figura, la red de tumor posee una mayor cantidad de interacciones miR-miR y una menor cantidad de interacciones miR-gen y gen-gen en comparación con la red control.

Analizamos la contribución de los miR y los genes sobre los nodos y las interacciones de las redes, los parámetros resultantes están descritos en la Tabla 1. La visualización de las redes completas puede encontrarse en la Figura 5-3. Decidimos enfocarnos en el Componente gigante (CG) para cada red (Figura 5-4) debido a que como puede observarse en la Tabla 1 este componente contiene casi todos los nodos e interacciones de la red.

Los componentes gigantes de las redes inferidas a partir de los datos control (Figura 5-4a, 5-5a) y de tumor (Figura 5-4b, 5-5b) poseen una cantidad similar de nodos, aunque la cantidad de nodos que corresponden a miRs en la red de tumor es casi cuatro veces mayor en comparación con la red control. Además como se observa en la representación de Hiveplot de la Figura 5-5, hay menos interacciones de asociaciones miR-gen y gen-gen en la red de datos de tumor (Figura 5-5 b), además de que la cantidad de interacciones de las asociaciones miR-miR es más de siete veces mayor en comparación con la red control (Figura 5-5 a, Tabla 1).

Tabla 1. Atributos de las redes inferidas a partir de los datos de tumor y los datos control.

Atributo	Red completa		Componente gigante		Subred CG gen-gen	
	Control	Tumor	Control	Tumor	Control	Tumor
Nodos totales	4,575	4,602	4,229	4,200	4,096	3,714
miR	241	514	133	486	0	0
gen	4,334	4,088	4,096	3,714	4,096	3,714
interacciones totales	33,879	29,186	33,388	28,913	14,486	11,010
miR-miR	482	1,775	240	1,769	0	0
miR-gen	18,760	16,173	18,662	16,134	0	0
gen-gen	14,637	11,238	14,486	11,010	14,486	11,010
Componentes	102	165	1	1	1,856	2,590
Nodos solos	0	0	0	0	1,814	2,524

CG: Componente gigante

Para probar la susceptibilidad del método a nuestra elección de punto de corte para mantener las asociaciones más significativas creamos una serie de redes que poseen cantidades de interacciones miRs y genes distintos a los propuestos originalmente, además de variaciones entre las proporciones de los mismos. Encontramos que las redes construidas partiendo de la misma cantidad inicial de miR o mRNA, o utilizando una cantidad menor o mayor del número de interacciones originales (por lo menos en al menos dos órdenes de magnitud) no presentan cambios en cuanto a sus atributos. Como puede

observarse en las Tablas del Apéndice A: A1-A4, las redes construidas mantienen una cantidad menor de miRs en comparación con la cantidad de genes.

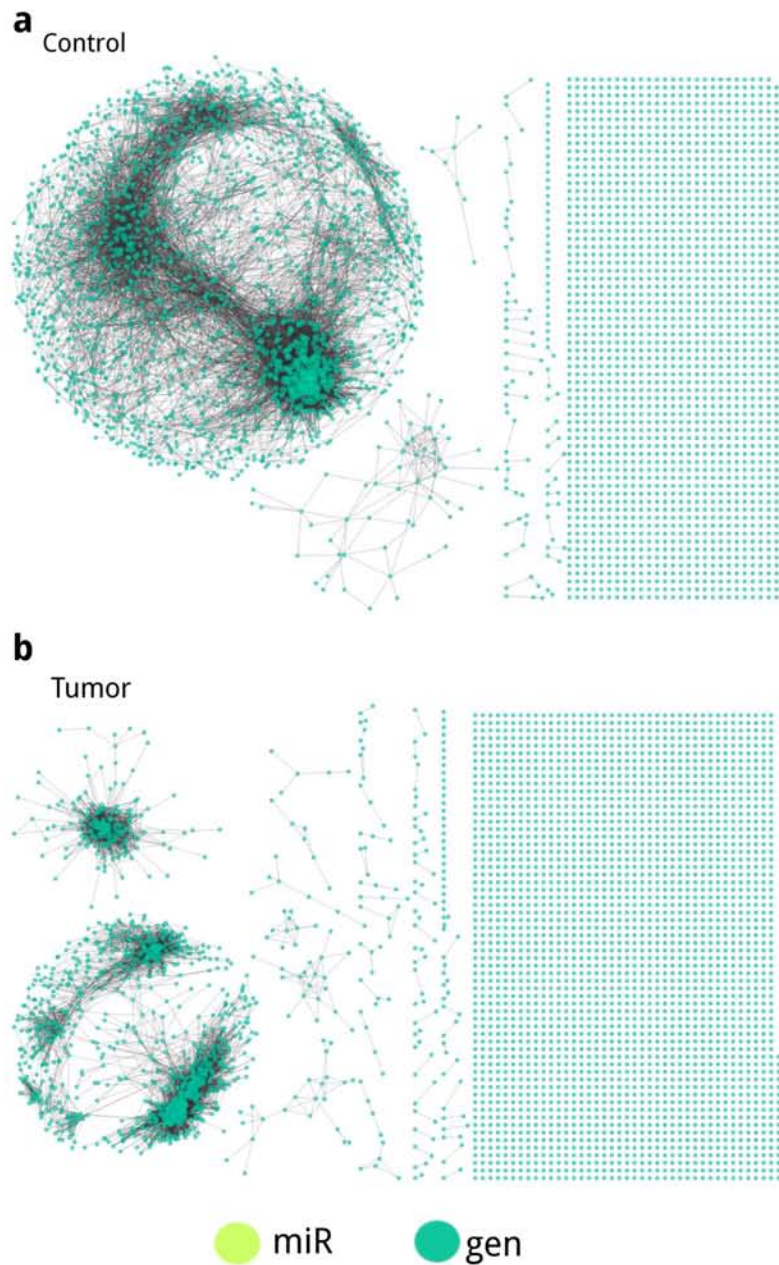


Figura 5-6: Representación de las redes de datos control a) y de tumor b) sin la presencia de miRs en su estructura. En esta representación sólo se pueden observar los nodos turquesa que corresponden a los genes y las interacciones entre ellos se representan de color gris. Como consecuencia de remover los miRs y sus interacciones las redes se han fragmentado y nodos que ya no forman parte de ningún componente debido a su ausencia de interacciones con otros genes han quedado aislados.

5.2.1. Los miRs mantienen la cohesión de las redes

Las redes tienen un componente asociado a las interacciones de los miR, y un componente asociado a las interacciones de los genes. Es bien conocido que los miR poseen propiedades biológicas diferentes a los genes, por lo que decidimos estudiar el efecto que tiene sobre la red el remover el componente asociado a los miR.

Si observamos las redes obtenidas por su componente gen-gen únicamente y removemos todos los miR y sus interacciones, podemos ver cómo la cantidad de componentes en la red aumenta y aparecen miles de nodos solos.

Este comportamiento está presente en la Tabla 1 y en las Tablas del apéndice A: A1-A4, en la sección de “Subred CG gen-gen” y en la Figura 5-6 a,b; indicando que los miRs son importantes para la cohesión de la red aún cuando los parámetros usados para su construcción sean modificados.

5.2.2. Las redes muestran enriquecimiento diferencial entre fenotipos

Para estudiar la importancia biológica de las redes, decidimos realizar un análisis de enriquecimiento de Ontología de Genes (GO). Incluimos las anotaciones para los tres aspectos usados para describir la función asociada a los genes: proceso biológico, función molecular y componente celular.

El enriquecimiento de GO para los componentes gigantes de las redes de los datos control y de tumor mostraron importantes diferencias en los procesos, funciones y componentes enriquecidos para cada red. Aunque la cantidad de genes de los que parte el análisis es similar entre las dos redes, obtuvimos 128 términos enriquecidos para el análisis del control y 446 del análisis de los datos de tumor.

Los 5 términos más enriquecidos que cubrían los criterios establecidos en los métodos se muestran en la Tabla 2; encontramos que los resultados del control están principalmente relacionados con la traducción, transcripción y transducción de señales. En cambio, el análisis de enriquecimiento de los tumores muestra procesos relacionados a la regulación de la respuesta inmune, la adhesión celular y biológica, funciones moleculares relacionadas a la unión de moléculas extracelulares (polisacáridos, citocinas, carbohidratos, patrones moleculares), y componentes celulares de la adhesión celular relacionados con la matriz extracelular y el espacio extracelular.

Tabla 2. Resultados del análisis de enriquecimiento, en la siguiente tabla se presentan los 5 términos más enriquecidos para las categorías de GO. La columna de la izquierda muestra las categorías con sus respectivos términos enriquecidos, mientras que la columna de la derecha contiene la significancia estadística asociada al enriquecimiento. Se puede observar que el análisis control muestra una gran cantidad de términos generales con la actividad celular, en cambio el análisis de los datos de tumor muestra procesos mucho más específicos y con estadísticos mucho más significativos.

Controls		Tumor	
	FDR		FDR
Proceso biológico		Proceso biológico	
Elongación traduccional	5.4716×10^{-19}	Proceso del sistema inmune	1.3985×10^{-56}
Fosforilación	7.5397×10^{-6}	Respuesta inmune	1.9804×10^{-44}
Transcripción de rRNA	1.0708×10^{-4}	Proceso de Regulación del sistema inmune	4.0949×10^{-44}
Óxido reducción	1.4061×10^{-4}	Adhesión celular	2.0828×10^{-37}
Fosforilación de aminoácidos en proteínas	1.4996×10^{-4}	Adhesión biológica	2.4522×10^{-37}
Función Molecular		Función Molecular	
Actividad transferasa	6.2019×10^{-11}	Unión de citocinas	6.3558×10^{-16}
Transferencia de grupos fosfato			
Actividad cinasa	3.2346×10^{-10}	Constituyente estructural unión de carbohidratos	1.8109×10^{-13}
Actividad fosfotransferasa, grupo alcohol como aceptor	6.1591×10^{-8}	Unión de polisacáridos	2.2194×10^{-13}
Unión de enzimas	2.3155×10^{-7}	Unión de patrones	2.2194×10^{-13}
Actividad regulatoria GTPasa	9.8384×10^{-7}	Unión de glucosaminoglucano	2.8339×10^{-13}
Componente celular		Componente celular	
Ribosoma citosólico	4.2300×10^{-20}	Parte de la región extracelular	7.3601×10^{-50}
Unión célula-célula	1.1648×10^{-11}	Matriz extracelular	3.3204×10^{-36}
Parte citosólica	6.2019×10^{-11}	Matriz extracelular proteica	6.5129×10^{-34}
Subunidad ribosomal	1.2319×10^{-9}	Espacio extracelular	5.3344×10^{-31}
Subunidad ribosomal pequeña citosólica	1.5364×10^{-9}	Parte de la matriz extracelular	1.7167×10^{-16}

FDR: corrección de Benjamini-Hochberg p-valor < 0.01

5.3. miR-200 y miR-199 definen la estructura de la red

De estas redes decidimos estudiar a los nodos que tienen el grado más alto debido a que comparten una gran cantidad de interacciones con los otros componentes de la red y juegan un papel central en la comunicación entre los demás nodos, sean miRs o genes. Los nodos con el grado más alto resultaron ser miRs tanto para la red de tumor como para el control, los 10 nodos con el grado más alto de ambas redes se presenta en la Tabla 3.

Tabla 3. Los 10 nodos con grados más altos en las redes de los datos control y de tumor.

Control		Tumor	
Nodo	Grado	Nodo	Grado
hsa-miR-200b.MIMAT0000318	1,445	hsa-let-7c-MIMAT0000064	509
hsa-miR-200a.MIMAT0000682	1,437	hsa-miR-199a.MIMAT0000231	500
hsa-miR-141.MIMAT0004598	1,392	hsa-miR-199b.MIMAT0000263	498
hsa-miR-141.MIMAT0000432	1,363	hsa-miR-337.MIMAT0000754	446
hsa-miR-200a.MIMAT0001620	1,135	hsa-miR-99a.MIMAT0000097	419
hsa-miR-200c.MIMAT0000617	1,106	hsa-miR-134.MIMAT0000447	342
hsa-miR-193b.MIMAT0004767	1,086	hsa-miR-199a.MIMAT0000232	319
hsa-miR-652.MIMAT0003322	770	hsa-miR-199b.MIMAT0004563	318
hsa-miR-22.MIMAT0000077	769	hsa-miR-382.MIMAT0000737	311
hsa-miR-378a.MIMAT0000731	631	hsa-miR-223.MIMAT0000280	285

5.3.1. Los nodos de alto grado pertenecen a las familias miR-200 y miR-199

Para estudiar las implicaciones biológicas de los nodos con grados más altos en nuestras redes, decidimos analizarlos como familias de miRs. También decidimos centrarnos principalmente en los nodos con los que comparten interacciones de manera directa, es decir, sus primeros vecinos. Estudiar a los miRs como familias es útil para este tipo de análisis debido a que las familias de miR poseen un gran poder predictivo en cuanto a su capacidad regulatoria dada su similitud estructural Kamanu *et al.* [2013]. Nos enfocamos en miR-200 y miR-199 debido a que la mayoría de sus miembros se encuentran altamente conectados en las redes control y de tumor, respectivamente.

Es importante enfatizar que en las variaciones de las redes que construimos también se mantiene este comportamiento. Ya que como se puede observar en la Tabla 3 y en las tablas del apéndice B: B1-B4, los nodos de grado más alto son miRs en todas las redes, además de que se mantienen miembros de miR-200 y miR-199 entre estos nodos.

5.3.2. miR-200 es importante para la estructura de las redes independientemente de su fenotipo

La familia de miRs con el grado más alto en la red de controles es miR-200 que se compone por: hsa-miR-200a (MIMAT0000682, MIMAT0001620), hsa-miR-200b (MIMAT0000318, MIMAT0004571), hsa-miR-200c (MIMAT0000617, MIMAT0004657), hsa-miR-141 (MIMAT0004598, MIMAT0000432), and hsa-miR-429 (MIMAT0001536); todos estos miRs están presentes en tanto en la red control como en la inferida de los datos de tumor.

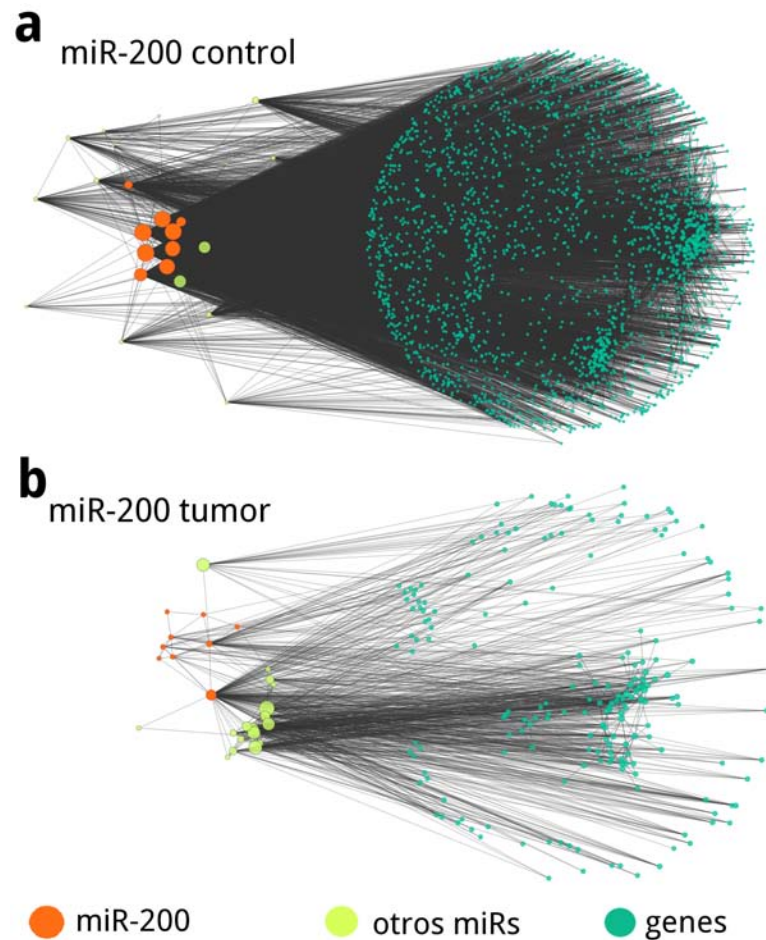


Figura 5-7: Subredes de primeros vecinos de miR-200 para los datos a) control y de b) tumor. En esta representación los nodos turquesa corresponden a los genes, los nodos verde claro corresponden a los miR y los nodos naranjas corresponden a los miR de la familia miR-200; el tamaño del nodo es proporcional a su grado de conexión y las interacciones se representan en color gris oscuro. La familia miR-200 se encuentra presente tanto en la red de tumor como en la red control; sin embargo, el grado de conectividad de miR-200 en controles es mucho mayor.

Para analizar a miR-200 en las redes de tumor y control seleccionamos a los nodos (miR y genes) con los que compartían asociaciones de manera directa para crear sub-redes de primeros vecinos. Las redes de primeros vecinos de miR-200 presentan 2,272 nodos y 16,923 interacciones en la red control (Figura 5-7 a), y 224 nodos con 1,046 interacciones en la red de tumor (Figura 5-7 b). Como era de esperarse, los grados de los miembros de miR-200 son mucho más altos en la red control que en la red de datos de tumor (Tabla 4). La red control tiene una mayor cantidad de genes y de interacciones que corresponden a asociaciones miR-gen (2,247 genes en la red control y 198 genes en la red de datos de tumor). A pesar de esto, casi no hay diferencia entre la cantidad de miRs entre las dos redes (25 miR en la red control y 26 miR en la red de datos de tumor).

Tabla 4. Grado de miR-200 en las redes de primeros vecinos.

miR	Control	Tumor
	Grado	
hsa-miR-200a.MIMAT0000682	1,437	16
hsa-miR-200a.MIMAT0001620	1,135	2
hsa-miR-200b.MIMAT0000318	1,445	18
hsa-miR-200b.MIMAT0004571	291	17
hsa-miR-200c.MIMAT0000617	1,106	53
hsa-miR-200c.MIMAT0004657	641	7
hsa-miR-141.MIMAT0004598	1,392	5
hsa-miR-141.MIMAT0000432	1,363	181
hsa-miR-429.MIMAT0001536	367	4

El análisis de enriquecimiento de los genes contenidos en las redes de primeros vecinos de miR-200 muestra procesos enriquecidos distintos para la red de datos de tumor en comparación con el control. Aunque miR-200 está presente en ambas redes los blancos de su regulación son distintos, ya que obtuvimos 33 términos GO enriquecidos para el análisis del control en comparación con los 62 términos enriquecidos para el análisis de los datos de tumor. A pesar de que la red control contiene 2,247 genes, y la red de tumor 198. Estas diferencias se pueden observar de manera evidente en la Tabla 5.

Tabla 5. Análisis de enriquecimiento para los genes de las redes de primeros vecinos de miR-200. En la tabla se muestran los términos más significativos obtenidos a partir de los genes del análisis control y de los datos de tumor. Es interesante que el análisis control posee valores de significancia estadística para su enriquecimiento que son menores a los del análisis de los datos tumor, ya que el enriquecimiento del control parte de una mayor cantidad de genes.

