UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
POSGRADO EN CIENCIAS BIOMÉDICAS

Identificación de cambios en los patrones de conectividad genética y epigenética en distintas etapas de progresión del cáncer.

TESIS
QUE PARA OPTAR POR EL GRADO DE:
Doctor en Ciencias
(Biomédicas)

PRESENTA:
Jose Maria Zamora Fuentes

TUTOR PRINCIPAL
Dr. Jesús Espinal Enríquez
Instituto de Ecología

COMITÉ TUTOR
Dr. Luis Mendoza Sierra
Instituto de Ciencias Biomédicas

Dra. Patricia García López,
Instituto de Cancerología

CIUDAD UNIVERSITARIA, NOVIEMBRE, 2023

JURADO ASIGNADO:

Presidente:	Dr. Felix Recillas Targa
Secretario:	Dr. Jesús Espinal Enriquez
Vocal:	Dra. Lorena Aguilar Arnal
Vocal:	Dra. Patricia García López
Vocal:	Dr. Alejandro Manuel García Carranca

La tesis se realizó en: Universidad Nacional Autónoma de México.

TUTOR DE TESIS:

Dr. Jesús Espinal Enríquez
Instituto de Ecología

Universidad Nacional Autónoma de México

Posgrado en Ciencias Biomédicas

Identificación de cambios en los patrones de conectividad genética y epigenética en distintas etapas de progresión del cáncer.

T E S I S

QUE PARA OPTAR POR EL GRADO DE:

Doctor en Ciencias
(Biomédicas)

PRESENTA:

Jose Maria Zamora Fuentes

TUTOR:
Dr. Jesús Espinal Enríquez
Instituto de Ecología

Ciudad Universitaria, Noviembre, 2023

Resumen

El cáncer es una enfermedad compleja y altamente heterogénea. La estructura celular de un tejido sufre cambios determinantes en cada una de las etapas de progresión del cáncer. Debido a la complejidad del problema utilizamos un enfoque de biología de sistemas para dilucidar procesos que pueden estar influyendo en la dinámica de la enfermedad. Principalmente utilizamos las redes de coexpresión como herramienta para estudiar los cambios dinámicos en el cáncer. También se ha visto que los alcances genómicos no son suficientes para explicar la complejidad en el avance de la enfermedad. Un paso muy importante en este sentido ha sido la incorporación de otras ómicas como la metilación, la expresión de ARNs no codificantes, ó la estructura espacial del ADN. La integración de estos datos puedan acercarnos más a los procesos que dirigen el desarrollo de la enfermedad. En este trabajo, desarrollamos una metodología completa para integrar datos genómicos y epigenéticos (metilación-miRNAs) para estudiar la progresión en las cuatro etapas del cáncer. Elegimos el Cáncer Renal de células claras (CRcc) como caso de estudio, por el alto número de muestras con datos disponibles en todas las ómicas estudiadas. Además, de ser un cáncer con una alta taza de mortalidad en etapas avanzadas. Nuestro trabajo muestra afectaciones en el programa transcripcional que cambian el comportamiento de la células en las cuatro etapas. Encontramos cuatro genes (*CXCL13*, *PLG*, *SAAC2-SAAC4*,*SLC6A19*) que cambian su expresión acorde a la progresión del cáncer. Por otro lado, propusimos un modelo de relaciones gen-miRNA. Este modelo nos reveló a *miR-217* como un miRNA que afecta genes y funciones biológicas específicas y diferentes en cada etapa. Estudiamos el metiloma para las cuatro etapas de CRcc. Construimos un algoritmo para determinar promotores hipometilados o hipermetilados que sufren este cambio significativamente. Aplicando este algoritmo encontramos oncogenes y supresores tumorales como *ITK* ó *RAB25*, entre otros, que fueron afectados de alguna manera por la metilación de ADN. Nuestros hallazgos indicaron cambios en el programa trascripcional de genes clave que afectan la estructura de las células normales o en etapas iniciales, en procesos biológicos esenciales como el ciclo celular y la apoptósis. Además, desde los primeros resultados encontramos indicios de un cambio en la infiltración de células estromales y del sistema inmune. Nuestros datos confirman posibles causas de estos eventos que ya han sido reportados y encuentran CRcc como un cáncer altamente infiltrado. Finalmente, proponemos algunas hipótesis que nos lleven a confirmar nuestros resultados experimentalmente.

Abstract

Cancer is a complex and highly heterogeneous disease. The cellular structure of a tissue undergoes decisive changes in each of the stages of cancer progression. Due to the complexity of the problem, we use a systems biology approach to elucidate processes that may be influencing the dynamics of the disease. We mainly use coexpression networks as a tool to study dynamic changes in cancer. It has been seen genomic landscape is not enough to explain the complexity in the progression of disease. A very important step in this sense has been the incorporation of other omics such as methylation, the expression of non-coding RNAs, or the spatial structure of DNA. The integration of these data can bring us a pointview closer to the processes of the direct disease development. In this work, we developed a complete methodology to integrate genomic and epigenetic data (methylation-miRNAs) to study the progression in the four stages of cancer. In this work, we developed a comprehensive methodology to integrate genomic and epigenetic data (methylation-miRNAs) to study progression in the four stages of cancer. We chose Clear Cell Renal Cancer (ccRC) as a case study, due to the high number of samples with data available in all the omics studied. In addition, it is a cancer with a high mortality rate in advanced stages. Our study revealed alterations in the transcriptional program that change the behavior of cells in four stages. We found four genes (*CXCL13*, *PLG*, *SAAC2-SAAC4*, *SLC6A19*) that change their expression according to cancer progression. We proposed a model of gene-miRNA relationships. This model revealed *miR-217* as a miRNA that affects specific and different genes and biological functions at each stage. We studied the methylome for all stages of ccRC. We built an algorithm to determine hypomethylated or hypermethylated promoters that significantly undergo this change. Applying this algorithm we found oncogenes and tumor suppressors such as *ITK* or *RAB25*, among others, that were affected in some way by DNA methylation. Our findings indicated changes in the transcriptional program of key genes that affect the structure of normal cells or in early stages, in essential biological processes such as the cell cycle and apoptosis. In addition, from the first results we found indications of a change in the infiltration of stromal cells and the immune system. Our data confirm possible causes of these events that have already been reported and find ccRC as a highly infiltrated cancer. Finally, we propose a set of hypotheses that lead us to confirm our results experimentally.

***A toda mi familia**:*
a todas mis seres queridos que ya no están físicamente;
a mi familia de cuatro patas que siempre estuvo conmigo;
y en general, a todas las personas que forman parte de mi vida;

Agradecimientos

Este documento es el compendio de tres artículos que representaron un esfuerzo de casi cinco años. En este periodo conocí investigadores, personal administrativo y compañeros que son personas extraordinarias y admirables. Todos ellos me apoyaron y dedicaron parte de su tiempo para que yo aprendiera las habilidades profesionales y académicas para optar por este grado. A todos ellos les estaré agradecido por siempre.

Especialmente, agradezco a mi tutor Jesús Espinal Enríquez por la paciencia, el apoyo y la obstinación para creer en que yo pudiera concluir este trabajo.

Doy las gracias a mis Padres y a mi hermano por todo el apoyo espiritual, anímico, logístico y económico que me brindaron durante todo este periodo. Sin ellos nada de esto sería posible.

Agradezco a las dos personas que mas cerca estuvieron de mi por todo su apoyo, consejos y ayuda en todos estos años.

Destaco el apoyo del Programa de Doctorado en Ciencias Biomédicas de la Universidad Nacional Autónoma de México y de las personas que en él laboran, especialmente por los retos enfrentados durante la pandemia de COVID-19. A quienes se encargan de los asuntos administrativos, agradezco por el acompañamiento y las facilidades otorgadas durante estos años.

Agradezco a los profesores de este programa por su esfuerzo y dedicación. Agradezco al Instituto Nacional de Medicina Genómica (INMEGEN), a la Dirección de Investigación, en particular al Área de Genómica Computacional. Gracias por permitirme pertenecer a este instituto y por el apoyo que me otorgaron.

Agradezco a CONACYT por la beca de doctorado asignada a mi CVU 267236.

Finalmente agradezco a todos mis amigos de la vida y a mis compañeros de Laboratorio por los buenos y malos momentos que hemos pasados juntos.

Declaración de autenticidad

Por la presente declaro que, salvo cuando se haga referencia específica al trabajo de otras personas, el contenido de esta tesis es original y no se ha presentado total o parcialmente para su consideración para cualquier otro título o grado en esta o cualquier otra Universidad. Esta tesis es resultado de mi propio trabajo y no incluye nada que sea el resultado de algún trabajo realizado en colaboración, salvo que se indique específicamente en el texto.

Jose Maria Zamora Fuentes. Ciudad Universitaria, Noviembre, 2023

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción.

1.1. Conceptos generales

1.1.1. Dogma Central de la Biología Molecular

Los últimos avances en biomedicina han partido de la confirmación experimental del Dogma Central de la Biología Molecular (DCBM). Hasta el momento el DCBM es la base para entender el funcionamiento subyacente de una célula. El concepto general del DCBM describe o modela el proceso de formación de las proteínas, en otras palabras, el proceso de transcripción-traducción. En esta sección describiremos brevemente conceptos importantes sobre el DCBM que forman la base para cumplir los objetivos de este trabajo.

Inicialmente se encontró que el código que aloja las instrucciones para producir las proteínas está encriptado en el ADN (*ácido desoxirribonucleico*) [23]. Dentro de la macromolécula lineal de ADN existen segmentos bien definidos que codifican para una proteína única, estos segmentos son llamados genes. Y en una capa superior el genoma agrupa al conjunto de todos los genes. Además, el genoma esta espacialmente organizado en cromosomas (23 en el caso de los seres humanos) y se encuentra confinado en el núcleo de las células. Por tanto, los genes contienen las instrucciones para construir las proteínas y se encuentran en un código secuencial de adedinas (A), citocinas (C), timinas (T) y guaninas(G) [24].

En general, el código de un gen es transcrito por una polimerasa llamada Pol-II la cual produce moléculas de *ácido ribonucleico mensajero* (ARNm) [146]. Este ARNm es exportado (o transportado) hacia fuera del núcleo, más precisamente al citoplasma. Posteriormente, en el retículo endoplasmático, los ribosomas traducen el ARNm en proteínas. El proceso de traducción es complejo y esta regido por un código de traducción de tripletes [151]. Finalmente, algunas proteínas pueden tener modificaciones que se consideran post-traduccionales, las cuales confieren características particulares o finas en sus funciones biológicas [125].

Se considera que el proceso de fabricación de proteínas es unidireccional. Sin em-

bargo, en algunos sistemas biológicos el proceso transcripción-traducción puede ser bidireccional. Un ejemplo de estas situaciones es la retrotranscripción viral [20].

Hay que destacar que el ADN puede presentar mutaciones en el código (secuencia) de los genes. Estas mutaciones pueden deberse a factores externos (rayos X, contaminantes, etc) o a factores de herencia (patologías hereditarias). Una mutación en al menos un par de bases puede desarrollar varios resultados; 1) la inhibición de la proteína (*deleción*) o 2) una proteína con un funcionamiento errado.

La viabilidad de las células depende de las proteínas y los genes. De tal manera, las células tienen rutas de procesos (circuitería) donde las proteínas interactúan de manera alostérica para activar o desactivar las funciones biológicas necesarias. Estos circuitos biológicos se denominan *vías de señalización*. El funcionamiento correcto de las vías de señalización, permite que la célula cumpla su ciclo de vida de manera adecuada. También permite su entrada al ciclo celular de manera correcta, dando paso a la división celular y a la herencia de los genomas. Se destaca que las mutaciones genéticas y sus consecuentes afectaciones epigenéticas (ver adelante, 1.1.3) también pueden ser heredables [38].

Finalmente, la expresión individual de cada gen puede ser asociada con la cantidad necesaria (adecuada) de producción de proteínas para que los procesos celulares funcionen correctamente [75]. Por ejemplo, podemos decir: "si el gen que codifica para la proteína X no se expresa, entonces, la vía de señalización Y puede ser afectada". Este marco biológico es importante porque nos permite hacer asociaciones y deducciones entre los genotipos y los fenotipos.

1.1.2. Regulación genética

El carácter específico o el fenotipo de una célula esta definido por la manera en que se regulan los genes, de manera individual (uno a uno) ó global (un gen regula varios genes)[15]. Podemos convenir que casi todas las células de un organismo tienen el mismo ADN. Sin embargo, la expresión de ciertos genes en cada célula confiere propiedades específicas que se reflejan macroscópicamente en el fenotipo. Este proceso es llamado: *diferenciación celular*. Por ejemplo, en las células endoteliales del riñón, algunos procesos de la diferenciación celular esta coordinada por la expresión de los genes *NRP1*, *CDH5*. *ELN*, mientras que en las células del cerebro *PAX6*, *TRB1*, *PROX1* y *CREB* son genes clave para su funcionamiento y desarrollo [10, 64].

La regulación genética es un proceso complejo, el cual depende de todos los factores involucrados en la trascripción de un gen (ver 1.1.1). En este sentido hay varios elementos importantes: 1) la presencia de los factores de transcripción en el promotor, 2) las condiciones y estructura de la cromatina en la posición del gen, 3) la presencia de la maquinaria transcripcional (ya sea *POL-II*, el spliceosoma, etc.), 4) las condiciones estructurales para la activación de los *enhancers* correspondientes y 5) la activación de las moléculas energéticas y de ARN necesarias para realizar el proceso. Es importante destacar, que cada gen sostiene su expresión individual a partir de la combinación de los factores antes mencionados [53]. Debido a la complejidad de este proceso biológico, dilucidar los mecanismos subyacentes requiere un estudio global del mismo.

En los últimos años se siguen utilizando técnicas experimentales como: la secuenciación masiva, Western Blot, RT-PCR, etc. que están ayudando para aproximar respuestas a preguntas biológicas básicas sobre el proceso de transcripción y sus implicaciones en algunas enfermedades [66]. Y a pesar de la complejidad, se han realizado aportaciones importantes en el campo. Por tomar un ejemplo, se ha descrito a detalle la relación que existe entre los factores de transcripción y los genes [9, 53].

En resumen, el proceso que regula la expresión de los genes es un estructura complicada de desarmar. Un enfoque para abordar este sistema biológico, es modelar las expresiones entre todos los genes (aprox. 20,000 genes en las células humanas) como una red de correlaciones. La correlación entre genes calculada de forma específica se ha denominado *coexpresión* (ver secciones 1.4 y 3.3). Este formato matemático nos ayuda para describir el fenómeno de manera global.

1.1.3. Regulación epigenética

Los estudios epigenéticos abordan elementos que no son parte de la macromolécula de ADN pero que pueden modificar la expresión genética. Estos factores van desde las modificaciones de las *histonas* hasta las alteraciones posteriores en la traducción de las proteínas. Como ejemplo, se han encontrado marcas en ciertas lisinas de las histonas H1/H5, H2, H3 y H4, que son indicativo de *eucromatina* o ADN abierto. Estas marcas representan secciones de DNA para ser transcritas [114]. En este contexto, definimos los reguladores epigenéticos como aquellos elementos moleculares fuera del ADN que tiene la capacidad de regular la expresión genética. Algunos ejemoplos son: 1) la metilación del ADN, 2) los ARNs pequeños, 3) la conformación espacial del ADN, 4) los ARNs largos no codificantes, etc. En este trabajo, estudiamos dos elementos esencialmente: 1) la metilación del ADN y 2) los micoARNs.

1.1.3.1. Los microARNs como elemento de regulación genética

Durante el proceso de transcripción de los genes (ver 1.1), existe un paso intermedio: el ayuste. Los genes están compuestos de *intrones* y *exones*. Por un lado, los intrones son segmentos de ARNm que no traducen al gen, mientras que los exones son segmentos de ARNm que son unidos por el spliceosoma (proceso de ayuste) para codificar posteriormente en la traducción. Las funciones de la mayoría de los intrones permanece incierta. Sin embargo, se ha encontrado que algunos intrones tienen multiples funciones celulares. Una de las funciones de los intrones es la biogénesis de ciertos segmentos cortos de ARN con características reguladoras [25]. Dentro de estos segmentos cortos de ARN se han destacado los *microARNs* (ó miRNAs, igualmente en este trabajo utilizamos la abreviatura miRs). Los *miRNAs* son segmentos con longitud de 19 a 22 pb, que ligan a transcritos (ARNm) en el citosplama como marca para su degradación. Este proceso puede ser visto como una represión pre-traduccional de un gen.

Como se mencionó, los miRNAs pueden surgir de transcritos cortos durante el proceso de ayuste. Sin embargo, también pueden ser transcritos por la polimerasa Pol-II

(forma canónica). Por ejemplo, ciertos microRNAs se pueden encontrar directamente en el promotor de un gen.

La biogénesis de los microARNs se inicia con la formación de pequeñas asas de ARNm que son cortadas por el microprocesador *DROSHA/DGCR6* formando los pri-miRNAs (microRNAs primarios). Después se une la proteína de transporte Exportina5 (*EXP5/RAN*) que saca el precursor del microRNA fuera del núcleo. Durante la etapa de maduración *DICER* (*DCR/TRP*) libera el asa. Finalmente, *HSP90/HSC70* desprende el microRNA [47] para ser presentado a su diana de ARNm en el citoplasma. Este proceso se describe en la Figura 1.1.

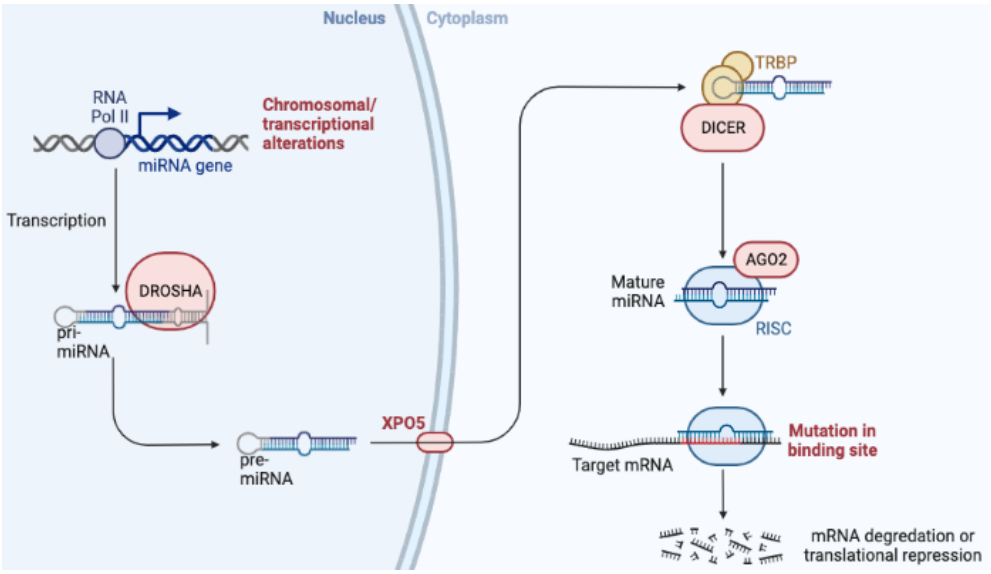

Figura 1.1: *Biogenesis de los miRNAs*. En la figura se muestra la forma canónica en la que los miRNAs dan paso a su actividad. En esta ruta los miRNAs son transcritos de exones y siguen su cadena de procesos hasta degradar un ARNm diana. Se resalta que los miRNAs transcritos de intrones no requieren ser adaptados por *DROSHA*.

En el proceso final, un miRNA maduro, en su mayor parte, se une a la región 3' no traducida del ARN mensajero (ARNm) y, dependiendo del grado de complementariedad con el ARN diana, se puede conducir a la degradación o bloqueo de la traducción del ARNm. Estudios recientes sugieren que el bloqueo de la traducción va acompañado de cierta degradación [131]. Específicamente, la regulación miRNA-gen puede reprimir o promover directamente la transcripción (ruta no canónica) o la traducción (ruta canónica) [103].

La importancia de la regulación por miRNAs se muestra en datos bioinformáticos. Por ejemplo, en el ADN humano, casi mil miRNAs son capaces de regular aproximadamente 3/4 del transcriptoma humano [11]. Además, es conocido que un solo miRNA

puede regular varios ARNm y que una sola transcripción de ARNm puede ser la diana de varios miRNAs [56]. Para comprender en términos generales la complejidad intrínseca de una interacción *miRNA-gen*, se vuelve obligatorio desarrollar enfoques integrales que combinen diferentes fuentes de información.

Por supuesto, estos efectos reguladores están involucrados en enfermedades y procesos biológicos anormales [99]. Por ejemplo, en cáncer los miRNAs pueden asociarse con el desarrollo de funciones oncogénicas y supresoras de tumores. Además, estos elementos pueden tener la capacidad de modular diferentes genes que al mismo tiempo dependen de otros contextos biológicos productos de una patología [150]. En secciones posteriores se reportan algunos miRNAs importantes en cáncer. Debido a estas condiciones, se vuelve imprescindible comprender el paisaje de regulación de los microARNs de forma dinámica y completa.

Finalmente hay que mencionar una observación sobre las microARNs: un miRNA se considera como elemento epigenético si y sólo si la diana a la que esta regulando es un elemento molecular que afecte el ADN directamente.

Además, vale la pena destacar que respecto a las definiciones sobre los elementos epigenéticos regulatorios existe una discusión abierta y no hay un consenso en el campo. Por ejemplo, en eucariontes, los miRNAs solo se han probado como mecanismo epigenético que se transmite a la descendencia en levadura, C. Elegans y parcialmente en moscas.

1.1.3.2. La metilación como un elemento epigenético y regulatorio

Otro factor regulatorio importante es la metilación de ADN. Una de las modificaciones epigenéticas más conocida en mamíferos es la metilación del lado 5' de las citosinas (5metilcitosina; *5mC*) en regiones genómicas con abundancia de dinucleótidos *CpG* (*citosina-fosfato-guanina*). Estas regiones son llamadas *islas CpG* (CGIs). Este rasgo epigenético de ADN es relativamente simple. Otras modificaciones epigenéticas como las alteraciones postraduccionales de las histonas son más diversas y complejas.

La maquinaria de metilación en los mamíferos está compuesta por dos elementos: 1) las ADN metiltransferasas (*DNMTs*), que establecen y mantienen los patrones de metilación del ADN, y 2) las proteínas de unión a metil-CpGs (*MBDs*), que están involucradas en el reconocimiento de las marcas de metilación [156] en el ADN.

Los patrones de metilación del ADN son esenciales para el desarrollo de los mamíferos y para el funcionamiento normal del organismo adulto [114]. La metilación es un potente mecanismo para dirigir la diferenciación correcta de una célula. De forma estructural, la metilación puede prevenir la recombinación ilegítima de segmentos repetitivos de ADN que pueden causar la desregulación de genes cercanos y producir fenotipos anómalos [99].

La metilación del ADN ocurre en hasta el 80 % de las CGIs, y los residuos de CpG no metilados restantes se enriquecen en CGIs ubicadas en los promotores de genes [156]. Sin embargo, se desconoce el estado de metilación de muchas CGIs asociadas a patologías como el cáncer [114, 120]. La metilación del ADN reprime la transcripción directamente,

al inhibir la unión de factores de transcripción específicos, e indirectamente, al reclutar proteínas de unión como Polycomb [77]. Esto deviene en actividades represivas asociadas a la remodelación de la cromatina.

Un vínculo entre la metilación del ADN y el cáncer se descubrió por primera vez en 1983, cuando se demostró que los genomas de las células cancerosas estaban relativamente hipometilados con respecto a sus contrapartes normales.[40]. De esta manera, se observó como la hipometilación de regiones de heterocromatina puede dar paso a recombinación mitótica e inestabilidad genómica [95].

En el otro extremo, la hipermetilación también puede generar cambios sustanciales en células somáticas normales [137]. Se ha visto como genes involucrados en la regulación del ciclo celular, la invasión de células tumorales, la reparación del ADN, la remodelación de la cromatina, la señalización celular, la transcripción y la apoptosis, se hipermetilan de forma anómala. Estos genes se encuentran silenciados en casi todos los tipos de tumores [120]. Este fenómeno puede conferir a las células tumorales la habilidad de crecimiento descontrolado y la inestabilidad genética necesaria para formar metástasis.

Los cambios en la metilación dentro de las etapas de cáncer son difíciles de rastrear. Una restricción es el número tan grande de islas CpG que se encuentran esparcidas en todo el genoma. Además, las dianas de las proteínas metiladoras (*DNMT3*) pueden actuar sobre el promotor o el cuerpo de los genes [120]. Por lo que construir las redes que relacionan los genes y los patrones de metilación es un problema complejo. En este trabajo se propone un enfoque sistemático para abordar este problema.

1.2. Biología del cáncer

El cáncer es una enfermedad multifactorial, donde el mal funcionamiento en la expresión de los genes da paso al inicio y progresión de los tumores. Además, la incorrecta actividad en la regulación de la expresión en los genes puede desarrollar diferentes patologías en el microambiente tumoral. En este contexto el microambiente se define como el cúmulo completo de células y moléculas que conforman el tumor.

En particular, el cáncer se caracteriza por el mal funcionamiento de ciertos genes. Por ejemplo, el desarrollo y progresión del cáncer de mama, está relacionado con afectaciones directas en los genes *HER2*, *MKI67*, *BRCA1* y *PD-L1* [91].

Diversas enfermedades han sido asociadas con un patrón de expresión en los genes de células afectadas [118]. En este sentido, podemos definir el cáncer como un resultado global de alteraciones en la expresión de genes que provocan cambios en los procesos biológicos. Algunos ejemplos de los procesos celulares impactados directamente por el cáncer son: 1) proliferación celular, 2) apoptosis, 3) metástasis y 4) el control del ciclo celular. [55].

Es importante señalar las consideraciones funcionales de algunos genes clave en las vías de señalización. Estos genes se pueden agrupar en tres tipos : 1) oncogenes (OG), 2) supresores tumorales (TSG) y 3) ambivalentes (BG). Los oncogenes son elementos genómicos que participan directamente en la iniciación del cáncer (*tumorigénesis*). De

forma general, los oncogenes codifican para proteínas que controlan la proliferación celular, la apoptosis o ambas. Con menos probabilidad, pero también pueden controlar otras funciones básicas de las células. Pueden ser activados por alteraciones estructurales resultantes de mutaciones o fusión de genes, por yuxtaposición a elementos *enhancer* o por amplificación [25].

Por otro lado, los TSGs son genes esenciales en el metabolismo y el desarrollo celular. Los TSGs deben estar completamente apagados para que se lleve a cabo la tumorigénesis. Estos genes pueden ser haploinsuficientes, es decir, requieren la expresión en dos alelos. Sin embargo, por ser genes clave pueden alterar dramáticamente los fenotipos con la pérdida de un solo alelo [107].

Los oncogenes y los TSGs son pleiotrópicos, en otras palabras, son esenciales en múltiples procesos celulares, y algunas de estas funciones pueden ser más sensibles a la expresión génica que otras. Este hecho revela que la tumorigenesis es un proceso complejo de interacción entre estos genes, por lo que sus funciones pueden ser ambivalentes y desconocidas, aún teniendo evidencia experimental de las propiedades oncogénicas o supresoras [107].

Otro aspecto importante en la biología del cáncer es la respuesta inmunológica. La relación entre el microambiente (contexto inmunológico) y la progresión del cáncer sigue siendo incierta. Se han realizado avances para encontrar los patrones funcionales de macrófagos y células T, caracterizados por una amplia diversidad, tanto en fenotipos como en respuestas [19]. Para superar este desafío, existen esfuerzos por encontrar marcadores de los diferentes elementos de la respuesta inmune; por ejemplo, en [148], se encontró que los individuos con respuestas inflamatorias enriquecidas para *BAP1* tienen un peor pronóstico. Otro ejemplo destaca que el gen *PBRM1* se encuentra disminuido, tanto en modelos animales como en muestras humanas de diferentes tumores. Estos resultados también se asocian con una infiltración inmune a la baja. Además, se ha observado que los tumores sin el gen *PBRM1* fueron más resistentes al anticuerpo anti-PD-1 [87].

En este sentido, hay tumores con diferentes patrones de infiltración del sistema inmune. Por ejemplo, el cáncer renal destaca por una alta infiltración, mientras que el cáncer cervico-uterino muestra una baja infiltración de células del sistema inmune [82].

Por tanto, la heterogeneidad del microambiente tumoral es solo parcialmente responsable de la complejidad detrás de las respuestas de algunos tipos de cáncer. Se sabe que los elementos reguladores y los moduladores epigenómicos también desempeñan funciones importantes; por ejemplo, se ha argumentado que los miRNAs parecen regular más del 80 % de los genes humanos [42]. Además, se han reportado patrones de expresión aberrantes de miRNAs en muchos cánceres humanos [80]. Varios de estos genes se consideran factores clave en las vías de desarrollo del cáncer [1]. Por mencionar solo algunos, miRNAs como *miR-646*, *miR-21* y *miR-204* se han implicado en el desarrollo y la progresión del carcinoma de células renales [154]. También se ha reportado que las familias de miRNAs, como la familia miR-200, están fuertemente desreguladas en metástasis y tumores primarios [34, 104]. Por otro lado, la metilación de las islas CpG ubicadas en las regiones promotoras de varios genes supresores de tumores y oncogenes también se ha considerado un *hallmark* epigenético importante en el proceso de carcinogénesis[25, 54].

1.2.1. Progresión del cáncer

En general, el estadio tumoral se considera como el parámetro de pronóstico más importante para el tratamiento clínico y la evolución de los carcinomas. El agrupamiento por estadios (TNM) del *American Joint Committee on Cancer (AJCC)* es el sistema de estadificación del cáncer más utilizado. En este momento, esta clasificación se encuentra en su Octava revisión. La agrupación TNM juega un papel importante en las decisiones de tratamientos de acuerdo con las pautas de la Red Nacional Integral del Cáncer (NCCN), especialmente en los criterios de selección para la terapia adyuvante [133, 138].

La clasificación TNM esta descrita de la siguiente manera:

1. (T1,T2,T3,T4) tamaño del tumor,
2. (N0-N1) presencia de nódulos linfáticos en el tumor,
3. (M0-M1) la evidencia de metástasis en otro órgano del paciente.

La agrupación TNM es **tejido-específica**, y puede contener subclasificaciones acorde a las condiciones de tipos de cáncer raros. En general, la clasificación clínica se considera como: Etapa I (T1N0M0), Etapa II (T2N0M0), Etapa III (T1-3N1M0 + T3N0M0) y Etapa IV (T*N*M1+T4N1M0) [138].

Es importante notar que la agrupación TNM esta sujeta a revisiones y adaptaciones continuamente. La evidencia muestra que la subclasificación puede tener diversos ajustes moleculares cuando se aplica en poblaciones y tumores diferentes. Por tanto, se siguen desarrollando herramientas teóricas basadas en evidencia clínica para mejorar el rendimiento de estas subclasificaciones [133].

1.3. Contexto tecnológico de los experimentos

1.3.1. Cáncer y experimentos de secuenciación

Como marco para el estudio del cáncer existen muchos experimentos biológicos que han sido desarrollados desde hace más de 50 años. Sin embargo, las tecnologías de secuenciación han generado una gran relevancia en los últmos años. Como describiremos en esta sección estas tecnologías nos dan una cantidad enorme de información con mucho detalle de fenómenos moleculares y biológicos clave. En este sentido, la pregunta de investigación de este trabajo está formulada a partir de tres tecnologías de secuenciación, a saber, RNAseq, miRNA-seq y Methyl-seq.

1.3.2. Contexto tecnológico de los experimentos

Gracias al desarrollo de las tecnologías de secuenciación de próxima generación (NGS), se ha mapeado el genoma humano en muchos individuos [58]. Sin embargo,

los últimos años han sido testigos de una avalancha de nuevos métodos para interrogar diferentes propiedades de una célula a escala de todo el genoma. Hoy en día, las nuevas tecnologías ofrecen mayor volumen (cantidad de pares de base secuenciadas) y precisión en los experimentos.

A nivel transcriptómico, el proceso de secuenciación resulta en la expresión de los genes dentro de una masa de células (*RNA-seq bulk*). La secuenciación de ARNm se desarrolló hace más de una década. El flujo de trabajo estándar comienza en el laboratorio con la extracción de ARNm, seguido por el enriquecimiento de ARNm, la síntesis de ADNc y la preparación de una biblioteca de secuenciación ligada a un adaptador. Posteriormente, la biblioteca se lleva a secuenciación con una profundidad de lectura de 10 a 30 millones de lecturas por muestra, este parámetro, según sea la plataforma de alto rendimiento considerada (por ej. Illumina).

Los pasos finales son computacionales: 1) alinear y/o ensamblar las lecturas de secuenciación en un transcriptoma, 2) cuantificar las lecturas que se superponen a las transcripciones, 3) filtrar y normalizar entre muestras. Y finalmente, 4) se realiza un modelado estadístico de cambios significativos en los niveles de expresión por genes individuales y/o transcripciones entre grupos de muestras. En este enfoque, el resultado es una matriz de expresión con genes en los renglones y muestras individuales de pacientes enfermos o sanos en las columnas.

Desde el punto de vista epigenómico, se han desarrollado grandes avances en varios tipos de secuenciación como son: reconocimiento de estados de la cromatina (ATAC-seq), de marcas en las histonas (CHIP-seq) o de puntos de contacto (HI-C). [58]. En general los pasos para un experimento de secuenciación epigenómico son: 1) preparación de muestras, 2) transposición, 3) preparación de bibliotecas, 4) secuenciación y 5) análisis de datos. Cada uno de estos análisis ofrece una visión única pero complementaria, de la organización del genoma y la función celular. Se espera que la integración de estos datos proporcione más conocimientos biológicos que el uso de un sólo estudio [136]. En el área biomédica, el desafío que enfrentamos es comprender un modelo completo (genómico y epigenómico) para determinar que cambios celulares pueden conducir a una enfermedad.

