UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

DOCTORADO EN CIENCIAS BIOMÉDICAS

INSTITUTO DE ECOLOGÍA

EVALUACIÓN DE LA REGULACIÓN TRANS EN CÁNCER: UN ENFOQUE DE BIOLOGÍA DE SISTEMAS

TESIS

QUE PARA OPTAR POR EL GRADO DE:

DOCTORA EN CIENCIAS

PRESENTA:

DIANA ELISA GARCÍA CORTÉS

TUTOR PRINCIPAL:

DR. JESÚS ESPINAL ENRÍQUEZ

INSTITUTO DE ECOLOGÍA

MIEMBROS DEL COMITÉ TUTOR:

DR. ALFREDO HIDALGO MIRANDA

FACULTAD DE MEDICINA

DR. GUSTAVO MARTÍNEZ MEKLER

INSTITUTO DE CIENCIAS FÍSICAS

CIUDAD UNIVERSITARIA, CIUDAD DE MÉXICO, ABRIL DE 2023

Agradecimientos

Agradezco a CONACYT por la beca de doctorado asignada a mi CVU 558985.

Agradezco al Programa de Doctorado en Ciencias Biomédicas de la Universidad Nacional Autónoma de México y a las personas que en él laboran, especialmente por los retos enfrentados durante la pandemia de COVID-19. A quienes se encargan de los asuntos directamente relacionados con la comunidad de estudiantes, agradezco por el acompañamiento y las facilidades otorgadas durante estos años. A los profesores de este programa agradezco 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 y a la Dirección de Enseñanza y Divulgación. Gracias por permitirme pertenecer a este instituto y por el apoyo que otorgan a sus estudiantes.

Índice general

Lista de abreviaciones

ADN Ácido Desoxirribonucleico
ARN Ácido Ribonucleico
IM Información Mutua
GO Gene Ontology
TCGA The Cancer Genome Atlas
GDC Genomic Data Commons
TAD Dominio Topológicamente Asociado
IDH Índice de Desarrollo Humano
OMS Organización Mundial de la Salud
ONU Organización de las Naciones Unidas
INMEGEN Instituto Nacional de Medicina Genómica

Índice de figuras

Índice de tablas

Resumen

Las células de cáncer muestran un conjunto de rasgos comunes y características habilitadoras que han sido compiladas y descritas en un marco conceptual llamado *Hallmarks of Cancer*. Estos rasgos aparecen en conjunto con alteraciones en los mecanismos regulatorios que controlan la transcripción genética. El análisis de perfiles de co-expresión y la construcción de redes de co-expresión genética permiten identificar conjuntos de genes que podrían compartir los mismos elementos regulatorios, posiblemente alterados, en su proceso de transcripción. En dichas redes, los nodos representan genes y las interacciones entre ellos simbolizan relaciones de alta co-expresión. La pérdida de co-expresión en cáncer de mama es el antecedente principal de esta tesis, en la que se investiga si dicho fenómeno prevalece en otros tipos de cáncer mediante el estudio de perfiles de co-expresión en quince tejidos diferentes de cáncer y sus contrapartes normales. En los perfiles de cáncer se encontró que las interacciones con valores de co-expresión más altos se dan entre genes dentro del mismo cromosoma y que la fuerza de las interacciones decae de acuerdo a la distancia entre estos genes, medida en pares de bases. Por lo tanto, hay una pérdida de co-expresión inter-cromosómica y un decaimiento en la co-expresión intra-cromosómica dependiente de la distancia física entre genes. Esto no ocurre en los tejidos del fenotipo normal. La estructura topológica de las redes se asoció con procesos funcionales mediante un análisis de comunidades y posterior análisis de enriquecimiento con Gene Ontology. Se encontró que en cáncer, estas comunidades están asociadas a procesos relacionados con la tumorigénesis, en particular con la activación de la respuesta inmune adaptativa. En cambio, en las redes de tejido normal, las comunidades se relacionan con procesos metabólicos y procesos asociados al mantenimiento celular. Las redes de múltiples tejidos y en ambas condiciones comparten un conjunto de genes que codifican para riboproteínas, lo que sugiere una alta relevancia de estos genes para la viabilidad celular. La pérdida de co-expresión a larga distancia no se observa en otras enfermedades crónicas como la Diabetes tipo 2 o Alzheimer. Los resultados de este proyecto sugieren que la pérdida de co-expresión a larga distancia es una característica común en cáncer.

1 Introducción

1.1. Cáncer como un problema de salud mundial

El cáncer es una de las principales causas de muerte a nivel mundial. Según la Organización Mundial de la Salud (OMS), en 112 de 183 países es la primera o segunda causa de muerte antes de los 70 años, mientras que en otros 23 países se encuentra en el tercer o cuarto lugar [1]. Su creciente importancia en la tasa de mortalidad está relacionada con la disminución de la mortalidad de enfermedades cardiovasculares y es parte de una transición epidemiológica que ocurre desde la segunda mitad del siglo pasado, en la que la preeminencia de las enfermedades infecciosas ha sido sustituida por enfermedades no transmisibles o enfermedades crónicas [2], al menos hasta antes de la pandemia de COVID-19.

Entre los principales factores de riesgo para el desarrollo de neoplasias tanto en hombres como en mujeres se encuentran: consumo de tabaco, consumo de alcohol, alto índice de masa corporal, malos hábitos alimenticios y contaminación del aire [3]. La distribución de éstos y otros factores de riesgo están asociados al desarrollo socioeconómico de los países e influyen en el grado de incidencia y mortalidad del cáncer [1]. Esto se ejemplifica en la Figura 1, donde el panel A muestra el Índice de Desarrollo Humano (IDH), reportado por la Organización de las Naciones Unidas (ONU), mientras que el panel B muestra la posición ocupada por el cáncer en términos de mortalidad prematura en cada país. Aquí puede observarse que países con IDH altos tienen en primer o segundo lugar como causa de mortalidad al cáncer

En el 2020 se estimaron alrededor de 19.3 millones de nuevos casos de cáncer y un total de 10 millones de muertes debido a esta enfermedad En términos de nuevos casos, los tipos de cáncer más común son: mama (2.26 millones de casos), pulmón (2.21 millones de casos) y cáncer colorrectal (1.93 millones de casos). Los tipos de cáncer que causaron mayor número de fallecimientos, fueron pulmón (1.8 millones), cáncer colorrectal (916 mil) y hepático (830 mil) [4].

La transición epidemiológica también se presenta en México. Sin embargo, la falta de un registro confiable sobre la incidencia y prevalencia del cáncer en nuestro país, debido en parte por un sistema de salud fragmentado, dificulta evaluar el alcance real de esta enfermedad y obstaculiza la creación de políticas públicas para el control del cáncer[5].

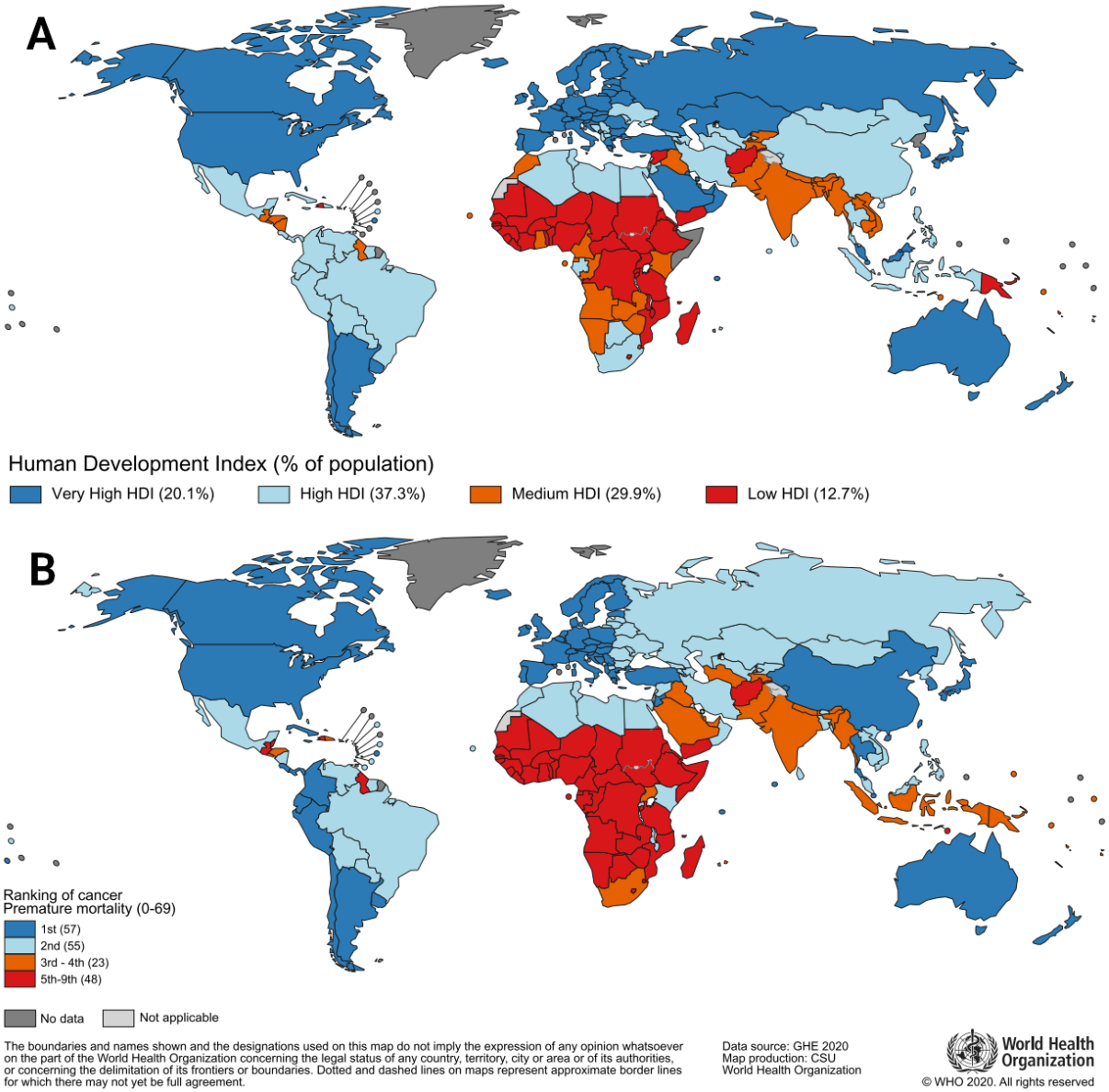

Figura 1: Índice de Desarrollo Humano (IDH) y mortalidad del cáncer. A) IDH de cada país reportado por la ONU. B) Lugar de mortalidad prematura ocupado por el cáncer en cada país, según datos de la OMS. Imagen modificada de [1]

Aunque los números absolutos varían según las fuentes, se estima que en el año 2013, alrededor de medio millón de personas en México vivían con cáncer y esta enfermedad fue causante de un 12.8 % de las muertes de dicho año después de las enfermedades cardiacas y la diabetes, con una tasa de mortalidad del 65.1 por 100 mil habitantes [6]. Proyecciones realizadas en el 2017 para las cifras del 2020 [5, 6], indican que la población de mexicanos con cáncer se elevaría a alrededor de 1 millón 200 mil habitantes, con una tasa de mortalidad de 79 por cada 100 mil habitantes.

En cuanto a los tipos de cáncer con mayores tasas de mortalidad en la población mexicana, se observa que a partir del 2013 y hasta las proyecciones del 2020, hay un aumento en la mortalidad de pacientes por cáncer de próstata, hígado, mama y cáncer colorrectal y un

descenso en las tasas de mortalidad de cáncer de pulmón, estómago y cáncer cérvico-uterino. Esto puede observarse en la figura 2.

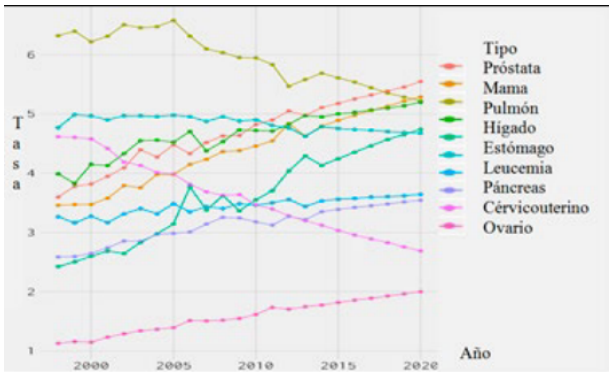

Figura 2: Tasas de mortalidad en los principales tipos de cáncer en México. Imagen de [5]

A pesar del incremento en la mortalidad y de que nuestro país tiene una alta prevalencia en los factores de riesgo asociados al cáncer, como consumo de alcohol, tabaco, además de altos índices de obesidad, México se ubica en puntos intermedios en cuanto a tasas de mortalidad en comparación con otros países en vías de desarrollo de la región [5]. Sin embargo, es indudable que el avance en la transición epidemiológica presentará retos importantes en todos los niveles del sector salud.

Hasta ahora, para afrontar dichos retos, las medidas implementadas se han enfocado en la prevención secundaria, intentando que la enfermedad sea detectada a tiempo para poder otorgar el tratamiento adecuado. Sin embargo, hay carencias en las políticas públicas de prevención primaria con un enfoque que permita reducir la incidencia o el desarrollo del cáncer, principalmente en términos de promoción de un estilo de vida saludable [7]. Además, aunque la implementación de un registro de cáncer de base poblacional se encuentra en proceso, la cobertura actual es baja, siendo éste un recurso indispensable para poder estimar indicadores poblacionales de incidencia, prevalencia y mortalidad, así como para poder evaluar la eficacia de las intervenciones de prevención [8].

Tanto el seguimiento de la carga del cáncer, como la ejecución de intervenciones de prevención y el reforzamiento de los sistemas de salud a nivel nacional y local, forman parte de los objetivos establecidos por la OMS en la Asamblea Mundial de la Salud en 2017 [4, 9]. Dentro de éstos también se encuentra el realizar estudios sobre los mecanismos de la carcinogénesis en el ser humano, que es uno de los intereses principales del área de Genómica Computacional del Instituto Nacional de Medicina Genómica (INMEGEN), donde fue desarrollada esta tesis.

1.2. Signos distintivos del cáncer

Cáncer es un término que engloba un amplio grupo de enfermedades que pueden afectar a cualquier parte del organismo. El desarrollo de un tumor de cáncer es un proceso multifacético y heterogéneo. Sin embargo, existen rasgos comunes en el desarrollo de tumores en diferentes tejidos, siendo el principal la multiplicación de células anormales que pueden propagarse más allá de sus límites habituales e invadir otros órganos [4].

A nivel celular, dichos rasgos comunes en las neoplasias han sido compiladas y descritas a lo largo de dos décadas por Hanahan y Weinberg, dando lugar a los *Hallmarks of Cancer*[10]. Como marco conceptual, estos rasgos distintivos pretenden explicar los orígenes de las neoplasias y son definidos como capacidades funcionales adquiridas que permiten que las células del cáncer sobrevivan, proliferen y se diseminen [11-13].

Los signos distintivos comunes a la mayoría de las neoplasias, descritos en [10, 14] y desplegados en la Figura 3 son:

- **Mantenimiento de señalización proliferativa.** Esta es una de las características principales de las células de cáncer. En tejidos normales la producción de señales que promueven el crecimiento está cuidadosamente regulada. Se asegura la homeostasis y el mantenimiento de la arquitectura y función del tejido mediante el control del ciclo de crecimiento y división celular. En tejidos tumorales la desregulación de estas señales permite que las células progresen en el ciclo celular, proliferen e incrementen su tamaño.
- **Evasión de supresores tumorales.** Además de mantener una señal proliferativa, las células de cáncer tienen que evadir los mecanismos que regulan de forma negativa la proliferación celular, muchos de los cuales están a cargo de genes supresores de tumores. Dos ejemplos muy importantes los encontramos en las proteínas *RB* y *TP53*, que operan como nodos centrales en los circuitos regulatorios que controlan la proliferación y los mecanismos de senescencia y apoptosis.
- **Evasión de apoptosis.** La muerte programada por apoptosis puede ser desencadenada por diversas situaciones de estrés a las que se enfrentan las células de cáncer. Sin embargo, éstas desarrollan una variedad de estrategias para eludir este proceso. La principal es la pérdida de función de *TP53*, aunque también se observa un incremento en las señales anti-apoptóticas o señales de supervivencia y disminución de las señales pro-apoptóticas.
- **Potencial replicativo ilimitado.** Los telómeros son secuencias repetidas en los extremos de los cromosomas que protegen el Ácido Desoxirribonucleico (ADN) y que se desgastan durante la división celular. Su desgaste sucesivo genera estrés y posterior senescencia y muerte celular, limitando el número de veces que una célula puede pasar a través del ciclo de crecimiento y división. Uno de los mecanismos mediante los cuales las células de cáncer pueden evadir dichas barreras, es gracias a la acción de la telomerasa, la ADN polimerasa especializada que añade segmentos en los telómeros y cuya actividad está ausente en células normales.
- **Inducción de angiogénesis.** Al igual que los tejidos normales, los tumores requieren nutrientes y oxígeno, así como un mecanismo para eliminar desechos metabólicos y dióxido de carbono. Estas necesidades se satisfacen mediante la angiogénesis o generación de nuevos vasos sanguíneos. Este proceso está presente en adultos de forma temporal

y en situaciones específicas como la cicatrización. Las células tumorales lo mantienen activo gracias a la señalización aberrante de los mecanismos pro-angiogénicos.

- **Capacidad de invadir tejidos y generar metástasis.** La molécula de adhesión E-cadherina es clave para la formación de las uniones celulares y el mantenimiento estructural del tejido. En carcinomas humanos se observa una disminución en la expresión de ésta y otras moléculas involucradas en la adhesión celular y el establecimiento de la matriz extracelular, promoviendo un fenotipo metastático.
- **Evasión de destrucción por el sistema inmune.** Entre las funciones del sistema inmune se encuentra el constante monitoreo de células y tejidos, mediante el cual se reconocen y eliminan células que presentan alteraciones iniciales asociadas a cáncer, previniendo la formación de tumores. Las neoplasias que logran formarse han evadido exitosamente dicho sistema de vigilancia gracias a sus procesos de señalización aberrante.
- **Desregulación del metabolismo celular.** Para sostener la proliferación, las células de cáncer ajustan su metabolismo energético. La producción de *ATP* se realiza mediante glucólisis en lugar de por fosforilación oxidativa en la mitocondria. Aunque las razones de este cambio no se conocen del todo, se sabe que las células de cáncer aprovechan los metabolitos intermedios que se producen en este proceso.

Este marco conceptual sigue evolucionando y dos *hallmarks* más han sido recientemente añadidos en [15]:

- **Desbloqueo de plasticidad fenotípica.** La diferenciación celular, resultado final del proceso de organogénesis, reprime la proliferación en tejidos normales. Las células de cáncer sortean dicha barrera con mecanismos como la desdiferenciación a estados progenitores, el bloqueo de la diferenciación o la transdiferenciación a otros linajes.
- **Células senescentes.** La senescencia celular involucra un arresto proliferativo que pretende ser un mecanismo de protección para mantener homeostasis, ya que inactiva células que han sufrido algún daño o se encuentran en condiciones de estrés. Sin embargo, las proteínas que estas células secretan estimulan el fenotipo tumoral en las células que las rodean, contribuyendo al desarrollo de otras características distintivas del cáncer.

Hanahan y Weinberg han identificado también lo que ahora son cuatro propiedades que facilitan el establecimiento de los signos distintivos, llamadas características habilitadoras, descritas en [14, 15]:

- **Inestabilidad genómica y mutaciones.** El establecimiento de muchos de los signos distintivos del cáncer depende principalmente de alteraciones en el genoma de las células neoplásicas. Algunas mutaciones proveen de ventajas selectivas a grupos de subclonas

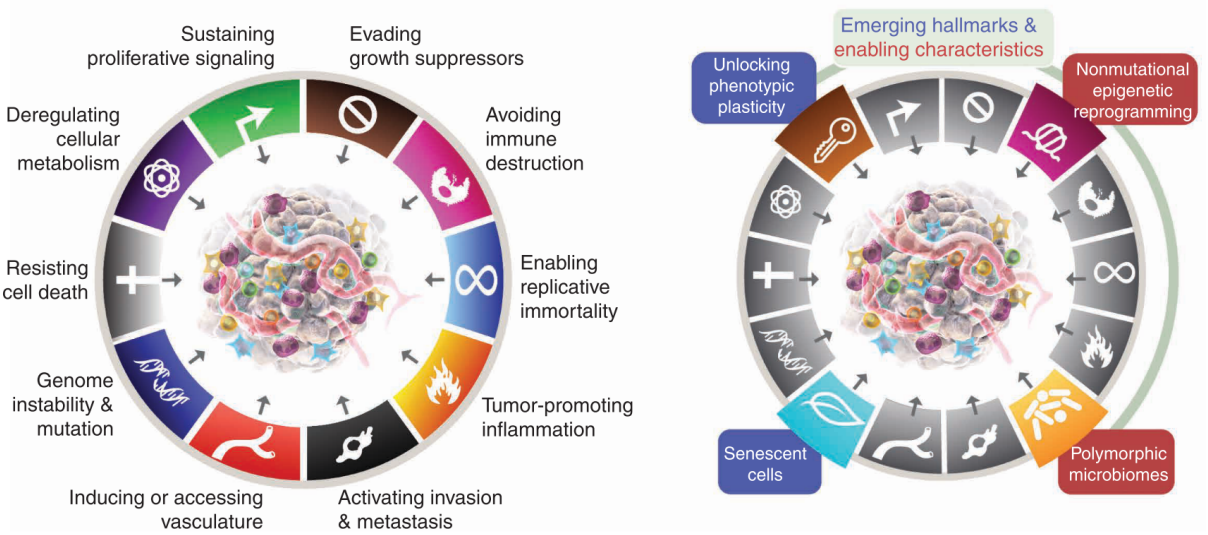

Figura 3: *Hallmarks of cancer*. El panel izquierdo muestra los ocho signos distintivos del cáncer y dos características habilitadoras, descritas en [10, 14]. El lado derecho contiene los procesos recién incorporados a este marco conceptual en 2022. Imagen tomada de [15]

celulares que crecen y eventualmente dominan el tejido. Para promover la adquisición de dichas alteraciones, las células de cáncer incrementan la tasa de mutación con mecanismos que involucran mayor sensibilidad a agentes mutagénicos o la corrupción de los sistemas que monitorean la integridad del genoma.

- **Inflamación tumor-promotora.** La presencia de células neoplásicas activa la respuesta del sistema inmune que, al ser incapaz de erradicar un tumor naciente, genera un estado de inflamación en el microambiente tumoral debido a los factores de crecimiento, factores pro-angiogénicos, las enzimas que alteran la matriz extracelular y demás biomoléculas, contribuyendo al establecimiento del fenotipo tumoral.

- **Reprogramación epigenética no-mutacional.** Además de las alteraciones a nivel del genoma, las modificaciones epigenéticas contribuyen también a cambios en la expresión genética que promueven el fenotipo tumoral. Estas alteraciones generan otro nivel de heterogeneidad en la población celular.

- **Microbiomas polimórficos.** Los tejidos que componen la barrera de nuestro organismo interactúan con una gran cantidad de microorganismos que pueden impactar en el desarrollo de enfermedades como el cáncer, con efectos tanto benéficos como desfavorables. Además, la presencia de estos microbiomas está asociada con otras características habilitadoras como la inflamación tumor-promotora, por su interacción con el sistema inmune y la inestabilidad genómica, debido a la producción de toxinas y otras moléculas que presentan algunas bacterias y que son capaces de dañar el ADN directamente, vulnerar los sistemas que mantienen la integridad genómica o causar estrés celular que de forma indirecta daña la fidelidad con la que el ADN se replica y repara.

1.3. Regulación de la transcripción

Las disrupciones en las características habilitadoras logran su realización en un fenotipo oncogénico que manifiesta los signos distintivos funcionales, principalmente mediante el proceso de transcripción genética. Es decir, las alteraciones genéticas y epigenéticas en las células de cáncer, junto con las señales que provienen del microambiente tumoral, promueven la expresión de genes involucrados en los procesos asociados a los signos distintivos del cáncer. Este complejo escenario, con algunos de sus elementos ejemplificados en la Figura 4, resulta en cambios funcionales coordinados y complementarios en diversos procesos biológicos [16].

La regulación transcripcional, es decir, el conjunto de procesos que controlan la transcripción genética, requieren de una compleja sinergia entre múltiples mecanismos regulatorios. A continuación se detallan algunos de sus mecanismos principales.

En el núcleo de una célula eucarionte el genoma está compartamentalizado principalmente en dos niveles. En el primero se observan regiones cromosómicas con propiedades funcionales y bioquímicas similares: la heterocromatina y la eucromatina. En el segundo nivel encontramos a los dominios topológicamente asociados o Dominio Topológicamente Asociado (TAD)s, que son regiones con alto número de interacciones internas en la cromatina, pero con menos contactos en las regiones vecinas [17]. La eucromatina y la heterocromatina se caracterizan por modificaciones particulares en las histonas, las proteínas que forman las unidades estructurales del genoma llamadas nucleosomas. Dichas modificaciones propician una estructura de cromatina más laxa en la eucromatina, asociada a mayor actividad transcripcional; mientras que la heterocromatina se encuentra transcripcionalmente silenciada y es más compacta. A su vez, las regiones que delimitan los TADs se estabilizan gracias a la presencia de complejos de cohesina y otras proteínas estructurales como *CTCF* [18].

Esta organización estructural en el genoma es dinámica. Es decir, las marcas en las histonas pueden alterarse y los dominios topológicamente asociados pueden reestructurarse. Además, hay asociaciones entre los componentes de la maquinaria transcripcional capaces de definir algunas características de la eucromatina. Por lo tanto, es posible la retroalimentación entre los cambios en la expresión de un gen y las alteraciones estructurales en la región que ocupa [17].

A nivel de ADN se encuentran los promotores y los *enhancers*, también llamados secuencias -*cis* reguladoras. Para iniciar la transcripción, la Polimerasa II, encargada de sintetizar ARN mensajeros en células eucariontes, debe ganar acceso a la región del promotor al inicio de un gen. A su vez, los *enhancers* regulan este proceso a larga distancia y típicamente en regiones aisladas dentro de los TADs, usualmente flanqueadas por sitios de unión de la proteína estructural *CTCF* [19]. La comunicación entre *enhancers* y sus promotores blanco requiere

proximidad y depende de la arquitectura dinámica de la cromatina, así como también de la actividad de los factores de transcripción [17].

Los factores de transcripción son proteínas que se unen a secuencias específicas del ADN tanto en promotores como en *enhancers* y guían a la polimerasa a sus promotores blanco. Algunos factores controlan el inicio de la transcripción, como los factores pioneros encargados de abrir la cromatina al inicio del proceso, mientras que otros controlan la elongación. Además, unen cofactores, que son complejos proteicos que contribuyen a la activación o represión transcripcional, pero no tienen la capacidad de unirse al ADN [20]. El proceso está también acompañado del reclutamiento de complejos remodeladores de la cromatina que tienen que desplazar los nucleosomas para permitir el acceso de la maquinaria transcripcional. A su vez, la regulación de la elongación de la transcripción con sus correspondientes factores afina la cantidad de ARN sintetizada por unidad de tiempo [19].

El conjunto de genes transcritos definen en gran medida la identidad celular y en las células de cáncer se observan alteraciones que afectan cada componente del programa regulatorio, modificando la expresión genética y generando cambios en el metabolismo y la actividad de señalización en las células tumorales y su microambiente [21], descritos ya en los signos distintivos del cáncer.

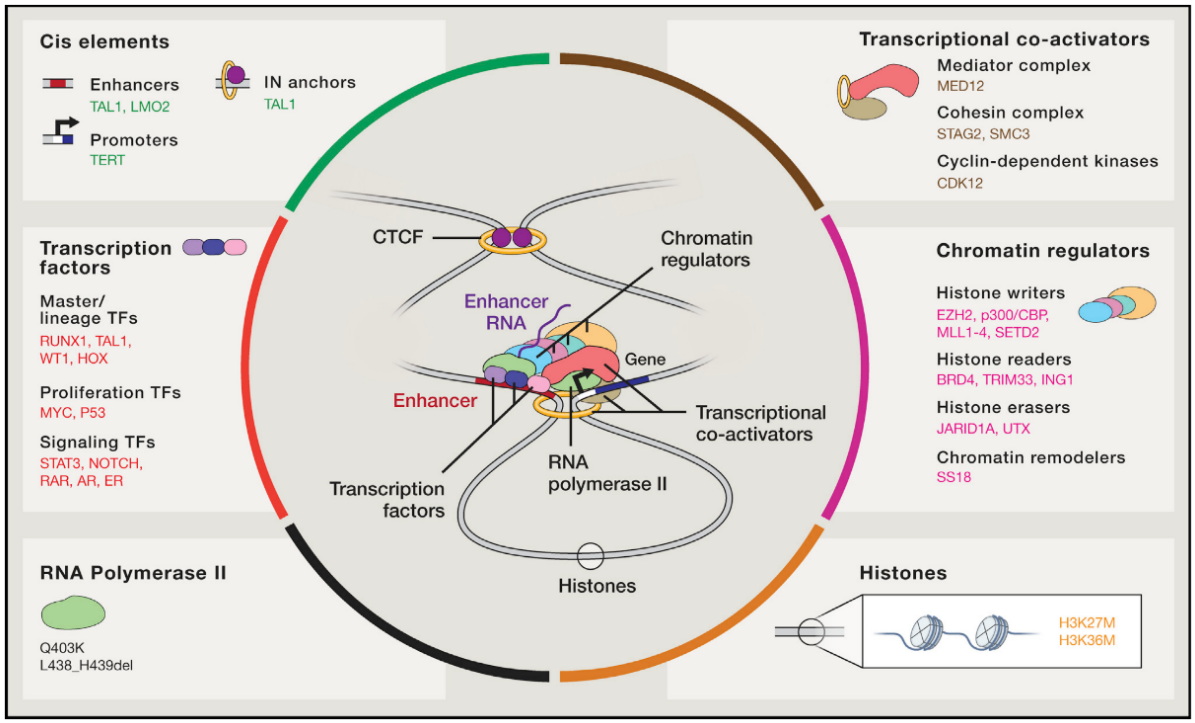

Figura 4: Elementos regulatorios de la transcripción y algunas alteraciones encontradas en cáncer. Imagen tomada de [22]

Las alteraciones en las células tumorales pueden ocurrir directamente en los elementos -*cis*, *enhancers* y promotores, cuyas mutaciones promueven típicamente la expresión de oncogenes o pueden ser indirectas, como las mutaciones en factores de transcripción, proteínas de señalización, cofactores, reguladores de la cromatina o proteínas estructurales [22]. Por ejemplo, el promotor de *TERT*, parte del complejo que forma la telomerasa, presenta mutaciones que promueven su expresión en cáncer [23]. En cuanto a alteraciones indirectas, el factor *P-TEFb* o *positive transcription elongation factor b*, que controla las transiciones entre pausa y elongación de la Polimerasa II, puede ser estimulado por el oncogen *MYC* para amplificar la transcripción y producir un incremento en los niveles de transcritos [22].

Los efectos del programa regulatorio descrito y sus estados alterados en cáncer pueden estudiarse analizando el transcriptoma, ya que los valores de expresión de los genes y de las secuencias no codificantes son un resultado medible del proceso de transcripción, gracias a la secuenciación de ARN. Estos análisis y sus tecnologías asociadas forman parte de la transcriptómica, disciplina en cuyos objetivos están: catalogar las diferentes especies de transcritos, determinar la estructura transcripcional de los genes, sus patrones de *splicing* y otras modificaciones post-transcripcionales y cuantificar los cambios de expresión de los transcritos en diferentes condiciones [24]. Además, los datos de expresión genética pueden servir de entrada para diversos análisis funcionales y estadísticos como el análisis de co-expresión genética [21], una de las herramientas principales en este proyecto.

1.4. Co-expresión genética

El análisis de perfiles de co-expresión genética permite identificar conjuntos de genes cuya expresión se encuentra altamente correlacionada o con patrones de expresión coordinada [25], sugiriendo su asociación a mecanismos regulatorios que controlan su expresión de manera conjunta [26]. Este tipo de estudios pertenecen al área de biología de sistemas, que integra información biológica que puede provenir de diferentes fuentes o de diferentes tecnologías asociadas a las ciencias ómicas, con las cuales se construyen modelos matemáticos para estructurar hipótesis que pueden ser probadas experimentalmente, generando nuevo conocimiento [27].

Los perfiles de co-expresión genética se obtienen calculando la correlación entre los valores de expresión de pares de genes. La correlación indica la dependencia entre dos variables, la cual puede ser o no, lineal. En este caso, las dos variables son los valores de expresión de dos genes en una matriz de expresión que contiene múltiples muestras de un fenotipo particular. Por lo tanto, la correlación en los valores de expresión de un par de genes, también puede ser llamada co-expresión y para calcularla pueden utilizarse, por ejemplo: el coeficiente de correlación de Pearson, Spearman y el cálculo de Información Mutua (IM).