Control		Tumor	
	FDR		FDR
Proceso biológico		Proceso Biológico	
Adhesión celular	2.8480×10^{-4}	Regulación de la respuesta a estímulo	1.6702×10^{-6}
Adhesión biológica	2.9344×10^{-4}	Proceso de regulación del desarrollo	1.6702×10^{-6}
		Movimiento de componente celular	3.3965×10^{-6}
		Regulación de la proliferación celular	3.3965×10^{-6}
		Regulación de proceso del sistema inmune	9.4147×10^{-6}
Función Molecular		Función Molecular	
Proteína receptor transmembra	3.6305×10^{-5}	Unión de receptor	1.9975×10^{-6}
Actividad cinasa			
Proteína receptor transmembrana	9.7399×10^{-5}	Unión de carbohidratos	1.2562×10^{-5}
Actividad proteína tirosina cinasa			
Actividad proteína tirosina cinasa	3.6087×10^{-3}	Unión de polisacáridos	6.1321×10^{-4}
Actividad proteína cinasa	1.0257×10^{-2}	Unión de patrones	6.1321×10^{-4}
		Unión de factores de crecimiento	9.2541×10^{-4}
Componente celular		Componente celular	
Unión celular	8.6796×10^{-13}	Parte de la región extracelular	9.1753×10^{-9}
Unión de anclaje	6.3052×10^{-11}	Espacio extracelular	9.4147×10^{-6}
Unión célula-célula	6.3052×10^{-11}	Matriz extracelular	3.3656×10^{-4}
Unión adherente	1.0953×10^{-7}	Superficie celular	3.3656×10^{-4}
Membrana plasmática basolateral	1.0953×10^{-7}	Matriz extracelular proteica	5.0306×10^{-4}

FDR: corrección de Benjamini-Hochberg p-valor < 0.01

5.3.3. Las redes de primeros vecinos de miR-200 muestran un centro común relacionado a EMT/MET

Intersección de las redes de miR-200

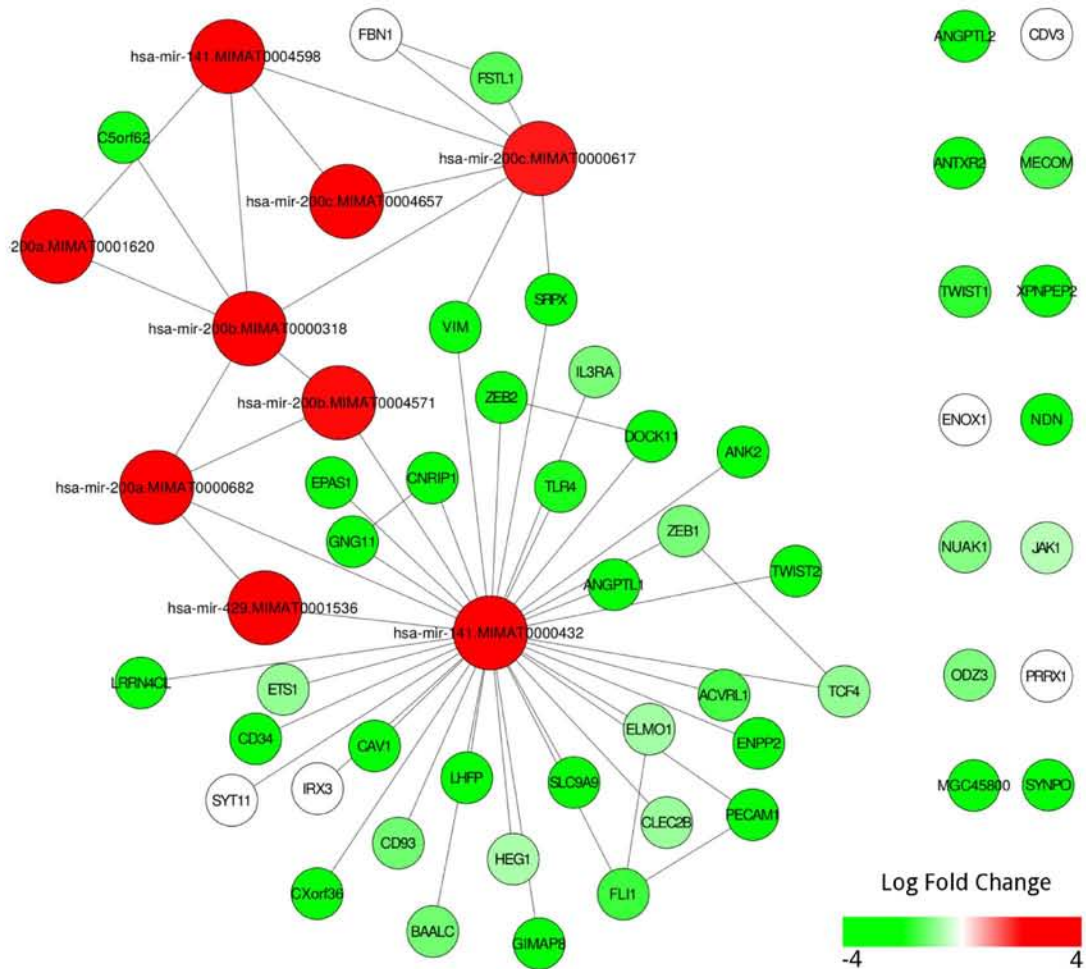


Figura 5-8: Intersección de las redes de primeros vecinos de miR-200 control y de tumor. En esta figura se representan aquellos nodos y aquellas interacciones que aparecen tanto en la red de tumor como en la red control, lo que quiere decir que se encuentran conservadas entre los fenotipos a pesar de sus diferencias estructurales. En esta representación los nodos están coloreados de acuerdo a su expresión diferencial. Es interesante notar que en esta representación basada en los datos experimentales utilizados para este análisis los miRNAs están sobre expresados y los genes sub expresados.

Aunque las redes de miR-200 difieren en gran manera en cuanto a su número de nodos y su conectividad (Figura 5-7 a,b), existen 59 genes comunes entre ellas y algunos hasta mantienen las mismas interacciones (Figura 5-8). En la intersección representada en la Figura 5-8 podemos encontrar a genes asociados a la regulación de la transición epitelio mesénquima y mesénquima epitelio, tales como VIM, ZEB1/2 y TWIST1/2 [Zhu *et al.*, 2016; Park *et al.*, 2008; Mani *et al.*, 2008]. Es importante resaltar que aunque las interacciones entre miR-200 y los genes se conservan entre los fenotipos sus valores de expresión difieren en gran manera con la sobreexpresión de miR-200 y la subexpresión de los genes asociados en el fenotipo tumoral, y el comportamiento inverso en el control (Figura 5-8).

5.3.4. El comportamiento de miR-199 es determinante para la estructura de la red de datos de tumor

La familia con grado más alto en la red de datos de tumor fue la familia miR-199, compuesta por: hsa-miR-199b (MIMAT0000263, MIMAT0004563), hsa-miR-199a-1 and hsa-miR-199a-2 (MIMAT0000231, MIMAT0000232). Los miembros de la familia miR-199 están presentes en el componente gigante de la red inferida a partir de los datos de tumor, siendo de los nodos que poseen los grados más altos (Tabla 6).

Tabla 6. Grado de miR-199 en las redes de primeros vecinos. Los miembros de la familia miR-199 poseen grados altos de manera consistente en la red de tumor, sin embargo en la red control ni siquiera están presentes en el componente gigante y algunos de los miembros están ausentes.

miR	Control Tumor	
	Grado	
hsa-miR-199a.MIMAT0000231		500
hsa-miR-199a.MIMAT0000232	1	319
hsa-miR-199b.MIMAT0000263		498
hsa-miR-199b.MIMAT0004563	1	318

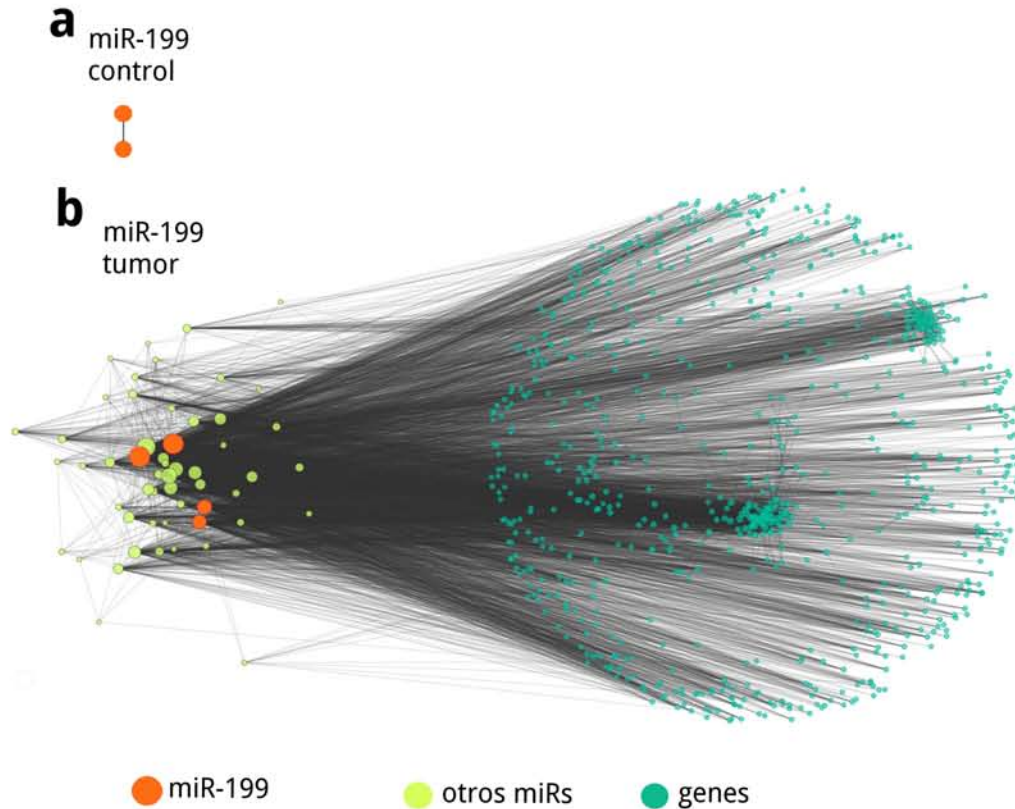


Figura 5-9: Subredes de primeros vecinos de miR-199 para los datos a) control y de b) tumor. Los nodos turquesa representan a los genes, los nodos verde claro representan a los miR y los nodos naranjas representan a los miR de la familia miR-199, el tamaño de los nodos es proporcional a su grado; las interacciones se representan de color gris oscuro. Se puede observar que esta familia prácticamente no está presente en la red control.

En cambio en la red inferida a partir de los datos control estos miRs sólo se encuentran presentes como una red aislada de dos nodos (grado = 1, Figura 5-9 a). Al seleccionar a miR-199 y sus primeros vecinos (miR y genes) de la red de datos de tumor, obtuvimos una red con 834 nodos y 7,053 interacciones (5-9 b). Los grados que presentan los miembros de la familia miR-199, y de hecho, todos los miR de grado más alto en la red de datos de tumor presentan grados más bajos que los miR más conectados en la red control.

Los genes presentes en la red de primeros vecinos de miR-199 enriquecieron para 119 términos de GO. Los 5 términos más significativos pertenecientes a las tres categorías de GO se pueden encontrar en la Tabla 7. Los procesos biológicos enriquecidos por los genes de la red de primeros vecinos de miR-199 en la red de tumor están principalmente relacionados a la adhesión celular, la organización

extracelular y el desarrollo. Las funciones enriquecidas incluyen la unión de moléculas y la participación de la matriz extracelular. Los componentes celulares enriquecidos también están centrados en la región extracelular. La red de controles no contiene ningún gen, por lo que no hay términos GO enriquecidos. Es importante mencionar que los resultados de los términos enriquecidos son similares a los obtenidos para los genes de la red de primeros de miR-200 de la red de tumor.

Tabla 7. Análisis de enriquecimiento para los genes de las redes de primeros vecinos de miR-199. En la tabla se muestran los términos más significativos obtenidos a partir de los genes del análisis de los datos de tumor.

Tumor	
	FDR
Proceso Biológico	
Adhesión biológica	8.5209×10^{-29}
Adhesión Celular	8.5209×10^{-29}
Organización de la estructura extracelular	1.0724×10^{-13}
Desarrollo del sistema esquelético	3.9744×10^{-13}
Oranización de la matriz extracelular	1.9525×10^{-12}
Función Molecular	
Unión de iones de calcio	1.1340×10^{-16}
Constituyente de la matriz extracelular	2.8499×10^{-11}
Unión de integrinas	7.4108×10^{-9}
Unión de glucosaminoglucano	1.0312×10^{-6}
Unión de patrones	5.5217×10^{-6}
Componente celular	
Matriz extracelular	4.0521×10^{-42}
Matriz extracelular proteica	7.9803×10^{-41}
Parte de la región extracelular	1.2162×10^{-27}
Unión de iones de calcio	1.1340×10^{-16}
Parte de la matriz extracelular	1.8404×10^{-13}

FDR < 0.01

5.4. Los miRs de la red de datos de tumor muestran una tendencia a localizarse en el cluster *DLK1-DIO3*

Analizamos las localizaciones cromosómicas de los miRs de las redes de primeros vecinos de miR-200 y miR-199. Encontramos que los miRs de la red control de miR-200 principalmente mapean a localizaciones en el cromosoma 1 y 12, las cuales corresponden a la localización de los miRs de la familia miR-200.

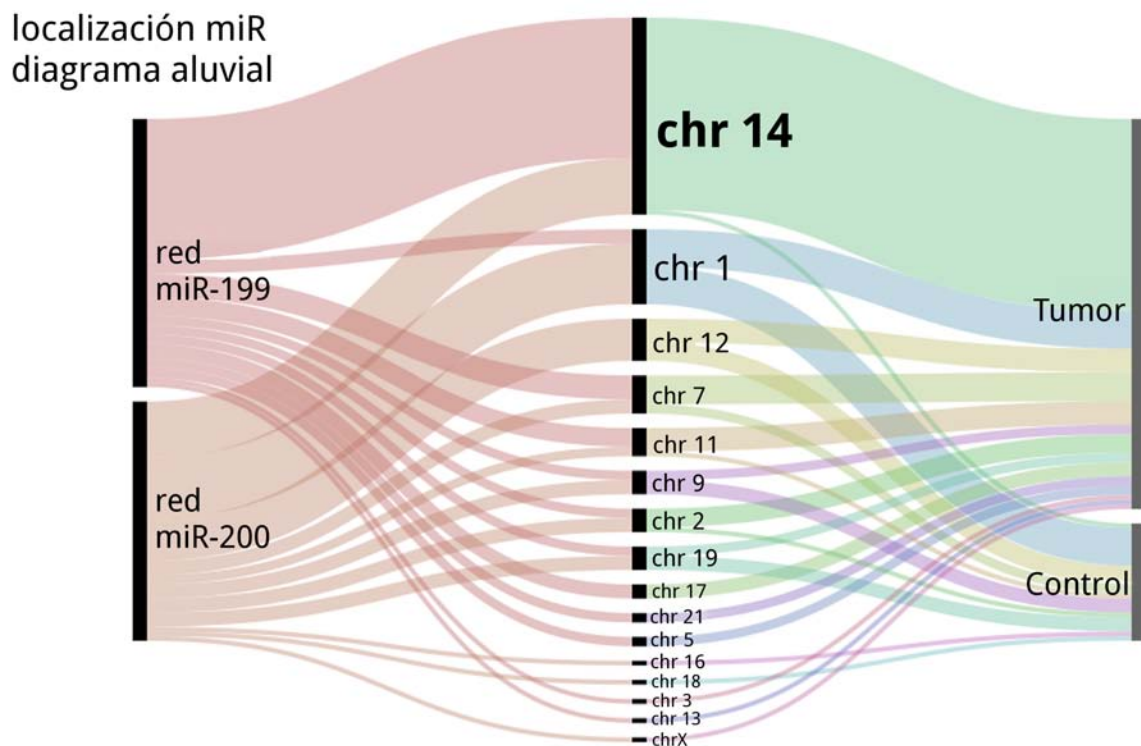
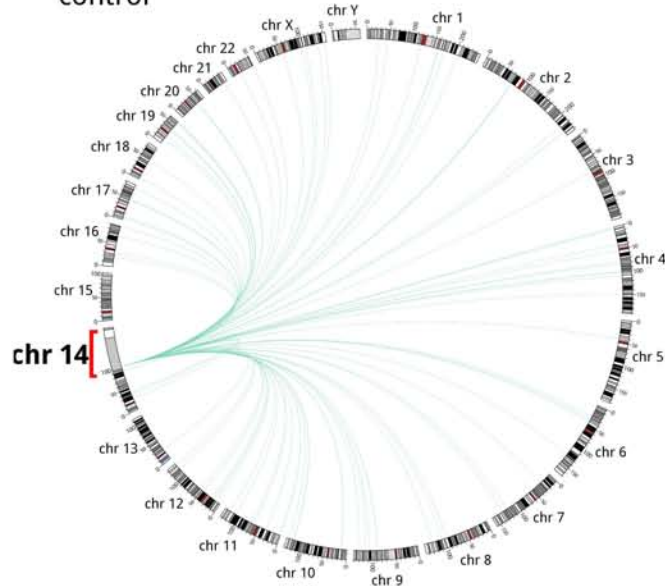


Figura 5-10: Diagrama aluvial de la posición genómica de los miR presentes en las redes de miR-200 y miR-199. En este diagrama se representa la proporción de miRs de las subredes de primeros vecinos que mapean a un determinado cromosoma. Se puede observar que una gran cantidad de los miRs presentes en las subredes de la red de tumor se localizan en el cromosoma 14.

También observamos que en la red inferida a partir de los datos control miR-200 tiende a asociarse con un número restringido de miRs que se localizan en distintos cromosomas, sin ninguna preferencia por alguno de ellos. En cambio, en la red inferida a partir de los datos de tumor los miRs tienden a mapear a posiciones en el cromosoma 14 (Figura 5-10).

a Interacciones de los miR del cluster *DLK1-DIO3* control



b Interacciones de los miR del cluster *DLK1-DIO3* tumor

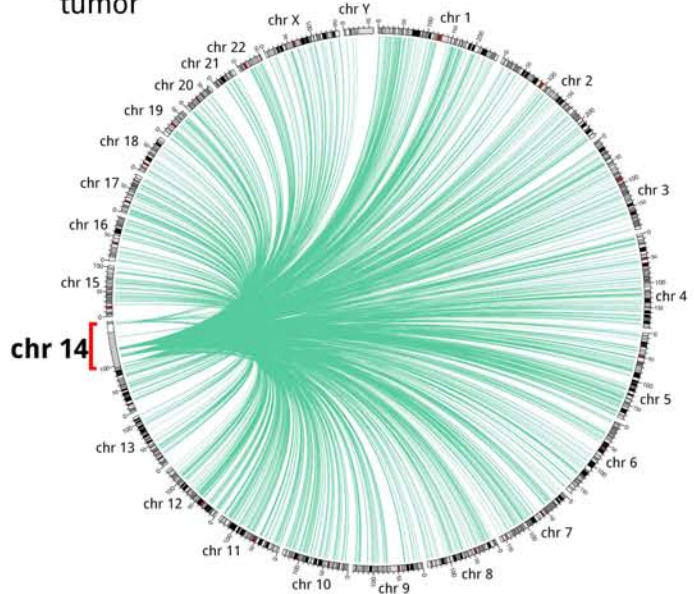


Figura 5-11: Gráficas de circo para las interacciones directas con miRs del cluster *DLK1-DIO3* para las redes de los datos de control a), y tumor b). En estas gráficas se representan en color verde las interacciones de los miR del cluster con genes y miRs de acuerdo a posición en el genoma. Todos los cromosomas se representan de forma circular, y la región *DLK1-DIO3* (cromosoma 14q32) presenta un aumento de 100x para facilitar la visualización. Se puede observar que en la red de tumor hay una cantidad mucho mayor de interacciones de los miR del cluster *DLK1-DIO3* al resto del genoma en comparación de la red control.