Un método computacional estándar para integrar diferentes fuentes de datos es el aprendizaje no supervisado (ANS). Este método es escalable en cuanto al volumen de estudios por integrar. Una característica importante del ANS es que los datos se abordan sin sesgos, conocimientos o hipótesis previos (análisis *data-driven*). En un enfoque no supervisado simplemente se plantea la pregunta: ¿qué tipos de patrones existen en un conjunto de datos? Una suposición común hecha por los enfoques no supervisados es que las características interesantes de los datos son las que ocurren con frecuencia y, por lo tanto, el objetivo es encontrar patrones comunes. Generalmente en estos modelos, un gran conjunto de datos se parte en grupos más pequeños que puedan interpretarse más fácilmente. Por los motivos antes mencionados, en este trabajo se utilizaron varias técnicas de ANS.

Vale la pena destacar que los pasos experimentales descritos en esta sección están cubiertos con extrema precaución por los consorcios (TCGA-GDC) que administran las

entradas en las bases de datos usadas en en este trabajo.

1.3.3. Secuenciación de miRNAs

Uno de los primeros éxitos en la aplicación de la NGS al estudio de ARN fue la secuenciación de los fragmentos de miRNAs (miRNA-seq). Los protocolos para preparar bibliotecas de secuenciación de miRNAs son relativamente simples y generalmente se realizan en una sola reacción. El hecho de que los miRNAs se encuentren en su estado nativo con un fosfato terminal 5' permite el uso de ligasas para apuntar selectivamente a los miRNAs.

Sin importar la secuenciación de ARNm (mRNA-Seq) o de ARN pequeño (smRNA-Seq), la preparación de muestras generalmente incluye tres pasos: 1) aislamiento de ARN total, 2) enriquecimiento del ARN objetivo y 3) transcripción inversa de ARN en ADN complementario (ADNc).

Los ARN pequeños son una clase de ARN no codificantes que tienen menos de 200 nucleótidos. Las especies más comunes y mejor estudiadas son los miRNAs, los cuales desempeñan un papel fundamental en la regulación de genes (ver 1.1.3.1). La preparación de la biblioteca de smRNA-Seq es simple debido a un fosfato 5' terminal presente en el estado nativo de los miRNAs. La preparación de la biblioteca comienza con una ligadura en dos pasos. Primero, un adaptador de ADN adenilado del extremo 3' se liga a los ARN utilizando una ARN ligasa 2. Después, se usa un segundo adaptador de ARN del extremo 5' el cual se liga con la ARN ligasa. Posteriormente a las ligaduras, se realiza una PCR de transcripción inversa para convertir los miRNAs ligados en ADNc. Finalmente, el producto de la amplificación y la selección por tamaño en gel, se procesa para posteriormente concluir con la secuenciación [66].

Una limitación importante en la construcción de bibliotecas de miRNAs surge cuando la cantidad de ARN de entrada es baja (p. ej., $<$200 ng de ARN total); Los dímeros adaptadores cortos compiten en la reacción de RT-PCR con el producto deseado, los adaptadores y las inserciones de miRNAs. Cuando hay demasiados dímeros adaptadores, estos fluyen hacia arriba del gel durante el paso de selección de tamaño y contaminan las bandas del producto. Para minimizar este problema, muchos kits comerciales de preparación de bibliotecas de miRNA ahora incorporan varias estrategias para suprimir la formación de dímeros adaptadores [60].

1.3.4. Secuenciación de sitios CpG (metilación)

En el año 2000, se analizó el estado de metilación de 1,184 islas CpG de 98 muestras de tumores, donde se mostró que la metilación *de novo* de las islas CpG está muy extendida en las células tumorales. En este estudio, el grado de metilación varió entre los tipos de tumores y los individuos. Un promedio de 608 islas CpG estaban anormalmente hipermetiladas [22].

Para estudiar la metilación de todo el genoma (metiloma) se experimenta con secuenciaciones del tipo *WGBS - whole-genome bisulfite sequencing*. En esta tecnología

se realiza el tratamiento del ADN genómico con bisulfito de sodio que convierte las citosinas no metiladas en uracilos (U), mientras que las citosinas no metiladas no se ven afectadas. Posteriormente, los uracilos se convierten en timinas (T) por una reacción en cadena de la polimerasa (PCR). Los métodos basados en la conversión de bisulfito brindan una resolución de base única y se usan comúnmente para investigar secuencias de ADN específicas cuando se combinan con la secuenciación.

El protocolo Infinium HumanMethylation450 BeadChip (HM450K) de Illumina, utilizado en este trabajo, implica la conversión con bisulfito del ADN genómico y su amplificación. En este experimento, el ADN se hibrída en matrices que contienen sondas prediseñadas para distinguir entre citosinas metiladas y no metiladas. Cada HM450K BeadChip puede interrogar a más de 450 000 sitios de metilación que cubren el 96 % de los CGIs. Hasta la fecha, las matrices HM450K dominan los estudios que investigan el metiloma del cáncer y otros estudios de todo el epigenoma [156].

1.4. Coexpresión y construcción de redes ómicas

Tomando en cuenta las tecnologías de secuenciación revisadas en la sección 1.3, y retomando la biología de los sistemas estudiados en este trabajo, entonces podemos introducir el estudio de la coexpresión genética.

Como se ha mencionado, en las células la expresión génica está regulada por proteínas. Estas proteínas son en sí mismas productos de genes. Por otro lado, los fenotipos celulares están determinados por la actividad dinámica de enormes redes de genes co-regulados [44]. Este concepto se puede entender como la retroalimentación de un sistema complejo, por lo que es posible aproximar un modelo basado en las asociaciones estadísticas entre los niveles de abundancia de ARNm en cada gen [75]. Y aunque las correlaciones entre genes no son directamente proporcionales a las concentraciones de proteínas activadas, deberían proporcionar pistas para descubrir los mecanismos reguladores de genes. En este sentido, la llegada de las tecnologías de secuenciación de alto rendimiento para medir los niveles de abundancia de ARNm en todo el genoma ha generado muchas investigaciones destinadas a utilizar estos datos para construir modelos conceptuales basados en *redes de genes* con el fin de describir de manera concisa las influencias reguladoras que los genes ejercen entre sí [94].

Estrictamente, la coexpresión es una cuantificación de la correlación que existe entre la expresión de dos genes. Esta correlación puede ser medida y ajustada de forma lineal o no lineal. Existen medidas de correlación lineales como son *Pearson* o *Spearman*. La aplicación de estos métodos para construir redes de coexpresión se ha denominando como *ingeniería reversa de redes celulares*. Por otro lado, hay métodos de ajuste más estrictos y robustos (no lineales) que se han desarrollado, uno de estos es la correlación por *información mutua*. El objetivo es producir una representación de alta fidelidad de la topología (estructura) de la red celular. Esta representación es un grafo donde los genes se representan como vértices (nodos) y están conectados por aristas que representan interacciones reguladoras directas.

Para ver detalles sobre el contexto y cómo se calcula la coexpresión por información mutua ver sección A.3. Vale la pena destacar, que durante este trabajo se desarrolló una adaptación para ejecutar en paralelo (*multi-thread*) este tipo de cálculos. Este código reduce el tiempo de cómputo exponencialmente y se encuentra disponible en línea con acceso libre (ver sección 3.4.5). Actualmente, esta aportación de código se utiliza ampliamente en al menos un par de grupos de investigación y en diferentes trabajos de ciencia básica.

Finalmente, el estudio de la coexpresión también nos permite tener una base biológica sistemática para encarar un desafío mayor: la integración de datos multiómicos (RNA-seq, miRNA-seq, METHYL-seq).

1.4.1. Relaciones *-cis* y *-trans* en cáncer

En estudios previos se ha mostrado que las redes de coexpresión tienen distintas proporciones de interacciones *-cis* y *-trans* [29, 31, 32, 37, 44]. Las interacciones *-cis* (intracromosomal) son aquellas que conectan dos genes en el mismo cromosoma. Mientras que las interacciones *-trans* (intercromosomal) relacionan dos genes de distinto cromosoma.

Estudios muy completos sobre la perdida de coexpresión en cáncer de mama y pulmón, muestran la importancia estructural de este fenómeno [6, 44]. En este trabajo se pretende complementar la evidencia en este sentido. Por ejemplo, todavía no se ha explicado con claridad cómo se desarrolla el fenómeno de perdida de comunicación intercromosomal durante la progresión del cáncer. Las relaciones *-cis*/*-trans* en cada fenotipo por etapa debería tener sus condiciones particulares.

Con las bases establecidas hasta este punto de la introducción, podemos entonces abordar un caso de estudio. En la sección siguiente, describimos a detalle los motivos por los que utilizamos cáncer renal de células claras. Además describiremos a detalle antecedentes y la importancia molecular de esta patología.

1.5. Caso de estudio: Cáncer de Riñón

La incidencia global de los carcinomas de células renales (CCR) ha aumentado notoriamente desde 2008, ejerciendo una carga importante tanto en los individuos como en los sistemas de salud [128]. Actualmente, se ha puesto en marcha una serie de acciones básicas y clínicas para tratar de paliar esta situación. Se han realizado importantes esfuerzos en la búsqueda de reguladores clave en el desarrollo de esta enfermedad. Los oncogenes y los supresores de tumores como los genes: *VHL* (3p26), *FH* (1q42.1), *MET* (7q34) o *FLCN* (17p11.2) se han estudiado en diferentes tipos de CCR. Estos genes también están asociados con diferentes síndromes y patrones de herencia [52].

El cáncer renal (CR) se clasifica, de forma general, en tres subtipos: cáncer renal papilar, cáncer renal cromófobo y cáncer renal de células claras. Hasta el 85 % de los casos de CR corresponden al subtipo de células claras (CRcc) [17]. La progresión en este

subtipo de tumor es comúnmente iniciada por mutaciones en *VHL*. Algunos factores de transcripción se acumulan debido a la inactivación de *VHL*, lo que induce la expresión del factor de crecimiento endotelial vascular (*VEGF*). Por lo tanto, los CRcc a menudo están altamente vascularizados y responden a la terapia antiangiogénica [150]. Las mutaciones posteriores suelen surgir en *BAP1*/*PBRM1*/*SET2*/*KDM5C*, lo que da lugar a defectos de reparación del ADN (ver sección A.2). Estos genes se han considerado como impulsores para la evolución del CRcc. Como punto adicional, la activación de la vía PI3K puede promover la metástasis [69]. Teniendo en cuenta que hasta un tercio de los casos presentarán metástasis, es innegable la importancia de encontrar con mayor precisión los factores moleculares que subyacen en la progresión del CRcc.

Cabe destacar que en recientes estudios se ha visto que la inactivación de *VHL* en humanos y ratones no induce directamente la tumorigénesis de CRcc [65]. La importancia de esta evidencia nos habla del espacio de conocimiento que sigue abierto en el estudio de estos carcinomas.

En el contexto epigenético, cada vez hay más pruebas del papel de los miRNAs en el aumento y desarrollo del carcinoma renal de células claras. Las alteraciones en las funciones regulatorias de los miRNAs pueden ser factores clave en el desarrollo y progresión de CRcc a etapas más avanzadas. Sin embargo, aún se desconoce el papel específico de los miRNAs y sus familias durante las etapas de progresión de CRcc [99].

También se ha visto un alto porcentaje de CGIs que están metiladas en CRcc. Estos tumores, que tienen un fenotipo *metilador* de islas CpG son agresivos y tienen una actividad energética aumentada [57]. Además, los tumores marcados por la metilación se asocian con una supervivencia deficiente [114]. Como se mencionó, la presencia de grupos metilo en regiones específicas del ADN depende de *DNMT1*, *DNMT3A* y *DNMT3B*, que comúnmente se sobreexpresan en diferentes cánceres incluyendo, el CRcc [99].

Capítulo 2

Objetivos

2.1. Hipótesis

A partir de los antecedentes, podemos establecer que el cáncer es un resultado de los cambios genéticos y epigenéticos dentro de las células normales. Estos cambios en el tiempo son progresivos y convierten los mecanismos celulares en fenómenos anormales que se incrementan y degeneran en un mal funcionamiento de las células. Entonces, nuestra hipótesis de trabajo es que:

Existen patrones de regulación llevados a cabo por miRNAs ó por metilación de ADN que afectan la coexpresión de genes involucrados en la progresión del cáncer; además, estos patrones son identificables a través del enfoque de redes.

2.2. Pregunta de investigación

Por lo tanto, planteamos la siguiente pregunta: ¿Cuáles son los posibles patrones de control regulatorio llevados a cabo por miRNAs ó por metilación de ADN que afectan la coexpresión de genes involucrados en la progresión del cáncer, y que son identificables a través de un análisis de estructura de redes?

2.3. Objetivo general

Encontrar, a través de un enfoque de redes, genes **clave** que son afectados por miRNAs o en su metilación, los cuales cambian su programa regulatorio y de co-expresión durante la progresión del cáncer. Los aspectos genéticos subyacentes como las mutaciones o las variantes en el número de copias, quedan fuera del marco de este trabajo.

2.4. Objetivos particulares

- Construir redes de coexpresión de cada fenotipo (control y etapas de progresión del cáncer). En estas redes determinaremos los cambios en el programa regulatorio de los genes durante la progresión del cáncer. Las relaciones de coexpresión serán soportadas por una significancia estadística.
- Diseñar un protocolo bioinformático para encontrar genes clave afectados en su expresión como efecto de un cambio en la expresión de al menos un miRNA asociado. Los resultados de este protocolo se determinarán por una relación estadísticamente significativa.
- Diseñar un protocolo bioinformático para encontrar genes clave afectados en su expresión debido un cambio en la metilación de su promotor. Los resultados de estos algoritmos y filtros se determinarán por una relación estadísticamente significativa.

Capítulo 3

Métodos.

3.1. Descripción general de los métodos

En este trabajo se proponen diferentes enfoques bioinformáticos para abordar la regulación genética y epigenética en la progresión del CRcc. Se puede considerar que se realizaron tres estudios: 1) regulación gen-gen, 2) regulación miRNA-gen y 3) regulación metilación-gen. En general, los pasos a seguir en común son los siguientes:

1. Obtención de datos
2. Limpieza y control de calidad
3. Procesamiento de datos
    a) Filtrado de datos
    b) Construcción de Redes
4. Enriquecimiento funcional
5. Análisis de resultados

Cada uno de estos pasos esta adaptado y contiene modificaciones dependiendo del estudio bioinformático y las preguntas biológicas. En las siguientes subsecciones se describe a detalle cada una de los procedimientos y la metodología general.

Para proporcionar una manera clara y fácil de reproducir los resultados mostrados en este trabajo, se publicaron las matrices de expresión y todo el código de cada uno de los protocolos. En los siguientes repositorios se encuentra todo el flujo de trabajo, desde la descarga de datos hasta el enriquecimiento funcional de cada protocolo.

1. Regulación gen-gen. https://github.com/josemaz/kidney-stages
2. Regulación miRNA-gen. https://github.com/josemaz/kirc-mirna
3. Regulación metilación-gen. https://github.com/josemaz/kirc-methyl

3.2. Obtención de datos

Como se menciona en la sección 1.5 el caso de estudio para mostrar el potencial de esta metodología es el CRcc. El conjunto de datos completo fue descargado del repositorio de GDC (Genome Data Commons - https://portal.gdc.cancer.gov/repository) para CRcc. El consorcio The Cancer Genome Atlas (TCGA) es una fuente de datos de expresión genética (secuenciación de ARN), expresión de miRNAs (secuenciación de de miRNAs) y metilación de ADN (mircroarreglos de alta densidad por conversión de bisulfito).

Para esta tarea, desarrollamos un conjunto de *scripts* que utilizan como entrada los datos epigenómicos y metadatos transcriptómicos del proyecto TCGA (en este caso, CRcc). El proyecto asociado en la base de datos se denomina *TCGA-KIRC*. Estos programas recopilan todas las muestras con sus perfiles transcriptómicos y epigenéticos. Los índices de todos los conjuntos de datos (trasncriptómicos y epigenéticos) se armonizaron para que coincidieran con los códigos de los pacientes en cada uno de los experimentos computacionales. Este último punto se vuelve relevante debido a que no todas las tecnologías se implementaron en todos los pacientes.

En todos los casos, utilizamos archivos de expresión de nivel 3 (RNA-Seq) para muestras de CRcc. Dividimos a estos pacientes por etapa de progresión del cáncer, así como control de tejido no tumoral (tejido normal adyacente al tumor, de acuerdo al análisis del propio consorcio).

3.3. Análisis de datos transcriptómicos

En la Figura 3.1 se introduce el flujo de trabajo llevado acabo en el protocolo de análisis de relaciones gen-gen.

El número de casos para cada etapa se muestra en la Tabla 3.1. El total de muestras para los estudios de relaciones gen-gen fue de 608 pacientes.

Tabla 3.1: *Número de muestras de los datos transcriptómicos.* Datos de muestras por etapa para los estudios de regulación genética (interaciones gen-gen). El total son 608 muestras de CRcc.

Control	**Etapa** I	**Etapa II**	**Etapa III**	**Etapa IV**
72	272	59	123	82

Las muestras tumorales se separaron en estadios según la variable *tumor_stage*. Esta variable es proporcionada por TCGA como etiqueta en cada expediente clínico. En el caso de que no se informara el valor *tumor_stage*, decidimos descartar esa muestra.

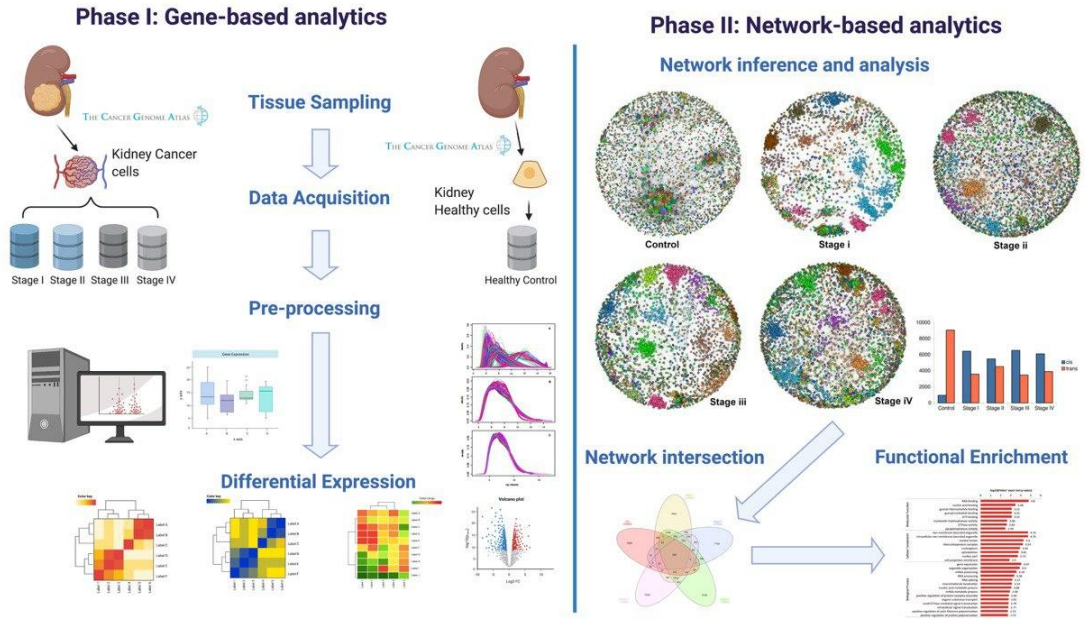

Figura 3.1: *Flujo de trabajo para analizar los datos transcriptómicos.* La base de datos origen para este análisis es TCGA con especificación del proyecto KIRC para CRcc. Vale la pena destacar que el procesamiento de datos es un estándar en este trabajo. El resultado final de este proceso son las redes gen-gen las cuales se analizan en la sección de resultados. Finalmente, el análisis de enriquecimiento funcional nos da las pistas biológicas para generar las hipótesis correspondientes.

3.3.1. Procesamiento de datos

Construimos un pipeline de pre-procesamiento de datos en tres fases:

1. Control de calidad previo a la normalización,
2. correcciones por lotes y de sesgo (normalización) y
3. control de calidad posterior a la normalización.

El pre-procesamiento de datos se realizó con la guía de otros trabajos realizados en el grupo ([29, 31, 34, 37, 44, 129]). En detalle, evaluamos las siguientes condiciones: 1) la abundancia de biotipos, para asegurar que las muestras contenían genes codificadores de proteínas, 2) se revisaron los datos de expresión crudos (*raw counts*) por biotipo, con el fin de confirmar que la expresión mediana más alta correspondía a genes que codifican para proteínas. Y finalmente, 3) calculamos la cantidad de genes detectados

por muestra utilizando gráficos de saturación. Estos pasos se realizaron con el paquete `NOISeq` de `R` estándar [139].

El método de normalización para corregir el sesgo por longitud (*full*) y por contenido de GCs (Guaninas-citosinas) (*full*) fue *within-lane*. Además, aplicamos una normalización *TMM* para eliminar los sesgos por composición de ARN entre bibliotecas. Con esto, los datos están preparados para encontrar genes diferencialmente expresados [119]. Tanto los análisis de componentes principales (PCA) como los gráficos antes mencionados son material suplementario en [158]. Los genes se filtraron por valores de expresión promedio ($mean > 10$). La normalización para corregir los efectos por lotes fue realizada con el método ARSyN [102]. Este método está implementado en el paquete `NOISeq`. Los *scripts* para realizar el análisis de preprocesamiento se pueden encontrar en https://github.com/josemaz/kidney-stages.

3.3.2. Expresión diferencial

Realizamos un análisis de expresión diferencial (DE) para comparar los cambios entre cada etapa de CRcc y el control (tejido adyacente). Este análisis se validó a través de la prueba empírica de Bayes sobre los errores estándar utilizando el paquete *edgeR* [121]. Para considerar un gen expresado diferencialmente, se tomó un corte de $LogFoldChange$ ($|LFC| > 2.0$).

3.3.3. Significancia estadística

Con el fin de comparar múltiples perfiles de expresión, implementamos cálculos con corrección por *False Discovery Rate* (FDR) de *Benjamini & Hochberg*. El límite del valor-p ajustado por FDR se estableció en 0.05 para cada comparación.

También realizamos una comparación multigrupo basada en el método de prueba de razón de verosimilitud (LRT) para obtener todos los contrastes entre grupos (4 etapas y control) [92]. Este método está implementado en el paquete `DEseq2` de `R`. Usamos la desviación de cada grupo para el cálculo de los valores p y este paso lo repetimos en cada contraste. Filtramos los genes con un valor de significancia corregido que fuera inferior a 0.05 y un cambio logarítmico de $-0.5 > |LFC| > 0.5$ para cada contraste. Lo que buscamos fueron genes expresados diferencialmente, no solo entre genes de estadios de cáncer y muestras control, sino también entre estadios consecutivos y progresivos.

Dado que los datos de CRcc se separan en etapas, nos enfocamos en aquellos genes que sufren un cambio consecutivo, es decir, ¿la expresión diferencial aumenta o disminuye progresivamente con las etapas?. Para determinar la importancia de esas diferencias, realizamos una prueba de Wilcoxon con signo entre la expresión génica individual en las cuatro etapas. Ver más detalle de los resultados en la sección 4.1.

3.3.4. Construcción de redes de coexpresión genética

Como se mencionó anteriormente, inferir las correlaciones existentes entre distintos pares de genes, o entre genes y otros elementos, es crucial para determinar los procesos asociados a la evolución del CRcc. La inferencia global de dichas interacciones en conjunto, forman una **red de coexpresión**. Para construir las redes utilizamos la medida de dependencia estadística de información mutua (IM). Con esta relación podemos cuantificar la coexpresión entre genes. Para realizar los cálculos de IM usamos la implementación ARACNe [94] como se ha descrito previamente ([3, 4, 37, 44]). Para más detalle del contexto sobre IM ver A.3. En este trabajo determinamos todas las interacciones gen-gen en el transcriptoma completo. Con este procedimiento inferimos cinco redes, una para cada etapa de progresión y otra para el fenotipo control.

Con el fin de obtener las interacciones de mayor relevancia (dada por sus valores de IM) en cada fenotipo se tiene que establecer un umbral de corte. Aquí hay que considerar el problema de dispersión de redes, es decir, determinar el número de aristas significativas que representan mejor la estructura de red consistente con los datos. En este sentido hay un trabajo previo para soportar nuestras consideraciones ([33]) . Para evitar los posibles efectos de tamaño de red, decidimos efectuar cortes de interacciones mayores que abarcan varias escalas muy por encima de nuestros umbrales de trabajo. Los umbrales de corte van desde las 100 interacciones más altas hasta 1M de interacciones principales, es decir, cinco órdenes de magnitud en el tamaño de la red. Usamos esos puntos de corte para evaluar si los fenómenos bajo estudio, como las tasas *-cis*/*-trans*, se debían realmente al tamaño de la red.

Las visualizaciones de red se realizaron en `Cytoscape` versión 3.8.1 [132], así como la biblioteca `iGraph` implementada en `Python` [26].

3.3.5. Intersección entre las etapas

Con el fin de identificar las estructuras de red que se conservaban a lo largo de las etapas, así como aquéllas que se perdían entre etapas, comparamos las diferencias e intersecciones entre la red de control y cada etapa de progresión. Primero, observamos las diferencias de red, es decir, aquellas interacciones gen-gen que no se comparten entre fenotipos. Al mismo tiempo, observamos las interacciones genéticas compartidas entre la red de control y cualquier etapa de progresión de CRcc. Adicionalmente, buscamos las interacciones que se conservan entre todos los fenotipos (intersección), y también, entre los estadios de cáncer únicamente. Realizamos una intersección multigrupo con el fin de obtener la subred integrada por aquellos enlaces compartidos por todos los fenotipos, y también la subred de solo CRcc. Ver detalles en 4.1.4.

3.3.6. Inferencia de comunidades

La estructura de las redes de coexpresión de CRcc para las 100K interacciones más fuertes en términos de IM, es altamente compleja: miles de genes correlacionados entre

sí, formando estructuras a menudo difíciles de interpretar.

Para facilitar el análisis de este tipo de redes, una herramienta sumamente utilizada es la descomposición de las redes en sub-redes o *comunidades*. Esto es, conjuntos de nodos fuertemente conectados entre ellos y débilmente conectados con el resto de elementos de la red. Se ha visto que la estructra de las comunidades a menudo refleja también detalles funcionales [3, 4, 45]. Para la detección de comunidades en redes se utilizó el algoritmo Infomap [122], tal como se implementa en [3, 4, 5].

3.3.7. Inferencia biológica

Si bien es cierto que el análisis de las redes de coexpresión, así como la estructura de las comunidades dentro de dichas redes nos arroja información relevante sobre los fenotipos bajo estudio, esta información resulta insuficiente para poder integrarla con el conocimiento biológico que se tiene del CRcc. Para subsanar esta situación, una de las herramientas más utilizadas es el *enriquecimiento funcional*. Esto es, el cálculo de la probabilidad de que una red o comunidad obtenida bajo nuestro flujo de trabajo, se asocie o no a un proceso biológico conocido. Los procesos biológicos, así como los genes asociados a dichos procesos fueron obtenidos de la base de datos *Gene Ontology*, en particular del árbol *Biological Process*.

El análisis de enriquecimiento funcional se realizó con la API de `g:profiler` [112] en `Python`. El paquete `g:Profiler` utiliza una prueba hipergeométrica para medir la importancia de un término funcional (biológico) en la lista de genes de entrada [116]. Además, se realizaron varias correcciones estadísticas mediante el algoritmo `g:SCS` implementado en `g:Profiler` con un nivel de significancia de 0.05; y una tasa de FDR de 0.05. Vale la pena notar que para considerar la estructura de la red en el enriquecimiento funcional, se aplicó el algoritmo `g:SCS` sobre las comunidades en las redes, y no sobre las redes completas.

Para proporcionar una manera clara y fácil de reproducir los resultados mostrados en este sección, se publicaron las cinco matrices de expresión y todo el código en `https://github.com/josemaz/kidney-stages`. En este repositorio se encuentra todo el flujo de trabajo, desde la descarga de datos hasta el enriquecimiento funcional.

3.4. Análisis de datos gen-microRNA

Para llevar a cabo este protocolo, hemos implementado una metodología analítica simplificada. En la Figura 3.2 se muestra una representación gráfica del flujo de trabajo. En las siguientes subsecciones, expondremos los diferentes aspectos y detalles detrás del flujo de trabajo que acabamos de presentar.

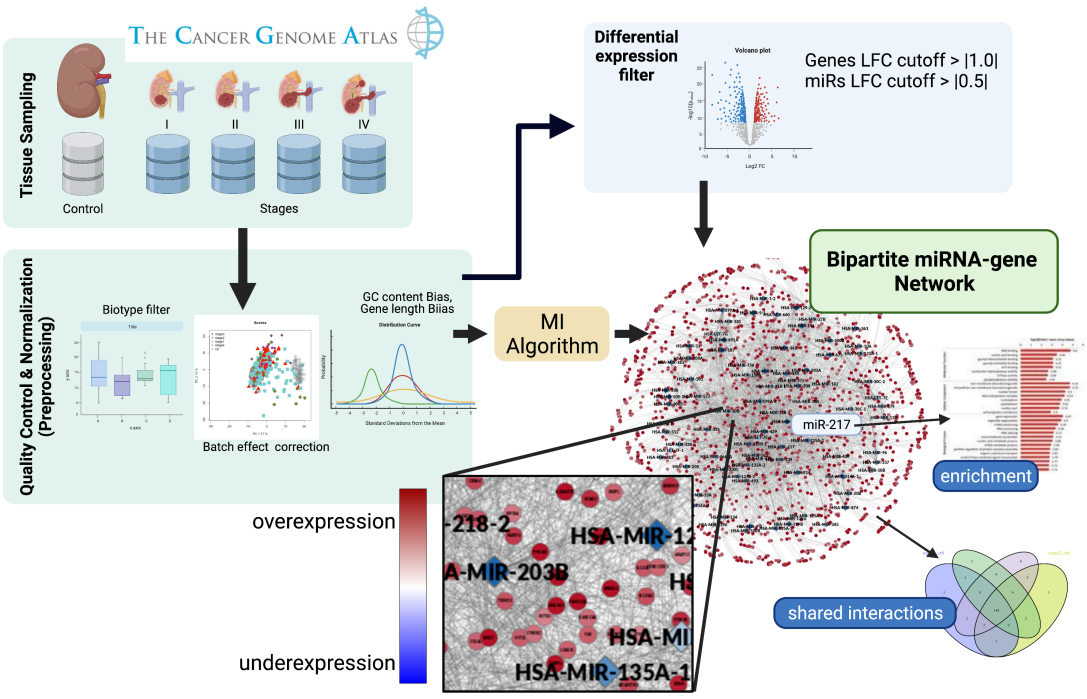

Figura 3.2: *Flujo de trabajo para el análisis de datos de miRNA-seq.* En la figura se muestran los diferentes protocolos que se realizaron durante este análisis. El proceso general se puede dividir en tres etapas: 1) descarga, filtrado y normalización de los datos, 2) generación de las redes de coexpresión y 3) armonización de las datos para el análisis biológico

3.4.1. Procesamiento de datos

Descargamos los perfiles de expresión con miRNA-seq para las muestras originales de RNA-seq (ver tabla 3.2). Posteriormente, compilamos dos conjuntos de datos principales: 1) cuantificación de la expresión de miRNAs (lecturas por millón) y 2) cuantificación de la expresión de isoformas (miRNAs), que contiene información detallada sobre las especies transcritas (como coordenadas asignadas) para cada transcripción. Esto se puede utilizar para obtener información a nivel de miRNA maduro.

Los índices de ambos conjuntos de datos se armonizaron para que coincidieran con los códigos de los pacientes. Usamos una *clave maestra* para aglomerar cuentas de lecturas (*counts*) sin procesar, tanto para RNAseq como miRNAs. En la Tabla 3.2 se puede ver un resumen de los datos preprocesados.

Tabla 3.2: *Número de muestras en el análisis de miRNAs.* Los datos armonizados se cuantifican en esta tabla, siendo la etapa I en CRcc el estadio más abundante de muestras correctamente etiquetadas y la etapa II, se observa como el estadio menos poblado.

Control	**Etapa** I	**Etapa II**	**Etapa III**	**Etapa IV**
71	251	55	122	81

3.4.2. Información clínica

Como en los datos de RNAseq, aquí procesamos la información clínica directamente del proyecto TCGA-KIRC. Clasificamos todas las muestras por su variable *tumor_ stage*. Se eliminaron las muestras con estadios no reconocidos (por ejemplo, *NAs*, *not reported*, etc). La biblioteca que se utilizó para recuperar los datos del repositorio de TCGA fue `TCGAbiolinks` (V 2.24.1).

3.4.3. Pre-procesamiento de datos

Preprocesamos los datos de genes y miRNAs de la siguiente manera: 1) eliminamos genes sin anotaciones en la base de datos de BioMart, 2) eliminamos genes con más del 50 % de ceros en sus cuentas por muestra y 3) genes con una expresión media de menos de 10 cuentas (*counts*) también fueron eliminados. Para la corrección de sesgos, utilizamos el paquete `EDASeq` en `R` (V 2.30.0) (36). En resumen, eliminamos los sesgos por el contenido de GC, por longitud del gen y por el biotipo. Finalmente, para corregir posibles efectos de lote, utilizamos el método ARSyN, implementado en la biblioteca `NOIseq` (V 2.40.0) en `R` [102].

Después de que se aplicaron todos los filtros y se eliminó el sesgo, el número total de miRNAs para el análisis fue de 275; en tanto, el total de genes fue de 16,224. Esas fueron las entidades utilizadas para inferir las redes gen-miRNA y para realizar el análisis de expresión diferencial correspondiente. En la Tabla 3.3 se muestra un resumen de los fenotipos, las unidades de conteo y la cantidad de genes y miRNAs.

3.4.4. Expresión diferencial de genes y miRNAs

El análisis de expresión diferencial se implementó utilizando el paquete DESeq en `R` (V 1.8.3) [92]. Aquí consideramos genes expresados diferencialmente (GDE) con los siguientes filtros: $|LogFC| > 1.0$ y valor-$p < 1.0^{-5}$ corregido por FDR. Por otro lado, para los miRNAs expresados diferencialmente (mDE), los filtros fueron $|LogFC| > 0.5$ y valor-$p < 1.0^{-5}$. Vale la pena notar que los cortes de *Log Fold Change* (LFC) dependen de las distribuciones de datos empíricos y los rangos dinámicos asociados de las mediciones de las variables. Aunque la secuenciación de RNAseq y miRNAseq fue con

Tabla 3.3: *Características de genes y miRNAs.*. Los fenotipos fueron filtrados con el fin de conservar sólo los genes que codifican para proteínas, en el caso de los miRNAs se unificaron en miRNAs con capacidad de ser maduros.

