Universidad Nacional
Autónoma de México

Facultad de Ciencias

Redes de regulación transcripcional asociadas a procesos inflamatorios en cáncer de mama

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

BIÓLOGO

P R E S E N T A:

Tadeo Enrique Velázquez Caldelas

DIRECTOR DE TESIS:

Dr. Enrique Hernández Lemus

2015

Ciudad Universitaria, D. F.

Universidad Nacional Autónoma de México
UNAM
Biblioteca Central
Dirección General de Bibliotecas de la UNAM

UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

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

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

1.- Datos del alumno
Velázquez
Caldelas
Tadeo Enrique
55838890
Universidad Nacional Autónoma de México
Facultad de Ciencias
Biología
099333128

2.- Datos del tutor
Dr
Enrique
Hernández
Lemus

3.- Datos del sinodal 1
Dr
Pedro Eduardo
Miramontes
Vidal

4.- Datos del sinodal 2
Dr
Jesús
Espinal
Enríquez

Datos del sinodal 3
Dr
Arturo Carlos II
Becerra
Bracho

Datos del sinodal 4
Biol
Saúl
Cano
Colín
Datos del trabajo escrito
Redes de regulación transcripcional asociadas a procesos inflamatorios en cáncer de mama
80p
2015

Dedicatoria

Para Julian y Valentina

Agradecimientos

Al grupo de biología de sistemas computacional del INMEGEN, por ser siempre pacientes. A Enrique, por darme la oportunidad de trabajar en su laboratorio, a Jesús por su interminable disposición de ayudar, a Hugo por sus opiniones siempre honestas y al resto del laboratorio: Guillermo, Daniel, Raúl, Rodrigo, Aldo, Ana, Sergio, Estefanía, Diana, Miguel, a todos gracias por hacer un gran equipo de trabajo.

A mis papás, a mi tía Rocío, a mi tia Car, a mis abuelitos por insistirme siempre en terminar.

Resumen

En este trabajo presentamos una metodología con un enfoque computacional para investigar procesos biológicos asociados a nivel de regulación transcripcional con el proceso inflamatorio en cáncer primario de mama. El análisis incluye el procesamiento de conjuntos de datos biológicos de gran volumen y la integración de conocimiento de bases de datos como la ontología de genes.

El análisis parte de un proceso biológico de interés: inflamación. Utilizando la información de los niveles de expresión de RNA mensajero del genoma completo, los genes son asociados por medio de la correlación en sus niveles de expresión en una red de regulación transcripcional. La estructura de la red permite encontrar grupos de genes coexpresados, que pueden representar módulos transcripcionales. Los análisis de enriquecimiento estadístico de conjuntos de genes en la red permiten encontrar los procesos biológicos representados en ésta. Finalmente, integrando el análisis de expresión diferencial de los genes en los tejidos de cáncer primario de mama, con respecto a tejido mamario normal, investigamos el tipo de regulación, positiva o negativa, que se realiza sobre los procesos encontrados.

Los resultados obtenidos sugieren que el proceso inflamatorio en el cáncer primario de mama se encuentra activo y asociado a procesos como la cicatrización y la modulación de la respuesta inmune, mismos que se han sugerido ampliamente en la literatura relaciónados a la tumorigénesis. La estructura de la red revela grandes conjuntos de genes agrupados con una expresión diferencial similar, con una predominancia de genes subexpresados. Los grupos de genes estrechamente

coexpresados muestran un enriquecimeinto en más de un proceso, mientras que los procesos individuales localizados dentro de la estructura de la red forman más de un grupo, lo que sugiere patrones de regulación complejos.

Índice general

Capítulo 1

Análisis computacional de fenómenos biológicos

El cáncer de mama ocupa, a nivel mundial, el primer lugar en número de casos de cáncer diagnosticados en mujeres, con más de un millón de casos nuevos anuales y medio millón de defunciones anuales en todos los grupos de edad [1]. En México el cáncer de mama es reconocido por el INEGI como la primera causa de muerte en mujeres entre 35 y 44 años de edad (INEGI 2012).

El uso de ensayos de alto contenido de información en experimentos de gran envergadura ha mostrado la heterogeneidad molecular del cáncer de mama, tanto en la presencia de mutaciones en genes específicos como en los niveles de expresión de genes individuales [2]. La disponibilidad de conjuntos de datos de gran volumen nos posibilita el ir más allá del manejo simultáneo de una gran cantidad de genes individuales, integrando información biológica con el fin de entender la estructura, dinámica, mecanismos de control y de diseño de sistemas biológicos, marco que toma en cuenta la biologá de sistemas [3].

A partir de microarreglos de expresión de genoma completo puede extraerse información global acerca de las relaciones funcionales entre genes [4]. Éstas relaciones a su vez nos permiten formular modelos que explican de manera plausible

el comportamiento (fenotipo) observado en las muestras biológicas. La inferencia o *ingeniería reversa* de redes de genes es el proceso de identificar interacciones entre genes a partir de datos experimentales por medio del análisis computacional [5]. En éste tipo de enfoque que se denomina como *dirigido por los datos*(en inglés *data driven*) tenemos una suposición sobre la naturaleza de las interacciones, sin embargo la identidad de los elementos que interactúan solamente se conoce después de analizar los datos.

En este trabajo proponemos el análisis computacional para investigar la biologá de la respuesta inflamatoria en el cáncer primario de mama.

El análisis consiste de los siguientes pasos:

- **Inferencia de redes de regulación transcripcional**. Las redes están formadas por nodos, que representan a los genes, y aristas que representan la correlación en sus niveles de expresión. Las redes se calculan a partir de una lista de genes de inflamación (sección 4.1). Cada gen en la lista es comparado con cada uno de los genes en el genoma y cuando existe una correlación por arriba de un valor de referencia, se añade una arista (sección 4.2).
- **Identificación de comunidades** definidas como grupos de genes en la red altamente conectados entre sí y poco conectados al exterior (sección 5.2.2). Este análisis de la estructura de la red permite encontrar conjuntos de genes estrechamente coexpresados.
- **Análisis de enriquecimiento estadístico** de conjuntos de genes, que permite obtener una lista de procesos biológicos conocidos, representados con un buen nivel de confianza estadísica (sección 4.3.4). Este análisis se realiza sobre listas de genes de prueba que se encuentran al analizar la estructura de la red, lo que quiere decir que forman un grupo de genes correlacionados a nivel de su expresión.
- **Integración de los datos de expresión diferencial y los procesos biológicos representados por los genes en la red**, lo que nos ayuda a

entender de qué manera se encuentran regulados los procesos biológicos en los tejidos analizados .

La finalidad de este análisis en el contexto del proceso inflamatorio en cáncer primario de mama es la de entender cómo los patrones de expresión de genes influyen en la función biológica del tejido, función que se relaciona con los procesos biológicos que llevan a cabo las proteńas codificadas por los genes en la red. Debido a que esta red contiene genes asociados por sus patrones de expresión, independientemente de si se consideran inflamatorios o no, esto nos permite recuperar otros procesos biológicos asociados (corregulados a nivel transcripcional) con la inflamación (sección 5.2)(sección 5.2.3).

Capítulo 2

Biología del proceso inflamatorio

2.1. Cáncer y su relación con el proceso inflamatorio

El término cáncer abarca un conjunto de enfermedades caracterizadas por el desarrollo anormal de tejido, particularmente con un crecimiento descontrolado y agresivo, en muchos casos con la capacidad de invadir sitios distantes dentro del organismo (metástascris). El desarrollo del cáncer se ha asociado al estilo de vida, la edad y factores de riesgo como la exposición a agentes ambientales, por ejemplo suspensiones de particulas no degradables por el organismo como los asbestos y ollin, así como mutágenos químicos y físicos, tales como radiaciones ionizantes, toxinas, solventes y radicales libres entre otros [6].

Los factores biológicos asociados al desarrollo del cáncer incluyen la cantidad de divisiones celulares que muestran correlación con la prevalencia de ciertos tipos de cáncer no explicados por factores de riesgo ambientales, posiblemente por errores acumulados durante la replicación repetida del DNA [7]. Así mismo se ha visto la asociación en algunos tipos de cáncer con la presencia de virus entre los que se encuentran el virus del papiloma humano, herpesvirus, virus de la hepatitis tipo B, virus de Epstein Barr y virus de la hepatitis tipo C, que contienen en su genoma la secuencia de proteínas oncogénicas o de proteñas que interaccionan con

oncogenes, o bien (en el caso de los retrovirus) tienen la capacidad de insertarse en el genoma del hospedero [8].

La relación entre el cáncer y las células del sistema inmunológico puede rastrearse hasta 1863 cuando Rudolf Virchow observó que existían leucocitos en los tumores. Virchow sugirió que este *infiltrado linforreticular* reflejaba el origen del cáncer en sitios de inflamación crónica [9]. El papel de la inflamación como facilitadora del proceso de tumorigenesis se ha respaldado con numerosos experimentos realizados en modelos animales, principalmente en ratones, donde la manipulación de la presencia de células del sistema inmune, tanto innato como adaptativo tiene efectos en el desarrollo de tumores y establecimiento de metástasis [10] [11] [12]. El proceso de *cicatrización* se ha propuesto como uno de los mecanismos clave para el crecimiento tumoral, asociando la inflamación, angiogénesis y reparación del tejido. La cicatrización que no es capaz de resolverse puede llevar a una inflamación crónica, a la formación sostenida de vasos sanguíneos que permiten el paso del plasma sanguíneo (angiogénesis defectuosa [13]) y la proliferación constante de células mesenquimales [14].

Dos artículos seminales de Douglas Hanahan y Robert A. Weinberg [15, 16] proponen una serie de características que deben ser adquiridas por las células del organismo para poder convertirse de manera exitosa en células cancerosas. La primera versión de Hallmarks of Cancer en el año 2000 menciona como características: *la autosuficiencia de señales de crecimiento*, *insensibilidad a las señales anticrecimiento*, *evasión de la apoptosis*, *angiogénesis sostenida*, potencial replicativo ilimitado y la invasión de tejido y metástasis. La versión revisada de 2011 Hallmarks of Cancer: The next generation incluye dos características emergentes: *desregulación del metabolismo energético* y *evasión de la destrucción por el sistema inmune*. También son incluidas dos características facilitadoras: *inestabilidad genómica y mutación* e *inflamación* promotoras de tumores.

De las características mencionadas las señales de crecimiento, evasión de la

apoptosis, angiogénesis sostenida y evasión de la destrucción por el sistema inmune tienen una relación estrecha con el proceso inflamatorio [17, 12, 18, 19, 20, 21]. La Invasión de tejido y metástasis también se ha asociado con células del sistema inmunlógico en modelos experimentales[22].

El uso de ensayos de alto contenido de información aunado a los resultados de experimentos de gran envergadura han mostrado la heterogeneidad molecular del cáncer de mama [2]. Es deseable ir un paso mas allá del manejo simultáneo de una gran cantidad de genes individuales con el fin de entender la estructura, dinámica, mecanismos de control y de diseño de sistemas biológicos, marco que toma en cuenta la biologá de sistemas [3]. En este contexto, los modelos de redes permiten realizar análisis estructurales y la generación de hipótesis de caracter global, en contraste con los análisis de genes individuales. Particularmente las redes de coexpresión que se infieren a partir de datos de microarreglos de RNA mensajero revelan los conjuntos de genes que actúan de manera coordinada en la forma de comunidades.

2.2. Fisiopatología del proceso inflamatorio

La inflamación es una respuesta a la alteración de la homeostasis, que ocurre debido a la infección y al daño en células y tejidos [23]. Las características evidentes de la inflamación incluyen: la hinchazón, el enrojecimiento de la zona afectada, el aumento de la temperatura local, dolor y la reducción de la funcionalidad normal del tejido [24].