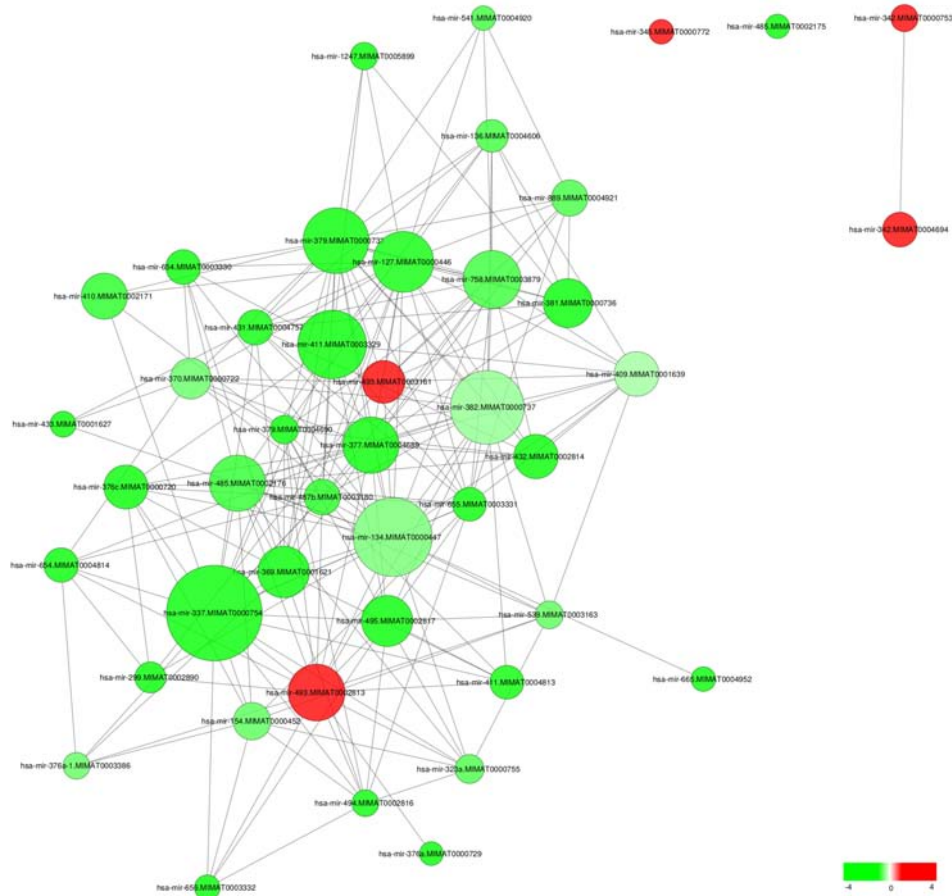


Figura 5-12: Subred de miRs del cluster *DLK1-DIO3* presentes en la red inferida a partir de los datos de tumor. Sólo se muestran las interacciones que corresponden a las asociaciones miR-miR, y los nodos están coloreados de acuerdo a su expresión diferencial. En esta representación se puede observar que la mayoría de los miR están subexpresados.

La localización cromosómica de los miRs que mapean al cromosoma 14 muestra que estos miR están localizados en una región definida conocida como el cluster *DLK1-DIO3*. Esta región en el Chr14q32 se caracteriza por tener una densidad alta de miRs y otros genes de RNAs no codificantes. Una gráfica de circos (<http://circos.ca/>) es una manera de visualizar información en una forma circular que permite explorar relaciones entre objetos o posiciones, por lo que ha resultado especialmente útil para representar las interacciones de una red con sus posiciones genómicas; por lo que decidimos realizar este tipo de visualización para estudiar a los miR del cluster *DLK1-DIO3*. Las gráficas de circos en la Figura 5-11 a,b, muestran las interacciones asociadas a los miRs localizados en el cluster *DLK1-DIO3* y sus blancos directos (primeros vecinos); en estas figuras se puede observar que en la red inferida a partir de datos de tumor (Figura 5-11 b) presenta una cantidad de asociaciones

mucho mayor en comparación del control (Figura 5-11 a). En la Figura 5-12 se encuentra una representación visual de el perfil de expresión de los miR del cluster *DLK1-DIO3* presentes en la red inferida a partir de los datos de tumor.

5.5. Análisis de vías: miRs y la deregulación de las vías asociadas a EMT

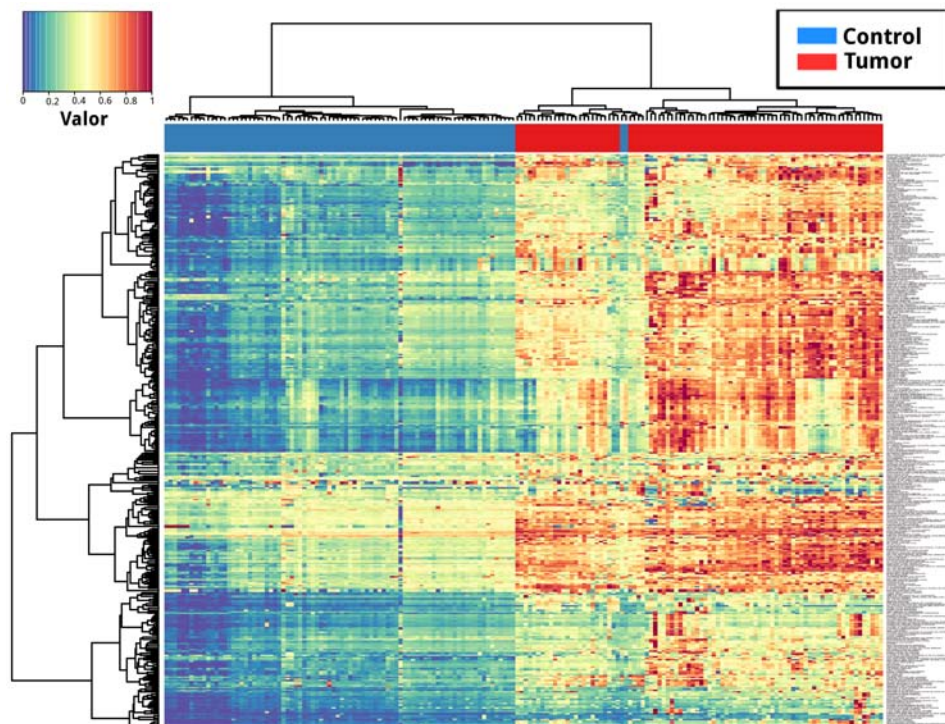


Figura 5-13: Mapa de calor de los Score de deregulación de Vías (PDS) de las vías de Reactome que contienen genes asociados a los miR del cluster *DLK1-DIO3* en las redes inferidas a partir de los datos de tumor. En esta representación las filas corresponden a las vías de señalización seleccionadas para el análisis, y las columnas corresponden a las muestras sean tumorales o control. Los PDS obtenidos para las 393 vías de Reactome y las 172 muestras que trabajamos (86 muestras de tejido tumoral y 86 muestras de tejido control) se encuentran en las celdas del mapa de calor, coloreados de acuerdo a la escala de colores en la esquina superior izquierda. Lo que significa que las celdas en colores azules y verdes representan vías que se encuentran menos dereguladas con respecto al grupo control, mientras que las celdas rojas representan vías que se encuentran altamente dereguladas. La barra superior indica si las muestras son casos o controles. Se utilizaron distancias euclidianas y el método de agrupamiento jerárquico de Ward para crear el dendrograma. La flecha roja señala a la vía: señalización del receptor TGF-Beta en EMT (Transición epitelio mesénquima).

Seleccionamos las vías de señalización de las bases de datos de Reactome, WikiPathways y KEGG que contenían por lo menos un gen que corresponde a alguno de los nodos de las redes de primeros vecinos de miR-200 y los miR del cluster *DLK1-DIO3* en las redes de tumor para analizar las vías en las que participan los genes asociados a los miRs de interés (ver métodos). Utilizando Pathifier [Drier *et al.*, 2013] estimamos el Score de Deregulación de Vías (PDS) de las vías seleccionados para cada una de las muestras.

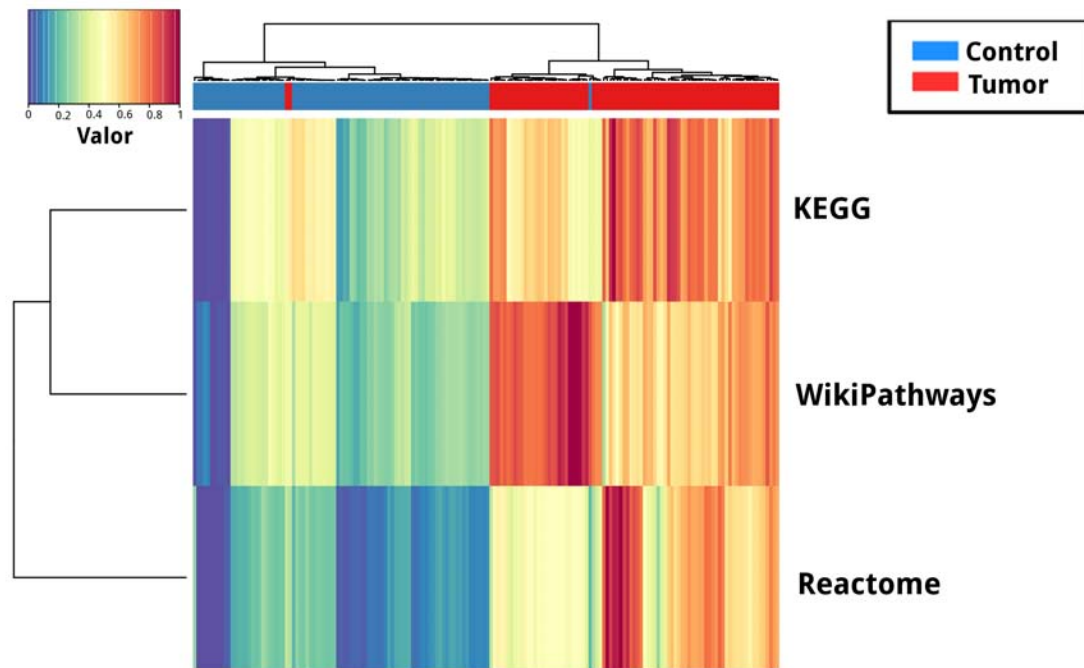


Figura 5-14: Mapa de calor de deregulación de las vías asociadas a EMT, estas corresponden a la vía de señalización del receptor TGF-Beta en EMT (Transición epitelio mesénquima) en Reactome, la vía de señalización de TGF-Beta en células de tiroides para la Transición Epitelio Mesénquima en WikiPathways y la vía de señalización de TGF-Beta en KEGG. Estas vías muestran importante deregulación en comparación con las muestras control.

Elegimos Pathifier debido a que nos permite integrar la información de la expresión de genes y la información de las vías de cada muestra en una métrica específica al contexto (PDS), reflejando las alteraciones con respecto a nuestro control. Obtuvimos matrices de PDS para 393 vías de Reactome (Figura 5-13), 237 vías de WikiPathways (Figura C-1) y 133 vías de KEGG (Figura C-2) que contenían genes relacionados a los miR del cluster *DLK1-DIO3*. También obtuvimos matrices de 193 vías de Reactome (Figura C-3), 159 vías de WikiPathways (Figura C-4) y 79 vías de KEGG (Figura C-5) que

contenían genes relacionados a la red de primeros vecinos de miR-200 en los datos de tumor.

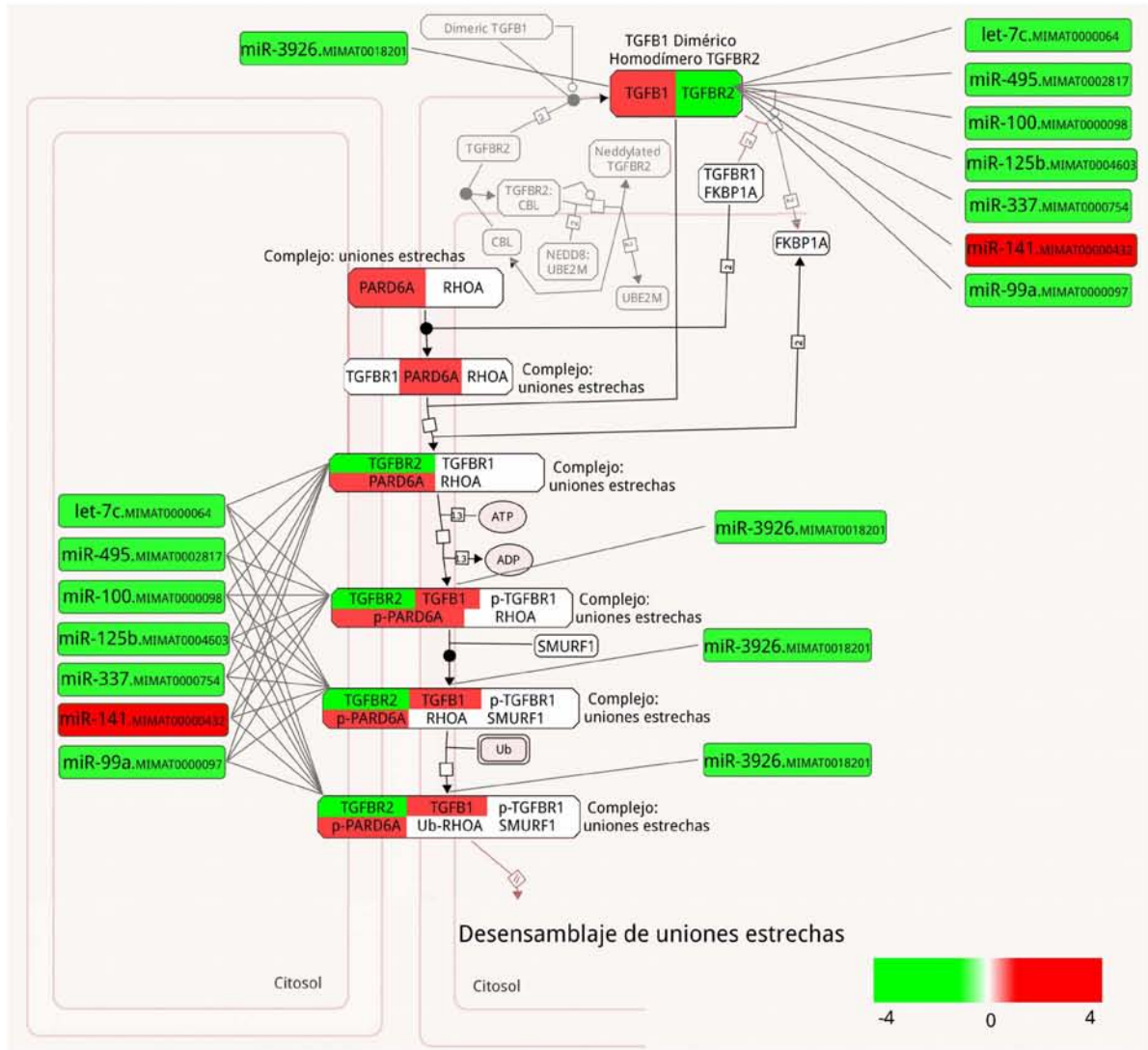


Figura 5-15: Representación de las asociaciones inferidas en la red de datos de tumor que se relacionan con los genes de vía de vía de señalización del receptor TGF-Beta en EMT (Transición epitelio mesénquima) en Reactome. En esta figura se puede observar que hay una cantidad importante de interacciones entre miR y genes en nuestras redes que se asocian con genes que participan en la vía de señalización de EMT. Entre los miR que se asocian a esta vía se encuentran miRs que aparecen de manera recurrente en los resultados del análisis de las redes, como es el caso miR-141 (MIMAT00000432) que también se encuentra presente en el núcleo de interacciones conservadas entre las redes de miR-200 para las redes control y de tumor.

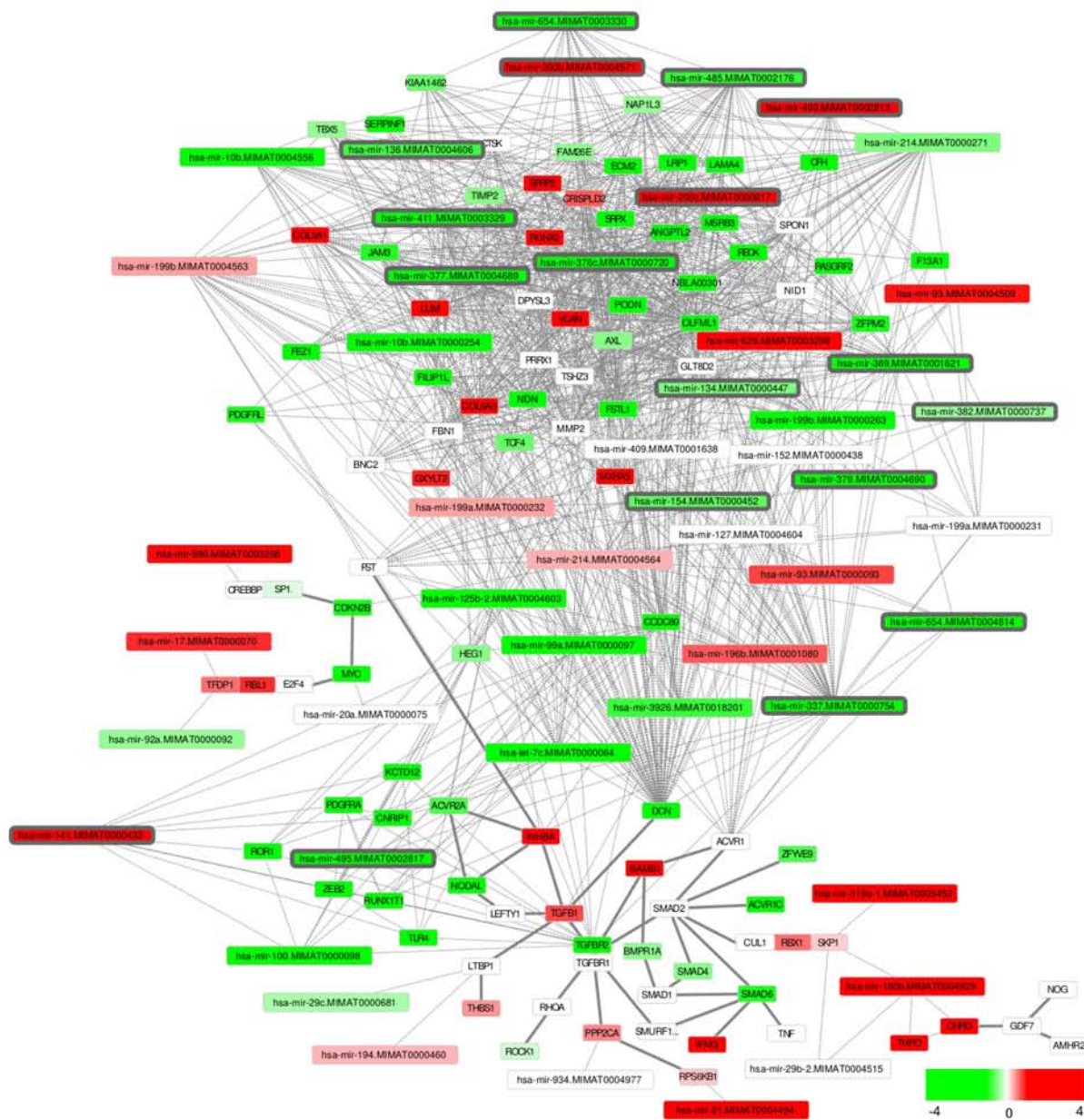


Figura 5-16: Representación en forma de red de la vía de señalización de TGF-Beta de acuerdo a KEGG adicionando las asociaciones directas de los genes de la vía con nodos presentes en la red de datos de tumor. Los nodos están coloreados de acuerdo a su expresión diferencial; las interacciones lisas representan las interacciones de la vía, y las interacciones quebradas representan las interacciones inferidas a partir de nuestro análisis. Los nodos con bordes resaltados representan a miR-200 y miRs del cluster *DLK1-DIO3*. En esta figura podemos observar que hay una gran cantidad de genes y miRs que interactúan de manera directa con los genes que participan en la vía. La expresión diferencial de la red muestra que las relaciones de regulación son muy complejas y se asocian con la sobreexpresión y subexpresión de miRs relacionados principalmente a miR-200 o al cluster *DLK1-DIO3*.