Los valores de co-expresión más significativos pueden ser extraídos para construir una red de co-expresión genética donde los nodos representan genes y los enlaces indican la presencia de una relación de alta co-expresión [25] entre dos genes. Estas redes proveen de una caracterización topológica al transcriptoma y nos permiten comparar propiedades de conectividad global, como el número de componentes conexos en las redes y conectividad local, mediante la identificación de comunidades o comparando directamente los nodos que participan en las interacciones. Existen muchos métodos para construir redes de co-expresión, que dependen del enfoque y los datos disponibles [28].

En investigación en cáncer las redes de co-expresión se han utilizado para estudiar conjuntos de genes asociados a procesos funcionales importantes en la enfermedad [29], para identificar genes biomarcadores asociados a prognosis o respuesta a tratamiento [30, 31] y, en el caso particular del laboratorio, al estudio del fenómeno de la pérdida de co-expresión inter-cromosómica en cáncer [32], que será detallado en la siguiente sección.

1.5. Pérdida de la co-expresión -*trans* en cáncer de mama

Los *circos plots* en la Figura 5, una modificación de la figura del artículo *RNA-Seq based genome-wide analysis reveals loss of inter-chromosomal regulation in breast cancer*, publicado en el año 2017 por Espinal y colaboradores [32], muestra las interacciones de más alta co-expresión entre pares de genes en un conjunto de muestras de tejido sano y muestras de tejido de cáncer de mama. Sobre la parte externa de la circunferencia se aprecian los cromosomas del genoma humano y las diferentes regiones dentro del cromosoma, llamadas citobandas. Cada arco o enlace, de color azul o anaranjado, une a un par de genes con alto valor de co-expresión (en este caso el top $0.01\,\%$ de las interacciones). Los arcos azules conectan genes dentro del mismo cromosoma, llamadas interacciones -*cis* o intra-cromosómicas, mientras que los arcos anaranjados unen genes en diferentes cromosomas, interacciones -*trans* o inter-cromosómicas.

Aunque a simple vista no parezca, ambas gráficas contienen el mismo número de interacciones. Sin embargo, se observa una gran diferencia entre el tipo de interacciones más abundantes en cada tejido. En tejido sano se muestra un alto número de interacciones -*trans*, mientras que en el tejido de cáncer las interacciones -*trans* son escasas y la mayoría de las interacciones de alta co-expresión son -*cis*. Además, en cáncer de mama, las interacciones -*cis* parecen aglomerarse en vecindarios cercanos. A este fenómeno, reportado por primera vez en el laboratorio, se le ha llamado pérdida de co-expresión -*trans* en cáncer de mama y es la motivación principal para este proyecto.

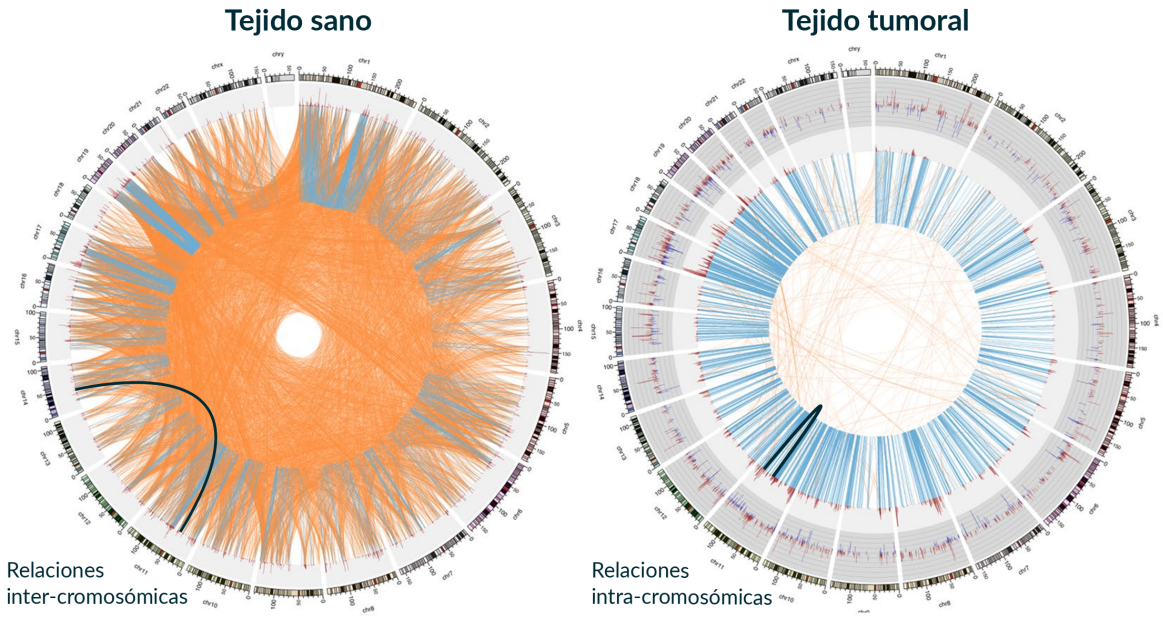

Figura 5: *Circos plots* representando la pérdida de co-expresión inter-cromosómica en cáncer de mama. Ambas gráficas contienen el 0.01 % de interacciones con valores más altos de Información Mutua (IM). Figura modificada de [32].

Las redes del estudio citado anteriormente fueron inferidas a partir de datos de secuenciación de ARN del proyecto The Cancer Genome Atlas (TCGA), una iniciativa estadounidense que tuvo como objetivo compilar un catálogo de las alteraciones genómicas en tumores humanos, mediante diferentes técnicas de secuenciación [33]. Los *datasets* de TCGA son accesibles de forma pública y gratuita, por lo que se han convertido en una fuente de datos para diversos proyectos.

Altos valores de co-expresión local han sido reportados en tejidos normales, tanto en datos de secuenciación de ARN *bulk* [34] como en secuenciación de célula única [35], reforzando la conclusión de que el orden de los genes en el genoma de células eucariontes no es aleatorio [36, 37]. Sin embargo, este fenómeno ha sido observado en pares de genes en una vecindad local abarcando distancias menores a una mega base [38, 39]. En cuanto a los mecanismos regulatorios asociados, el mismo grupo de investigación concluyó que no es posible relacionar de manera principal un único mecanismo (sitios de unión a *CTCF*, presencia de factores de transcripción, etc.), sino que es una contribución coordinada de múltiples elementos [34].

Para continuar investigando sobre la pérdida de co-expresión *-trans* en cáncer de mama y además establecer las bases del análisis computacional que permitiría ampliar la evaluación hacia otros tipos de cáncer, decidimos analizar los perfiles de co-expresión en subtipos moleculares de cáncer de mama, lo que dio lugar a la publicación: *Gene Co-expression Is Distance-Dependent in Breast Cancer* [40].

Los cuatro principales subtipos en cáncer de mama: Luminal A, Luminal B, Her2 positivo y Basal, son un ejemplo de la heterogeneidad que se observa en diversas manifestaciones del fenotipo tumoral. Estos subtipos no solamente presentan diferencias a nivel molecular [41], con características particulares en términos de patrones de expresión y alteraciones genómicas, sino que también son diferentes a nivel clínico, con pronósticos más desfavorables en el subtipo Basal [42]. Estos contrastes sugerían la posibilidad de que los perfiles de co-expresión de los subtipos tumorales fueran también distintos.

Las redes de co-expresión construidas con las diez mil interacciones de IM más altas de cada subtipo, presentadas en la Figura 6 donde los genes están coloreados de acuerdo al cromosoma en el cual se localizan, muestran que, a este valor de corte, la red del subtipo Luminal A mantiene un componente formado por interacciones inter-cromosómicas, identificado con un círculo gris. En cambio, las redes de los subtipos restantes tienen componentes conexos más chicos formados principalmente por interacciones dentro de un mismo cromosoma.

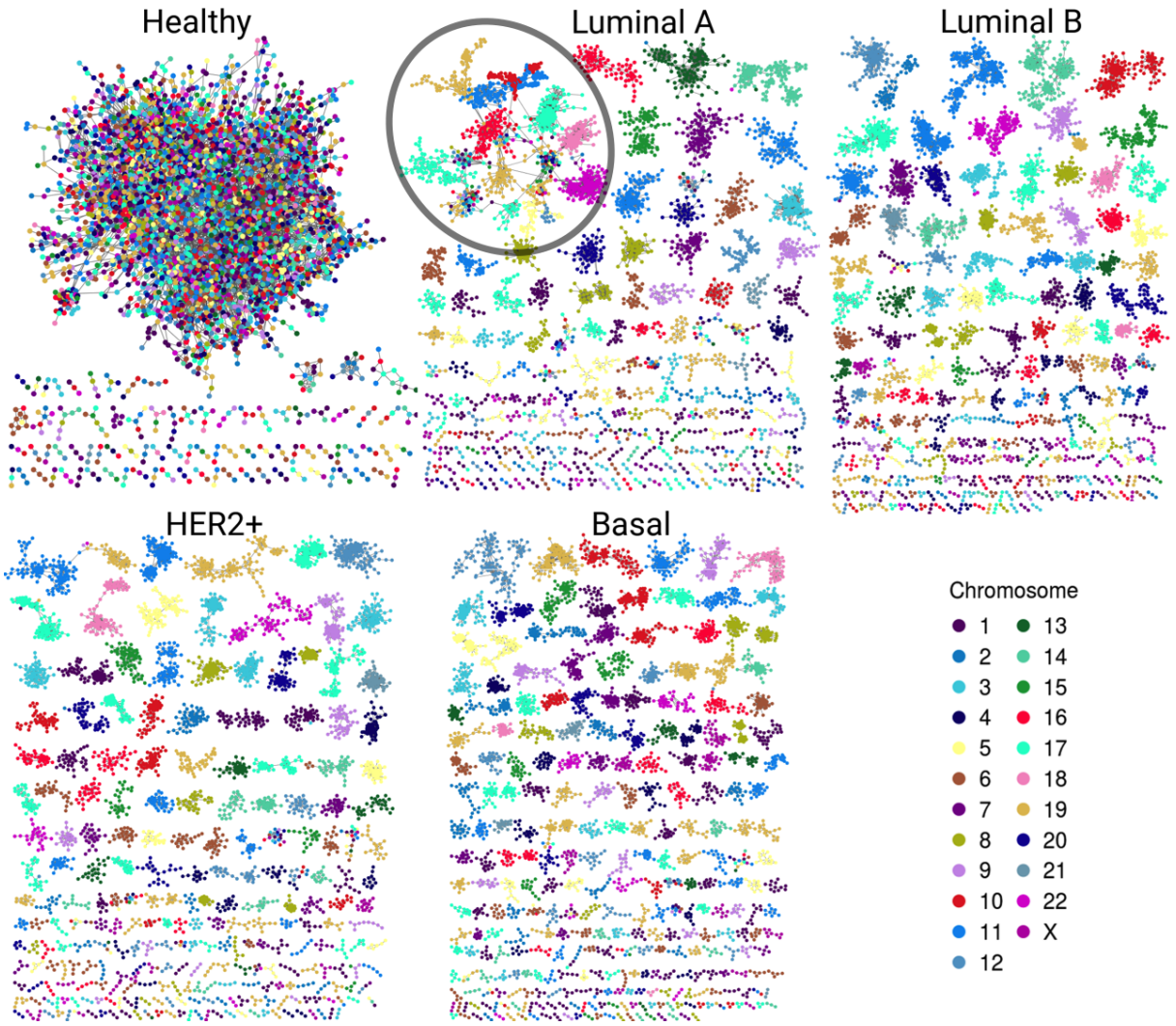

Figura 6: Redes de co-expresión en subtipos moleculares de cáncer de mama construidas con las diez mil interacciones más altas de Información Mutua (IM).

La red de tejido sano mantiene la estructura reportada en la publicación del tejido sin subtipificar [32]: un componente conexo conformado por un alto número de interacciones, la mayoría *-trans* y múltiples componentes más pequeños, de tres o cuatro nodos, con interacciones también *-trans*.

Además, también evaluamos el comportamiento de los valores de co-expresión en las interacciones que se dan entre genes del mismo cromosoma. Encontramos que, en cáncer, dichos valores son más altos en parejas de genes que están a distancias cortas, donde la distancia es medida en pares de bases entre las secuencias de ADN que los codifican y al incrementarse la distancia, los valores de IM establecen una meseta. Esto no sucede en el fenotipo normal y nuevamente, los subtipos moleculares presentan diferencias en estos perfiles, como puede observarse en la Figura 7.

Finalmente, encontramos también que esas interacciones *-cis* no se dan de forma regular a lo largo de los cromosomas, sino que existen regiones con una alta aglomeración de interacciones, mientras que otras regiones están prácticamente desiertas de enlaces de alta co-expresión. La distribución de estas regiones es diferente en cada cromosoma y para cada subtipo. Como ejemplo, la Figura 8 presenta las interacciones dentro del cromosoma 8, diferenciando interacciones dentro de las citobandas con color azul y enlaces que unen genes en diferentes citobandas de color morado. Puede observarse que el subtipo Luminal A concentra un alto número de interacciones en el extremo del brazo p; mientras que los cuatro subtipos, en menor medida el Basal, tienen un alto número de interacciones en la citobanda q24.3.

Una vez descrita la pérdida de co-expresión a larga distancia en los subtipos moleculares de cáncer de mama, decidimos ahondar en los mecanismos que podrían dar lugar a dicho fenómeno.

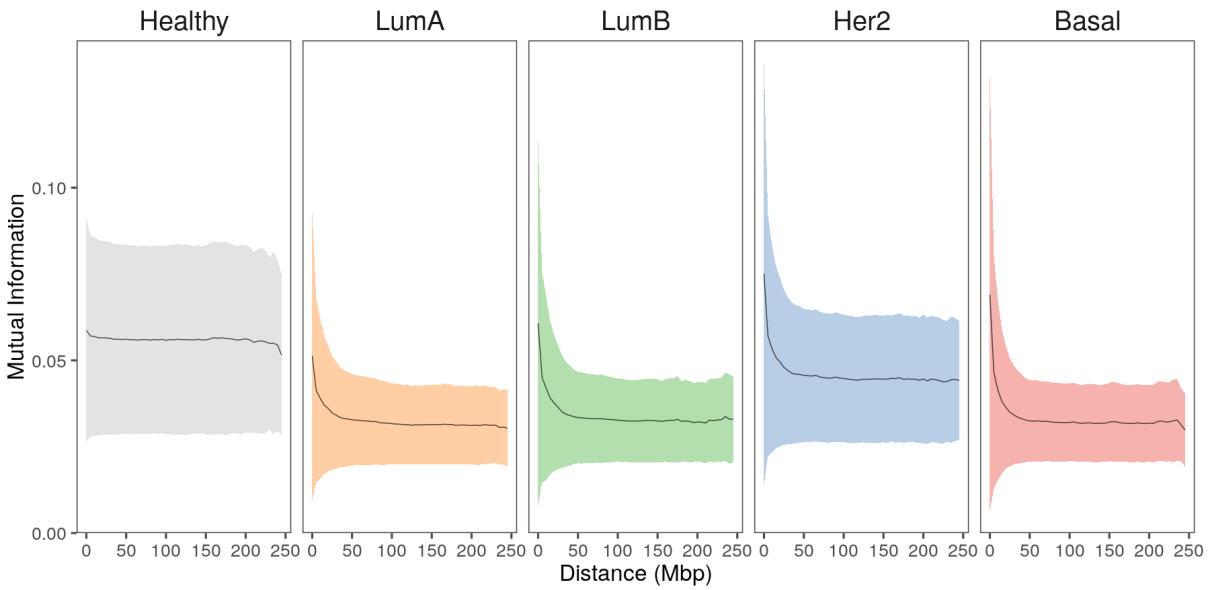

Figura 7: Valores promedio de Información Mutua (IM) y su desviación estándar, graficada contra la distancia promedio entre genes en grupos de mil interacciones.

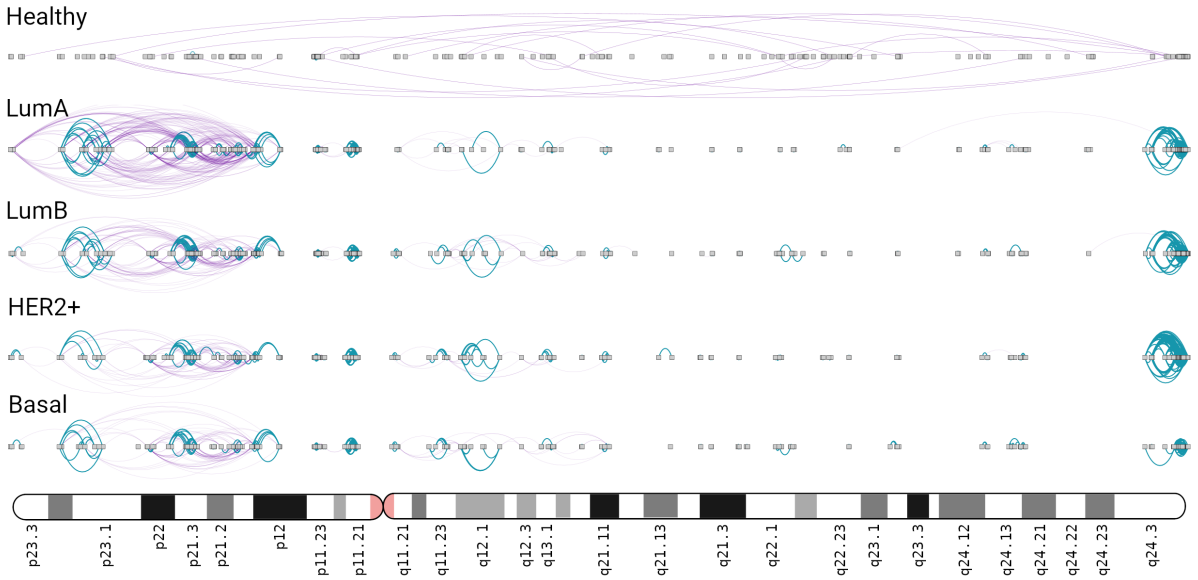

Figura 8: Interacciones dentro del cromosoma 8 en los subtipos moleculares de cáncer de mama.

Esto llevó a la publicación: *Luminal A Breast Cancer Co-expression Network: Structural and Functional Alterations* [43]. En esta publicación nos concentramos en la red de co-expresión del subtipo Luminal A, por ser el que conserva mayor similitud al fenotipo sano. Realizamos una caracterización funcional de la red a partir de sus comunidades y analizamos diversos elementos asociados a la regulación de la transcripción, como la identificación de interacciones entre factores de transcripción y genes blanco, la presencia de sitios de unión a *CTCF*, que es una proteína estructural relacionada con la formación de bucles en el ADN que aíslan secuencias reguladoras e identificamos regiones con alteración en número de copias.

El primer paso en este análisis fue la identificación de comunidades en la red de co-expresión. En su definición más clásica, las comunidades son grupos de nodos densamente conectados entre ellos y escasamente conectados a otros nodos fuera de la comunidad [44]. La partición de la red en comunidades permite su posterior asociación a características funcionales u otro tipo de evaluaciones en donde se toman en cuenta los genes que pertenecen a cada una de ellas.

En este estudio, para cada comunidad, realizamos un cómputo de dos características a las que llamamos asortatividad cromosomal y asortatividad de expresión diferencial. La asortatividad permite cuantificar la preferencia con la que los nodos con un atributo similar tienden a conectarse entre ellos [44]. La asortatividad cromosomal nos indica si los nodos de una comunidad tienen la tendencia a pertenecer a un mismo cromosoma. A su vez, la asortatividad de expresión diferencial evalúa si los genes comparten el mismo signo de expresión diferencial, es decir, si están conjuntamente sobre o subexpresados. Ambos valores tienen un intervalo de $[-1, 1]$, donde -1 indicaría interacciones entre genes de diferentes cromosomas o diferente signo de expresión diferencial, mientras que un valor de 1 se asignaría a comunidades formadas por

genes totalmente similares en dichas características.

Las comunidades fueron también relacionadas a características funcionales, mediante un análisis de sobre-representación utilizando los términos de la categoría de Procesos Biológicos de Gene Ontology (GO). Este análisis no toma en cuenta las interacciones, sino los genes de cada comunidad para evaluar si el traslape entre dicho conjunto y el conjunto de genes anotados en cada proceso en GO es significativo, dado el número total de genes en la red y el tamaño del conjunto de genes anotados.

Ambos valores de asortatividad y el número de términos de GO asociados por comunidad se muestran en la Figura 9. El tamaño de los círculos está asociado al número de nodos en la comunidad mientras que su color indica el número de términos de Gene Ontology asociados. Se observa que las comunidades con mayor número de términos biológicos asociados tienden a estar formadas por enlaces entre genes de diferentes cromosomas (asortatividad cromosomal negativa) y con genes conjuntamente sobre o subexpresados (asortatividad de expresión diferencial positiva).

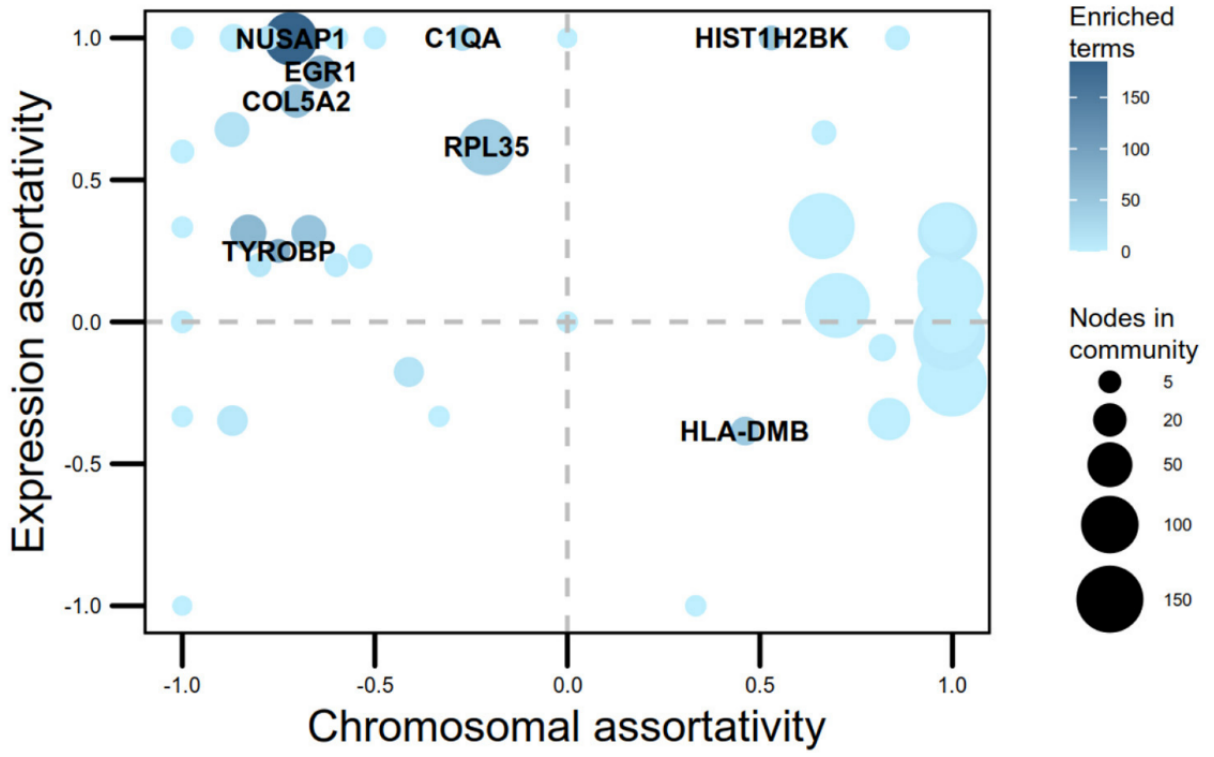

Figura 9: Valores de asortatividad cromosomal y asortatividad de expresión diferencial para cada comunidad en la red.

Para evaluar si la actividad de factores de transcripción podría explicar la formación de comunidades, identificamos los genes con dicha actividad en la red y determinamos si sus interacciones de alta co-expresión corresponden a relaciones de factor de transcripción y gen blanco. Las comunidades se nombran a partir del gen asociado al nodo cuyo valor de *page rank* [45], una medida de centralidad en ciencia de redes, es mayor.

En el caso de la comunidad de *NUSAP1*, donde todos los genes se encuentran sobreexpresados, encontramos un alto número de interacciones entre el factor de transcripción *FOXM1* y sus genes blancos y, en menor medida, pero también importante, interacciones del factor *CENPA* (Figura 10 páneles A y B). El color de los nodos está asociado al valor de expresión diferencial. Los nodos con circunferencia amarilla son genes con actividad de factores de transcripción. Las interacciones identificadas con color verde son enlaces de co-expresión regulatorios, es decir, que van de un factor de transcripción hacia su gen blanco. Aunque la actividad de factores de transcripción es importante en esta comunidad en particular, no es un comportamiento generalizable para toda la red.

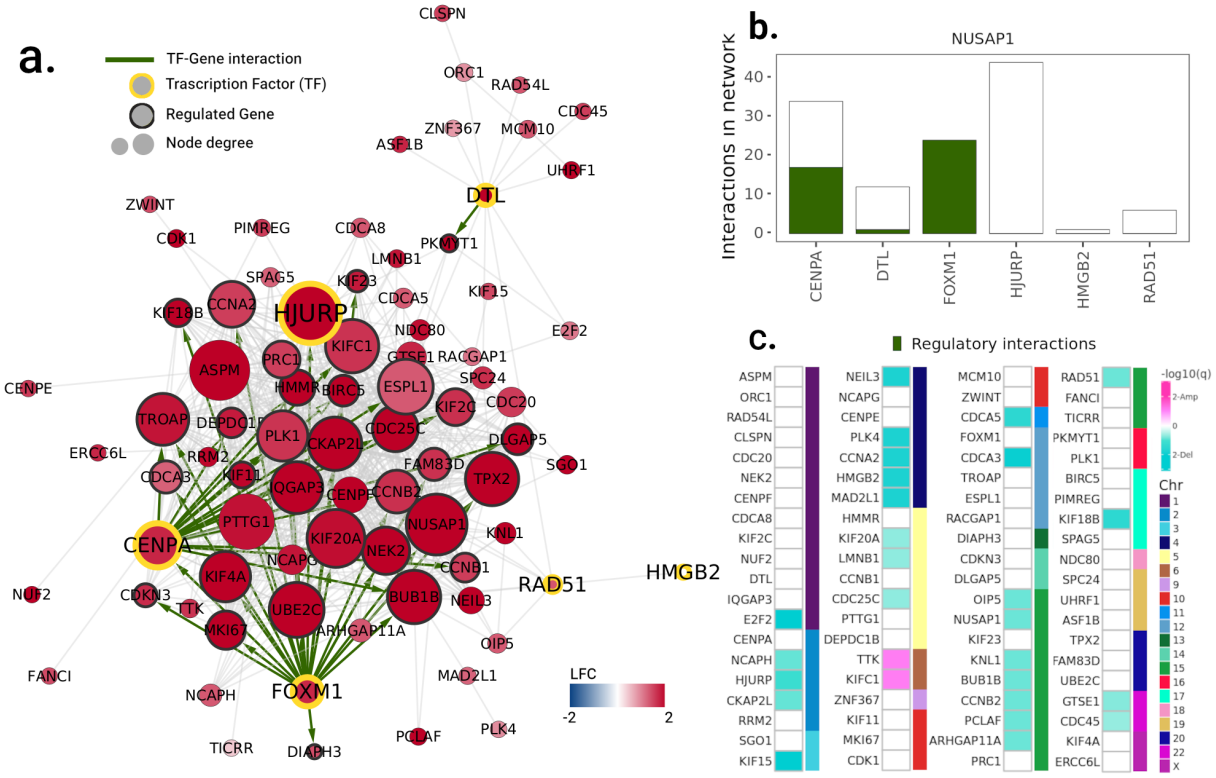

Figura 10: A) Comunidad de genes en la red de co-expresión del subtipo Luminal A identificada por el nombre de *NUSAP1*. B) Número de interacciones que unen a factores de transcripción con algún otro gen. C) Genes en la comunidad y su posible localización en picos de deleción (verde agua) o amplificación (rosa).

Como se mencionó previamente, la inestabilidad genómica es una característica habilitadora del cáncer. Por lo tanto, decidimos evaluar si las regiones que presentan alteración en el número de copias de los genes, podrían estar asociadas con la formación de *clusters* con alta densidad en el número de interacciones intra-cromosómicas. Esto es una explicación plausible ya que secuencias que abarcan múltiples genes y que están amplificadas en el genoma, podrían transcribirse de forma coordinada, afectando los valores de expresión de los genes en la región y por lo tanto, incrementando su co-expresión. Lo mismo para regiones deletadas que contienen múltiples genes. Para la comunidad de *NUSAP1*, los resultados se muestran en la Figura 10. Nuevamente, aunque algunas comunidades concuerdan con dichas alteraciones genéticas, su presencia no es suficiente para explicar el fenómeno a nivel de todo el genoma.

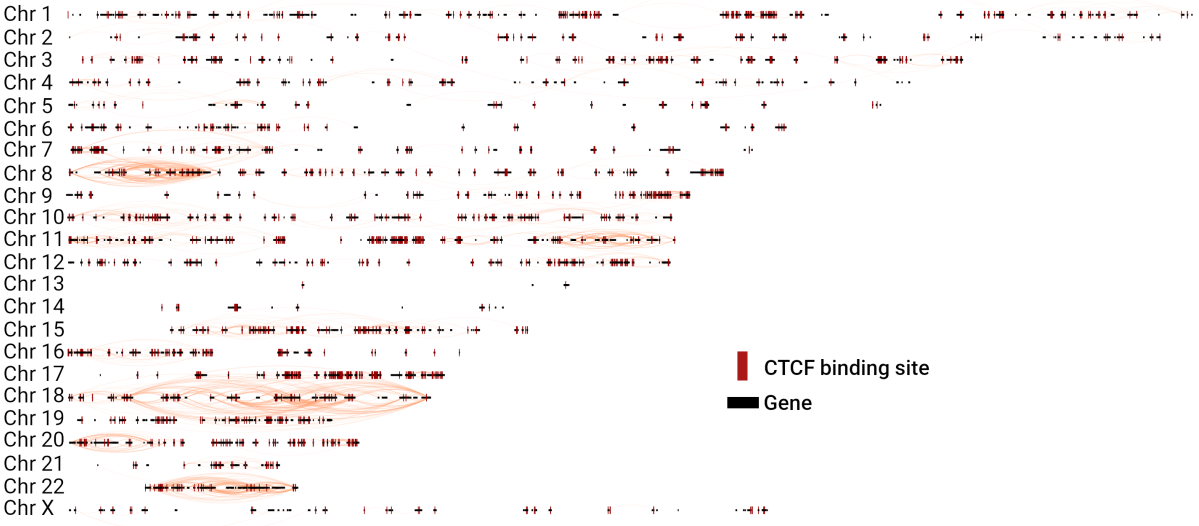

Figura 11: Sitios de unión a *CTCF* asociados a comunidades intra-cromosómicas.

Finalmente, para evaluar si los conglomerados con alta densidad de interacciones *-cis* se encuentran dentro de bucles en regiones regulatorias insuladas, analizamos la presencia de sitios de unión a *CTCF*. Para ello utilizamos solamente las interacciones intra-cromosómicas, identificamos comunidades formadas por estas interacciones y comparamos el número de sitios de unión a *CTCF* dentro de la comunidad con el número de sitios en las 50 kilo bases fuera de sus extremos. Las diferencias no fueron significativas. Los sitios de unión a *CTCF* en las diferentes comunidades pueden apreciarse en la Figura 11.

En resumen, el análisis realizado en la red de co-expresión del subtipo Luminal A sugiere que las comunidades asociadas a procesos biológicos suelen estar mayormente formadas por genes de diferentes cromosomas. Además, indica que, aunque algunos mecanismos alterados de la regulación transcripcional influyen en la pérdida de co-expresión a larga distancia, ninguno es capaz de explicar el fenómeno en su totalidad.

Estos dos primeros artículos asociados a esta tesis permitieron una evaluación preliminar de los conceptos de ciencia de redes y las herramientas computacionales que se requerirían para la investigación de la pérdida de co-expresión a larga distancia en otros tipos de cáncer. Además, los resultados obtenidos por compañeros del laboratorio en cáncer de pulmón [46], y cáncer de riñón [47], apoyan la hipótesis y el planteamiento sugerido para esta tesis, detallado a continuación.

1.6. Descripción del proyecto de tesis

Pregunta de investigación

Si la co-expresión inter-cromosómica se pierde en cáncer de mama. ¿Se pierde también en otros tipos de cáncer?