Aunque la respuesta inflamatoria se identifica sobre todo con una respuesta aguda (una respuesta intensa y que aparece en un tiempo corto) también se reconoce la inflamación crónica como un estado de estimulación inflamatoria constante y que presenta una respuesta comparativamente menos intensa. La inflamación crónica está asociada a numerosas enfermedades crónico degenerativas como la aterosclerosis, la diabetes y la artritis [23].

Los componentes moleculares de una respuesta inflamatoria pueden clasificarse en cuatro grupos con función biológica específica: *inductores, sensores, mediadores y efectores*. Describiremos cada uno de ellos a continuación.

2.2.1. Inductores de la respuesta inflamatoria

Los *inductores* son moléculas indicadoras de un funcionamiento anormal, de daño estructural o funcional y de infección que funcionan activando vías de señalización específicas en la célula. Los inductores se clasifican en dos grandes grupos dependiendo de su origen. Inductores *endógenos* son aquellas moléculas propias de los tejidos pero que se encuentran aisladas dentro de compartimentos o bien en cantidades limitadas. Si existe un daño en el tejido, estas moléculas son producidas en cantidades anormales, o bien son liberadas de sus compartimentos habituales quedando expuestas a receptores específicos. Los inductores *exógenos* son moléculas de origen externo al organismo, que pueden ser de naturaleza microbiana o no microbiana.

Los inductores microbianos se clasifican a su vez en *PAMP*s (patrones moleculares asociados a patógenos del inglés: *pathogen-associated molecular patterns*) y *factores de virulencia*. Los PAMPs son patrones moleculares conservados dentro de un grupo de microorganismos y no están restringidos a patógenos, ya que están presentes también en microorganismos comensales. Pueden incluir proteínas bacterianas como flagelina, componentes de membrana plasmática como el ácido lipoteicoico y lipopolisacáridos, componentes de la pared celular como peptidoglicanos, así como productos de la ruptura celular incluyendo DNA de doble cadena.

Los factores de virulencia están asociados a la patogénesis e incluyen numerosas toxinas y proteasas que ayudan en el proceso de invasión al hospedero. La detección de factores de virulencia se da principalmente en forma indirecta, gracias a sus efectos disruptivos en las células y tejidos, lo que conduce a la producción de numerosos inductores endógenos, incluyendo fragmentos celulares y de matriz

extracelular. Existen numerosos agentes que pueden causar efectos similares a los patógenos en el organismo, incluyendo alérgenos como el polen, excretas de ácaros y esporas de moho que contienen proteasas, así como compuestos químicos irritantes y tóxicos y partículas no degradables como el hollín del humo de tabaco.[23]

Los inductores endógenos denominados en conjunto como *DAMP*s (del inglés *danger associated molecular patterns*) pueden ser derivados intracelulares, de la membran plasmática y de la matriz extracelular. Un tema común en la detección de DAMPs es la pérdida de compartamentalización en células y tejidos que en condiciones normales mantiene numerosas enzimas y sustratos separados por barreras físicas, pero que interactúan cuando la integridad del tejido se encuentra comprometida. Al sufrir daño o por el efecto de toxinas, la membrana plasmática permite la salida de moléculas pequeñas como ATP, iónes como K^+ y algunas proteínas como HMGB1 [25]. La concentración de Ca^{++} intracelular es un indicador de la integridad de las membranas. El citoplasma tiene normalmente una concentracion baja de Ca^{++} pero éste puede ingresar al formarse poros en la membrana plasmática o escapar del retículo endoplásmico con la consecuente activación de vías de señalización de daño [26].

Al ocurrir una herida, la ruptura de vasos sanguíneos en el tejido permite que el plasma y células sanguíneas entren en contacto con la matriz extracelular [21]. Moléculas presentes en el plasma sanguíneo como el factor de Hageman son activadas por la colágena y otros componentes de la matriz extracelular desencadenando cascadas enzimáticas: la cascada de coagulación, de fibrinólisis, cascada de kallikreina-cinina y cascada del complemento, todas éstas capaces de generar mediadores inflamatorios [23]. Al perderse la compartamentalización del tejido, se produce la activación de las plaquetas que liberan citocinas y factores de crecimiento contenidas en sus gránulos de secreción. Otros derivados de la degradación de la matriz extracelular incluyen polisacáridos derivados del ácido hialurónico y proteínas extracelulares glicosiladas por vías no enzimáticas (ibídem).

2.2.2. Sensores y activación de vías que conducen a la activación transcripcional

El *sistema inmune innato* es el principal disparador de la respuesta inflamatoria aguda. La tarea de identificar la presencia de patógenos o de daños al tejido se lleva a cabo por células especializadas entre las que se encuentran macrófagos y células dendríticas [27]. Otros tipos celulares no profesionales (esto es, que no están especializados en la detección y respuesta a antígenos) como células endoteliales, epiteliales y fibroblastos también contribuyen a la respuesta inmune innata [28] [29].

Los receptores de reconocimiento de patrones o *PRR*s son los principales sensores en el sistema inmune innato. Se han identificado cuatro tipos de PRRs: Receptores tipo Toll o *TLR*s, receptores tipo-C de lectina o *CLR*s, receptores tipo RIG-I-like o *RLR*s y receptores tipo NOD o *NLR*s [30]. La detección de PAMPs o DAMPs por medio de los PRRs activa vías de señalización que favorecen la transcripción de genes involucrados en la respuesta inflamatoria. Éstos genes son expresados de manera coordinada formando módulos transcripcionales [31].

Los *TLR*s son receptores transmembranales ubicados en la membrana plasmática y en la membrana de endolisosomas. Entre sus ligandos que pueden clasificarse como PAMPs y DAMPs se encuentran componentes de la membrana plasmática de bacterias: triacil lipoproteínas, diacil lipoproteínas, lipoproteínas y lipopolisacáridos. El componente del flagelo bacteriano, flagelina (TLR5) y componentes asociados a protistas como moléculas parecidas a profilina son reconocidos por TLRs de membrana. Varios tipos de ácidos nucléicos incluyendo RNA de cadena sencilla, RNA de doble cadena y CpG DNA que pueden ser resultado de la actividad de virus, bacterias y protistas intracelulares son reconocidos por TLRs de endolisosomas [30].

Proteínas que se encuentran fuera de sus compartimentos habituales, como el caso de HSP60, son capaces de activar receptores como TLR4; HMGB1 se ha asociado como ligando de TLR4 y TLR2. En condiciones patológicas es liberado DNA

mitocondrial que activa los receptores TLR9 endosomales [32](para una descripción más detallada de los ligandos de TLRs se recomienda [33].

La unión de ligandos a los receptores TLR permite el reclutamiento de las moléculas adaptadoras que inician sus respectivas cascadas de señalización. Existen al menos cuatro adaptadores que interactúan directamente con TLRs: MyD88, TIRAP, TRIF y TRAM. MyD88 es el primer adapatador descrito para los receptores TLR [33].

La cascada de señalización iniciada por MyD88 requiere de la formación de una serie de complejos intermediarios que llevan a la activación del factor de transcripción NF-κB. TAK1 (*transforming-growth-factor -β-activated kinase*) que es activado luego de la ubiquitinación de TRAF6 en el citoplasma tiene la capacidad de fosforilar al complejo IKK que regula la actividad de NF-κB. TAK1 también participa en la fosforilación de MAP cinasas como JNK (*JUN-N terminal kinase*) que está involucrada en la activación del factor de transcripción AP-1 cuyos blancos transcripcionales incluyen citocinas como RANTES (CCL-5), IL-8 y GM-CSF [34]. TRIF por su parte tiene importancia en la activación de NF-κB independiente de MyD88. TRIF es también un intermediario importante en la activación de la transcripción de Interferones tipo I (que incluye IFNα y β) a través de IRF3 (Figura 2.1).

Los receptores tipo NOD o *NLR*s son una familia de receptores citoplásmicos. Algunos NLRs se han implicado en la formación de inflamasomas que son plataformas moleculares que permiten la activación de Caspasa-1 y la maduración de IL-Iβ. El ejemplo modelo es el inflamasoma NLRP3. Después de su activación, NLRP3 se dimeriza asociándose con el adaptador ASC. ASC se une con CASP1 que se activa autocatalíticamente formando Caspasa-1. Caspasa-1 procesa la pro-IL-1β para formar IL-1β. La activación del inflamasoma NLRP3 se ha asociado con PAMPs y señales de daño en la célula incluyendo desbalance iónico por entrada de Ca^{2+} o salida de K^{+} y la producción de ROS (especies reactivas de oxígeno) [35].

NOD1 y NOD2 son receptores citosólicos que responden a peptidoglicanos. Tras su activación, NOD1 y NOD2 pueden reclutar a RIPK2 que es una cinasa relacionada en la fosforilación de TAK1 y en la ubiquitinación de NEMO (IKKγ), lo que lleva a la activación corriente abajo de MAP cinasas y NF-κB y subsecuentemente a la inducción de peptidos antimicrobiales, citocinas y quimiocinas. El inflamasoma NLRP3 se ha asociado con el procesamiento de IL-1β, IL-18 y a una muerte celular pro-inflamatoria denominada *piroptosis* [36].

Entre las moléculas que controlan la activación del inflamasoma NLRP3 se encuentra BCL2 cuya pérdida de función se ha asociado con el proceso de tumorigénesis. Existe evidencia a favor de que la inflamación mediada por el inflamasoma NLRC4 tiene un papel protector antitumoral y está relacionado como uno de los genes activadores de muerte celular regulados por p53 [37]. CIITA, es un NLR que funciona como regulador transcripcional de las moléculas MHC de clase II [35] [36].

Los receptores tipo RIG o *RLRs* son una familia de receptores citoplásmicos que reconocen RNA de doble cadena derivado de la replicación viral. Se reconocen al menos tres miembros: RIG-I, MDA5 y LGP2[33]. La señalizacion de RLRs lleva a la transcripción de interferones tipo-I por medio de la activación de IRF3 e IRF7, así como de otras citocinas por medio de NF-κB. RIG-I es regulado por medio de poliubiquitinación [33]. RIG-I y MDA5 interactúan con la proteína mitocondrial MAVS. MAVS tiene un papel como intermediario en la activación de las vías de NF-κB y de IRF3/IRF7 a través de TRAF3. La E3 ubiquitin ligasa Triad3A regula negativamente a TRAF3 por medio de poliubiquitinación [38].

Los receptores de lectina tipo C o *CLR*s se expresan principalmente en células dendríticas (DCs). La señalización por CLRs se lleva a cabo principalmente en respuesta a carbohidratos (de ahí el nombre de lectinas) que contienen manosa, fucosa o D-glucosa presentes en componentes virales, hongos y micobacterias. El

reconocimiento por parte de CLRs conduce a la internalización del patógeno y a la subsecuente presentación de antígenos [38].

El proceso de presentación de antígenos en células dendríticas implica la captura de antígenos, transporte intracelular y degradación de éstos, la unión de los péptidos resultantes a moléculas MHC tipo I y II y la subsecuente exposición en la membrana plasmática de los complejos [39]. La presentacion de antígenos por las células dendríticas tiene un papel primordial en la activación de células T y células B del sistema inmunológico durante el desarrollo de la respuesta inmune adaptativa [40].

Los CLRs asociados a membrana se clasifican en dos grupos: el grupo I de receptores de manosa y el grupo II de receptores de asialoglicoproteína que incluyen a las subfamilias dectina 1 (lectina tipo C 1 asociada a células dendríticas o CLEC7A) y el inmunoreceptor de células dendríticas (CLEC4A).

La señalización por CLRs puede tener efectos tanto en la inducción de genes proinflamatorios como en la regulación de otras vías, particularmente las de TLRs [33]. Dectina 1 al unirse a carbohidratos de origen fúngico se dimeriza y es fosforilada, lo que induce el reclutamiento de SYK (tirosina cinasa del bazo). SYK conduce a la supresión de NF-κB a través de NIK, o bien reclutando a CARD9, BCL-10 y MALT1 que regulan el complejo IKKγ/IKKβ/IKKα que ubiquitina y conduce a la degradación de IκBα quien es el represor de NFκB.