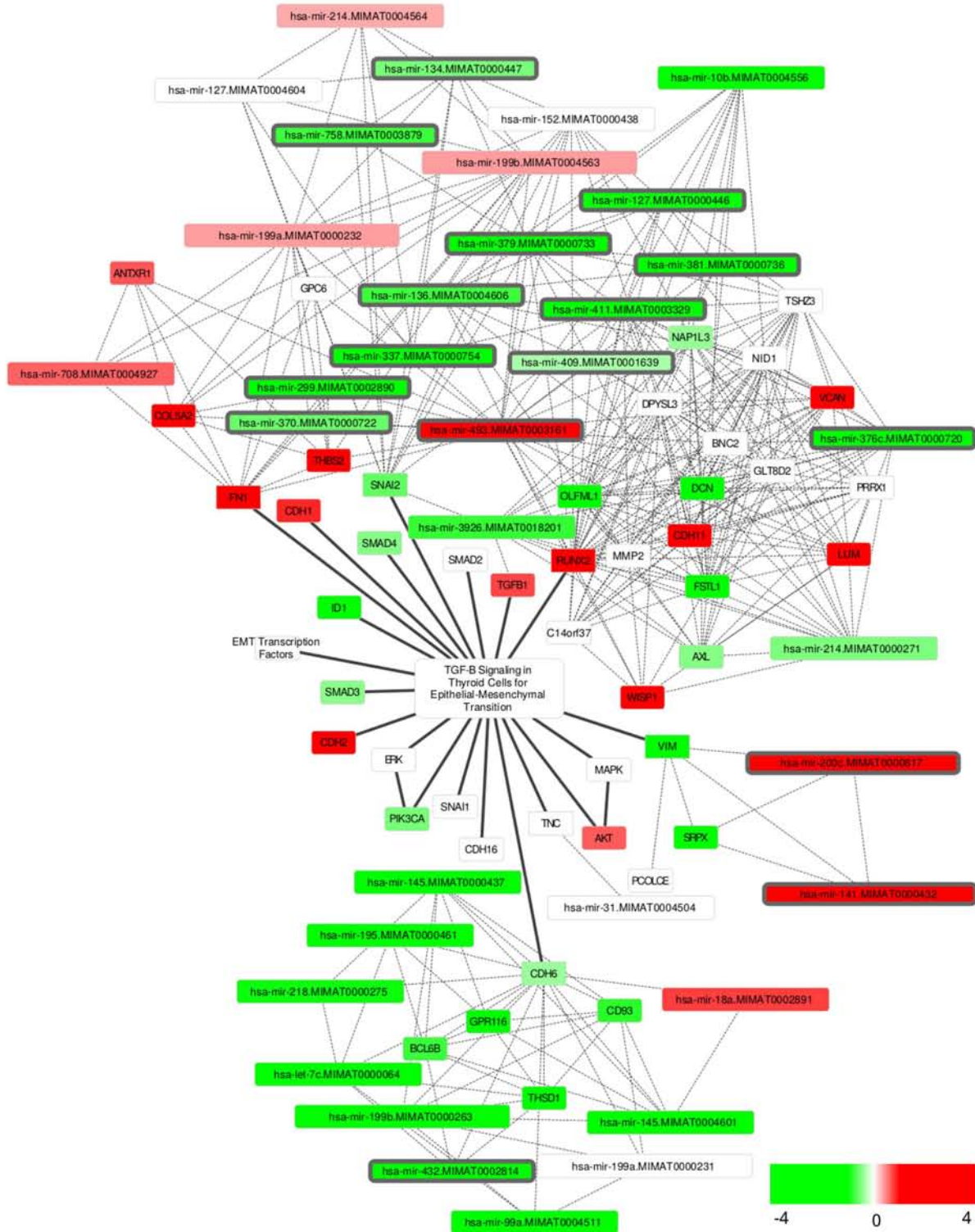


Figura 5-17: Visualización en forma de red de la vía de señalización de TGF-Beta en células de tiroides para la Transición Epitelio Mesénquima en WikiPathways con las asociaciones relacionadas de la red inferida de datos de tumor. En esta representación se puede observar que una cantidad importante de los genes y los miR que interactúan de forma directa con los genes descritos en la vía se encuentran subexpresados.

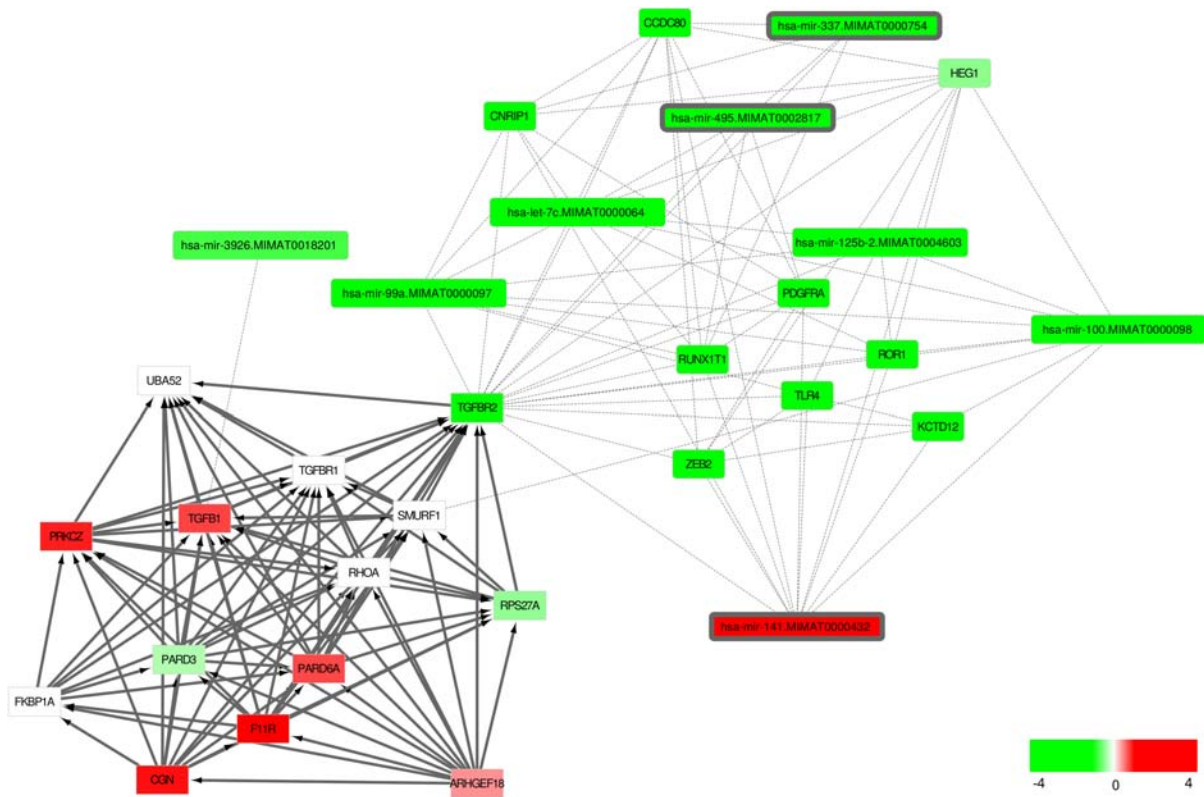


Figura 5-18: Representación en forma de red de la vía de señalización del receptor TGF-Beta en EMT (Transición epitelio mesénquima) en Reactome con las asociaciones relacionadas de la red inferida de datos de tumor. Se puede observar que esta vía es más pequeña en relación con las vías descritas en otras bases de datos, además posee un número menor de interacciones directas que mapean a interacciones presentes en la red de tumor que fue inferida a partir de los datos de TCGA. Es interesante notar que mientras miembros de la vía muestran sobreexpresión los miR y genes asociados por medio de las redes están principalmente subexpresados.

Buscamos vías comunes entre los análisis de los genes en la red de primeros vecinos de los miR del cluster y el análisis de los genes en la red de primeros vecinos de miR-200, y encontramos vías asociadas a EMT dereguladas en cada base de datos. Estas vías corresponden a la vía de señalización del receptor TGF-Beta en EMT (Transición epitelio mesénquima) en Reactome, la vía de señalización de TGF-Beta en células de tiroides para la Transición Epitelio Mesénquima en WikiPathways y la vía de señalización de TGF-Beta en KEGG (Figura 5-14). De manera específica, la vía de Reactome de EMT sugiere la existencia de una relación regulatoria entre miR-141 (MIMAT0000432) y TGFBR2, así como la participación de los miR del cluster *DLK1-DIO3* (Figura 5-15). En las Figuras 5-18, 5-17 y 5-16 se representan las vías asociadas a EMT para cada una de las bases de datos junto con las

interacciones que reconstruimos a partir de los datos control y de tumor. Algunas de estas vías han sido anotadas para un tipo de tejido en específico (como por ejemplo: tiroides), sin embargo esto se debe a que los mecanismos asociados a EMT se mantienen principalmente desconocidos y aún se encuentra en proceso de anotación.

Tabla 8. miRs relevantes en la red inferida a partir de datos de tumor y su relación en el cáncer. Esta tabla muestra a miRs presentes en las redes de miR-200, miR-199 y de los miRs en el cluster *DLK1-DIO3*. Las funciones anotadas de los miR son consistentes con su perfil de expresión y el fenotipo a partir del cual la red fue inferida.

miR	Participación en cáncer	Expresión		Referencia
		5-p	3-p	
Red de datos de tumor de miR-200				
miR-381	Supresor de la migración indirecto, Asociado a la auto renovación		-1.49	Ming <i>et al.</i> [2015] Boo <i>et al.</i> [2016]
miR-379	Sub expresado en cáncer de mama, tiene a la ciclina B1 como blanco	-1.19	-1.40	Khan <i>et al.</i> [2013]
miR-100	Supresión de la migración de las células cancerosas de mama por medio de la inhibición de la vía Wnt/beta-catenina activador de EMT, tumorigenesis e inhibidor de la invasión	-1.69		Jiang <i>et al.</i> [2015]
miR-96	Altamente sobre expresado en cáncer de mama, importante para el crecimiento celular y la migración	3.16		Li <i>et al.</i> [2014]
Red de datos de tumor de miR-199				
miR-145	Inhibe crecimiento y migración de células de cáncer de mama	-2.27	-1.52	Zheng <i>et al.</i> [2015]
miR-656	Sub expresado en múltiples tipos de cáncer, asociado a supresión tumoral		-1.29	Laddha <i>et al.</i> [2013]
miR-655	Relacionado a la inhibición de MET en cáncer de mama sobre expresión asociada a inhibición de la migración e invasión		-1.01	Lv <i>et al.</i> [2016]
miR-493	Reducida supervivencia de pacientes con cáncer de mama tumores agresivos y resistencia a fármacos	1.35	1.49	Tambe <i>et al.</i> [2016]
Red inferida de datos de tumor de los miR del Cluster <i>DLK1-DIO3</i>				
miR-379	Relacionado con EMT/MET en modelo de cáncer de próstata	-1.19	-1.40	Gururajan <i>et al.</i> [2014]
miR-495	Relacionado a la represión de la señalización entre TWIST-1, SMI-1, ZEB-1/2 y miR-200		-1.50	Haga y Phinney [2012]

5.6. La función de los miRs es consistente con su participación en la red de datos de tumor

Las redes inferidas a partir de los datos de tumor poseen muchos miRs con funciones reportadas, y que tienen relación a la promoción de los tumores. Algunos de ellos se han estudiado incluso, en el

contexto específico del cáncer de mama. Los ejemplos más representativos de estos miRs se encuentran en la Tabla 8. Es importante recalcar que sus funciones estudiadas muestran ser consistentes con el fenotipo en el que se reconstruyeron sus asociaciones y con su perfil de expresión.

5.7. Las asociaciones de los miRs y genes son consistentes con interacciones en TargetScan y miRTarBase

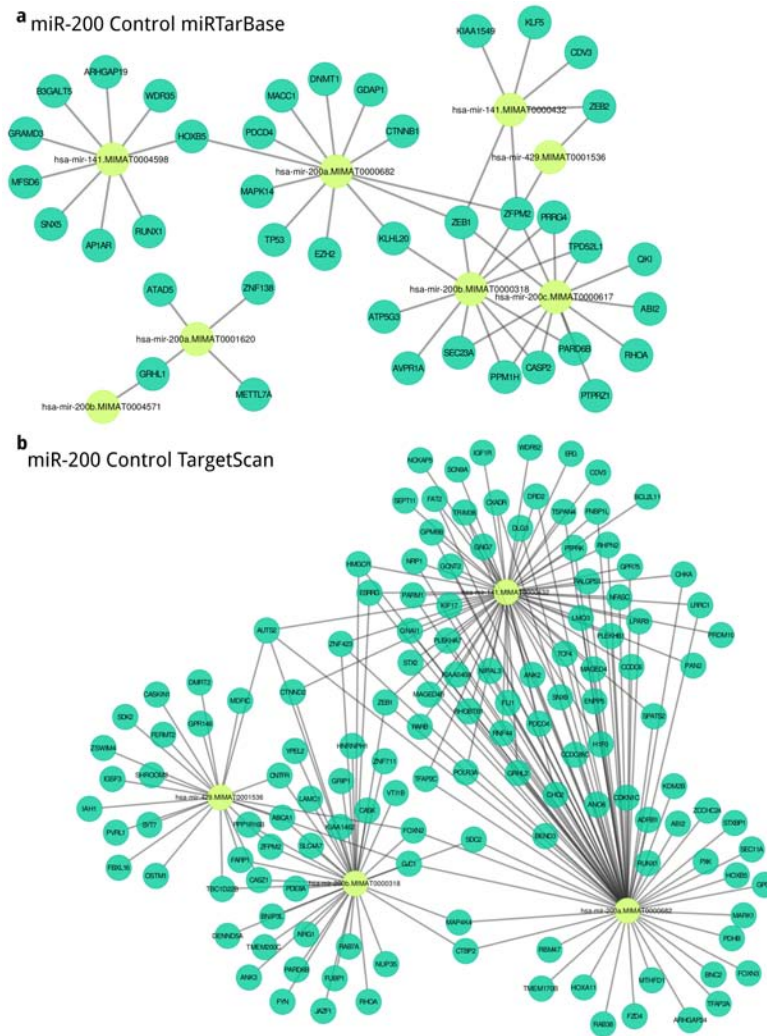


Figura 5-19: Asociaciones de las redes de primeros vecinos de miR-200 que empalman con interacciones de las bases de datos miRTarBase para la red control a), y la base de datos TargetScan para la red b) control. Las redes control están coloreadas de acuerdo al tipo del nodo (miR = verde claro, gen = turquesa).

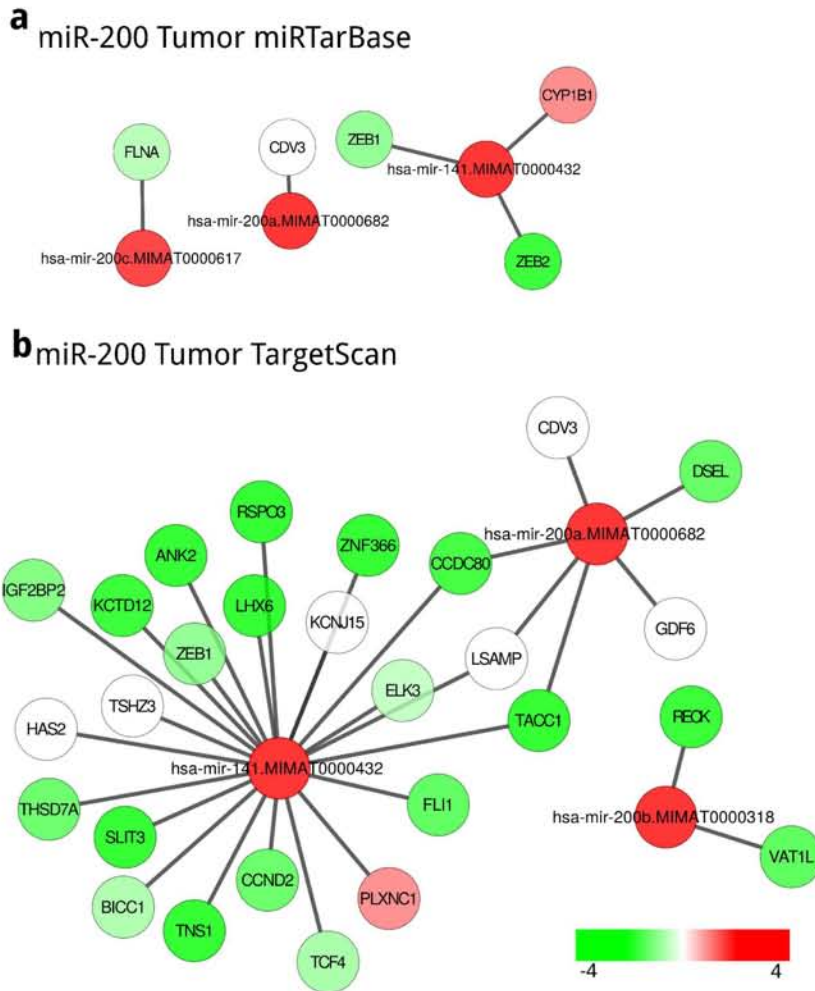


Figura 5-20: Asociaciones de las redes de primeros vecinos de miR-200 que empalman con interacciones de las bases de datos miRTarBase para la red de tumor a), y la base de datos TargetScan para la red de b) tumor. Las redes de tumor están coloreadas de acuerdo a su expresión diferencial. También en este análisis aparece miR-141 (MIMAT0000432) en los resultados, las interacciones que se conservan entre la red y las bases de datos incluyen a miR-141 y factores de transcripción como *ZEB1* y *ZEB2*.

Usando toda la información disponible en las bases de datos de interacciones experimentalmente validadas miRTarBase [Chou *et al.*, 2016] y la base de datos de predicción miR-blanco TargetScan [Friedman *et al.*, 2009], exploramos la presencia de interacciones comunes con las asociaciones que reconstruimos en nuestras redes. De nuestra red de primeros vecinos construida a partir de los datos de tumor, evaluamos los nodos y las interacciones que mapean a una interacción validada o predicha en las bases de datos antes mencionadas. Los resultados de ambos análisis se pueden encontrar en la Tabla 9 y 10, y las asociaciones con miR-200 también se representan en las Figuras 5-19, 5-20. En

consistencia con los demás resultados, también en este análisis aparece miR-141 (MIMAT0000432) al analizar las interacciones de miR-200.

Tabla 9. Nodos e interacciones comunes entre las bases de datos TargetScan y miRTarBase, y las asociaciones obtenidas por medio de la metodología propuesta.