	Genes	miRs
Tamaño	16,224	275
Unidades	HTSeq Counts	Reads per-million-miRNA-mapped
Fenotipos	5	5

aproximadamente la misma tecnología (Illumina NGS Sequencing), existen diferencias en las tasas de captura, las llamadas de variantes (*calling*) y las anotaciones de variantes. Aún más importante, puede haber diferencias en la abundancia natural de estos dos tipos de transcritos en las muestras. El impacto de estas consideraciones se intenta disminuir en la etapa de corrección de sesgos.

Comparamos el conjunto de datos no tumorales (NT o Control) con todas las etapas de progresión (Etapa I, Etapa II, Etapa III y Etapa IV). Además, para rastrear la evolución de la progresión del tumor, también realizamos un análisis de expresión diferencial entre etapas contiguas (contraste de progresión) NT-EtapaI, EtapaI-EtapaII, etc. Para visualizar los GDE y mDE, construimos gráficos de volcanes para cada contraste con las especificaciones predeterminadas del paquete `EnhancedVolcano` (V. 1.14.0) (https://github.com/kevinblighe/EnhancedVolcano).

Cuantificamos el número de GDE y mDE que aparecían para cada contraste. También calculamos esos GDE y mDE únicos en cada contraste, así como los GDE/mDE compartidos en todos los contrastes. El código para reproducir estos análisis se puede encontrar en el siguiente repositorio: `https://github.com/josemaz/kirc-mirna`.

3.4.5. Inferencia de red

Para analizar el papel potencial que desempeñan los miRNAs en el programa de expresión génica, inferimos cinco redes gen-miRNA, una para muestras de tejido control adyacente al tumor y una para cada etapa de progresión del tumor. Todas las redes se infirieron utilizando información mutua (IM) como medida de dependencia estadística. Los valores de IM se calcularon para todas las parejas gen-miRNA ($275 \times 16,227 \approx 4.5$ *millones* de interacciones por pares) en cada fenotipo. Debido al costo computacional de estos cálculos, implementamos un código de computadora que explota las características paralelas de las máquinas y disminuye el tiempo de este proceso exponencialmente. Este código *multihilo* acelera las inferencias acorde al número de procesadores en la máquina de forma casi lineal. Además, esta basado en el algoritmo original de ARACNe [94] dando compatibilidad completa con la metodología. Destacamos la importancia de

esta aportación, porque este código se ha seguido manteniendo y actualizando dentro del grupo de investigación, siendo bastante útil para minimizar el tiempo de desarrollo en nuevos trabajos. El código en paralelo para inferir redes basadas en IM se puede encontrar en `https://github.com/josemaz/aracne-multicore`.

3.4.6. Filtrado y visualización de redes

Para encontrar genes desregulados que son posibles dianas de los miRNAs, utilizamos GDE y mDE como filtros de red. Conservamos las 100K interacciones miRNA-gen con IM más alta para capturar las relaciones de coexpresión más relevantes para cualquier fenotipo dado. Conservamos solo aquellas interacciones miRNA-gen en las que el micro-ARN y su objetivo tienen expresiones diferenciales opuestas: miRNA sobreexpresado y gen subexpresado, así como el caso contrario: miRNA subexpresado y gen sobreexpresado, recordando que el objetivo es buscar interacciones canónicas entre miRNAs y genes.

Finalmente, analizamos la fracción de interacciones miRNA-gen conservadas y la fracción de interacciones únicas para cada fenotipo. Las visualizaciones de red se realizaron con `Cytoscape` 3.8.2 (39).

3.5. Análisis sobre los datos de metilación

En esta sección, analizamos conjuntamente la metilación del ADN para 383,862 sitios CpG y los datos de expresión en 16,170 genes provenientes tanto del CRcc como del tejido adyacente normal. Dividimos las muestras de CRcc según la etapa de progresión: 24 no tumorales, 158 muestras para la Etapa I, 31 para la Etapa II, 72 para la Etapa III y finalmente 57 para la Etapa IV. En la figura se muestra una representación gráfica de este flujo de trabajo.

3.5.1. Adquisición de datos

Utilizamos datos del consorcio The Cancer Genome Atlas (TCGA) como fuente. Descargamos datos de expresión genética (secuenciación de ARN) y metilación de ADN (arreglos de metilación de alta densidad por conversión de bisulfito). Más detalle en 1.3.4.

Los datos de metilación en TCGA fueron descargados como valores beta (β), que miden el nivel de metilación del ADN en sitios CpG conocidos por sondas arregladas en matrices dentro del chip *Illumina HumanMethylation450* (HM450). Estos valores se calculan a partir de las intensidades del arreglo (datos de nivel 2) como $\frac{M}{M+U}$, donde M corresponde a las sondas metiladas (marcadas con la conversión de bisulfito) y U a las no metiladas [165].

Con el fin de realizar las comparaciones adecuadas, se armonizó el conjuntos de datos por el código de los pacientes, tanto para los valores beta de metilación como para RNA-

seq. Esta es la razón por la cual el número de muestras no corresponde con el número de muestras de RNA-Seq original. La descarga, la anotación y el análisis de bajo nivel

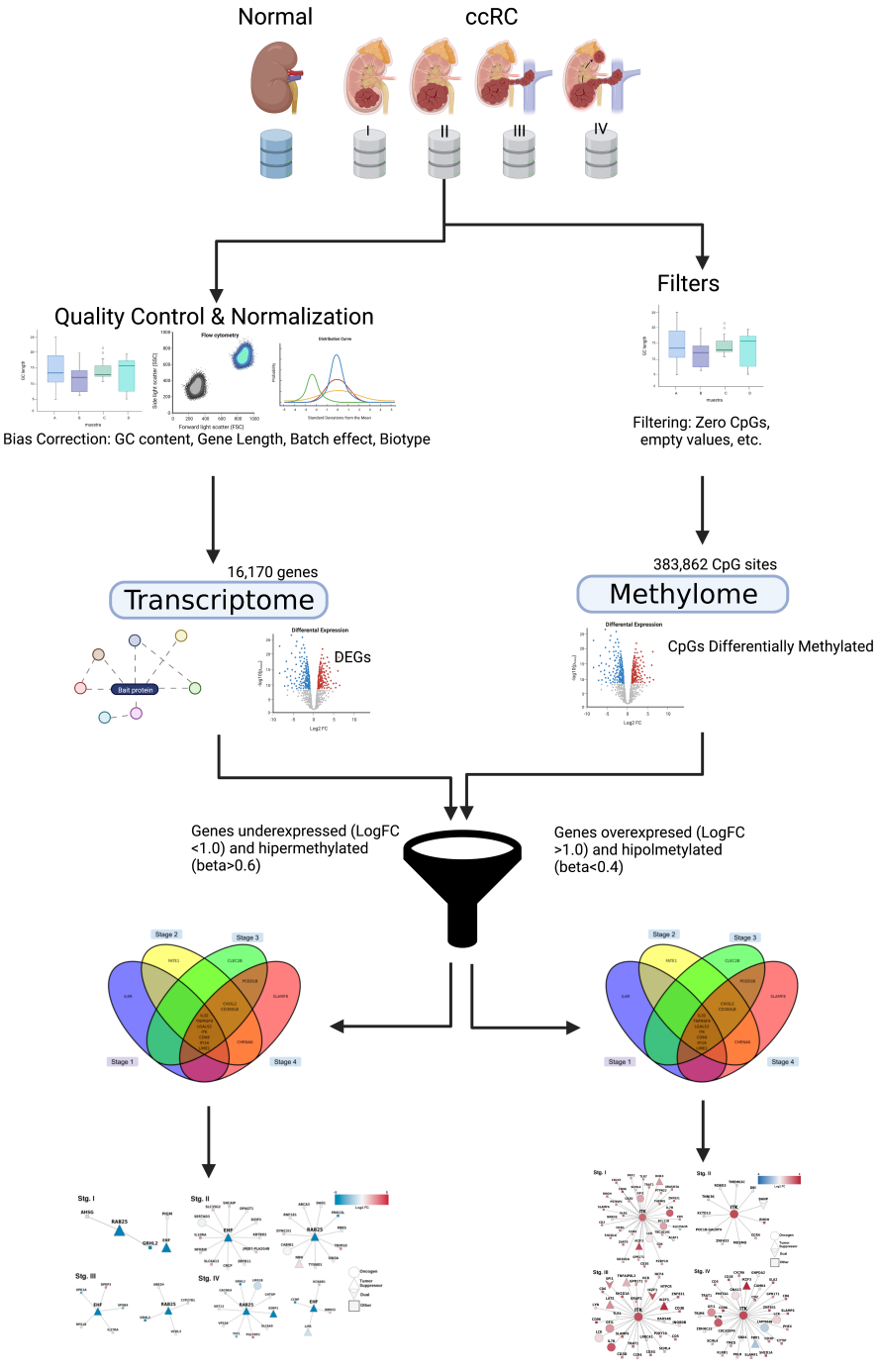

Figura 3.3: *Flujo de trabajo para el análisis metilación-gen.*

Figura 3.3: En la figura se muestra como se integraron dos fuentes de datos, a saber, el transcriptoma y el metiloma de CRcc. Además, se destacan algunos de los parámetros que se utilizaron para el algoritmo de filtrado. Finalmente, este flujo concluye con los diagramas de Venn y en las redes de los genes metilados. Estos resultados se analizan en secciones posteriores (4.3). Vale la pena señalar, que en el modelo planteado en esta sección, las relaciones de los genes fueron determinadas también por su capacidad oncogénica ó supresora de tumores.

se realizó utilizando la biblioteca `TCGAbiolinks` de `R` [21]. Igualmente, procesamos la información clínica directamente del proyecto TCGA-KIRC. Clasificamos todas las muestras por la variable *tumor_stage*. Las muestras se limpiaron para excluir aquellas muestras con etapas o valores no informados. En este caso, la biblioteca `TCGAbiolinks` también se utilizó para recuperar datos clínicos de los pacientes.

3.5.2. Pre-procesamiento de datos

Preprocesamos los datos de RNA-seq de la siguiente manera: 1) Eliminamos genes sin anotación en BioMart [135], 2) Eliminamos genes con más del 50 % de recuentos cero por muestra, 3) Eliminamos genes con expresión media menor a 10 conteos. Para las correcciones de sesgo en la secuenciación, usamos `EDASeq` en `R` [119], filtrando sesgos en el contenido de GC, la longitud del gen y el biotipo. Finalmente, para corregir los posibles efectos por lotes, utilizamos el método ARSyn, implementado en R como una función de la biblioteca `NOIseq` [102]. Los datos de metilación se limpiaron eliminando los sitios CpG a los que les faltaba al menos un valor beta.

La asociación entre CpGs y genes se realizó manualmente con la primera aparición en una anotación predefinida creada por `TCGAbiolinks`. Después de aplicar todos los filtros y procedimientos de eliminación de sesgos, el número total de CpGs para la evaluación fue de 383,862; mientras que el total de genes fue 16,170. Esas entidades se utilizaron para inferir diferentes relaciones metilación-gen y realizar los análisis correspondientes.

3.5.3. Expresión diferencial

El análisis de expresión diferencial se calculó con el paquete `DESeq` de `R` [92]. Consideramos GDE con los siguientes filtros: $LogFC > 2.0$ y $FDR < 0.05$. Comparamos el conjunto de datos no tumorales (NT) con todas las etapas de progresión (Etapa I, Etapa II, Etapa III y Etapa IV). Además, contrastamos GDE en etapas consecutivas de CRcc (ver 4.3.2).

3.5.4. Sitios CpG diferencialmente metilados

El análisis de metilación diferencial (DM) se realizó mediante un método basado en la media implementado en el paquete `TCGAbiolinks` de `R`. Consideramos a los CpGs metilados diferencialmente (CpGs-DM) con un límite de diferencia media de 0.15 y un valor de p de 0.05 (prueba de Wilcoxon). Todos los gráficos de volcanes para cada contraste (NT_Stage1, NT_Stage2, NT_Stage3, NT_Stage4) en la progresión de CRcc se pueden calcular con el código fuente proporcionado en este trabajo. El umbral para identificar CpGs-DM era un valor beta por debajo de 0.4 ó por encima de 0.6.

3.5.5. Genes dirigidos por metilación

Agrupamos todos los CpGs para cada gen en la posición de los promotores. Evaluamos si al menos un sitio CpG para un gen dado estaba metilado diferencialmente. Después, etiquetamos este gen como candidato para ser un gen diferencialmente metilado (DMG). El otro criterio inclusivo en estos candidatos fue que el gen mismo se expresara diferencialmente. Consideramos genes hipermetilados cuando su mediana de metilación en todos sus CpGs fue > 0.6. Por el contrario, un gen con una mediana de metilación de < 0.4 se consideró hipometilado.

Vale la pena notar que este procedimiento de filtrado para los CpGs-DM se toma como *proxy* preliminar. Sin embargo, las condiciones suficientes para determinar el estado de metilación en un gen se dieron tomando los valores medianos de todos los CpGs del promotor dentro de cada gen.

3.5.6. Oncogenes y Supresores tumorales

Una pregunta fundamental con respecto al papel de las marcas de metilación en la progresión de CRcc es si esas marcas alteran o no los genes relacionados con el cáncer. Por lo tanto, investigamos aquellos genes con la propiedad de ser oncogenes o genes supresores de tumores. Se obtuvo una lista completa de oncogenes de la base de datos *Human OncoGene* [88]. El catálogo correspondiente a los supresores tumorales se descargó de la base de datos *TSGene*[163]. Combinamos los genes que presentaban funciones de oncogen y supresor de tumores, para después etiquetarlos como *both* en la base de datos de funciones tumorales [160].

3.5.7. Inferencia de redes

Para analizar el papel de los GDM y su programa de expresión, construimos cuatro redes GDM-GDM para cada contraste de progresión. Todas las redes se infirieron utilizando el valor de información mutua (IM) como medida de correlación. El valor de IM se calculó sobre los valores de expresión de todos los pares de GDM para cada fenotipo.

Además, para validar nuestros resultados, obtuvimos redes utilizando un corte de 1K, 10K y 100K aristas superiores. Anteriormente hemos descrito las consideraciones para

establecer este parámetro [33, 158]. La visualización de redes se realizó con `Cytoscape` (V. 3.9.1) [132].

3.5.8. Análisis de enriquecimiento

En este caso, para identificar funciones biológicas estadísticamente enriquecidas empleamos *Gene Onthology* (GO). Filtramos los resultados con valores-$p < 0.01$. Ejecutamos este análisis usando tanto el servidor web `Gprofiler2` como el paquete de `R` llamado `Gprofiler2` ([160]).

El código completo para utilizar ó repetir este flujo de trabajo fue publicado en `https://github.com/josemaz/kirc-methyl`. También agregamos un protocolo `snakemake` para lograr la reproducibilidad y mejorar las prácticas de escritura de código fuente científico.

Capítulo 4

Resultados y discusión

4.1. Panorama de expresión genética en Cáncer Renal de Células Claras

4.1.1. La expresión diferencial es similar entre las etapas de CRcc

Como primer acercamiento a los datos de expresión genética en este tumor revisamos el panorama de expresión diferencial. La idea principal fue tener una cuantificación de los genes que se distinguen inicialmente entre el control y las cuatro etapas de progresión en CRcc. De manera interesante, esta cantidad es similar en los cuatro contrastes (etapas).

La Figura 4.1 muestra los gráficos de volcanes para genes expresados diferencialmente en las cuatro etapas. Es visible una gran similitud en la distribución de genes y el rango de valores para las cuatro etapas. La cantidad de genes expresados diferencialmente también es similar. La tabla 4.2 muestra la correlación de rangos de Spearman entre las cuatro etapas. Como se puede observar, la ρ de correlación de Spearman es mayor a 0.9 en todos los casos. Esto evidencia la similitud entre la cantidad de genes expresados diferencialmente en las cuatro etapas.

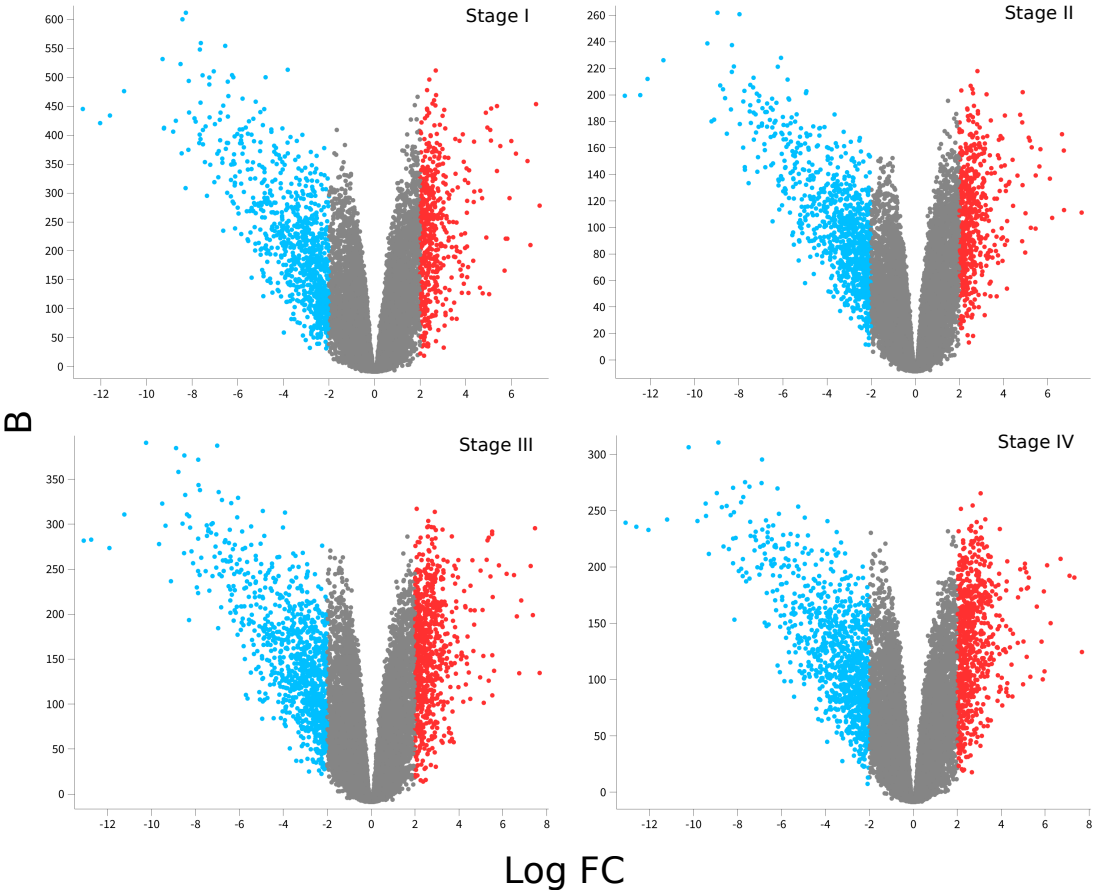

Figura 4.1: *Expresión génica diferencial para cada etapa de CRcc.* En estas gráficas de volcanes, se representa la expresión diferencial entre cada etapa y las muestras de control. Los puntos rojos representan genes sobreexpresados, mientras que los subexpresados están en azul. Es preciso tomar en cuenta que los genes subexpresados están más ampliamente distribuidos que los sobreexpresados, y los valores de $LogFC$ son similares en las cuatro figuras; sin embargo, el estadístico B cambia según la etapa de CRcc.

4.1.2. Los genes *SLC6A19* y *PLG* muestran una expresión progresivamente decreciente

A pesar de que los gráficos de los cuatro volcanes son similares y la correlación de Spearman entre todos los estadios es alta, algunos genes parecen expresarse según los estadios de progresión del tumor, como es el caso de los genes observados en la Figura 4.2. Curiosamente, *SLC6A19* y *PLG*, ambos muestran una disminución notable en su

expresión durante las etapas de progresión (Figura 4.2).

4.1.3. Los genes *SAAC2-SAAC4* y *CXCL13* muestran una expresión progresivamente creciente

Con respecto a la sobreexpresión de genes durante la progresión de CRcc, encontramos que solo dos genes, a saber, los genes *SAAC2-SAAC4* y *CXCL13*, se sobreexpresan acorde a las etapas de progresión del tumor. Esta condición se puede observar en el lado izquierdo de la Figura 4.2. Vale la pena tener en cuenta que en los cuatro casos, esos genes se expresan diferencialmente entre el control y cualquier etapa, pero también entre etapas consecutivas.

Además, realizamos un análisis de expresión diferencial multigrupo, para observar si dicha diferencia en la expresión génica también aparecía entre etapas. En todos los casos, estos genes se expresan diferencialmente. Sin embargo, entre la etapa III y IV, el *Log Fold Change* se estableció en 0.5. Esto significa que los valores de expresión de los cuatro genes son diferentes, pero, no tan diferentes como en las etapas anteriores. Esto podría deberse a las características clínicas e histopatológicas que pueden compartir ambas etapas.

Hasta donde sabemos, no se ha reportado previamente que el gen *SLC6A19* esté significativamente subexpresado en el cáncer renal; sin embargo, en el Human Protein Atlas, la subexpresión de *SLC6A19* se reporta como probable biomarcador para el cáncer renal ([110]). *SLC6A19* se expresa en gran cantidad en el tejido renal ([39]). Por lo tanto, su baja expresión puede traer consecuencias funcionales relevantes.

El gen *PLG* también presenta una disminución notable a medida que avanzan en las etapas (Figura 4.2). Anteriormente, se ha mostrado que *PLG* ha disminuido y es un posible biomarcador para el carcinoma renal ([93]; [162]).

En el caso de la sobreexpresión de *CXCL13* ([68]), se ha encontrado que está relacionada con células inmunes infiltrantes de tumores (células T, por ejemplo), así como con mal pronóstico en CRcc. En nuestro caso, no solo encontramos el gen sobreexpresado, sino que también aumentó progresivamente a través de las cuatro etapas. Este gen destaca en este trabajo por su posible participación en la formación de folículos inmunológicos dentro de los tumores, siendo que el CRcc se había etiquetado como un tumor con alta infiltración de células del sistema inmune.

En cuanto al gen *SAA2-SAA4*, se ha observado su sobreexpresión desfavorable en cáncer renal, pero a la vez favorable en cáncer de mama [109]. *SAA2-SAA4* es una fusión natural entre dos genes de amiloide sérico (A2 y A4). La sobreexpresión de *SAA2-SAA4* también se ha asociado con tumores cerebrales metastásicos derivados del carcinoma papilar de tiroides ([127]). También se ha asociado con metástasis hepática de tumor colorectal ([126]). El hecho de que la sobreexpresión de *SAA2-SAA4* se haya asociado con metástasis de un tumor primario vecino es materia de investigación adicional. Sin embargo, vale la pena mencionar que la expresión de este gen aumenta progresivamente a través de las etapas de progresión de CRcc.

Hasta donde sabemos, esta es la primera vez que se informa sobre la expresión de *SAA2-SAA4*, *CXCL13*, *PLG* y *SLC6A19* y cómo se expresa de manera diferencial a través de etapas de progresión en el carcinoma renal de células claras, lo que muestra una posible nueva línea de investigación relacionada con el progreso genómico de las alteraciones en CRcc.

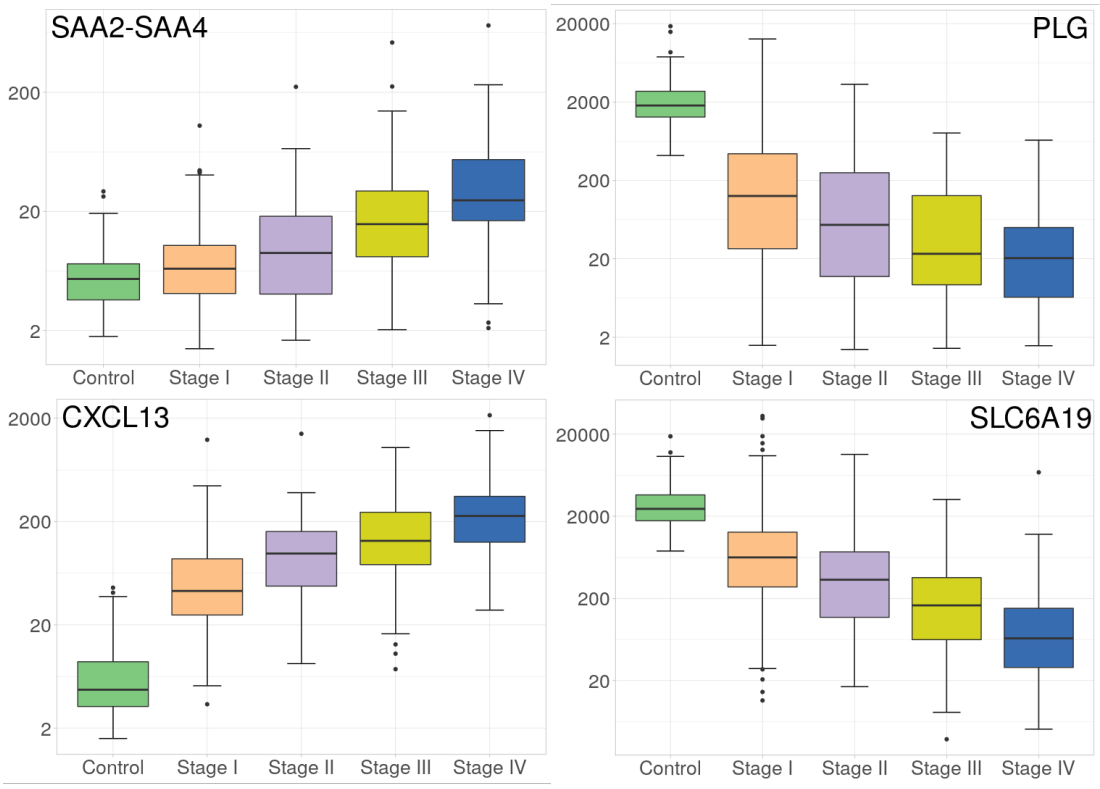

Figura 4.2: *Aumento y disminución progresiva de la expresión de cuatro genes en las diferentes etapas de CRcc.* Estos diagramas de caja muestran la expresión promedio de los genes *SAAC2-SAAC4* y *CXCL13* (izquierda) y los genes *SLC6A19* y *PLG* (derecha). Diferentes colores representan las etapas de progresión. Observe que el eje Y (expresión génica) está, en todos los casos, representado en escala logarítmica.

4.1.4. La red de *control* es topológicamente diferente a cualquier red tumoral

Posterior al enfoque de los genes individuales comenzamos a estudiar el sistema biológico global. Entonces, construimos las redes de coexpresión y encontramos que todas

son sustancialmente diferentes entre ellas. Sin embargo, la red de control presenta una diferencia más llamativa en términos de sus características topológicas. La red de control tiene una mayor proporción de interacciones *-trans* (intercromosómicas), mientras que para cualquier estadio de cáncer la cantidad de interacciones intracromosómicas (*-cis*) son más abundantes (Figura 4.4).

Entre los parámetros de red más importantes a examinar se encuentra la distribución de grados $p(k)$. Es bien sabido que el gráfico k contra $p(k)$ y sus parámetros para el ajuste de curvas pueden reflejar varias propiedades relacionadas con el propio sistema. En la Figura 4.3 se muestra la distribución de grados por gen. En el caso del corte de las 10K aristas superiores, podemos observar cómo en todos los casos la distribución se ajusta bien a una distribución de ley de potencias ($y = ax^b$). Las diferencias se observan en las distintas pendientes de los ajustes de las curvas, así como a nivel de parámetros. La pendiente de la distribución de grados de la red de control (verde claro) es la más baja (-1.842), en comparación con las etapas CRcc. La Tabla 4.1 contiene los parámetros para el ajuste de curvas no lineales de las cinco redes. Este resultado puede explicar que la comunicación de largo alcance sea una característica de un fenotipo saludable.

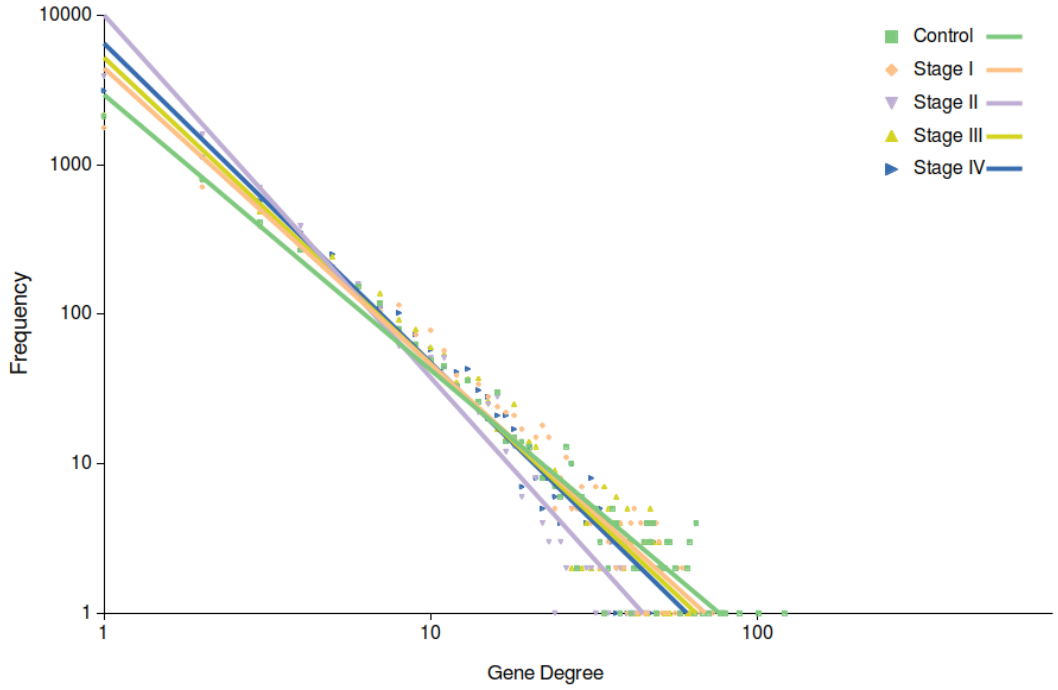

Figura 4.3: *Distribución de grados (de nodo) en las cinco redes.* En esta gráfica, los puntos corresponden a la distribución de grados para cada fenotipo. El código de colores es el mismo que el de la Figura 4.2. También se muestra el ajuste de la curva ($y = ax^b$) para cada distribución de grados. Observe que la pendiente de distribución de la red de control (verde claro) es la más baja.

Tabla 4.1: *Modelo de ajuste en la relación de interacciones principales.* En esta tabla se resumen los estadísticos calculados para el ajuste lineal de la curva sobre la distribución de grados

Parametro	**Control**	**Etapa I**	**Etapa II**	**Etapa III**	**Etapa IV**
a	2941	4430	10047	5215	6490
b	-1.842	-1.982	-2.426	-2.052	-2.132
Correlation	0.992	0.981	0.973	0.98	0.987
R-square	0.935	0.931	0.973	0.98	0.987

4.1.5. Diferencias estadísticas en las redes de coexpresión

4.1.5.1. Existe una coexpresión preferencial -*cis* en las redes de CRcc

Los componentes gigantes conectados de cada red se representan en la Figura 4.4. Los genes están coloreados según el cromosoma al que pertenece cada gen. En la red de control, los genes se coexpresan con genes de cualquier cromosoma, con una alta prevalencia de interacciones -*trans*. Por el contrario, para las etapas tumorales, en todos los casos hay expresión intracromosomal (-*cis*) de manera preferencial. Esto también se refleja en los gráficos de barras en la parte inferior derecha de la Figura 4.4. Las barras naranjas representan el número de interacciones -*trans*, mientras que los enlaces -*cis* están representados por barras azules.

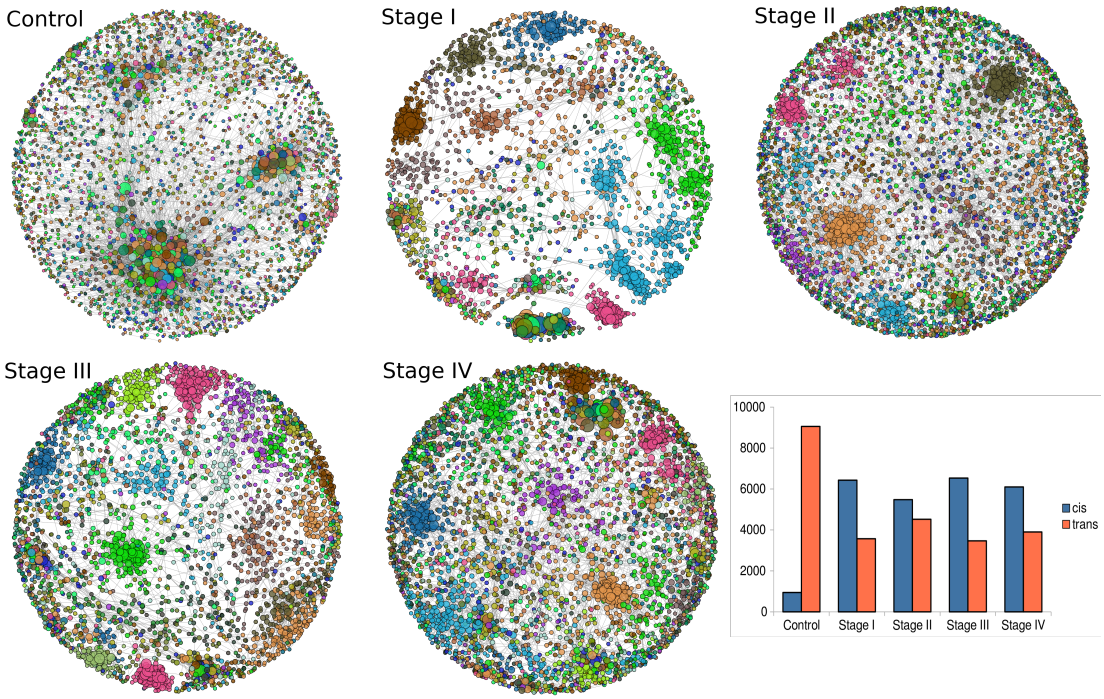

Figura 4.4: *Topologías de red de CRcc por etapa.* Las cifras de arriba a abajo corresponden al mayor componente conectado de control, etapa I, etapa II, etapa III y etapa IV, respectivamente. El color de los nodos corresponde al cromosoma al que pertenece cada gen. El gráfico de barras representa la proporción de interacciones -*cis* (azul) y -*trans* (naranja).