La interacción de los CLRs con la señalización por TLRs se ha asociado con la supresión de la transcripción de genes proinflamatorios entre los que se encuentran interferones tipo I (IFNa), IL-12a, IL-6 y TNF. La proteína SHP1 es un adaptador clave en la inhibición de TLR8 y TLR9 por DCIR y de TLR4 por MICL sin embargo la ruta precisa de señalización aún no se ha resuelto [41].

2.2.3. Mediadores y efectores de la respuesta inflamatoria

La vía del factor nuclear NF-κB ha sido considerada como el prototipo de vía de señalización proinflamatoria, principalmente en la expresión de genes entre los que se encuentran citocinas, quimiocinas y moléculas de adhesión [42]. Estas moléculas funcionan como mediadores inflamatorios y su acción está asociada a vías de señalización que inducen un cambio en el comportamiento de tipos celuares presentes en el microambiente afectado y la expresión de genes en respuesta a la integración de estímulos.

Las quimiocinas funcionan como quimioatrayentes para los leucocitos y su efecto se lleva a cabo por medio de receptores de tipo GPCR (receptores acoplados a proteínas G). Estas vías de transducción de señales se caracterizan por la producción de segundos mensajeros como AMPc y lípidos derivados de la membrana que a su vez activan cinasas involucradas en eventos de remodelación del citoesqueleto de actina y el tráfico vesicular, lo que puede conducir a la secreción del contenido de vesículas, a la exposición de ligandos y receptores membranales secuestrados y al movimiento celular [43].

El efecto de citocinas se lleva a cabo por medio de receptores de tirocina cinasa u otros asociados a cinasas que conducen a la activación de factores de transcripción y la subsecuente expresión de genes. En el caso de los interferones tipo I la unión al receptor formado por dos subunidades IFNAR1/IFNAR2 permite el reclutamiento de JAK1 (*janus kinase 1*) y TYK2 (tirosina cinanasa 2). La fosforilacion del receptor permite el reclutamiento de proteinas STAT que de manera dependiente del contexto pueden formar diversos complejos con efectos regulatorios diferentes: STAT1/STAT2/IRF9 activan genes de respuesta antiviral, el homodímero STAT1/STAT1 puede unise a secuencias GAS en el DNA y favorecer la transcripción de genes proinflamatorios y el homodimero STAT3/STAT3 induce la trascripción de genes que codifican para proteínas involucradas en la represión de vías inflamatorias [44].

La expresión de estos conjuntos de genes se ve reflejada en el fenotipo celular. En el caso de los leucocitos, la combinación de estímulos recibidos permite reclutar células precursoras en circulación a través de las paredes de los capilares y su posterior diferenciación. La señalización mediada por quimiocinas está involucrada en la activación de integrinas que a su vez median la interacción y diferenciación de leucocitos entre los que se encuentran las células T [45]. Otro subtipo celular denominado MDSCs (*myeloid derived suppressor cells* o células supresoras derivadas de precursores mieloides) tiene un papel inmunoregulatorio controlando la expansión y diferenciación de células T y es inducido por señales que incluyen IFNγ, ligandos de TLRs y citocinas que incluyen a IL-13, IL-4 y TGFβ. Éstas moléculas ejercen su función través de vías de señalización asociadas a varios STATs y NF-κB [46].

Los macrófagos constituyen un grupo de células mieloides de gran plasticidad en respuesta a la integración de señales. La estimulación con IFNγ y TNF favorece el desarrollo de macrófagos de tipo M1 que producen IL-12, iNOS y quimiocinas entre éstas CCL15, CCL20, CXCL9, 10 y 11. La estimulación con IL-4 favorece la diferenciación a un fenotipo M2 productor de CCL18, RELMα, CCL17, IL-27rα, IGF1 y Factor XIII-A que coordinan la cicatrización por medio de la deposición de matriz extracelular, el reclutamiento de otras células del sistema inmune y la inhibición de citocinas pro-inflamatorias [47]. (para una descripción más detallada de la diferenciación de células mieloides y su relevancia en cáncer se recomienda [48]).

En un punto intermedio entre las respuestas innata y adaptativa, el sistema del complemento puede explotar la presencia de anticuerpos sobre la superficie de patógenos o de células que presentan antígenos por medio del complejo mayor de histocompatibilidad MHC para montar una respuesta no celular. La vía clásica del complemento es iniciada por el complejo C1Q al unirse con inmunoglobulinas IgM o IgG asociadas con atígenos en la membrana. C1Qa y C1Qb promueven la asociación de varias moléculas en una cascada de proteólisis. Al final de la cascada C5b puede unirse a C6 C7 C8 y C9 para formar el llamado complejo de ataque

a la membrana o MAC que es capaz de lisar a las células formando poros en la membrana (y posiblemente estimulando la producción de DAMPs)[49].

Varios de los productos proteolíticos de la cascada del complemento funcionan como quimioatrayentes reclutando y activando fagocitos hacia el sitio de activación [50]. C3a es una anafilatoxina (péptido del sistema del complemento capaz de causar la desgranulación de las células cebadas y liberación de histamina) que actúa a través de su receptor C3AR1, capaz de atraer e inducir la degranulación de células cebadas en un modelo de asma en ratón [51]. En un modelo experimental de ratón se ha observado la dependencia de C5a para el desarrollo de tumores asociado a la inmunosupresión de los linfocitos CD8+ por MDSCs [49]. Alternativamente, lectinas que se unen a manosa también conocidas como MBL reconocen patrones de glicosilación encontrados en patógenos como bacterias, levaduras y algunos virus. Las MBL forman trímeros que se asocian con proteasas de serina denominadas MASP (MASP1, MASP2 y MASP3), formando una C3 convertasa que continúa con la cascada del complemento.

La respuesta inflamatoria implica una supervisión permanente del funcionamiento de células y tejidos. Esta supervisión se lleva a cabo por medio de una gran cantidad de mecanismos de seguridad que pueden detectar señales de peligro internas y externas y responder de manera apropiada. Dependiendo del contexto, la detección de señales de peligro puede llevar a la activación de programas específicos que desencadenan las señales de alarma en forma de mensajeros proinflamatorios, con efectos que pueden incluir la muerte de células funcionalmente afectadas por medio de apoptosis, o bien poner un alto en la destrucción de tejido por medio de la inmunosupresión y el inicio del programa de cicatrización [23].

Se ha observado que los tumores tienen un componente inmunológico en forma de infiltración celular por leucocitos [20] [9]. Además de poseer características como angiogénesis defectuosa [13] y metabolismo alterado. Dado que la respuesta inflamatoria involucra programas de regulación genética en respuesta a las señales

de peligro en los tejidos [31], explorar las redes de regulación transcipcional basadas en genes que codifican proteínas del proceso inflamatorio será de utilidad para entender la relación entre éste y el desarrollo de los tumores primarios de mama.

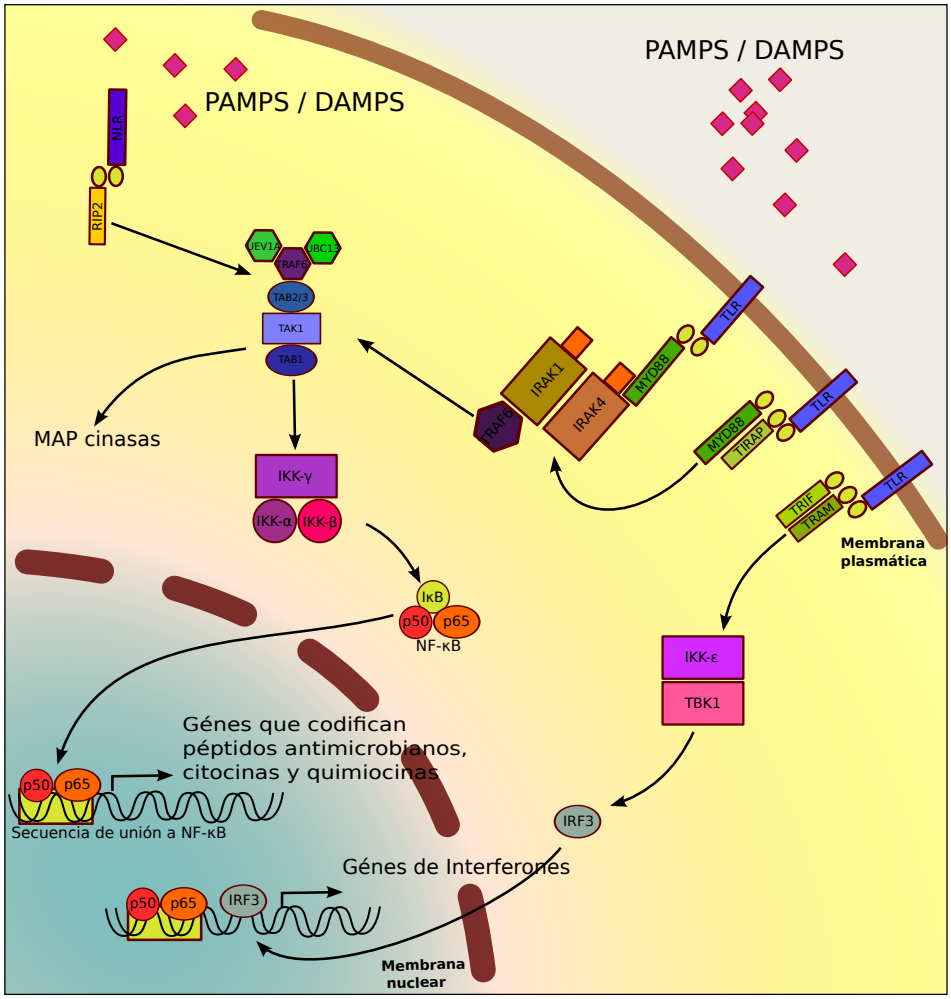


Figura 2.1: Despues de la union a sus ligandos (DAMPs o PAMPs), los receptores de reconocimiento de patrones inician cascadas de señalización que conducen a la activación transcripcional de genes. Los TLRs utilizan diversos adaptadores con dominios TIR como MYD88, TRAF,TRAM, TRIF y TIRAP. Estos adptadores reclutan enzimas que permiten la modificación covalente de complejos moleculares involucrados en la activación o represión de factores de transcripción. TRAF6 es una ubiquitin ligasa que una vez activada forma un complejo con TAK1/TAB1/TAB2. Una vez translocado al citosol se une a las ubiquitin ligasas UEV1A y UBC13 lo que le permite activar la vía de MAP cinasas y por otro lado a IKK-γ. IKK-γ tambien conocido como NEMO se asocia a IKK-α e IKK-β en un complejo que fosforila a IκB lo que favorece su ubiquitinación y degradación. Esto libera a NF-κB quien se transloca al núcleo donde se une a secuencias específicas en el DNA iniciando la transcripción de genes de respuesta inflamatoria. Otros receptores como NLRs señalizan a través de vías comunes en lo que se conoce como *entrecruzamiento*, en este caso por medio de moléculas compartidas entre vías como TRAF6.

Capítulo 3

Objetivos

- Calcular las redes de regulación transcripcional basadas en genes que codifican proteínas del proceso inflamatorio, a partir de datos de expresión de genoma completo de muestras de cáncer primario de mama.
- Identificar los procesos biológicos representados por los genes en la red.
- Identificar módulos transcripcionales (comunidades) en la red y su relación con procesos biológicos.
- Identificar los patrones de expresión diferencial de los procesos encontrados, es decir, si los genes en los procesos están siendo sobreexpresados o subexpresados.
- Identificar de qué manera se encuentran distribuidos procesos biológicos conocidos, que son importantes en la respuesta inflamatoria, dentro de la red.