Control				Tumor			
miRTarBase		TargetScan		miRTarBase		TargetScan	
Nodos	2,987	Nodos	2,330	Nodos	3,484	Nodos	2,692
Nodos con int.	189	Nodos con int.	239	Nodos con int.	263	Nodos con int.	250
Int. en común	186	Int. en común	298	Int. en común	202	Int. en común	243

Nodo: se refiere a la cantidad de genes y miRs que están presentes tanto en las bases de datos como en las redes inferidas.

Nodos con int.: nodos que además de estar presentes en las redes y las bases de datos presentan interacciones comunes.

Int. en común: número de interacciones comunes entre la red y las bases de datos.

Tabla 10. Nodos y interacciones comunes entre las redes de primeros vecinos de miR-200 y las bases de datos.

Control				Tumor			
miRTarBase		TargetScan		miRTarBase		TargetScan	
Nodos con int.	48	Nodos con int.	140	Nodos con int.	8	Nodos con int.	30
Int. en común	57	Int. en común	203	Int. en común	5	Int. en común	30

Nodo: se refiere a la cantidad de genes y miRs que están presentes tanto en las bases de datos como en las redes inferidas.

Nodos con int.: nodos que además de estar presentes en las redes y las bases de datos presentan interacciones comunes.

Int. en común: número de interacciones comunes entre la red y las bases de datos.

Parte IV

Discusión

Para entender la relación de regulación entre los miRs y los genes en cáncer de mama, usando datos pareados de RNA-Seq y miR-Seq de muestras de tejido tumoral y tejido adyacente de los mismos 86 pacientes, construimos y analizamos redes de regulación génica. Estas redes las inferimos usando un abordaje basado en MI; en el cual los nodos corresponden a los miR y los genes, y las interacciones a la dependencia estadística entre los perfiles de expresión para cada pareja de nodos. Las redes obtenidas muestran diferencias en sus distribuciones de MI y en su estructura. La comparación de las distribuciones de MI entre las redes inferidas a partir de los datos control (Figura 5-1 a) y de tumor (Figura 5-1 b) muestran que las interacciones de la red de tumor tienden a tener valores de MI más bajos, y las diferencias entre las redes de tumor y control que se observan en las Figuras 5-4, 5-5 son evidentes. En las redes que construimos se puede observar una propiedad cohesiva de mis miRs sobre la red (Figura 5-6 a,b). Además, en dichas redes encontramos que los nodos de más alto grado corresponden a la familia miR-200 (Figura 5-7 a,b) para la red control y a miR-199 (Figura 5-9 a,b) en la red de tumor.

A partir de las asociaciones de los nodos en la red con las familias de miR más altamente conectadas construimos redes de primeros vecinos para miR-200 y miR-199, las cuales demostraron tener enriquecimiento funcional diferencial (Tablas 13 y 14). Sin embargo, a pesar de las diferencias funcionales las redes de miR-200 demostraron poseer un núcleo común de nodos y interacciones directamente asociado a miR-200 (Figura 5-8) que se comparte tanto en la red control (Figura 5-7 a) como en la red de tumor (Figura 5-7 b). En cuanto a la localización cromosómica, en nuestros resultados encontramos que un cluster de miRs en el Chr14q32 (Figura 5-10) presenta una importante diferencia en conectividad entre las redes control (Figura 5-11 a) y de tumor (Figura 5-11 b).

Un análisis de deregulación de vías de señalización indicó que las vías asociadas a la transición epitelio mesénquima y TGF-beta son procesos cruciales involucrados en el cáncer de mama que se encuentran deregulados en nuestras muestras. Finalmente analizamos la presencia de miRs relevantes con base en información de bases de datos de predicción de asociaciones miR-blanco y bases de datos de interacciones de miRs experimentalmente validadas. Dados los resultados antes descritos y congruencia a la evidencia previamente reportada proponemos una serie de hipótesis que sugieren que los miRs tienen un papel central en la regulación transcripcional del cáncer de mama.

Aún si las redes inferidas de los datos control o de tumor fueron construidas de la misma manera utilizando datos de muestras de tejido provenientes de los mismos 86 pacientes, observamos diferencias importantes entre dichas redes. Estas diferencias parecen reflejar la deregulación transcripcional

presente en las muestras tumorales. Por ejemplo, la mayor cantidad de miRs y de interacciones miR-miR en la red inferida a partir de los datos de tumor podrían considerarse como una ganancia de la actividad regulatoria de los miR. Sin embargo, éste parece no ser el caso, debido a que las interacciones en las redes inferidas a partir de los datos control tienden a poseer valores de MI más altos (Figura 5-1 a) que los obtenidos por las artistas de la red de datos de tumor (Figura 5-1 b). Es importante resaltar que el MI es una medida de la dependencia estadística, y por lo tanto, valores de MI mayores podrían implicar una relación de regulación más restrictiva entre un número más reducido de miRs sobre una mayor cantidad de genes. Una mayor cantidad de miRs y de interacciones miR-miR en la red de los datos de tumor podría afectar la especificidad de las relaciones de regulación que en otro caso serían altamente organizadas [Vidigal y Ventura, 2015], favoreciendo la señalización asociada a la plasticidad y heterogeneidad característica de las células cancerosas [Banerji *et al.*, 2013].

Los parámetros de las redes sugieren la participación de los miR en la cohesión de las mismas. Los miR unen componentes pequeños e incorporan genes que sólo se mantienen unidos a la red por medio de interacciones del tipo miR-gen. Al remover los miR presentes en la red, miles de genes se convierten en nodos aislados y muchos componentes aparecen (Figura 5-6 a,b). Tener menos interacciones del tipo gen-gen y valores de MI menores para dichas asociaciones aumenta la susceptibilidad de la red a desintegrarse cuando los miR son removidos, ya que en las redes inferidas a partir de los datos de tumor se obtienen casi el doble de componentes en comparación con el control.

La especificidad afectada debido al aumento de interacciones entre miRs y genes, y la susceptibilidad a la desintegración presente en la red de datos de tumor, parecen relacionarse con la gran cantidad de términos GO que presentaron enriquecimiento en los datos asociados a estas redes en comparación con el análisis control. Los términos enriquecidos en el análisis de las redes de tumor se asocian con mecanismos como la promoción y supervivencia tumoral, especialmente aquellos términos relacionados a las interacciones con la matriz extracelular [Gilkes *et al.*, 2014]. por otro lado, los resultados del enriquecimiento asociados al análisis de la red de datos control muestran términos asociados a la homeostasis tisular y el mantenimiento celular (Tabla 2). El trabajo aquí presentado intenta resaltar la importancia de la inclusión de estos RNAs pequeños en la construcción de las redes de regulación génica debido a que sus propiedades parecen ser cruciales para entender los mecanismos de la regulación transcripcional que pueden influir al cáncer de mama.

Un análisis más detallado, basado en los nodos con los grados más altos, reveló que miR-200 y miR-199 son determinantes para la estructura de las redes inferidas a partir de los datos control y de

tumor, respectivamente. A pesar de que los miR que conforman la familia miR-200 están presentes tanto en las redes de datos control y de tumor, éstos muestran importantes diferencias en cuanto a sus asociaciones (Figura 5-7 a,b). Por un lado, los primeros vecinos de miR-200 determinan la estructura global en la red de datos control, y por el otro, la red obtenida de los datos de tumor, aunque contiene a todos los miR que constituyen a miR-200 su estructura está determinada por las asociaciones de miR-199.

Con respecto a su funcionalidad, las diferencias de conectividad antes mencionadas también se ven reflejadas en los resultados del análisis de enriquecimiento. Para la red de miR-200 inferida a partir de los datos control podemos observar que los términos están asociados a la adhesión celular, la actividad tirosina cinasa y las uniones celulares (Tablas 13). Mientras tanto, los procesos enriquecidos para el análisis de la red de datos de tumor incluyen términos relacionados a la regulación del espacio extracelular, la unión de moléculas, respuesta inmune y procesos del desarrollo. También en la red de primeros vecinos de miR-200 de datos de tumor enriqueció para una mayor cantidad de términos que el análisis con los datos de la red control, aunque el número de genes en el análisis de los datos de tumor es mucho menor. Además, los miR de miR-199 y sus primeros vecinos mostraron resultados de enriquecimiento similares a los obtenidos para miR-200, en cuanto a los términos relacionados con el desarrollo y la matriz extracelular. Esto podría sugerir una relación cooperativa entre miR-199, miR-200 y sus primeros vecinos en la red de datos de tumor.

Nuestras redes de primeros vecinos de miR-200 inferidas a partir de los datos control y de tumor presentan un núcleo común de interacciones conservadas entre los dos fenotipos (Figura 5-8). En la Figura 5-8 se puede observar que el núcleo común está formado por los miRs de miR-200 y una serie de factores de crecimiento entre los que se encuentran: *TWIST-1*, *TWIST-2*, *ZEB-1* y *ZEB-2*. La presencia de este núcleo conservado entre las diferencias de los fenotipos parecer ser fundamental debido a que los miR y genes que lo componen han sido asociados con la supresión de la ganancia de características mesenquimales en células epiteliales mediante un proceso conocido como transición Epitelio-Mesénquima (EMT) [Zhu *et al.*, 2016; Park *et al.*, 2008; Mani *et al.*, 2008], y el proceso contrario conocido como transición Mesénquima-Epitelio (MET) [Park *et al.*, 2008]. La EMT y sus procesos relacionados han demostrado ser importantes para la metastásis, la invasividad, y la supresión inmune en el cáncer [Zheng y Kang, 2014], así como para adquirir funcionalidad troncal [Mani *et al.*, 2008].

En el núcleo común entre las redes de miR-200 están altamente sobreexpresados mientras que la mayoría de los genes y factores de transcripción presentes están subexpresados (Figura 5-8). Este

perfil de expresión parece sugerir la interacción canónica entre miR-200 y sus posibles blancos, sin embargo nuestras redes son de naturaleza no dirigida por lo que no es posible obtener la dirección de las interacciones únicamente por medio de la inferencia de la red. El perfil de expresión particular que muestra este núcleo sugiere un papel crucial para estos miR y sus asociaciones en la regulación de los mecanismos relacionados al cáncer de mama.

La sobreexpresión de miR-200 que presentan nuestras muestras mantiene una relación con MET [Park *et al.*, 2008] y la adquisición de características epiteliales, los cuales son pasos clave para la colonización tumoral [Korpal *et al.*, 2011; Gunasinghe *et al.*, 2012], pluripotencialidad y la expresión de genes de auto-regeneración [Celià-Terrassa *et al.*, 2012]. El hecho de que nuestros datos provienen de muestras de tumores primarios de mama exclusivamente (donde existe evidencia de que miR-200 se encuentra sobreexpresado) y que miR-200 se encuentra subexpresado en células cancerosas que han adquirido características mesenquimales [Park *et al.*, 2008], nos sugiere que por medio de un abordaje de redes fuimos capaces de inferir un programa de regulación en el cual miR-200 promueve la tumorigénesis y colonización al favorecer las características epiteliales en el tumor (MET) [Becker *et al.*, 2014; Dykxhoorn *et al.*, 2009]. Lo que resalta la importancia de la regulación de miR-200 en el cáncer de mama.

Los resultados obtenidos sugieren un comportamiento dual entre miR-200 y sus moléculas relacionadas en la regulación de los procesos de EMT y MET. En la Figura 5-8 se puede observar que miR-141 contribuye con la mayoría de las interacciones entre miR-200 y los factores de transcripción asociados. También se puede observar que marcadores clásicos de EMT, tales como *VIM*, *ZEB-1* y *ZEB-2* están sub expresados, por lo que miR-141 podría estar relacionado con la adquisición de características epiteliales. El núcleo y su perfil de expresión son instancias de la dualidad de EMT/MET, la cual también se ve reflejada en las diferencias estructurales que presentan las redes de primeros vecinos de miR-200 que se infirieron a partir de los datos control y de tumor. Esto también podría estar asociado con el programa conocido como Plasticidad Epitelio Mesénquima [Ye *et al.*, 2014; Tsai y Yang, 2013] y su participación en cáncer, el cual incluye la relación entre EMT, MET y sus estados intermedios.

Junto con miR-200, miR-199 parece estar involucrado con la promoción del cáncer. Se ha reportado que miR-199 es regulado por *Twist-1* durante el desarrollo [Lee *et al.*, 2009], esta asociación está presente únicamente en la red inferida a partir de los datos de tumor junto asociaciones con genes como *ZEB-1*, *SNAI-2* y *VIM*. Además de estar relacionado a los marcadores tradicionales de EMT antes

mencionados, hay evidencia de que miR-199a/b se asocia con la proliferación, migración y adhesión [Mudduluru *et al.*, 2011]; sugiriendo un rol central en la regulación de EMT/MET para miR-199a/b [Suzuki *et al.*, 2014].

Otra característica relevante obtenida a partir de este análisis, es que la red inferida a partir de los datos de tumor contiene una cantidad importante de miRs que mapean al cluster *DLK1-DIO3* (Delta-like 1 homolog-deiodinase, iodothyronine 3) en el locus 14q32. El cluster *DLK1-DIO3* se encuentra regulado bajo impronta, esta región contiene a los genes paternalmente expresados *DLK1*, *RTL1* y *DIO3* y a una serie de ncRNAs maternalmente expresados (lncRNA, miR, snoRNA y pseudogenes) [Benetatos *et al.*, 2013]. Además de ser sujeto a impronta parental, ésta región está asociada al complejo de represión polycomb 2 [Kaneko *et al.*, 2014]. La región *DLK1-DIO3* se encuentra conservada entre mamíferos y sus transcritos relacionados con procesos del desarrollo son capaces de afectar tanto el crecimiento y diferenciación celular [Mo *et al.*, 2015], como afectar procesos que modulan enfermedades degenerativas [Stelzer *et al.*, 2014], neurológicas y metabólicas [Benetatos *et al.*, 2013]. Por lo que la función y regulación de la región *DLK1-DIO3* ha demostrado estar implicada en múltiples patologías humanas y el cáncer [Benetatos *et al.*, 2013; Lehner *et al.*, 2014; Valdmanis *et al.*, 2015].

Es conocido que los miR presentes en la región *DLK1-DIO3* participan en la supresión de EMT. El mecanismo que ha sido propuesto incluye la represión de la señalización que involucra a la familia miR-200, *ZEB-1/2* y *TWIST-1* [Haga y Phinney, 2012]. Por esta razón los miRs del cluster *DLK1-DIO3* han sido etiquetados principalmente como supresores de tumor. Sin embargo, también se ha reportado que estos miRs se encuentran sobreexpresados en los líneas celulares de cáncer de mama y muestras de cáncer de mama [Haga y Phinney, 2012]. Además también se ha reportado la subexpresión de estos miR en los tumores epiteliales [Zhang *et al.*, 2008], y en etapas tempranas de la reprogramación celular [Henzler *et al.*, 2013]; por lo que se han asumido que su subexpresión podría ser importante para mejorar la eficiencia de la reprogramación. En concordancia con lo reportado en la literatura, estos miRs están principalmente sub expresados en nuestros datos (Figura 5-12), sugiriendo de nuevo una relación importante entre el perfil de expresión de los miRs en la red inferida a partir de los datos de tumor y EMT/MET.

Nuestros resultados sugieren que los miRs de del cluster *DLK1-DIO3* y miR-200 están relacionados con la regulación de EMT/MET. Decidimos extender nuestro análisis de resultados obtenidos a partir de información a nivel de genes a información de vías de señalización relevantes. Utilizamos Pathifier [Drier *et al.*, 2013], debido a que esta herramienta nos permite obtener información

del estado de deregulación de las vías por medio de los datos de expresión de los genes asociados a los pacientes en nuestro análisis de una manera contexto-dependiente. Basándonos en los resultados obtenidos por nuestras redes, encontramos vías de señalización de reguladas directamente relacionadas a EMT en tres bases de datos distintas. Estos resultados muestran la presencia de un proceso específico contenido en los datos de los pacientes, que logramos reconstruir a partir de nuestro análisis basado en redes.

La presente metodología integra diferentes niveles de información para crear un modelo robusto que describe el paisaje de regulación de los miR en el cáncer de mama. Aunque no podemos inferir la dirección de las asociaciones o asegurar que en efecto son interacciones entre miR-blanco, logramos reconstruir de manera exitosa asociaciones que son críticas para la promoción del crecimiento y diseminación tumoral. Como muestran nuestros resultados, además de las relaciones evidentes entre los nodos más altamente conectados junto con los miR en cluster y EMT/MET, el comportamiento general de los miRs en la red también sugieren mecanismos responsables por las diferencias fenotípicas presentes entre las muestras control y tumorales.

Este enfoque nos permitió identificar diferentes oncomiRs relacionados con la promoción y supresión del cáncer, dependiendo de su perfil de expresión. Algunos de ellos están presentes junto con sus respectivas referencias en la Tabla 8. Aunque para algunas de las interacciones con los oncomiRs existe evidencia de su papel en el cáncer de mama (como en el caso de miR-145, miR-100, miR-379 y miR-493), para la mayoría, su participación particular aún no ha sido estudiada. La consistencia entre el perfil de expresión y la asociación de los blancos que se muestra en la Tabla 8, junto con los resultados de los miRs de mayor grado de nuestras redes inferidas resulta importante para el estudio de las asociaciones de regulación de los miR, debido a que las bases de datos de predicción de secuencia pueden estar sesgadas y las bases de datos de interacciones validadas experimentalmente aún no están completas. Aun así fuimos capaces de encontrar interacciones reportadas previamente en dichas bases de datos, reforzando la utilidad de aplicar este enfoque para el estudio de la regulación transcripcional en cáncer de mama (Tablas 17 y 18).

Parte V

Conclusiones

Las redes presentadas son una reconstrucción probabilística de las asociaciones entre miRs y genes por medio de sus perfiles de expresión. Considerando que las asociaciones que estamos reportando se obtienen a través de un algoritmo por medio de datos de secuenciación, el hecho de que pudiéramos encontrar asociaciones que se han reportado en bases de datos le confiere confianza a nuestros resultados. Tenemos que considerar que debido a las limitaciones en la información disponible no es posible hacer un análisis de falsos positivos para evaluar la inferencia de nuestras redes. Sin embargo, dados los resultados obtenidos éstas han resultado útiles para describir las asociaciones no sólo entre los genes y los miR, sino su comportamiento entre ellos.

En las redes inferidas las interacciones entre los genes y los miR representan asociaciones que son independientes de sus niveles de expresión individuales sino que reflejan el paisaje regulatorio presente en cada fenotipo. Las interacciones entre miRs se relacionan con su expresión por cluster y muchas de ellas mapean a su clasificación por familias, siendo miR-200 el ejemplo más claro de este comportamiento. Las interacciones miR-gen pueden representar relaciones directas e indirectas miR-blanco e interacciones del tipo factor de transcripción-miR. Aunque las asociaciones que obtenemos no tienen direccionalidad, el hecho de que hayamos encontrado interacciones que son consistentes con la información de las bases de datos nos muestra que tienen relevancia funcional, por lo que introducir información genómica que capture otros mecanismos de regulación génica podría enriquecer aun más la información contenida en este tipo de redes.

La expresión alterada de los miR en nuestra redes inferidas tiene un efecto complejo sobre la regulación de los genes. Estos efectos parecen impactar la topología de la red resaltando el papel central que juegan la EMT/MET y sus miR relacionados en el cáncer de mama, entrando en un nivel de descripción más profundo, la información que obtuvimos a partir de el análisis de las vías de señalización junto con el análisis de expresión diferencial nos muestra un panorama en el cual todos estos mecanismos asociados a invasividad, metastásis, plasticidad fenotípica y resistencia convergen en el análisis.