4.1.5.2. Las proporciones *-cis/-trans* no vuelven a guiar las etapas de progresión

En trabajos previos del grupo ([44]), se ha demostrado que la relación *-cis/-trans* aumenta con la severidad de los subtipos de cáncer de mama, siendo Luminal A, Luminal B, HER2+ y Basal el orden de tasas *-cis/-trans*. En ese momento, también se mostró que la red de control es el único gráfico que contiene más interacciones *-trans* que *-cis*.

Intuitivamente, uno podría esperar (basado en nuestra experiencia previa con el cáncer de mama) una disminución progresiva en el número de interacciones *-trans*, comenzando desde el número más grande en la red de control y disminuyendo a lo largo de las etapas de CRcc. Sin embargo, este no es el caso, como se puede apreciar también en los gráficos de barras, así como en las redes. La red de CRcc con menos enlaces de coexpresión *-trans* es la etapa III, seguida de la etapa I, la etapa IV y finalmente la etapa II. Sin embargo, la diferencia entre el control y cualquier etapa también es evidente.

4.1.5.3. Las tasa de conexiones *-cis* específicas de un cromosoma son diferentes entre los fenotipos

Una vez obtenida la proporción de interacciones *-cis/-trans* globales, se calcularon las tasas de *-cis* cromosómicas aisladas. Definimos la tasa de *-cis* como el número de aristas *-cis* dividido por el número total de aristas en cada red. Como se puede observar en el diagrama de barras de la Figura 4.5, para la red de control, todos los cromosomas excepto $ChrY$ tienen una tasa de *-cis* < 1, pero en el caso de $ChrY$, todos los fenotipos tienen una tasa de *-cis* > 1. En general, la red en etapa III tiene las tasas más altas de *-cis* a nivel cromosómico.

4.1.6. Las diferencias topológicas no siguen las etapas de progresión

Como primer enfoque, el corte de la red se estableció en las 10K aristas superiores, clasificadas por valores de IM. Cada red contiene un número diferente de genes. Dado que estas redes se obtienen a partir de la expresión génica del tejido renal, uno puede esperar similitudes en términos de genes e incluso interacciones. Además, dado que las redes de estudio se separaron en etapas de progresión, también sería de esperar que las etapas consecutivas fueran más similares entre ellas que con el resto de redes.

En la Figura 4.7, mostramos el número de interacciones compartidas entre fenotipos, así como sus diferencias. Como era de esperar, la red de control es la más diferente en términos de número de enlaces compartidos con las redes tumorales. El porcentaje de divergencia es del 94 % en el caso más similar (etapa I).

Las redes tumorales también difieren enormemente entre ellas, más del 60 % de diferencia en cualquier caso. La red de Etapa II es la más diferente, en términos de número de aristas compartidas. Por el contrario, las redes de etapa I y etapa III son el par

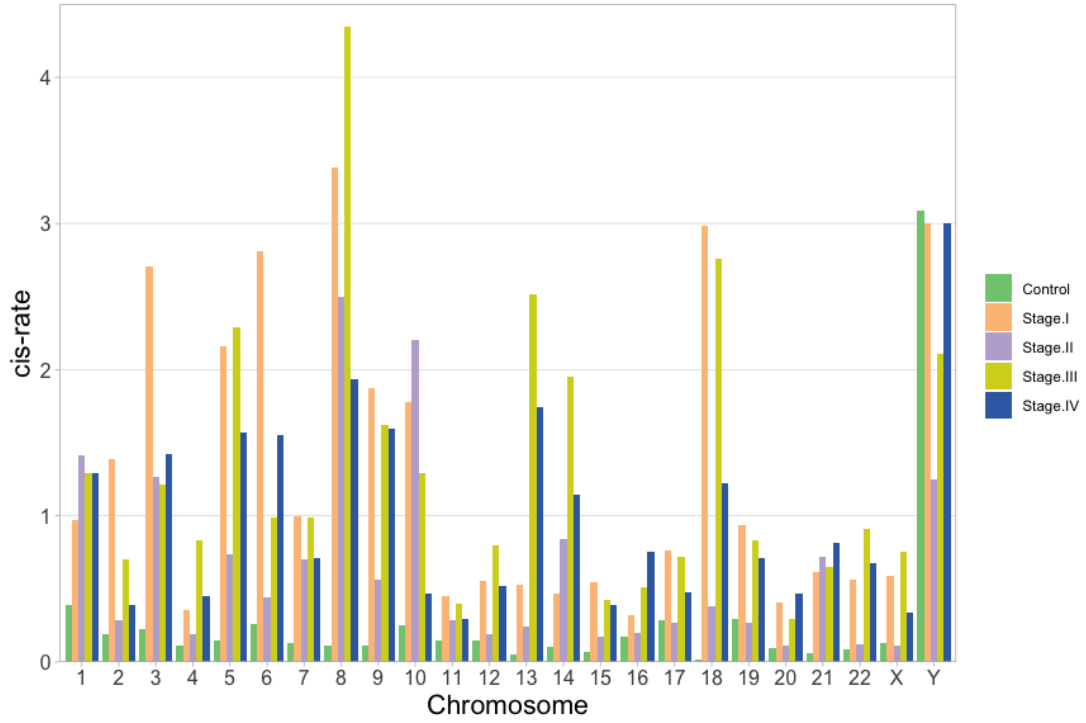

Figura 4.5: *La taza de relaciones -cis (aristas -cis/# de genes) por cromosoma en las 5 redes.* El código de colores es: verde, naranja, violeta, amarillo y azul para control, etapa I, etapa II, etapa III y etapa IV, respectivamente. En todos los casos excepto para ChrY, la relación es inferior a 1 para la red de control.

más similar, incluso la etapa I y la etapa IV mantienen más interacciones compartidas entre ellas (74 %) que con la etapa II.

Este último resultado es sorprendente, teniendo en cuenta la alta similitud en términos de expresión genética diferencial en los cuatro fenotipos (Tabla 4.2). Podemos concluir que las topologías de red y los programas de coexpresión concomitantes no coinciden con las firmas de expresión génica de las etapas de progresión de CRcc.

Tabla 4.2: *Correlación entre el rango de genes expresados diferencialmente para todas las etapas.* En el cálculo de las correlaciones se utilizó el método de Spearman. Como se observa los valores son altos (mayores a 0.9), lo que demuestra las similitudes importantes entre los diagramas de volcanes de la figura 4.1.

CRcc Etapa	Etapa I	Etapa II	Etapa III	Etapa IV
Etapa I	1	0.995	0.974	0.948
Etapa II	0.995	1	0.995	0.958
Etapa III	0.974	0.994	1	0.997
Etapa IV	0.948	0.958	0.997	1

Además, el pequeño número de interacciones genómicas compartidas entre el control y las redes de CRcc también refleja una reorganización radical del programa transcripcional entre la etapa sana y la enfermedad.

Biológicamente, la disminución de las interacciones comunes entre fenotipos es un claro indicativo de que cada una de las etapas de CRcc se comporta de manera diferente. Esto podría ser importante, ya que cada red mapea una instantánea específica del paisaje de coexpresión en diferentes momentos del proceso cancerígeno.

4.1.6.1. La mayoría de las interacciones son especificas del fenotipo

En la Figura 4.6, se representa la intersección de las interacciones de coexpresión para las cinco redes (control y etapas tumorales). Como se puede observar, la mayor cantidad de enlaces pertenecen a los conjuntos no compartidos en las cinco redes. Esto indica que, independientemente del fenotipo, las redes son estructuralmente diferentes. Al igual que en la figura anterior, la mayor diferencia se da en la red de control (9,295 aristas únicas). Se comparten 533 aristas entre los cuatro fenotipos de CRcc. Este es el conjunto de interacciones de coexpresión que aparecen en cualquier estadio del carcinoma renal de células claras.

4.1.7. Topologías de red en diferentes cortes de IM

Dado que la elección de cortes sigue siendo un problema no cerrado en ciencia de redes (el llamado problema de dispersión de redes), decidimos cubrir una amplia gama de cortes para evaluar el resultado observado en las secciones anteriores. Las redes originales fueron podadas (16,000 genes, 130 millones de aristas) en pequeños conjuntos clasificados de información mutua, desde los 100 interacciones hasta el millón de aristas, es decir, cubriendo cinco órdenes de magnitud. En el material complementario de [158], se encuentran las figuras correspondientes a estos resultados.

4.1.7.1. Disminución en proporción de la intersección de redes con respecto a los tamaños de red

En la Figura 4.7 se puede apreciar que la proporción de intersecciones se mantiene en una amplia gama de cortes, entre todos los fenotipos (control y CRcc), así como en

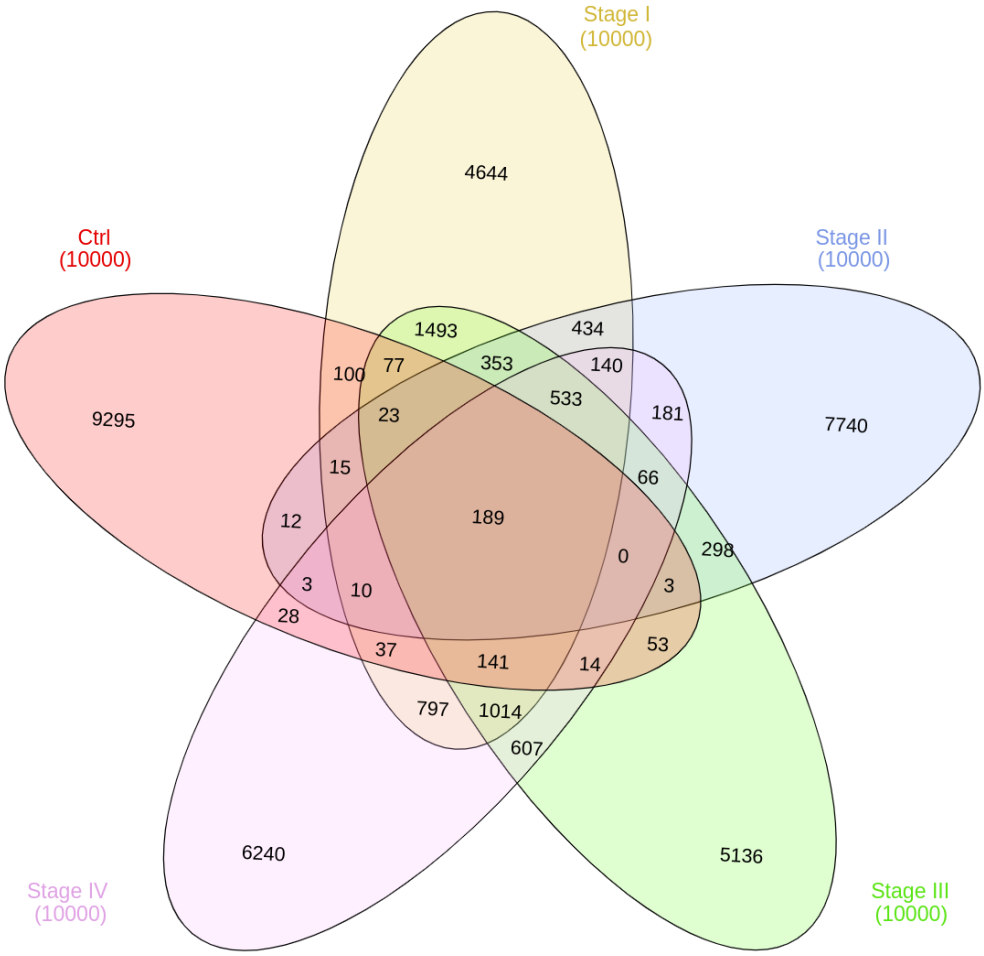

Figura 4.6: *Intersección de aristas en todas las redes.* El diagrama de Venn muestra, en cada conjunto, el número de aristas por fenotipo. El número refleja los genes compartidos entre las redes, así como las interacciones específicas de la red. Observe que de 10K interacciones, solo se comparten 189 bordes entre las cinco redes.

redes de sólo CRcc. También, se puede apreciar claramente la disminución consecutiva de la proporción de enlaces compartidos a medida que las redes crecen en tamaño.

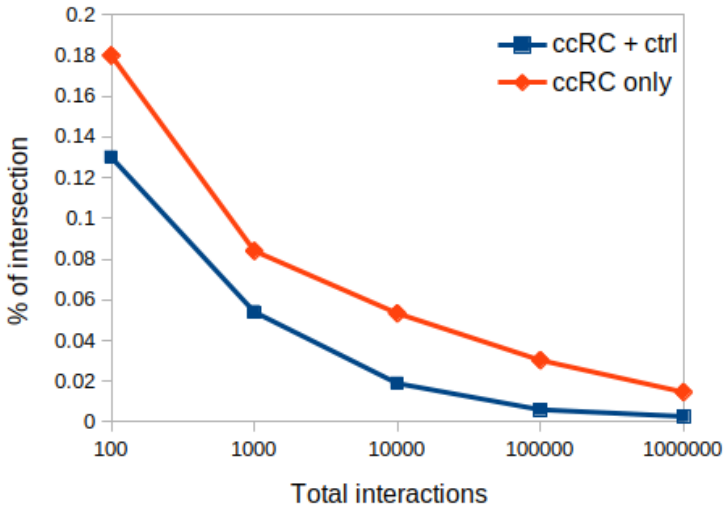

Figura 4.7: *Proporción de la intersección de redes en diferentes cortes de red.* En este gráfico, se representa la proporción de la intersección de la red entre las cuatro etapas CRcc (rombos naranjas) y aquéllas con red de control (cuadrados azules). El eje X representa diferentes valores de corte de red.

4.1.7.2. Las diferencias de conectividad cromosómica entre las redes de control y de cáncer son independientes del límite de IM

Con respecto a la diferencia *-cis* y *-trans* entre las redes de control y cáncer, en la Figura 4.8 podemos observar que las interacciones *-trans* en el control son siempre más altas que en cualquier red tumoral a pesar del valor de corte de IM.

También se puede apreciar que las interacciones *-trans* tienden a converger de acuerdo al aumento de tamaño. Se esperaba este resultado ya que cuantas más aristas aparecen en la red, más aristas *-cis* se han "cargado" en los cortes anteriores. Estos resultados también coinciden con un hallazgo reciente en redes de cáncer de mama, donde capas consecutivas no superpuestas de 100K aristas (clasificadas de arriba a abajo por valor de IM) contienen más interacciones *-cis* en las capas superiores, y disminuyen a medida que se acercan a la capa de ruido (1M de interacciones). ([32]).

4.1.7.3. Las redes de cáncer presentan un cambio en el orden de la tasa *-cis* en una pequeña variedad de interacciones

En la Figura 4.8 también se puede observar que desde el rango inicial (100) hasta aproximadamente 3,000 aristas, el rango de interacciones *-trans* es etapa I → III → IV → II. Sin embargo, en el rango de 3Ka 10K aristas este rango en redes de CRcc cambia de I → III → IV → II a II → IV → III → I. Ese orden adquirido se conserva hasta la convergencia ya comentada en 1 millón de aristas.

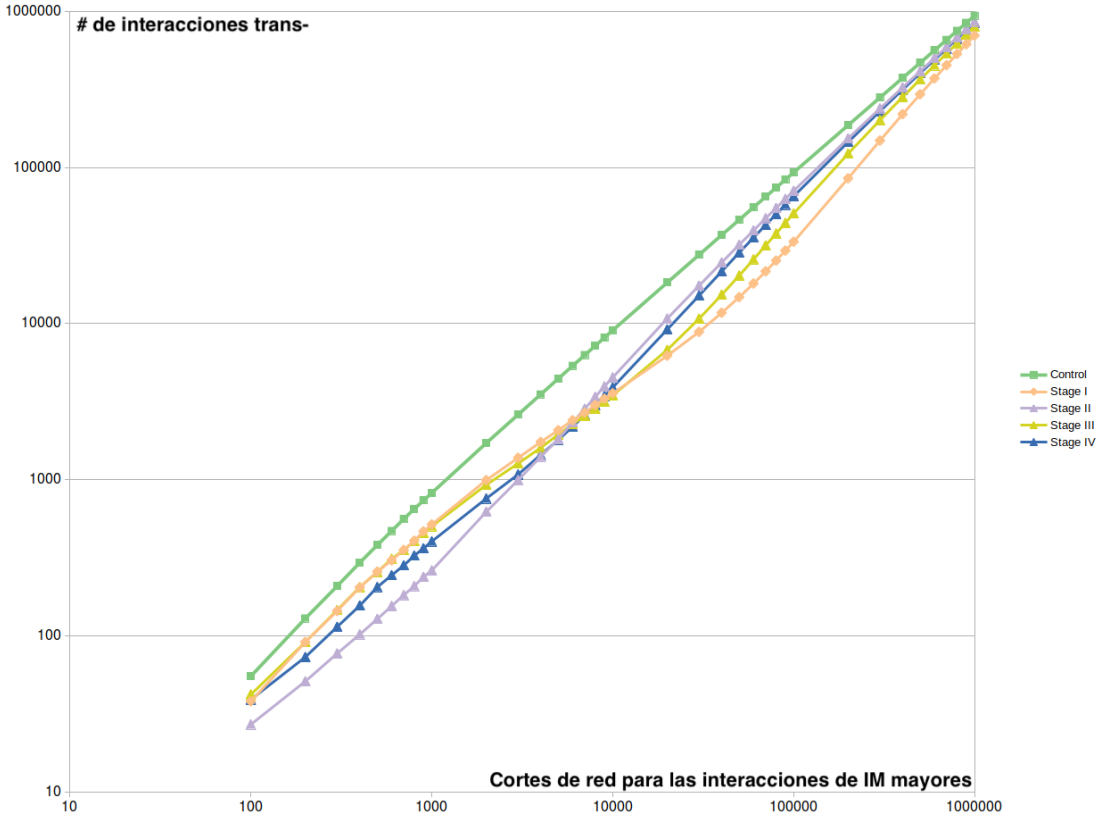

Figura 4.8: *Tasa de interacciones -trans en diferentes puntos de corte para cada etapa de CRcc y control.* En este gráfico, el eje X representa el valor de corte (interacciones mayores) en cada red para las cinco etapas, a saber, control y las cuatro etapas de CRcc. El eje Y muestra el número de interacciones entre cromosomas por cada corte de red. Tenga en cuenta que los enlaces *-trans* de la red de control son más grandes que cualquier etapa de progresión CRcc en cualquier valor de red de corte.

Como se mencionó anteriormente, el rango de proporción *-cis/-trans* no sigue la progresión de CRcc en ningún valor de corte. Por lo tanto, podemos concluir que las diferencias en las interacciones de la red intra/inter cromosoma no son un parámetro informativo para evaluar la progresión en CRcc. Se necesita más investigación sobre el cambio antes mencionado para tener una idea más completa del fenómeno.

4.1.8. 189 aristas relevantes se comparten en los cinco fenotipos

Se comparten 189 interacciones de coexpresión entre las cinco redes. Esas interacciones se representan en la Figura 4.9. La red resultante se compone de 230 genes y 189 interacciones. Los genes se colorean de acuerdo con su expresión génica diferencial.

Curiosamente, los componentes de esta subred común se agrupan en su mayoría de acuerdo con la tendencia de expresión diferencial: hay grupos compuestos solo por genes sobreexpresados, así como solo por genes subexpresados. Cabe mencionar que la correlación de Spearman entre el rango de genes diferencialmente expresados es superior a 0.9 para cualquier etapa (Tabla 4.2).

Además, los pequeños componentes conectados se enriquecen para procesos biológicos particulares y específicos. Por ejemplo, el primer componente, que contiene genes como *KIF20A*, *KIF18B* o *UBE2C*, está enriquecido para características de la matriz extracelular (MEC). Este es un componente altamente sobreexpresado, lo que indica que para cualquier etapa, los elementos de la MEC están involucrados en procesos exacerbados. Por el contrario, el tercer componente, con genes como *EGR2*, *EGR3*, *ATF* o *FOSB*, está completamente subexpresado y está enriquecido para los procesos relacionados con la respuesta inmune, lo que podría significar que la respuesta inmune se afecta en cualquier etapa de CRcc. Estos resultados nos sugirieron pistas importantes a seguir en este trabajo, destacando la respuesta inmune y la transformación del tejido por alteraciones en la MEC.

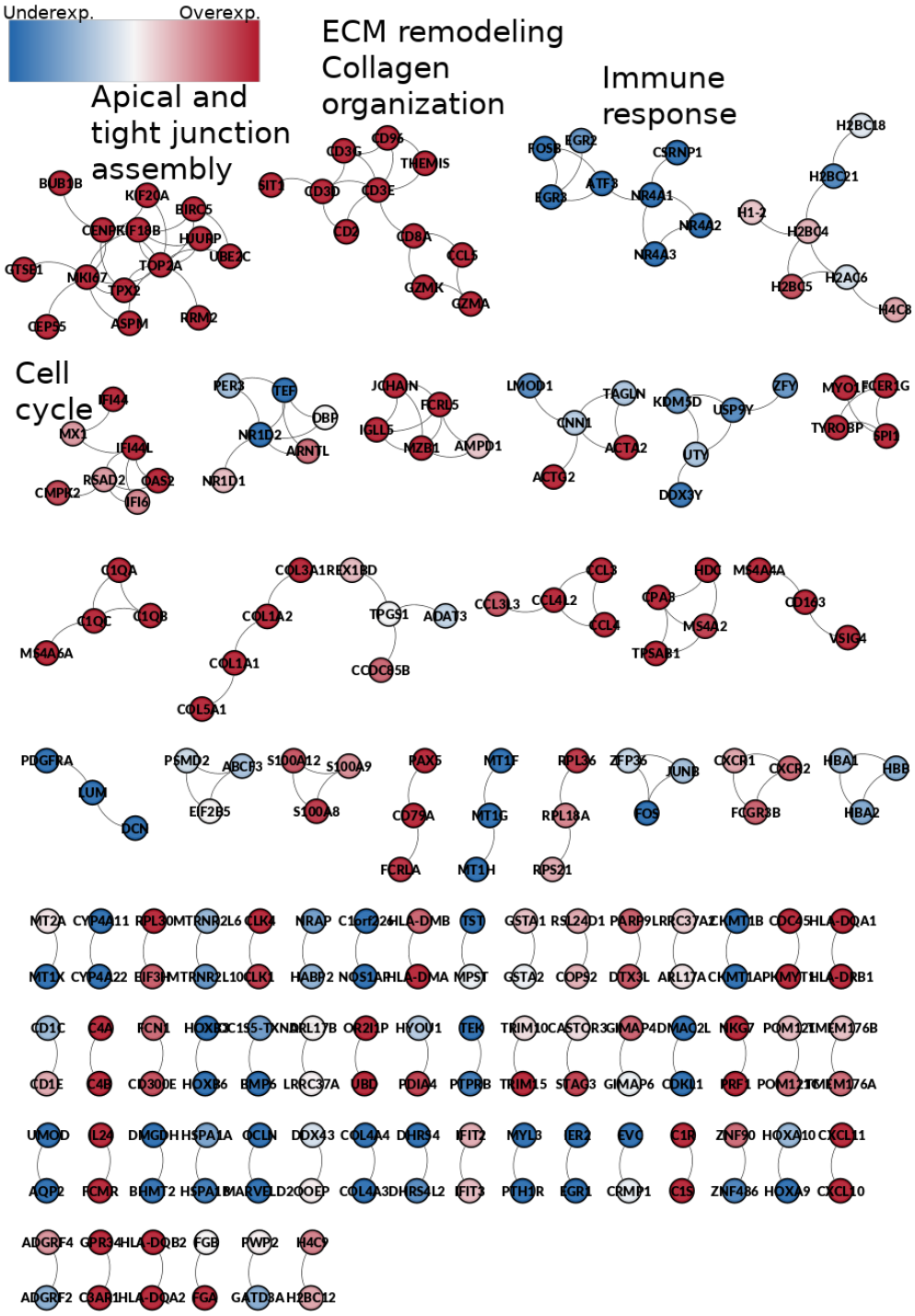

Figura 4.9: *Enriquecimiento biológico en las redes de interacciones compartidas entre los cinco fenotipos.*

Figura 4.9: La red resultante está compuesta por 189 aristas y 230 genes. Estos están coloreados de acuerdo con la expresión diferencial en comparación con el grupo de control. Observe que los componentes más pequeños de la red tienen un patrón de expresión similar. Algunos componentes se enriquecen en categorías GO específicas, lo que significa que esos procesos aumentan o disminuyen durante todo el proceso de progresión de CRcc.

4.1.9. Las funciones biológicas enriquecidas son independientes del valor de corte

La Figura 4.10 muestra las categorías enriquecidas obtenidas al cruzar las cuatro etapas de progresión (y excluyendo las interacciones de control). Para detalles del enriquecimiento ver la sección 3.3.7. De manera análoga a la Figura 4.9, en este caso (en 533 aristas) tenemos genes coloreados por sus valores de expresión diferencial, mientras que las categorías enriquecidas están pintadas con diferentes colores según el componente al que pertenecen esos procesos. Vale la pena mencionar que esta cifra solo incluye procesos con un valor-$p < 10^{-10}$. La lista completa de procesos enriquecidos para todos los fenotipos, en los cinco valores de red de corte, se incluye en el material suplementario de [158]. Las visualizaciones de red de procesos enriquecidos en la intersección de los fenotipos tumorales con 100K y 1M de interacciones también se incluyen en el mismo trabajo.

Otra característica compartida entre las Figuras 4.10 y 4.9 es que los grupos de genes con la misma tendencia de expresión diferencial tienen categorías enriquecidas. Entre las categorías más enriquecidas podemos encontrar procesos relacionados con el Ciclo Celular (amarillo), con *IFN-*γ (azul oscuro) y con procesos de las células T (verde).

En este trabajo una de las preguntas más relevantes que hicimos fue: ¿Cual es la relación entre la estructura de la redes y las etapas de progresión en el CRcc?. Buscando la respuesta, calculamos las *comunidades* de red mediante el algoritmo *infomap* ([123]).Posteriormente, realizamos el análisis de enriquecimiento sobre conjuntos separados de genes según la comunidad a la que pertenecen los mismos.

Dado el hecho de que las redes grandes a menudo contienen más comunidades que las redes pequeñas, realizamos el análisis de enriquecimiento para diferentes valores de corte de las intersecciones de la red. Independientemente del corte de la red, las intersecciones de las redes de sólo CRcc siempre presentan este conjunto de categorías enriquecidas, asociadas con el ciclo celular, el sistema inmunológico, la estructura tridimensional del ADN y la cromatina, o la regulación de la transcripción. Los detalles de estos cálculos se encuentran en [158].

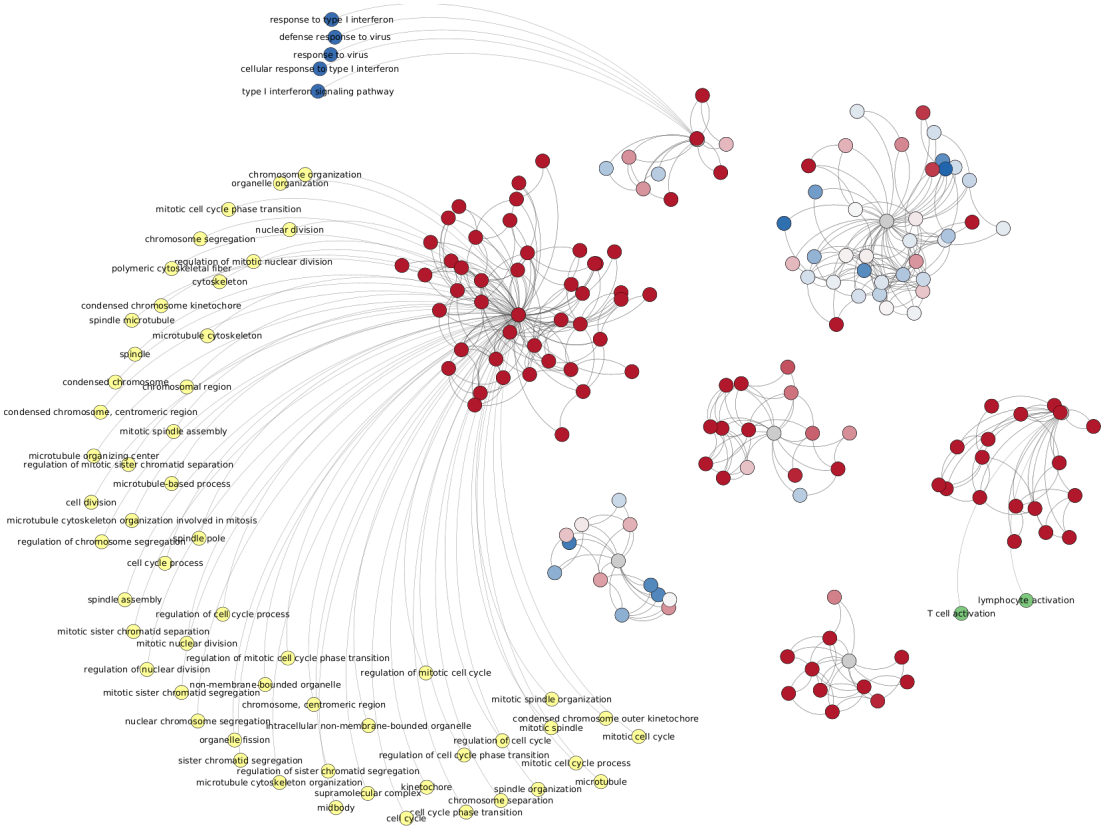

Figura 4.10: *Red construida con las 10K interacciones compartidas entre las redes de CRcc.* Destacar que el umbral fué establecido por las consideraciones de las secciones 3.3.4, 4.1.9 La red resultante se compone de 533 aristas y 148 genes. Éstos se colorean según su expresión diferencial en comparación con el grupo de control. Como en el caso de la Figura 4.9, los grupos expresados diferencialmente se enriquecen para categorías específicas.

4.2. Modelo de regulación genética por miRNAs en las cuatro etapas de CRcc

Con el fin de determinar la distribución entre la expresión de genes que codifican para proteínas y miRNAs, realizamos una comparación multigrupo (Ver sección 3.4) entre el control y cada etapa de progresión. Se observa un mayor número de genes y miRNAs sobreexpresados que subexpresados. La Tabla 4.3 muestra la comparación entre miRNAs y genes diferencialmente expresados entre no tumorales (NT-Control) y cada etapa de

progresión de CRcc (Etapa I, Etapa II, Etapa III y Etapa IV). Curiosamente, el número de GDEs aumenta con las etapas de progresión; esto puede sugerir que todo el programa regulador de genes se interrumpe en mayor medida cuando el tumor evoluciona hacia etapas posteriores.

Tabla 4.3: *Resumen cuantitativo de genes y miRNAs en cada contraste.* Cada uno de estos elementos también fue cuantificado en subexpresados (subexp) y sobreexpresados (sobrexp). Las muestras control se denotan como TN (Tejido Normal). Se destaca la poca cantidad de miRNAs válidos con respecto a los genes en su totalidad del experimento.

	TN-EtapaI	**TN-EtapaII**	**TN-EtapaIII**	**TN-EtapaIV**
Genes subexp.	1,946	2,012	2,106	2,187
Genes sobrexp.	2,187	2,238	2,587	2,630
miRNAS subexp.	87	87	88	88
miRNAS sobrexp.	88	87	96	91
Total de genes	4,133	4,250	4,693	4,817
Dif. consecutiva	-	117	443	124
Total de miRNAs	175	174	184	179

La mayor diferencia consecutiva entre GDEs está entre NT y Etapa I (Tabla 4.3). Esta observación es clara si se compara TN-Etapa I con el resto de contrastes. Este resultado puede deberse al reclutamiento y acumulación de varios tipos de células diferentes asociadas al cáncer, además de las alteraciones genómicas intrínsecas de las células cancerosas con respecto a las normales.

A pesar de la gran cantidad de GDEs y mDEs, los genes únicos expresados diferencialmente o miRNAs son bastante escasos. La Tabla 4.4 muestra el número de GDEs y mDEs **únicos** por cada contraste. Como se observa, la cantidad de GDE/mDE únicos por contraste es casi 40 veces menor que la cantidad total de GDE/mDE.

Estos resultados cobran especial interés porque parece que la mayoría de los GDEs/mDEs se conservan a lo largo de toda la evolución de la enfermedad. Sin embargo, como hemos observado previamente en las redes de coexpresión gen-gen (sección 4.1 y [158]), la expresión diferencial no es suficiente para explicar la evolución de las primeras etapas a las más avanzadas.

Tabla 4.4: *Genes y miRNAs* ***únicos***. Resumen que relaciona las características de expresión y la cuantificación de los genes y los miRNAs. También fueron cuantificados por subexpresión (subexp) o sobreexpresión (sobrexp). Las muestras control se denotan como TN (Tejido Normal). En este resultado se puede apreciar la poca cantidad de genes y miRNAs que aportan un cambio en los fenotipos. Desde luego, siempre establecida la relación miRNA-gen.

	TN-EtapaI	**TN-EtapaII**	**TN-EtapaIII**	**TN-EtapaIV**
Genes sobrexp.	52	54	61	189
Genes subexp.	56	58	30	141
miRNAS sobrexp.	2	2	4	4
miRNAS subexp.	1	1	1	3

4.2.1. Los GDEs y mDEs son más abundantes entre el control y la etapa I que en cualquier otro contraste

Si bien la expresión diferencial entre etapas de control y progresión proporciona información sobre aquellos genes y miRNAs que pueden ejercer influencia en la adquisición de rasgos oncogénicos, una comparación entre etapas contiguas puede ser, en cierto sentido, más reveladora ya que representa la evolución del programa de expresión genética a lo largo de la progresión del tumor.