Objetivo General

Evaluar la co-expresión inter e intra-cromosómica en diferentes tipos de cáncer mediante un enfoque de biología de sistemas.

Objetivos Específicos

- Construir redes de co-expresión genética con datos de secuenciación de ARN de tejido tumoral y tejido sano para diferentes tipos de cáncer.
- Identificar la localización cromosómica de los genes que resulten relacionados en dichas redes.
- Comparar la proporción de interacciones inter e intra-cromosómicas en las redes obtenidas.
- Analizar si la pérdida de regulación inter-cromosómica es un fenómeno generalizado en cáncer.

Hipótesis

En redes asociadas a tejido sano se observará un componente conexo formado por genes que interactúan con genes de otros cromosomas. En cambio, en redes asociadas a tejido tumoral la mayoría de las interacciones unirán genes pertenecientes al mismo cromosoma.

2 Metodología. Análisis de perfiles de co-expresión

Para obtener y posteriormente analizar los perfiles de co-expresión en el fenotipo normal y los tejidos de cáncer, se implementaron dos flujos de trabajo. El primero se encarga del procesamiento de datos de secuenciación de ARN, incluida su obtención, control de calidad, normalización, expresión diferencial y finaliza con el cálculo de Información Mutua (IM) como medida de correlación en parejas de genes para obtener el perfil de co-expresión de cada fenotipo. El segundo *pipeline* ejecuta el análisis de los perfiles y las redes de co-expresión, incluyendo la asociación a procesos biológicos de las redes. Ambos utilizan *Snakemake* [48] como sistema para el manejo de flujos de trabajo e integran *scripts* desarrollados en R. La implementación de estos *pipelines* tuvo como finalidad generar un análisis reproducible, altamente automatizado y común para todos los tejidos. El código fuente de se encuentra en: `https://github.com/ddiannae/tcga-xena-pipeline` y `https://github.com/ddiannae/distance-analysis`. Además, para la integración de datos y la generación de figuras se desarrolló otro conjunto de *scripts* que pueden encontrarse en: `https://github.com/ddiannae/pan-loss-correlation`.

Los pasos principales en el análisis se muestran en la Figura 12.

2.1. Bases de datos

TCGA [33] es un proyecto estadounidense que inició en el año 2006 liderado por el *National Cancer Institute* y el *National Human Genome Research Institute* de Estados Unidos. A través del portal Genomic Data Commons (GDC), `https://portal.gdc.cancer.gov`, este proyecto ofrece de forma pública y gratuita, datos de genómica, transcriptómica, epigenómica y proteómica, de alta calidad por lo que es una excelente fuente de datos para proyectos de biología de sistemas en cáncer. Para los tejidos de vejiga, mama, colon, riñón, hígado, pulmón, tiroides y útero, los datos de secuenciación de ARN se obtuvieron de TCGA. El conjunto de datos del fenotipo normal lo constituyen las muestras etiquetadas como *Solid Tissue Normal*, mientras que para el conjunto de cáncer se utilizaron las muestras de *Primary Tumor*.

Sin embargo, para algunos tejidos, el número de muestras del fenotipo normal en TCGA es bajo y deben buscarse otras opciones. En este caso, para los tejidos de cerebro, esófago, ovario, páncreas, próstata, piel y testículo, los datos se obtuvieron del proyecto Xena del *Genomics Institute* de la Universidad de California en Santa Cruz (UCSC) [49]. El conjunto

completo está disponible en la página de descargas de UCSC Xena: `https://xenabrowser.net/datapages/?cohort=TCGA%20TARGET%20GTEx`. Este proyecto integra, desde pasos iniciales del análisis de secuenciación de ARN, datos de expresión de TCGA y del proyecto Genotype-Tissue Expression (GTEx) [50], a través de un *pipeline* implementada en la plataforma Toil [51]. GTEx contiene muestras de tejidos no asociados a alguna enfermedad, lo que permite contar con más muestras para el fenotipo normal. Dichos datos están etiquetados como *Normal Tissue*, en GTEx-Xena, mientras que los datos de TCGA-Xena que conforman el conjunto de cáncer están clasificados como *Primary Tumor*.

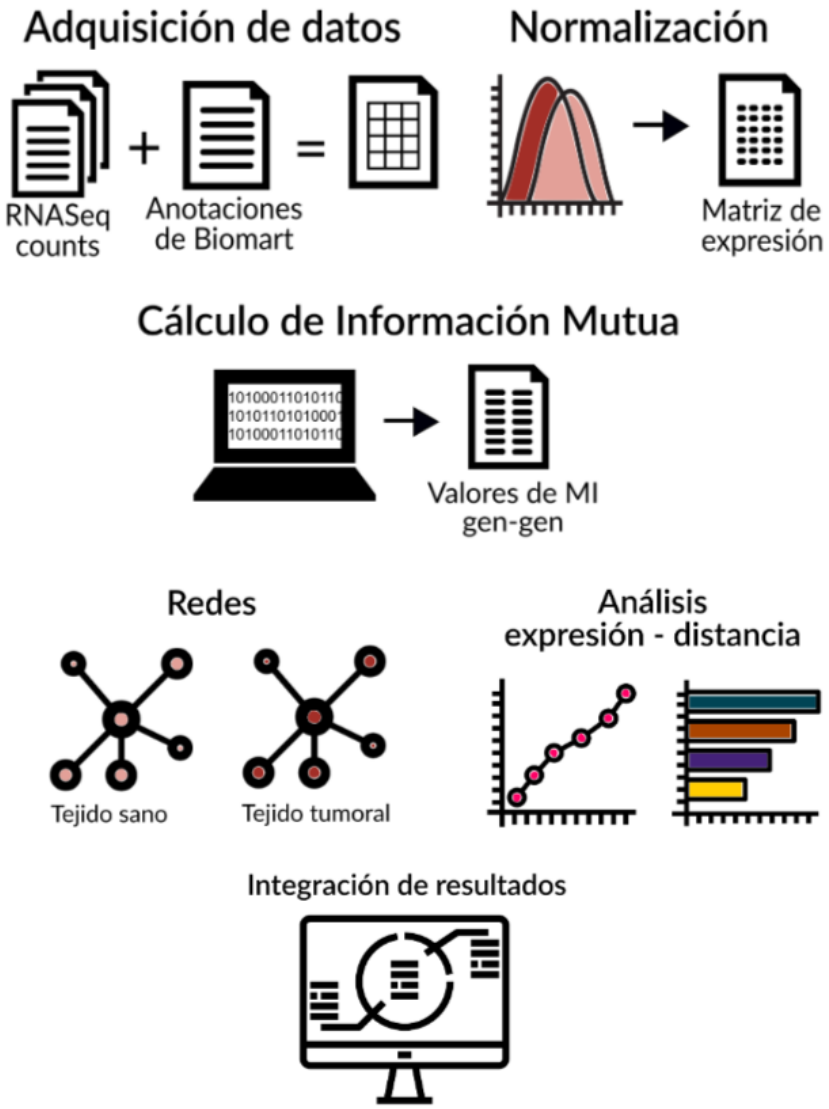

Figura 12: Flujo de trabajo para la obtención y análisis de perfiles de co-expresión.

Tabla 1: Número de muestras y fuente de datos para cada tejido analizado

Tejido	Fuente	Condición	Muestras	Genes
Cerebro	UCSC Xena	Normal Cáncer	88 508	10608
Colon	UCSC Xena	Normal Cáncer	163 287	15571
Esófago	UCSC Xena	Normal Cáncer	269 178	10438
Hígado	UCSC Xena	Normal Cáncer	107 358	14756
Mama	TCGA	Normal Cáncer	111 1047	14326
Ovario	UCSC Xena	Normal Cáncer	88 413	10236
Páncreas	UCSC Xena	Normal Cáncer	165 177	10386
Piel	UCSC Xena	Normal Cáncer	231 88	9908
Próstata	UCSC Xena	Normal Cáncer	93 473	10001
Pulmón	TCGA	Normal Cáncer	101 995	14750
Riñón	TCGA	Normal Cáncer	123 839	14314
Testículo	UCSC Xena	Normal Cáncer	162 138	11342
Tiroides	TCGA	Normal Cáncer	56 471	14637
Útero	TCGA	Normal Cáncer	35 591	13335
Vejiga	TCGA	Normal Cáncer	19 398	13390

El número de genes en la matriz de expresión (Genes) cambia por tejido debido a los filtros aplicados en el procesamiento de datos.

2.2. Procesamiento de datos de secuenciación de ARN

Preprocesamiento

Los datos de TCGA, para los cuales se utilizó el algoritmo HTSeq [52], contienen únicamente Ensembl ID y conteos por gen. Sus respectivas anotaciones (cromosoma y posición del gen, tipo de transcrito, símbolo, etc.) se añadieron usando el archivo de referencia GENCODE v22, provisto por GDC en `https://gdc.cancer.gov/about-data/gdc-data-processing/`

`gdc-reference-files`. En cambio, Los datos de UCSC Xena se encuentran como conteos esperados obtenidos por el algoritmo Expectation-Maximization (RSEM) [53], el algoritmo de cuantificación de transcritos utilizado por este proyecto, a partir de los datos de secuenciación de ARN. Fueron transformados de $log_2(expected\_count + 1)$ a $expected\_count$ en valores enteros. Estos datos se anotaron usando GENCODE v23, como se sugiere en [51].

Para los pasos restantes en el *pipeline*, solamente se mantuvieron genes que codifican para proteínas, cromosomas convencionales y secuencias que aparecen en el archivo de anotación GENCODE v37 (April, 2021). Una vez que las matrices de conteos crudos se integran con sus anotaciones correspondientes, los conjuntos de datos de ambas fuentes comparten los pasos restantes en el análisis, el cual se ejecuta de forma independiente para cada tejido.

Control de calidad

La librería `NOISeq` [54] en R se utilizó para el control de calidad de los datos. Se eliminaron los genes con bajos valores de expresión, identificados como genes con un promedio de expresión < 10 y genes con valores de expresión iguales a cero en más de la mitad de las muestras por tejido. Diagramas de caja con las distribuciones de conteos de expresión revelaron que algunas de las muestras de UCSC Xena contenían valores extremadamente bajos, así que se removieron muestras si su expresión promedio caía dos desviaciones estándar por debajo o por arriba del promedio por fenotipo.

Normalización

Se obtuvieron diferentes gráficas para identificar la influencia del tamaño de los transcritos y su contenido de GC, ya que se sabe que estas propiedades afectan en la secuenciación [55]. Para corregir dicha influencia se probaron diversas combinaciones de estrategias de normalización utilizando los paquetes `EDASeq` [56] y `NOISeq`. La mejor alternativa por tejido se seleccionó mediante inspección visual de las gráficas. Después de la normalización se utilizó la función de `ARSyNSeq`, nuevamente en la librería `NOISeq` para remover efectos de lote y para reducción de ruido.

2.3. Cálculo de Información Mutua (IM)

El cálculo de Información Mutua se computó utilizando el algoritmo ARACNE [57]. La IM mide la dependencia estadística entre dos variables aleatorias; es decir, mide la reducción de entropía de una de las variables debido al conocimiento del valor de la otra. A diferencia del coeficiente de correlación de Pearson, IM no se limita a correlaciones lineales.

El conjunto completo de valores de IM para todas las parejas de genes se obtuvo de forma independiente para cada tejido y para cada fenotipo. Para ejecutar ARACNE dentro del *pipeline* de Snakemake, se creó un contenedor utilizando la plataforma Singularity. El código fuente de este proyecto se encuentra en el siguiente repositorio: `https://github.com/ddiannae/ARACNE-multicore`.

2.4. Análisis de expresión diferencial

El análisis de expresión diferencial permite identificar genes cuya expresión se encuentra alterada de forma consistente al comparar dos fenotipos. En este caso, los genes diferencialmente expresados en las muestras de cáncer comparadas con el fenotipo normal en cada tejido se identificaron utilizando el paquete de R `limma` [58]. Los valores de significancia estadística o valores p asociados a la expresión diferencial en log_2 *fold change* se obtuvieron con pruebas de hipótesis basadas en el método de Bayes empírico y se utilizó el método de Benjamini y Hochberg para ajustar por comparaciones múltiples. Se estableció un umbral de $p_{adj} < 0.05$ para considerar genes sobre- o sub-expresados.

2.5. Análisis de perfiles de co-expresión

Identificación de fracciones de interacciones -*cis* a diferentes puntos de corte

Para evaluar la pérdida de interacciones inter-cromosómicas en cáncer se tomaron los perfiles completos de co-expresión para cada tejido y cada fenotipo; es decir, los valores de co-expresión o IM de todas las parejas posibles de genes. Esto permite eliminar el sesgo de tomar un solo punto de corte, sin conocer el comportamiento de los valores de co-expresión en cada tejido.

Los valores de IM se ordenaron de forma descendente y se calculó la fracción de interacciones entre genes del mismo cromosoma o fracción de enlaces -*cis*, a los diferentes puntos de corte establecidos. Estos puntos de corte van de mil en mil hasta diez mil, de diez mil en diez mil hasta cien mil, etc., hasta 1e8. Para comparar las fracciones observadas en tejidos normales y en cáncer, se ejecutaron pruebas de Kolmogorov-Smirnov a partir de cada punto de corte y hasta el total de pares de IM. La prueba de Kolmogorov-Smirnov se utiliza para comparar dos muestras bajo la hipótesis nula de que ambas provienen de la misma distribución [59]. En el segundo caso, que es el que aplica en esta situación, se cuantifica la distancia entre ambas distribuciones empíricas para evaluar la diferencia de ambas condiciones en la sucesión de fracciones -*cis* asociadas a los distintos umbrales.

Análisis de co-expresión contra distancia

Para evaluar la pérdida de co-expresión a larga distancia es necesario identificar únicamente el conjunto de pares de genes dentro de un mismo cromosoma o enlaces *-cis*, ya que en este caso hablamos de distancia física entre genes calculada en pares de bases y dicha distancia no puede ser definida para genes que se encuentran en diferentes cromosomas. Por lo tanto, todos los enlaces *-cis*, sin establecer punto de corte, se agruparon en conjuntos de mil interacciones formando *bins*, ordenados de menor a mayor distancia entre genes y se obtuvieron los valores promedio de IM, su desviación estándar, así como la distancia promedio. Al no establecer un punto de corte para los valores de co-expresión, los *bins* en los fenotipo normal y cáncer contienen las mismas parejas en un tejido en particular. La distribución de valores de IM en cada *bin* se comparó contra los demás *bins* en el mismo tejido y fenotipo usando la prueba de *Mann–Whitney–Wilcoxon*, también llamada *Wilcoxon rank-sum*. Esta es una prueba no paramétrica que evalúa si dos muestras provienen de poblaciones equidistribuidas. La hipótesis nula en esta prueba dice que los valores de una muestra no tienden a ser mayores a los valores de la otra. Es decir, al seleccionar los valores X y Y de ambas poblaciones, la probabilidad de que X sea más grande que Y es igual a la probabilidad de que Y sea más grande que X [60].

Para identificar regiones compartidas con alta densidad de interacciones se utilizaron las citobandas o bandas citogenéticas. Estas bandas delimitan regiones con configuraciones particulares asociadas al empaquetamiento de la cromatina y aparecen al teñir los cromosomas con metodologías como la tinción de Giemsa durante la metafase en la división celular [61]. En este análisis, en cada tejido, se tomó en cuenta el número total de interacciones dentro de cada citobanda, dados los genes en su matriz de expresión. Se seleccionaron las cien mil interacciones con valores de IM más altos y se identificaron las citobandas con más de la mitad del total de interacciones en la citobanda dentro del top cien mil, para clasificarse como regiones de alta densidad de interacciones *-cis*. La aparición de estas regiones de alta densidad fue comparada con un modelo nulo que asigna al azar el número de interacciones encontradas en cada cromosoma. Esta asignación se generó 10 mil veces para comparar el número real de interacciones intra-citobanda en esta distribución aleatoria y asignarle así un valor p.

2.6. Análisis de redes de co-expresión

Detección de comunidades

Con las cien mil interacciones de valores más altos de IM se construyeron redes de co-expresión para cada tipo de tejido y para cada fenotipo. Para detectar módulos de genes altamente co-expresados en las redes, se probaron cuatro algoritmos de detección de comunidades: Louvain, Fast Greedy, Infomap, y Leading Eigenvalue, a través de sus implementaciones en el paquete

`igraph` para R [62]. Para seleccionar el mejor algoritmo se calcularon los valores de modularidad. La modularidad es una métrica que indica qué tan buena es la partición obtenida dada la densidad de las comunidades resultantes, al ser comparada con un modelo nulo que toma en cuenta los grados de los vértices en la red original [44]. El algoritmo de Louvain obtuvo los valores de modularidad más altos en todos los casos por lo que solamente se consideró esa partición. Además, este algoritmo ha sido evaluado previamente demostrando ser efectivo para encontrar comunidades en redes biológicas que sean posteriormente asociadas a procesos funcionales [63]. Los resultados de los valores de modularidad se muestran en la Tabla 2. El nombre del gen dentro de cada comunidad con valor mayor de *page rank* [44] fue asignado como nombre de la comunidad.

Asortatividad nominal

La asortatividad, también llamada homofilia en redes, es la tendencia de los nodos a estar conectados con otros nodos con los que comparten atributos similares [44]. En este caso, la utilizamos para calcular dos tipos de asortatividades: cromosomal y de expresión diferencial. Para la asortatividad cromosomal se utiliza el cromosoma en el que cada gen está localizado. En este caso, para cada comunidad se calculó el número de enlaces que unen genes en el mismo cromosoma (interacciones -*cis*) menos el número de enlaces que unen genes en diferentes cromosomas (interacciones -*trans*), dividido entre el total de enlaces en la comunidad. La asortatividad de expresión diferencial se obtuvo de forma similar, pero solo es calculada en las redes de cáncer ya que toma en cuenta el signo del log_2 *fold change* en los genes sobre- o sub-expresados como el atributo de asortatividad, para contar los enlaces que unen a genes con el mismo signo y restar las interacciones que unen genes con signo diferente, dividiendo por el total de enlaces en cada comunidad.

$$AS_{chr} = \frac{|\{\{x,y\} \mid x,y \in C_i \text{ and } x.\text{chr } = y.\text{chr }\}| - |\{\{x,y\} \mid x,y \in C_i \text{ and } x.\text{chr } \neq y.\text{chr }\}|}{|\{\{x,y\} \mid x,y \in C_i\}|}$$

$AS_{chr} =$ Asortatividad cromosomal

$C_i =$ Comunidad i en la red

$x, y =$ Nodos en la red

Tabla 2: Valores de modularidad obtenidos por los cuatro algoritmos de detección de comunidades evaluados en las redes de co-expresión.

Tejido	Condición	Louvain	Fast Greedy	Infomap	Leading Eigenvalue
Cerebro	Normal	0.505	0.452	0.445	0.449
	Cáncer	0.796	0.758	0.752	0.193
Colon	Normal	0.550	0.513	0.486	0.002
	Cáncer	0.956	0.948	0.935	0.025
Esófago	Normal	0.624	0.584	0.579	0.002
	Cáncer	0.896	0.872	0.868	0.002
Hígado	Normal	0.498	0.459	0.422	0.446
	Cáncer	0.948	0.937	0.926	0.212
Mama	Normal	0.515	0.459	0.448	0.334
	Cáncer	0.965	0.960	0.930	0.109
Ovario	Normal	0.448	0.421	0.401	0.406
	Cáncer	0.959	0.952	0.917	0.005
Páncreas	Normal	0.506	0.458	0.466	0.003
	Cáncer	0.746	0.711	0.706	0.306
Piel	Normal	0.578	0.497	0.549	0.002
	Cáncer	0.620	0.594	0.580	0.000
Próstata	Normal	0.436	0.382	0.395	0.372
	Cáncer	0.697	0.656	0.655	0.203
Pulmón	Normal	0.545	0.500	0.477	0.386
	Cáncer	0.971	0.968	0.933	0.122
Riñón	Normal	0.556	0.492	0.497	0.477
	Cáncer	0.894	0.845	0.868	0.222
Testículo	Normal	0.630	0.518	0.580	0.511
	Cáncer	0.747	0.713	0.704	0.001
Tiroides	Normal	0.477	0.445	0.410	0.001
	Cáncer	0.689	0.632	0.645	0.014
Útero	Normal	0.542	0.433	0.465	0.456
	Cáncer	0.946	0.936	0.915	0.174
Vejiga	Normal	0.554	0.471	0.466	0.369
	Cáncer	0.967	0.964	0.935	0.020

La partición con el algoritmo de Louvain tiene los mejores puntajes en todas las redes, por lo que esas comunidades son las que se toman en cuenta en el estudio.

Análisis de sobre-representación de GO

Gene Ontology es una base de datos que integra información sobre la caracterización de proteínas y los genes que las codifican en términos de su funcionalidad, localización dentro de la célula e interacciones en diferentes organismos [64]. Está formada por tres grandes conjuntos de datos:

- Función molecular: actividades moleculares de las proteínas y otros productos génicos.

- Componente celular: indica el componente celular donde los productos génicos realizan su función.
- Proceso biológico: vías de señalización y otros procesos funcionales que requieren la actividad coordinada de varios procesos génicos.

Estos conjuntos están conformados, a su vez, por diversos *términos* o procesos organizados en una gráfica dirigida acíclica donde las aristas entre los términos representan relaciones jerárquicas.

Para conocer los términos de GO que están asociados o enriquecidos en las comunidades en las redes de co-expresión, se utiliza el análisis de sobre-representación [65]. En él, los valores p para la asociación de cada término en una comunidad de la red, se calculan utilizando la distribución hipergeométrica:

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{i}}$$

Donde N es el número total de genes en la matriz de expresión asociada al tejido, M son los genes de la matriz de expresión que están anotados en el término de GO que se está analizando, n es el tamaño de la comunidad de la red para la cual se está haciendo el cálculo y k son los genes en la comunidad que se encuentran anotados en el término de GO. En este caso, solamente se analizaron los términos en el conjunto de Procesos biológicos.

Para realizar esos cálculos se utilizó la función de `enrichGO` del paquete de R `clusterProfiler` [66], tomando en cuenta solamente las comunidades que contienen más de cinco genes y los términos de GO con más de 10 elementos. Después de corregir con el método de Benjamini and Hochberg para comparaciones múltiples se seleccionaron únicamente los términos con $p_{adj} < 1e^{-10}$.

Descomposición de *coreness*

La descomposición de k-core o *coreness* en las redes de co-expresión se utilizó como medida para aproximar la posición de las comunidades dentro de la red dependiendo si se encuentran hacia el centro o hacia la periferia. El k-core de una red se refiere a un conjunto de vértices donde todos tienen un grado de valor de al menos k [44]. Los nodos con valores bajos de k-core forman conjuntos de nodos con grado bajo, es decir, están unidos por pocas interacciones, mientras que los vértices con alto k-core están más conectados entre ellos y por lo tanto, representan la parte central de la red.

Esta descomposición se obtuvo utilizando la función de `coreness` en la librería de `igraph` para encontrar los valores asociados a cada vértice y posteriormente calcular el `coreness` promedio por comunidad.

3 Resultados. Los perfiles de co-expresión en el tejido normal y cáncer

A través de las diversas herramientas de análisis de datos descritas en la sección anterior, se obtuvo la caracterización de los perfiles de co-expresión en las dos condiciones, sanos y cáncer, y los quince tejidos elegidos. Sus principales propiedades se detallan a continuación.

3.1. La pérdida de co-expresión -*trans* y la pérdida de co-expresión a larga distancia son fenómenos recurrentes en cáncer

Los perfiles de co-expresión de los diferentes tejidos, obtenidos a partir del cálculo de Información Mutua (IM) entre pares de genes, son los *datasets* principales para evaluar si la pérdida de co-expresión inter-cromosómica o -*trans* es un fenómeno presente en diferentes tipos de cáncer.

Debido a los diferentes filtros en el control de calidad y la normalización de los datos de secuenciación de ARN, las matrices de expresión obtenidas contienen un conjunto diferente de genes entre los diferentes tejidos. Sin embargo, para cada tejido, el conjunto de genes es el mismo en la matriz de expresión del fenotipo normal y el de cáncer. Esto significa que el conjunto final de parejas en los perfiles de co-expresión es diferente para cada tejido, pero es el mismo para ambos fenotipos y tiene un intervalo de tamaños que va de los 46.442 millones de parejas en el tejido de piel a los 102.95 millones en pulmón.

Si tomamos en cuenta el conjunto completo de parejas de IM, formadas por los genes presentes en las matrices de expresión en los quince tejidos analizados, encontramos que, en promedio, un 5.33 % está formado por parejas de genes que se encuentran dentro del mismo cromosoma (interacciones intra-cromosómicas o -*cis*). Sin embargo, al tomar conjuntos de diferentes tamaños de valores máximos de IM, (es decir: las mil, diez mil, cien mil, etc., interacciones de IM más fuertes), en el fenotipo de cáncer se observa que los primeros conjuntos y hasta las cien mil interacciones de co-expresión más fuertes (y en algunos casos hasta quinientos mil interacciones), tienen fracciones de interacciones -*cis* cercanas al 1, en todos los tejidos. En cambio, en el fenotipo normal, aunque el porcentaje de interacciones -*cis* es también mayor en los primeros conjuntos de IM llegando a un máximo de 23.32 % en el tejido de tiroides, dicho

porcentaje se mantiene estable y alrededor del 5 % en muchos tejidos. Esto se puede observar en el *heatmap* de fracciones -*cis* en la Figura 13 y en las gráficas de línea de la Figura 14.

Para evaluar si las distribuciones de fracciones intra-cromosómicas en perfiles normales y cáncer en cada tejido son diferentes, se utilizó la prueba de Kolmogorov-Smirnov tomando las fracciones a partir de cada uno de los umbrales de IM establecidos (cada mil hasta diez mil, cada diez mil hasta cien mil, etc.) y hasta las $1e^8$ interacciones. Las pruebas indicaron que el conjunto de fracciones es significativamente diferente (con valores p que van desde $1e^{-16}$ a 0.005) cuando se consideran hasta el primer millón de interacciones y después son cada vez más similares. Estos valores se muestran en la Tabla 3.

Entonces, podemos afirmar que en cáncer las interacciones de mayor co-expresión se dan entre genes del mismo cromosoma, característica que no se observa en los tejidos del fenotipo normal. Por lo tanto, existe una pérdida de co-expresión inter-cromosómica al comparar los perfiles de alta co-expresión en el fenotipo normal y cáncer, en todos los tejidos estudiados.

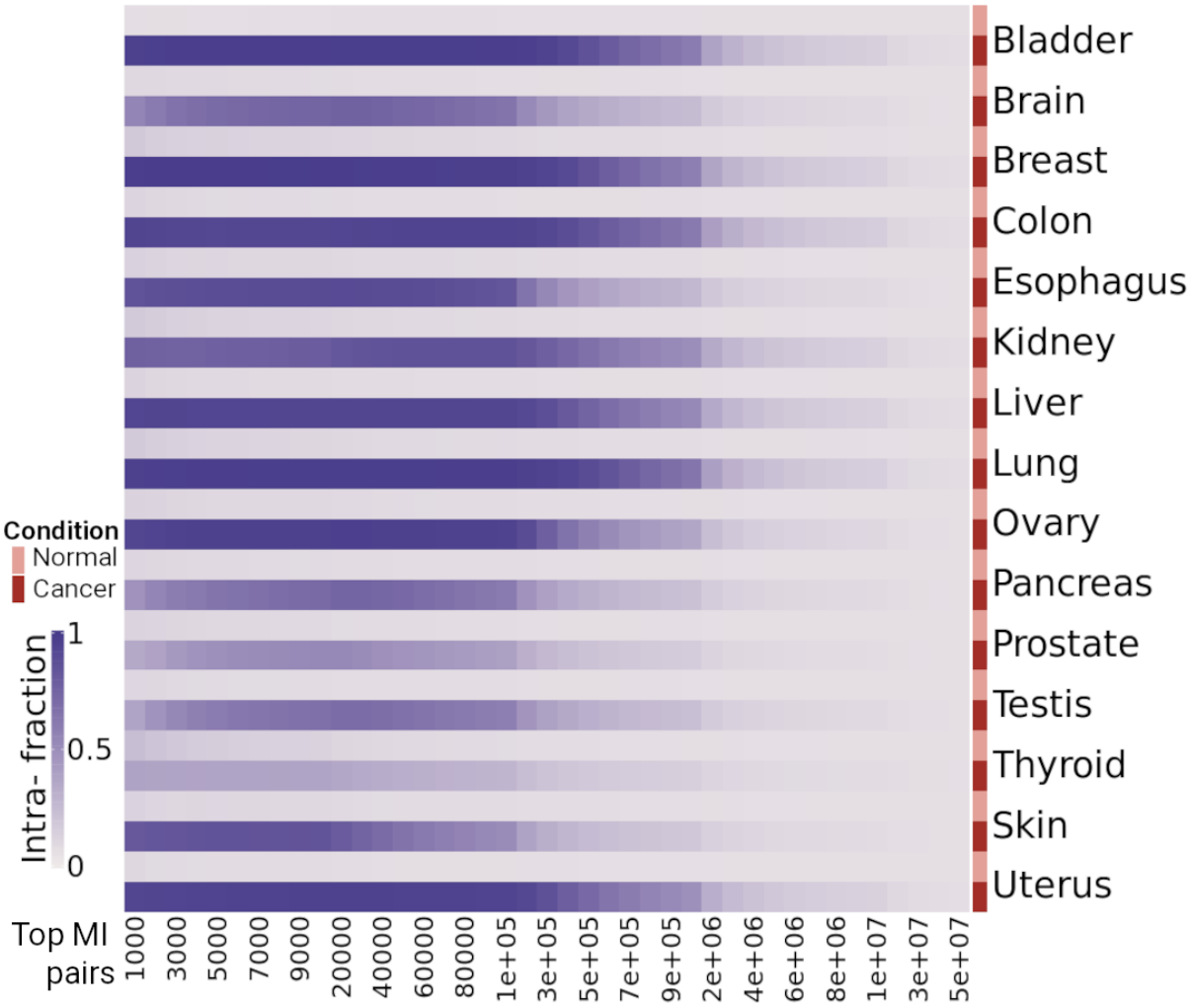

Figura 13: *Heatmap* de fracciones de interacciones *cis*- en diferentes cortes de Información Mutua (IM) en los quince tejidos analizados y ambas condiciones: normal y cáncer.

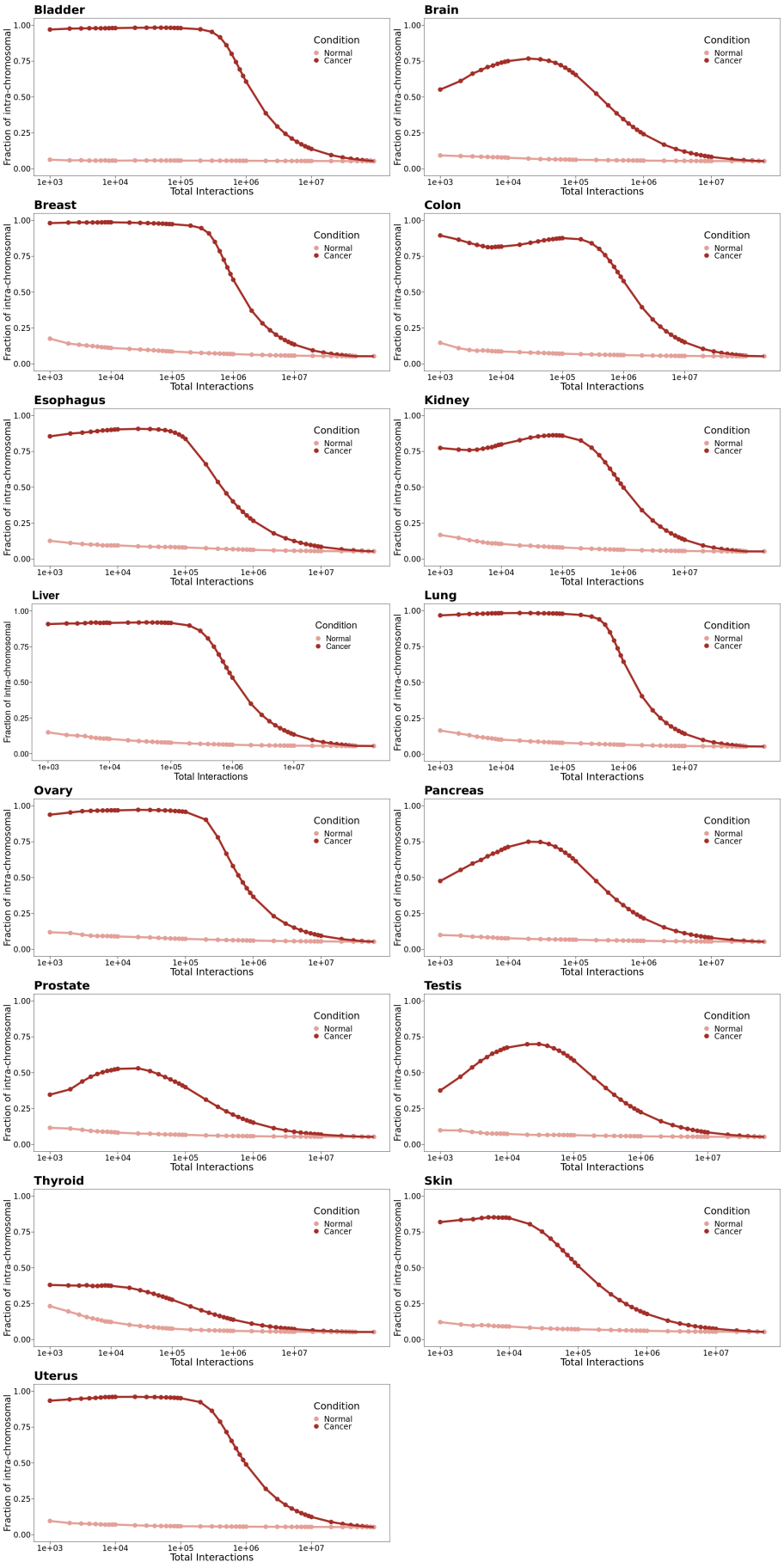

Figura 14: Gráficas de línea de fracciones de interacciones *cis*- en diferentes cortes de Información Mutua (IM) en los quince tejidos analizados y ambas condiciones: normal y cáncer.