Este enfoque basado en redes de Información Mutua resalta cómo un análisis centrado en la reconstrucción del panorama celular a partir de experimentos de expresión es útil para entender la regulación de los miR sobre procesos relacionados al cáncer. Lo cual resulta especialmente útil debido a que no requiere de ninguna información previa sobre posibles blancos y/o datos de reconocimiento de secuencia, para proporcionar información valiosa en cuanto a la regulación de los miR que podría ser difícil de obtener de otra manera. Lo que convierte esta estrategia en una herramienta beneficiosa para el diseño de experimentos en el futuro.

Dados los resultados obtenidos se espera aplicar la metodología aquí propuesta para estudiar los mecanismos regulatorios en los que participan los miRs, ahora se propone continuar con otro análisis computacional buscando un tipo de regulación más fina centrándose en los subtipos moleculares de cáncer de mama. Una de las principales limitaciones de la inferencia de redes por información mutua es la cantidad de muestras necesaria para el análisis, por lo que esperamos que conforme aumente la cantidad de datos públicos disponibles de secuenciación sea posible analizar diferentes conjuntos de datos y distintos tipos de cáncer bajo la misma perspectiva, permitiéndonos entender los mecanismos de regulación transcripcional en estas enfermedades.

Parte VI

Apéndices

Apéndice A

Tablas con las propiedades de las redes con variaciones en su construcción

Tabla A1. Atributos de una variación de las redes inferidas a partir de los datos de tumor y los datos control. La red aquí presentada se construyó ajustando el p-valor original de las asociaciones de los genes para corresponder a una cantidad igual de interacciones de asociaciones miR-miR y miR-gen y de interacciones de interacciones gen-gen.

Atributo	Red completa		Componente gigante		Subred CG gen-gen	
	Control	Tumor	Control	Tumor	Control	Tumor
Nodos totales	5,000	4,966	4,624	4,375	4,501	3,888
miR	241	514	123	487	0	0
gen	4,759	4,452	4,501	3,888	4,501	3,888
interacciones totales	41,572	33,186	41,036	32,747	24,127	16,106
miR-miR	482	1,775	230	1,769	0	0
miR-gen	16,777	14,908	16,679	14,872	0	0
gen-gen	24,313	16,503	24,127	16,106	24,127	16,106
Componentes	102	241	1	1	1,489	2,325
Nodos solos	0	0	0	0	1,449	2,238

CG: Componente gigante

Tabla A2. Atributos de una variación de las redes inferidas a partir de los datos de tumor y los datos control. La red aquí presentada se construyó ajustando el p-valor original de las asociaciones de los miR para corresponder a una cantidad igual de interacciones de asociaciones miR-miR y miR-gen y de interacciones de interacciones gen-gen.

Atributo	Red completa		Componente gigante		Subred CG gen-gen	
	Control	Tumor	Control	Tumor	Control	Tumor
Nodos totales	3,875	3,363	3,501	2,855	3,419	2,461
miR	202	455	82	394	0	0
gen	3,673	2,908	3,419	2,461	3,419	2,461
interacciones totales	25,905	21,231	25,502	20,874	14,476	10,964
miR-miR	327	1,174	155	1,140	0	0
miR-gen	10,941	8,819	10,871	8,770	0	0
gen-gen	14,637	11,238	14,476	10,964	14,476	10,964
Componentes	122	204	1	1	1,189	1,379
Nodos solos	0	0	0	0	1,157	1,338

CG: Componente gigante

Tabla A3. Atributos de una variación de las redes inferidas a partir de los datos de tumor y los datos control. La red aquí presentada se construyó ajustando el p-valor original de las asociaciones de los genes y los miRs para corresponder a la disminución del número de interacciones en un orden de magnitud.

Atributo	Red completa		Componente gigante		Subred CG gen-gen	
	Control	Tumor	Control	Tumor	Control	Tumor
Nodos totales	1,096	1,036	823	682	794	610
miR	89	250	29	72	0	0
gen	1,007	786	794	610	794	610
interacciones totales	3,897	3,566	3,651	3,191	1,315	1,362
miR-miR	103	310	52	158	0	0
miR-gen	2,304	1,770	2,284	1,671	0	0
gen-gen	1,490	1,486	1,315	1,362	1,315	1,362
Componentes	84	108	1	1	480	380
Nodos solos	0	0	0	0	460	372

CG: Componente gigante

Tabla A4. Atributos de una variación de las redes inferidas a partir de los datos de tumor y los datos control. La red aquí presentada se construyó ajustando el p-valor original de las asociaciones de los genes y los miRs para corresponder al aumento del número de interacciones en un orden de magnitud.

Atributo	Red completa		Componente gigante		Subred CG gen-gen	
	Control	Tumor	Control	Tumor	Control	Tumor
Nodos totales	10,538	15,428	10,439	15,422	9,925	14,789
miR	519	514	514	633	0	0
gen	10,019	14,795	9,925	14,789	9,925	14,789
interacciones totales	229,249	251,726	229,192	251,723	121,354	65,051
miR-miR	3,588	8,934	3,587	8,934	0	0
miR-gen	104,254	177,741	104,251	177,741	0	0
gen-gen	121,407	65,051	121,354	65,048	121,354	65,048
Componentes	43	4	1	1	3,646	7,404
Nodos solos	0	0	0	0	3,569	7,077

CG: Componente gigante

Apéndice B

Tablas con nodos con mayor grado para las redes con variaciones en su construcción

Tabla B1. Los 10 nodos con grados más altos en las variaciones de las redes inferidas a partir de los datos control y de tumor. Los grados aquí presentados pertenecen a la red que construyó ajustando el p-valor original de las asociaciones de los genes para corresponder a una cantidad igual de interacciones de asociaciones miR-miR y miR-gen y de interacciones de interacciones gen-gen.

Control		Tumor	
Nodo	Grado	Nodo	Grado
hsa-miR-200b.MIMAT0000318	1,272	hsa-let-7c.MIMAT0000064	445
hsa-miR-200a.MIMAT0000682	1,247	hsa-miR-199b.MIMAT0000263	435
hsa-miR-141.MIMAT0004598	1,192	hsa-miR-199a.MIMAT0000231	433
hsa-miR-141.MIMAT0000432	1,170	hsa-miR-337.MIMAT0000754	387
hsa-miR-193b.MIMAT0004767	1,025	hsa-miR-99a.MIMAT0000097	366
hsa-miR-200a.MIMAT0001620	1,002	hsa-miR-134.MIMAT0000447	299
hsa-miR-200c.MIMAT0000617	997	hsa-miR-382.MIMAT0000737	293
hsa-miR-22.MIMAT0000077	681	hsa-let-7i.MIMAT0004585	267
hsa-miR-652.MIMAT0003322	640	hsa-miR-199a.MIMAT0000232	262
hsa-miR-200c.MIMAT0004657	625	hsa-miR-199b.MIMAT0004563	260

Tabla B2. Los 10 nodos con grados más altos en las variaciones de las redes inferidas a partir de los datos control y de tumor. Los grados aquí presentados pertenecen a la red que construyó ajustando el p-valor original de las asociaciones de los miR para corresponder a una cantidad igual de interacciones de asociaciones miR-miR y miR-gen y de interacciones de interacciones gen-gen.

Control		Tumor	
Nodo	Grado	Nodo	Grado
hsa-miR-200a.MIMAT0000682	972	hsa-miR-199b.MIMAT0000263	327
hsa-miR-200b.MIMAT0000318	955	hsa-miR-199a.MIMAT0000231	327
hsa-miR-141.MIMAT0004598	949	hsa-miR-337.MIMAT0000754	319
hsa-miR-141.MIMAT0000432	875	hsa-let-7c.MIMAT0000064	311
hsa-miR-200a.MIMAT0001620	723	hsa-miR-99a.MIMAT0000097	256
hsa-miR-193b.MIMAT0004767	691	hsa-miR-134.MIMAT0000447	238
hsa-miR-200c.MIMAT0000617	641	hsa-miR-199a.MIMAT0000232	234
hsa-miR-22.MIMAT0000077	496	hsa-miR-199b.MIMAT0004563	233
hsa-miR-652.MIMAT0003322	456	hsa-miR-223.MIMAT0000280	214
hsa-miR-378a.MIMAT0000731	399	hsa-miR-150.MIMAT0000451	203

Tabla B3. Los 10 nodos con grados más altos en las variaciones de las redes inferidas a partir de los datos control y de tumor. Los grados aquí presentados pertenecen a la red que construyó ajustando el p-valor original de las asociaciones de los genes y los miRs para corresponder a un disminución del número de interacciones en un orden de magnitud.

Control		Tumor	
Nodo	Grado	Nodo	Grado
hsa-miR-200a.MIMAT0000682	231	hsa-miR-150.MIMAT0000451	181
hsa-miR-224.MIMAT0000281	215	hsa-miR-142.MIMAT0000433	145
hsa-miR-652.MIMAT0003322	213	hsa-miR-155.MIMAT0000646	126
hsa-miR-141.MIMAT0004598	198	hsa-miR-199a.MIMAT0000232	111
hsa-miR-200b.MIMAT0000318	197	hsa-miR-199b.MIMAT0004563	110
hsa-miR-141.MIMAT0000432	191	hsa-miR-146a.MIMAT0000449	103
hsa-miR-452.MIMAT0001635	160	hsa-miR-134.MIMAT0000447	99
hsa-miR-378a.MIMAT0000731	148	hsa-miR-199b.MIMAT0000263	95
hsa-miR-200a.MIMAT0001620	120	hsa-miR-199a.MIMAT0000231	93
hsa-miR-224.MIMAT0009198	86	hsa-miR-337.MIMAT0000754	87

Tabla B4. Los 10 nodos con grados más altos en las variaciones de las redes inferidas a partir de los datos control y de tumor. Los grados aquí presentados pertenecen a la red que construyó ajustando el p-valor original de las asociaciones de los genes y los miRs para corresponder al aumento del número de interacciones en un orden de magnitud.

Control		Tumor	
Nodo	Grado	Nodo	Grado
hsa-miR-193b.MIMAT0004767	2,563	hsa-miR-190b.MIMAT0004929	1,771
hsa-miR-200a.MIMAT0000682	2,469	hsa-miR-199a.MIMAT0000231	1,668
hsa-miR-200b.MIMAT0000318	2,408	hsa-miR-29b-2.MIMAT0004515	1,590
hsa-miR-141.MIMAT0004598	2,379	hsa-miR-199b.MIMAT0000263	1,475
hsa-miR-200c.MIMAT0000617	2,348	hsa-miR-337.MIMAT0000754	1,392
hsa-miR-141.MIMAT0000432	2,280	hsa-miR-18a.MIMAT0002891	1,370
hsa-miR-200a.MIMAT0001620	2,260	hsa-miR-452.MIMAT0001636	1,306
hsa-miR-200c.MIMAT0004657	2,,193	hsa-miR-382.MIMAT0000737	1,246
hsa-miR-652.MIMAT0003322	2,014	hsa-miR-18a.MIMAT0000072	1,228
hsa-miR-146b.MIMAT0002809	1,984	hsa-miR-29c.MIMAT0004673	1,173

Apéndice C

**Mapas de calor de los análisis de
deregulación de vías para las vías de
señalización asociadas a las redes de
miR-200 y *DLK1-DIO3***

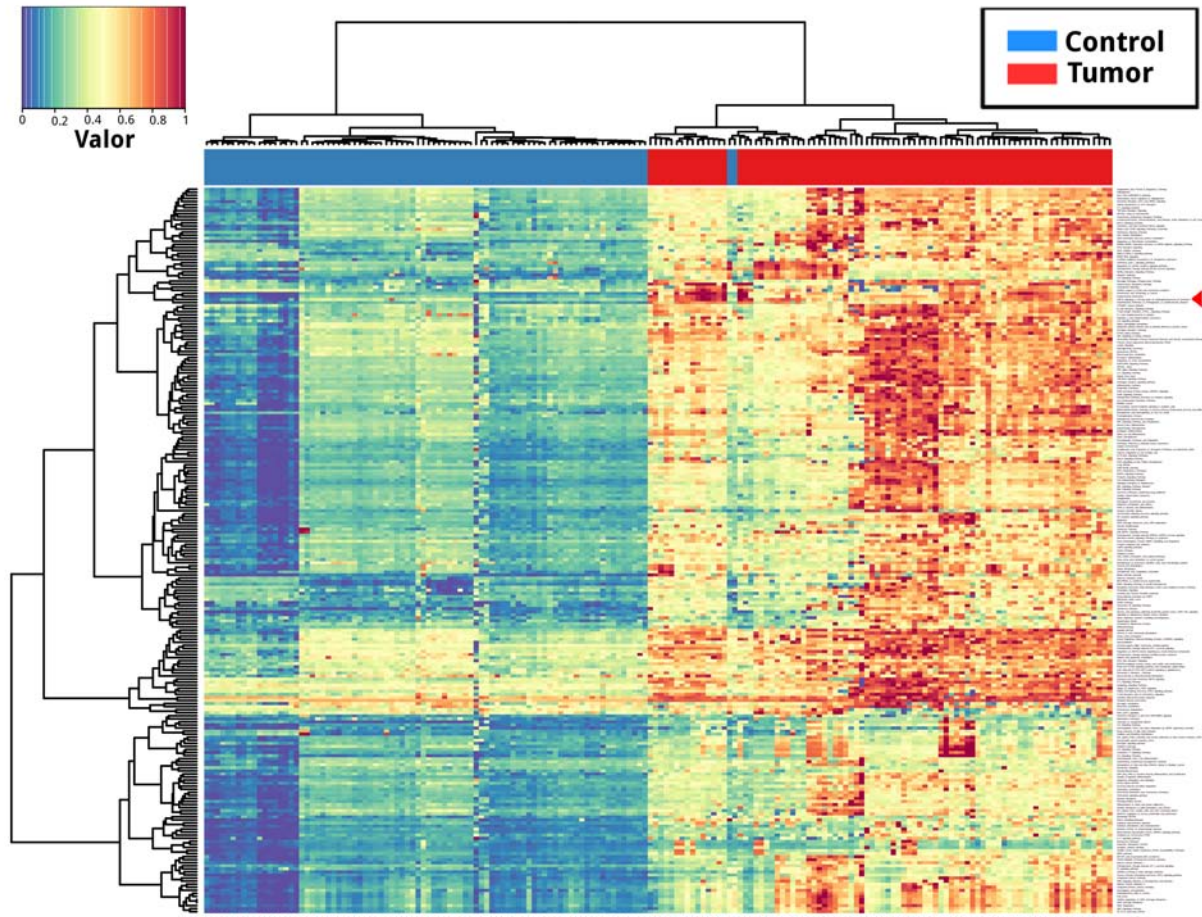


Figura C-1: Mapa de calor de los Score de deregulación de Vías (PDS) de las vías de Reactome que contienen genes asociados a los miR del cluster *DLK1-DIO3* en las redes inferidas a partir de los datos de tumor. En esta representación las filas corresponden a las vías de señalización seleccionadas para el análisis, y las columnas corresponden a las muestras sean tumorales o control. Los PDS obtenidos para las 237 vías de WikiPathways y las 172 muestras que trabajamos (86 muestras de tejido tumoral y 86 muestras de tejido control) se encuentran en las celdas del mapa de calor, coloreados de acuerdo a la escala de colores en la esquina superior izquierda. Lo que significa que las celdas en colores azules y verdes representan vías que se encuentran menos dereguladas con respecto al grupo control, mientras que las celdas rojas representan vías que se encuentran altamente dereguladas. La barra superior indica si las muestras son casos o controles. Se utilizaron distancias euclidianas y el método de agrupamiento jerárquico de Ward para crear el dendrograma. La flecha roja señala a la vía: señalización de TGF-Beta en células de tiroides para la Transición Epitelio Mesénquima.

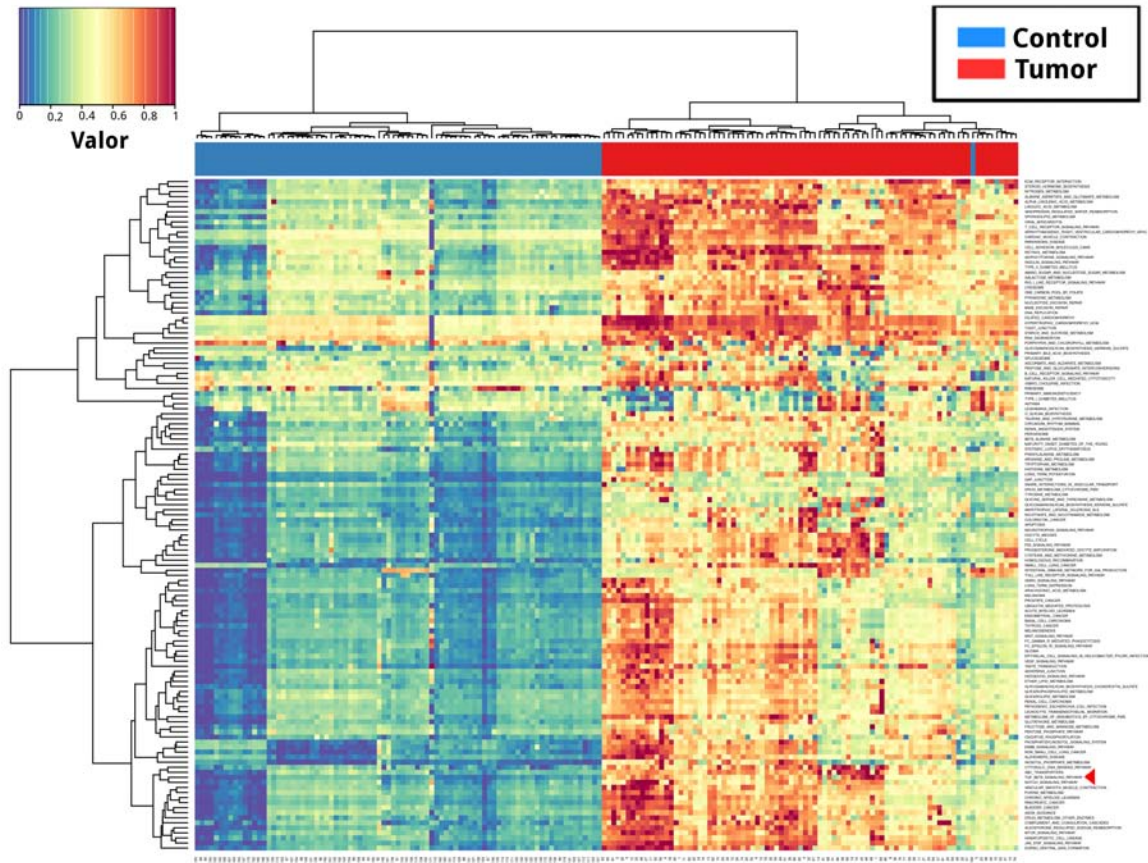


Figura C-2: Mapa de calor de los Score de deregulación de Vías (PDS) de las vías de Reactome que contienen genes asociados a los miR del cluster *DLK1-DIO3* en las redes inferidas a partir de los datos de tumor. En esta representación las filas corresponden a las vías de señalización seleccionadas para el análisis, y las columnas corresponden a las muestras sean tumorales o control. Los PDS obtenidos para las 133 vías de KEGG y las 172 muestras que trabajamos (86 muestras de tejido tumoral y 86 muestras de tejido control) se encuentran en las celdas del mapa de calor, coloreados de acuerdo a la escala de colores en la esquina superior izquierda. Lo que significa que las celdas en colores azules y verdes representan vías que se encuentran menos dereguladas con respecto al grupo control, mientras que las celdas rojas representan vías que se encuentran altamente dereguladas. La barra superior indica si las muestras son casos o controles. Se utilizaron distancias euclidianas y el método de agrupamiento jerárquico de Ward para crear el dendrograma. La flecha roja señala a la vía: señalización de TGF-Beta.