3.1. Hipótesis

El desarrollo del cáncer implica la formación de tejidos de manera anormal, los cuales se caracterizan por presentar angiogénesis anómala [13], y un metabolismo alterado. Dada la ubicuidad de los mecanismos de detección de peligro en el organismo [25], parece razonable pensar que en tejidos como los tumores exista una

estimulación constante de éstos mecanismos de defensa.

Los mecanismos de respuesta inflamatoria llevan a la activación transcripcional de genes, lo que significa cámbios en la cantidad de RNA mensajero disponible. Ya que el cáncer de mama ha mostrado ser altamente hterogéneo a nivel de genes individuales, proponemos la alternativa de buscar procesos biológicos y vías metabólicas, aprovechando las ventajas de las tecnologías de alto desempeño como los microarreglos de expresión de genoma completo [3].

Por medio de la inferencia de redes de regulación transcripcional y el análisis de las mismas, esperamos encontrar módulos transcripcionales [31](Comunidades en la red sección 5.2.2) que representen procesos biológicos que se encuentran regulados. Partiendo de los genes reconocidos como inflamatorios, buscamos a los procesos biológicos asociados a la inflamación, característicos del fenotipo del cáncer primario de mama.

Sobre los datos obtenidos con microarreglos de expresión de RNA mensajero

El desarrollo de los microarreglos de DNA se ha beneficiado de la tecnología de microcircuitos electrónicos, en especial los GeneChip© de Affymetrix hacen uso de técnicas de fotolitografía para fabricar lo que se conoce como microarreglos de alta densidad [52]. Las plataformas de Affymetrix como GPL96 usan colecciones de sondas de DNA que capturan transcritos específicos. GPL96 utiliza una parte (22284 sondas) del conjunto de sondas Human Genome U133 que contempla alrededor de 33000 transcritos (Para una explicación detallada del funcionamiento de los microarreglos de expresión se recomienda consultar [4]).

La tecnología de microarreglos ha facilitado el tránsito desde el estudio de unos pocos genes, o en el mejor de los casos de vías biológicas aisladas, hacia investigaciones de caracter global sobre la actividad celular [53]. Entre las aplicaciones prácticas de los microarreglos de expresión se encuentran el perfilado transcripcio-

nal, genotipificación, análisis de variantes de splicing, interacciones entre DNA y proteínas (*ChIP-on-chip*) y otras (ver, por ejemplo [53] y [52]).

Los datos resultantes de estudios realizados con estas plataformas de alto contenido de información se encuentran en muchos casos disponibles de manera pública en bases de datos como el *Gene Expression Omnibus* (GEO) [54]. Los resultados en crudo procedentes de varios experimentos pueden reunirse, estandarizarse y ser utilizados en la inferencia de redes donde un gran número de muestras otorga robustez estadística [55].

Para este trabajo los datos de los microarreglos de expresión icluyeron todos los subtipos moleculares reconocidos de cáncer de mama. Las muestras son de cáncer primario, es decir, en casos clínicos que hasta ese momento no presentaban metástasis. Las muestras son de pacientes individuales y no se realiza ningun tipo de ordenamiento cronológico que nos permita realizar series de tiempo o estudios de progresión de la enfermedad. En lugar de esto damos un vistazo a la variabilidad del cáncer primario de mama para buscar patrones generales de regulación transcripcional.

Es importante reconocer que en el caso de los genes que codifican proteínas la función biológica final está sujeta a mecanismos de control transcripcional [31], posttranscripcional [56], traduccional y a modificaciones posttraduccionales que no pueden ser medidas por ésta tecnología de microarreglos de expresión, sin embargo la cantidad de RNA mensajero es una aproximación ya que la producción de proteínas depende de la presencia de su RNA mensajero.

Adicionalmente debemos considerar que los tumores están formados de una variedad de tipos celulares que incluyen células cancerosas, células del sistema inmunológico, células vasculares y células mesenquimales, cada una de las cuáles sigue programas regulatorios diferentes [20]. El análisis de muestras de tumor contiene potencialmente estos tipos celulares por lo que las redes regulatorias obtenidas

corresponen al tumor y no necesariamente a tipos celulares específicos.

Capítulo 4

Materiales y métodos

4.1. Datos de expresión y lista de genes del proceso inflamatorio

Datos. Se utilizó una base de datos curada de microarreglos de expresión de genoma completo basados en la plataforma HGU133a de Affymetrix [55]. Estos datos corresponden a 819 muestras de cáncer primario de mama y 61 muestras de tejido mamario normal. Todas las muestras provienen de pacientes que presentaron tumores primarios invasivos de mama que no fueron tratadas previamente con ningún medicamento. Los datos crudos fueron descargados del *Gene Expression Omnibus* (GEO) [54], que es una base de datos pública mantenida por el NCBI. Esta base de datos fue procesada, haciendo análisis de control de calidad, normalización, corrección de efecto de lote y análisis de expresión diferencial. Aquí aprovechamos los datos normalizados y de expresión diferencial para generar y analizar las redes resultantes.

Análisis de expresión diferencial. El análisis de expresión diferencial consiste en la comparación de los niveles de expresión de genes entre condiciones experimentales [57]. Para encontrar cuáles son los genes que cambian su nivel de expresión utilizamos dos estadísticos: *log-fold change* y estadístico B definidos como sigue:

- ***log-fold change*** ($logFCh$). Es la proporcion de cambio en el nivel de expresión entre condiciones experimentales expresado en logaritmo base 2. Permite conocer la magnitud del cambio en el nivel de expresión. Para dos condiciones H y K con niveles de expresión en un mismo gen g_H y g_K respectivamente $logFCh = \log_2(g_h) - \log_2(g_K)$

- **Estadístico** B. Es un indicador de consistencia estadística en los niveles de expresión de un gen. La posibilidad (odds) a favor de un evento es la proporcion entre la probabilidad e que este ocurra p respecto a que no ocurra $1 - p$. B se define como

$$B = \log\left(\frac{p}{1-p}\right)$$

  La probabilidad asociada a cada valor de B se calcula de la siguiente manera:

$$p = \frac{e^B}{e^B + 1}$$

  B=6 equivale a un nivel de confianza de 99.8 %

Inferencia estadística y valores de probabilidad. Existen varios enfoques para la prueba de hipótesis. Entre los métodos mas usados se encuentra el uso del valor de probabilidad o $p - value$ propuesto por Ronald Fisher (por ejemplo en [58]) y el uso de la significancia estadística α de Neyman y Pearson [59]. El uso del $p - value$ se basa en una prueba por contradicción, en donde se calcula la probabilidad de obtener un resultado al menos tan extremo como el observado por medio de un modelo nulo, siendo rechazado si esta probabilidad es razonablemente baja (usualmente un 5 % o menor) . La significancia α por otro lado es utilizada como una manera de ajustar el poder estadístico de una prueba, de manera que la probabilidad de rechazar un resultado positivo sea minimizada [60].

Gene Ontology. El Consorcio de Ontología de Genes (*Gene Ontology Consortium*) es una organización que tiene el propósito de producir un vocabulario común, preciso y estructurado para la descripción del papel de los genes y sus productos en cualquier organismo [61]. La ontología de genes (GO) está consti-

tuida por tres categorías independientes: Proceso biológico, componente celular y función molecular. Las categorías de GO forman una red con una estructura jerárquica de términos. Proceso biológico indica el *objetivo biológico* con el cual el gen o su proteína contribuye. Función molecular se refiere a la actividad bioquímica, incluyendo la unión específica a ligandos o estructuras del producto de ese gen. Componente celular refiere a la parte de la célula donde el producto de ese gen se encuentra activo (para una información más detallada se sugiere consultar `geneontology.org/page/ontology-documentation`).

Genes del proceso inflamatorio. Para determinar la asociación con los procesos inflamatorios se extrajo una lista de genes anotados como inflamatorios en el archivo de anotación de la plataforma. El archivo de anotación para la plataforma HGU133a se obtuvo de la página web del fabricante (Affymetrix). Para procesar el archivo de anotación de la plataforma, se elaboró un script de R que lee la categoría *GO Biological Process* para encontrar la palabra clave *inflammatory*, lo que incluye procesos inflamatorios agudos, crónicos y cualquier otro reportado como relacionado a respuesta inmune e inflamación. El script extrae de las columnas ID, Gene Symbol y GO biological process únicamente los renglones del archivo de anotación que contengan la palabra clave. La lista obtenida cuenta con 626 identificadores que corresponden a 418 genes.

Los datos de expresión diferencial de los genes fueron obtenidos previamente comparando las muestras de tejido sano contra las muestras de tumores. El nivel de expresión (fold change) está expresado en logaritmo base 2. Se utilizaron como puntos de referencia valores de expresión de la mitad o menor que el normal ($logFCh \leq -1$) para considerar un gen subexpresado y el doble o mayor que el normal ($logFCh \geq 1$) para considerar un gen sobreexpresado. El estadístico B se tomó como medida de confiabilidad a partir de un valor de 6 o mayor como se describe en [57].

4.2. Inferencia de redes de regulación transcripcional a partir de datos de expresión de RNA mensajero de genoma completo

Una red G en su definición más simple contiene un conjunto de elementos V (nodos) y un conjunto de relaciones E entre ellos (aristas) $G = \{V, E\}$. Una arista puede representar cualquier tipo de relación, incluyendo interacciones moleculares y correlaciones estadísticas. Las aristas pueden describirse como presentes/ausentes (por ejemplo con los valores 0 y 1) o bien pueden tener una cuantificación asociada tambien denominada *peso*. Si añadimos un signo, también pueden indicar direccionalidad. Una red donde las aristas tienen un valor asociado se denomina *pesada*.

Una red en la que las aristas tienen una dirección se denomina *dirigida*. El número de nodos al que un nodo se conecta directamente se denomina grado (k). Al conjunto de nodos y aristas que forman una unidad conectada se le denomina componente. Una red puede constar de uno o varios componentes. Se denomina *componente principal* a aquel que posee la mayor cantidad de nodos. Para una descripción extensiva del concepto de redes y sus medidas asociadas se recomienda consultar [62].

La inferencia o *ingeniería reversa* de redes de genes es el proceso de identificar interacciones entre genes a partir de datos experimentales por medio del análisis computacional [5]. Los métodos de la Teoría de la Información resultan de especial utilidad cuando existen dependencias no lineales y en el contexto de experimentos en los que el numero de variables es mucho mayor al numero de muestras disponibles; condiciones que se cumplen al utilizar datos de microarreglos de expresión y que representan el núcleo de este trabajo [63].

4.2.1. Información mutua

En el campo de la investigación biomédica los experimentos a nivel de genoma completo se caracterizan por un gran numero de variables (nivel de expresion de muchos genes), que no necesariamente se encuentran distribuidas uniformemente, con un numero reducido de réplicas experimentales y ruido inherente a la obtención y procesamiento de los datos. Los análisis estadísticos como la correlación de Pearson que están basados en dependencias lineales resultan inadecuados [64] [65].