Para investigar más sobre esto, realizamos un análisis de expresión diferencial entre etapas secuencialmente contiguas. En las Figuras 4.11 y 4.12 podemos observar gráficos de volcanes que muestran los GDEs y mDEs entre las etapas consecutivas de la evolución de CRcc: NT-stI (NT-Etapa I), stI-stII (Etapa I-Etapa II), stII-stIII (Etapa II-Etapa III) y stIII-stIV (Etapa III-Etapa IV). El material suplementario de [159] muestra los genes/miRNAs compartidos y únicos para cada contraste.

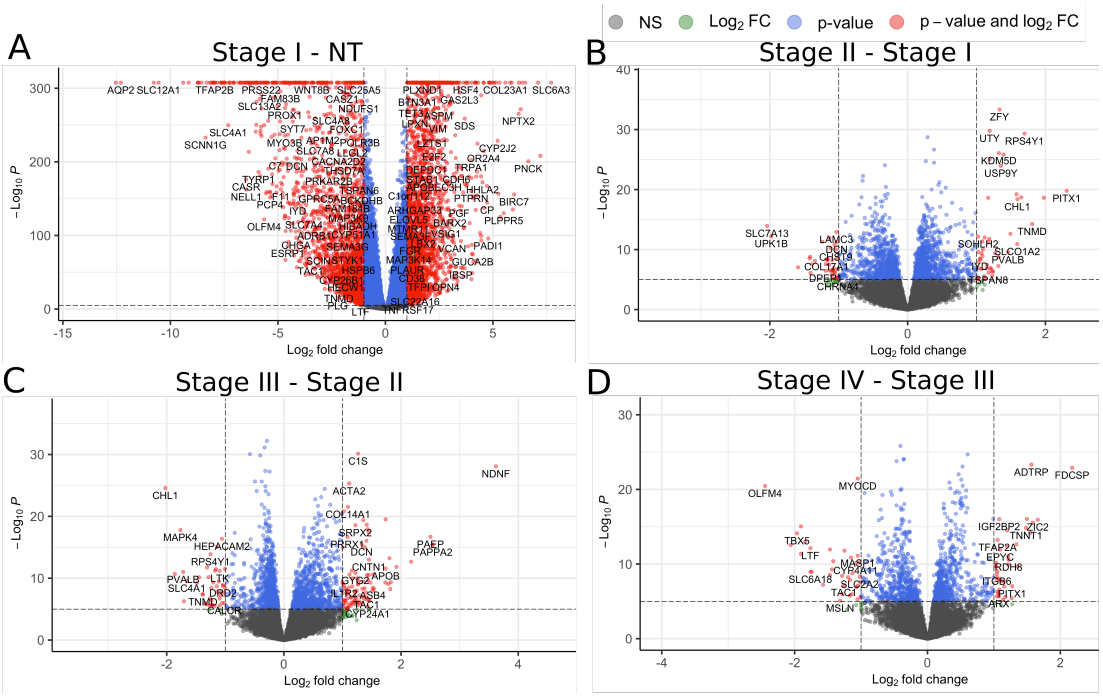

Figura 4.11: *Genes expresados diferencialmente para cada etapa contigua de CRcc.* (A) Contraste entre el control y la etapa I; (B) etapa I y etapa II; (C) etapa II y etapa III; (D) etapa III y etapa IV. Los círculos rojos representan genes con un $|LogFC| > 1$ y un valor de p $<$ 1e-5; los círculos representados en verde tienen en cuenta aquellos genes con un $|LogFC| > 1$ pero valor p $<$ 1e-5; los genes con un $|LogFC| < 1$ pero un valor-p $< 10^{-5}$ se representan en azul. Finalmente, aquellos genes con valores inferiores a esos umbrales se representan en gris. Se hace evidente que el contraste con más DEGs es el que existe entre Control y etapa 1.

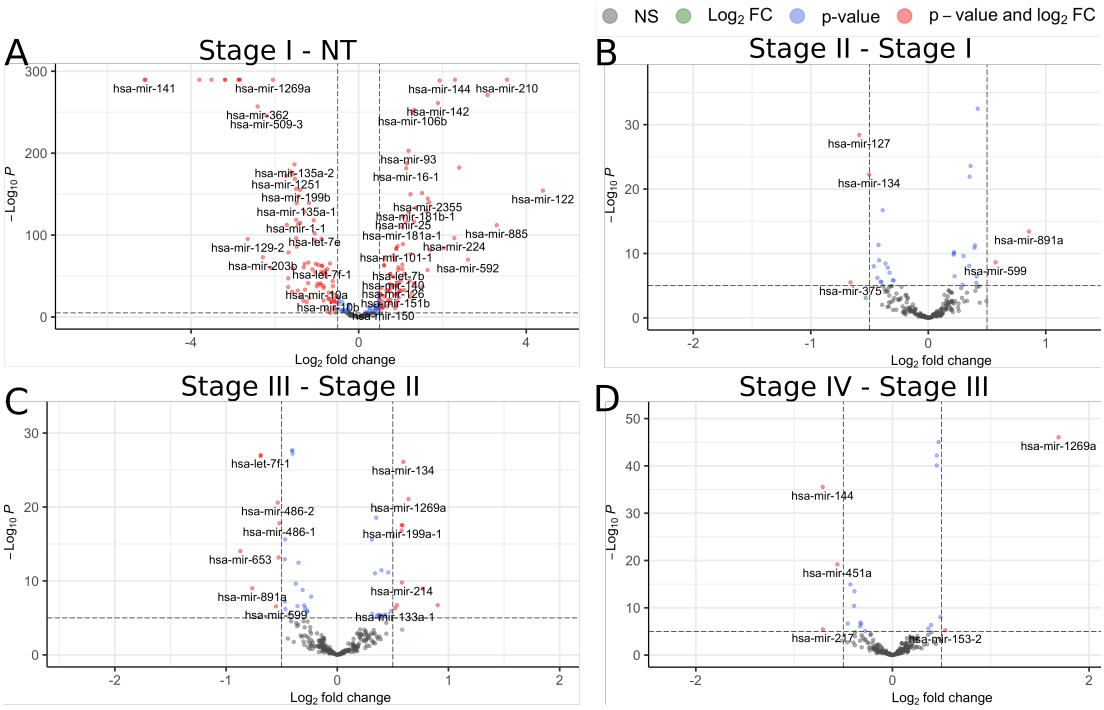

Figura 4.12: *miRNAs expresados diferencialmente para cada etapa contigua de CRcc.* (A) Contraste entre el control y la etapa I; (B) etapa I y etapa II; (C) etapa II y etapa III; (D) etapa III y etapa IV. El código de color es el mismo que el de la Figura 4.11.

La Figura 4.11 muestra los diagramas de volcanes para los GDEs. Como se puede observar, el contraste con el mayor número de genes se da entre Control y Etapa I, con un total de 2,187 genes sobreexpresados y 1.946 subexpresados. Los siguientes contrastes (Etapa I-Etapa II,...) tenían un número de GDEs más de 100 veces menor que el primero. Análogamente, en la Figura 4.12, podemos observar un comportamiento similar para los mDEs. De acuerdo con estos resultados, en CRcc, los principales cambios en los programas reguladores de genes y miRNAs ocurren en la fase inicial de la evolución tumoral.

En estos últimos contrastes, tanto GDEs como mDEs muestran diferencias específicas en su expresión; por ejemplo, *miR-155* (considerado como miR regulador de *VHL* [74], [101]) está sobreexpresado en la comparación Control- Etapa I, y como se muestra, en los siguientes contrastes no se expresa diferencialmente.

Para notar, en los últimos contrastes, tanto para los casos de miRNAs como de genes, la lista de GEDs y mDEs es diferente en cada contraste. En el Material suplementario de [159], se adjuntan todos los contrastes entre Control y todas las etapas de progresión de CRcc, así como entre etapas secuencialmente contiguas.

Los miRNAs y los genes sobre y subexpresados resultantes se usaron luego para construir las redes gen-miRNA para cada fenotipo (Control y las cuatro etapas). Sólo

conservamos aquellas interacciones miRNA-gen entre GDEs y mDEs con una tendencia de expresión diferencial opuesta (que potencialmente corresponde a los mecanismos canónicos de la regulación miRNA-gen).

4.2.2. Las redes miRNA-gen son en su mayoría específicas de la etapa

La Figura 4.13 muestra un gráfico *upset* de las interacciones compartidas entre un gen y un miRNA para cada etapa. El contraste marcado con amarillo contiene la mayor cantidad de genes compartidos entre GDEs y mDEs. En el caso de las redes miRNA-gen, solo hay un pequeño subconjunto de interacciones compartidas entre redes. Más del 90 % de las interacciones miRNA-gen son únicas para cada red.

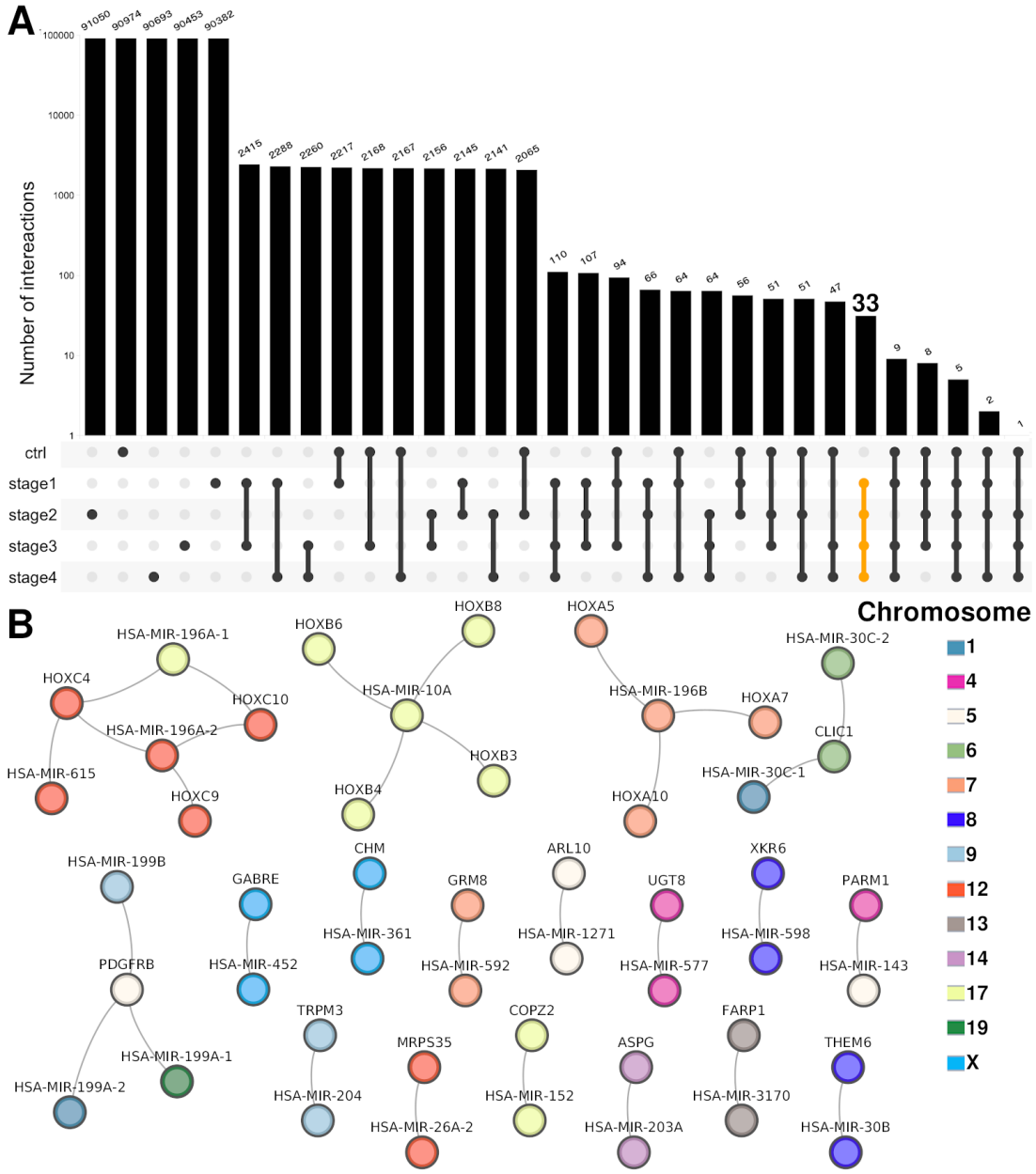

Figura 4.13: *Intersección de las redes de coexpresión gen-miRNA*. (A) Cada barra en el gráfico UpSet muestra el número de interacciones en el conjunto seleccionado, representado por puntos vinculados debajo de las barras (escala logarítmica). Encima de cada barra, se muestra el número de interacciones. Las primeras cinco barras representan interacciones únicas.

Figura 4.13: A partir de la sexta barra, cada una de ellas muestra el número de interacciones compartidas entre dos o más redes. En el lado derecho, el conjunto de interacciones compartidas entre las cuatro etapas de progresión de CRcc (pero no Control) se resalta en amarillo. (B) Se representan las 33 interacciones compartidas entre las cuatro etapas de progresión pero no compartidas con la red no tumoral. En la figura, el color de los nodos representa el cromosoma donde se encuentran los miRNAs y los genes.

Este resultado fue aparentemente contrario a la intuición, ya que la cantidad de genes compartidos y miRNAs entre los contrastes era muy alta. Sin embargo, los programas reguladores miRNA-gen, representados por redes de coexpresión de alta confianza, son altamente específicos para cada etapa de progresión.

Un resultado concomitante derivado de la singularidad de las interacciones miRNA-gen para cada etapa de progresión es, que existe, un pequeño conjunto de interacciones compartidas entre las etapas del cáncer, pero no compartidas con la red normal (ó no tumoral). Sólo 33 interacciones son comunes para los cuatro estadios y no se presentan en el fenotipo control.

Al observar esas interacciones, se puede apreciar que prácticamente todas ellas corresponden a miRNAs y genes que pertenecen al mismo cromosoma. Además, pertenecen a la misma citobanda (Material suplementario en [159]).

Curiosamente, los miRNAs más conectados corresponden a la familia *miR-196* listados como: *miR-10A* y *miR-196A-1* (Chr17q21.32), *miR-196A-2* (Chr12q13.13) y *miR-196B* (Chr7p15.2). Como se puede observar en la Figura 4.13.B, esos miRNAs (parte superior de la red) están asociados a genes HOX. Estos genes pertenecen a la misma localización que dichos miRNAs: *HOXC9*, *HOXC10* y *HOXC11* que están situados en Chr12q13.13. Lo mismo se presenta en el caso de *HOXA5*, *HOXA7* y *HOXA10* (Chr7p15.2) o *HOXB3*, *HOXB4*, *HOXB6* y *HOXB8* (17q21.32).

El papel de los genes HOX en el surgimiento y desarrollo de varios tipos de Cáncer ha sido ampliamente informado [14, 79, 130]. Además, también se ha descrito el papel de la familia miR-196 en diferentes carcinomas [96, 113, 153]. El hecho de que los complejos de genes miR-196-HOX se compartan entre todas las etapas de progresión pero estén ausentes en la red no tumoral, puede indicar el papel biológico-funcional de estas relaciones en la progresión de CRcc.

Vale la pena notar, que los genes compartidos por los conjuntos no tumorales (control) y los estadios fueron: *HOXA9*, *MEST*, *TENM4*, *ARPP21*, *DIO3*. Como se ha mencionado, los genes *HOXA* tienen un papel importante en el cáncer. La pérdida de impronta del gen *MEST* se ha relacionado con ciertos tipos de cáncer y puede deberse al cambio de promotor. Sin embargo, todos esos genes juegan un papel crítico en el desarrollo de los mamíferos como una característica común.

Finalmente, la ubicación vecina de miRNAs y genes observados en la 4.13.B se ha

descrito previamente en redes de coexpresión gen-gen para cáncer de mama ([35, 44, 45]), Cáncer de pulmón [6] y también en la progresión por etapas de CRcc [158]. En este caso, donde las redes inferidas se obtienen correlacionando la expresión del miRNA con la expresión génica, el efecto de pérdida de la coexpresión a larga distancia no se aprecia en el conjunto de las redes. Sin embargo, en las 33 (de 100K para cada etapa) interacciones compartidas por el cáncer, observamos no solo interacciones miRNA-gen con moléculas del mismo cromosoma, sino también la misma citobanda y, además, ubicaciones contiguas en términos de posiciones de inicio (Material suplementario en [159]).

Después de observar la ubicación de miRNAs y genes en la red compartida, podemos argumentar que la aparición de interacciones intra-citobanda en fenotipos exclusivos de cáncer podría estar relacionada con un evento transcripcional anómalo que permite tener patrones de expresión similares entre miRNAs y transcritos de genes. Sin embargo, se necesita corroboración experimental para dilucidar completamente el papel de esas interacciones. Además, el complejo HOX-miR-196 también debe investigarse para proporcionar una posible explicación de esos genes relacionados con el desarrollo en la progresión de CRcc.

4.2.3. Las redes miRNA-gen son diferentes entre etapas, tanto en tamaño como en composición

Como se indicó anteriormente, inferimos cinco redes (una para cada fenotipo), una red para Control y una red para cada etapa de progresión de CRcc. Para construir todas las redes, calculamos la medida de información mutua entre miRNA y genes usando las matrices de expresión para cada ómica en todas las etapas (ver Tabla 4.4).

Conservamos las 100K interacciones mas fuertes miRNA-gen para cada una de las cinco redes (Material suplementario en [159]). De estas 100K interacciones, filtramos solo aquellas coexpresiones entre GDEs y mDEs con una tendencia de expresión diferencial opuesta (miR sobreexpresado, gen subexpresado y viceversa). Las redes resultantes se representan en la Figura 4.14. La diferencia de tamaño entre las redes de etapa I y el resto de redes es evidente.

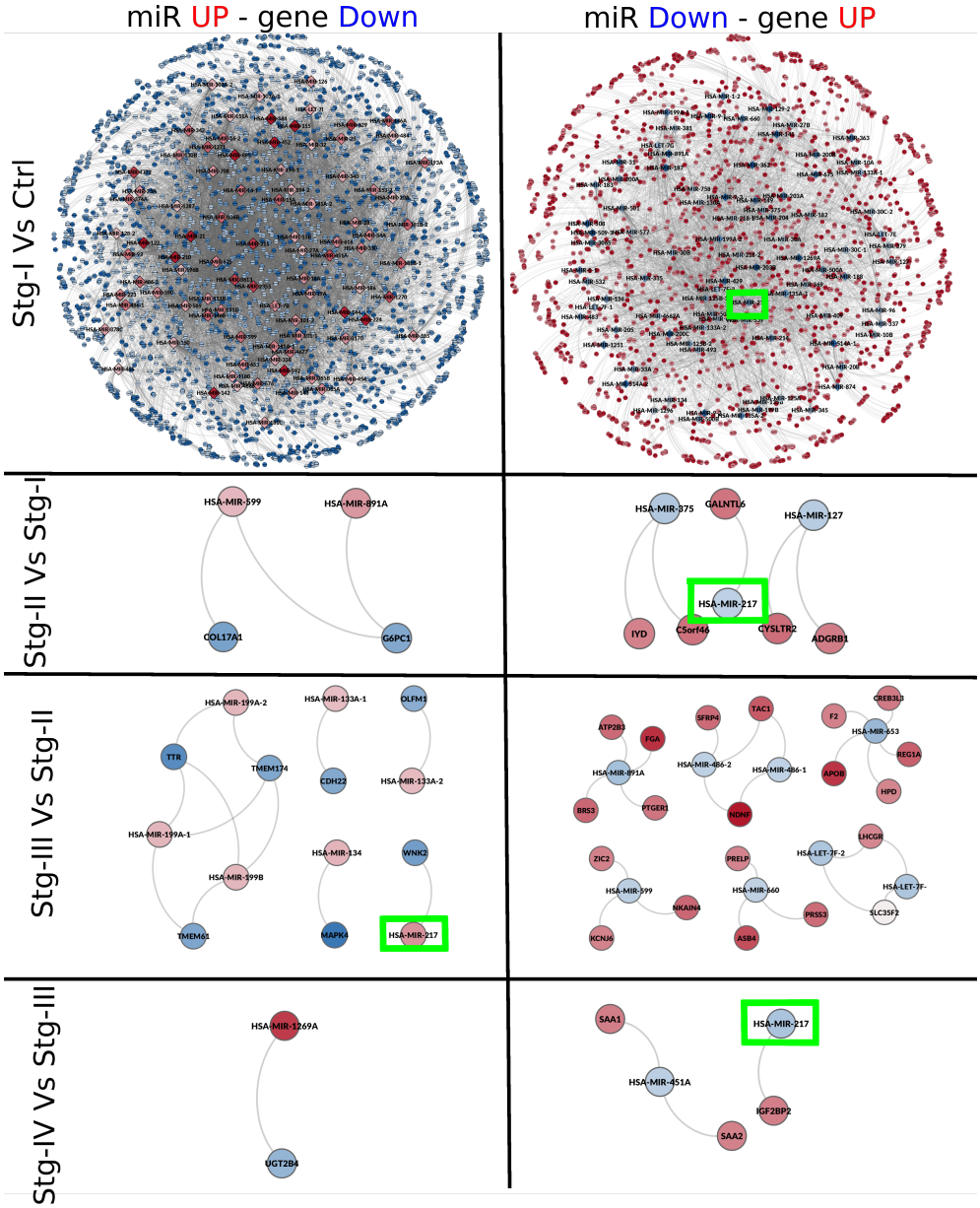

Figura 4.14: *Redes gen-miRNA para cada etapa de progresión.*

Figura 4.14: En esta figura, podemos observar redes inferidas por información mutua entre la expresión de miRNAs y genes en cada etapa de progresión de CRcc. Las redes se colocaron de arriba a abajo según la etapa de progresión. El contraste utilizado para representar cada red se coloca a la izquierda. Los nodos rojos representan miRNAs ó genes sobreexpresados; mientras que las moléculas subexpresadas se representan en azul. En el lado izquierdo, se pueden encontrar redes construidas con miRNAs sobreexpresados y genes subexpresados. La parte derecha de las figuras contiene redes con miRNAs subexpresados y genes sobreexpresados. Los cuadrados verdes marcan la ubicación de *miR-217*, el único microRNA presente en las cuatro redes.

4.2.4. *miR-217* se expresa de manera diferencial en todos los contrastes secuencialmente contiguos, pero muestra diferentes genes diana para cada etapa

En la Figura 4.14 se puede apreciar como, en cada red, *miR-217* aparece DE y además tiene un gen diana diferente en todos los casos. En el contraste entre el estadio I y Control, *miR-217* está subexpresado ($LogFC$ = -1.32). En esta etapa, este microRNA potencialmente regula hasta 60 genes objetivo (Material suplementario en [159]). Entre los genes diana de *miR-217* podemos encontrar los genes *BIRC7* ($LogFC$ = 5.9), *LAMA4* ($LogFC$ = 4.0) o *E2F2* ($LogFC$ = 2.2) (tabla 4.5).

Para el contraste entre el estadio I y NT, *BIRC7* fue el gen más sobreexpresado. Se ha informado que *BIRC7* es crucial en el desarrollo del cáncer al inhibir la apoptosis [117]. Esto se ha observado en varios tipos de cáncer, como tiroides [87], leucemia [67] ó neuroblastoma [27]. En particular, para el carcinoma de células renales, la sobreexpresión de *BIRC7* se ha asociado con malignidad relacionada a *PTEN* y tener peor pronóstico [18] con comportamiento metastásico [143].

Tabla 4.5: *Estadísticas de expresión para miR-217 y sus genes diana.*. En la tabla se denotan las muestras control como TN (Tejido normal), etapa I (I), etapa II (II), etapa III (III) y etapa IV (IV). Este resultado es muy importante porque supone el proceso que puede seguir *miR-217* durante la progresión consecutiva del cáncer.

	TN-I	**I-II**	**II-III**	**III-IV**
mi-217 *logFC*	-1.322	-0.5326	0.9033	-0.7965
Número de dianas	60	*GALNTL6*	*WNK2*	*IG2BP2*
logFC de las dianas	1.899 (promedio)	1.1678	-1.2373	1.0798

LAMA4 también se sobreexpresa fuertemente en la etapa I en comparación con Control. Su sobreexpresión se ha relacionado con metástasis en cáncer de páncreas [164]. Además, se ha observado que *miR-200b* regula a la baja *LAMA4* y disminuye la metástasis del carcinoma de células renales [84].

En cuanto a la red de la etapa II, *GALNTL6* (polipéptido Nacetil-galactosaminil-transferasa 6) es la única diana presente de *miR-217*. Este gen está relacionado con el metabolismo de las proteínas y la glicosilación O-ligada [12]. Las anotaciones de Gene Ontology (GO) relacionadas con este gen incluyen la unión de carbohidratos y la actividad del polipéptido N-acetilgalactosaminiltransferasa. La familia GALNT normalmente inician la O-glicosilación en el aparato de Golgi, pero en modelos de cultivo celular estas enzimas pueden trasladarse al retículo endoplasmático (RE) a través de un proceso que implica la señalización aberrante de Src, lo que conduce a una mayor densidad de O-glicosilaciones en MUC1 [115]. Se ha informado que *GALNTL6* está amplificado en carcinomas papilares de tiroides [106].

Para la red de etapa III, el único objetivo de *miR-217* es *WNK2* (proteína quinasa 2 deficiente en lisina de WNK). Las vías relacionadas con *WNK2* son el transporte de glucosa y otros azúcares, sales biliares y ácidos orgánicos, iones metálicos y compuestos de amina y transporte de canales iónicos. Las anotaciones GO relacionadas con este gen incluyen la actividad transferasa, la transferencia de grupos que contienen fósforo y la actividad de proteínas tirosina-cinasa.

Debemos notar que en la red de Etapa III, *WNK2* está subexpresado y *miR-217* está regulado a la alza. *WNK2* se considera un gen supresor de tumores porque inhibe la proliferación celular [98], regulando negativamente la señalización del receptor del factor de crecimiento epidérmico a través de la inhibición de MEK1 [70].

Tomando en cuenta este contexto, el hecho de que *miR-217* resultara sobreexpresado y su única diana en la red de etapa III fue *WNK2*, respalda la siguiente hipótesis: *WNK2* puede ser un gen supresor de tumores específico de etapa III regulado a la baja por *miR-217*.

Finalmente, encontramos a *IGF2BP2* como la única diana de *miR-217* en la red de

etapa IV. *IGF2BP2* es un regulador postranscripcional de IGF2 (factor de crecimiento de insulina 2). Este gen afecta a otros genes como: *MYC* y *PTEN*, dos participantes cruciales en vías asociadas con la tumorigénesis [13]. *IGF2BP2* fue considerado como un regulador del metabolismo. Modula el metabolismo celular en diabetes, obesidad o esteatosis hepática mediante regulación génica postranscripcional [145]. Recientemente se ha demostrado que la sobreexpresión de este gen es un factor pronóstico en varios tipos de cáncer, como la leucemia [59], de mama [83], de pulmón [134], colorrectal [155] o hepatocarcinoma [111].

En la Figura 4.15 se muestra un modelo resumido de los cambios regulatorios que pudieran estar surgiendo en *miR-217* y sus genes diana dentro del CRcc.

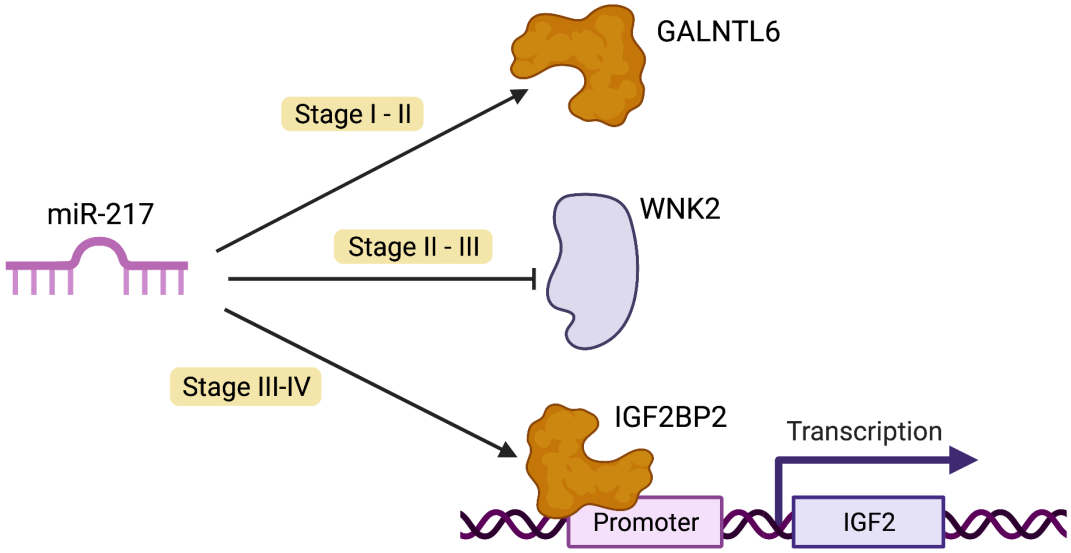

Figura 4.15: *Probable papel oncogénico de miR-217.* En la transición uno del cáncer (etapa I-II), *miR-217* permite la sobreexpresión de *GALNTL6*. Esta proteína normalmente inicia modificaciones postraduccionales en el aparato de Golgi. Además, en modelos de cultivo celular, estas enzimas afectan el retículo endoplasmático a través de la señalización aberrante de Src. En las etapas II-III (transición dos), *miR-217* reprime la expresión de *WNK2*, un supresor tumoral que inhibe la proliferación celular al modular negativamente la activación de la vía MEK1. En la última transición (etapas III-IV), *miR-217* permite la sobreexpresión de *IGF2BP2*. Este gen promueve la progresión tumoral en varios tipos de cáncer, como el glioblastoma multiforme y el cáncer de vesícula biliar. *IGF2BP2* también promueve la proliferación de células tumorales a través de la vía PI3K-Akt.

En la red de etapa IV, *miR-217* está subexpresado y su único objetivo es *IGF2BP2*, que está sobreexpresado ($LogFC = 1.0798$). La sobreexpresión de este gen puede deberse a la subexpresión de *miR-217* en esta etapa de CRcc. Y en definitiva esta sugerencia debería probarse experimentalmente.

Vale la pena notar que la expresión diferencial de todos los genes antes mencionados ocurre entre etapas secuencialmente contiguas, es decir, el contraste entre esos genes lo realiza la fase previa de CRcc. Estos resultados son notables ya que el conjunto de datos de "control" es una etapa anterior de CRcc; ese conjunto de datos de expresión génica de control ya está alterado por el cáncer. Por lo tanto, GDEs y mDEs están "más

diferenciados" que en la red de control, considerando así, que control es el contraste tradicionalmente seleccionado para comparar.

Como se muestra en las Figuras 4.11 y 4.12, el número de interacciones estadísticamente significativas en la red NT-Etapa I es mucho mayor que en cualquier otro contraste. Esto implica que las mayores alteraciones que ocurren entre estas etapas de progresión contiguas están dadas por el alto número de genes y miRNAs expresados diferencialmente, lo que da pie a la desregulación de varios procesos biológicos, que a su vez están asociados con cambios radicales de todo el fenotipo.

Por otro lado, el bajo número de interacciones en los contrastes posteriores puede implicar que la desregulación gen-miRNA observada en las etapas avanzadas es un proceso complementario. Además, este proceso es concomitante con varios otros fenómenos que impulsan la progresión del carcinoma renal de células claras.

Cabe destacar un resultado aparentemente contraintuitivo; como se muestra en la Figura 4.13A, se utilizó el mismo conjunto de 16,227 genes y 275 miRs para construir cada red. Sin embargo, la cantidad de interacciones compartidas es muy baja en comparación con las interacciones únicas por red (más de 90K de 100K para cualquier fenotipo dado). Este efecto de unicidad en las interacciones de la red probablemente obedece a las especificidades de regulación por microARN en cada contexto. A pesar de que las cinco redes contienen los mismos genes y miRNAs, la forma en que los miRNAs y los genes se coexpresan es exclusiva. La progresión de CRcc aparentemente modifica los procesos de regulación genética mediados por microARNs.

No obstante, la red compuesta por las interacciones compartidas entre las cuatro etapas de progresión de CRcc también es informativa. A partir de esa red, podemos observar que casi todas las interacciones ocurren entre genes y miRNAs del mismo cromosoma.

Nuestro grupo ha informado previamente sobre el sesgo de las interacciones intracromosómicas en las redes de coexpresión de genes para el cáncer de mama. [32, 33, 35, 45, 49], cáncer de pulmón [6], y también CRcc [158]. Estos resultados muestran una clara tendencia a favorecer estrechas correlaciones de genes en términos de distancia de pares de bases. Sin embargo, para las redes de coexpresión gen-miRNA en cáncer de mama [28, 30, 34], no observamos una tendencia con más interacciones intracromosómicas miRNA-gen que las intercromosómicas. Hasta donde sabemos, esta es la primera vez que se observa un sesgo en las interacciones intracromosómicas gen-miRNA en el contexto del cáncer de riñón.

El hallazgo de esas 31 interacciones intracromosómicas gen-miRNA puede estar relacionado con el mismo mecanismo detrás del sesgo que favorece las correlaciones locales sobre las de larga distancia.

Sin embargo, el mecanismo por el cual este fenómeno surge en el cáncer, pero no en las redes de control, sigue siendo confuso. Hemos investigado el papel de otros procesos biomoleculares, como los sitios de unión del factor de transcripción, los sitios de unión de CTCF [44] o las alteraciones del número de copias genómicas [61]. Vale la pena notar que ninguno de ellos ha demostrado estar significativamente relacionado con la pérdida de interacciones intercromosómicas.

Con respecto a las diferencias entre las redes durante la progresión, el bajo número de genes regulados por miRNAs es intrigante ya que los informes de genes *regulados* por miRNAs en el contexto del carcinoma renal han aumentado en los últimos años [para una revisión sistemática, ver [81]]. Esto último podría deberse a la forma en que se construyeron las redes. Estas redes se obtuvieron mediante tres filtros diferentes: (a) las 100K interacciones principales de coexpresión de miRNA-gen, (b) aquellos miRNAs y genes que resultaron expresados diferencialmente entre etapas contiguas, y (c) las relaciones de coexpresión entre miRNAs y genes con signo opuesto en sus valores de expresión diferencial.