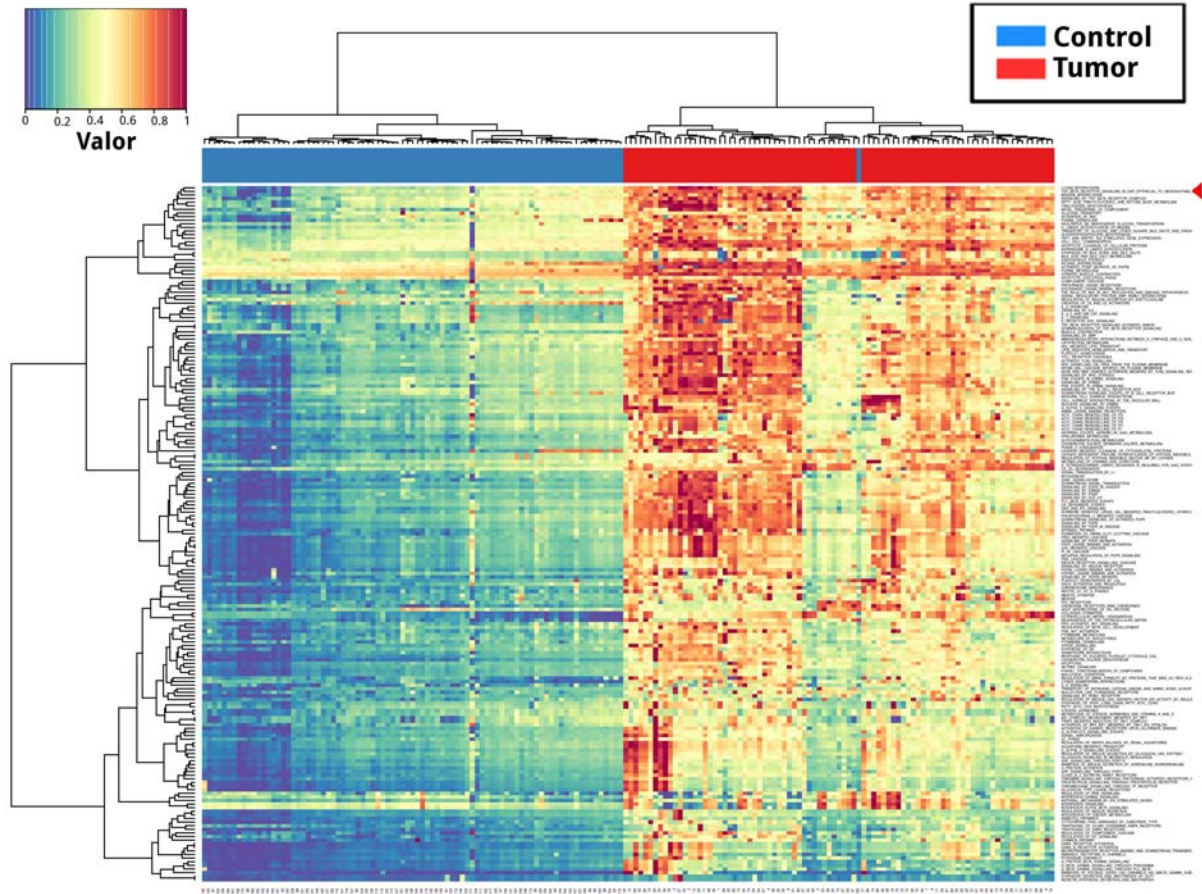


Figura C-3: Mapa de calor de los Score de deregulación de Vías (PDS) de las vías de Reactome que contienen genes presentes en la subred de primeros vecinos de miR-200 construidas a partir de las redes inferidas a partir de los datos de tumor. En esta representación las filas corresponden a las vías de señalización seleccionadas para el análisis, y las columnas corresponden a las muestras sean tumorales o control. Los PDS obtenidos para las 193 vías de Reactome y las 172 muestras que trabajamos (86 muestras de tejido tumoral y 86 muestras de tejido control) se encuentran en las celdas del mapa de calor, coloreados de acuerdo a la escala de colores en la esquina superior izquierda. Lo que significa que las celdas en colores azules y verdes representan vías que se encuentran menos dereguladas con respecto al grupo control, mientras que las celdas rojas representan vías que se encuentran altamente dereguladas. La barra superior indica si las muestras son casos o controles. Se utilizaron distancias euclidianas y el método de agrupamiento jerárquico de Ward para crear el dendrograma. La flecha roja señala a la vía: señalización del receptor TGF-Beta en EMT (Transición epitelio mesénquima).

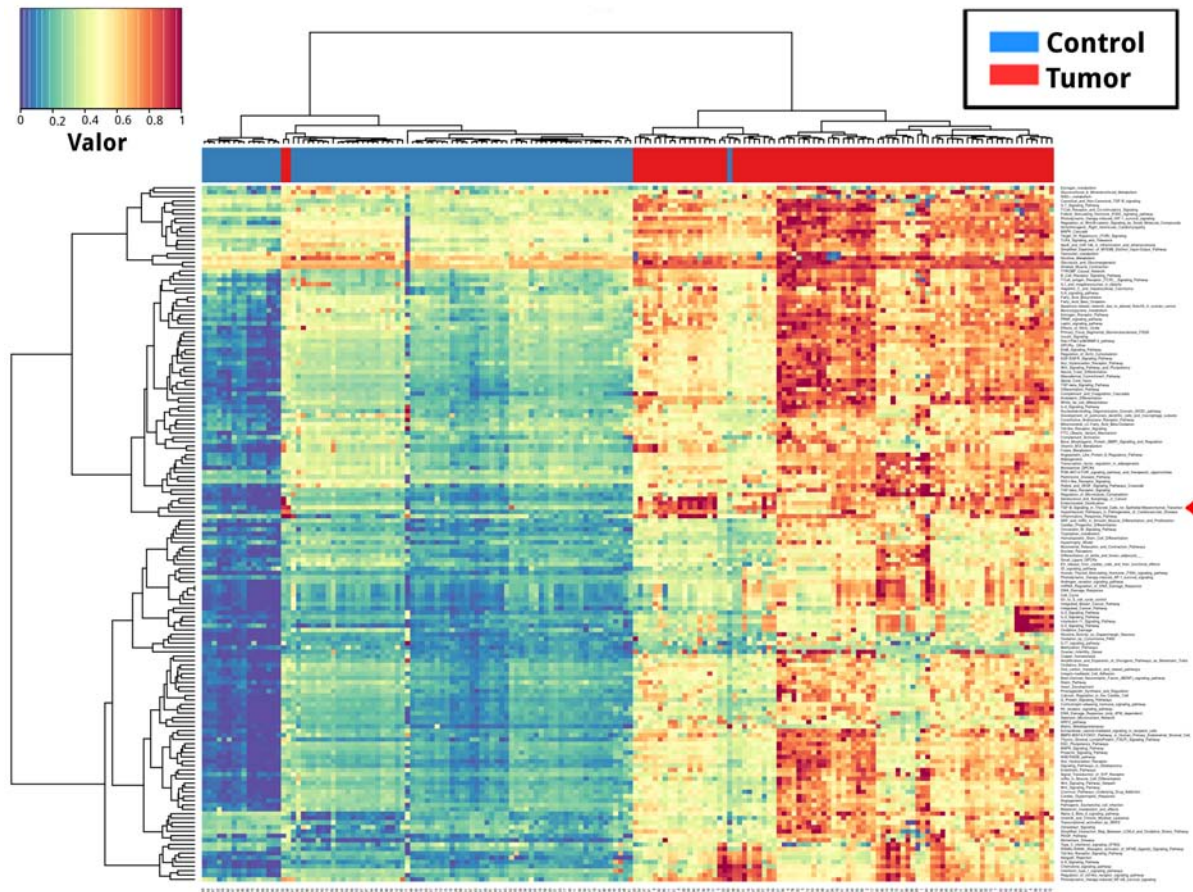


Figura C-4: Mapa de calor de los Score de deregulación de Vías (PDS) de las vías de Reactome que contienen genes presentes en la subred de primeros vecinos de miR-200 construidas a partir de las redes inferidas a partir de los datos de tumor. En esta representación las filas corresponden a las vías de señalización seleccionadas para el análisis, y las columnas corresponden a las muestras sean tumorales o control. Los PDS obtenidos para las 59 vías de WikiPathways y las 172 muestras que trabajamos (86 muestras de tejido tumoral y 86 muestras de tejido control) se encuentran en las celdas del mapa de calor, coloreados de acuerdo a la escala de colores en la esquina superior izquierda. Lo que significa que las celdas en colores azules y verdes representan vías que se encuentran menos dereguladas con respecto al grupo control, mientras que las celdas rojas representan vías que se encuentran altamente dereguladas. La barra superior indica si las muestras son casos o controles. Se utilizaron distancias euclidianas y el método de agrupamiento jerárquico de Ward para crear el dendrograma. La flecha roja señala a la vía: señalización de TGF-Beta en células de tiroides para la Transición Epitelio Mesénquima.

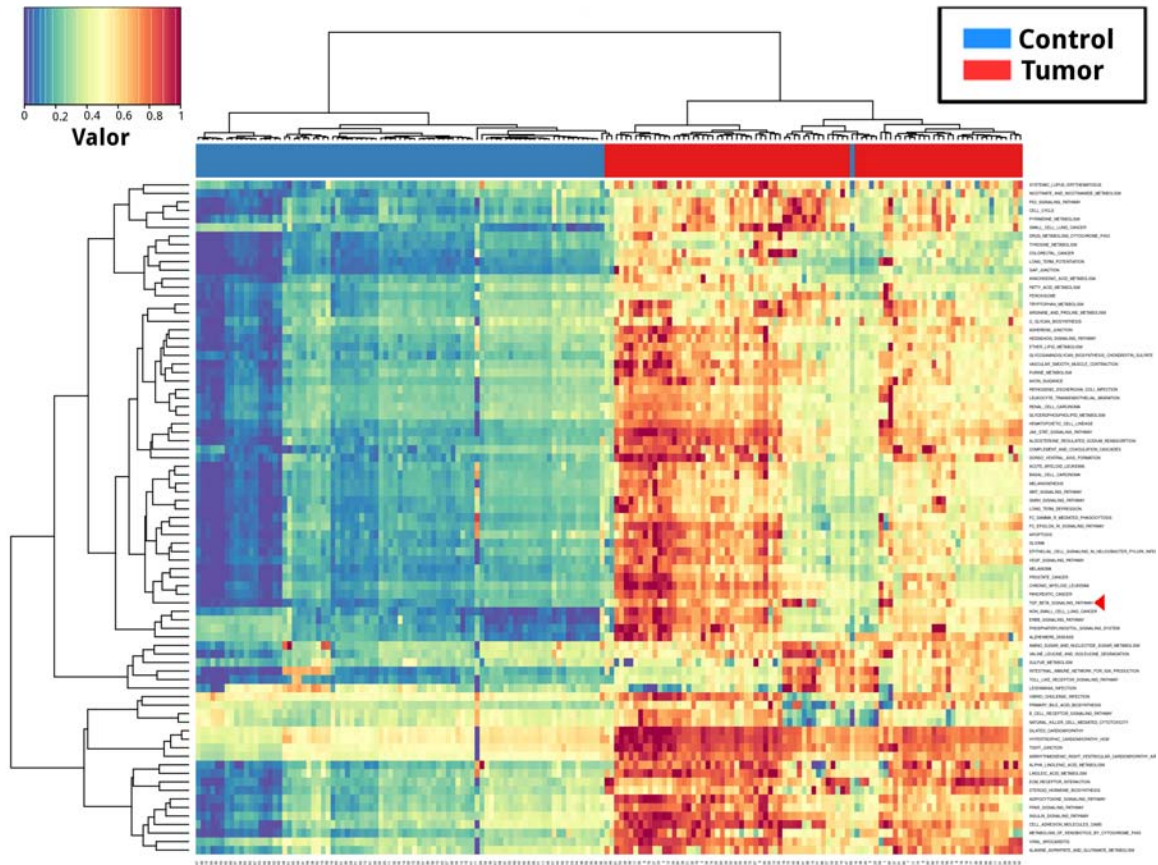


Figura C-5: Mapa de calor de los Score de deregulación de Vías (PDS) de las vías de Reactome que contienen genes presentes en la subred de primeros vecinos de miR-200 construidas a partir de las redes inferidas a partir de los datos de tumor. En esta representación las filas corresponden a las vías de señalización seleccionadas para el análisis, y las columnas corresponden a las muestras sean tumorales o control. Los PDS obtenidos para las 79 vías de KEGG y las 172 muestras que trabajamos (86 muestras de tejido tumoral y 86 muestras de tejido control) se encuentran en las celdas del mapa de calor, coloreados de acuerdo a la escala de colores en la esquina superior izquierda. Lo que significa que las celdas en colores azules y verdes representan vías que se encuentran menos dereguladas con respecto al grupo control, mientras que las celdas rojas representan vías que se encuentran altamente dereguladas. La barra superior indica si las muestras son casos o controles. Se utilizaron distancias euclidianas y el método de agrupamiento jerárquico de Ward para crear el dendrograma. La flecha roja señala a la vía: señalización de TGF-Beta.

Bibliografía

- ANDREWS, M.C., CURSONS, J., HURLEY, D.G., ANAKA, M., CEBON, J.S., BEHREN, A., Y CRAMPIN, E.J. Systems analysis identifies miR-29b regulation of invasiveness in melanoma. *Molecular cancer* **15**:72 (2016)
- BAEK, D., VILLÉN, J., SHIN, C., CAMARGO, F.D., GYGI, S.P., Y BARTEL, D.P. The impact of microRNAs on protein output. *Nature* **455**(7209):64–71 (2008)
- BANERJI, C.R., MIRANDA-SAAVEDRA, D., SEVERINI, S., WIDSCHWENDTER, M., ENVER, T., ZHOU, J.X., Y TESCHENDORFF, A.E. Cellular network entropy as the energy potential in Waddington's differentiation landscape. *arXiv preprint arXiv:1310.7083* (2013)
- BANSAL, M., BELCASTRO, V., AMBESI-IMPIOMBATO, A., Y DI BERNARDO, D. How to infer gene networks from expression profiles. *Molecular systems biology* **3**(1):78 (2007)
- BECKER, L.E., TAKWI, A.A.L., LU, Z., Y LI, Y. The role of miR-200a in mammalian epithelial cell transformation. *Carcinogenesis* pág. bgu202 (2014)
- BENETATOS, L., HATZIMICHAEL, E., LONDIN, E., VARTHOLOMATOS, G., LOHER, P., RIGOUTSOS, I., Y BRIASOULIS, E. The microRNAs within the DLK1-DIO3 genomic region: involvement in disease pathogenesis. *Cellular and Molecular Life Sciences* **70**(5):795–814 (2013)
- BOHNSACK, M.T., CZAPLINSKI, K., Y GÖRLICH, D. Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *Rna* **10**(2):185–191 (2004)
- BOO, L., HO, W.Y., ALI, N.M., YEAP, S.K., KY, H., CHAN, K.G., YIN, W.F., SATHARASINGHE, D.A., LIEW, W.C., TAN, S.W. *et al.* MiRNA Transcriptome Profiling of Spheroid-Enriched Cells with Cancer Stem Cell Properties in Human Breast MCF-7 Cell Line. *International Journal of Biological Sciences* **12**(4):427 (2016)

- CANCER GENOME ATLAS NETWORK AND OTHERS. Comprehensive molecular portraits of human breast tumours. *Nature* **490**(7418):61–70 (2012)
- CELIÀ-TERRASSA, T., MECA-CORTÉS, Ó., MATEO, F., DE PAZ, A.M., RUBIO, N., ARNAL-ESTAPÉ, A., ELL, B.J., BERMUDO, R., DÍAZ, A., GUERRA-REBOLLO, M. *et al.* Epithelial-mesenchymal transition can suppress major attributes of human epithelial tumor-initiating cells. *The Journal of clinical investigation* **122**(5):1849–1868 (2012)
- CHEN, C.Y., CHEN, S.T., FUH, C.S., JUAN, H.F., Y HUANG, H.C. Coregulation of transcription factors and microRNAs in human transcriptional regulatory network. *BMC bioinformatics* **12**(1):1 (2011)
- CHIN, K., DEVRIES, S., FRIDLAND, J., SPELLMAN, P.T., ROYDASGUPTA, R., KUO, W.L., LAPUK, A., NEVE, R.M., QIAN, Z., RYDER, T. *et al.* Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer cell* **10**(6):529–541 (2006)
- CHO, W.C. OncomiRs: the discovery and progress of microRNAs in cancers. *Molecular cancer* **6**(1):1 (2007)
- CHOU, C.H., CHANG, N.W., SHRESTHA, S., HSU, S.D., LIN, Y.L., LEE, W.H., YANG, C.D., HONG, H.C., WEI, T.Y., TU, S.J. *et al.* miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic acids research* **44**(D1):D239–D247 (2016)
- DAI, X., XIANG, L., LI, T., Y BAI, Z. Cancer Hallmarks, Biomarkers and Breast Cancer Molecular Subtypes. *Journal of Cancer* **7**(10):1281 (2016)
- DIAZ, G., ZAMBONI, F., TICE, A., Y FARCI, P. Integrated ordination of miRNA and mRNA expression profiles. *BMC genomics* **16**(1):1 (2015)
- DRIER, Y., SHEFFER, M., Y DOMANY, E. Pathway-based personalized analysis of cancer. *Proceedings of the National Academy of Sciences* **110**(16):6388–6393 (2013)
- DYKXHOORN, D.M., WU, Y., XIE, H., YU, F., LAL, A., PETROCCA, F., MARTINVALET, D., SONG, E., LIM, B., Y LIEBERMAN, J. miR-200 enhances mouse breast cancer cell colonization to form distant metastases. *PloS one* **4**(9):e7181 (2009)
- ESQUELA-KERSCHER, A. Y SLACK, F.J. Oncomirs - microRNAs with a role in cancer. *Nature Reviews Cancer* **6**(4):259–269 (2006)
- FEINBERG, A.P., KOLDOBSKIY, M.A., Y GÖNDÖR, A. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nature Reviews Genetics* (2016)

- FERLAY, J., SOERJOMATARAM, I., DIKSHIT, R., ESER, S., MATHERS, C., REBELO, M., PARKIN, D.M., FORMAN, D., Y BRAY, F. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *International journal of cancer* **136**(5):E359–E386 (2015)
- FORBES, S.A., BEARE, D., GUNASEKARAN, P., LEUNG, K., BINDAL, N., BOUTSELAKIS, H., DING, M., BAMFORD, S., COLE, C., WARD, S. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic acids research* **43**(D1):D805–D811 (2015)
- FRIEDMAN, R.C., FARH, K.K.H., BURGE, C.B., Y BARTEL, D.P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome research* **19**(1):92–105 (2009)
- GARCÍA-CAMPOS, M.A., ESPINAL-ENRÍQUEZ, J., Y HERNÁNDEZ-LEMUS, E. Pathway analysis: state of the art. *Frontiers in physiology* **6** (2015)
- GAROFALO, M. Y CROCE, C.M. microRNAs: Master regulators as potential therapeutics in cancer. *Annual review of pharmacology and toxicology* **51**:25–43 (2011)
- GILKES, D.M., SEMENZA, G.L., Y WIRTZ, D. Hypoxia and the extracellular matrix: drivers of tumour metastasis. *Nature Reviews Cancer* **14**(6):430–439 (2014)
- GOLDHIRSCH, A., WOOD, W., COATES, A., GELBER, R., THÜRLIMANN, B., SENN, H.J. *et al.* Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. *Annals of oncology* *pag.* *mdr304* (2011)
- GRIFFITHS-JONES, S., GROCOCK, R.J., VAN DONGEN, S., BATEMAN, A., Y ENRIGHT, A.J. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic acids research* **34**(suppl 1):D140–D144 (2006)
- GRIFFITHS-JONES, S., SAINI, H.K., VAN DONGEN, S., Y ENRIGHT, A.J. miRBase: tools for microRNA genomics. *Nucleic acids research* **36**(suppl 1):D154–D158 (2008)
- GUNASINGHE, N.D., WELLS, A., THOMPSON, E.W., Y HUGO, H.J. Mesenchymal–epithelial transition (MET) as a mechanism for metastatic colonisation in breast cancer. *Cancer and Metastasis Reviews* **31**(3-4):469–478 (2012)
- GURURAJAN, M., JOSSON, S., CHU, G.C.Y., LU, C.L., LU, Y.T., HAGA, C.L., ZHAU, H.E., LIU, C., LICHTERMAN, J., DUAN, P. *et al.* miR-154* and miR-379 in the DLK1-DIO3 microRNA mega-cluster regulate epithelial to mesenchymal transition and bone metastasis of prostate cancer. *Clinical Cancer Research* **20**(24):6559–6569 (2014)