La información mutua (MI) se define como $I(x,y) = S(x)+S(y)-S(x,y)$ donde $S(t)$ es la entropía de Shannon para una variable arbitraria t. Para una variable discreta, la entropía se define como $S(t) = -\langle logp(t_i)\rangle = -\sum_i p(t_i)logp(t_i)$ donde $p(t_i) = Prob(t = t_i)$ es la probabilidad de cada estado de la variable. La entropía es una medida de *incertidumbre* o *elección*, donde el caso para el cual todas las opciones son igualmente probables el valor de S es máximo. La información mutua para un par de variables puede ser evaluada de la manera siguiente:

$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\frac{p(x,y)}{p(x)\,p(y)}$$

La información mutua es una medida de la desviación que existe entre la distribucion de probabilidad conjunta $p(x,y)$ y las marginales $p(x)$ y $p(y)$ para un par de variables, siendo cero sí y solo sí se cumple la condición de independencia estadística $p(x,y) = p(x)p(y)$

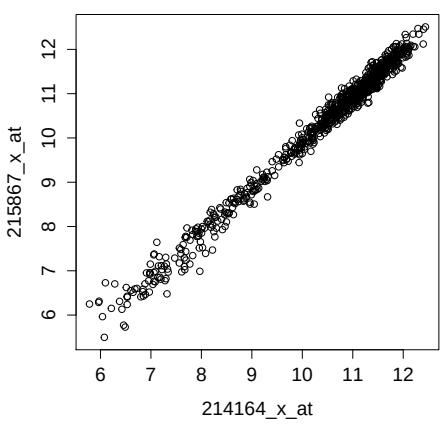
En el caso de los datos obtenidos por medio de microarreglos de expresión, contamos solamente con *vistas instantáneas* del estado de las células a lo largo de un gran numero de variables. Dado que el numero de muestras es pequeño en relación al numero de variables y la posibilidad de valores faltantes en la distribución de probabilidad discreta, es deseable poder estimar la distribución de probabilidad contínua, para lo cual el valor de MI se expresa como:

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

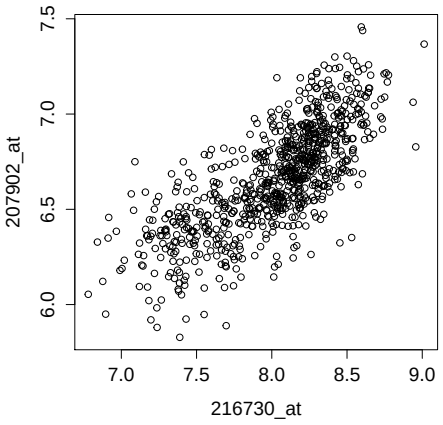
La inferencia de las redes de regulación transcripcional se llevó a cabo por medio del algoritmo basado en información mutua ARACNE [65]. Para aproximar la estrucutura de interacción el algoritmo basado en información mutua ARACNE realiza una estimación de la distribucion de probabilidad conjunta (JPD) para obtener el valor de MI [65].

El algoritmo funciona en dos pasos consecutivos: Primero calcula el valor de MI para cada par posible de genes a partir de las distribuciones de probabilidad estimadas. Si el valor obtenido sobrepasa un umbral predefinido se registra la interacción de ese par de genes con su valor de MI asociado (figura 4.1. En el caso de las interacciones de genes con un valor de MI alto (cercano a 1) al graficar los niveles de expresión de un gen contra el otro es posible observar el agrupamiento de los datos (figura 4.1(a)), mientras que valores de MI mas bajos ocurren con valores dispersos de expresión conjunta (figuras 4.1(b) 4.1(c)).

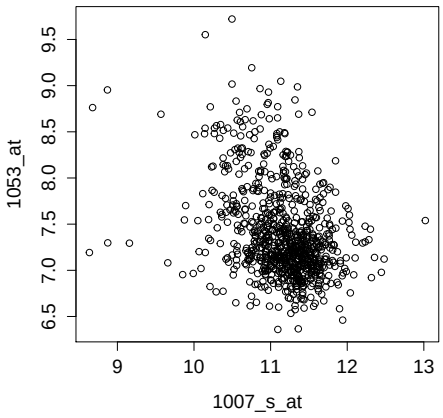
En un segundo paso ARACNE puede deshacerse de interacciones que sean sospechosas de representar falsos positivos utilizando la Desigualdad de Procesamiento de Datos (DPI)[64] buscando triángulos de interacciones en la red y eliminando la interacción con valor de MI más pequño si se encuentra debajo de un punto de corte arbitrario establecido por el usuario. En el caso particular de nuestras redes la tolerancia para el uso de DPI fue de MI ≥ 0 debido a que utilizamos un valor de por sí estricto en el primer paso.



(a) MI=1.03686



(b) MI=0.40775



(c) Sondas no asociadas en la red por tener un MI por debajo del punto de corte

Figura 4.1: Gráficos que muestran la distribución conjunta de los niveles de expresión de tres pares distintos de genes. El agrupamiento de los valores de expresión aparece en relación al valor calculado de MI para ese par de genes. Los ejes x y y representan los valores de expresión normalizados de cada gen y están rotulados con el identificador (AffyID) del gen para la plataforma. Para que dos sondas compartan una arista en la red es necesario que $MI \geq 0.4$, lo que significa una confianza estadística de $p \leq 1 \times 10^{-100}$ para este conjunto de datos. Las interacciones (a) y (b) forman parte de la red mientras que (c) no fué incluida por estar debajo del punto de corte de MI.

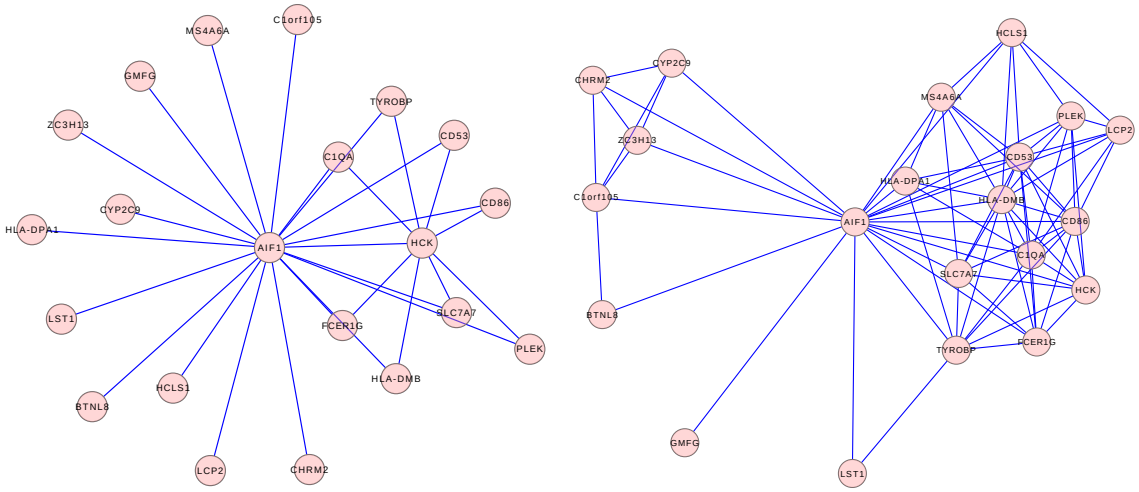
El algoritmo ARACNE tiene un escalamiento computacional del orden de $O(N^3 + N^2M^2)$ donde N es el número de genes y M el número de muestras [65]. Para reducir el tiempo de cálculo de las redes se utilizó un servidor Dell *Power Edge* T620 de 32 procesadores intel Xeon a 2.6GHz con 384GB en RAM. Se implementó una paralelización del cómputo sacando ventaja de que ARACNE toma de manera independiente el cálculo de información mutua para cada gen en la lista de entada. La paralelización se llevó a cabo mediante el software HT Condor [66]. HT Condor utiliza un esquema de administración de trabajo que permite utilizar recursos computacionales distribuidos, en este caso los 32 procesadores del servidor.

Parámetros. La redes fueron inferidas utilizando un $p \leq 1 \times 10^{-100}$ con valores de $MI \geq 0.4$. Debido al nivel de significancia estadística con el que se incluyeron las interacciones en la red no se consideró necesaria la eliminación de interaciones indirectas por lo que la tolerancia de DPI se tomó con valor de $p = 1$.

Las redes fueron calculadas en dos pasos sucesivos (dos iteraciones): En un primer paso se tomó la lista de genes inflamatorios contra todos los genes en el genoma, lo que produce numerosos vecinos con una sola interacción en la red. En un segundo paso, los genes obtenidos en la primera red fueron usados para calcular una nueva red que recupera todas las interacciones no consideradas por el primer paso (figura 4.2).

Red de primera iteración. Primero, una red de la lista de conjuntos de sondas anotados en Gene Ontology como de inflamación contra el genoma completo (Figura 4.2(a)). La red resultante contiene las relaciones al interior de la lista de conjuntos de sondas de inflamación y los primeros vecinos aunque no estén anotados en esa categoría. Las redes obtenidas por este método son redes pesadas,donde el peso de la arista corresponde al valor de información mutua entre los nodos conectados y son no dirigidas.

Red de segunda iteración. Para conocer la relación entre los genes que son primeros vecinos, pero que no estaban en la lista inicial de genes inflamatorios, se generó una segunda red usando todos los conjuntos de sondas (Affy IDs) de la primera red (Figura 4.2(b)). Esta aproximación permite obtener las interacciones entre genes sin necesidad de realizar el cálculo de la red de genoma completo que resulta computacionalmente más costoso. Los análisis de enriquecimiento estadístico y la extracción de comunidades fueron realizados solamente en esta red.



(a) (b)

Figura 4.2: Redes de primera y segunda iteración. Subredes correspondientes a los primeros vecinos de AIF1 que constan de los mismos 21 genes. La red de primera iteración (a) tiene 28 aristas. Asocia los genes de inflamación contra todos los genes en el genoma con los que tengan una correlación estadística robusta. La segunda iteración (b) tiene 78 aristas. las aristas adicionales permiten conocer la asociación existente entre los primeros vecinos que no son genes inflamatorios (en la lista inicial)

4.3. Procesamiento de las redes de regulación transcripcional

Las redes calculadas con ARACNE fueron procesadas para facilitar su análisis en las diferentes herramientas, lo que incluye traducir el formato de archivo

generado en ARACNE (`.adj`) a un formato legible por Cytoscape (`.sif`)[67], la asignación de nombres y otros atributos a los nodos, incluyendo los niveles de expresión y el mapeo visual de los mismos.

4.3.1. Curado de las redes

Para realizar la interpretación biológica de las redes es deseable contar con el nombre de los genes [68]. La traducción de los identificadores de la plataforma (Affy IDs) se hizo con un script de Python y una tabla de correspondencias extraída del archivo de anotacion de la plataforma (ultima version disponible del archivo de anotación). Los conjuntos de sondas (probesets) restantes sin anotación se tradujeron con la herramienta en línea Webgestalt [69]. Los identificadores que no pudieron ser traducidos y sus interacciones se descartaron de la red. Finalmente se eliminaron las interacciones duplicadas y las autointeracciones (bucles).

4.3.2. Análisis topológico y visualizaciones para las redes

Las visualizaciones de las redes se generaron utilizando el programa Cytoscape [67] en su versión 2.8. Cytoscape es una plataforma de código abierto (open source) centrada en el análisis de redes biológicas que permite la integración de información, incluyendo anotaciones visualización y análisis topológico. Cuenta con numerosas aplicaciones que pueden ser integradas a la interfaz de usuario y que permiten extender la funcionalidad por medio de herramientas de análisis, visualización, soporte de formatos de archivo y generación de *scripts*. Para mayor información se recomienda visitar la página oficial `www.cytoscape.org`

Como parte de las visualizaciones se tomó en cuenta el perfil de expresión de los genes respecto a las muestras control, marcandose los diferencialmente expresados con un estadístico B ≥ 6 y un log fold change $logFCh \leq -1$ para los genes subexpresados y $logFCh \geq 1$ para los sobreexpresados (logaritmos en base 2)[57]. El agrupamiento espacial de los nodos se llevó a cabo por medio del algoritmo

dirigido por fuerzas y pesado por aristas (edge-weighted force-directed layout). El algoritmo calcula una posición de equilibrio asumiendo un valor de repulsión entre nodos y atracción de las aristas lo que produce que los nodos más conectados entre sí sean cercanos y los nodos que comparten menos conecciones se alejen [70] (más información sobre las distribuciones de nodos en cytoscape está disponible en *http://wiki.cytoscape.org/Cytoscape_User_Manual/Navigation_Layout*).