Resulta relevante comentar que ninguno de los blancos de *miR-217* es un remodelador de la cromatina. Este hecho nos lleva a reevaluar el papel epigenético de los miRNAs, en particular en el modelo propuesto para *miR-217*. En las conclusiones ahondaremos en este punto con mayor detalle.

4.3. Regulación epigenética por metilación

4.3.1. Primera vista de la metilación en las etapas de CRcc

Para rastrear los cambios en la progresión del cáncer asociados con las etapas (y los contrastes con el tejido normal - NT), realizamos una agrupación jerárquica de los datos por fenotipo de cáncer. Observamos un grupo bien definido de muestras de NT (no tumoral), que está claramente separado de las muestras en etapa de cáncer (consulte la Figura 4.16A). Para tener en cuenta el sesgo potencial del tamaño de la muestra, seleccionamos subconjuntos de valores de metilación de diferentes tamaños: 10^2, 10^3, 10^4, 10^5 (los gráficos se pueden ver en el material suplementario de [160]). Confirmamos el mismo patrón, donde las únicas muestras que se agruparon consistentemente fueron las no tumorales. Curiosamente, al considerar solo las distribuciones de valores de metilación por sitio CpG para cada fenotipo, no es posible obtener diferencias estadísticamente significativas entre etapas (ver Figura 4.16C).

Teniendo en cuenta los resultados del agrupamiento, calculamos las distribuciones de metilación por contraste, es decir, una distribución de sitios CpG hipermetilados e hipometilados para cada comparación entre fenotipos. Utilizamos un método de *bootstrap* para mitigar los efectos del tamaño de la muestra (consulte la Figura 4.16B). Para tener en cuenta los desequilibrios de grupo, utilizamos 24 muestras para cada fenotipo en el proceso de *bootstrap*, reduciéndolo al grupo más pequeño. Las principales diferencias se observaron entre el control y los diferentes estadios tumorales. No se observaron diferencias estadísticamente significativas en las comparaciones de progresión (NT-etapa1, etapa1-etapa2, etapa2-etapa3, etapa3-etapa4). Con base en estos resultados, decidimos usar solo el contraste de control versus cada etapa para desarrollar un *modelo de progresión versus línea base*.

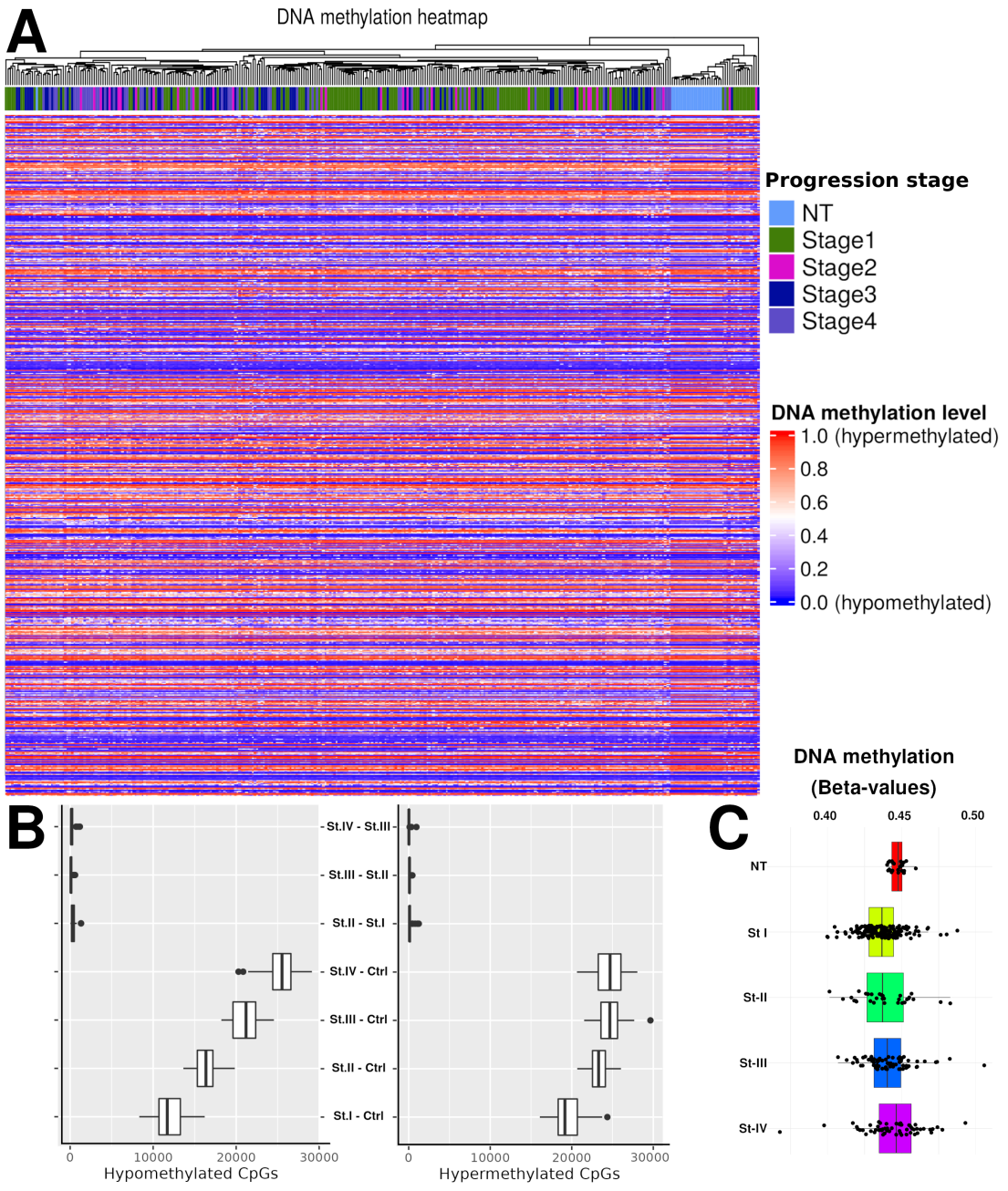

Figura 4.16: *Vista general del metiloma en CRcc.* A partir de la cuantificación de estos datos se logró plantear una estrategia para abordar el fenómeno.

Figura 4.16: A) Mapa de calor que muestra el nivel de metilación para cada sitio CpG. El agrupamiento jerárquico se realizó por etapa de progresión, incluido el tejido normal (NT). B) Cuantificación de CpGs por contraste. Como primera aproximación consideramos dos modelos de progresión: 1) secuencial (Etapa I Vs. NT, Etapa II Vs. Etapa I,...) y 2) comparado con control (Etapa I Vs. NT, Etapa II Vs. NT,...). La figura muestra que la mayor cantidad de metilación diferencial en sitios CpG(MD-CpG) en los contrastes está dada por el segundo modelo. Por lo tanto, adoptamos esta segunda estrategia. C) Distribuciones de valores β en CpG por estadio tumoral y para tejido normal (NT).

4.3.2. Genes metilados por contraste

En la Fig. 4.16B, podemos observar sitios CpG diferencialmente metilados (CpG) de forma significativa. Por construcción, estos sitios CpG están asociados con regiones promotoras de varios genes. Esta evidencia nos llevó a seguir un procedimiento sistemático para obtener genes con claras diferencias entre fenotipos (Etapa I vs NT, Etapa II vs NT, ...).

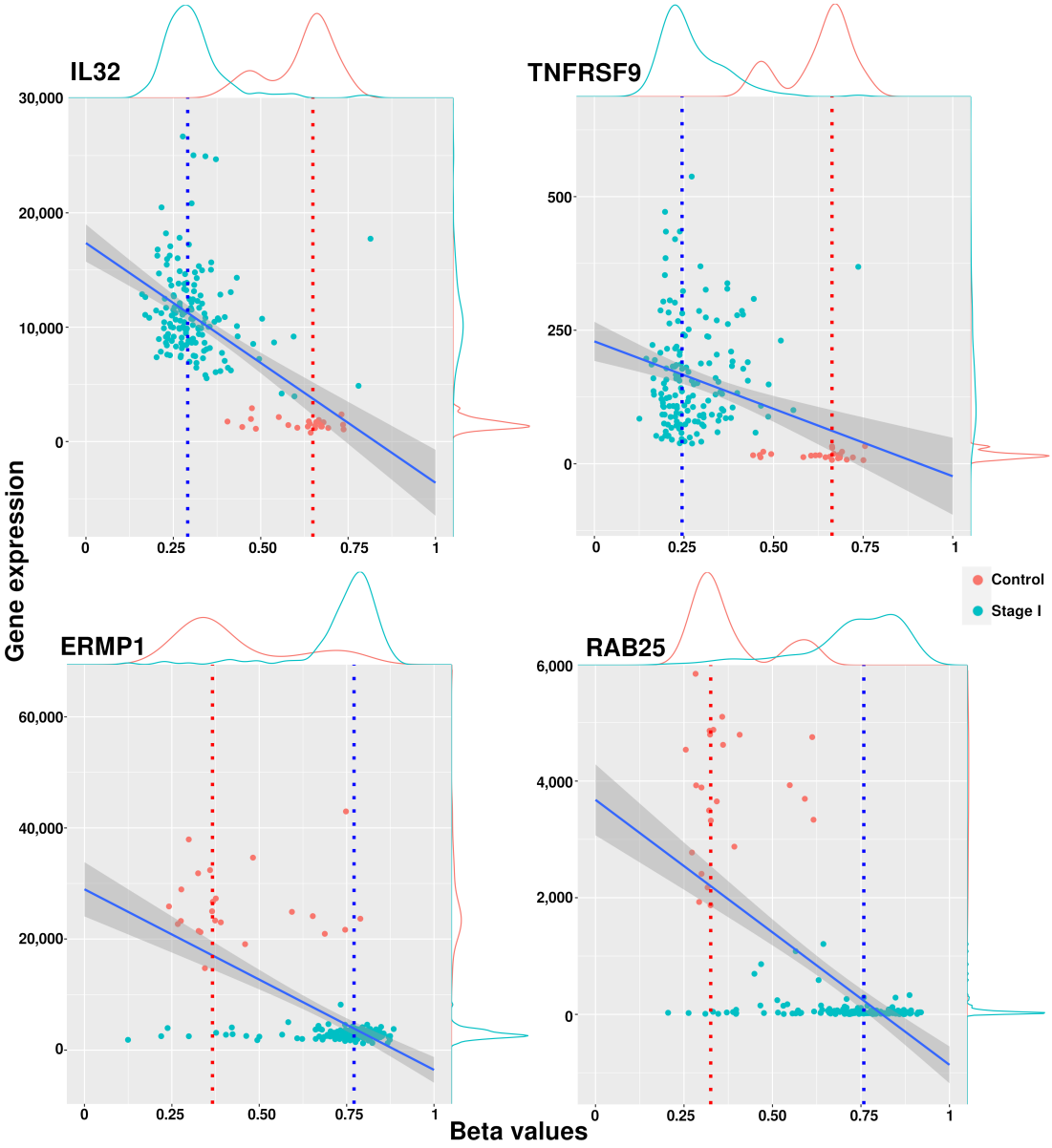

Figura 4.17: *Gráficos de dispersión de valores de metilación y expresión.* Aquí listamos dos ejemplos de genes hipometilados y dos ejemplos de genes hipermetilados. *IL32* y *TNFRSF9* tienen una condición hipometilada en cáncer, mientras que *ERMP1* y *RAB25* resultaron hipermetilados en etapas de cáncer. Estos gráficos son la base para que el algoritmo construido en este trabajo discrimine y agrupe los genes acorde a su metilación.

Observamos que existe una diferencia entre los sitios CpG hipo e hipermetilados sobre las etapas de progresión, aumentando su número de CpG metilados diferencialmente según la etapa de progresión. Este resultado nos proporciona evidencia de claras diferencias de metilación entre las etapas del cáncer. Para avanzar en nuestra comprensión de este fenómeno, diseñamos un método de filtrado con una mayor granularidad para obtener genes con diferencias en la metilación del promotor en comparación con el contraste evaluado.

Los genes con un valor medio de metilación inferior a 0.4 y sobreexpresados se consideraron *hipometilados*, mientras que los genes con un valor medio de metilación superior a 0.6 y subexpresados se etiquetaron como *hipermetilados*. Asociamos estas modificaciones de metilación con cambios fenotípicos y asumimos que fueron impulsadas por un mecanismo celular subyacente que debe explorarse más a fondo. En el suplementario S4 de [160] se incluye una lista con todos los genes filtrados para cada fenotipo. La Figura 4.17 muestra ejemplos del comportamiento de metilación en dos genes hipometilados (*IL32* y *TNFRSF9*) y dos hipermetilados (*ERMP1* y *RAB25*). El conjunto completo de diagramas de dispersión para todos los genes filtrados por este enfoque se puede generar con el código citado en 3.1.

En la Figura 4.18 se muestran los genes diferencialmente metilados (GDMs) obtenidos con el filtro anterior. Consideramos tanto los genes específicos de la etapa para cada contraste como los genes compartidos en todas las etapas. Dado que no hay genes hipermetilados exclusivos de la etapa en las etapas I, II y III, entonces analizamos sólo genes compartidos en las cuatro etapas.

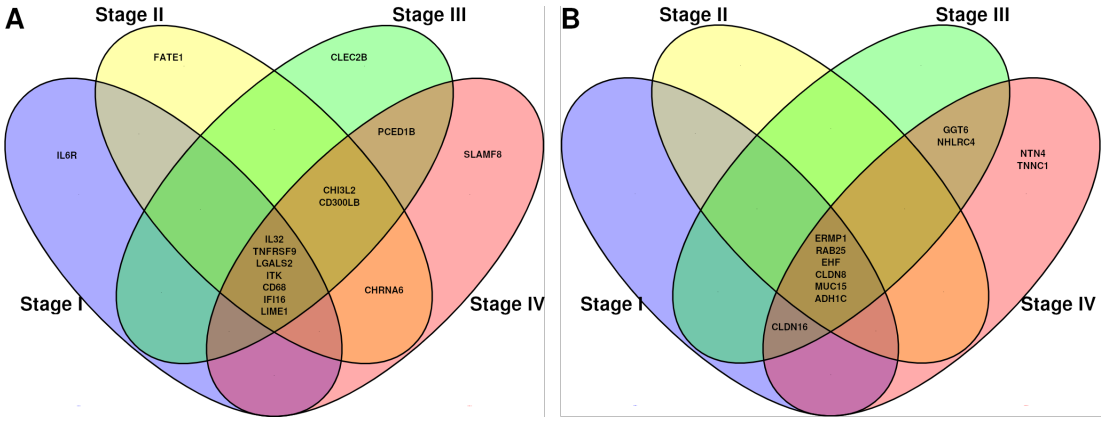

Figura 4.18: *Diagramas de Venn que muestran genes comunes manejados por metilación.* A) genes que disminuyeron su patrón de metilación de tejido normal a tejido tumoral (hipometilados). B) genes que aumentaron su patrón de metilación de tejido normal a tejido tumoral (hipermetilados).

4.3.3. ITK es un oncogén relacionado con la metilación; *RAB25* y *EHF* son supresores de tumores relacionados con la metilación

Identificamos genes específicos relacionados con la metilación en las cuatro etapas de progresión de CRcc, incluidos *IL32*, *CD68*, *EHF*, *MUC15* y otros. Estos genes se etiquetaron como oncogenes (OG), genes supresores de tumores (TSG) o ambos (si la evidencia respaldaba ambas características). Al final, sólo tres genes cumplieron los criterios de tener estas propiedades, a saber, *ITK*, *RAB25* y *EHF*. Para obtener una comprensión integral del fenómeno, construimos una red de coexpresión que involucra estos genes y sus genes vecinos (Fig. 4.19). Además, examinamos los patrones de expresión diferencial de estos genes. En las conclusiones de este trabajo profundizaremos en estos resultados (ver 5.2.2)

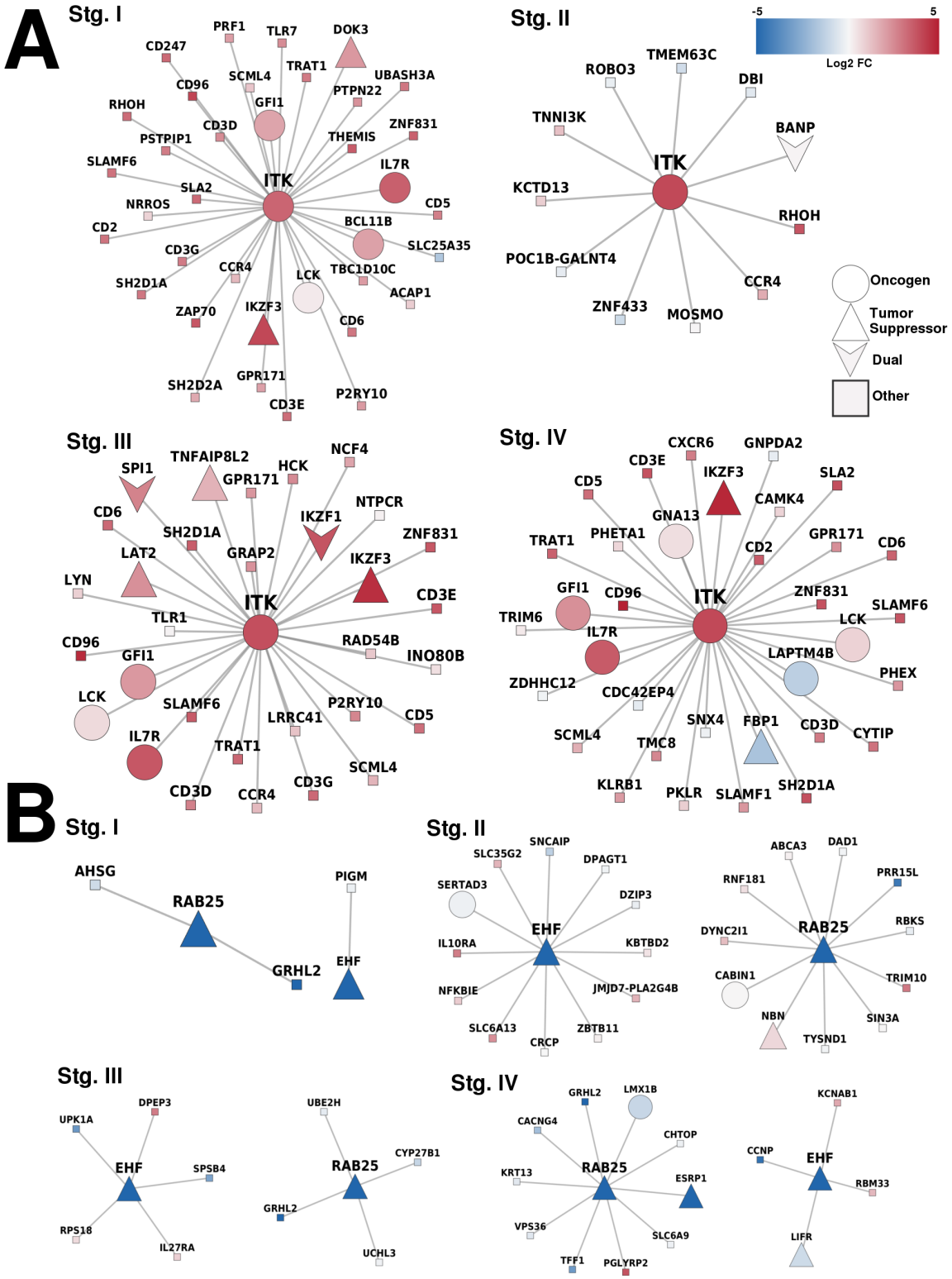

Figura 4.19: *Redes de coexpresión para los genes manejados por metilación.*

Figura 4.19: A) Genes sobreexpresados-hipometilados. *ITK* fue el único gen encontrado que cumplía con nuestros criterios. B) Redes para genes encontrados subexpresados e hipermetilados; en este caso, *EHF* y *RAB25*. Vale la pena notar la consistencia entre la tendencia de expresión diferencial de los GDMs y sus primeros vecinos.

Con este enfoque, identificamos posibles genes regulados por la metilación del ADN en el carcinoma renal de células claras; sin embargo, sigue existiendo incertidumbre con respecto a la correlación de *RAB25* e *ITK* con el pronóstico. De hecho, la correlación entre las características moleculares y el pronóstico del cáncer es una piedra angular en la investigación del cáncer.

4.3.4. La expresión de *RAB25* y *FOXP3* se asocia con mal pronóstico en CRcc

En el caso de *RAB25*, identificamos una disparidad sustancial en el pronóstico entre los niveles de expresión altos y bajos (Fig. 4.20A). Como se mencionó anteriormente, *RAB25* exhibe doble funcionalidad en la carcinogénesis. Actúa como un oncogén en algunos tipos de cáncer [76, 85] mientras funciona como un gen supresor de tumores en otros, como el cáncer colorrectal, el carcinoma de células escamosas de esófago y el carcinoma de células escamosas de cabeza y cuello [48, 140] . Una curva de Kaplan-Meier para *RAB25* demuestra que el grupo de alta expresión muestra un peor pronóstico en comparación con el grupo de baja expresión (valor de $p = 0.017$).

En cuanto a la expresión del gen *ITK*, es importante señalar su importante correlación con el pronóstico en otros tipos de cáncer. Estudios anteriores informaron una fuerte asociación entre la expresión de *ITK* y el mal pronóstico en adenocarcinoma de pulmón, cáncer de mama, carcinoma hepatocelular y linfoma [89, 105]. Sin embargo, en estos datos, no observamos una diferencia significativa en el pronóstico entre los niveles de *ITK* de expresión alta y baja (Fig. 4.20B). Para proporcionar información adicional, hemos incluido un diagrama de Kaplan-Meier para *FOXP3*, una molécula río abajo bien conocida en la vía de señalización de *ITK* [149], que de hecho exhibe un comportamiento distinto en relación con el pronóstico: la alta expresión de *FOXP3* está relacionada con un mal pronóstico en CRcc (Fig. 4.20C).

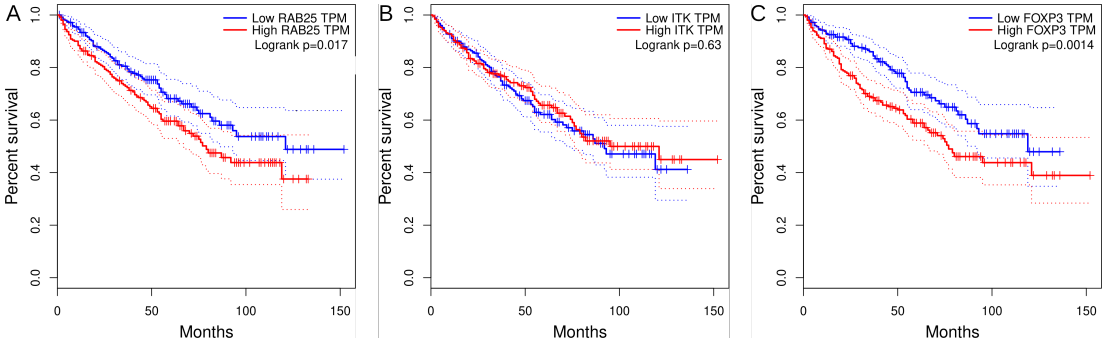

Figura 4.20: *Gráficos de Kaplan-Meier que correlacionan los genes relacionados con la metilación y la supervivencia de CRcc.* A) gen *RAB25*. B) *ITK*. C) *FOXP3*. Aquí se muestra el caso de *FOXP3*, ya que este gen es una molécula corriente abajo en la vía de señalización de *ITK*, pero no está modulada por la metilación.

4.3.5. Discusión sobre los efectos de la metilación en la coexpresión de los genes

Con el flujo de trabajo diseñado en este trabajo, hemos demostrado que algunos genes presentan una asociación significativa entre sus patrones de expresión y los perfiles de metilación. Este resultado era esperado, sin embargo, nos parecía importante cuantificarlo y sobre todo generar un modelo que nos orientara para proponer experimentos dirigidos. Se ha informado que algunos de esos genes están relacionados con el proceso oncogénico. Además, hemos mostrado evidencia de la correlación entre la expresión génica y el pronóstico. A continuación, discutiremos algunas ideas a la luz de los resultados antes mencionados.

En términos de los genes coexpresados con *ITK*, vale la pena notar que sus primeros vecinos en la red de coexpresión, tienen el mismo patrón de sobreexpresión. *ITK* codifica para una tirosina quinasa intracelular expresada en las células T. También tiene un papel fundamental en el crecimiento, la señalización y la función de las mismas. La activación de las células T y la regulación del sistema inmunitario fueron los dos procesos más significativamente enriquecidos. Estos resultados demuestran la relevancia de *ITK*, ya que este gen mantiene sus funciones independientemente de sus vecinos.

El resultado del análisis de sobrerrepresentación donde la activación de células T es el proceso más significativo involucrado en la red de primeros vecinos de *ITK*, refleja un hecho importante con respecto a la respuesta inmune en CRcc: independientemente de la etapa de progresión, se sugiere que la hipometilación de *ITK* promueve la activación de células T en el fenotipo canceroso. La célula de origen en la que se produce esta hipometilación es materia de investigación adicional.

El caso de la expresión de *ITK* y su correlación con el pronóstico es intrigante.

ITK normalmente no se expresa en células claras dentro del riñón, lo que sugiere que este gen puede originarse a partir de células inmunitarias. Además, como dijimos líneas arriba, el CRcc se caracteriza por una infiltración inmune y estromal significativas, lo que enfatiza aún más la relevancia potencial de los genes río abajo influenciados por *ITK*, como *FOXP3* [149].

En el caso de *RAB25* y *EHF*, ambos genes son conocidos supresores de tumores que cambiaron su estado de metilación. En el tejido normal, ambos estaban hipometilados, pero su metilación aumentaba en el cáncer en cualquier etapa. Como se observa, los patrones de expresión en los primeros vecinos de estos genes son similares, ya sea subexpresados o sin cambios. Este fenómeno puede deberse a una especie de efecto de *anclaje* en las redes, en las que los vecinos siguen patrones de coexpresión de genes que afectan su expresión individual a través de algún mecanismo regulador como factores de transcripción, metilación cercana o efectos conformacionales [142].

La función oncogénica de *RAB25* probablemente se atribuya a su papel en la regulación del tráfico vesicular, lo que aumenta el reciclaje de integrinas a la membrana plasmática y estimula las vías de señalización intracelular asociadas con las funciones oncogénicas [2]. En particular, la pérdida de Rab25 en los cánceres de colon humanos se ha relacionado con un peor pronóstico de los pacientes [100].

Además, se ha demostrado que la expresión reducida de *RAB25* se correlaciona con una supervivencia general disminuida y se ha documentado en líneas celulares de carcinoma de células escamosas de esófago (EScc) en comparación con tejidos normales agrupados [140]. También se ha observado que la expresión de *RAB25* tanto en líneas celulares de EScc como en muestras clínicas está asociada con la hipermetilación del promotor [51]. La proteína codificada por *RAB25* es miembro de la superfamilia RAS de pequeñas GTPasas [97] y está involucrada en el tráfico de membranas y la supervivencia celular [147]. Hay evidencia de que este gen actúa como supresor de tumores y también como oncogen, dependiendo del contexto [97]. Se han identificado dos variantes, una codificante de proteína y otra no codificante, para este gen [2].

Hemos demostrado que los procesos relacionados con el epitelio se enriquecieron, los datos son reportados en [160]. Este resultado agrega evidencia de los impactos sobre las modificaciones de la matriz extracelular en la evolución del tumor [36, 41, 108, 152]. En este caso, *RAB25* de alguna manera está perdiendo su funcionalidad debido a su subexpresión acompañada de sus componentes coexpresados.

Es importante destacar la subexpresión de *RAB25* en las muestras de CRcc. En términos de valor pronóstico, la expresión alta de *RAB25* se asocia con un resultado desfavorable, pero su expresión está regulada por el perfil de metilación dentro de esas muestras. Se puede plantear la hipótesis de que la metilación de *RAB25* puede impedir su sobreexpresión, lo que influye en el pronóstico.

Como se mencionó en la sección 4.3.3, identificamos genes relacionados con la metilación que se sobreexpresaron e hipometilaron en asociación con la etapa de progresión del CRcc. Estos genes incluyen *IL32*, *TNFRSF9*, *LGALS2*, *CD68*, *IFI16* y *LINE1*. Estos genes exhibieron una sobreexpresión significativa a lo largo de todas las etapas de progresión, mientras que también estaban significativamente hipometilados con valores

β por debajo de 0.4. En particular, estos genes están asociados con procesos del sistema inmunitario, alineándose con el papel de *ITK* en la progresión del CRcc.

Por ejemplo, la sobreexpresión de IL32 se ha identificado como un factor pronóstico en pacientes con CRcc localizado [78]. De manera similar, se ha sugerido que *IL32* muestra una correlación positiva entre su expresión y el estado de metilación correspondiente en el melanoma cutáneo de la piel [71].

En cuanto al gen *TNFRSF9*, su sobreexpresión se ha asociado con la progresión y el pronóstico en CRcc [86]. Además, se ha encontrado que está inversamente correlacionado con la metilación del ADN en varios sitios CpG en melanoma. La expresión elevada del ARNm de *TNFRSF9* y la hipometilación de *TNFRSF9* se relacionaron con una mayor supervivencia en general [43].

En el caso del gen *CD68*, los niveles altos de *CD68* se asocian con un grado tumoral más alto, un tamaño tumoral más grande, positividad para Ki67 y otras características malignas, lo que indica progresión y agresividad del tumor [161]. El perfil de metilación y su relación con la expresión se han asociado con el pronóstico en el carcinoma papilar de células renales [90]. Sin embargo, su relación con la progresión en este tipo de cáncer no ha sido reportada previamente.

Por último, *IFI16* promueve la progresión del cáncer de cuello uterino a través de la vía NF-kB [16]. La expresión de este gen también se ha correlacionado con el estado de metilación en líneas celulares de cáncer de mama [73].

A pesar de que estos genes se observan en relación con diferentes tipos de cáncer y su perfil de metilación muestra una correlación con la expresión génica, no se ha informado una correlación distinta entre la metilación y la expresión génica durante la progresión del cáncer.

En este trabajo, hemos demostrado que las redes de coexpresión formadas por genes relacionados con la metilación difieren constantemente entre las etapas de progresión, como se muestra en la Fig. 4.19. Por lo tanto, se puede inferir que los genes relacionados con la metilación observados durante las etapas de CRcc actúan de manera diferente en cada etapa de progresión, y cada etapa se ve afectada de manera diferente por estos genes relacionados con la metilación.

Argumentamos que los cambios en los genes metilados pueden estar asociados con la progresión del cáncer en al menos dos formas generales: 1) genes que cambian su estado de metilación/expresión en cada etapa del cáncer, o 2) genes cuyo estado de metilación/expresión no se ve afectado en todas las etapas. Estas huellas dactilares epigenéticas se pueden estudiar como biomarcadores en un análisis prospectivo [141]. Sugerimos una relación entre la coexpresión y los genes metilados durante la progresión del cáncer. Dado que la metilación puede reprimir la expresión génica, podemos inferir redes de genes diferencialmente metilados para cada estadio/fenotipo tumoral. Como resultado, podemos relacionar algunas funciones biológicas con eventos marcados por modificaciones epigenéticas.

Con este enfoque sistemático y automatizado, pudimos identificar CpG individuales asociados con genes candidatos como genes relacionados con la metilación. Esta asociación se basa en una prueba que compara genes expresados diferencialmente con CpG

diferencialmente metilados. Esto respalda aún más la evidencia de la metilación del ADN como uno de los principales factores que afectan los cambios entre las etapas del tumor y la carcinogénesis [99]. Con base en esto último, analizamos cómo la metilación afecta el programa de coexpresión en su conjunto. Encontramos evidencia de que los grupos de genes coexpresados pueden activar mecanismos de defensa antitumorales, funciones celulares específicas como la activación de células T y la regulación del sistema inmunitario [144]. Por otro lado, identificamos genes subexpresados e hipermetilados que resultaron coexpresados. Esto puede desactivar las funciones celulares y, por lo tanto, alterar la morfología y la diferenciación [99].

En suma a estas funciones celulares clave afectadas, encontramos que un gen supresor de tumores reportado (*RAB25*) estaba hipermetilado y subexpresado en tres de las cuatro etapas del cáncer, a través de una vía de señalización FAK-Raf-MEK1/2-ERK desregulada [50].

También mostramos evidencia de que la expresión de *ITK* está impulsada por la metilación, ya que su hipometilación en el cáncer resultó en una sobreexpresión. Esta modificación epigenética puede estar impulsando una respuesta antitumoral en cuatro etapas, activando funciones de respuesta inmune [124]. Actualmente, *ITK* no se reporta como un oncogen, pero similar al caso de *RAB25*, nosotros proponemos *ITK* como un biomarcador epigenético.

Según [159], se ha observado un aumento progresivo de varias quimiocinas [158] en la progresión de CRcc. En este caso, *CXCL13* destaca por aprovechar la migración de células del sistema inmunitario. Esta molécula desencadena vías intracelulares que conducen a la migración celular en los ganglios linfáticos, los tejidos endoteliales y epiteliales [72]. La importancia del microambiente tumoral es bien conocida [54]. El hecho de que el proceso más enriquecido asociado con los primeros vecinos de *ITK* sea la activación de células T nos sugiere la relevancia que ejerce la infiltración inmune en este carcinoma. Con este trabajo, se han desarrollado enfoques bioinformáticos para cuantificar la infiltración celular en tumores, basados en firmas moleculares [7, 82, 157]. Curiosamente, utilizando datos derivados de TCGA, el carcinoma renal fue el tumor con más infiltración celular entre 14 tejidos [157].

Capítulo 5

Conclusiones

El carcinoma renal de células claras es una enfermedad en la que se encuentran involucradas varias capas de complejidad. Con el objetivo de estudiar esta patología se debe diseccionar el fenómeno para tener un panorama completo que nos permita mejorar el entendimiento de su progresión. En este trabajo utilizamos tres enfoques: la coexpresión (regulación genética), la regulación de la transcripción por miRNAs y la metilación.