Tabla 3: Valores p de pruebas Kolmogorov-Smirnov

Bin inicial	Bladder	Brain	Breast	Colon	Esophagus	Kidney	Liver	Lung
1.00E+03	0.00e+00	3.60e-14	5.36e-13	0.00e+00	2.21e-11	5.36e-13	3.33e-14	5.36e-13
2.00E+03	0.00e+00	8.10e-14	3.40e-13	1.14e-13	1.01e-11	3.40e-13	1.14e-13	3.40e-13
3.00E+03	0.00e+00	5.51e-14	2.35e-14	0.00e+00	3.06e-11	2.35e-14	2.35e-14	2.35e-14
4.00E+03	0.00e+00	0.00e+00	3.30e-13	2.96e-14	1.34e-11	3.30e-13	3.30e-13	3.30e-13
5.00E+03	6.31e-14	0.00e+00	9.95e-13	6.65e-14	4.13e-11	9.95e-13	9.95e-13	9.95e-13
6.00E+03	7.66e-15	0.00e+00	3.27e-12	4.22e-13	1.72e-11	3.27e-12	3.27e-12	3.27e-12
7.00E+03	0.00e+00	3.73e-14	1.01e-11	1.44e-12	5.41e-11	1.01e-11	1.01e-11	1.01e-11
8.00E+03	0.00e+00	1.62e-13	3.06e-11	4.32e-12	1.68e-10	3.06e-11	3.06e-11	3.06e-11
9.00E+03	0.00e+00	7.40e-13	9.24e-11	1.34e-11	5.20e-10	9.24e-11	9.24e-11	9.24e-11
1.00E+04	0.00e+00	2.44e-12	2.77e-10	5.46e-12	1.59e-09	2.77e-10	2.77e-10	2.77e-10
2.00E+04	0.00e+00	8.31e-12	8.23e-10	1.72e-11	6.78e-10	8.23e-10	1.27e-10	1.27e-10
3.00E+04	1.10e-14	2.78e-11	2.42e-09	5.41e-11	2.64e-10	3.85e-10	3.85e-10	3.85e-10
4.00E+04	0.00e+00	9.24e-11	7.07e-09	1.68e-10	8.47e-10	1.16e-09	1.16e-09	1.16e-09
5.00E+04	0.00e+00	2.82e-11	2.04e-08	6.82e-11	3.05e-10	3.46e-09	3.46e-09	3.46e-09
6.00E+04	0.00e+00	9.60e-11	1.03e-08	2.16e-10	9.99e-10	1.03e-08	1.03e-08	1.03e-08
7.00E+04	6.39e-14	3.25e-10	3.01e-08	6.78e-10	3.25e-09	3.01e-08	4.83e-09	4.83e-09
8.00E+04	1.10e-13	1.09e-09	8.72e-08	2.11e-09	1.05e-08	8.72e-08	1.45e-08	1.45e-08
9.00E+04	4.08e-13	3.64e-09	2.50e-07	6.53e-09	3.35e-08	2.50e-07	4.33e-08	4.33e-08
1.00E+05	4.76e-13	1.21e-08	7.08e-07	2.00e-08	1.06e-07	7.08e-07	1.28e-07	1.28e-07
2.00E+05	5.37e-12	3.96e-08	1.98e-06	8.49e-09	3.33e-07	3.71e-07	6.06e-08	3.71e-07
3.00E+05	1.10e-11	1.26e-08	1.07e-06	3.25e-09	1.03e-06	1.07e-06	1.82e-07	1.82e-07
4.00E+05	7.00e-11	4.27e-08	3.03e-06	1.05e-08	3.16e-06	5.40e-07	5.40e-07	5.40e-07
5.00E+05	1.90e-10	1.43e-07	8.49e-06	3.20e-08	9.55e-06	1.58e-06	2.51e-07	1.58e-06
6.00E+05	8.99e-10	4.77e-07	2.34e-05	1.06e-07	4.18e-06	4.57e-06	7.61e-07	4.57e-06
7.00E+05	1.86e-09	1.57e-06	6.34e-05	3.96e-08	1.30e-05	1.30e-05	2.28e-06	2.28e-06
8.00E+05	1.13e-08	5.13e-06	3.64e-05	1.03e-07	3.97e-05	3.64e-05	6.71e-06	6.71e-06
9.00E+05	3.98e-08	1.65e-05	1.00e-04	4.16e-07	1.20e-04	1.95e-05	3.16e-06	1.95e-05
1.00E+06	1.39e-07	5.23e-05	2.70e-04	9.91e-07	3.53e-04	5.57e-05	9.55e-06	5.57e-05
2.00E+06	4.81e-07	1.88e-05	1.56e-04	3.40e-07	1.63e-04	1.56e-04	4.18e-06	2.84e-05
3.00E+06	1.06e-07	6.25e-05	4.29e-04	1.57e-06	5.00e-04	8.31e-05	1.30e-05	8.31e-05
4.00E+06	3.87e-07	1.56e-04	2.38e-04	4.81e-07	1.50e-03	2.38e-04	5.13e-06	3.97e-05
5.00E+06	7.48e-07	6.55e-04	6.70e-04	1.65e-06	4.37e-03	1.20e-04	1.65e-05	1.20e-04
6.00E+06	5.00e-06	1.36e-03	1.84e-03	5.61e-06	2.06e-03	3.53e-04	5.23e-05	3.53e-04
7.00E+06	1.78e-05	6.29e-03	1.02e-03	1.40e-06	6.29e-03	1.02e-03	1.63e-04	1.63e-04
8.00E+06	3.40e-05	2.49e-03	2.87e-03	5.00e-06	1.86e-02	2.87e-03	6.25e-05	5.00e-04
9.00E+06	2.17e-04	8.16e-03	7.86e-03	1.78e-05	5.30e-02	7.86e-03	2.04e-04	1.50e-03
1.00E+07	7.40e-04	2.60e-02	2.07e-02	6.24e-05	1.43e-01	4.37e-03	6.55e-04	4.37e-03
2.00E+07	2.49e-03	7.94e-02	1.23e-02	2.17e-04	7.94e-02	1.23e-02	2.06e-03	2.06e-03
3.00E+07	8.16e-03	2.29e-01	6.29e-03	7.40e-04	2.29e-01	6.29e-03	7.40e-04	6.29e-03
4.00E+07	1.52e-02	4.00e-01	1.86e-02	2.49e-03	4.00e-01	1.86e-02	2.49e-03	2.49e-03
5.00E+07	7.94e-02	1.00e+00	4.14e-02	8.16e-03	1.00e+00	4.14e-02	8.16e-03	8.16e-03
6.00E+07	1.43e-01	1.00e+00	1.43e-01	2.60e-02	1.00e+00	1.43e-01	2.60e-02	2.60e-02
7.00E+07	4.00e-01	NA	2.86e-01	7.94e-02	NA	2.86e-01	7.94e-02	7.94e-02
8.00E+07	1.00e+00	NA	7.71e-01	1.43e-01	NA	6.57e-01	1.43e-01	2.29e-01
9.00E+07	1.00e+00	NA	6.00e-01	6.00e-01	NA	4.00e-01	6.00e-01	6.00e-01
1.00E+08	NA	NA	1.00e+00	1.00e+00	NA	1.00e+00	1.00e+00	1.00e+00
2.00E+08	NA	NA	1.00e+00	1.00e+00	NA	1.00e+00	1.00e+00	1.00e+00

Las pruebas Kolmogorov-Smirnov comparan las fracciones de interacciones -*cis* entre ambas condiciones a partir de un umbral o *bin* de Información Mutua (IM) inicial y hasta el total de interacciones en cada tejido

Continuación. Valores p de pruebas Kolmogorov-Smirnov

Bin inicial	Ovary	Pancreas	Prostate	Testis	Thyroid	Skin	Uterus
1.00E+03	4.22e-13	3.27e-12	3.72e-09	6.65e-14	4.85e-05	3.90e-10	0.00e+00
2.00E+03	2.31e-13	1.44e-12	2.06e-09	3.60e-14	3.77e-05	1.95e-10	0.00e+00
3.00E+03	6.68e-14	5.47e-13	5.86e-09	8.10e-14	2.90e-05	5.72e-10	0.00e+00
4.00E+03	0.00e+00	1.70e-12	3.19e-09	1.32e-14	2.20e-05	1.66e-09	0.00e+00
5.00E+03	0.00e+00	6.22e-13	8.51e-09	0.00e+00	1.64e-05	8.19e-10	6.46e-14
6.00E+03	0.00e+00	1.91e-13	2.51e-08	5.06e-14	1.21e-05	2.42e-09	1.32e-14
7.00E+03	4.65e-14	4.65e-14	1.37e-08	1.51e-13	8.77e-06	6.93e-09	0.00e+00
8.00E+03	1.62e-13	1.62e-13	3.47e-08	6.89e-13	2.16e-05	3.36e-09	5.06e-14
9.00E+03	7.40e-13	7.40e-13	1.08e-07	1.76e-12	4.71e-05	1.03e-08	1.65e-13
1.00E+04	2.44e-12	2.44e-12	5.84e-08	7.70e-12	1.11e-04	4.83e-09	6.89e-13
2.00E+04	8.31e-12	8.31e-12	3.01e-08	2.44e-12	8.32e-05	2.11e-09	1.62e-13
3.00E+04	2.78e-11	2.78e-11	1.45e-08	7.12e-12	6.13e-05	8.47e-10	4.47e-13
4.00E+04	9.24e-11	9.24e-11	6.53e-09	2.78e-11	4.42e-05	3.05e-10	5.83e-14
5.00E+04	3.05e-10	3.05e-10	2.69e-09	9.24e-11	1.01e-04	7.72e-10	6.83e-13
6.00E+04	9.99e-10	9.99e-10	7.69e-09	3.05e-10	7.28e-05	2.33e-10	1.09e-12
7.00E+04	3.25e-09	3.25e-09	2.65e-08	9.99e-10	5.14e-05	1.09e-09	8.29e-12
8.00E+04	1.05e-08	1.05e-08	1.05e-08	3.25e-09	3.52e-05	2.89e-09	2.22e-11
9.00E+04	3.35e-08	3.35e-08	3.35e-08	1.05e-08	2.33e-05	1.21e-08	9.60e-11
1.00E+05	1.06e-07	1.06e-07	1.06e-07	3.35e-08	1.49e-05	3.96e-08	2.55e-11
2.00E+05	3.96e-08	3.96e-08	3.33e-07	1.21e-08	9.09e-06	1.29e-07	7.75e-11
3.00E+05	1.29e-07	1.29e-07	1.03e-06	3.96e-08	2.46e-05	4.16e-07	3.10e-10
4.00E+05	4.16e-07	4.16e-07	3.16e-06	1.29e-07	6.06e-05	1.33e-06	7.44e-10
5.00E+05	1.33e-06	1.33e-06	1.33e-06	4.16e-07	3.70e-05	4.18e-06	3.69e-09
6.00E+05	4.18e-06	4.18e-06	4.18e-06	1.33e-06	1.03e-04	1.30e-05	8.81e-09
7.00E+05	1.57e-06	1.30e-05	1.30e-05	4.18e-06	2.44e-04	5.13e-06	4.27e-08
8.00E+05	5.13e-06	3.97e-05	3.97e-05	1.30e-05	6.50e-04	1.65e-05	1.13e-08
9.00E+05	1.65e-05	1.65e-05	1.20e-04	5.13e-06	1.50e-03	5.23e-05	3.98e-08
1.00E+06	5.23e-05	5.23e-05	3.24e-04	1.65e-05	3.67e-03	1.63e-04	1.39e-07
2.00E+06	1.63e-04	1.63e-04	1.41e-04	5.23e-05	2.50e-03	5.00e-04	4.81e-07
3.00E+06	6.25e-05	5.00e-04	5.00e-04	1.88e-05	1.61e-03	2.04e-04	1.65e-06
4.00E+06	2.04e-04	2.04e-04	1.14e-03	6.25e-05	9.61e-04	6.55e-04	4.27e-06
5.00E+06	6.55e-04	6.55e-04	4.37e-03	2.04e-04	2.76e-03	2.06e-03	1.88e-05
6.00E+06	2.06e-03	2.06e-03	9.69e-03	6.55e-04	6.02e-03	6.29e-03	5.00e-06
7.00E+06	6.29e-03	6.29e-03	3.36e-02	2.06e-03	3.58e-03	1.86e-02	1.78e-05
8.00E+06	1.86e-02	1.86e-02	1.86e-02	6.29e-03	1.26e-02	8.16e-03	6.24e-05
9.00E+06	5.30e-02	5.30e-02	5.30e-02	1.86e-02	2.19e-02	2.60e-02	2.17e-04
1.00E+07	1.43e-01	1.43e-01	1.43e-01	5.30e-02	1.34e-02	7.94e-02	7.40e-04
2.00E+07	3.57e-01	7.94e-02	2.86e-01	2.60e-02	7.95e-03	2.29e-01	2.49e-03
3.00E+07	2.29e-01	2.29e-01	6.57e-01	7.94e-02	3.36e-02	6.00e-01	8.16e-03
4.00E+07	4.00e-01	6.00e-01	1.00e+00	1.43e-01	1.86e-02	1.00e+00	1.52e-02
5.00E+07	1.00e+00	1.00e+00	1.00e+00	4.00e-01	3.85e-02	1.00e+00	7.94e-02
6.00E+07	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.13e-01	NA	2.29e-01
7.00E+07	NA	NA	NA	1.00e+00	4.76e-02	NA	4.00e-01
8.00E+07	NA	NA	NA	NA	2.29e-01	NA	1.00e+00
9.00E+07	NA	NA	NA	NA	4.00e-01	NA	1.00e+00
1.00E+08	NA	NA	NA	NA	1.00e+00	NA	NA
2.00E+08	NA	NA	NA	NA	1.00e+00	NA	NA

Las pruebas Kolmogorov-Smirnov comparan las fracciones de interacciones -*cis* entre ambas condiciones a partir de un umbral o *bin* de Información Mutua (IM) inicial y hasta el total de interacciones en cada tejido

El perfil de co-expresión de cáncer de mama sugería que las interacciones de mayor co-expresión se daban entre genes cercanos dentro de los cromosomas [32] y efectivamente, para los cuatro subtipos principales de cáncer de mama reportamos un decaimiento en los valores de co-expresión con respecto a la distancia física entre parejas de genes medida en pares de bases [40]. Este fenómeno se presenta también en los tejidos aquí analizados.

La distancia en pares de bases está definida solamente para parejas de genes dentro del mismo cromosoma. Por lo tanto, para evaluar el decaimiento se seleccionaron solamente las interacciones intra-cromosómicas o *-cis* y se agruparon en conjuntos de mil interacciones, tomadas de menor a mayor distancia. La Figura 15 muestra los valores promedio de IM en dichos conjuntos contra sus respectivos valores promedio de distancia en los quince tejidos.

En los conjuntos de datos del fenotipo de cáncer, los pares de genes con menor distancia presentan valores de co-expresión más altos y, a medida que la distancia incrementa, los valores de IM disminuyen, llegando a establecerse en una meseta. En el fenotipo normal no se observa dicho decaimiento; los valores de co-expresión son también más altos para distancias muy cortas pero se establecen en un valor regular de forma muy rápida.

Para evaluar dicho decaimiento se ejecutaron pruebas de Mann–Whitney–Wilcoxon entre cada uno de los conjuntos de mil interacciones en el mismo tejido y el mismo fenotipo. Los valores p obtenidos se muestran en la Figura 16. En el tejido normal, solamente los primeros conjuntos de mil interacciones son significativamente diferentes del resto, mientras que en cáncer los conjuntos vecinos son similares entre ellos, pero significativamente diferentes a los demás.

Por lo tanto, además de pertenecer al mismo cromosoma, los pares de genes con alta co-expresión en el fenotipo de cáncer se encuentran físicamente cercanos entre ellos y esto no se observa en el tejido normal, lo que nos habla de una pérdida de co-expresión a larga distancia presente en todos los tejidos de cáncer estudiados.

Ahora, para conocer si existen regiones con mayor aglomeración de interacciones de alta co-expresión en los cromosomas e identificar si dichas regiones aparecen en múltiples tejidos, decidimos extraer las citobandas que cuentan con más de la mitad de sus enlaces posibles en los cien mil valores más altos de IM. Las citobandas son una buena aproximación para delimitar vecindarios en los cromosomas ya que representan regiones con configuraciones particulares de la cromatina [61]. Las regiones con alta densidad de enlaces *-cis* que además están compartidas en, al menos, cinco tejidos, se muestran en la Figura 17.

En el conjunto de datos del fenotipo normal, solamente hay una región con las características previamente descritas: la región q11.221 en el cromosoma Y. Esta región contiene cinco genes: *DDX3Y*, una helicasa de ARN; *NLGN4Y*, miembro de la familia de las neuroliginas; *TMSB4Y*, gen asociado a la región específicamente masculina del cromosoma Y; *UTY*, un antígeno menor de histocompatibilidad y una peptidasa llamada *USP9Y* [67].

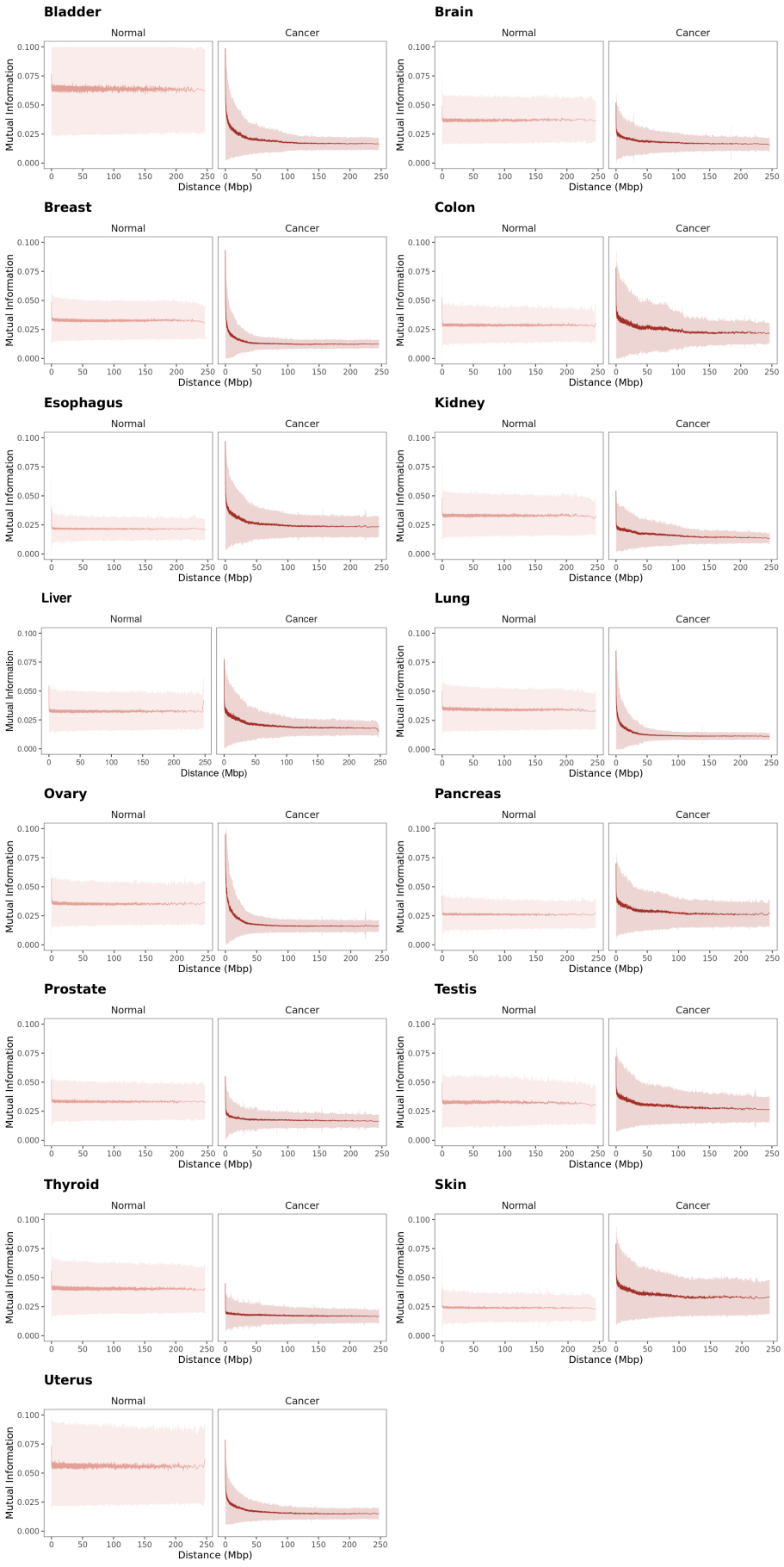

Figura 15: Distancia promedio en pares de bases contra valores promedio de IM para *bins* de mil interacciones intra-cromosómicas.

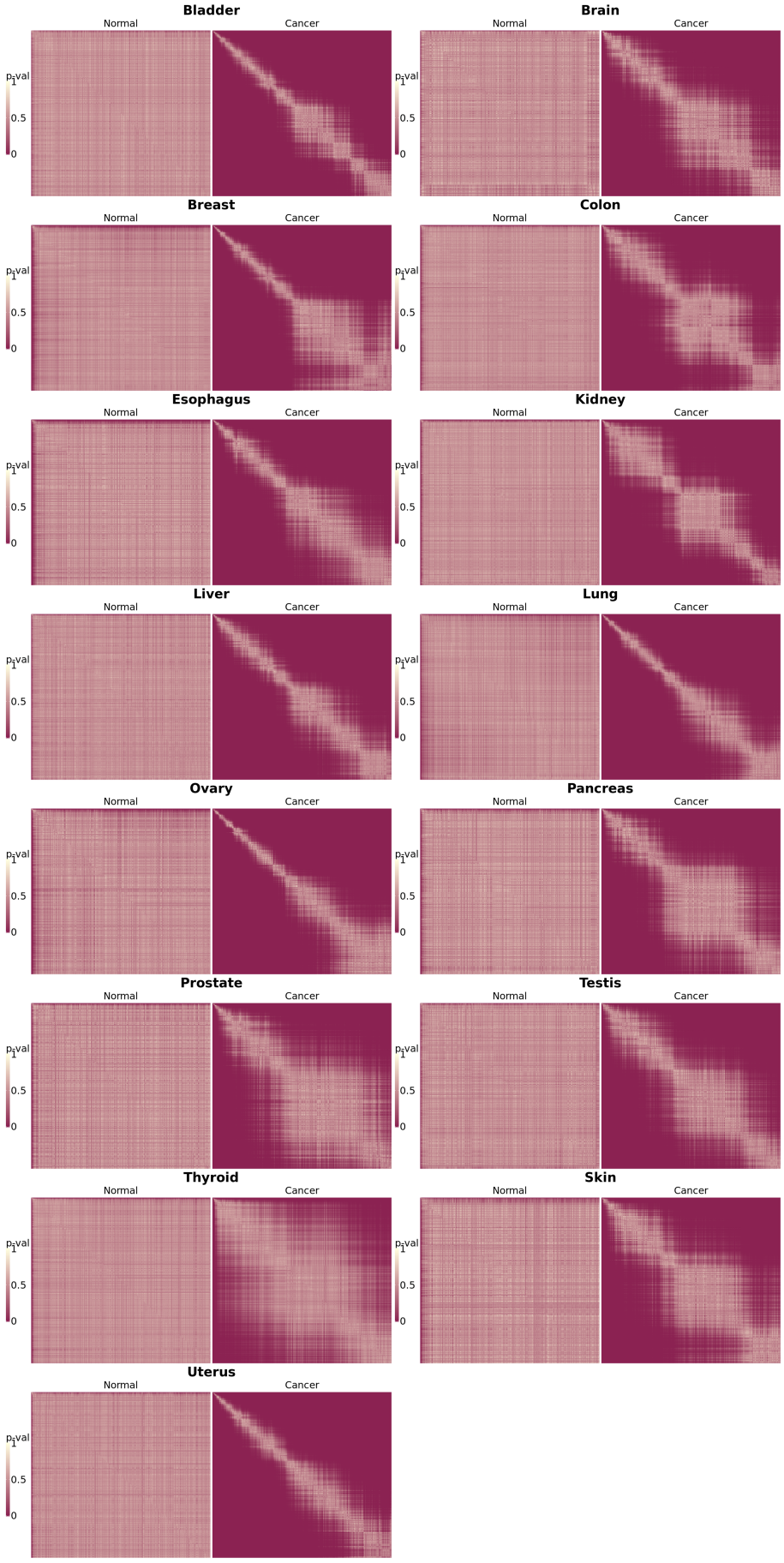

Figura 16: Valores p resultado de pruebas de Mann–Whitney–Wilcoxon comparando las distribuciones de IM en *bins* de mil interacciones que aparecen en la Figura 15.

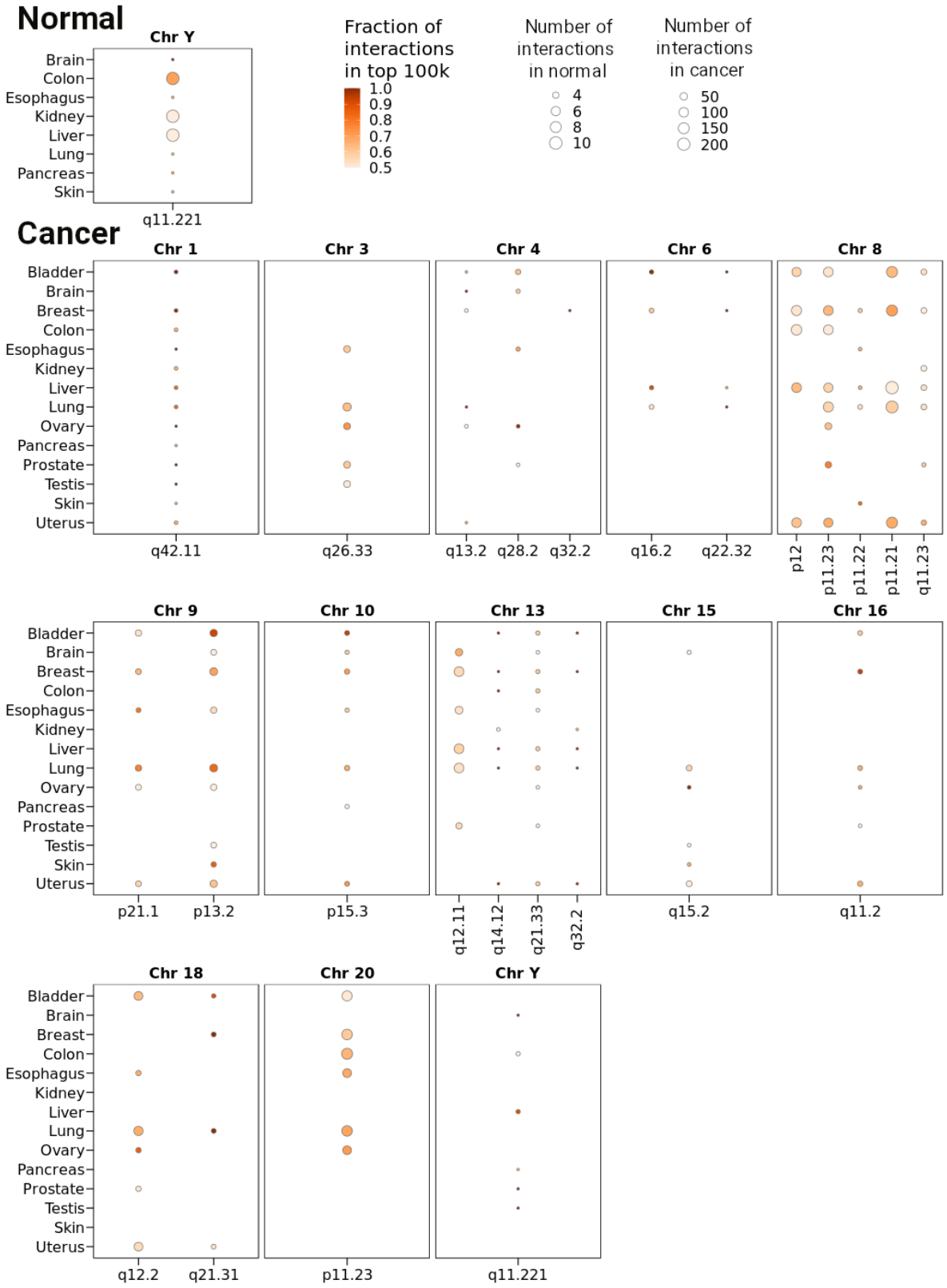

Figura 17: Citobandas con más de la mitad de sus enlaces posibles presentes en el top cien mil de valores de IM y compartidas en, al menos, cinco tejidos.

En cambio, en cáncer encontramos 25 regiones similares, con un máximo de cinco en el cromosoma 8 y en todos los tejidos, a excepción de tiroides. El cromosoma 8 presenta tres citobandas contiguas de alta densidad de co-expresión (p11.23, p11.22 y p11.21) en tejido de mama, hígado y pulmón. Esta región ha sido asociada con amplificación en número de copias en múltiples tipos de cáncer, con *ZNF703*, *FGFR1* y *PLPP* propuestos como genes *drivers* [68] y también con una región de corte asociada al gen *ADAM32* [69]. La región con más repeticiones es la 1q42.11, presente en trece tejidos, seguida de q21.33 en el cromosoma 13 repetida en diez tejidos. Estas dos regiones no han sido previamente asociadas al fenotipo tumoral.

Para conocer la relevancia asociada a encontrar estas regiones de alta densidad de co-expresión, se comparó el número de enlaces intra-citobanda en estas regiones con las encontradas dado un modelo nulo que asigna las interacciones dentro de cada cromosoma al azar. Las regiones compartidas alcanzaron un valor de $p < 1e^{-16}$.

Esto indica que las interacciones de alta co-expresión en el fenotipo de cáncer se encuentran físicamente cercanas, formando vecindarios con alta densidad de enlaces, y además, algunas de ellas aparecen en múltiples tejidos de cáncer estudiados.

3.2. Las redes de co-expresión del fenotipo normal y el tejido de cáncer tienen diferentes características topológicas

Las cien mil interacciones con valores más altos de IM se utilizaron para construir redes de co-expresión. En las redes de cáncer, como se mencionó previamente, a este punto de corte la mayoría de los enlaces unen genes dentro del mismo cromosoma, mientras que las redes de tejido normal están formadas por un componente con enlaces -*trans*.

La Tabla 4 presenta las principales características de las redes, mientras que la Figura 18 muestra una representación visual de las redes de pulmón, ovario y útero y la Figura 19 las redes de todos los tejidos.

Si tomamos en cuenta la fracción de interacciones intra-cromosómicas o -*cis* en las redes de cáncer, se pueden identificar dos grupos de tejidos. Por un lado, las redes de vejiga, mama, esófago, riñón, hígado, pulmón, ovario y útero, que tienen más del 83 % de enlaces -*cis* y, por otro lado, las redes de cerebro, páncreas, próstata, piel, testículo y tiroides con menos del 66 %. La estructura de las redes de cáncer de estos dos grupos es diferente cuando se aplica una visualización con *force directed layout*. Sin embargo, en las redes de tejido normal no existe dicha distinción. El promedio de las fracciones de interacciones -*cis* en las redes de cáncer es de 0.7582 ± 0.2327, mientras que las redes de tejido normal tienen una fracción promedio de 0.0716 ± 0.0085.