El análisis topológico de las redes permite obtener parámetros que describen características estructurales de estas, tales como: la cantidad de componentes conectados, la cantidad de vecinos que tiene cada nodo, la tendencia al agrupamiento de los nodos, la cantidad de caminos mínimos entre nodos y su longitud. Se realizó el análisis de la red de segunda iteración por medio del plugin Network Analyzer [71] para Cytoscape. Una definición detallada de los parámetros puede encontrarse en [62] y [72], así como una discusión de algunas de estas medidas en redes biológicas [73].

Los parámetros reportados incluyen:

- **Número de nodos.** Es el número de nodos incluido en todos los componentes de la red.
- **Número de nodos aislados.** Nodos que no poseen ninguna arista.
- **Número de autointeracciónes.** Las autointeracciónes son aristas que tienen como destino al mismo nodo en que se originan. En estas redes fueron eliminados en el paso de limpieza por lo que su valor será cero.
- **Pares de nodos con aristas múltiples.** Se cuentan si un par de nodos comparte más de una arista. Estas redes sólo consideran un único valor por par de nodos por lo que esta medida es cero.
- **Número de componentes conectados.** Un componente conectado es un conjunto de nodos alcanzables por medio de aristas. Si dos nodos se encuentran en componentes diferentes no existe ningún camino que los una.

- **Número promedio de vecinos.** Es la conectividad promedio de los nodos de la red.
- **Caminos de longitud mínima.** Un camino dentro de la red se forma tomando el número de aristas recorridas para alcanzar un nodo a partir de otro. Dos nodos pueden ser alcanzados por uno o más caminos. El camino de longitud mínima es el camino más corto entre dos nodos. La longitud del camino más corto tambien se denomina *distancia*.
- **Longitud característica de caminos.** Es la distancia esperada entre dos nodos.
- **Diámetro de la red.** Es el tamaño máximo de los caminos mínimos en todos los componentes de la red.
- *Excentricidad.* Se define para un nodo como la longitud del camino mínimo mas largo que empieza en ese nodo.
- **Radio de la red.** Es el valor mínimo de excentricidad diferente de cero en la red.
- **Coeficiente de agrupamiento.** Se define como

$$C_n = \frac{2e_n}{k_n(k_n - 1)}$$

  donde k_n es el número de vecinos de n y e_n es el número de pares conectados entre todos los vecinos de n. C_n es la proporción entre el número de aristas entre los vecinos de un nodo y el número máximo posible de de aristas para ese número de vecinos.
- **Densidad de la red.** Es la proporción entre el número de aristas en la red con respecto a al máximo posible. Su valor es 0 si no existen aristas y 1 si se trata de un clique.
- **Hetrogeneidad de la red.** Refleja la tendencia de la red a tener *hubs*(nodos con una conectividad muy alta).

- **Centralización de la red.** Describe la tendencia de la red a tener estructura de *estrella* (nodos conectados solamente a un nodo central. Se expresa en valores de 0 (poco centralizado) a 1 (muy centralizado).
- **Distribución del grado de conectividad.** Es una medida de la diversidad de conectividad de la red completa. La distribucion de grado $P(k)$ da la fracción de nodos $N(k)$ que tienen un grado $k = 1, 2, 3...$ dividido por el número total de nodos N.

Para numerosas redes que modelan sistemas reales como las redes metabólicas, el internet o redes de interacciones sociales la distribucion de grado puede describirse con una ecuación de la forma $P(k) = Ak^{-\gamma}$, donde A es una constante que permite la suma de los $P(k)$ a 1, lo que se conoce como una ley de potencia. Esta función indica una gran diversidad de grados sin un valor característico o promedio. Estas redes son denominadas *libres de escala* [73][74].

Las redes reales poseen además características como un diámetro pequeño combinado con un alto coeficiente de agrupamiento, lo que se conoce como *propiedad de mundo pequeño* [75], la sobrerepresentación estadística de patrones (denominados *motifs*) [76] y estructura de comunidades [77]. Si se cumplen estas propiedades, las redes son denominadas *redes complejas* [78].

4.3.3. Identificación de comunidades de nodos en la red completa

Una comunidad es un subgrupo de nodos con una alta conectividad al interior y comparativamente baja al exterior. Un grupo de nodos completamente conectado se denomina *clique* o *grafo completo* [62]. La presencia de estos grupos altamente conectados se ha asociado comunmente a redes de sistemas reales donde se ha sugerido que pueden representar módulos funcionales [77].

En el contexto de una red de regulación transcripcional, una comunidad corresponde a un grupo de nodos estrechamente correlacionados en sus niveles de

expresión, lo que sugiere una regulación transcripcional coordinada [31]. La búsqueda de comunidades en la red de segunda iteración se llevó a cabo por medio de la herramienta para Cytoscape MCODE [79].

MCODE utiliza un algoritmo que identifica las regiones densamente conectadas en la red por medio del pesado de los nodos para lo que define las medidas siguientes:

- En una red $G = \{V, E\}$ con úmero de nodos $|V|$ y número de aristas $|E|$, la densidad de conexiones D_G es el cociente entre el número de aristas $|E|$ dividido por el número máximo de aristas posibles $|E|_{max}$.

$$D_G = \frac{|E|}{|E|_{max}}$$

- Para redes sin bucles (aristas que unen un vertice con sí mismo).

$$|E|_{max} = \frac{|V|(|V|-1)}{2}$$

- El k-núcleo es un subconjunto de nodos máximo (esto es, que no se pueda añadir ningún nodo extra conservando la propiedad definida) con una conectividad igual o mayor a k.
- El coeficiente de agrupamiento de núcleos (*core-clustering coefficient*) para un nodo v es la densidad del k-núcleo de valor mas alto en la vecindad de v incluyéndolo.

En un primer paso, MCODE asigna un peso a cada nodo de la red multiplicando el coeficiente de agrupamiento de núcleos por el nivel del k-núcleo más alto en la vecindad del nodo. Una vez pesados los nodos se toma como semilla el nodo con el mayor peso y se incluyen aquellos vecinos que estén por arriba de un umbral

determinado como un porcentaje del peso del nodo inicial. De manera recursiva son inspeccionados los vecinos de los nodos incluidos en cada paso, hasta que no se puedan añadir más nodos. El proceso es repetido para los nodos con una alta puntuación que no fueron vistos por los grupos anteriores. El algoritmo incluye un tercer paso de ajuste para eliminar o añadir nodos basado en la conectividad local.

4.3.4. Análisis de enriquecimiento estadístico de conjuntos de genes

El análisis de enriquecimiento estadístico nos permite relacionar información de conjuntos de genes con funciones biológicas. Se basa en la comparación entre listas de genes de procesos conocidos con una lista de interés (en este caso las listas de genes obtenidas en la red). La cantidad de genes que correspondan tanto al proceso como a la lista constituye el enriquecimiento. El análisis consiste en determinar con un modelo estadístico (comúnmente una prueba hipergeométrica) la probabilidad de que ese enriquecimiento se presente al azar dados nuestros datos.

La formula de la distribución hipergeométrica se define de la siguiente forma (Los N y k usados aquí son distintos de los descritos anteriormente en el contexto de la topología de redes):

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Donde

N es el tamaño de la población
K es el numero de éxitos en la población
n es el número de intentos, en este caso el tamaño de la muestra
k es el número de éxitos en la muestra
$\binom{a}{b}$ es el número de combinaciones de a objetos tomados en grupos de tamaño b

Si la probabilidad de encontrar un determinado nivel de enriquecimiento es

pequeña (a partir de un punto de corte arbitrario como $p = 0.05$ o $p = 0.0001$) se puede considerar como razonable la hipótesis de que el proceso se encuentra afectado. Para un conjunto de genes en una red de regulación transcripcional el análisis de enriquecimiento estadístico nos permite averiguar cuáles son los procesos que están siendo regulados y de que manera se relacionan entre sí en el cáncer primario de mama.

Para el análisis de enriquecimiento estadístico se utilizó el plugin para Cytoscape BINGO [80]. El punto de corte se eligió de manera arbitraria con un valor de $p \leq 0.001$ para la red completa y de $p \leq 0.05$ para las comunidades. La razón de utilizar puntos de corte diferentes es que el número de genes considerados es mayor en la red completa, por lo que esperamos obtener valores de probalidad p más restrictivos. La lista de genes de la red completa tambien se utilizó en un análisis de Pathway Commons con la herramienta WebGestalt [69] para obtener las vías metabólicas estadísticamente enriquecidas.

Capítulo 5

Resultados

Las redes obtenidas con información mutua (MI) a partir de los niveles de expresión de RNA mensajero nos permiten identificar grupos de genes que genes están correlacionados a nivel transcripcional [5]. El análisis de la red de segunda iteración mostró características topológicas compartidas con las redes libres de escala [73], sin embargo no se encontró evidencia de que la distribución del grado de conectividad de los nodos siga una ley de potencia, según el procedimiento descrito en [81]. Se identificaron cuatro comunidades en el componente principal de la red de segunda iteración, que agrupan genes con patrones de expresión similares.

Los análisis de enriquecimiento estadístico mostraron sobrerrepresentación en los procesos: inflamatorio, inmunológico y morfogénetico, principalmente. Los procesos enriquecidos en las comunidades fueron diferentes entre éstas. Con excepción de la primera, cada comunidad presentó un enriquecimiento estadístico significativo en más de un proceso biológico. El nivel de enriquecimiento de las comunidades no corresponde a una proporción directa número de nodos que contienen, siendo comparativamente grande (con un *p-value* más significativo) en la cuarta comunidad, que cuenta con menos genes.

De los procesos biológicos con mayor enriquecimiento estadístico se tomaron el *proceso del sistema inmune*, la *señalización* y la *respuesta inflamatoria* para inves-

tigar la distribución de los genes involucrados en estos procesos en la estructura de la red. Encontramos que en ninguno de los tres casos los genes forman solo un grupo conectado, lo que sugiere que los procesos biológicos tal como los define Gene Ontology no forman módulos únicos a nivel transcripcional, al menos en el cáncer primario de mama. Este resultado se complementa con el hallazgo de que las comunidades en la red no corresponden a procesos únicos, lo que será discutido más adelante.

5.1. Análisis topológico de las redes de regulación completas

La red de primera iteración (Figura 5.1) consta de 689 nodos y 1,632 aristas. El componente más grande consta de 619 nodos. La red cuenta con 13 componentes pequeños de al menos dos nodos y una interacción. En esta red se encuentran 78 genes de los 418 genes inflamatorios de la lista de entrada. La primera red muestra las interacciones entre los genes explícitamente inflamatorios y sus primeros vecinos a nivel transcripcional en el resto del genoma. La red de segunda iteración (Figura 5.2) contiene 1,280 nodos y 21,332 aristas. Tiene un componente principal con 1252 nodos y seis componentes pequeños. Es interesante notar que al calcular la segunda red, el número de nodos cambió en un factor de 2 mientras que el número de interacciones aumentó en un factor de 20, lo que se vió reflejado en el análisis topológico de las redes (Cuadro 5.1) con un aumento en el coeficiente de agrupamiento, en el número promedio de vecinos y la disminución en el tamaño promedio y el porcentaje de caminos de longitud mínima.