5.1. Coexpresión y regulación genética

En trabajos previos observamos un incremento importante en la correlación *-cis* en los subtipos moleculares de cáncer de mama, de acuerdo con la malignidad de esos fenotipos. Dado que la pérdida de la coexpresión de largo alcance se observó en el cáncer de mama y más notablemente en el subtipo Basal (el de peor pronóstico). Una variante de la hipótesis de trabajo es que cuanto más avanzado es el estadio del cáncer, mayor es la proporción de coexpresiones *-cis*.

Después del análisis de la red de cáncer de mama y pulmón, el carcinoma renal de células claras es el tercer cáncer en el que observamos una diferencia notable entre las interacciones *-cis* y *-trans*, esto confirma una disminución importante en la coexpresión intracromosomal de genes en las redes de cáncer.

Inesperadamente, el aumento de las coexpresiones *-cis* no coincide con la etapa de progresión en el carcinoma renal de células claras. Esto se observó no solo en las redes de 10K enlaces, sino también en las redes con un rango de más de cinco órdenes de magnitud. Esto podría implicar que la proporción de enlaces *-cis* no es un parámetro para distinguir etapas de progresión, al menos para CRcc.

Al observar la discrepancia entre la tasa de enlaces *-cis* en la progresión de CRcc con las observadas en los subtipos moleculares de cáncer de mama [44], donde hay una alta proporción de interacciones intracromosómicas en aquellos fenotipos con peor pronóstico podemos argumentar lo siguiente:

- El hecho de que la tasa de enlaces *-cis* no coincida con las etapas de progresión,

puede reflejar que la alta proporción de interacciones intracromosómicas no es un parámetro a tener en cuenta para diferenciar la progresión del cáncer, al menos en el carcinoma renal de células claras.

- Una alta tasa de enlaces -*cis* no implica malignidad o peor pronóstico en una red cancerosa, sino un programa de coexpresión diferente en el que, probablemente, se favorezcan las interacciones entre genes físicamente cercanos.
- Los mecanismos detrás de la coexpresión preferencial en genes vecinos deben implicar factores epigenéticos, como microRNAs con dianas relacionadas a: 1) la remodelación de la cromatina, 2) lncRNA, 3) perfiles de metilación, 4) estructura tridimensional del ADN, 5) sitios de unión a CTCF, etc. (Para una revisión profunda de la regulación espacial del ADN en el proceso oncogénico, véase [62]).

Queremos enfatizar que los cánceres de riñón son fundamentalmente diferentes de los cánceres de mama en muchas formas [63]. En este sentido, las similitudes topológicas entre el cáncer de mama y las redes de coexpresión de CRcc deben tomarse con cuidado. Sin embargo, es notable que en ambos tejidos, así como en instancias separadas (etapas de progresión y subtipos moleculares), el efecto de la pérdida de coexpresión de largo alcance es una característica común del cáncer (incluyendo pulmón [6]).

En este primer enfoque, nos hemos centrado en dos firmas moleculares principales, a saber, la expresión y los paisajes de coexpresión. En la primera capa, hemos observado que el perfil de expresión diferencial es muy similar entre etapas de progresión, incluso entre la etapa I y la etapa IV, lo que puede indicar que el perfil de expresión se adquiere de alguna manera una vez que el cáncer ha comenzado. Sin embargo, ciertos genes parecen replicar la progresión del proceso oncogénico, como es el caso de *SLC6A19* y *PLG* (subexpresión), y *SAAC2-SAAC4* y *CXCL13* (sobreexpresión). Vale la pena mencionar que ninguno de estos genes había sido reportado como progresivamente diferenciado en el carcinoma renal.

Por otro lado, la similitud observada a nivel de expresión, no se replicó a nivel de red de coexpresión. En realidad, el número de enlaces compartidos es realmente bajo. Argumentamos que los perfiles de expresión diferencial son de hecho insuficientes para describir adecuadamente la regulación de la expresión génica, pero la forma en que esos genes interactúan en el tiempo y el espacio es lo que finalmente determina el establecimiento del fenotipo tumoral.

5.2. Modelo epigenético de Metilación y regulación por miRNAs

5.2.1. Sobre la actividad de miRNAs y su relación con los genes

En este trabajo, hemos construido un conjunto de redes para proporcionar un marco descriptivo sobre la evolución del paisaje de coexpresión miRNA-gen durante la pro-

gresión del carcinoma renal de células claras. Como resumen de los hallazgos, podemos establecer lo siguiente:

- Con este enfoque, pudimos encontrar genes y miRNAs expresados diferencialmente para cada etapa de progresión. Al mismo tiempo, fuimos capaces de inferir redes filtradas para buscar interacciones reguladoras canónicas miRNA-gen.
- La diferencia más grande en cuanto al número de genes expresados diferencialmente, así como en el número de interacciones entre un gen y un miRNA ocurren entre el control y la etapa I.
- Cada red se comporta de manera diferente en términos de miRNAs y genes involucrados. Esas redes no comparten interacciones, y la gran mayoría de los enlaces gen-miRNA son únicos para cada red de etapa de progresión.
- *miR-217* se expresa diferencialmente en todas las redes. Es el único microARN que se expresa diferencialmente en cada etapa con blancos genéticos expresados opuestamente.
- *miR-217* se correlaciona con un conjunto de genes completamente diferente según la etapa de progresión. Además, la expresión diferencial de todos estos genes diana coincide con su papel como oncogenes o genes supresores de tumores, reportados en otras bases de datos.
- El hallazgo de *LAMA4*, *BIRC7*, *GALNTL6*, *WNK2* e *IGF2BP2* como blancos potenciales de *miR-217* en diferentes momentos de la evolución del tumor puede ayudar a desarrollar estrategias etapa-específicas, teniendo en cuenta la expresión diferencial de *miR-217* en cada estadio de progresión del carcinoma renal de células claras.
- Hasta donde sabemos, esta es la primera vez que se rastrea la evolución de los patrones de expresión de un microARN durante todos los pasos de la progresión del carcinoma y, al mismo tiempo, se observa su capacidad para regular diferentes objetivos según la evolución de la enfermedad. A este comportamiento lo denominamos modelo de "switch".

Vale la pena destacar que a la luz de estos resultados, el presente trabajo plantea un detalle conceptual fino pero muy importante. Los miRNAs canónicamente realizan su función reguladora fuera del núcleo (ver sección 1.1.3.1), más específicamente dentro de un escenario *pre-traduccional* cercano a los ribosomas. Como los resultados lo muestran, esto no estaría siendo estrictamente una regulación epigenética. Por ejemplo, *miR-217* actúa sobre genes con funciones biológicas fuera del núcleo que esencialmente no interactúan con la cromatina (*IGF2BP2*, *WNK2* y *GALNTL6*, por ejemplo). Sin embargo, si un miRNA estuviera marcando un regulador epigenético, por ejemplo, el ARNm de DNMTs, ó de subunidades de Policomb, ó algún lncARN regulador, ó modificadores de histonas. Si fuera así, entonces, podemos considerar que el miRNA tiene una función

epigenética. Para un mejor entendimiento de este concepto mostramos la figura 5.1. Este concepto también nos habla de la complejidad para definir la interdependencia "Génetica ⇔ Epigéntica". Es muy difícil, poder establecer cuál es el origen de esta interdependencia. Pero además los participantes pueden tener un origen epigenético o génetico. Por ejemplo, los miRNAs también pueden ser regulados por algún miRNA [8]. Estos conceptos también subrayan la necesidad de abordar estas funciones biológicas como un sistema complejo.

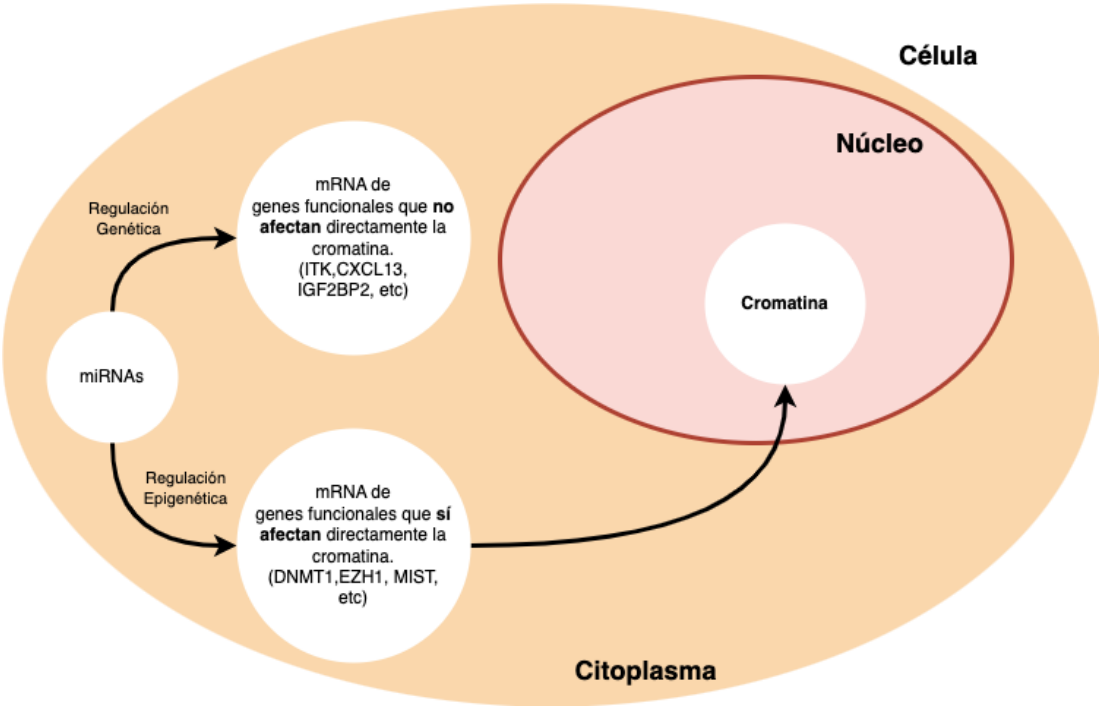

Figura 5.1: *Diagrama que muestra los diferentes caminos de regulación llevados a cabo por los microRNAs.* En términos generales, se considera un microRNA epigenético como aquel que regula genes que tiene impacto directo sobre la cromatina. De forma canónica todos los miRNAs tienen la capacidad de afectar a cualquier gen.

Varias de las hipótesis que éste y otros estudios han generado deben probarse experimentalmente en diferentes condiciones para capturar completamente los mecanismos potenciales y sus implicaciones. En la sección 6 se proponen diferentes caminos para darle continuidad experimental a este trabajo.

5.2.2. Efectos de la metilación sobre los cambios en la expresión de los genes

El carcinoma renal de células claras es una enfermedad altamente heterogénea. Por lo tanto, para obtener una imagen completa y comprender mejor su progresión, se deben diseccionar los orígenes, la evolución y las características asociadas. En trabajos previos se han encontrado diferencias importantes en las etapas de progresión del cáncer, destacando fenómenos como el claro sesgo a la coexpresión entre genes de un mismo cromosoma [6, 31, 37, 44, 46, 158, 159], o las diferencias en los procesos de enriquecimiento a lo largo de las etapas de progresión en el carcinoma renal de células claras [158]. Sin embargo, la conexión con los mecanismos epigenéticos sigue siendo una cuestión clave en el campo [99].

Si bien la reproducibilidad es una piedra angular de la investigación científica, validar nuestros hallazgos con otro conjunto de datos similar al que obtuvimos de The Cancer Genome Atlas plantea desafíos importantes. En primer lugar, nuestro estudio utilizó dos tecnologías diferentes de alto rendimiento, a saber, RNA-Seq (expresión genética) e Illumina HumanMethylation450 (HM450) para datos de metilación. Ambas tecnologías se combinaron cuidadosamente para cada individuo en nuestro conjunto de datos. Además, las muestras se estratificaron según el estadio de progresión, lo que requirió la integración de información clínica como el estadio de progresión y el estado vital. Nos aseguramos de que cada grupo contuviera una cantidad sustancial de muestras para análisis posteriores a fin de obtener resultados estadísticamente significativos. Hasta donde sabemos, ningún otro conjunto de datos actualmente disponible posee todas las características antes mencionadas. Sin embargo, teniendo en cuenta la cantidad de muestras coincidentes para cada fenotipo, la tecnología utilizada para la secuenciación y los altos estándares para el manejo de muestras, nos permite tener un marco sólido para estudiar los cambios en la expresión génica dependiendo de la metilación. Al mismo tiempo, las estadísticas astringentes, así como la reproducibilidad del flujo de trabajo computacional, sugieren realizar este análisis en otros tejidos cancerosos de TCGA.

Esta investigación ha destacado el importante papel de los genes relacionados con la metilación en la modulación de las funciones biológicas y la contribución a la progresión de varios carcinomas, incluido el carcinoma renal de células claras.

Si bien actualmente no existe un método definitivo para establecer relaciones claras entre los factores genéticos y epigenéticos que afectan la progresión del cáncer, hemos desarrollado un enfoque bioinformático para identificar genes relacionados con la metilación y establecer su relación con la coexpresión génica y la regulación de la metilación en todo el genoma.

Este análisis ha identificado varios genes, incluidos *ITK* y *TSFRN9*, que aparecen hipometilados y fuertemente involucrados en las funciones de la respuesta inmunitaria a lo largo de las cuatro etapas de la progresión del CRcc. Análogamente, el gen supresor de tumores *RAB25*, está hipermetilado y potencialmente evita funciones reprimidas en la vía de señalización de AKT durante la evolución del CRcc. Los detalles de las relaciones

biológicas de estos genes se revisan en la sección 4.3.3. Estos hallazgos brindan información importante sobre los mecanismos epigenético-genéticos subyacentes involucrados en la progresión del cáncer.

5.3. Trabajos publicados en revistas internacionales arbitreadas

Los resultados derivados de esta tesis formaron parte integra en la publicación de los siguientes tres artículos:

- Zamora-Fuentes, J. M., Hernández-Lemus, E., and Espinal-Enríquez, J. Gene expression and co-expression networks are strongly altered through stages in clear cell renal carcinoma. Frontiers in Genetics 11 (Nov. 2020).
- Zamora-Fuentes, J. M., Hernández-Lemus, E., and Espinal-Enríquez, J. Oncogenic role of miR-217 during clear cell renal carcinoma progression. Frontiers in Oncology 12 (2022),
- Zamora-Fuentes, J. M., Hernández-Lemus, E., and Espinal-Enríquez, J. Methylation-related genes involved in renal carcinoma progression. Frontiers in Genetics 14 (2023).

Adicionalmente, en esta etapa se publicaron otros trabajos teóricos y experimentales que utilizan la metodología y el conocimiento generado en esta tesis (por ejemplo, [6], [49]).

Capítulo 6

Perspectivas

6.1. Propuestas experimentales

Uno de los valores más importantes de este trabajo es el planteamiento teórico de nuevas hipótesis experimentales. Es decir, con los resultados sistemáticos de este trabajo se pueden plantear nuevas preguntas de investigación dentro de un marco experimental. A continuación se listan algunas de las ideas clave para darle continuidad a este trabajo.

- Como concluimos en las secciones 4.1.1 y 4.1, los resultados sobre la perdida de coexpresión *-trans* en cáncer, representan uno de los efectos biológicos más interesantes para conocer los mecanismos de la enfermedad. En este sentido, uno de los experimentos que hemos planteado es la utilización de la técnica HI-C. Con este experimento podemos demostrar si estas afectaciones en la coexpresión, son estrictamente, un efecto dentro de la organización espacial de la cromatina.

  En particular, este experimento no se ha planteado en CRcc específicamente por etapas, mucho menos para una población latina. Y aunque estamos conscientes que la recuperación de muestras es complicada, con este trabajo ya se tiene el marco teórico para plantear el experimento en colaboración con otros grupos de investigación.

- Respecto al estudio de miRNAs, una extensión de este trabajo es integrar el análisis con otras fuentes ómicas. Como se señaló, la pregunta biológica que quedó fuera de este marco, es la respuesta epigenómica de los miRNAs altamente conectados (*miR-10A*, *miR-196A-1*, *miR-196A-2* y *miR-196B*, por ejemplo). Con experimentos como: ChIP-seq, ATAC-seq, Hi-C, METHYL-seq, etc., se puede probar cuáles de ellos tienen como blancos genes involucrados en el remodelaje de la cromatina y determinar cuáles son sus funciones en el epigenóma.

  Por otra parte, la expresión de este conjunto de miRNAs se puede cruzar con bases de datos de ATAC-seq o CHIP-seq, con el fin de rastrear su papel como regulador epigenético. Por ejemplo, con CHIP-Seq podemos encontrar efectos en

marcas de histonas de posibles genes diana de los miRNAs hipotéticos. Con la técnica ATAC-seq podemos evaluar los efectos en la cromatina directamente (es decir, si se observa abierta o cerrada) de los genes reprimidos por estos miRNAs.

Asimismo, resulta llamativo estudiar el papel de los ARN largos no codificantes como factores de regulación de los propios miRNAs. Hay que recordar que sólo un pequeño porcentaje del genoma transcribe a gen, el resto es ARN no codificante. Actualmente ya existen técnicas experimentales más finas para dilucidar este proceso (como RNAseq *bulk*).

Hoy en día las bases de datos como GEO (Gene Expression Omnibus) contienen más experimentos con muestras suficientes de tumores, incluyendo CRcc. Por tanto, la idea de integrar varias fuentes ómicas proporciona un modelo más realista de la regulación transcriptómica y los efectos epigenéticos en el cáncer. Estos avances serán importantes para una comprensión más completa de los programas de regulación en enfermedades como el cáncer.

- En este trabajo se propone un modelo novedoso de funcionamiento sobre los miRNAs (ver Figura 4.15 y sección 4.2.4). Este modelo biológico es dinámico en el tiempo, es decir, a través de las cuatro etapas del cáncer. Una de las implicaciones más importantes de este trabajo es el modelo de "*switch*" para un miRNA.

  La demostración experimental de un modelo con estas características es compleja, sin embargo, creemos que un punto de partida puede ser utilizar líneas celulares de CRcc. Inicialmente, hemos planteado tomar una línea celular en cualquier etapa y verificar el resultado de apagar (*knockout*) *miR-217*. Posteriormente, realizar un experimento de secuenciación de mRNA y verificar la expresión de los genes diana (*GALNTL6*, *WNK2* y *IGF2BP2*) revisados en nuestro modelo.

  Durante el experimento sería importante monitorear posibles efectos en las vías de señalización afectadas por estos genes, a saber, mal funcionamiento en la glicosilación ó cambios celulares en las vías MEK1 ó PI3K-Akt. Eventualmente, se podría repetir el experimento con muestras de CRcc en humanos, de ser posible divididas por etapa, con la finalidad de darle las características dinámicas al resultado. Estamos conscientes de las dificultados clínicas para obtener las muestras de tumores adecuadas para el experimento, y debido a esto destacamos la importancia de un antecedente en líneas celulares.

- Nuestros resultados en metilación relacionados con la expresión genética (el *eje CXCL13-ITK*), nos dio un indicio claro del papel fundamental del sistema inmune en el cáncer. Y en este sentido, nos inclinamos por plantear experimentos de secuenciación de célula única (sC-RNAseq). Con estos experimentos podemos tener un panorama más preciso e íntegro de las funciones inmunológicas que están produciendo cada uno de los grupos celulares dentro del tumor; por ejemplo, a nivel de grupos celulares, incluyendo subpoblaciones de células T.

  Por un lado, nos interesaría investigar sobre las funciones "atacantes" de las células T (CD8, CD4 y subtipos) que pudieran mostrar sobreexpresión de *ITK*. Esto po-

dría ser el antecedente de un gen que interviene en la reprogramación de las células tumorales. Por otro lado, nos interesa también dilucidar el funcionamiento de las citocinas "contaminantes" en el ambiente tumoral. Un experimento de scRNA-seq podría indicarnos las subpoblaciones que producen citocinas específicas.

Con el experimento anterior podemos mejor el entendimiento del microambiente tumoral, incluso, a nivel del desarrollo folicular. Aunado a esto, el antecedente de CXCL13 como formador de folículos inmunológicos en los nodos linfáticos en condiciones normales, nos puede proveer las bases para investigar la reprogramación del sistema inmune intratumoral.

En este sentido, *CXCL13* puede ser la clave. Rastrear la función de este gen a nivel de célula única nos puede proporcionar el panorama funcional de las células que están produciendo las citocinas atractoras en la formación de folículos tumorales por *CXCL13*.

Considerando lo complicado que puede ser obtener muestras de tumores humanos, este experimento puede ser planeado en carcinomas humanos o en modelos animales.

Vale la pena destacar que todas las extensiones adicionales de este trabajo podrían estar relacionadas con la clasificación de las muestras en función de otras características clínicas y no solo en la etapa de progresión, como la edad, el sexo o la sobrevida.

Bibliografía

[1] N. Abd-Aziz, N. I. Kamaruzman, and C. L. Poh. Development of MicroRNAs as potential therapeutics against cancer. *Journal of Oncology*, 2020:1–14, July 2020. 7

[2] R. Agarwal, I. Jurisica, G. B. Mills, and K. W. Cheng. The emerging role of the rab25 small gtpase in cancer. *Traffic*, 10(11):1561–1568, 2009. 71

[3] S. A. Alcalá-Corona, G. De Anda Jáuregui, J. Espinal-Enríquez, and E. H.-L. Hernández-Lemus. Network modularity in breast cancer molecular subtypes. *Frontiers in Physiology*, 8:915, 2017. 21, 22

[4] S. A. Alcalá-Corona, J. Espinal-Enríquez, G. De Anda Jáuregui, and E. H.-L. Hernández-Lemus. The hierarchical modular structure of her2+ breast cancer network. *Frontiers in Physiology*, 9, 2018. 21, 22

[5] S. A. Alcalá-Corona, T. E. Velázquez-Caldelas, J. Espinal-Enríquez, and E. Hernández-Lemus. Community structure reveals biologically functional modules in MEF2C transcriptional regulatory network. *Frontiers in physiology*, 7, 2016. 22

[6] S. D. Andonegui-Elguera, J. M. Zamora-Fuentes, J. Espinal-Enríquez, and E. Hernández-Lemus. Loss of long distance co-expression in lung cancer. *Frontiers in Genetics*, 12, mar 2021. 12, 55, 61, 76, 79, 80

[7] D. Aran, Z. Hu, and A. J. Butte. xcell: digitally portraying the tissue cellular heterogeneity landscape. *Genome biology*, 18:1–14, 2017. 73

[8] R. O. Bak and J. G. Mikkelsen. miRNA sponges: soaking up miRNAs for regulation of gene expression: MicroRNA sponges. *Wiley Interdisciplinary Reviews: RNA*, 5(3):317–333, May 2014. 78

[9] A. Balsalobre and J. Drouin. Pioneer factors as master regulators of the epigenome and cell fate. *Nature Reviews Molecular Cell Biology*, Mar. 2022. 3

[10] M. S. Balzer, T. Rohacs, and K. Susztak. How Many Cell Types Are in the Kidney and What Do They Do? *Annual Review of Physiology*, 84(1):507–531, 2022. 2

[11] C. L. Bartels and G. J. Tsongalis. MicroRNAs: Novel biomarkers for human cancer. *Clinical Chemistry*, 55(4):623–631, Apr. 2009. 4

[12] A. Bateman, M.-J. Martin, S. Orchard, M. Magrane, and R. Agivetova. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, nov 2020. 58

[13] J. L. Bell, K. Wächter, B. Mühleck, and N. Pazaitis. Insulin-like growth factor 2 mRNA-binding proteins (IGF2bps): post-transcriptional drivers of cancer progression? *Cellular and Molecular Life Sciences*, 70(15):2657–2675, oct 2012. 59

[14] S. Bhatlekar, J. Z. Fields, and B. M. Boman. HOX genes and their role in the development of human cancers. *Journal of Molecular Medicine*, 92(8):811–823, jul 2014. 54

[15] N. P. Blackledge and R. J. Klose. The molecular principles of gene regulation by Polycomb repressive complexes. *Nature Reviews Molecular Cell Biology*, 22(12):815–833, Dec. 2021. 2

[16] H. Cai, L. Yan, N. Liu, M. Xu, and H. Cai. Ifi16 promotes cervical cancer progression by upregulating pd-l1 in immunomicroenvironment through sting-tbk1-nf-kb pathway. *Biomedicine & pharmacotherapy*, 123:109790, 2020. 72

[17] A. Caliò, S. Marletta, M. Brunelli, and G. Martignoni. WHO 2022 Classification of Kidney Tumors: What is relevant? An update and future novelties for the pathologist. *Pathologica*, pages 1–9, Jan. 2023. 12

[18] T. Cheng, J.-G. Zhang, Y.-H. Cheng, Z.-W. Gao, and X.-Q. Ren. Relationship between PTEN and livin expression and malignancy of renal cell carcinomas. *Asian Pacific Journal of Cancer Prevention*, 13(6):2681–2685, jun 2012. 57

[19] S. Chevrier, J. H. Levine, V. R. T. Zanotelli, and K. Silina. An immune atlas of clear cell renal cell carcinoma. *Cell*, 169(4):736–749.e18, May 2017. 7

[20] M. Cobb. 60 years ago, Francis Crick changed the logic of biology. *PLOS Biology*, 15(9):e2003243, Sept. 2017. 2

[21] A. Colaprico, T. C. Silva, C. Olsen, L. Garofano, and C. Cava. TCGAbiolinks: an r/bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*, 44(8):e71–e71, Dec. 2015. 28

[22] J. F. Costello, M. C. Frühwald, D. J. Smiraglia, L. J. Rush, and G. P. Aberrant CpG-island methylation has non-random and tumour-type–specific patterns. *Nature Genetics*, 24(2):132–138, Feb. 2000. 10

[23] F. Crick. The origin of the genetic code. *Journal of Molecular Biology*, 38(3):367–379, Dec. 1968. 1

[24] F. H. Crick. On protein synthesis. In *Symp Soc Exp Biol*, volume 12, page 8, 1958. 1

[25] C. M. Croce. Oncogenes and Cancer. *The New England Journal of Medicine*, 2008. 3, 7

[26] G. Csardi, T. Nepusz, et al. The igraph software package for complex network research. *InterJournal, complex systems*, 1695(5):1–9, 2006. 21

[27] A. Dasgupta, C. S. Alvarado, Z. Xu, and H. W. Findley. Expression and functional role of inhibitor-of-apoptosis protein livin (BIRC7) in neuroblastoma. *Biochemical and Biophysical Research Communications*, 400(1):53–59, sep 2010. 57

[28] G. de Anda-Jáuregui, J. Espinal-Enríquez, D. Drago-García, and E. Hernández-Lemus. Nonredundant, highly connected MicroRNAs control functionality in breast cancer networks. *International Journal of Genomics*, 2018:1–10, May 2018. 61

[29] G. de Anda-Jáuregui, J. Espinal-Enriquez, and E. Hernández-Lemus. Spatial organization of the gene regulatory program: An information theoretical approach to breast cancer transcriptomics. *Entropy*, 21(2):195, Feb. 2019. 12, 19

[30] G. de Anda-Jáuregui, J. Espinal-Enríquez, and E. Hernández-Lemus. Highly connected, non-redundant microRNA functional control in breast cancer molecular subtypes. *Interface Focus*, 11(4):20200073, June 2021. 61

[31] G. de Anda-Jáuregui, C. Fresno, D. García-Cortés, J. E. Enríquez, and E. Hernández-Lemus. Intrachromosomal regulation decay in breast cancer. *Applied Mathematics and Nonlinear Sciences*, 4(1):223–230, Jan. 2019. 12, 19, 79

[32] R. Dorantes-Gilardi, D. García-Cortés, E. Hernández-Lemus, and J. Espinal-Enríquez. Multilayer approach reveals organizational principles disrupted in breast cancer co-expression networks. *Applied Network Science*, 5(1), Aug. 2020. 12, 42, 61

[33] R. Dorantes-Gilardi, D. García-Cortés, E. Hernández-Lemus, and J. Espinal-Enríquez. k-core genes underpin structural features of breast cancer. *Scientific Reports*, 11(1), Aug. 2021. 21, 30, 61

[34] D. Drago-García, J. Espinal-Enríquez, and E. Hernández-Lemus. Network analysis of emt and met micro-rna regulation in breast cancer. *Scientific reports*, 7(1):13534, 2017. 7, 19, 61

[35] J. Espinal-Enriquez, C. Fresno, G. Anda-Jáuregui, and E. Hernández-Lemus. Rna-seq based genome-wide analysis reveals loss of inter-chromosomal regulation in breast cancer. *Scientific reports*, 7:1760, May 2017. 55, 61

[36] J. Espinal-Enriquez, S. Munoz-Montero, I. Imaz-Rosshandler, A. Huerta-Verde, C. Mejia, and E. Hernandez-Lemus. Genome-wide expression analysis suggests a crucial role of dysregulation of matrix metalloproteinases pathway in undifferentiated thyroid carcinoma. *BMC Genomics*, 16(1), Mar 2015. 71

[37] J. Espinal-Enríquez, D. A. Priego-Espinosa, A. Darszon, C. Beltrán, and G. Martínez-Mekler. Network model predicts that catsper is the main ca 2+ channel in the regulation of sea urchin sperm motility. *Scientific Reports*, 7(1):4236, 2017. 12, 19, 21, 79

[38] M. Esteller. CpG island hypermethylation and tumor suppressor genes: A booming present, a brighter future. *Oncogene*, 21(35):5427–5440, Aug. 2002. 2

[39] L. Fagerberg, B. M. Hallström, P. Oksvold, C. Kampf, and D. Djureinovic. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular & Cellular Proteomics*, 13(2):397–406, 2014. 33

[40] A. P. Feinberg and B. Tycko. The history of cancer epigenetics. *Nature Reviews Cancer*, 4(2):143–153, Feb. 2004. 6

[41] C. Frantz, K. M. Stewart, and V. M. Weaver. The extracellular matrix at a glance. *Journal of cell science*, 123(24):4195–4200, 2010. 71

[42] R. C. Friedman, K. K.-H. Farh, C. B. Burge, and D. P. Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19(1):92–105, Oct. 2008. 7

[43] A. Fröhlich, S. Loick, and Bawden. Comprehensive analysis of tumor necrosis factor receptor tnfrsf9 (4-1bb) dna methylation with regard to molecular and clinicopathological features, immune infiltrates, and response prediction to immunotherapy in melanoma. *EBioMedicine*, 52, 2020. 72

[44] D. García-Cortés, G. de Anda-Jáuregui, C. Fresno, E. Hernandez-Lemus, and J. Espinal-Enriquez. Gene co-expression is distance-dependent in breast cancer. *Frontiers in Oncology*, 10, 2020. 11, 12, 19, 21, 38, 55, 61, 75, 79

[45] D. García-Cortés, E. Hernández-Lemus, and J. Espinal-Enríquez. Luminal a breast cancer co-expression network: Structural and functional alterations. *Frontiers in Genetics*, 12, apr 2021. 22, 55, 61

[46] D. Garcia-Cortes, E. Hernandez-Lemus, and J. Espinal-Enriquez. Loss of long-range co-expression is a common trait in cancer. *bioRxiv*, pages 2022–10, 2022. 79

[47] S. Garza-Manero, I. Pichardo-Casas, C. Arias, L. Vaca, and A. Zepeda. Selective distribution and dynamic modulation of miRNAs in the synapse and its possible role in Alzheimer's Disease. *Brain Research*, 1584:80–93, Oct. 2014. 4

[48] J. Goldenring and K. Nam. Rab25 as a tumour suppressor in colon carcinogenesis. *British journal of cancer*, 104(1):33–36, 2011. 69

[49] A. González-Espinoza, J. Zamora-Fuentes, E. Hernández-Lemus, and J. Espinal-Enríquez. Gene co-expression in breast cancer: A matter of distance. *Frontiers in Oncology*, 11, Nov. 2021. 61, 80

[50] P. D. Gopal Krishnan, E. Golden, and E. A. Woodward. Rab gtpases: emerging oncogenes and tumor suppressive regulators for the editing of survival pathways in cancer. *Cancers*, 12(2):259, 2020. 73

[51] Y. Gu, Y. M. Zou, D. Lei, Y. Huang, and W. Li. Promoter dna methylation analysis reveals a novel diagnostic cpg-based biomarker and rab25 hypermethylation in clear cell renel cell carcinoma. *Scientific reports*, 7(1):1–11, 2017. 71

[52] N. B. Haas and K. L. Nathanson. Hereditary kidney cancer syndromes. *Advances in Chronic Kidney Disease*, 21(1):81–90, Jan. 2014. 12

[53] V. Haberle and A. Stark. Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews Molecular Cell Biology*, 19(10):621–637, Oct. 2018. 2, 3

[54] D. Hanahan. Hallmarks of cancer: new dimensions. *Cancer discovery*, 12(1):31–46, 2022. 7, 73

[55] D. Hanahan and R. A. Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011. 6

[56] Y. Hashimoto, Y. Akiyama, and Y. Yuasa. Multiple-to-multiple relationships between MicroRNAs and target genes in gastric cancer. *PLoS ONE*, 8(5):e62589, May 2013. 5

[57] M. R. Hassler and G. Egger. Epigenomics of cancer – emerging new concepts. *Biochimie*, 94(11):2219–2230, Nov. 2012. 13

[58] R. D. Hawkins, G. C. Hon, and B. Ren. Next-generation genomics: An integrative approach. *Nature Reviews Genetics*, 11(7):476–486, July 2010. 8, 9

[59] X. He, W. Li, X. Liang, X. Zhu, L. Zhang, Y. Huang, T. Yu, S. Li, and Z. Chen. IGF2bp2 overexpression indicates poor survival in patients with acute myelocytic leukemia. *Cellular Physiology and Biochemistry*, 51(4):1945–1956, 2018. 59

[60] S. R. Head, H. K. Komori, S. A. LaMere, and Whisenant. Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques*, 56(2):61–77, Feb. 2014. 10

[61] C. Hernández-Gómez, E. Hernández-Lemus, and J. Espinal-Enríquez. The role of copy number variants in gene co-expression patterns for luminal b breast tumors. *Frontiers in Genetics*, 13, Apr. 2022. 61

[62] E. Hernández-Lemus, H. Reyes-Gopar, J. Espinal-Enríquez, and S. Ochoa. The many faces of gene regulation in cancer: A computational oncogenomics outlook. *Genes*, 10(11):865, Oct. 2019. 76