Tabla 4: Características principales de las redes de co-expresión del fenotipo normal y cáncer

Tejido	Fenotipo	Genes	Fracción *-cis*	Comunidades	Sobre-	Sub-
Cerebro	Normal	8897	0.06249	68		
	Cáncer	8738	0.65396	129	4735	3436
Colon	Normal	10645	0.07128	176		
	Cáncer	8958	0.87639	373	5111	3264
Esófago	Normal	8325	0.08036	85		
	Cáncer	9117	0.83717	139	4580	3856
Hígado	Normal	11952	0.07664	144		
	Cáncer	9622	0.91591	357	4984	3551
Mama	Normal	11902	0.08626	93		
	Cáncer	10155	0.97363	423	5571	3435
Ovario	Normal	9071	0.07302	44		
	Cáncer	7729	0.95778	193	3626	3496
Páncreas	Normal	8057	0.06685	97		
	Cáncer	9659	0.61381	51	4604	4469
Piel	Normal	7370	0.07258	97		
	Cáncer	9841	0.51385	29	5164	4098
Próstata	Normal	9614	0.06704	26		
	Cáncer	8645	0.39908	108	4406	3629
Pulmón	Normal	12469	0.07914	80		
	Cáncer	10116	0.97844	424	6950	2321
Riñón	Normal	10955	0.08052	89		
	Cáncer	9612	0.85921	394	5168	3486
Testículo	Normal	6825	0.06484	93		
	Cáncer	10174	0.58501	73	5605	4128
Tiroides	Normal	14095	0.07634	50		
	Cáncer	10259	0.27775	372	4509	4189
Útero	Normal	11935	0.05935	91		
	Cáncer	9956	0.95163	330	4778	3312
Vejiga	Normal	13221	0.05709	34		
	Cáncer	9403	0.98007	321	4618	1995

Las redes se construyen con las primeras 100 mil interacciones de Información Mutua y difieren en el número de nodos (Genes). Fracción *-cis*: Fracción de enlaces que unen a genes del mismo cromosoma. Comunidades: número de comunidades en la red usando el algoritmo de Louvain (ver Métodos). Sobre-, Sub-: Genes sobre- y sub-expresados en la red ($pval_{adj} < 0.05$) .

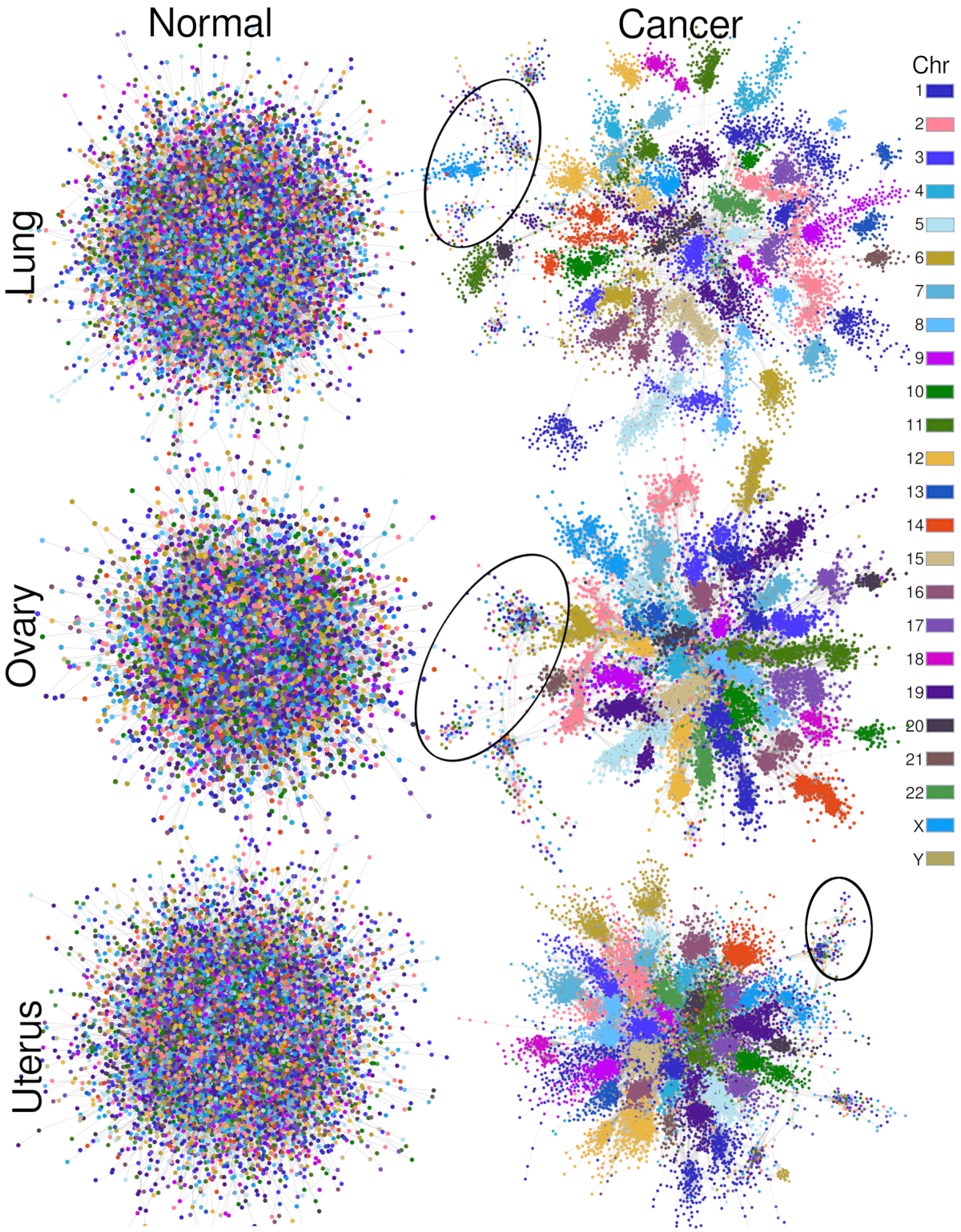

Figura 18: Redes de co-expresión para tejido de pulmón, ovario y útero, formadas por los cien mil pares de valores de IM más altos en el fenotipo normal y los tejidos de cáncer analizados. Se dibujan utilizando *force directed layout*. Los genes están coloreados de acuerdo al cromosoma al que pertenecen. Los círculos en negro identifican comunidades asociadas con el proceso de respuesta inmune adaptativa de Gene Ontology. Las redes para los tejidos restantes se presentan en la Figura 19.

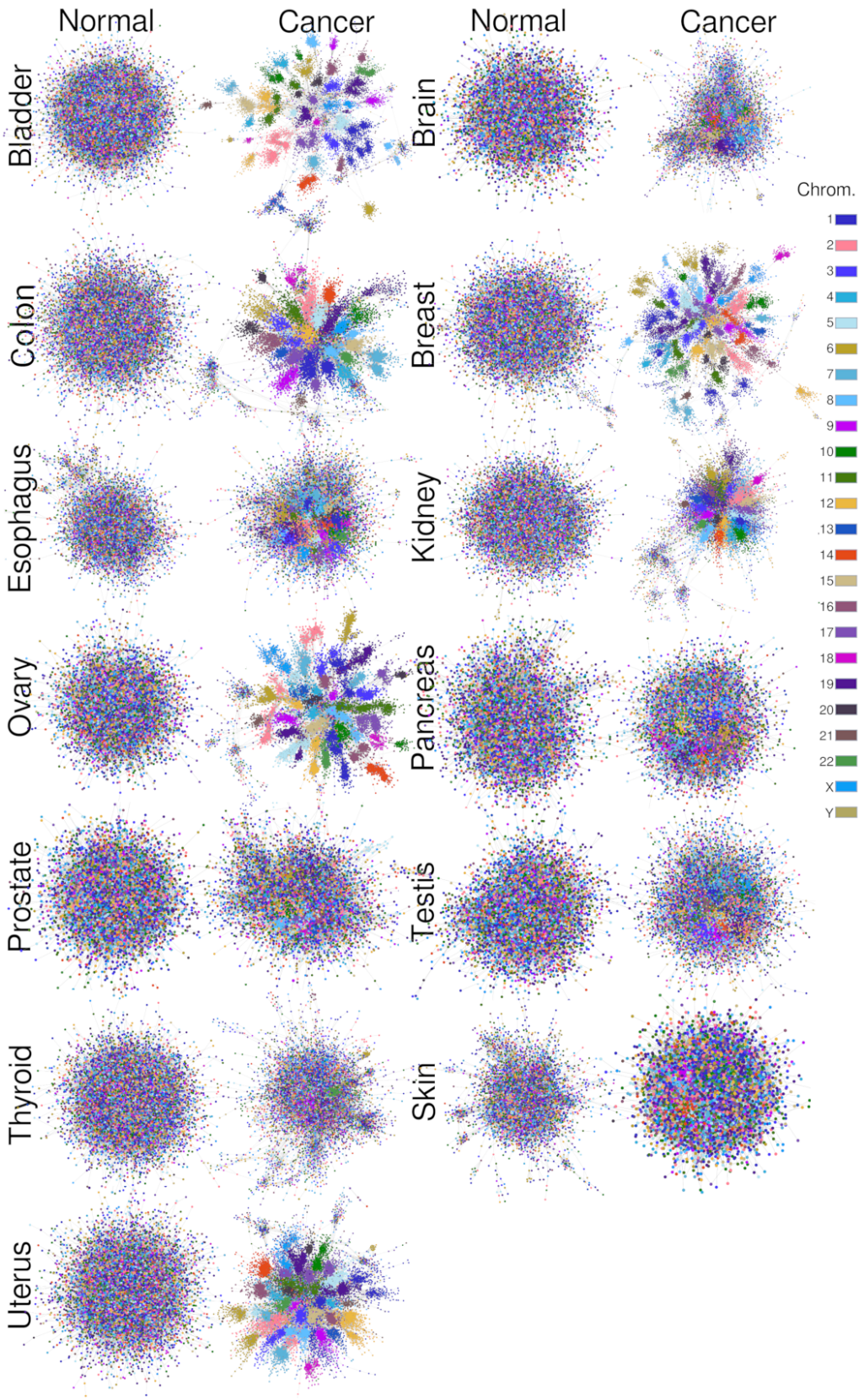

Figura 19: Redes de co-expresión formadas por los cien mil pares de valores de IM más altos en el fenotipo normal y los tejidos de cáncer analizados.

La fracción de interacciones -*cis* no es una métrica suficientemente rigurosa para establecer similitudes entre las redes. Es por eso que evaluamos la presencia de cada interacción en las redes de tejidos normales y cáncer y con dicha información realizamos un agrupamiento jerárquico. Los conjuntos de redes con más de mil enlaces compartidos se muestran en el *heatmap* de la Figura 20A. El conjunto más grande de interacciones compartidas se da en las redes de tejido normal de esófago y piel, como puede observarse en el cuadro de color verde oscuro del *heatmap* de tejido normal. Aunque en las visualizaciones gráficas las redes normales parecen tener una topología similar y además, mantienen fracciones similares de interacciones -*cis*, comparten menos enlaces entre ellas que las redes de cáncer. Es por eso que su *heatmap* se compone de menos renglones y tiene más espacios en blanco. El agrupamiento jerárquico no supervisado divide a los tejidos de cáncer en tres conjuntos principales, donde las redes de esófago y ovario están situadas junto con las redes de menor fracción de interacciones intra-cromosómicas. Las redes de mayor fracción de interacciones -*cis* se dividen en dos conjuntos: uno incluye las redes de mama, vejiga, pulmón y útero y el otro se conforma por las redes de colon, riñón e hígado.

Las gráficas de barra de la Figura 20B muestran los 20 conjuntos con el mayor número de interacciones compartidas. En conformidad con su composición, las redes de co-expresión de tejido normal comparten principalmente enlaces -*trans*, mientras que las redes de tejido de cáncer comparten uniones en el mismo cromosoma. Las redes de cáncer de pulmón, mama y vejiga son las redes con el menor número de interacciones únicas (5,621, 7,894 y 8,607, respectivamente), mientras que la red de cáncer de tiroides presenta el mayor número de interacciones únicas con 67,077.

El promedio de interacciones únicas en las redes de tejido normal es de 87,053 ($\pm 9,025$).

La gráfica de barras resalta un aspecto asociado a la influencia de la fuente de los datos sobre el número de enlaces compartidos en las redes: los primeros catorce conjuntos de interacciones en las redes de tejido normal provienen de la base de datos de USCS Xena. Esta situación no ocurre en los tejidos de cáncer, donde el conjunto más grande contiene redes provenientes de ambas fuentes de datos.

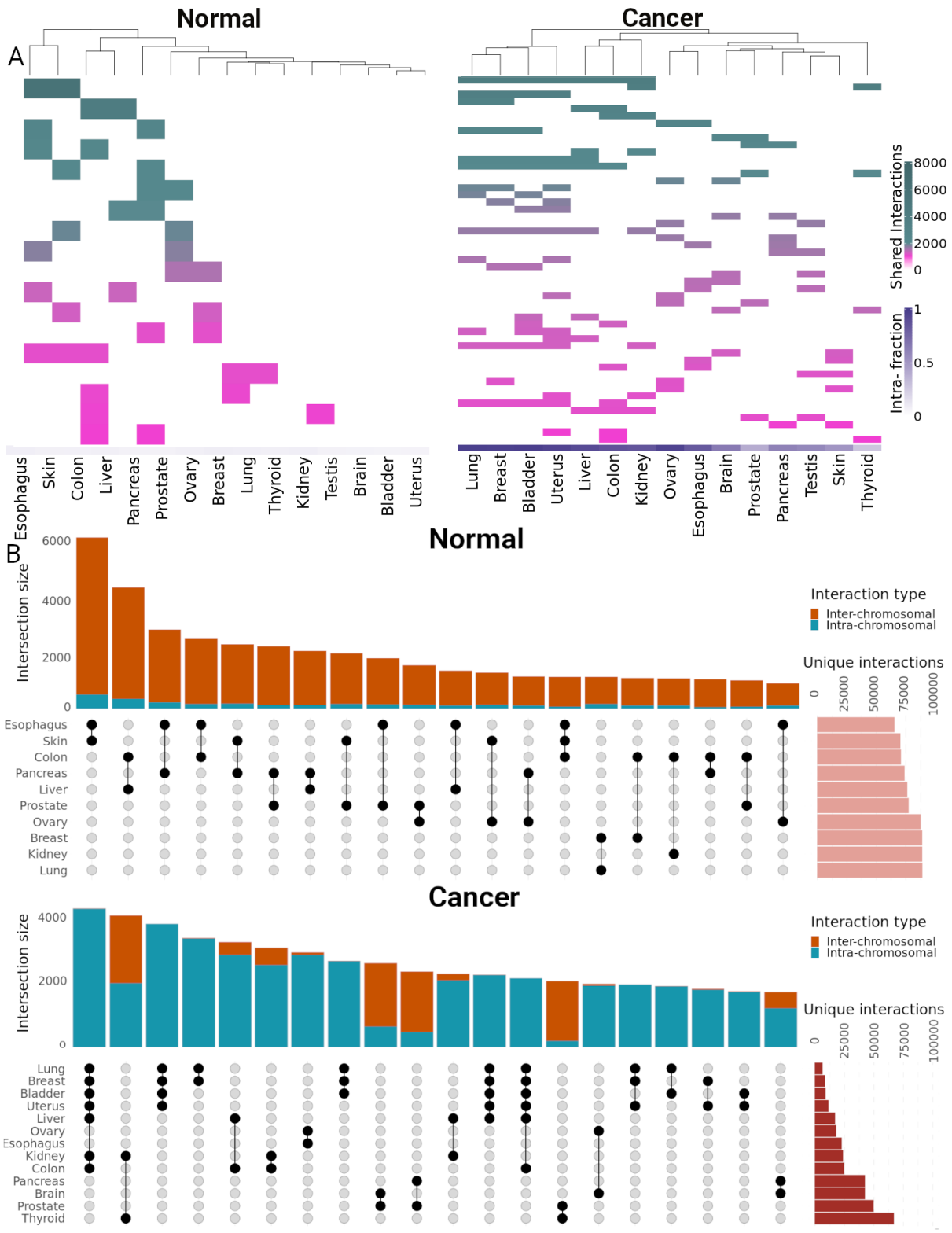

Figura 20: A) Agrupamiento jerárquico de los enlaces compartidos en las redes de co-expresión mostrando conjuntos con más de mil interacciones. B) Gráfica de barras y gráfica *upset* con los veinte conjuntos con más interacciones.

3.3. Las redes de co-expresión en cáncer y tejido normal comparten un *cluster* de genes que codifican para riboproteínas

Para identificar patrones de co-expresión que aparecen en múltiples tejidos en ambos fenotipos, se integró una red de interacciones compartidas en cáncer y un red de interacciones compartidas en el fenotipo normal, definidas como aquellos enlaces presentes en más de 10 tejidos para cada fenotipo. La Figura 21 muestra ambas redes.

Existen 12,802 interacciones presentes en más de 10 redes de co-expresión en cáncer, mientras que en las redes normales solo hay 355 enlaces. La red común de cáncer está compuesta principalmente por enlaces -*cis*, aunque también aparecen algunos conjuntos de interacciones -*trans*. En la intersección de estas dos redes compartidas se encuentran el 92 % de de los genes de la red de tejidos normales y el 59 % de sus enlaces.

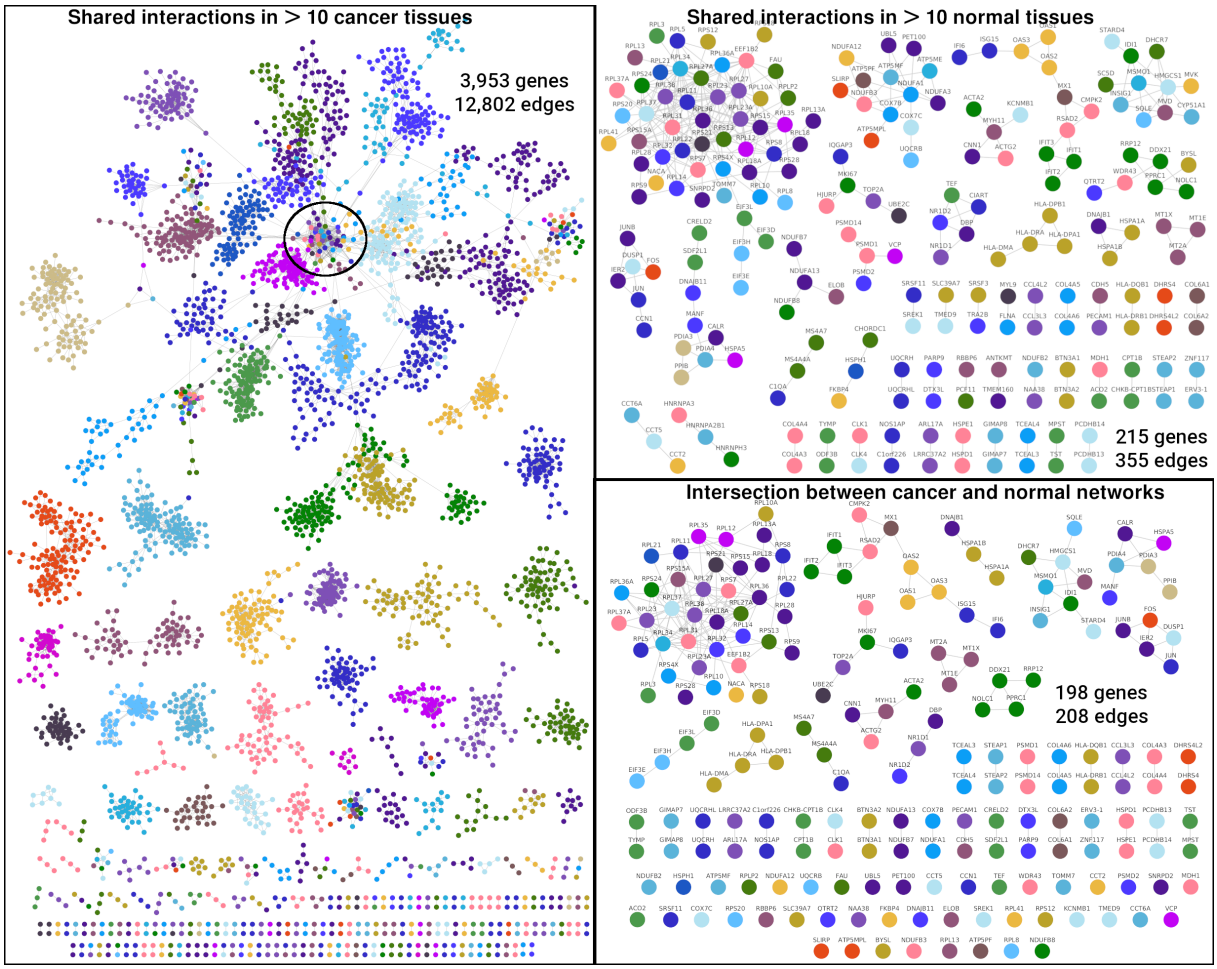

Figura 21: Redes de interacciones compartidas en más de diez tejidos en cada fenotipo y su intersección. Los genes que codifican para proteínas ribosomales forman el componente más grande en la intersección y están señalados con una circunferencia negra en el centro de la red compartida de cáncer.

El componente conexo más grande en la intersección de las redes comunes de tejido normal y cáncer consiste de 39 vértices, que se encuentran justo al centro del componente más grande en la red de cáncer (identificado con un círculo negro en la Figura 21). 37 de estos genes codifican para moléculas en la familia de proteínas ribosomales, con miembros tanto de la subunidad pequeña como de la grande. Los dos genes restantes son *EEF1B2* y *NACA*. De acuerdo a RefSeq [67], el primero está involucrado en la transferencia de tARNs a los ribosomas y el segundo previene la translocación incorrecta de proteínas aberrantes nacientes al retículo endoplasmático.

La presencia de este conjunto de vértices en la intersección de las redes comunes de tejido normal y cáncer indica que las riboproteínas mantienen un patrón de alta co-expresión en todos los tejidos analizados y en ambos fenotipos. Estas proteínas son encargadas de formar el ribosoma, un organelo esencial para todos los organismos y sus genes se encuentran altamente conservados entre diferentes especies, con una homología del 63 % entre genes humanos, de *Drosophila melanogaster*, *C. elegans* y *Saccharomyces cerevisiae* [70].

3.4. La estructura de las redes de co-expresión está asociada a procesos biológicos

En las redes de co-expresión de tejidos de cáncer, el *force directed layout* aplicado para las visualizaciones sugiere la existencia de una estructura modular guiada por el cromosoma donde se localiza cada gen, al menos en las redes con fracciones intra-cromosómicas altas. Dada esta observación, se realizó un análisis de detección de comunidades para identificar la existencia de dicha estructura de forma sistemática.

Una vez identificadas dichas comunidades y con el fin de obtener una mejor caracterización de los enlaces que las componen, se calcularon dos tipos de asortatividades nominales: asortatividad cromosomal y asortatividad de expresión diferencial, esta última solo para las redes de cáncer. Estas métricas nos permiten cuantificar la frecuencia con la que los enlaces en la comunidad unen genes con características comunes [44]. En este caso, el sesgo de los enlaces dentro de las comunidades hacia un solo cromosoma o hacia una sola dirección de log_2 *fold change*.

La distribución de asortatividad cromosomal para las redes de co-expresión del fenotipo normal y los tejidos de cáncer se presentan como gráficas de violín en la Figura 22. Aquí se muestra que las comunidades en las redes de cáncer tiene una tendencia hacia valores positivos cercanos a 1, lo que significa que las interacciones en las comunidades unen habitualmente genes en el mismo cromosoma. Las redes de co-expresión normales presentan una tendencia opuesta, con valores cercanos al -1 en tejidos como vejiga, mama, ovario, próstata, tiroides, etc. y con valores tanto en -1 y 1, en otros tejidos. Esto indica un sesgo a enlaces entre diferentes

cromosomas o una tendencia mixta de comunidades con enlaces ya sean -*cis* o -*trans*, pero no ambos.

Las distribuciones de asortatividad de expresión diferencial se muestran en la Figura 23. En ellas hay mayor variabilidad que la observada en la asortatividad cromosomal. Muchos tejidos tienen comunidades con asortatividad de expresión cercana a 1, lo que significa que las interacciones se dan entre genes sobre- o sub-expresados. Sin embargo, otros tejidos, como páncreas y piel, tienen la mayoría de sus valores alrededor del 0 o esófago, que tiene un número similar de comunidades con asortatividad de expresión diferencial en -1, 0 y 1.

Las comunidades encontradas en las redes fueron asociadas a procesos biológicos anotados en Gene Ontology (GO) mediante un análisis de sobre-representación. Solamente se tomaron en cuenta los procesos con significancia estadística con $p_{adj} < 1e^{-10}$. La fracción de comunidades enriquecidas en procesos es diferente para cada tejido y para cada condición. Para las redes de co-expresión normales, el número total de términos de GO asociados es de 300, mientras que las redes de cáncer están asociadas a 360 procesos biológicos y de estos, 146 términos están en la intersección ($62\,\%$ de los términos en normal y $44\,\%$ de los de cáncer). Para el fenotipo normal, las redes de colon, esófago y piel son las que tienen más procesos enriquecidos con 121, 113 y 97, mientras que las redes de vejiga y útero no tienen procesos de GO asociados. En cáncer, los tejidos de colon, hígado y tiroides son los de mayor número de términos enriquecidos con 135, 128 y 127, respectivamente.

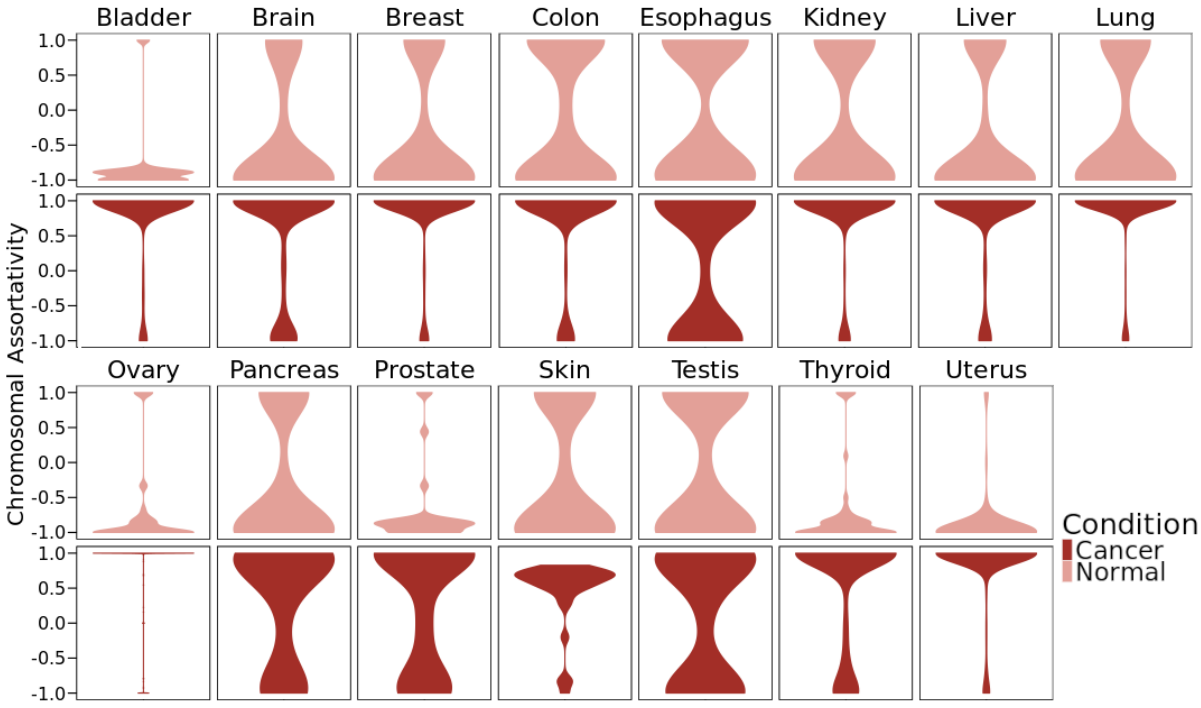

Figura 22: Asortatividad cromosomal en las comunidades de las redes de co-expresión en los quince tejidos en el estudio.

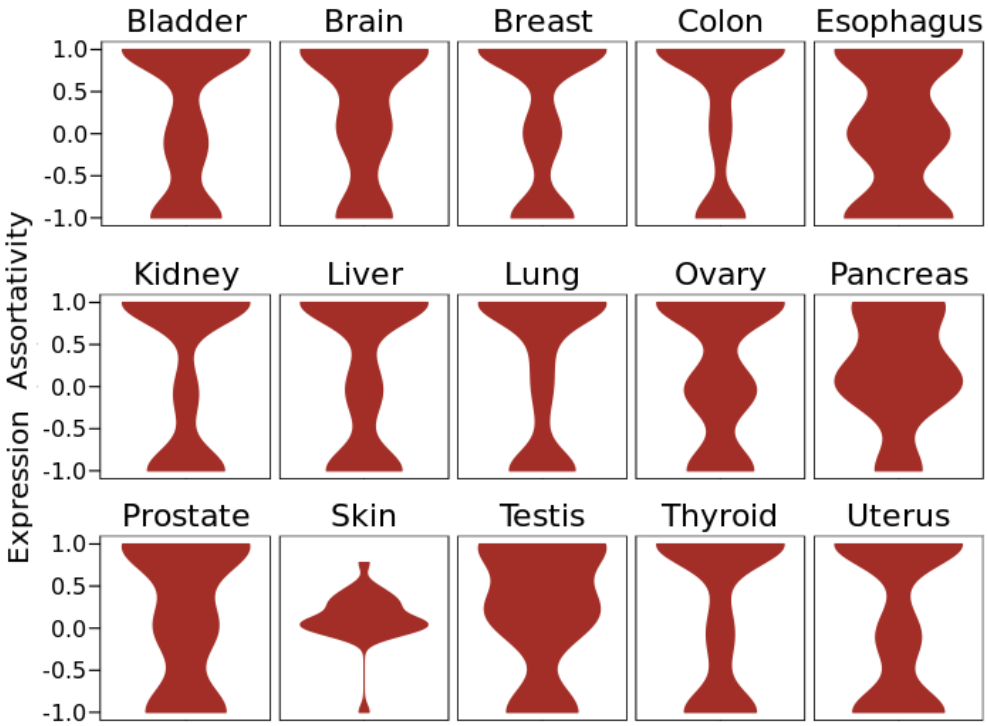

Figura 23: Asortatividad de expresión diferencial en las comunidades de las redes de co-expresión en cáncer.

Al evaluar nuevamente la asortatividad de expresión diferencial en cáncer, pero ahora separando entre comunidades asociadas a procesos de GO y comunidades sin enriquecimientos, se observan diferencias en varios tejidos. El promedio de asortatividad de expresión diferencial en comunidades asociadas a procesos es mayor, en la mayoría de los tejidos, que el de las comunidades no enriquecidas. Las distribuciones están desplegadas en la Figura 24. Esta observación sugiere que los genes asociados a procesos biológicos dentro de las comunidades suelen compartir una dirección en la expresión diferencial de sus genes.

La asortatividad cromosomal también presenta diferencias entre comunidades enriquecidas a procesos de Gene Ontology y comunidades sin asociación biológica, visibles en la Figura 25. En el caso de las redes de cáncer, las comunidades no enriquecidas muestran una asortatividad cromosomal cercana al 1, en particular en redes con altas fracciones de interacciones intra-cromosómicas, mientras que las comunidades enriquecidas tienen un intervalo más amplio de valores de asortatividad cromosomal, con preferencia hacia valores negativos. En cambio, en el fenotipo normal no hay diferencia entre comunidades enriquecidas y no enriquecidas en términos de asortatividad cromosomal ya que la gran mayoría de interacciones en estas redes son inter-cromosómicas.