La estructura de la red (Figura 5.2) muestra una asociación estrecha entre genes con expresión diferencial similar, destacando un gran cúmulo de genes subexpresados dentro del componente conectado principal. Solamente 83 de los 1280 nodos presentan un nivel de expresión igual o mayor al doble, respecto al tejido normal, mientras que 490 genes se encuentran subexpresados con al menos la mitad del nivel de expresión respecto al tejido normal. Aquellos genes con un cambio en el

Cuadro 5.1: Resultados del análisis topológico de las redes de inflamación

Parámetro	Red primera iteración	Red segunda iteración
Coeficiente de agrupamiento	0.312	0.631
Número de componentes conectados	14	7
Diámetro de la red	13	12
Radio de la red	1	1
Centralización de la red	0.359	0.293
Caminos de longitud mínima	383054 (80 %)	1566410 (95 %)
Longitud característica de caminos	4.212	3.72
Número promedio de vecinos	4.737	33.331
Número de nodos	689	1280
Densidad de la red	0.007	0.026
Heterogeneidad de la red	3.132	1.682
Número de nodos aislados	0	0
Bucles	0	0
Pares de nodos con aristas múltiples	0	0

nivel de expresión no significativo ($-1 \leq logFCh \leq 1$) se vieron agrupados con genes de la misma tendencia. Ésto es, genes ligeramente subexpresados son vecinos de genes francamente subexpresados y genes ligeramente sobreexpresados se asocian a genes francamente sobreexpresados (Figura 5.2).

Pese a esta tendencia a agrupar genes de expresión diferencial similar, también se encontraron grupos de genes que combinan patrones de expresión opuestos (se puede observar en los dos agrupamientos de la parte superior derecha en la figura 5.2). Estos grupos pueden darnos pistas respecto de los diferentes niveles de regulación transcripcional a nivel de células específicas, vías específicas y genes específicos [31].

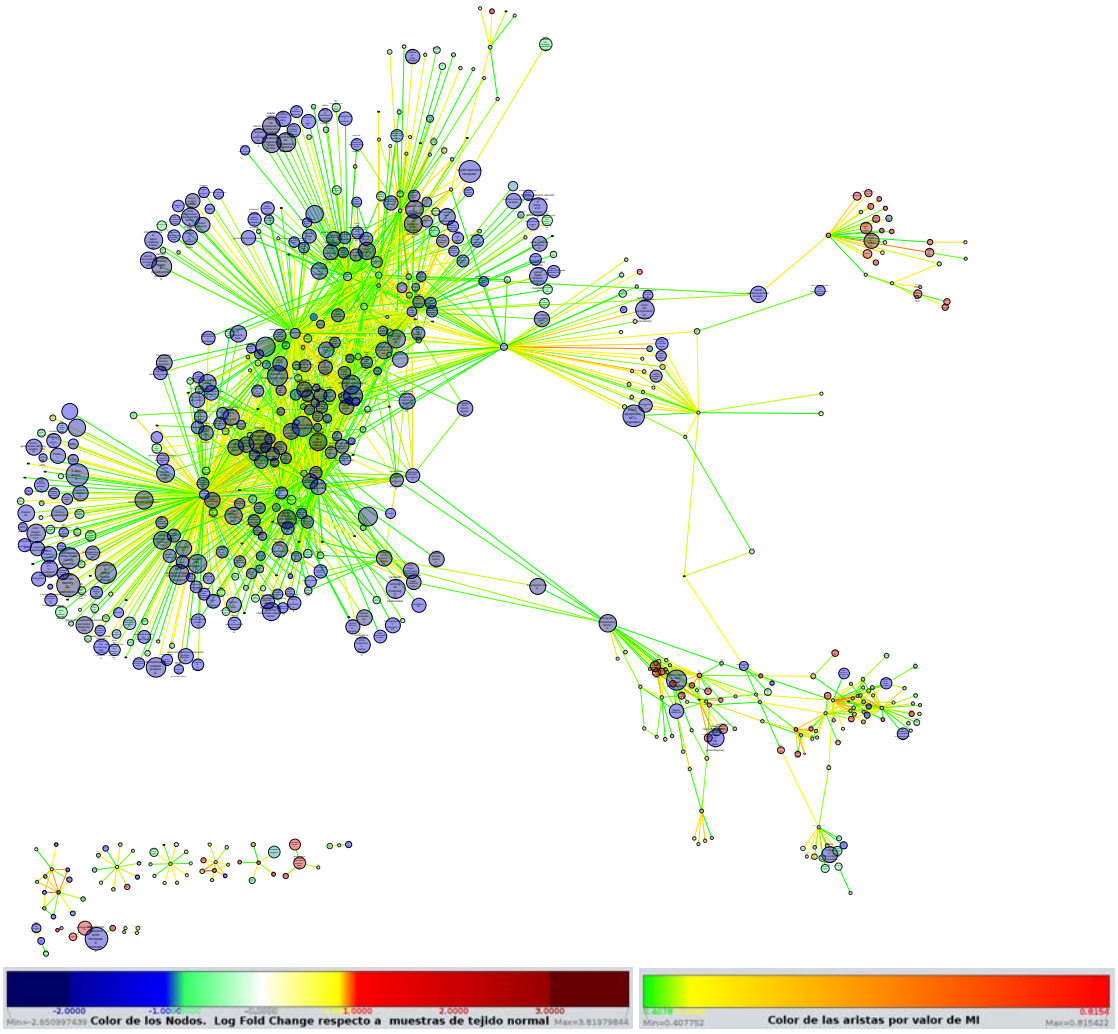


Figura 5.1: Primera red, obtenida con la lista de moléculas asociadas a inflamacion vs genoma completo. Esta red consta de 689 nodos y 1,632 aristas. Se puede observar que el componente principal de la red se encuentra unido por un esqueleto de unas pocas interacciones y la mayoría de los nodos se encuentran unidos a la red por una sola interacción. Su estructura consiste primordialmente de estrellas en donde un solo gen se asocia a numerosos genes no conectados entre sí (se aprecian en los componentes de la parte inferior izquierda). Se puede observar la prevalencia de genes subexpresados. El tamaño de los nodos es proporcional al estadístico B. Mayor tamaño indica consistencia en los niveles de expresión.

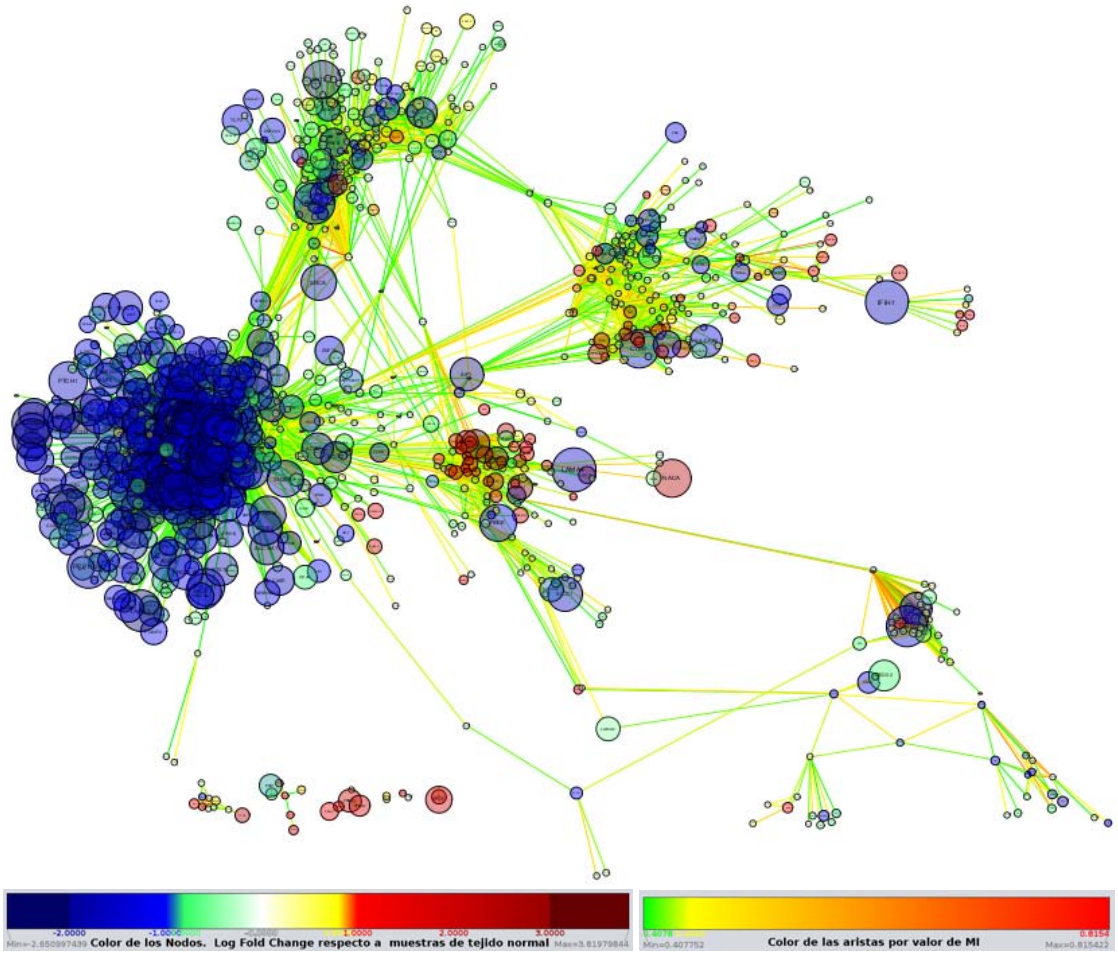


Figura 5.2: Red de segunda iteración. Se observa el aumento de complejidad respecto a la primera red, con aproximadamente el doble de nodos (1280) y veinte veces el número de aristas (21332). Esta red consiste de grupos de nodos altamente interconectados y se puede observar una tendencia al agrupamiento, donde patrones de expresión similares forman cúmulos de genes sobreexpresados y un gran grupo de genes subexpresados. En algunos de los grupos (parte superior del componente mayor) se logra apreciar la asociación entre genes con expresión diferencial opuesta, esto es, genes subexpresados asociados directamente con genes sobreexpresados. El tamaño de los nodos es proporcional al estadístico B. Mayor tamaño indica consistencia en los niveles de expresión.

5.1.1. Comunidades encontradas en la red de segunda iteración.

Las comunidades en una red corresponden a grupos de nodos altamente conectados. Las comunidades encontradas con MCODE no corresponden a cliques, sin embargo, tienen un alto grado de interconexión.

Las comunidades de la red de inflamación nos permiten encontrar grupos de genes con una correlacion estrecha entre ellos. En términos topológicos se trata de subredes con alta densidad y un coeficiente de agrupamiento alto (Cuadro 5.2) que sugieren módulos funcionales [77]. Las comunidades encontradas son mutuamente excluyentes y no comparten ningún gen (corresponden a particiones diferentes del conjunto de nodos). En las cuatro comunidades encontradas se pudo observar que la tendencia es la de agrupar genes con patrones de expresión similares. Los genes francamente subexpresados tienden a asociarse con otros genes subexpresados (Figura 5.3(b)). Los genes sobreexpresados se asocian preferentemente con genes sobreexpresados (Figura 5.3(d)).