- HA, M. Y KIM, V.N. Regulation of microRNA biogenesis. *Nat Rev Mol Cell Biol* **15**(8):509–524 (2014)
- HAGA, C.L. Y PHINNEY, D.G. MicroRNAs in the imprinted DLK1-DIO3 region repress the epithelial-to-mesenchymal transition by targeting the TWIST1 protein signaling network. *Journal of Biological Chemistry* **287**(51):42695–42707 (2012)
- HAN, J., LEE, Y., YEOM, K.H., KIM, Y.K., JIN, H., Y KIM, V.N. The Drosha-DGCR8 complex in primary microRNA processing. *Genes & development* **18**(24):3016–3027 (2004)
- HANAHAN, D. Y WEINBERG, R.A. The hallmarks of cancer. *cell* **100**(1):57–70 (2000)
- HANAHAN, D. Y WEINBERG, R.A. Hallmarks of Cancer: The Next Generation. *Cell* **144**(5):646 – 674 (2011)
- HASTIE, T. Y STUETZLE, W. Principal curves. *Journal of the American Statistical Association* **84**(406):502–516 (1989)
- HENZLER, C.M., LI, Z., DANG, J., ARCILA, M.L., ZHOU, H., LIU, J., CHANG, K.Y., BASSETT, D.S., RANA, T.M., Y KOSIK, K.S. Staged miRNA re-regulation patterns during reprogramming. *Genome biology* **14**(12):R149 (2013)
- HERNÁNDEZ-LEMUS, E. Y RANGEL-ESCARREÑO, C. The role of information theory in gene regulatory network inference. *Information Theory: New Research* págs. 109–144 (2011)
- HUA, L., LI, L., Y ZHOU, P. Identifying breast cancer subtype related miRNAs from two constructed miRNAs interaction networks in silico method. *BioMed research international* **2013**:798912 (2013)
- HUANG, G.T., ATHANASSIOU, C., Y BENOS, P.V. mirConnX: condition-specific mRNA-microRNA network integrator. *Nucleic acids research* **39**:W416–W423 (2011)
- HUANG, J.C., MORRIS, Q.D., Y FREY, B.J. Bayesian inference of MicroRNA targets from sequence and expression data. *Journal of Computational Biology* **14**(5):550–563 (2007)
- INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM AND OTHERS. Finishing the euchromatic sequence of the human genome. *Nature* **431**(7011):931–945 (2004)
- JANG, I.S., MARGOLIN, A., Y CALIFANO, A. hARACNe: improving the accuracy of regulatory model reverse engineering via higher-order data processing inequality tests. *Interface focus* **3**(4):20130011 (2013)

- JIANG, Q., HE, M., GUAN, S., MA, M., WU, H., YU, Z., JIANG, L., WANG, Y., ZONG, X., JIN, F. *et al.* MicroRNA-100 suppresses the migration and invasion of breast cancer cells by targeting FZD-8 and inhibiting Wnt/ β -catenin signaling pathway. *Tumor Biology* págs. 1–11 (2015)
- JONES, P.A. Overview of cancer epigenetics. En *Seminars in hematology*, tomo 42, págs. S3–S8. Elsevier (2005)
- JUNG, D., KIM, B., FREISHTAT, R.J., GIRI, M., HOFFMAN, E., Y SEO, J. miRTarVis: an interactive visual analysis tool for microRNA-mRNA expression profile data. *BMC proceedings* **9**:S2 (2015)
- KAMANU, T.K., RADOVANOVIC, A., ARCHER, J.A., Y BAJIC, V.B. Exploration of miRNA families for hypotheses generation. *Scientific reports* **3** (2013)
- KANEKO, S., BONASIO, R., SALDAÑA-MEYER, R., YOSHIDA, T., SON, J., NISHINO, K., UMEZAWA, A., Y REINBERG, D. Interactions between JARID2 and noncoding RNAs regulate PRC2 recruitment to chromatin. *Molecular cell* **53**(2):290–300 (2014)
- KHAN, S., BROUGHAM, C.L., RYAN, J., SAHRUDIN, A., O'NEILL, G., WALL, D., CURRAN, C., NEWELL, J., KERIN, M.J., Y DWYER, R.M. miR-379 regulates cyclin B1 expression and is decreased in breast cancer. *PLoS one* **8**(7):e68753 (2013)
- KIM, V.N. MicroRNA biogenesis: coordinated cropping and dicing. *Nature reviews Molecular cell biology* **6**(5):376–385 (2005)
- KNAUL, F.M., NIGENDA, G., LOZANO, R., ARREOLA-ORNELAS, H., LANGER, A., Y FRENK, J. Cáncer de mama en México: una prioridad apremiante. *Salud pública de México* **51**:s335–s344 (2009)
- KORPAL, M., ELL, B.J., BUFFA, F.M., IBRAHIM, T., BLANCO, M.A., CELIÀ-TERRASSA, T., MERCATALI, L., KHAN, Z., GOODARZI, H., HUA, Y., WEI, Y., HU, G., GARCIA, B.A., RAGOISSIS, J., AMADORI, D., HARRIS, A.L., Y KANG, Y. Direct targeting of Sec23a by miR-200s influences cancer cell secretome and promotes metastatic colonization. *Nat Med* **17**(9):1101–1108 (2011)
- KOZOMARA, A. Y GRIFFITHS-JONES, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic acids research* **42**(D1):D68–D73 (2014)
- KREEGER, P.K. Y LAUFFENBURGER, D.A. Cancer systems biology: a network modeling perspective. *Carcinogenesis* **31**(1):2–8 (2010)
- KRZYWINSKI, M., BIROL, I., JONES, S.J.M., Y MARRA, M.A. Hive plots—rational approach to visualizing networks. *Brief Bioinform* **13**(5):627–644 (2012)

- KUTMON, M., LOTIA, S., EVELO, C.T., Y PICO, A.R. WikiPathways App for Cytoscape: making biological pathways amenable to network analysis and visualization. *F1000Research* **3** (2014)
- LADDHA, S.V., NAYAK, S., PAUL, D., REDDY, R., SHARMA, C., JHA, P., HARIHARAN, M., AGRAWAL, A., CHOWDHURY, S., SARKAR, C. *et al.* Genome-wide analysis reveals downregulation of miR-379/miR-656 cluster in human cancers. *Biol Direct* **8**(10) (2013)
- LAL, E.O.A. Y O'DAY, E. MicroRNAs and their target gene networks in breast cancer. *Breast Cancer Research* **12**:201 (2010)
- LE, T.D., ZHANG, J., LIU, L., LIU, H., Y LI, J. miRLAB: An R Based Dry Lab for Exploring miRNA-mRNA Regulatory Relationships. *PloS one* **10**:e0145386 (2015)
- LEE, Y., KIM, M., HAN, J., YEOM, K.H., LEE, S., BAEK, S.H., Y KIM, V.N. MicroRNA genes are transcribed by RNA polymerase II. *The EMBO journal* **23**(20):4051–4060 (2004)
- LEE, Y.B., BANTOUNAS, I., LEE, D.Y., PHYRACTOU, L., CALDWELL, M.A., Y UNEY, J.B. Twist-1 regulates the miR-199a/214 cluster during development. *Nucleic acids research* **37**(1):123–128 (2009)
- LEHNER, B., KUNZ, P., SAEHR, H., Y FELLEBERG, J. Epigenetic silencing of genes and microRNAs within the imprinted Dlk1-Dio3 region at human chromosome 14.32 in giant cell tumor of bone. *BMC cancer* **14**(1):1 (2014)
- LI, B. Y DEWEY, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12**(1):1 (2011)
- LI, P., SHENG, C., HUANG, L., ZHANG, H., HUANG, L., CHENG, Z., Y ZHU, Q. MiR-183/-96/-182 cluster is up-regulated in most breast cancers and increases cell proliferation and migration. *Breast Cancer Res* **16**(6):473 (2014)
- LIEBERMAN-AIDEN, E., VAN BERKUM, N.L., WILLIAMS, L., IMAKAEV, M., RAGOCZY, T., TELLING, A., AMIT, I., LAJOIE, B.R., SABO, P.J., DORSCHNER, M.O. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science* **326**(5950):289–293 (2009)
- LOVE, M.I., HUBER, W., Y ANDERS, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**(12):1–21 (2014)
- LV, Z.D., KONG, B., LIU, X.P., JIN, L.Y., DONG, Q., LI, F.N., Y WANG, H.B. miR-655 suppresses epithelial-to-mesenchymal transition by targeting Prrx1 in triple-negative breast cancer. *Journal of cellular and molecular medicine* (2016)

- MACRAE, I.J., MA, E., ZHOU, M., ROBINSON, C.V., Y DOUDNA, J.A. In vitro reconstitution of the human RISC-loading complex. *Proceedings of the National Academy of Sciences* **105**(2):512–517 (2008)
- MAERE, S., HEYMANS, K., Y KUIPER, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**(16):3448–3449 (2005)
- MANI, S.A., GUO, W., LIAO, M.J., EATON, E.N., AYYANAN, A., ZHOU, A.Y., BROOKS, M., REINHARD, F., ZHANG, C.C., SHIPITSIN, M. *et al.* The epithelial-mesenchymal transition generates cells with properties of stem cells. *Cell* **133**(4):704–715 (2008)
- MARGOLIN, A.A., NEMENMAN, I., BASSO, K., WIGGINS, C., STOLOVITZKY, G., FAVERA, R.D., Y CALIFANO, A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics* **7**(Suppl 1):S7 (2006a)
- MARGOLIN, A.A., WANG, K., LIM, W.K., KUSTAGI, M., NEMENMAN, I., Y CALIFANO, A. Reverse engineering cellular networks. *Nature protocols* **1**(2):662–671 (2006b)
- MEYER, P.E., LAFITTE, F., Y BONTEMPI, G. minet: AR/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC bioinformatics* **9**(1):461 (2008)
- MING, J., ZHOU, Y., DU, J., FAN, S., PAN, B., WANG, Y., FAN, L., Y JIANG, J. miR-381 suppresses C/EBP α -dependent Cx43 expression in breast cancer cells. *Bioscience reports* **35**(6):e00266 (2015)
- MO, C.F., WU, F.C., TAI, K.Y., CHANG, W.C., CHANG, K.W., KUO, H.C., HO, H.N., CHEN, H.F., Y LIN, S.P. Loss of non-coding RNA expression from the DLK1-DIO3 imprinted locus correlates with reduced neural differentiation potential in human embryonic stem cell lines. *Stem cell research & therapy* **6**(1):1 (2015)
- MUDDULURU, G., CEPPI, P., KUMARSWAMY, R., SCAGLIOTTI, G., PAPOTTI, M., Y ALLGAYER, H. Regulation of Axl receptor tyrosine kinase expression by miR-34a and miR-199a/b in solid cancer. *Oncogene* **30**(25):2888–2899 (2011)
- NAGALAKSHMI, U., WANG, Z., WAERN, K., SHOU, C., RAHA, D., GERSTEIN, M., Y SNYDER, M. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**(5881):1344–1349 (2008)
- PALAZZO, A.F. Y LEE, E.S. Non-coding RNA: what is functional and what is junk? *Frontiers in genetics* **6**:2 (2015)

- PARK, S.M., GAUR, A.B., LENGYEL, E., Y PETER, M.E. The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2. *Genes & development* **22**(7):894–907 (2008)
- PARKER, J.S., MULLINS, M., CHEANG, M.C., LEUNG, S., VODUC, D., VICKERY, T., DAVIES, S., FAURON, C., HE, X., HU, Z. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology* **27**(8):1160–1167 (2009)
- PASQUINELLI, A.E. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nature Reviews Genetics* **13**(4):271–282 (2012)
- PENG, X., LI, Y., WALTERS, K.A., ROSENZWEIG, E.R., LEDERER, S.L., AICHER, L.D., PROLL, S., Y KATZE, M.G. Computational identification of hepatitis C virus associated microRNA-mRNA regulatory modules in human livers. *BMC genomics* **10**:373 (2009)
- REUTER, J.A., SPACEK, D.V., Y SNYDER, M.P. High-throughput sequencing technologies. *Molecular cell* **58**(4):586–597 (2015)
- ROBINSON, M.D., MCCARTHY, D.J., Y SMYTH, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**(1):139–140 (2010a)
- ROBINSON, M.D., OSHLACK, A. *et al.* A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**(3):R25 (2010b)
- SALES, G., COPPE, A., BISOGNIN, A., BIASIOLO, M., BORTOLUZZI, S., Y ROMUALDI, C. MAGIA, a web-based tool for miRNA and Genes Integrated Analysis. *Nucleic acids research* **38**:W352–W359 (2010)
- SELBACH, M., SCHWANHÄUSSER, B., THIERFELDER, N., FANG, Z., KHANIN, R., Y RAJEWSKY, N. Widespread changes in protein synthesis induced by microRNAs. *nature* **455**(7209):58–63 (2008)
- SEPHTON, C.F., CENIK, C., KUCUKURAL, A., DAMMER, E.B., CENIK, B., HAN, Y.H., DEWEY, C.M., ROTH, F.P., HERZ, J., PENG, J. *et al.* Identification of neuronal RNA targets of TDP-43-containing ribonucleoprotein complexes. *Journal of Biological Chemistry* págs. jbc–M110 (2010)
- SHANNON, P., MARKIEL, A., OZIER, O., BALIGA, N.S., WANG, J.T., RAMAGE, D., AMIN, N., SCHWIKOWSKI, B., Y IDEKER, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**(11):2498–2504 (2003)

- STELZER, Y., SAGI, I., YANUKA, O., EIGES, R., Y BENVENISTY, N. The noncoding RNA IPW regulates the imprinted DLK1-DIO3 locus in an induced pluripotent stem cell model of Prader-Willi syndrome. *Nature genetics* **46**(6):551–557 (2014)
- SUZUKI, T., MIZUTANI, K., MINAMI, A., NOBUTANI, K., KURITA, S., NAGINO, M., SHIMONO, Y., Y TAKAI, Y. Suppression of the TGF- β 1-induced protein expression of SNAI1 and N-cadherin by miR-199a. *Genes to Cells* **19**(9):667–675 (2014)
- TAM, S., TSAO, M.S., Y MCPHERSON, J.D. Optimization of miRNA-seq data preprocessing. *Briefings in bioinformatics* pág. bbv019 (2015)
- TAMBE, M., PRUIKKONEN, S., MÄKI-JOUPPIA, J., CHEN, P., ELGAAEN, B.V., STRAUME, A.H., HUHTINEN, K., CÁRPEN, O., LØNNING, P.E., DAVIDSON, B. *et al.* Novel Mad2-targeting miR-493-3p controls mitotic fidelity and cancer cells' sensitivity to paclitaxel. *Oncotarget* (2016)
- TOMCZAK, K., CZERWINSKA, P., WIZNEROWICZ, M. *et al.* The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)* **19**(1A):A68–A77 (2015)
- TOVAR, H., GARCÍA-HERRERA, R., ESPINAL-ENRÍQUEZ, J., Y HERNÁNDEZ-LEMUS, E. Transcriptional master regulator analysis in breast cancer genetic networks. *Computational biology and chemistry* **59**:67–77 (2015)
- TSAI, J.H. Y YANG, J. Epithelial–mesenchymal plasticity in carcinoma metastasis. *Genes & development* **27**(20):2192–2206 (2013)
- VALDMANIS, P.N., ROY-CHAUDHURI, B., KIM, H.K., SAYLES, L.C., ZHENG, Y., CHUANG, C.H., CASWELL, D.R., CHU, K., ZHANG, Y., WINSLOW, M. *et al.* Upregulation of the microRNA cluster at the Dlk1-Dio3 locus in lung adenocarcinoma. *Oncogene* **34**(1):94–103 (2015)
- VASUDEVAN, S., TONG, Y., Y STEITZ, J.A. Switching from repression to activation: microRNAs can up-regulate translation. *Science* **318**(5858):1931–1934 (2007)
- VIDIGAL, J.A. Y VENTURA, A. The biological functions of miRNAs: lessons from in vivo studies. *Trends in cell biology* **25**(3):137–147 (2015)
- WANG, K., SINGH, D., ZENG, Z., COLEMAN, S.J., HUANG, Y., SAVICH, G.L., HE, X., MIECZKOWSKI, P., GRIMM, S.A., PEROU, C.M. *et al.* MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic acids research* pág. gkq622 (2010)

- WANG, Z., GERSTEIN, M., Y SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics* **10**(1):57–63 (2009)
- WEIGELT, B., PETERSE, J.L., Y VAN'T VEER, L.J. Breast cancer metastasis: markers and models. *Nature reviews cancer* **5**(8):591–602 (2005)
- WINTER, J., JUNG, S., KELLER, S., GREGORY, R.I., Y DIEDERICH, S. Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nature cell biology* **11**(3):228–234 (2009)
- WISHART, D.S. Is cancer a genetic disease or a metabolic disease? *EBioMedicine* **2**(6):478–479 (2015)
- WU, G., DAWSON, E., DUONG, A., HAW, R., Y STEIN, L. ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis. *F1000Research* **3** (2014)
- YANG, M., SOGA, T., Y POLLARD, P.J. Oncometabolites: linking altered metabolism with cancer. *The Journal of clinical investigation* **123**(9):3652–3658 (2013)
- YE, F., TANG, H., LIU, Q., XIE, X., WU, M., LIU, X., CHEN, B., Y XIE, X. miR-200b as a prognostic factor in breast cancer targets multiple members of RAB family. *Journal of translational medicine* **12**(1):17 (2014)
- ZHANG, L., VOLINIA, S., BONOME, T., CALIN, G.A., GRESHOCK, J., YANG, N., LIU, C.G., GIANNAKAKIS, A., ALEXIOU, P., HASEGAWA, K. *et al.* Genomic and epigenetic alterations deregulate microRNA expression in human epithelial ovarian cancer. *Proceedings of the National Academy of Sciences* **105**(19):7004–7009 (2008)
- ZHENG, H. Y KANG, Y. Multilayer control of the EMT master regulators. *Oncogene* **33**(14) (2014)
- ZHENG, M., SUN, X., LI, Y., Y ZUO, W. MicroRNA-145 inhibits growth and migration of breast cancer cells through targeting oncoprotein ROCK1. *Tumor Biology* págs. 1–8 (2015)
- ZHU, Q.Q., MA, C., WANG, Q., SONG, Y., Y LV, T. The role of TWIST1 in epithelial-mesenchymal transition and cancers. *Tumour Biol* **37**(1):185–197 (2016)