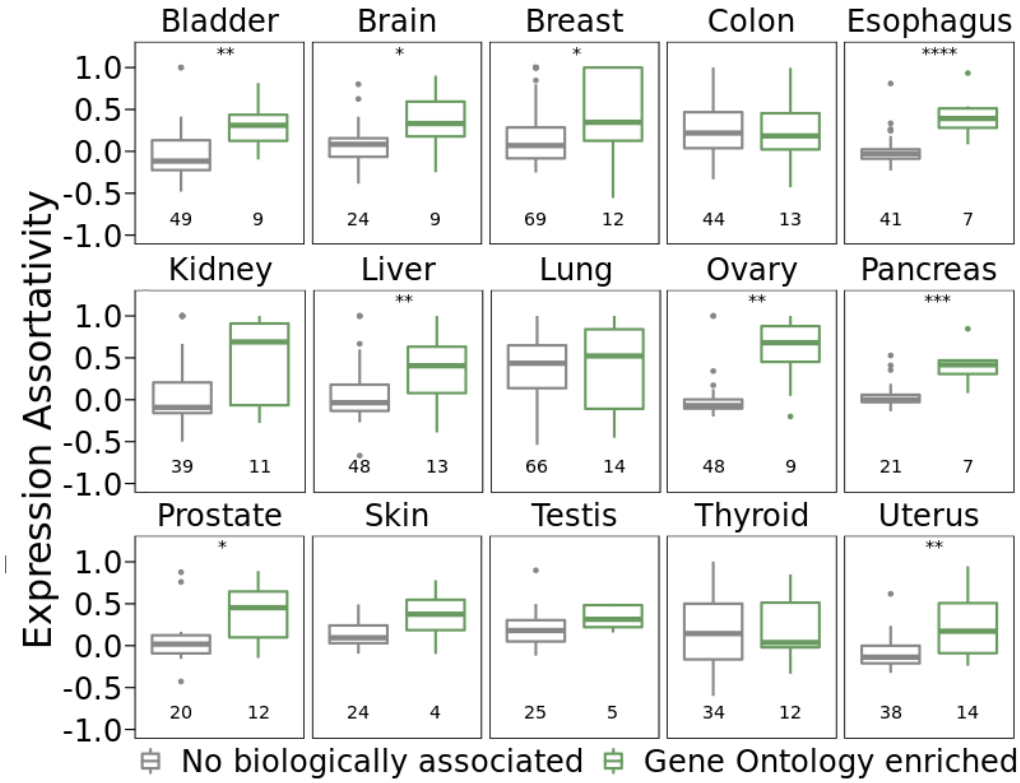

Figura 24: Asortatividad de expresión diferencial para comunidades asociadas a procesos de Gene Ontology en las redes de cáncer.

3.5. Procesos de Gene Ontology (GO) compartidos en las comunidades de las redes de co-expresión

Para identificar los procesos más comunes en el fenotipo normal y cáncer, se creó una red bipartita que une a las comunidades de las redes de co-expresión con sus procesos biológicos de Gene Ontology asociados y se filtraron los procesos con grado $D > 10$. Es decir, se toman en cuenta procesos asociados a más de diez comunidades en cada condición.

La red bipartita del fenotipo normal se muestra en la Figura 26. Los vértices en forma de diamantes representan procesos enriquecidos y los círculos representan las comunidades con colores acordes al tejido al que pertenecen. Esta red tiene dos componentes conexos asociados a la traducción de proteínas y su localización, con procesos de GO tales como biogénesis de ribonucleoproteínas, componentes principales del nucleoplasma (*GO:0022613, ribonucleoprotein complex biogenesis*) e iniciación de la traducción (*GO:0006413, translational initiation*). Varias de estas comunidades tienen genes que codifican para proteínas ribosomales como su gen con mayor *page rank*, como *RPL6* en pulmón o *RRP9* en piel.

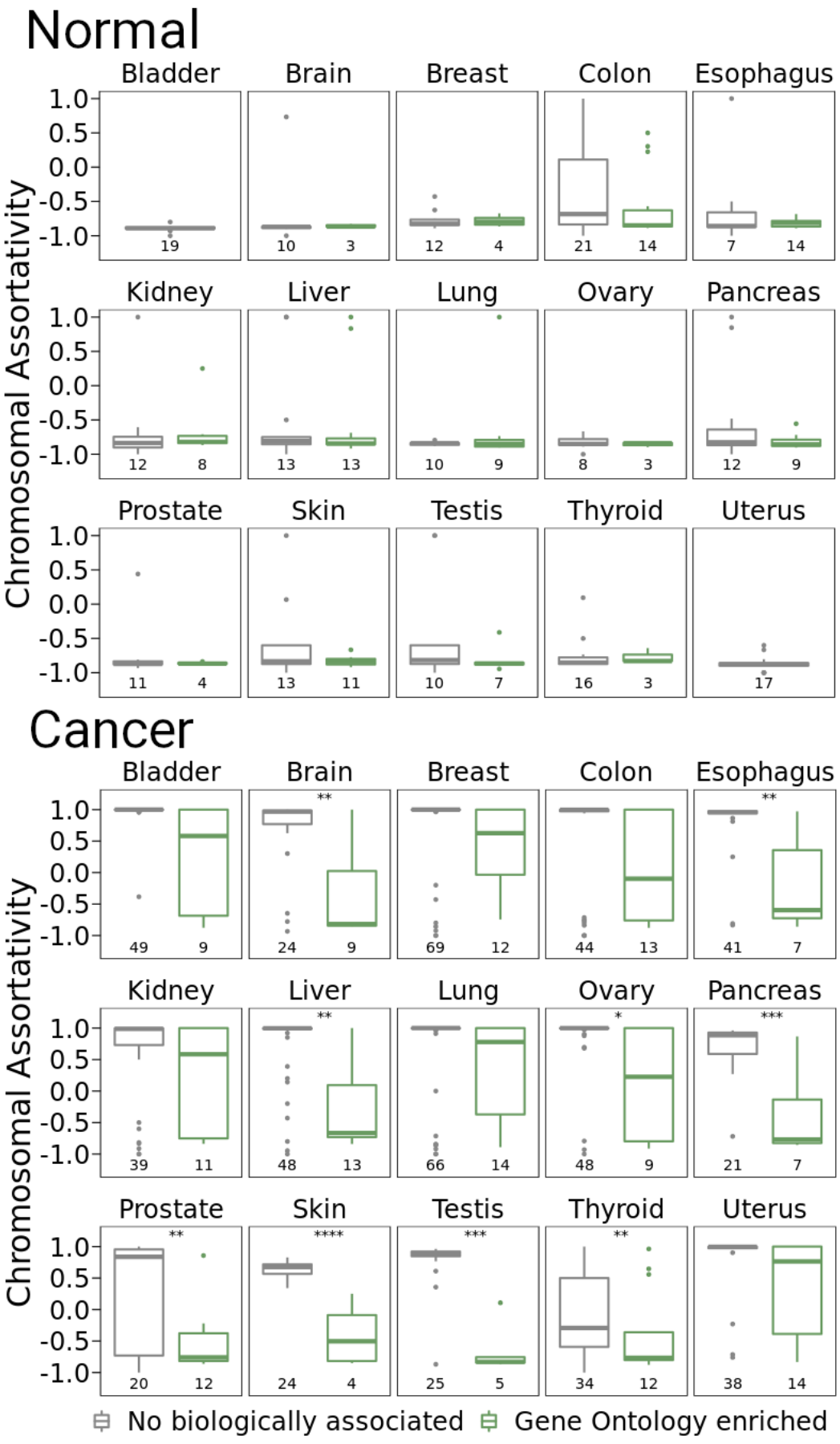

Figura 25: Asortatividad cromosomal en las comunidades de las redes de co-expresión en comunidades asociadas a procesos biológicos de Gene Ontology.

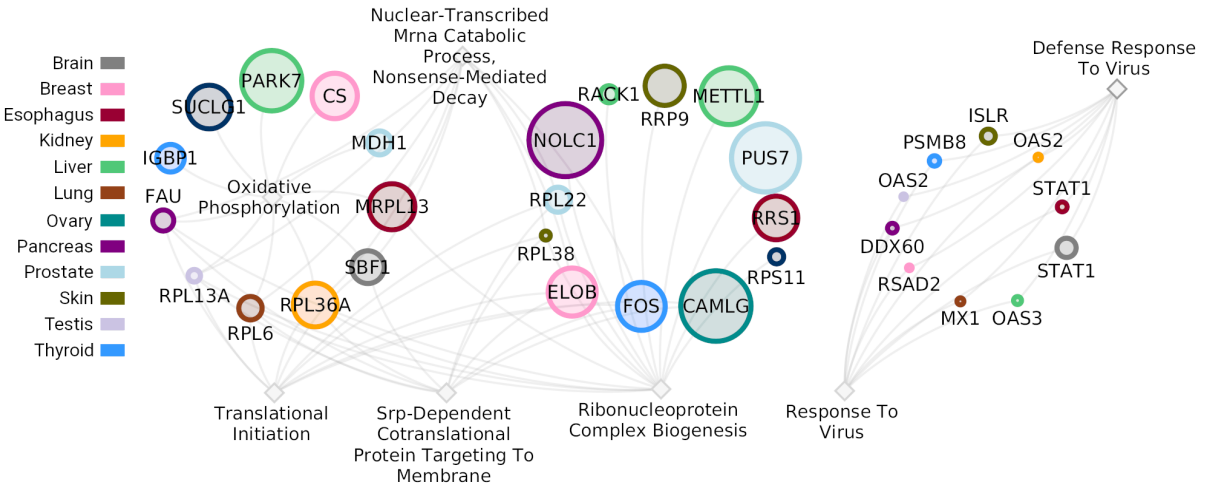

Figura 26: Red bipartita de procesos de Gene Ontology asociados a más de diez comunidades en las redes del fenotipo normal. Los nodos en forma de diamante representan procesos enriquecidos y los círculos representan comunidades con colores de acuerdo al tejido al que pertenecen.

La red bipartita de procesos comunes en cáncer en la Figura 27 está formada de cinco componentes. Uno de ellos se asemeja al componente asociado a la actividad de las riboproteínas en la red normal, donde también aparece iniciación de la traducción (*GO:0006413, translational initiation*) y establecimiento de la localización de las proteínas en la membrana (*GO:0090150, establishment of protein localization to membrane*), lo que nuevamente sugiere una prevalencia del patrón de alta co-expresión entre los genes que codifican para proteínas ribosomales. En este componente existen doce comunidades con un promedio de fracción de interacciones intra-cromosómicas de 0.4285 ± 0.282.

Los genes *HOX*, con las familias A, B, C, y D identificadas en mamíferos, son importantes para el desarrollo embrionario y su expresión mantiene una regulación espaciotemporal asociada a la posición de cada gen en el genoma y de acuerdo a la familia a la que pertenecen [71]. En la red de los procesos de GO comunes en cáncer aparecen doce comunidades con miembros de las familias *HOXA* y *HOXB* en cáncer de vejiga, mama, colon, riñón, hígado, pulmón y útero, desplegadas en la Figura 28. Estas comunidades son completamente *-cis*, con genes ubicados en el cromosoma 7 (*HOXA*) y el cromosoma 17 (*HOXB*).

Las comunidades de la familia *HOXA* presentan una tendencia de sub-expresión en tejidos de mama, colon y útero. En cáncer de mama y colon estos patrones han sido asociados con marcas de metilación aberrantes y silenciamiento epigenético [72, 73], mientras que la sub-expresión de *HOXA10* ha sido reportada en leiomiomas, el tipo más común de tumor uterino benigno. En cáncer de pulmón hay una tendencia mixta y la sobre-expresión de *HOXA10* ha sido asociada a la promoción de crecimiento y metástasis de adenocarcinoma [74] y también ha sido relacionada con la etapa clínica del carcinoma de células escamosas [75].

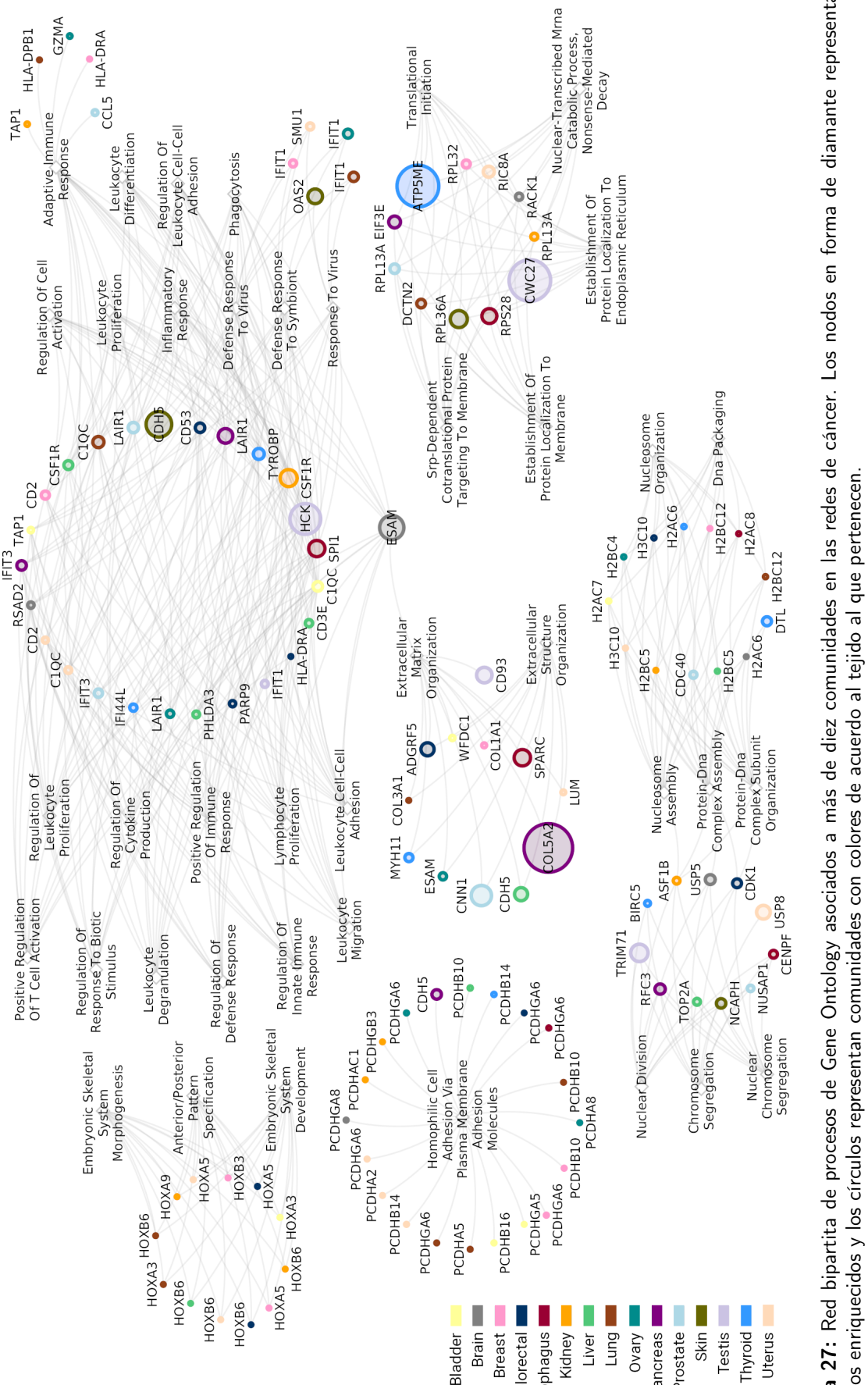

Figura 27: Red bipartita de procesos de Gene Ontology asociados a más de diez comunidades en las redes de cáncer. Los nodos en forma de diamante representan procesos enriquecidos y los círculos representan comunidades con colores de acuerdo al tejido al que pertenecen.

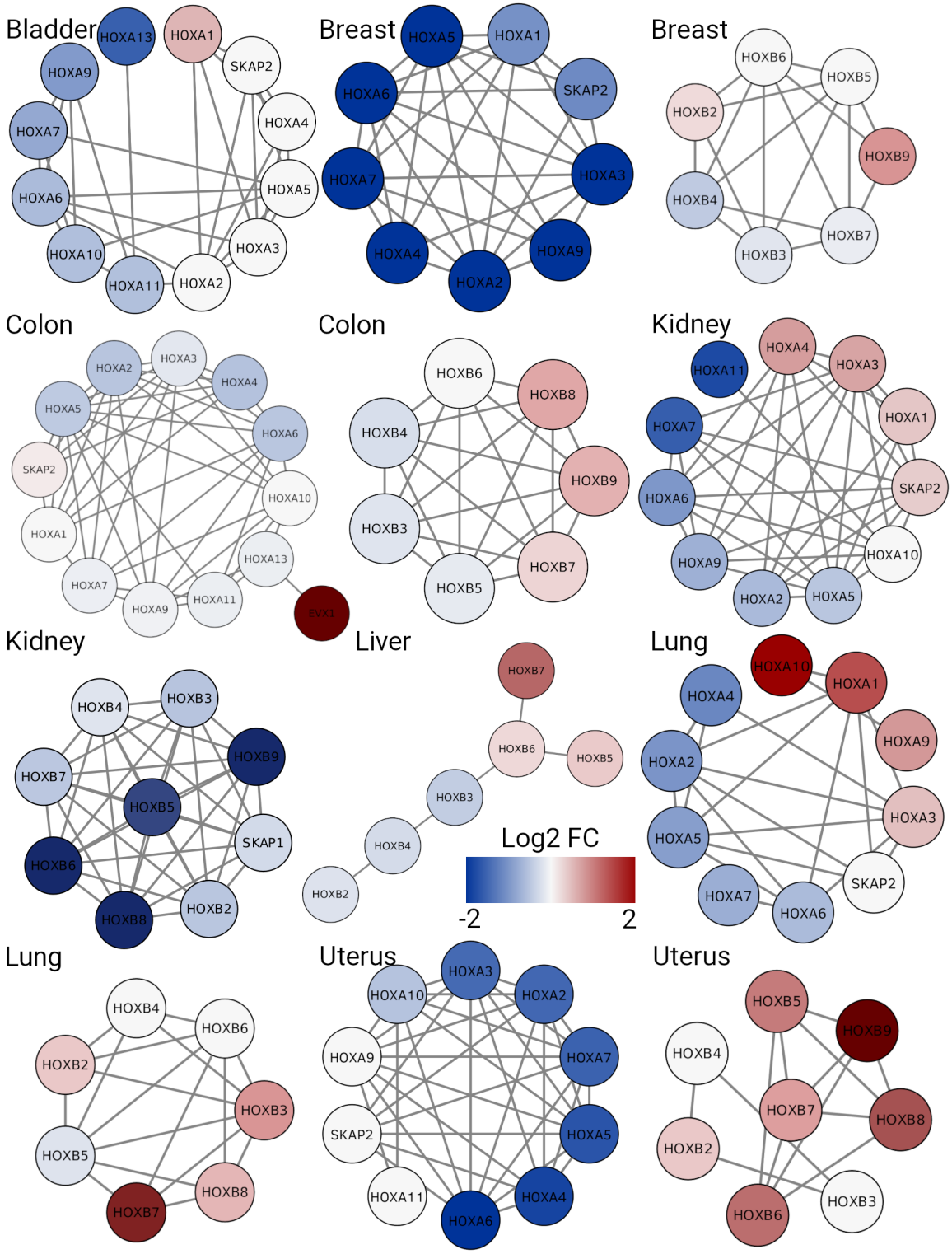

Figura 28: Comunidades asociadas al proceso de Gene Ontology: desarrollo y morfogénesis del sistema esquelético embrionario. El color de los nodos está asignado según los valores de expresión diferencial de los genes.

Los patrones de sobre-expresión que presentan los genes *HOXB* han sido reportados en múltiples tipos de cáncer. La alta expresión de *HOXB9* en cáncer de útero ha sido relacionada con el grado histológico y la metástasis a ganglios linfáticos [76]. En cáncer de colon, la sobre-expresión de los genes *HOXB*, en particular de *HOXB8*, ha sido asociada con la unión de Myc a un *super-enhancer* [77]. En cáncer de hígado se ha observado que la alta expresión de *HOXB7* resulta en proliferación, migración y progresión de cáncer [78]. Además, la expresión aberrante de los genes *HOXB* en cáncer de mama ha sido asociada con desarrollo tumoral y heterogeneidad [79].

Otro módulo está altamente asociado a actividad de histonas, con procesos de GO como ensamblaje del nucleosoma (*GO:0006334, nucleosome assembly*), ensamblaje del complejo proteína-ADN (*GO:0065004, protein-dna complex assembly*) y organización de la cromatina involucrada en la regulación negativa de la transcripción (*GO:0097549, chromatin organization involved in negative regulation of transcription*). El módulo está compuesto por 13 comunidades con una fracción promedio de interacciones intra-cromosómicas de 0.6442 ± 0.2228, con una tendencia a interacciones dentro del cromosoma 6. Los genes con *page rank* mayor pertenecen a la familia de las histonas *H2A* y *H2B*, aunque algunas comunidades incluyen miembros de las familias *H3C*, *H4C*, *H1* y *H4*. La Figura 29 muestra a las comunidades con menos de cien vértices en este módulo y su expresión diferencial asociada. Los genes tienen una tendencia de sobre-expresión en cáncer de vejiga, cerebro, mama, esófago, riñón, hígado, pulmón, ovario y útero. La expresión aberrante de las variantes de histonas ha sido reportada en diferentes tipos de cáncer [80] y su sobre-expresión en cáncer de mama se ha asociado con resistencia a tratamiento y disminución en supervivencia [81, 82].

Genes en la familia de las protocadherinas (*PCDHG* and *PCDHB*) forman otro componente, asociado al término de Gene Ontology: adhesión celular homofílica via moléculas de adhesión en la membrana plasmática (*GO:0007156 , homophilic cell adhesion via plasma membrane adhesion molecules*). Este conjunto incluye a comunidades de doce tipos de cáncer como cerebro, vejiga, mama, riñón, tiroides, ovario, entre otros. Solo tres de las 20 comunidades en este componente no son completamente intra-cromosómicas. Los genes, la gran mayoría dentro del cromosoma 5, presentan un patrón de sub-expresión, especialmente en cáncer de vejiga, mama y útero, el cual también fue reportado previamente por nosotros en el subtipo Luminal A de cáncer de mama [43]. Estas comunidades se muestran en la Figura 30. La sub-expresión de estos genes ha sido asociada con alteraciones epigenéticas, en un mecanismo llamado silenciamiento epigenético de largo alcance (*long-range epigenetic silencing (LRES)*) debido a hipermetilación en diferentes tipos de cáncer incluidos cáncer de mama, colon, riñón y pulmón y promoviendo procesos como migración, proliferación y disrupción de la adhesión celular [83-85].

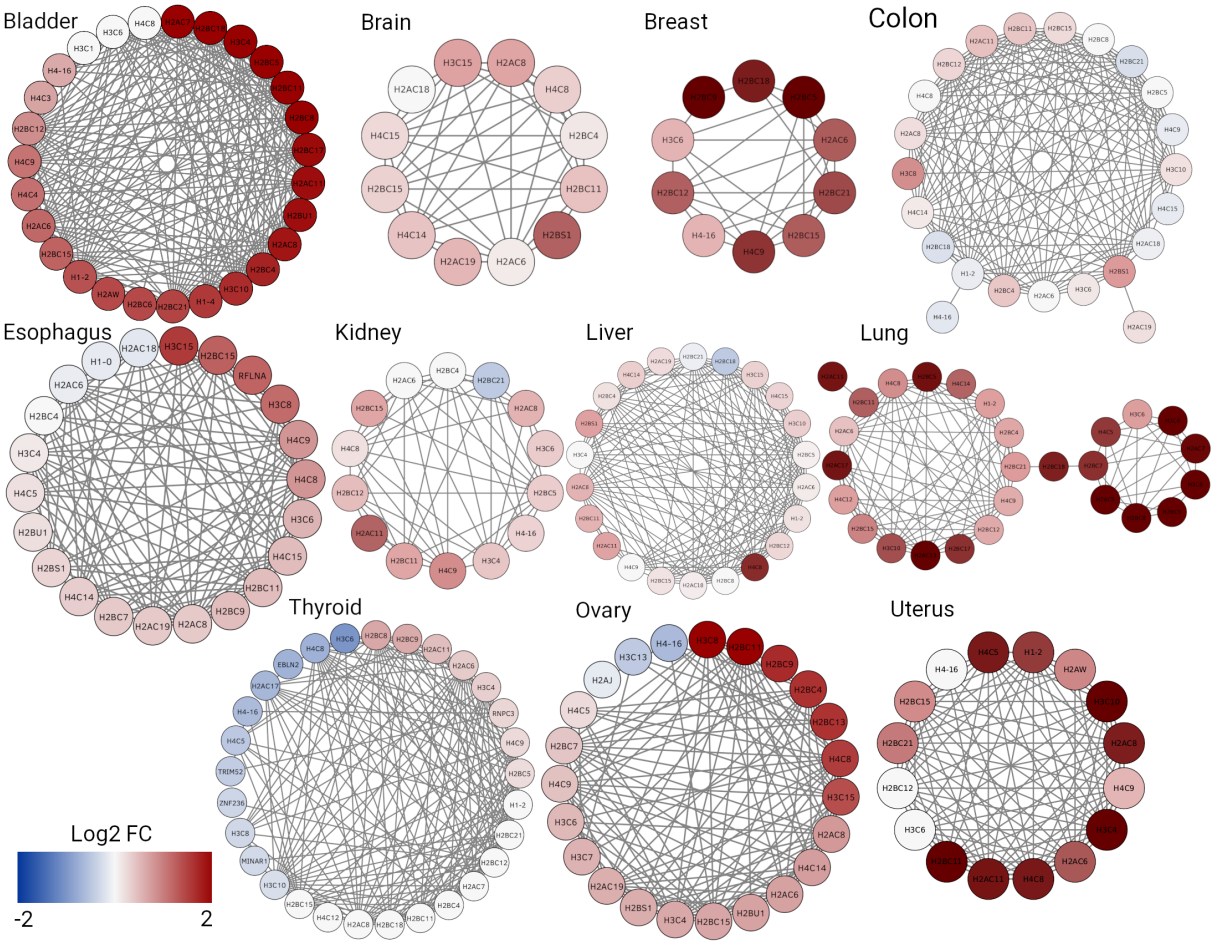

Figura 29: Comunidades asociadas a procesos de GO relacionados con actividad de histonas y con menos de cien nodos. El color de los nodos está asignado según los valores de expresión diferencial de los genes.

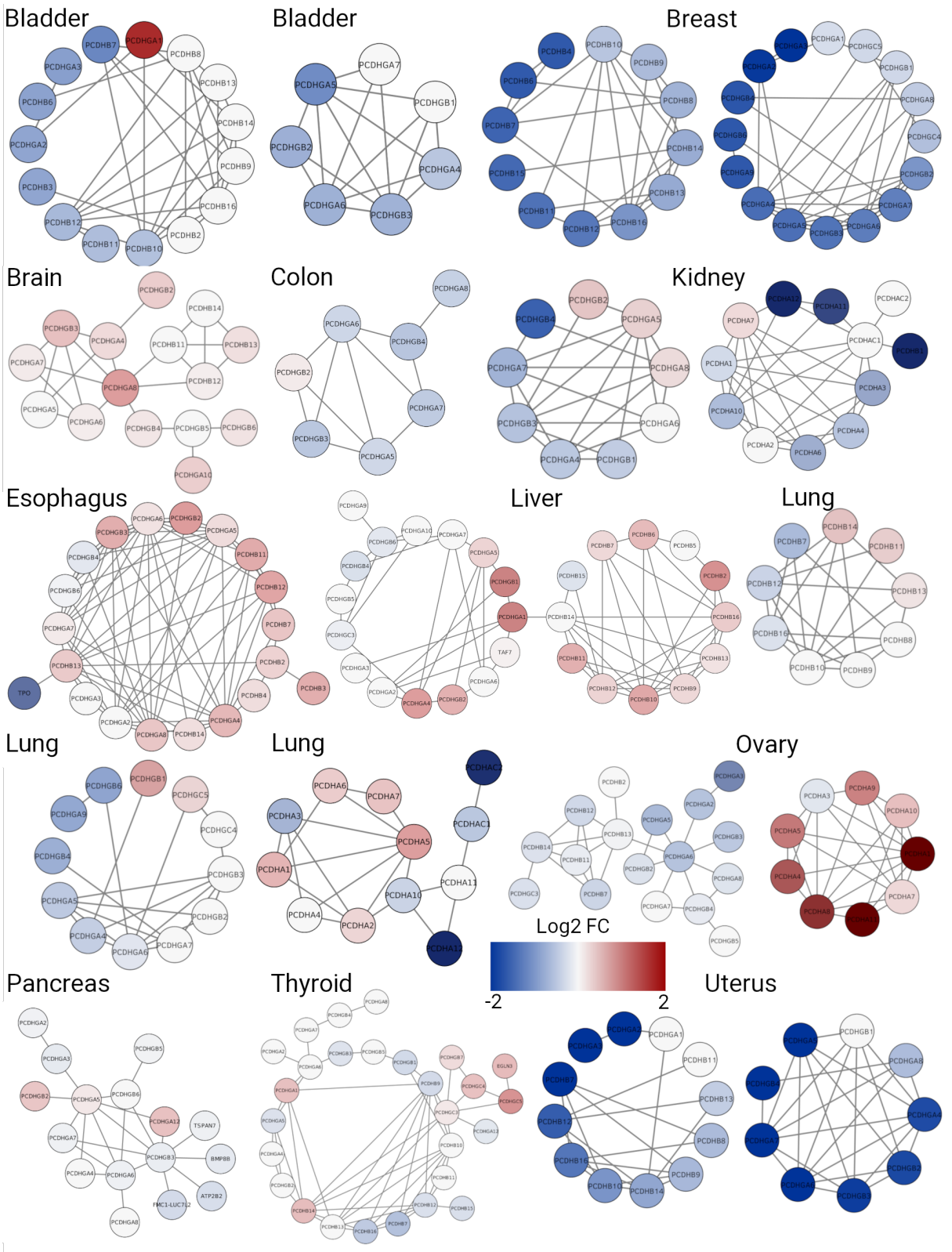

Figura 30: Comunidades asociadas al proceso de Gene Ontology: adhesión celular homofílica via moléculas de adhesión en la membrana plasmática. El color de los nodos está asignado según los valores de expresión diferencial de los genes.

3.6. Procesos asociados únicamente a una condición

A partir de la red bipartita de procesos y comunidades se extrajeron los procesos asociados a las redes de co-expresión de tejido normal o cáncer, que no aparecen en las redes de la otra condición. El conjunto de procesos similares que están únicamente en la red normal aparece en la Figura 31, mientras que los procesos de cáncer se muestran en la Figura 32. Nuevamente, los diamantes representan procesos enriquecidos ($p_{adj} < 1e^{-10}$) y los círculos sus comunidades correspondientes.

En las comunidades del fenotipo normal se identifican cinco conjuntos. El más grande: metabolismo y respiración celular incluye comunidades de ocho diferentes tejidos con procesos como traducción en la mitocondria (*GO:0032543, mitochondrial translation*), regulación de procesos metabólicos de lípidos (*GO:0019216, regulation of lipid metabolic process*) y homeostasis de colesterol (*GO:0042632, cholesterol homeostasis*). El siguiente conjunto, al cual llamamos desarrollo muscular, también incluye ocho tejidos y agrupa procesos como desarrollo de cardiomiocitos (*GO:0055006, cardiac cell development*) y ensamblaje de miofibrillas (*GO:0030239, myofibril assembly*). Los conjuntos restantes son: espermatogénesis, metabolismo energético y transporte nuclear. Todos estos procesos son necesarios para el mantenimiento celular.

En la red de procesos de cáncer el módulo más grande, encerrado en la elipse azul, contiene cincuenta y ocho procesos con comunidades de trece tejidos diferentes, fuertemente relacionados con la respuesta inmune. Este conjunto incluye procesos como: procesamiento y presentación

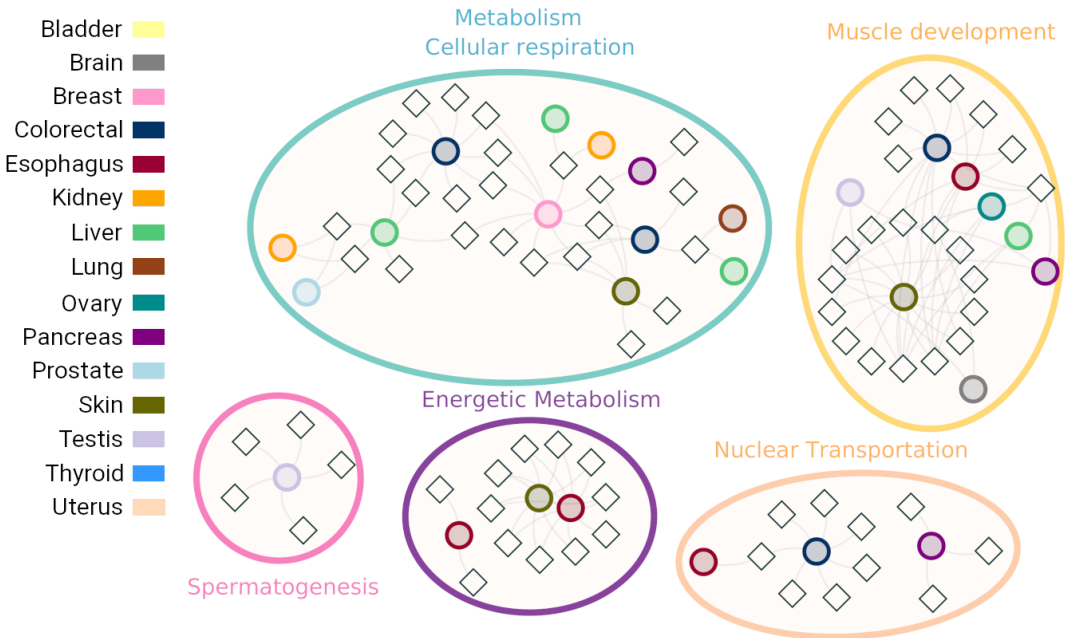

Figura 31: Red bipartita de procesos de Gene Ontology asociados a comunidades que aparecen únicamente en las redes de los tejidos del fenotipo normal. Los nodos en forma de diamante representan procesos enriquecidos y los círculos representan comunidades con colores de acuerdo al tejido al que pertenecen.

de antígenos *(GO:0019882, antigen processing and presentation)*, producción de interleucina 1 *(GO:0032612, interleukin-1 production)*, diferenciación de linfocitos (*GO:0030098, lymphocyte differentiation*) y respuesta a quimiocinas (*GO:1990868, response to chemokine*). Además, encontramos otros dos conjuntos asociados con la activación de la respuesta inmune: activación de células T y respuesta a virus.