[63] K. A. Hoadley, C. Yau, T. Hinoue, and D. M. W. et al. Cell-of-origin patterns dominate the molecular classification of 10, 000 tumors from 33 types of cancer. *Cell*, 173(2):291–304.e6, Apr. 2018. 76

[64] J. Hsieh. Orchestrating transcriptional control of adult neurogenesis. *Genes & Development*, 26(10):1010–1021, 2012. 2

[65] J. J. Hsieh, M. P. Purdue, S. Signoretti, C. Swanton, and L. Albiges. Renal cell carcinoma. *Nature Reviews Disease Primers*, 3(1), Mar. 2017. 13

[66] T. Hu, N. Chitnis, D. Monos, and A. Dinh. Next-generation sequencing technologies: An overview. *Human Immunology*, 82(11):801–811, Nov. 2021. 3, 10

[67] L. Ibrahim, D. Aladle, A. Mansour, A. Hammad, A. A. A. Wakeel, and S. A. A. El-Hameed. Expression and prognostic significance of livin/BIRC7 in childhood acute lymphoblastic leukemia. *Medical Oncology*, 31(5), apr 2014. 57

[68] F. Jiao, H. Sun, Q. Yang, H. Sun, and Z. Wang. Association of cxcl13 and immune cell infiltration signature in clear cell renal cell carcinoma. *Int J Med Sci*, 17:1610–1624, 2020. 33

[69] E. Jonasch, C. L. Walker, and W. K. Rathmell. Clear cell renal cell carcinoma ontogeny and mechanisms of lethality. *Nature Reviews Nephrology*, 17(4):245–261, Nov. 2020. 13

[70] P. Jun, C. Hong, A. Lal, J. M. Wong, and M. W. McDermott. Epigenetic silencing of the kinase tumor suppressor wnk2 is tumor-type and tumor-grade specific. *Neuro-oncology*, 11(4):414–422, 2009. 58

[71] J. Y. Kang and K. E. Kim. Prognostic value of interleukin-32 expression and its correlation with the infiltration of natural killer cells in cutaneous melanoma. *Journal of Clinical Medicine*, 10(20):4691, 2021. 72

[72] M. G. Kazanietz, M. Durando, and M. Cooke. Cxcl13 and its receptor cxcr5 in cancer: inflammation, immune response, and beyond. *Frontiers in endocrinology*, 10:471, 2019. 73

[73] M. I. Khan, S. M. Nur, and W. H. Abdulaal. A study on dna methylation modifying natural compounds identified egcg for induction of ifi16 gene expression related to the innate immune response in cancer cells. *Oncology Letters*, 24(1):1–10, 2022. 72

[74] W. Kong, L. He, E. J. Richards, S. Challa, C.-X. Xu, and J. Permuth-Wey. Upregulation of miRNA-155 promotes tumour angiogenesis by targeting VHL and is associated with poor prognosis and triple-negative breast cancer. *Oncogene*, 33(6):679–689, jan 2013. 51

[75] P. K. Kreeger and D. A. Lauffenburger. Cancer systems biology: a network modeling perspective. *Carcinogenesis*, 31(1):2–8, Jan. 2010. 2, 11

[76] L. A. Lapierre, C. M. Caldwell, J. N. Higginbotham, and Avant. Transformation of rat intestinal epithelial cells by overexpression of rab25 is microtubule dependent. *Cytoskeleton*, 68(2):97–111, 2011. 69

[77] A. Laugesen, J. W. Højfeldt, and K. Helin. Molecular mechanisms directing PRC2 recruitment and h3k27 methylation. *Molecular Cell*, 74(1):8–18, Apr. 2019. 6

[78] H.-J. Lee, Z. L. Liang, S. M. Huang, J.-S. Lim, D.-Y. Yoon, H.-J. Lee, and J. M. Kim. Overexpression of il-32 is a novel prognostic factor in patients with localized clear cell renal cell carcinoma. *Oncology Letters*, 3(2):490–496, 2012. 72

[79] B. Li, Q. Huang, and G.-H. Wei. The role of HOX transcription factors in cancer predisposition and progression. *Cancers*, 11(4):528, apr 2019. 54

[80] M. Li, C. Marin-Muller, U. Bharadwaj, K.-H. Chow, Q. Yao, and C. Chen. MicroRNAs: Control and loss of control in human physiology and disease. *World Journal of Surgery*, 33(4):667–684, Nov. 2008. 7

[81] M. LI, Y. WANG, Y. SONG, R. BU, B. YIN, X. FEI, Q. GUO, and B. WU. MicroRNAs in renal cell carcinoma: A systematic review of clinical implications (review). *Oncology Reports*, 33(4):1571–1578, Feb. 2015. 62

[82] T. Li, J. Fu, Z. Zeng, D. Cohen, and J. Li. Timer2. 0 for analysis of tumor-infiltrating immune cells. *Nucleic acids research*, 48(W1):W509–W514, 2020. 7, 73

[83] X. Li, Y. Li, and H. Lu. [ARTICLE WITHDRAWN] miR-1193 suppresses proliferation and invasion of human breast cancer cells through directly targeting IGF2bp2. *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics*, 25(4):579–585, apr 2017. 59

[84] Y. Li, B. Guan, J. Liu, Z. Zhang, and S. He. MicroRNA-200b is downregulated and suppresses metastasis by targeting LAMA4 in renal cell carcinoma. *EBioMedicine*, 44:439–451, jun 2019. 58

[85] Y. Li, Q. Jia, Q. Zhang, and Y. Wan. Rab25 upregulation correlates with the proliferation, migration, and invasion of renal cell carcinoma. *Biochemical and biophysical research communications*, 458(4):745–750, 2015. 69

[86] Y. Li, Z. Wang, W. Jiang, and Zeng. Tumor-infiltrating tnfrsf9+ cd8+ t cells define different subsets of clear cell renal cell carcinoma with prognosis and immunotherapeutic response. *Oncoimmunology*, 9(1):1838141, 2020. 72

[87] X.-D. Liu, W. Kong, C. B. Peterson, D. J. McGrail, and A. Hoang. PBRM1 loss defines a nonimmunogenic tumor phenotype associated with checkpoint inhibitor resistance in renal carcinoma. *Nature Communications*, 11(1), May 2020. 7, 57

[88] Y. Liu, J. Sun, and M. Zhao. Ongene: a literature-based database for human oncogenes. *Journal of Genetics and Genomics*, 44(2):119–121, 2017. 29

[89] Y. Liu, X. Wang, L. Deng, L. Ping, and Shi. Itk inhibition induced in vitro and in vivo anti-tumor activity through downregulating tcr signaling pathway in malignant t cell lymphoma. *Cancer Cell International*, 19:1–19, 2019. 69

[90] Z. Liu, Y. Wan, M. Yang, and Qi. Identification of methylation-driven genes related to the prognosis of papillary renal cell carcinoma: a study based on the cancer genome atlas. *Cancer Cell International*, 20(1):1–12, 2020. 72

[91] S. Loibl, P. Poortmans, M. Morrow, C. Denkert, and G. Curigliano. Breast cancer. *The Lancet*, 397(10286):1750–1769, May 2021. 6

[92] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), Dec. 2014. 20, 24, 28

[93] T. Luo, X. Chen, S. Zeng, B. Guan, B. Hu, and Y. Meng. Bioinformatic identification of key genes and analysis of prognostic values in clear cell renal cell carcinoma. *Oncology letters*, 16(2):1747–1757, 2018. 33

[94] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera, and A. Califano. ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(S1), Mar. 2006. 11, 21, 25, 101, 102, 103

[95] F. Mertens, B. Johansson, M. Höglund, and F. Mitelman. Chromosomal Imbalance Maps of Malignant Solid Tumors: A Cytogenetic Survey of 3185 Neoplasms1. *Cancer Research*, 57(13):2765–2780, 07 1997. 6

[96] S. E. Meyer, D. E. Muench, A. M. Rogers, T. J. Newkold, and E. Orr. miR-196b target screen reveals mechanisms maintaining leukemia stemness with therapeutic potential. *Journal of Experimental Medicine*, 215(8):2115–2136, jul 2018. 54

[97] S. Mitra, K. W. Cheng, and G. B. Mills. Rab25 in cancer: a brief update. *Biochemical Society Transactions*, 40(6):1404–1408, 2012. 71

[98] S. Moniz, F. Veríssimo, P. Matos, R. Brazão, and E. Silva. Protein kinase WNK2 inhibits cell proliferation by negatively modulating the activation of MEK1/ERK1/2. *Oncogene*, 26(41):6071–6081, jul 2007. 58

[99] M. R. Morris and F. Latif. The epigenetic landscape of renal cancer. *Nature Reviews Nephrology*, 13(1):47–60, 2017. 5, 13, 73, 79

[100] K. T. Nam, H.-J. Lee, J. J. Smith, and Lapierre. Loss of rab25 promotes the development of intestinal neoplasia in mice and is associated with human colorectal adenocarcinomas. *The Journal of clinical investigation*, 120(3):840–849, 2010. 71

[101] C. S. Neal, M. Z. Michael, L. H. Rawlings, M. B. V. der Hoek, and J. M. Gleadle. The VHL-dependent regulation of microRNAs in renal cancer. *BMC Medicine*, 8(1), oct 2010. 51

[102] M. j. Nueda, A. Ferrer, and A. Conesa. Arsyn: a method for the identification and removal of systematic noise in multifactorial time course microarray experiments. *Biostatistics*, 13(3):553–566, 2012. 20, 24, 28

[103] J. O'Brien, H. Hayder, Y. Zayed, and C. Peng. Overview of MicroRNA biogenesis, mechanisms of actions, and circulation. *Frontiers in Endocrinology*, 9, Aug. 2018. 4

[104] P. Olson, J. Lu, H. Zhang, A. Shai, M. G. Chun, Y. Wang, S. K. Libutti, E. K. Nakakura, T. R. Golub, and D. Hanahan. MicroRNA dynamics in the stages of tumorigenesis correlate with hallmark capabilities of cancer. *Genes & Development*, 23(18):2152–2165, Sept. 2009. 7

[105] B. Pan, M. Yang, X. Wei, W. Li, and Wang. Interleukin-2 inducible t-cell kinase: a potential prognostic biomarker and tumor microenvironment remodeling indicator for hepatocellular carcinoma. *Aging (Albany NY)*, 13(14):18620, 2021. 69

[106] N. Passon, E. Bregant, M. Sponziello, and M. Dima. Somatic amplifications and deletions in genome of papillary thyroid carcinomas. *Endocrine*, 50(2):453–464, apr 2015. 58

[107] S. R. Payne and C. J. Kemp. Tumor suppressor genetics. *Carcinogenesis*, 26(12):2031–2045, Dec. 2005. 7

[108] M. W. Pickup, J. K. Mouw, and V. M. Weaver. The extracellular matrix modulates the hallmarks of cancer. *EMBO reports*, 15(12):1243–1253, 2014. 71

[109] Protein Atlas. Gene saa2-saa, Accesed 2022. https://www.proteinatlas.org/ENSG00000255071-SAA2-SAA4/pathology. 33

[110] Protein Atlas. Gene slc6a19, Accesed 2022. https://www.proteinatlas.org/ENSG00000174358-SLC6A19/pathology. 33

[111] J. Pu, J. Wang, Z. Qin, A. Wang, and Y. Zhang. IGF2bp2 promotes liver cancer growth through an m6a-FEN1-dependent mechanism. *Frontiers in Oncology*, 10, nov 2020. 59

[112] U. Raudvere, L. Kolberg, I. Kuzmin, T. Arak, P. Adler, H. Peterson, and J. Vilo. g:profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47(W1):W191–W198, May 2019. 22

[113] V. P. S. Rawat, M. Götze, A. Rasalkar, N. M. Vegi, and S. Ihme. The microRNA miR-196b acts as a tumor suppressor in cdx2-driven acute myeloid leukemia. *Haematologica*, 105(6):e285–e289, sep 2019. 54

[114] F. Recillas-Targa. Cancer Epigenetics: An Overview. *Archives of Medical Research*, 53(8):732–740, 2022. 3, 5, 13

[115] C. Reily, T. J. Stewart, M. B. Renfrow, and J. Novak. Glycosylation in health and disease. *Nature Reviews Nephrology*, 15(6):346–366, Mar. 2019. 58

[116] J. Reimand, T. Arak, P. Adler, L. Kolberg, S. Reisberg, H. Peterson, and J. Vilo. g:profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Research*, 44(W1):W83–W89, Apr. 2016. 22

[117] D. B. Rigato, P. C. Branco, and C. S. Mateus Reis Silva. Birc7 (baculoviral iap repeat containing 7). *Atlas of Genetics and Cytogenetics in Oncology and Haematology*, 2020. 57

[118] T. J. Rintala, A. Ghosh, and V. Fortino. Network approaches for modeling the effect of drugs and diseases. *Briefings in Bioinformatics*, 23(4):bbac229, July 2022. 6

[119] D. Risso, K. Schwartz, G. Sherlock, and S. Dudoit. Gc-content normalization for rna-seq data. *BMC bioinformatics*, 12(1):1–17, 2011. 20, 28

[120] K. D. Robertson. DNA methylation and human disease. *Nature Reviews Genetics*, 6(8):597–610, Aug. 2005. 5, 6

[121] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, Nov. 2009. 20

[122] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008. 22

[123] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, Jan. 2008. 46

[124] I. Sagiv-Barfi, H. E. Kohrt, and D. K. Czerwinski. Therapeutic antitumor immunity by checkpoint blockade is enhanced by ibrutinib, an inhibitor of both btk and itk. *Proceedings of the National Academy of Sciences*, 112(9):E966–E972, 2015. 73

[125] P. E. Saw, X. Xu, J. Chen, and E.-W. Song. Non-coding RNAs: The new central dogma of cancer biology. *Science China Life Sciences*, 64(1):22–50, Jan. 2021. 1

[126] J. M. Sayagués and L. A. Corchete. Genomic characterization of liver metastases from colorectal cancer patients. *Oncotarget*, 7(45):72908–72922, Sept. 2016. 33

[127] H.-J. Schulten, D. Hussein, F. Al-Adwani, and S. Karim. Microarray expression profiling identifies genes, including cytokines, and biofunctions, as diapedesis, associated with a brain metastasis from a papillary thyroid carcinoma. *American Journal of Cancer Research*, 6(10):2140–2161, 2016. 33

[128] SEER. Cancer stat facts: Kidney and renal pelvis cancer. https://seer.cancer.gov/statfacts/html/kidrp.html, accessed January, 2023). 12

[129] E. A. Serrano-Carbajal, J. Espinal-Enríquez, and E. Hernández-Lemus. Targeting metabolic deregulation landscapes in breast cancer subtypes. *Frontiers in Oncology*, 10, Feb. 2020. 19

[130] N. Shah and S. Sukumar. The hox genes and their roles in oncogenesis. *Nature Reviews Cancer*, 10(5):361–371, apr 2010. 54

[131] R. Shang, S. Lee, G. Senavirathne, and E. C. Lai. microRNAs in action: biogenesis, function and regulation. *Nature Reviews Genetics*, June 2023. 4

[132] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, and J. T. Wang. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003. 21, 30

[133] N. Shao, H.-K. Wang, Y. Zhu, and D.-W. Ye. Modification of American Joint Committee on cancer prognostic groups for renal cell carcinoma. *Cancer Medicine*, 7(11):5431–5438, Nov. 2018. 8

[134] R. sheng Huang, Y. liang Zheng, C. Li, C. Ding, C. Xu, and J. Zhao. MicroRNA-485-5p suppresses growth and metastasis in non-small cell lung cancer cells by targeting IGF2bp2. *Life Sciences*, 199:104–111, apr 2018. 59

[135] D. Smedley, S. Haider, S. Durinck, L. Pandini, P. Provero, J. Allen, and O. Arnaiz. The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Research*, 43(W1):W589–W598, Apr. 2015. 28

[136] R. Stark, M. Grzelak, and J. Hadfield. RNA sequencing: The teenage years. *Nature Reviews Genetics*, jul 2019. 9

[137] L. Z. Strichman-Almashanu, R. S. Lee, P. O. Onyango, and E. Perlman. A genome-wide screen for normally methylated human CpG islands that can identify novel imprinted genes. *Genome Research*, 12(4):543–554, Mar. 2002. 6

[138] U. Swami, R. H. Nussenzveig, B. Haaland, and N. Agarwal. Revisiting AJCC TNM staging for renal cell carcinoma: Quest for improvement. *Annals of Translational Medicine*, 7(S1):S18–S18, Mar. 2019. 8

[139] S. Tarazona, F. García, A. Ferrer, J. Dopazo, and A. Conesa. NOIseq: a RNA-seq differential expression method robust for sequencing depth biases. *EMBnet.journal*, 17(B):18, Feb. 2012. 20

[140] M. Tong, K. W. Chan, J. Y. Bao, K. Y. Wong, and J.-N. Chen. Rab25 is a tumor suppressor gene with antiangiogenic and anti-invasive activities in esophageal squamous cell carcinomarab25 in esophageal cancer. *Cancer research*, 72(22):6024–6035, 2012. 69, 71

[141] N. S. Vasudev, P. J. Selby, and R. E. Banks. Renal cancer biomarkers: the promise of personalized care. *BMC medicine*, 10(1):1–10, 2012. 72

[142] D. Vipin, L. Wang, G. Devailly, T. Michoel, and A. Joshi. Causal transcription regulatory network inference using enhancer activity as a causal anchor. *International journal of molecular sciences*, 19(11):3609, 2018. 71

[143] N. Wagener, I. Crnković-Mertens, C. Vetter, S. Macher-Göppinger, and J. Bedke. Expression of inhibitor of apoptosis protein livin in renal cell carcinoma and non-tumorous adult kidney. *British Journal of Cancer*, 97(9):1271–1276, oct 2007. 57

[144] A. D. Waldman, J. M. Fritz, and M. J. Lenardo. A guide to cancer immunotherapy: from t cell basic science to clinical practice. *Nature Reviews Immunology*, 20(11):651–668, 2020. 73

[145] J. Wang, L. Chen, and P. Qiang. The role of IGF2bp2, an m6a reader gene, in human metabolic diseases and cancers. *Cancer Cell International*, 21(1), feb 2021. 59

[146] M. D. Wang, M. J. Schnitzer, H. Yin, R. Landick, J. Gelles, and S. M. Block. Force and velocity measured for single molecules of RNA polymerase. *Science*, 282(5390):902–907, Oct. 1998. 1

[147] S. Wang, C. Hu, F. Wu, and S. He. Rab25 gtpase: Functional roles in cancer. *Oncotarget*, 8(38):64591, 2017. 71

[148] T. Wang, R. Lu, P. Kapur, B. S. Jaiswal, R. Hannan, and Z. Zhang. An empirical approach leveraging tumorgrafts to dissect the tumor microenvironment in renal cell carcinoma identifies missing link to prognostic inflammatory factors. *Cancer Discovery*, 8(9):1142–1155, Sept. 2018. 7

[149] S. Weeks, R. Harris, and M. Karimi. Targeting ITK signaling for T cell-mediated diseases. *iScience*, 24(8):102842, Aug. 2021. 69, 71

[150] U. H. Weidle and A. Nopora. Clear Cell Renal Carcinoma: MicroRNAs With Efficacy in Preclinical *In Vivo* Models. *Cancer Genomics - Proteomics*, 18(3 Suppl):349–368, 2021. 5, 13

[151] J.-D. Wen, L. Lancaster, C. Hodges, A.-C. Zeri, S. H. Yoshimura, H. F. Noller, C. Bustamante, and I. Tinoco. Following translation by single ribosomes one codon at a time. *Nature*, 452(7187):598–603, Mar. 2008. 1

[152] J. Winkler, A. Abisoye-Ogunniyan, K. J. Metcalf, and Z. Werb. Concepts of extracellular matrix remodelling in tumour progression and metastasis. *Nature communications*, 11(1):5120, 2020. 71

[153] F. Xu, F. Zhu, W. Wang, W. Gao, X. Chen, and C. Yu. Down-regulation of miRNA-196b expression inhibits the proliferation, migration and invasiveness of HepG2 cells while promoting their apoptosis via the PI3k/akt signaling pathway. *Cellular and Molecular Biology*, 66(3):159–164, jun 2020. 54

[154] M. Xu, M. Gu, K. Zhang, J. Zhou, Z. Wang, and J. Da. miR-203 inhibition of renal cancer cell proliferation, migration and invasion by targeting of FGF2. *Diagnostic Pathology*, 10(1), Apr. 2015. 7

[155] S. Ye, W. Song, X. Xu, X. Zhao, and L. Yang. IGF2bp2 promotes colorectal cancer cell proliferation and survival through interfering with RAF-1 degradation by mir-195. *FEBS Letters*, 590(11):1641–1650, may 2016. 59

[156] W.-S. Yong, F.-M. Hsu, and P.-Y. Chen. Profiling genome-wide DNA methylation. *Epigenetics & Chromatin*, 9(1):26, Dec. 2016. 5, 11

[157] K. Yoshihara, M. Shahmoradgoli, E. Martínez, R. Vegesna, and H. Kim. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature Communications*, 4(1):2612, Dec. 2013. 73

[158] J. M. Zamora-Fuentes, E. Hernández-Lemus, and J. Espinal-Enríquez. Gene expression and co-expression networks are strongly altered through stages in clear cell renal carcinoma. *Frontiers in Genetics*, 11, Nov. 2020. 20, 30, 40, 46, 48, 55, 61, 73, 79

[159] J. M. Zamora-Fuentes, E. Hernández-Lemus, and J. Espinal-Enríquez. Oncogenic role of mir-217 during clear cell renal carcinoma progression. *Frontiers in oncology*, 12:934711–934711, 2022. 49, 51, 54, 55, 57, 73, 79

[160] J. M. Zamora-Fuentes, E. Hernández-Lemus, and J. Espinal-Enríquez. Methylation-related genes involved in renal carcinoma progression. *Frontiers in Genetics*, 14, 2023. 29, 30, 62, 66, 71

[161] J. Zhang, S. Li, F. Liu, and K. Yang. Role of cd68 in tumor immunity and prognosis prediction in pan-cancer. *Scientific Reports*, 12(1):7844, 2022. 72

[162] Z. Zhang, E. Lin, H. Zhuang, L. Xie, X. Feng, J. Liu, and Y. Yu. Construction of a novel gene-based model for prognosis prediction of clear cell renal cell carcinoma. *Cancer cell international*, 20(1):1–18, 2020. 33

[163] M. Zhao, P. Kim, R. Mitra, J. Zhao, and Z. Zhao. Tsgene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. *Nucleic acids research*, 44(D1):D1023–D1031, 2016. 29

[164] B. Zheng, J. Qu, K. Ohuchida, H. Feng, and S. J. F. Chong. LAMA4 upregulation is associated with high liver metastasis potential and poor survival outcome of pancreatic cancer. *Theranostics*, 10(22):10274–10289, 2020. 58

[165] W. Zhou, P. W. Laird, and H. Shen. Comprehensive characterization, annotation and innovative use of infinium DNA methylation BeadChip probes. *Nucleic Acids Research*, page gkw967, Oct. 2016. 26

Apéndice A

Apendices

A.1. Abreviaturas

- **CRcc** = Carcinoma renal de células claras.
- **CCR** = Carcinoma de células renales.
- **NT** = Tejido normal.
- **ADN** = Ácido desoxirribonucleico
- **ARNm** = Ácido ribonucleico mensajero
- **pb** = pares de base
- **microARN** ó **miRNA** ó **miR**= Segmentos cortos de ácido ribonucleico.
- **ANS** = Aprendizaje no supervisado
- **GDE** = Genes diferencialmente expresados
- **mDE** = miRNAs diferencialmente expresados
- **IM** = Información mutua
- **CGIs** = Islas CpG
- **CpG** = Combinación de un dinucleotido citosina-guanina unido por un fosfato
- **OG** = Oncogenes
- **TSG** = Genes supresores tumorales
- **BG** = Genes con funciones oncogénicas y supresoras de tumores
- **ADNc** = ADN complementario

- **TCGA** = The Cancer Genome Atlas
- **GDC** =Genomic Data Commons
- **mRNA-Seq** ó **RNAseq** = Secuenciación de ARN mensajero
- **smRNA-Seq** = Secuenciación de ARN pequeño
- **DE** = Expresión diferencial
- **FDR** = False Discovery Rate
- **LFC** = Log Fold Change
- **CpGs-DM** = Sitios CpG diferencialmente metilados
- **GO** = Gene ontolology
- **GDM** = Gen diferencialmente metilado
- **MEC** = Matriz extracelular
- **EScc** = Carcinoma de células escamosas de esófago
- **lncARN** = Largos no codificantes de ácido ribonucleíco
- **CHIP-seq** = Secuenciación combinada con inmunoprecipitación de la cromatina
- **ATAC-seq** = Secuenciación combinada con acceso de la transposas a la cromatina
- **HI-C** = Experimento que captura la conformación de la cromatina

A.2. Mutaciones en CRcc

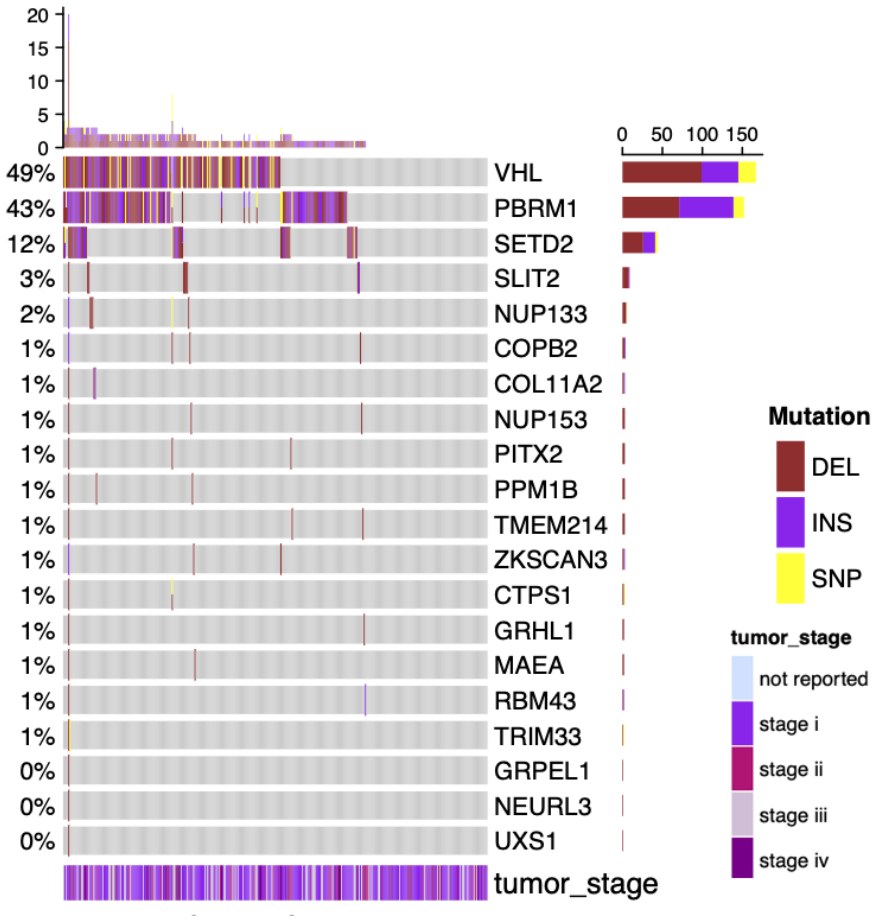

Figura A.1: *Porcentage de mutaciones en CRcc.* Esta gráfica muestra las mutaciones (deleciones, inserciones y alteraciones de un sólo nucleótido) para cada una de las muestras etiquetadas por etapa de progresión del cáncer. Los genes más mutados en la mayor cantidad de muestras son: *VHL*, *PBRM1* y *SETD2*

A.3. Contexto del cálculo de coexpresión genética

Es debido mencionar que la información integra de esta sección es tomada directamente de [94]. Se hacen adaptaciones a partir de la traducción y las condiciones de este trabajo.

Primeramente, tenemos que considerar la distribución de probabilidad conjunta

(JPD) de cada gen ($P(\{g_i\}), i = 1, ..., N,$) en la matriz de expresión. Esta probabilidad se puede ver como:

$$P(\{g_i\}) = \frac{1}{Z} exp[-\sum_{i}^{N} \phi_i(g_i) - \sum_{i,j}^{N} \phi_{ij}(g_i, g_j) - \sum_{i,j}^{N} \phi_{ijk}(g_i, g_j, g_k) - ...] \equiv e^{-H(\{g_i\})} \tag{A.1}$$

donde N es el número de genes, Z es un factor de normalización, también llamado función de partición. Las funciones ϕ son potenciales y $H(\{g_i\})$ es el *hamiltoniano* que define las estadísticas del sistema.

El modelo más simple es aquel en el que se supone que los genes son independientes, es decir, $H(\{g_i\}) = \sum \phi(g_i)$, de modo que los potenciales de primer orden pueden evaluarse a partir de las probabilidades marginales, $P(g_i)$, que se estiman a partir de experimentos y observaciones. Si truncamos la ecuación A.1 hasta la función potencial de segundo grado, obtenemos un nivel de interacciones por pares, descrito por la ecuación A.2.

$$H(\{g_i\}) = \sum_{i} \phi_i(g_i) + \sum_{ij} \phi_{ij}(g_i, g_j) \tag{A.2}$$

Dentro de esta aproximación, todos los genes para los que $\phi_{ij} = 0$ se declaran mutuamente no interactivos. Esto incluye genes que son estadísticamente independientes (es decir, $P(g_i, g_j) \approx P(g_i)P(g_j)$), así como genes que no interactúan directamente pero que son estadísticamente dependientes debido a su interacción a través de otros genes (es decir, $P(g_i, g_j) \neq P(g_i)P(g_j)$, pero $\phi_{ij} = 0$). Notamos que $P(g_i, g_j) = P(g_i)P(g_j)$ no es una condición suficiente para $\phi_{ij} = 0$. Es decir, los potenciales pueden ser cero y dos genes tener alta correlación ó ser dependientes.

Por lo tanto, identificamos las interacciones candidatas estimando la información mutua del perfil de expresión génica por pares, $I(g_i, g_j) \equiv I_{ij}$, una medida teórica de la información de la relación que es cero si y sólo si $P(g_i, g_j) = P(g_i)P(g_j)$. Es decir, la expresión de los genes es indipendiente. Luego filtramos los MI usando un umbral apropiado, I_0, calculado para un valor p específico, p_0, donde la hipótesis nula indica que dos genes son independientes. Una de las limitantes de este método surge cuando los genes separados por uno o más intermediarios (relaciones indirectas) pueden estar altamente co-regulados sin que ello implique una interacción *irreductible*, lo que da como resultado numerosos falsos positivos.

Por lo tanto, en su segundo paso, ARACNE elimina la gran mayoría de las interacciones candidatas indirectas ($\phi_{ij} = 0$) utilizando una propiedad teórica de la información bien conocida, la desigualdad en el procesamiento de datos (*DPI*, discutida en [94]).

A.3.1. Estimación de Información Mutua

La cantidad de *información mutua* (IM) para un par de variables aleatorias, X y Y , se define como $I(x, y) = S(x) + S(y) - S(x, y)$, donde $S(t)$ es la entropía de una

variable arbitraria t. Para una variable discreta, la entropía es: $S(t) = -\langle \log p(t_i) \rangle = \sum_i p(t_i) \log p(t_i)$, donde $p(t_i) = Prob(t = ti)$ y esta es la probabilidad de cada estado (valor) discreto de la variable. Para las variables continuas, la entropía es infinita, pero el valor de IM permanece bien definido y se puede calcular reemplazando $S(x)$ con la entropía diferencial, que promedia la densidad logarítmica de probabilidad en lugar de la masa logarítmica. Al igual que la correlación de Pearson más familiar, el valor de IM mide el grado de dependencia estadística entre dos variables. Sin embargo, mientras que los coeficientes de correlación no son invariantes bajo reparametrizaciones y pueden ser cero incluso para variables manifiestamente dependientes. IM es invariante de reparametrización y no es cero *si y sólo si* existe algún tipo de dependencia estadística.

Para estimar la IM usamos un estimador computacionalmente eficiente llamado: *estimador de kernel gaussiano*. Dado un conjunto de medidas de dos dimensiones $\overrightarrow{z_i} \equiv \{x_i, y_i\}, i = 1...M$, entonces la cantidad de IM se puede calcular como:

$$I(\{x_i\}, \{y_i\}) = \frac{1}{M} \sum_i \log \frac{f(x_i, y_i)}{f(x_i) f(y_i)} \tag{A.3}$$

Dado que MI es reparametrización invariante, se hace una copula-transformación (es decir, *rank-order*) x e y para la estimación de MI; el rango de estas variables transformadas está entre 0 y 1, y sus distribuciones marginales de probabilidad son manifiestamente uniformes. Esto disminuye la influencia de las transformaciones arbitrarias involucradas en el preprocesamiento de datos de *microarreglos* y elimina la necesidad de considerar anchos de kernel dependientes de la posición, h, que podrían ser preferibles para datos distribuidos de manera no uniforme.

A.3.2. Costo computacional de ARACNe

Para una red de N genes la complejidad de ARACNE es $O(N^3 + N^2M^2)$, donde M es el número de muestras y N es el número de genes. El primer término se relaciona con el análisis DPI [94] y el segundo con la estimación de información mutua. Esto se compara favorablemente con los métodos de optimización que deben explorar un espacio de búsqueda exponencial. En la práctica, el DPI se aplica a un pequeño subconjunto de tripletes para los cuales los tres bordes sobreviven al umbral de información mutua. Por lo tanto, para una M grande, la parte computacionalmente intensiva generalmente se asocia con el segundo término (información mutua computacional), que escala como $O(N^2M^2)$. Como resultado, ARACNE puede analizar eficientemente redes con decenas de miles de genes.

Considerando que en este trabajo se aporta una poderosa adaptación de este algoritmo para máquinas paralelas con P procesadores. Entonces, el costo computacional se puede aproximar como: $\frac{N^2}{P}$. Esta ventaja es muy importante porque abre la posibilidad de calcular redes de IM con *miles* de muestras y *miles* de genes. Este tipo de cálculos puede llegar a tener sentido en novedosas tecnologías como *SCRNA-seq (single cell rna-seq)*.