La red incluye también otros procesos que se sabe están alterados en cáncer como: migración celular, ciclo celular y organización de la cromatina. En cuanto al conjunto de procesos asociados con la respuesta a estrés por metales iónicos, se han reportado alteraciones en la metilación de metalotioneínas en cáncer de mama [86], riñón [87], pulmón [88, 89] y ovario [90], relacionadas con desarrollo desfavorable de la enfermedad.

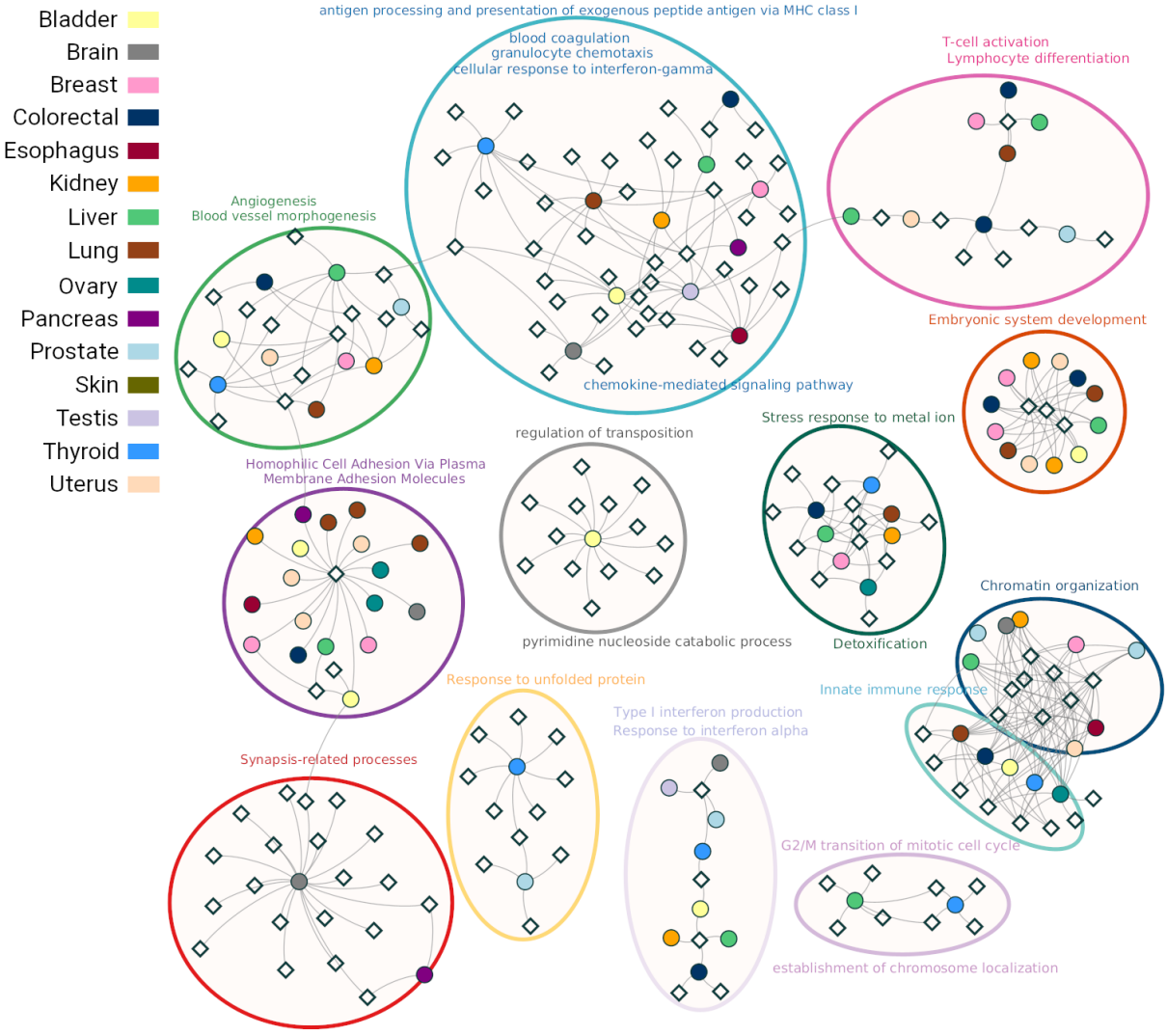

Figura 32: Red bipartita de procesos de Gene Ontology asociados a comunidades que aparecen únicamente en las redes de los tejidos de cáncer. Los nodos en forma de diamante representan procesos enriquecidos y los círculos representan comunidades con colores de acuerdo al tejido al que pertenecen.

3.7. Procesos asociados a un solo tejido

Finalmente, se extrajeron los términos de Gene Ontology (GO) asociados de manera única a un tejido en las redes de co-expresión. Estos términos únicos describen las características individuales de cada red de co-expresión y algunos están relacionados con procesos biológicos ya conocidos de los tejidos.

En la red de cáncer de cerebro la comunidad *CCKBR* está relacionada con procesos como secreción de neurotransmisores, señalización sináptica y regulación del potencial de membrana, entre otros. El conjunto completo de procesos asociados a esta comunidad, que solo están en la red de cerebro, se muestran en la Figure 33A. Los 257 genes que participan en los enriquecimientos de comunidad tienen una tendencia de subexpresión, con un promedio de log_2 de -0.703.

En la red de co-expresión del fenotipo normal de testículo, la comunidad llamada Boll, un gen necesario para el desarrollo de los espermas, está asociado a múltiples procesos específicos de

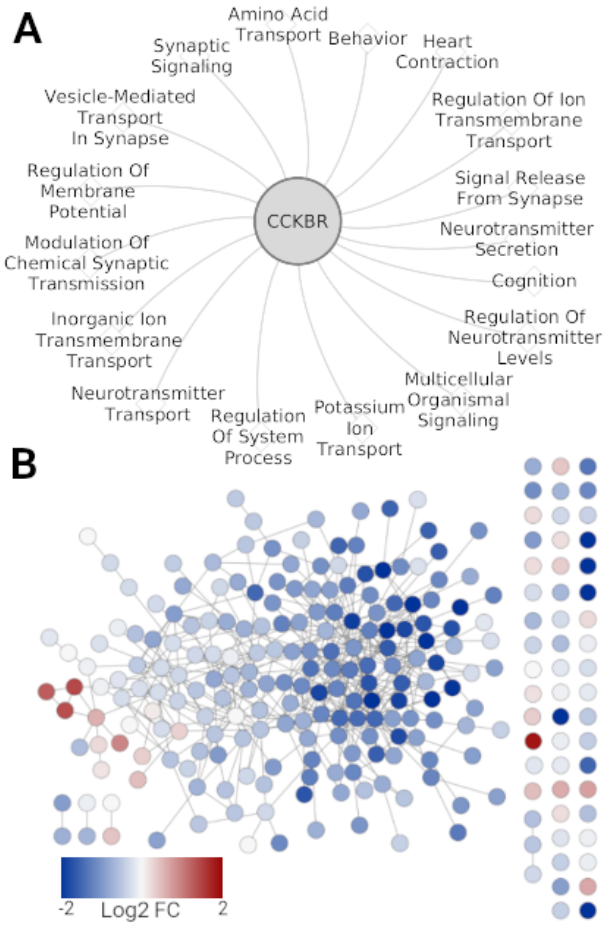

Figura 33: Términos asociados a la red de cáncer de cerebro. A) Pocesos de Gene Ontology (GO) enriquecidos en la comunidad *CCKBR*. B) Genes en la comunidad *CCKBR* que participan en los enriquecimientos con colores de acuerdo a sus valores de log_2FC.

los espermatozoides como: movimiento del cilium, espermatogénesis, reconocimiento célula-célula, etc. Esta comunidad tiene asortatividad cromosomal negativa (-0.887), lo que significa que la mayoría de sus interacciones son inter-cromosómicas. 133 genes en esta comunidad participan en procesos biológicos y de esos 133, 97 están también presentes en la red de co-expresión de cáncer, pero en diferentes comunidades. La mayoría de los genes están sub-expresados, con un promedio de log_2 *fold change* de -2.614.

En la red de co-expresión del fenotipo normal de piel, los términos asociados a la comunidad *MLANA* están relacionados con procesos metabólicos y biosintéticos de la melanina y pigmentación. La comunidad está formada por 19 genes y 7 de ellos participan en los enriquecimientos y también se encuentran en la red de co-expresión de cáncer, con una tendencia de sobre-expresión con un promedio de log_2 *fold change* de 1.3. El gen con el mayor *page rank*, *MLANA* o Melan-A, codifica para una proteína que está involucrada en la biogénesis del melanosoma y se utiliza para diagnóstico de melanoma [67]. Sin embargo, este gen no aparece anotado en ningún proceso de Gene Ontology y también está presente en la red de cáncer, con un log_2 *fold change* de 1.515.

Los términos de GO únicos para testículo y piel, así como los genes en las comunidades previamente mencionadas se muestran en la Figura 34.

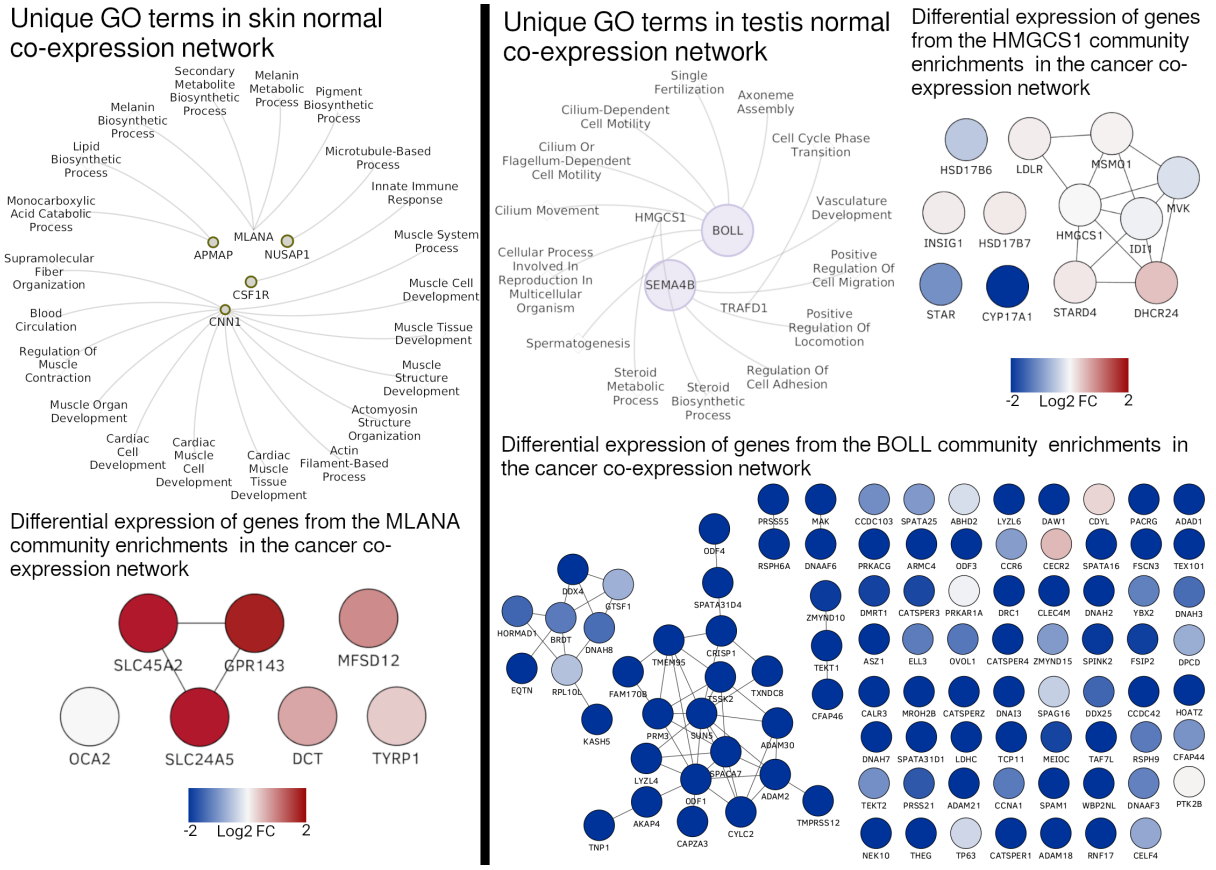

Figura 34: Procesos asociados de forma única a las redes de co-expresión de piel y testículo en el fenotipo normal y la expresión diferencial de los genes que se encuentran también en la red de cáncer.

3.8. Respuesta inmune adaptativa es un proceso aislado y común en cáncer

El proceso de Gene Ontology (GO) respuesta inmune adaptativa (*GO:0002250, adaptive immune response*) es el proceso con mayor grado en la red bipartita de comunidades a procesos en cáncer. Está enriquecido en veinticuatro comunidades en todos los tipos de cáncer a excepción de cáncer de piel. Algunas de estas comunidades incluyen a genes del cromosoma 6 que codifican para la mayor parte de los miembros del complejo principal de histocompatibilidad de clase II (*MHCII*), como *HLA-DP*, *HLA-DM*, *HLA-DOA*, *HLA-DQ*, and *HLA-DR*, además de *CIITA* (cromosoma 16) y *CD74* (cromosoma 5). Las moléculas del *MHCII* codifican para proteínas en la membrana que permiten que células del sistema inmune, como las células dendríticas y los macrófagos, llamadas células presentadoras de antígeno, presenten fracciones de antígenos a los linfocitos T [91]. *CIITA* regula la transcripción de genes *HLA* y actúa como un co-activador transcripcional [92], mientras que *CD74* media el ensamblaje y el tráfico subcelular del complejo *MHCII* [93].

Además, en las redes de cáncer las comunidades asociadas a este proceso parecen encontrarse en la periferia del componente gigante [44], es decir, el componente más grande formado por la mayoría de los vértices de la red, y los genes asociados tienen un grado bajo, lo que sugiere que estas comunidades no suelen tener enlaces de co-expresión con otras. En la Figura 18 se señalan dichas comunidades con una circunferencia negra. Para cuantificar esta observación se calculó la descomposición de k-core, también llamada *coreness*. Con este análisis, para cada vértice, se extrae la subgráfica maximal en la que los vértices que la componen tienen, al menos, grado k. Las comunidades asociadas a respuesta inmune adaptativa presentan un promedio de *coreness* menor que el resto de las comunidades, como puede verse en la Figura 35.

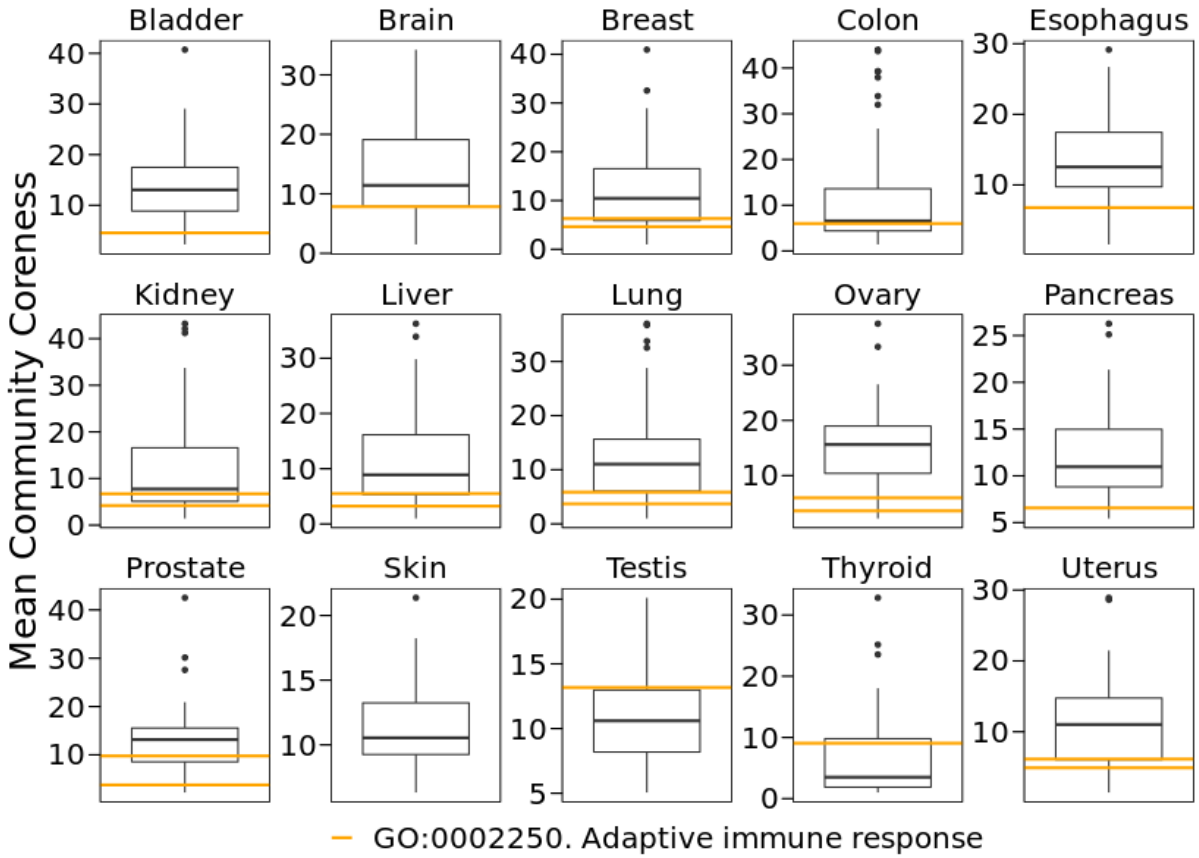

Figura 35: Distribución de *coreness* en las comunidades asociadas a respuesta inmune adaptativa. Estas comunidades, localizadas en la periferia de las redes de co-expresión están señaladas en la Figura 18.

3.9. La pérdida de co-expresión -*trans* no es una característica de otras enfermedades

Para evaluar si la pérdida de co-expresión inter-cromosómica o -*trans* es un fenómeno exclusivo de cáncer, se analizaron dos conjuntos de datos de secuenciación de ARN obtenidos de *Gene Expression Omnibus*. El primero contiene muestras de tejido de cerebro de pacientes de Alzheimer[94] y el segundo, proviene de sangre de pacientes de Diabetes tipo 2 [95].

Los valores de Información Mutua (IM) entre parejas de genes se calcularon con el mismo algoritmo utilizado para los datos de TCGA y UCSC Xena. Se extrajeron las cien mil interacciones con valores más altos de IM para construir redes de co-expresión.

En estos fenotipos, la fracción de enlaces -*cis* en las redes de co-expresión de casos y controles no presentan diferencias. Para Alzheimer, la red de controles tiene una fracción de interacciones intra-cromosómicas de 0.05997, mientras que para los casos, la fracción es de 0.05579. En la red de Diabetes tipo 2 del fenotipo normal, la misma fracción es de 0.07327, mientras que para los casos la fracción es de 0.07327.

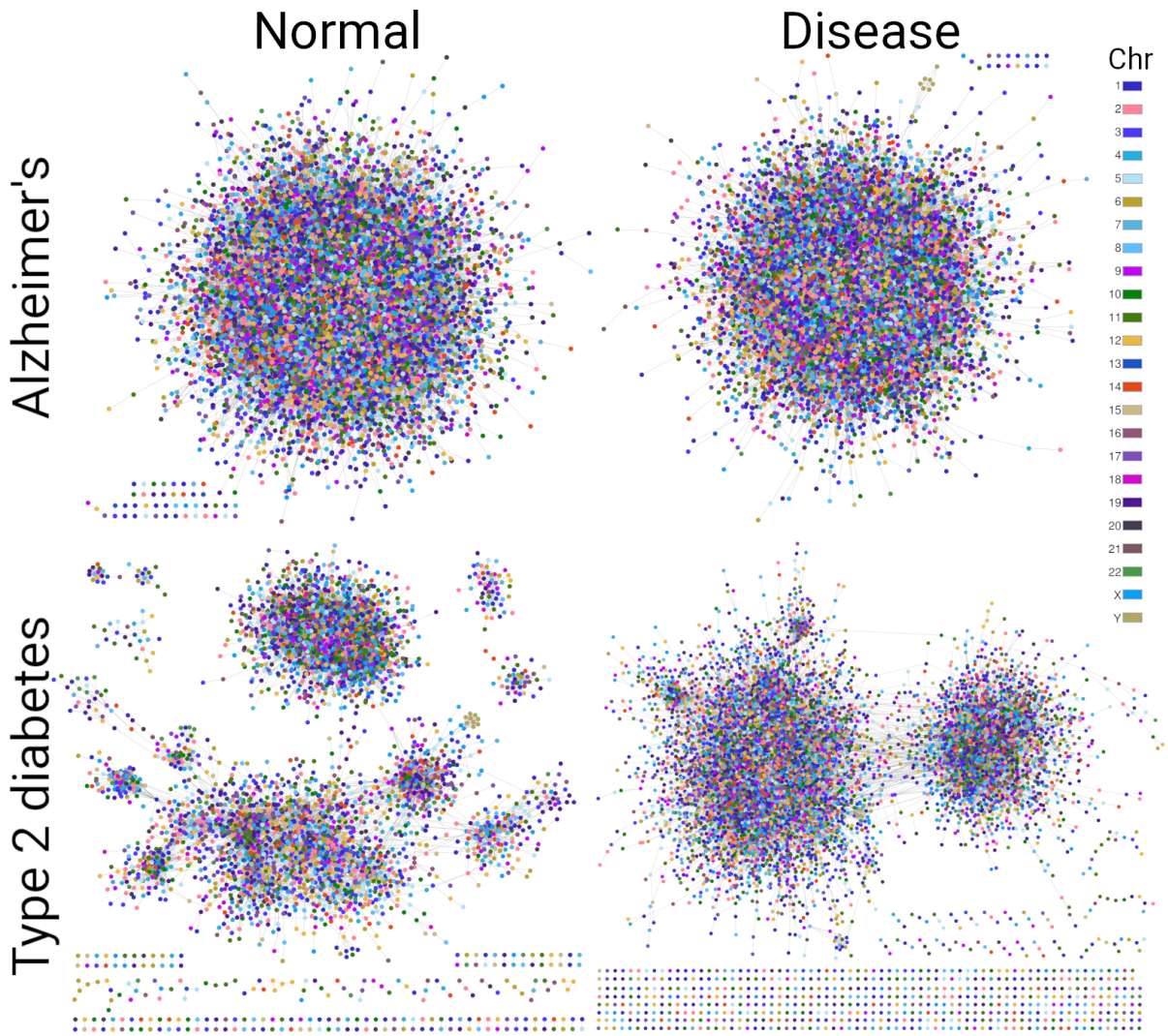

Figura 36: Redes de co-expresión en Alzheimer y Diabetes tipo 2. Formadas por las 100 mil interacciones de IM más fuertes en casos y controles. Estas redes no presentan pérdida de co-expresión inter-cromosómica.

4 Discusión. Implicaciones de la pérdida de co-expresión a larga distancia en cáncer

Este trabajo ha expuesto las características principales de dos fenómenos encontrados en cáncer asociados a la co-expresión genética: la pérdida de co-expresión inter-cromosómica y el decaimiento dependiente de la distancia física entre genes en las interacciones intra-cromosómicas. Estos fueron identificados en los perfiles de co-expresión de cáncer derivados de muestras de secuenciación de ARN de quince tejidos diferentes.

La pérdida de co-expresión inter-cromosómica se evaluó tomando diversos umbrales de corte en los valores de Información Mutua (IM) y observamos que las interacciones de co-expresión más fuertes se dan entre genes localizados en el mismo cromosoma. A su vez, el decaimiento dependiente de la distancia se analizó tomando solamente las interacciones intra-cromosómicas o *-cis* y evaluando su relación con la distancia física entre genes medida en pares de bases. Con esto concluimos que los valores de alta co-expresión forman vecindarios cercanos dentro de los cromosomas. Cabe recalcar que los resultados presentados fueron validados con pruebas estadísticas correspondientes.

Ambos fenómenos aparecen solamente en los perfiles de cáncer. En el fenotipo normal, conformado por el conjunto de muestras de tejido adyacente de TCGA y las muestras etiquetadas como *Normal Tissue* en GTEx-Xena, los valores de alta co-expresión no están sesgados a interacciones *-cis* o *-trans*. Tampoco están asociados a alguna distancia característica entre pares de genes, con excepción de distancias muy cercanas, como se muestra en el extremo izquierdo de las gráficas de distancia contra co-expresión en los perfiles de tejido normal desplegados en la Figura 15.

En dicha figura puede observarse también que todos los tejidos de cáncer presentan un decaimiento significativo en los valores promedio de IM con respecto a la distancia en pares de bases entre genes del mismo cromosoma. Por otro lado, los valores de IM en las redes normales alcanzan una meseta desde distancias muy cercanas. En ambos casos, la desviación estándar de las curvas refuerza las observaciones. Las principales interrogantes que propone la observación de este fenómeno son: ¿cómo puede interpretarse este decaimiento? y ¿qué mecanismos se encuentran tras este comportamiento en cáncer?

Niveles altos en la co-expresión local han sido reportados previamente entre pares de genes muy cercanos, con una distancia en pares de bases menor a 1 Mb, sin importar la cadena en

la que se encuentran o su orientación de transcripción [35, 36, 38, 39]. Esto muestra que el proceso de transcripción está influenciado por la ubicación física de los genes. Además, existe evidencia de que la transcripción contribuye a la formación de conjuntos de genes que pueden influir en la organización de la cromatina, al menos en una escala menor [96, 97]. En los perfiles de co-expresión de cáncer aquí analizados, los valores de alta co-expresión abarcan una mayor distancia, algunos de ellos llegan a su meseta de decaimiento alrededor de los 50 Mb, sugiriendo que los mecanismos que promueven la alta co-expresión local en fenotipos sanos se encuentran alterados en cáncer.

A propósito de dichos mecanismos, en la investigación de [34] se evaluó la contribución de diferentes elementos regulatorios tales como: *enhancers*, sitios de unión a factores de transcripción, sitios de unión de *CTCF* e interacciones de cromatina evaluados mediante Hi-C, a la alta co-expresión local en tejidos sanos. Reportaron que ningún mecanismo por sí solo puede explicar los patrones de co-expresión observados, pero se observa una contribución colectiva de éstos. Una observación importante radica en la disminución de elementos regulatorios, en particular de *enhancers* y sitios de unión a factores de transcripción, en las regiones locales de alta co-expresión, sugiriendo que este fenómeno podría estar relacionado con una especie de simplificación en la regulación transcripcional. Cabe mencionar que, durante el desarrollo de esta investigación, se analizaron elementos regulatorios en el subtipo Luminal A de cáncer de mama, y no se encontró una contribución importante de sitios de unión a *CTCF*, cambio en el número de copias o expresión diferencial en la formación de comunidades en las redes de co-expresión [43].

Esto no significa que dichas alteraciones, en particular las que están asociadas con cambios estructurales en el genoma, así como las alteraciones epigenéticas que modifican la transcripción de genes y son altamente representativas en cáncer, no estén asociadas al fenómeno de la pérdida de co-expresión a larga distancia. De hecho, existen en la literatura algunos casos que reportan alteraciones que podrían estar relacionadas al fenómeno. Por ejemplo, se sabe que debido a las modificaciones estructurales en la cromatina, en cáncer de próstata los TADs tienen un tamaño más pequeño, al compararse con tejido sano; es decir, se crean nuevas regiones insuladas usualmente asociadas al cambio en el número de copias. Además, se identifican regiones con ganancia en marcas represivas en las histonas y pérdida de marcas activas en regiones asociadas a genes represores de tumores, llamadas *long-range epigenetically silenced regions* [98]. Esto podría resultar en un incremento en la co-expresión de genes dentro de dichas regiones y/o dentro de las regiones insuladas. También se ha encontrado que en gliomas, alteraciones en la metilación de sitios de unión a la proteína estructural *CTCF* reducen su presencia, nuevamente modificando las regiones insuladas y permitiendo que elementos *-cis* reguladores como los *enhancers*, interactúen con genes en el mismo cromosoma, con los que usualmente no lo harían [99]. Estos resultados expresan la necesidad de continuar con la evaluación de la contribución de los mecanismos regulatorios que promueven la co-expresión local

en tejidos sanos y las alteraciones que permiten la extensión de estas regiones en cáncer.

Para calcular IM, cada conjunto con los valores de expresión de un gen en las diferentes muestras (en un tejido normal o de cáncer) es comparado con los valores de expresión de otro gen de forma independiente. Por lo tanto, no hay influencia debido a la posición cromosómica ni influencia del fenotipo. Es por eso que, el hecho de que genes vecinos en cáncer muestran mayor dependencia estadística que aquellos que están a mayor distancia, puede indicar que existe un mecanismo físico o mecánico en el genoma del cáncer que permite o promueve patrones de expresión con alta dependencia en largas regiones de los cromosomas.

Aunque los quince perfiles de los tejidos de cáncer presentan una pérdida de co-expresión a larga distancia, las Figuras 14 y 15 muestran que hay diferencias en la magnitud con la que el fenómeno se presenta en los diferentes tejidos. Por ejemplo, los perfiles de co-expresión del fenotipo normal y de cáncer en tejido de tiroides y de próstata presentan diferencias menos destacadas, mientras que los tejidos de vejiga, mama, pulmón y ovario muestran tanto mayores distancias entre las fracciones de interacciones -*cis*, como un decaimiento más pronunciado. Estas diferencias podrían estar asociadas a características particulares de la manifestación del cáncer en cada tejido.

Si nos enfocamos a cáncer de tiroides, la pérdida de co-expresión a larga distancia no es tan evidente como en el resto de los perfiles de cáncer: la pendiente con la que decaen los valores de IM dependientes de la distancia es menos pronunciada y su red de co-expresión tiene el mayor número de interacciones únicas. Además, aunque dicha red contiene comunidades asociadas a protocadherinas e histonas, no presenta otras características comunes al resto de las redes de cáncer, tales como la presencia de genes de la familia *HOX* o regiones con alta densidad de interacciones intra-citobanda. Según GLOBOCAN, el cáncer de tiroides se encuentra dentro de los tipos de cáncer con menores tasas de mortalidad: 0.5 en 100,000 mujeres y 0.3 en 100,000 hombres [1]. El subtipo más común de cáncer de tiroides es el papilar, seguido por el folicular y, en una menor proporción los subtipos más agresivos: medular y anaplásico [100]. Los primeros tres subtipos tienen una tasa de supervivencia de casi 100 % [101, 102]. Por lo tanto, la particular topología de la red de co-expresión de tiroides, podría estar asociada a una naturaleza menos agresiva de la enfermedad.

Las regiones de alta densidad de interacciones intra-citobanda que fueron identificadas en las cien mil interacciones más altas aparecen en todos los tejidos de cáncer, excepto en tiroides. Además, algunas de estas regiones, como 1q42.11 y 13q21.33, se comparten en más de diez tejidos y están presentes solo en perfiles de cáncer. Como citobandas, estas regiones no han sido asociadas a la enfermedad, aunque algunos genes sí se han identificado en este contexto. Por ejemplo, *NVL* en 1q42.11 se ha reportado como gen asociado a prognosis en cáncer de próstata [103] y en la misma región, *CNIH4* se ha encontrado que promueve metástasis en cáncer de colon [104]. La región q21.33 en el cromosoma 13 incluye a *BORA*, un activador

de las proteína cinasa Aurora A (*AURKA*), que participa en el desarrollo de cáncer de vejiga [105], cáncer de ovario [106], de mama, pulmón y adenocarcinomas gástricos [107].

Las redes de co-expresión, construidas con las cien mil interacciones de IM más altas son visualmente diferentes cuando se aplica un *force directed layout*, especialmente para las redes con altas fracciones de enlaces intra-cromosómicas como son: vejiga, mama, esófago, riñón, hígado, ovario y útero (Tabla 4 y Figura 19). Aunque la aparición de *clusters* de co-expresión derivados de datos de expresión genética y su asociación funcional ya ha sido previamente reportada [108-110], además de haberse reportado también una pérdida de conectividad en redes de co-expresión de cáncer provenientes de datos de microarreglos [111], la pérdida de interacciones de larga distancia junto con el incremento en los enlaces en regiones vecinas encontradas en las cien mil interacciones de co-expresión más altas, no había sido previamente descrita en cáncer.

Al obtener las intersecciones de las redes de co-expresión se observa una diferencia importante: las redes normales comparten un número menor de intersecciones que las redes de cáncer. Esto ya había sido reportado por nuestro laboratorio en cáncer de pulmón, donde las redes de cáncer de pulmón de células escamosas y adenocarcinoma de pulmón comparten más interacciones que sus contrapartes normales [46]. Además, el tipo de intersecciones compartidas es diferente, siendo congruente con el tipo de enlaces que componen cada red: inter-cromosómicas en normal e intra-cromosómicas en cáncer. Dado que el programa de co-expresión es una manifestación del estado de un fenotipo, el hecho de que este programa sea altamente similar entre tejidos de diferente origen y en diferente ubicación sugiere que existen similaridades funcionales entre ellos. Por lo tanto, proponemos que el conjunto de interacciones -*cis* que se comparten en cáncer podría estar asociado al proceso de reversión de la diferenciación celular que sufren las células durante el desarrollo del tumor [112-114]. En cambio, las interacciones -*trans* que conforman las redes normales y que no suelen compartirse entre ellas, podrían estar asociadas a las funciones particulares características de cada uno de los tejidos.

A pesar de que el número de interacciones compartidas por las redes normales es pequeño, aparece un grupo de interacciones en la red común de tejidos normales (>10 tejidos), que también está en la red común de tejidos de cáncer. Dicho *cluster* está compuesto por genes que codifican para riboproteínas. Esta observación es importante porque indica que estos genes preservan patrones de alta co-expresión en ambas condiciones y en todos los tejidos. Se sabe que el ribosoma es un componente celular esencial para todos los organismos y que las secuencias de los genes que codifican para estas proteínas están altamente conservadas entre especies [70]. Esto podría explicar la prevalencia de estos genes como conjunto de alta co-expresión en casi todas las redes. Su presencia refleja la importancia del proceso de traducción y exhibe la capacidad de las redes de co-expresión de capturar características biológicas importantes asociadas al fenotipo. Investigaciones posteriores podrían evaluar la presencia de estos genes

en redes de co-expresión en otras condiciones clínicas e inclusive otros organismos.

En cuanto a su estructura, las redes de cáncer presentan un visible comportamiento modular y, de hecho, sus valores de modularidad son más altos que los de las redes normales (Tabla 2), especialmente en las redes con alta fracción de interacciones intra-cromosómicas. Esto podría indicar que en los fenotipos normales, los genes interactúan de forma menos compartamentalizada, con eventos transcripcionales y valores de co-expresión dirigidos principalmente por funciones celulares y señales del ambiente. En contraste, los módulos segregados que encontramos en las redes de co-expresión en cáncer, podrían ilustrar una pérdida parcial de propagación de información o una ganancia en regiones específicas de alta co-expresión.

La asortatividad cromosomal (Figure 22) sirve como una métrica para resumir el siguiente fenómeno: en las comunidades de las redes normales hay una fuerte tendencia a las interacciones -*trans*, mientras que las comunidades de las redes de cáncer están principalmente conformadas por enlaces -*cis*. Además, no hay redes con valores de asortatividad tendientes a cero, lo que significa que las comunidades están compuestas ya sea por enlaces intra- o inter-cromosómicas. La importancia de ambas medidas de asortatividad se puede apreciar cuando se comparan las comunidades biológicamente asociadas y no-biológicamente asociadas: las comunidades enriquecidas tienen mayores valores de asortatividad de expresión diferencial, ilustrando cómo es que las características estructurales en las redes de co-expresión se relacionan con alteraciones funcionales, con genes en procesos biológicos expresados de forma coordinada y además, colectivamente sub o sobre-expresados.

La composición modular, el decaimiento de la co-expresión asociada a la distancia y la alta asortatividad cromosomal y de expresión en las comunidades enriquecidas en las redes sugieren una manifestación de eventos de co-expresión altamente orquestados que involucran a genes en vecindades cercanas en las redes de cáncer. La pérdida de comunicación a larga distancia a nivel transcriptómico y la aparición de *hotspots* de co-expresión puede derivar en la pérdida de funciones celulares relevantes. Esto puede apreciarse en las redes bipartitas de procesos de Gene Ontology (GO) compartidos en más de 10 tejidos ya sea normales o de cáncer y en las redes bipartitas de procesos solamente de tejidos normales y solamente de tejidos de cáncer (Figuras 26, 27, 31, 32). En las redes normales aparecen funciones celulares como la traducción, metabolismo y desarrollo, funciones clave para el mantenimiento de la viabilidad celular.

Por otro lado, aunque se encontró presencia importante de procesos asociados a respuesta inmune en comunidades formadas por enlaces -*trans* en redes de cáncer, también se encontraron comunidades con enlaces de genes en el mismo cromosoma que componen familias de genes como *HOXA*, *HOXB*, protocadherinas, variantes de histonas y metalotioneínas. Estos módulos presentan tendencias de expresión diferencial similares. Además, es importante resaltar que

algunas de estas familias fueron encontradas también en redes de alta co-expresión en tejidos normales, calculadas con un método diferente [34].

La estructura de algunas redes de co-expresión en cáncer revelan asociación tejido-específico. Por ejemplo, la comunidad observada en la Figura 33 tiene asociados un conjunto de procesos relacionados con el sistema nervioso formado principalmente por genes subexpresados. Este ejemplo indica una asociación entre expresión diferencial, co-expresión genética y características funcionales en una comunidad, demostrando una manifestación fenotipo-específica.

Las comunidades asociadas al proceso de *respuesta inmune adaptativa* aparecen en todos los tipos de cáncer, menos en piel y presentan valores bajos de *coreness* (lo que indica que los genes están localizados en las capas externas de las redes), sugiriendo bajos niveles de comunicación con el resto de la red. Adicionalmente, estas comunidades tienen genes específicos de células del sistema inmune. Esto puede indicar que la función asociada a dichas comunidades es adquirida gracias a la infiltración de células del sistema inmune en los tejidos de cáncer. En ese sentido, se requiere más información sobre la naturaleza y el origen de estas interacciones de co-expresión. Sin embargo, el resultado refleja la importancia que el análisis de co-expresión tiene en términos de hallazgos funcionales en la investigación en cáncer.

La afirmación de que la pérdida de co-expresión a larga distancia es un fenómeno específico del cáncer se refuerza con los resultados presentados en la Figura 36. Dos enfermedades crónico-degenerativas como el Alzheimer y la Diabetes tipo 2 presentan fuertes similitudes en sus paisajes de co-expresión, no solamente en una inspección visual sino también en sus valores de asortatividad cromosomal. Además, la presencia de las alteraciones estructurales del genoma previamente mencionadas, que suelen encontrarse en múltiples tipos de cáncer [115], pero no son características en Alzheimer [116] o Diabetes tipo 2, podría estar relacionada con la ausencia del fenómeno en estas enfermedades. Sin embargo, la existencia de pérdida de co-expresión a larga distancia en otras enfermedades debe ser evaluada con otros *datasets*.

A pesar de sus particularidades, los resultados presentados son consistentes en quince tejidos en cáncer, contrastados contra sus contrapartes normales. Además, no hay diferencias significativas asociadas a la fuente de los datos: tanto los datos de TCGA y UCSC Xena tienen tejidos con altos y bajos niveles de pérdida de co-expresión a larga distancia y el número de muestras aparentemente no influencia la estructura de la red, algo que ya se había evaluado en cáncer de mama [32] y cáncer de pulmón [46].

Estudiar la co-expresión genética comparando tejidos de cáncer y tejidos normales es una aproximación para mejorar el entendimiento sobre los elementos regulatorios que controlan la transcripción de grupos de genes de manera conjunta. En el caso particular de esta tesis, aunque los mecanismos moleculares no puedan ser dilucidados en este momento, la confirmación de la hipótesis que indica que las interacciones a larga distancia en el transcriptoma del

tejido neoplásico se pierden, sirve como una guía para continuar los esfuerzos del laboratorio por entender mejor las alteraciones genéticas y epigenéticas en las células de cáncer y su interrelación con el proceso de transcripción para promover el desarrollo de la enfermedad.

Finalmente es necesario hacer algunas consideraciones que podrían ser retomadas en análisis posteriores. Por ejemplo, el agrupamiento jerárquico mostrado en la Figura 20 no muestra asociaciones de acuerdo a la similitud del tejido o al origen del cáncer, por lo que sería importante indagar a profundidad sobre la naturaleza de las similitudes desplegadas mediante dicho agrupamiento. Además, en la misma figura, los conjuntos de intersección mayores, que incluyen interacciones de tejidos de esófago, piel y páncreas, pertenecen al conjunto de UCSC Xena, con datos de GTEx y por lo tanto, se necesita evaluar si es que los pasos del pre-procesamiento añadieron algún sesgo.

Esta investigación podría beneficiarse también de datos adicionales. Por ejemplo, datos de secuenciación de ARN de célula única podrían ayudarnos a identificar los tipos celulares en los que se presenta el fenómeno de la pérdida de co-expresión a larga distancia. Es importante mencionar que la secuenciación de ARN en *bulk* y el hecho de que la co-expresión es una cuantificación que compendia lo observado en múltiples muestras, nos muestran un panorama general del fenómeno. Sin embargo, ocultan características particulares, que son importantes para el desarrollo de la enfermedad, como son la existencia de subtipos tumorales, la presencia de múltiples subclonas y la importancia del microambiente tumoral [117].

Además, con *datasets* de otras ómicas en cáncer, podríamos evaluar la influencia de mecanismos regulatorios de la transcripción y de sus alteraciones en el fenómeno de pérdida de co-expresión a larga distancia. En particular, sería muy importante la evaluación de perfiles de metilación, de alteraciones estructurales a nivel de ADN y perfiles de contactos en la cromatina. La integración de estos datos y la evaluación de su contribución en el fenómeno representa retos notables. Sin embargo, es importante que las observaciones aquí reportadas sean eventualmente acompañadas de una explicación sobre el mecanismo molecular asociado para poder así entender las ventajas que adquieren las células de cáncer al mantener altos niveles de co-expresión en regiones cercanas y su relación con el desarrollo de las neoplasias.

4.1. Observaciones finales

La pérdida de co-expresión a larga distancia en cáncer consiste en dos manifestaciones principales: una pérdida de co-expresión inter-cromosómica y un decaimiento en la co-expresión intra-cromosómica dependiente de la distancia física entre genes. Se observó este fenómeno en quince tejidos de cáncer diferentes, pero no en tejidos normales. Además, su presencia no depende del umbral de corte en los valores de co-expresión, ni en el número de muestras y es independiente de la fuente de datos. La estructura que se observa en las redes de cáncer está

asociada con características funcionales comunes en la enfermedad. Genes que codifican para proteínas constituyen un conjunto particular de interacciones inter-cromosómicas compartidas en más de veinte tejidos en ambas condiciones, indicando la importancia de la co-regulación de estos genes para preservar la viabilidad celular. Las comunidades asociadas a procesos biológicos en cáncer presentan altos valores de asortatividad de expresión cromosomal; es decir, genes en la misma comunidad tienen la misma tendencia de expresión diferencial. Comunidades que se encuentran en las capas externas del componente más grande en la mayoría de las redes de cáncer están asociadas a la activación de la respuesta inmune, lo cual podría significar que estas comunidades no son parte estructural de la red y sugiere eventos de infiltración de células del sistema inmune. Finalmente, la pérdida de co-expresión a larga distancia no está presente en otras enfermedades crónico degenerativas como Diabetes tipo 2 o Alzheimer. Proponemos que este fenómeno es una manifestación de una característica común del cáncer que no ha sido previamente reportada. Se requiere más investigación, particularmente estudios experimentales para identificar los mecanismos detrás de estas alteraciones en el programa de co-expresión en las células. Sin embargo, creemos que la pérdida de correlación en la expresión genética a larga distancia podría ser reflejo de anomalías importantes en el genoma de cáncer que no han sido previamente exploradas.

Bibliografía

[1] Hyuna Sung y col. «Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries». En: *CA: a cancer journal for clinicians* 71.3 (2021), págs. 209-249.

[2] Freddie Bray y col. «The ever-increasing importance of cancer as a leading cause of premature death worldwide». En: *Cancer* 127.16 (2021), págs. 3029-3030.

[3] Khanh Bao Tran y col. «The global burden of cancer attributable to risk factors, 2010–19: a systematic analysis for the Global Burden of Disease Study 2019». En: *The Lancet* 400.10352 (2022), págs. 563-591.

[4] World Health Organization. *Cancer*. Feb. de 2022. URL: https://www.who.int/news-room/fact-sheets/detail/cancer.

[5] Nancy Reynoso-Noverón y Juan Alejandro Torres-Domínguez. «Epidemiología del cáncer en México: carga global y proyecciones 2000-2020». En: *Revista Latinoamericana de Medicina Conductual/Latin American Journal of Behavioral Medicine* 8.1 (2017), págs. 9-15.

[6] Alejandro Mohar-Betancourt y col. «Cancer trends in Mexico: essential data for the creation and follow-up of public policies». En: *Journal of global oncology* 3.6 (2017), págs. 740-748.

[7] Nicolás Padilla-Raygoza y col. «Cancer prevention programmes in Mexico: are we doing enough?» En: *ecancermedicalscience* 14 (2020).

[8] Hasan Brau-Figueroa, E Alejandra Palafox-Parrilla y Alejandro Mohar-Betancourt. «The National Cancer Registry in Mexico, a reality». En: *Gaceta mexicana de oncología* 19.3 (2020), págs. 107-111.

[9] World Health Organization y col. «WHO report on cancer: setting priorities, investing wisely and providing care for all». En: (2020).

[10] Douglas Hanahan y Robert A Weinberg. «The hallmarks of cancer». En: *cell* 100.1 (2000), págs. 57-70.

[11] Douglas Hanahan y Robert A Weinberg. «Hallmarks of cancer. An organizing principle for cancer medicine». En: *Primer of the molecular biology of cancer. 2nd ed. Philadelphia: Wolters Kluwer* (2015), págs. 28-57.

[12] Nadine Darwiche. «Epigenetic mechanisms and the hallmarks of cancer: An intimate affair». En: *American journal of cancer research* 10.7 (2020), pág. 1954.

[13] Levi A Garraway y Eric S Lander. «Lessons from the cancer genome». En: *Cell* 153.1 (2013), págs. 17-37.

[14] Douglas Hanahan y Robert A Weinberg. «Hallmarks of cancer: the next generation». En: *cell* 144.5 (2011), págs. 646-674.

[15] Douglas Hanahan. «Hallmarks of cancer: new dimensions». En: *Cancer discovery* 12.1 (2022), págs. 31-46.

[16] Pierre Hainaut y Amelie Plymoth. «Targeting the hallmarks of cancer: towards a rational approach to next-generation cancer therapy». En: *Current opinion in oncology* 25.1 (2013), págs. 50-51.

[17] Bas van Steensel y Eileen EM Furlong. «The role of transcription in shaping the spatial organization of the genome». En: *Nature reviews Molecular cell biology* 20.6 (2019), págs. 327-337.

[18] Malte Spielmann, Darío G Lupiáñez y Stefan Mundlos. «Structural variation in the 3D genome». En: *Nature Reviews Genetics* 19.7 (2018), págs. 453-467.

[19] Patrick Cramer. «Organization and regulation of gene transcription». En: *Nature* 573.7772 (2019), págs. 45-54.

[20] Tong Ihn Lee y Richard A Young. «Transcriptional regulation and its misregulation in disease». En: *Cell* 152.6 (2013), págs. 1237-1251.

[21] Enrique Hernández-Lemus y col. «The many faces of gene regulation in cancer: a computational oncogenomics outlook». En: *Genes* 10.11 (2019), pág. 865.

[22] James E Bradner, Denes Hnisz y Richard A Young. «Transcriptional addiction in cancer». En: *Cell* 168.4 (2017), págs. 629-643.

[23] Barbara Heidenreich y col. «TERT promoter mutations in cancer development». En: *Current opinion in genetics & development* 24 (2014), págs. 30-37.

[24] Zhong Wang, Mark Gerstein y Michael Snyder. «RNA-Seq: a revolutionary tool for transcriptomics». En: *Nature reviews genetics* 10.1 (2009), págs. 57-63.

[25] Sipko Van Dam y col. «Gene co-expression analysis for functional classification and gene–disease predictions». En: *Briefings in bioinformatics* 19.4 (2018), págs. 575-592.

[26] Paola Paci y col. «Gene co-expression in the interactome: moving from correlation toward causation via an integrated approach to disease module discovery». En: *NPJ systems biology and applications* 7.1 (2021), págs. 1-11.

[27] Miguel A García-Campos, Jesús Espinal-Enríquez y Enrique Hernández-Lemus. «Pathway analysis: state of the art». En: *Frontiers in physiology* 6 (2015), pág. 383.

[28] Abhijeet R Sonawane y col. «Network medicine in the age of biomedical big data». En: *Frontiers in Genetics* 10 (2019), pág. 294.

[29] Guillermo de Anda-Jáuregui y col. «Functional and transcriptional connectivity of communities in breast cancer co-expression networks». En: *Applied Network Science* 4.1 (2019), págs. 1-13.

[30] Yang Yang y col. «Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types». En: *Nature communications* 5.1 (2014), págs. 1-9.

[31] Rong Liu, Cheng-Xian Guo y Hong-Hao Zhou. «Network-based approach to identify prognostic biomarkers for estrogen receptor–positive breast cancer treatment with tamoxifen». En: *Cancer biology & therapy* 16.2 (2015), págs. 317-324.

[32] Jesús Espinal-Enríquez y col. «RNA-Seq based genome-wide analysis reveals loss of inter-chromosomal regulation in breast cancer». En: *Scientific Reports* 7.1 (2017), págs. 1-19. ISSN: 20452322.

[33] John N Weinstein y col. «The cancer genome atlas pan-cancer analysis project». En: *Nature genetics* 45.10 (2013), págs. 1113-1120.

[34] Diogo M Ribeiro y col. «The molecular basis, genetic control and pleiotropic effects of local gene co-expression». En: *Nature communications* 12.1 (2021), págs. 1-13.

[35] Diogo M Ribeiro, Chaymae Ziyani y Olivier Delaneau. «Shared regulation and functional relevance of local gene co-expression revealed by single cell analysis». En: *Communications Biology* 5.1 (2022), págs. 1-11.

[36] Laurence D Hurst, Csaba Pál y Martin J Lercher. «The evolutionary dynamics of eukaryotic gene order». En: *Nature Reviews Genetics* 5.4 (2004), págs. 299-310.

[37] Barak A Cohen y col. «A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression». En: *Nature genetics* 26.2 (2000), págs. 183-186.

[38] Avazeh T Ghanbarian y Laurence D Hurst. «Neighboring genes show correlated evolution in gene expression». En: *Molecular biology and evolution* 32.7 (2015), págs. 1748-1766.

[39] Miki Ebisuya y col. «Ripples from neighbouring transcription». En: *Nature cell biology* 10.9 (2008), págs. 1106-1113.

[40] Diana García-Cortés y col. «Gene Co-expression Is Distance-Dependent in Breast Cancer». En: *Frontiers in Oncology* 10.July (2020), págs. 1-13.

[41] Aleix Prat y Charles M Perou. «Deconstructing the molecular portraits of breast cancer.» En: *Molecular oncology* 5.1 (feb. de 2011), págs. 5-23. ISSN: 1878-0261.

[42] Saber Fallahpour y col. «Breast cancer survival by molecular subtype: a population-based analysis of cancer registry data». En: *CMAJ Open* 5.3 (2017), E734-E739.

[43] Diana García-Cortés, Enrique Hernández-Lemus y Jesús Espinal-Enríquez. «Luminal A Breast Cancer Co-expression Network: Structural and Functional Alterations.» En: *Frontiers in genetics* 12.July (2021), pág. 629475. ISSN: 1664-8021.

[44] Michele Coscia. «The Atlas for the Aspiring Network Scientist». En: *CoRR* abs/2101.00863 (2021). arXiv: 2101.00863.

[45] Sergey Brin y Lawrence Page. «The anatomy of a large-scale hypertextual web search engine». En: *Computer networks and ISDN systems* 30.1-7 (1998), págs. 107-117.

[46] Sergio Daniel Andonegui-Elguera y col. «Loss of long distance co-expression in lung cancer». En: *Frontiers in genetics* (2021), pág. 192.

[47] Jose María Zamora-Fuentes, Enrique Hernández-Lemus y Jesús Espinal-Enríquez. «Gene expression and co-expression networks are strongly altered through stages in clear cell renal carcinoma». En: *Frontiers in genetics* 11 (2020), pág. 578679.

[48] Felix Mölder y col. «Sustainable data analysis with Snakemake». En: *F1000Research* 10 (2021).

[49] Mary J Goldman y col. «Visualizing and interpreting cancer genomics data via the Xena platform». En: *Nature biotechnology* 38.6 (2020), págs. 675-678.

[50] John Lonsdale y col. «The genotype-tissue expression (GTEx) project». En: *Nature genetics* 45.6 (2013), págs. 580-585.

[51] John Vivian y col. «Toil enables reproducible, open source, big biomedical data analyses». En: *Nature biotechnology* 35.4 (2017), págs. 314-316.

[52] Givanna H Putri y col. «Analysing high-throughput sequencing data in Python with HTSeq 2.0». En: *Bioinformatics* 38.10 (2022), págs. 2943-2945.

[53] Bo Li y Colin N Dewey. «RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome». En: *BMC bioinformatics* 12 (2011), págs. 1-16.

[54] Sonia Tarazona y col. «Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package». En: *Nucleic acids research* 43.21 (2015), e140-e140.

[55] F Finotello y col. «RNA sequencing data: biases and normalization». En: *EMBnet. journal* 18.A (2012), pág. 99.

[56] Davide Risso y col. «GC-content normalization for RNA-Seq data». En: *BMC bioinformatics* 12.1 (2011), págs. 1-17.

[57] Adam A Margolin y col. «ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context». En: *BMC bioinformatics*. Vol. 7. BioMed Central. 2006, págs. 1-15.

[58] Matthew E Ritchie y col. «limma powers differential expression analyses for RNA-sequencing and microarray studies». En: *Nucleic acids research* 43.7 (2015), e47-e47.

[59] N Alan Heckert y col. «NIST/SEMATECH e-handbook of statistical methods». En: (2002).

[60] RA Sánchez Turcios. «Prueba de Wilcoxon-Mann-Whitney: mitos y realidades». En: *Rev Mex Endocrinol Metab Nutr* 2 (2015), págs. 18-21.

[61] Kuo-Ho Yen y col. «A precise and scalable method for querying genes in chromosomal banding regions based on cytogenetic annotations». En: *Bioinformatics* 21.17 (2005), págs. 3469-3474.

[62] Gabor Csardi y Tamas Nepusz. «The igraph software package for complex network research». En: *InterJournal* Complex Systems (2006), pág. 1695. URL: `https://igraph.org`.

[63] Sara Rahiminejad, Mano R. Maurya y Shankar Subramaniam. «Topological and functional comparison of community detection algorithms in biological networks». En: *BMC Bioinformatics* 20.1 (2019), págs. 1-25. ISSN: 14712105.

[64] Michael Ashburner y col. «Gene ontology: tool for the unification of biology». En: *Nature genetics* 25.1 (2000), págs. 25-29.

[65] Elizabeth I Boyle y col. «GO:: TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes». En: *Bioinformatics* 20.18 (2004), págs. 3710-3715.

[66] Guangchuang Yu y col. «clusterProfiler: an R package for comparing biological themes among gene clusters». En: *Omics: a journal of integrative biology* 16.5 (2012), págs. 284-287.

[67] Nuala A O'Leary y col. «Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation». En: *Nucleic acids research* 44.D1 (2016), págs. D733-D745.

[68] Ioannis A Voutsadakis. «Amplification of 8p11. 23 in cancers and the role of amplicon genes». En: *Life Sciences* 264 (2021), pág. 118729.

[69] Stefanie Marczok y col. «Comprehensive analysis of genome rearrangements in eight human malignant tumor tissues». En: *PloS one* 11.7 (2016), e0158995.

[70] Maki Yoshihama y col. «The human ribosomal protein genes: sequencing and comparative analysis of 73 genes». En: *Genome research* 12.3 (2002), págs. 379-390.

[71] Jordi Garcia-Fernández. «The genesis and evolution of homeobox gene clusters». En: *Nature Reviews Genetics* 6.12 (2005), págs. 881-892.

[72] Petr Novak y col. «Epigenetic inactivation of the HOXA gene cluster in breast cancer». En: *Cancer research* 66.22 (2006), págs. 10664-10670.

[73] Muhiddin Ishak y col. «Landscape of HOXA genes methylation in colorectal cancer». En: *Progress In Microbes & Molecular Biology* 3.1 (2020).

[74] Tong Lu y col. «Circular RNA circCSNK1G3 induces HOXA10 signaling and promotes the growth and metastasis of lung adenocarcinoma cells through hsa-miR-143-3p sponging». En: *Cellular Oncology* 44.2 (2021), págs. 297-310.

[75] Yi-Nan Guo y col. «Comprehensive clinical implications of homeobox A10 in 3,199 cases of non-small cell lung cancer tissue samples combining qRT-PCR, RNA sequencing and microarray data». En: *American Journal of Translational Research* 11.1 (2019), pág. 45.

[76] Junhu Wan y col. «HOXB9 promotes endometrial cancer progression by targeting E2F3». En: *Cell death & disease* 9.5 (2018), págs. 1-18.

[77] Ying Ying y col. «Oncogenic HOXB8 is driven by MYC-regulated super-enhancer and potentiates colorectal cancer invasiveness via BACH1». En: *Oncogene* 39.5 (2020), págs. 1004-1017.

[78] Kwei-Yan Liu y col. «Homeobox genes and hepatocellular carcinoma». En: *Cancers* 11.5 (2019), pág. 621.

[79] Simone Aparecida de Bessa Garcia y col. «HOX genes function in Breast Cancer development». En: *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1873.2 (2020), pág. 188358.

[80] Fátima Liliana Monteiro y col. «Expression and functionality of histone H2A variants in cancer». En: *Oncotarget* 5.11 (2014), pág. 3428.

[81] Shweta R Nayak y col. «A role for histone H2B variants in endocrine-resistant breast cancer». En: *Hormones and Cancer* 6.5 (2015), págs. 214-224.

[82] Wenting Xie y col. «Expression and potential prognostic value of histone family gene signature in breast cancer». En: *Experimental and therapeutic medicine* 18.6 (2019), págs. 4893-4903.

[83] Xinbing Sui y col. «Methylated promoters of genes encoding protocadherins as a new cancer biomarker family». En: *Molecular biology reports* 39.2 (2012), págs. 1105-1111.

[84] Ming Shan y col. «Aberrant expression and functions of protocadherins in human malignant tumors». En: *Tumor Biology* 37.10 (2016), págs. 12969-12981.

[85] Ana Florencia Vega-Benedetti y col. «Clustered protocadherins methylation alterations in cancer». En: *Clinical epigenetics* 11.1 (2019), págs. 1-20.

[86] Mary Ann Sens y col. «Metallothionein isoform 3 overexpression is associated with breast cancers having a poor prognosis». En: *The American journal of pathology* 159.1 (2001), págs. 21-26.

[87] Imad Alkamal y col. «An epigenetic screen unmasks metallothioneins as putative contributors to renal cell carcinogenesis». En: *Urologia Internationalis* 94.1 (2015), págs. 99-110.

[88] Bozena Werynska y col. «Correlation between expression of metallothionein and expression of Ki-67 and MCM-2 proliferation markers in non-small cell lung cancer». En: *Anticancer research* 31.9 (2011), págs. 2833-2839.

[89] Gui-You Liang y col. «Expression of metallothionein and Nrf2 pathway genes in lung cancer and cancer-surrounding tissues». En: *World Journal of Surgical Oncology* 11.1 (2013), págs. 1-5.

[90] JG Hengstler y col. «Metallothionein expression in ovarian cancer in relation to histopathological parameters and molecular markers of prognosis». En: *International journal of cancer* 95.2 (2001), págs. 121-127.

[91] Judith A Owen, Jenni Punt, Sharon A Stranford y col. *Kuby immunology*. WH Freeman New York, 2013, pág. 522.

[92] Dinah Singer y Ballachanda Devaiah. «CIITA and Its Dual Roles in MHC Gene Transcription». En: *Frontiers in Immunology* 4 (2013), pág. 476. ISSN: 1664-3224.

[93] Bernd Schröder. «The multifaceted roles of the invariant chain CD74 — More than just a chaperone». En: *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1863.6, Part A (2016), págs. 1269-1281.

[94] Samuel Morabito y col. «Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease». En: *Nature genetics* 53.8 (2021), págs. 1143-1155.

[95] Hung-Hsin Chen y col. «Novel diabetes gene discovery through comprehensive characterization and integrative analysis of longitudinal gene expression changes». En: *Human Molecular Genetics* (2022).

[96] Bas van Steensel y Eileen EM Furlong. «The role of transcription in shaping the spatial organization of the genome». En: *Nature reviews Molecular cell biology* 20.6 (2019), págs. 327-337.

[97] María E Soler-Oliva y col. «Analysis of the relationship between coexpression domains and chromatin 3D organization». En: *PLoS computational biology* 13.9 (2017), e1005708.

[98] Phillippa C Taberlay y col. «Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations». En: *Genome research* 26.6 (2016), págs. 719-731.

[99] William A Flavahan y col. «Insulator dysfunction and oncogene activation in IDH mutant gliomas». En: *Nature* 529.7584 (2016), págs. 110-114.

[100] Adalberto Miranda-Filho y col. «Thyroid cancer incidence trends by histology in 25 countries: a population-based study». En: *The lancet Diabetes & endocrinology* 9.4 (2021), págs. 225-234.

[101] American Cancer Society. *Thyroid cancer survival rates, by type and stage*. 2016.

[102] Luigino Dal Maso y col. «Survival of 86,690 patients with thyroid cancer: a population-based study in 29 European countries from EUROCARE-5». En: *European Journal of Cancer* 77 (2017), págs. 140-152.

[103] Shuang G Zhao y col. «The Landscape of Prognostic Outlier Genes in High-Risk Prostate CancerHigh-Risk Prostate Cancer Prognostic Outlier Gene Landscape». En: *Clinical cancer research* 22.7 (2016), págs. 1777-1786.

[104] Sonakshi Mishra y col. «The protein secretion modulator TMED9 drives CNIH4-TGFα-GLI signaling opposing TMED3-WNT-TCF to promote colon cancer metastases». En: *Oncogene* 38.29 (2019), págs. 5817-5837.

[105] Songtao Cheng y col. «BORA regulates cell proliferation and migration in bladder cancer». En: *Cancer cell international* 20.1 (2020), págs. 1-10.

[106] Alfonso Parrilla y col. «Aurora Borealis (Bora), which promotes Plk1 activation by Aurora A, has an oncogenic role in ovarian cancer». En: *Cancers* 12.4 (2020), pág. 886.

[107] Qiong-Xia Zhang y col. «Cell cycle protein Bora serves as a novel poor prognostic factor in multiple adenocarcinomas». En: *Oncotarget* 8.27 (2017), pág. 43838.

[108] Wencheng Yin y col. «Emergence of co-expression in gene regulatory networks». En: *PloS one* 16.4 (2021), e0247671.

[109] Pawel Michalak. «Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes». En: *Genomics* 91.3 (2008), págs. 243-248.

[110] Sergio Antonio Alcalá-Corona y col. «Network modularity in breast cancer molecular subtypes». En: *Frontiers in physiology* 8 (2017), pág. 915.

[111] Roberto Anglani y col. «Loss of connectivity in cancer co-expression networks». En: *PloS one* 9.1 (2014), e87075.

[112] Dinorah Friedmann-Morvinski e Inder M Verma. «Dedifferentiation and reprogramming: origins of cancer stem cells». En: *EMBO reports* 15.3 (2014), págs. 244-253.

[113] Arnatchai Maiuthed y col. «Nitric oxide promotes cancer cell dedifferentiation by disrupting an Oct4: caveolin-1 complex: A new regulatory mechanism for cancer stem cell formation». En: *Journal of Biological Chemistry* 293.35 (2018), págs. 13534-13552.

[114] Jinyang Li y Ben Z Stanger. «How tumor cell dedifferentiation drives immune evasion and resistance to immunotherapy». En: *Cancer research* 80.19 (2020), págs. 4037-4041.

[115] Deniz Demircioğlu y col. «A pan-cancer transcriptome analysis reveals pervasive regulation through alternative promoters». En: *Cell* 178.6 (2019), págs. 1465-1477.

[116] Jose V Sanchez-Mut y Johannes Gräff. «Epigenetic alterations in Alzheimer's disease». En: *Frontiers in behavioral neuroscience* 9 (2015), pág. 347.

[117] Ash A Alizadeh y col. «Toward understanding and exploiting tumor heterogeneity». En: *Nature medicine* 21.8 (2015), págs. 846-853.