Cuadro 5.2: Resultados del análisis topológico de las comunidades en la red de inflamación

Parámetro	**Primera comunidad**	**Segunda comunidad**	**Tercera comunidad**	**Cuarta comunidad**
Coeficiente de agrupamiento	0.858	0.684	0.940	0.884
Número de componentes conectados	1	1	1	1
Diámetro de la red	2	3	2	2
Radio de la red	1	2	1	1
Centralización de la red	0.222	0.583	0.083	0.170
Caminos de longitud mínima	10920 (100 %)	8556 (100 %)	650 (100 %)	380 (100 %)
Longitud característica de caminos	1.218	1.702	1.077	1.153
Número promedio de vecinos	81.352	28.516	23.077	16.1
Número de nodos	105	93	26	20
Densidad de la red	0.782	0.310	0.923	0.847
Heterogeneidad de la red	0.200	0.448	0.107	0.157

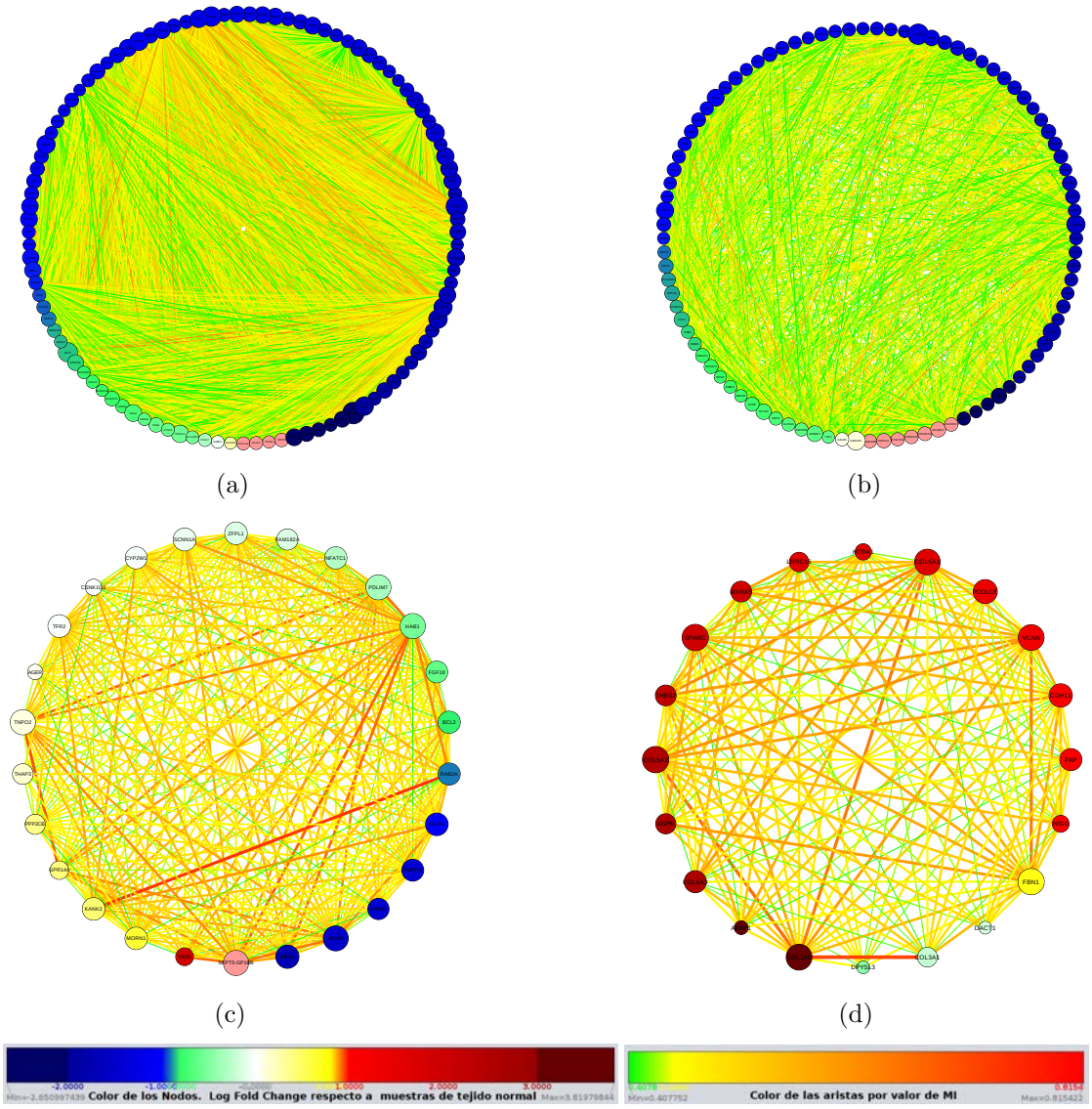


Figura 5.3: Comunidades en la red. Estas subredes consisten de grupos de genes con un alto grado de interconexión. Puede observarse que el patrón de expresión diferencial es similar al interior de cada comunidad. La primera comunidad (a) consta de genes subexpresados. La segunda comunidad (b) consta de genes subexpresados. La tercera comunidad (c) consta de genes de expresión mixta. La cuarta comunidad (d) consta de genes sobreexpresados. Tamaño de los nodos proporcional al número de conexiones (grado) en la red (las comunidades se muestran a escalas diferentes)

5.2. Análisis de enriquecimiento estadístico para el conjunto de genes en la red

Una proteína, codificada por un gen, puede participar en numerosas funciones biológicas. Al tomar un conjunto de genes que se coexpresan, a reserva de los mecanismos posttranscripcionales y traduccionales a los que se hallen sujetos, establecemos un contexto para el ambiente en el que éstos son expresados. El análisis de enriquecimiento estadístico, de los conjuntos de genes asociados en la red, tiene como propósito el formular hipótesis plausibles de las funciones biológicas que éstos llevan a cabo en el tejido analizado.

5.2.1. Procesos biológicos enriquecidos en la red completa de segunda iteración

Procesos de *Gene Ontology*

Los veinte procesos que resultaron con mayor nivel de enriquecimiento estadístico al tomatr la red completa se datallan en el cuadro 5.3. El mayor nivel de enriquecimiento se observó en procesos como: *proceso del sistema inmunológico*, *señalización*, *respuesta a heridas*, *respuesta inflamatoria*, *morfogénesis* y *regulación de procesos del sistema inmune*. Es de esperarse que se encontraran procesos inflamatorios estadísticamente enriquecidos dado que la lista inicial fué de genes del *proceso inflamatorio* y fue curada tomando al mismo *Gene Ontology* como base, sin embargo la respuesta inmune y la cicatrización se encuentraron representados con un mayor número de genes y valor de enriquecimiento.

Cuadro 5.3: Análisis de enriquecimiento estadístico de procesos biológicos en *Gene Ontology*. Se muestran los 20 procesos de enriquecimiento más significativo. BINGO con un $p - val \leq 0.05$

Descripción del proceso	p-value corregido	Genes en la red	Genes en el proceso
proceso del sistema inmune	**6.4046E-29**	**181**	**943**
respuesta de defensa	2.2273E-27	137	623
respuesta inmune	5.8980E-21	124	616
señalizacion	**3.1380E-16**	**364**	**3127**
respuesta a heridas	8.7625E-16	104	542
respuesta al estrés	1.6265E-15	235	1772
regulacion de procesos del sistema inmune	6.9432E-15	87	424
respuesta a estímulos	1.4555E-14	400	3624
respuesta inflamatoria	**4.1308E-14**	**71**	**316**
proceso multiorganismo	7.4817E-13	125	786
vías de señalización de receptores ligados a la superficie celular	1.2445E-11	173	1280
regulación de la respuesta a estímulos	1.2445E-11	92	524
regulación de los procesos de desarrollo	3.0331E-11	121	791
regulación de procesos organísmicos multicelulares	5.8645E-11	149	1066
regulación positiva de procesos del sistema inmune	2.0421E-10	57	265
desarrollo del sistema	2.1202E-10	275	2414
desarrollo de estructuras anatómicas	3.2816E-10	295	2648
regulación de la activación celular	4.5030E-10	46	191
regulación positiva de procesos biológicos	1.3970E-9	252	2201
desarrollo de organos	1.4148E-9	214	1789

Análisis de enriquecimiento de vías a nivel de la red completa de segunda iteración en *Pathway Commons*

El análisis de vías se hizo tomando por separado genes sobreexpresados y subexpresados (Cuadro 5.4). Entre las vías enriquecidas con genes sobreexpresados resaltan: el *sistema inmune*, *presentación de antígenos*, *señalización por integrinas* y *señalización de citocinas*. Los genes subexpresados mostraron un enriquecimiento significativo en la *señalización mediada por receptores acoplados a proteínas G*.

Cuadro 5.4: Análisis de enriquecimiento estadístico de *Pathway Commons*. Resultados obtenidos de la página de *Webgestalt* con un $p \leq 0.001$ para el conjunto de genes diferencialmente expresados en la segunda red. Se muestran los 10 resultados más significativos.

Descripción de la vía	**Genes en el conjunto de referencia**	**Genes en la lista**	**Genes esperados**	**proporción de enriquecimiento**	p **ajustado**
Conjunto de Genes sobreexpresados					
Sistema Inmune	476	33	7.83	4.21	6.10e-10
Vías del retículo endoplásmico/fagosoma	70	12	1.15	10.42	1.99e-07
Siste inmune adaptativo	211	19	3.47	5.47	1.99e-07
procesamiento y presentación cruzada de antígenos	76	12	1.25	9.60	2.88e-07
procesamiento y presentacion de antígenos mediados por MHC clase-I	91	12	1.50	8.01	1.57e-06
Interacciones de superficie celular mediadas por integrina Beta3	43	9	0.71	12.72	1.57e-06
Señalizacion por interferones	95	12	1.56	7.68	2.20e-06
Interacciones de superficie celular de la familia de integrinas	1267	47	20.85	2.25	2.69e-06
Interacciones de superficie celular de integrina Beta1	1241	46	20.42	2.25	3.51e-06
Señalización por citocinas en el sistema inmune	184	14	3.03	4.62	6.07e-05
Conjunto de Genes subexpresados					
Señalización por GPCRs	426	40	16.61	2.41	7.39e-05
Unión de ligandos de GPCRs	286	29	11.15	2.60	0.0004

Una comprobación adicional se realizó repitiendo el análisis de enriquecimiento, eliminando los genes explícitamente anotados como inflamatorios (los pertenecientes a la lista de genes con la que se calculó la primera red) (Cuadro 5.5). Tal como se esperaba, en esta prueba de enriquecimiento dejaron de aparecer procesos inflamatorios. Los procesos que se mantuvieron son de naturaleza inmunológica, e incluyen algunos relacionados con la activación de células T. Se encontraron también procesos asociados al desarrollo del organismo y formación de estructuras como la *osificación* y *formación de órganos hematopoiéticos y linfoides*. Resulta interesante la representación del proceso de *preñez* (female pregnancy) como uno de los que tienen mejor puntuación y que será discutido más adelante.

Cuadro 5.5: Análisis de enriquecimiento estadístico de procesos biológicos sin los genes inflamatorios en *Gene Ontology*. Se muestran los 20 procesos de enriquecimiento más significativo. BINGO con un $p - val \leq 0.05$

Descripción del proceso	p-value corregido	Genes en la red	Genes en el proceso
proceso del sistema inmune	5.1078E-13	135	943
respuesta inmune	3.8306E-8	89	616
desarrollo de sistema	1.2606E-7	244	2414
desarrollo de estructura anatómica	2.2856E-7	261	2648
señalización	6.0301E-7	296	3126
desarrollo de órgano	1.3818E-6	187	1788
preñez	1.7366E-6	29	122
desarrollo organísmico multicelular	2.5520E-6	279	2961
proceso multi-organismo	2.8005E-6	98	785
proceso organísmico multicelular	4.5564E-6	383	4362
procesamiento t presentación de antígenos	9.8172E-6	18	57
regulación de procesos del sistema inmune	1.1390E-5	61	423
proceso de desarrollo	2.2257E-5	293	3224
regulación de procesos de desarrollo	7.1385E-5	93	790
osificación	7.1385E-5	25	116
regulación de procesos organísmicos multicelulares	7.1385E-5	117	1065
organización de la matriz extracelular	1.0088E-4	23	103
vías de señalización de receptores ligados a la superficie celular	1.2189E-4	134	1279
proceso catabólico de drogas	1.4930E-4	7	10
regulación positiva de procesos biológicos	1.4930E-4	208	2200

Enriquecimientos en la categoría de *función molecular*

Los análisis de enriquecimiento estadístico para la categoría *molecular function* (función molecular)(Figura 5.4) mostraron una participación importante de la actividad de citocinas y sus receptores, en la actividad de receptores acoplados a proteínas G y en la interacción con componentes extracelulares, tanto de la matriz extracelular como de plasma sanguíneo.

En términos generales, se observó un enriquecimiento en funciones involucradas con la comunicación celular parácrina (y autócrina) y de la interacción con el espacio extracelular.

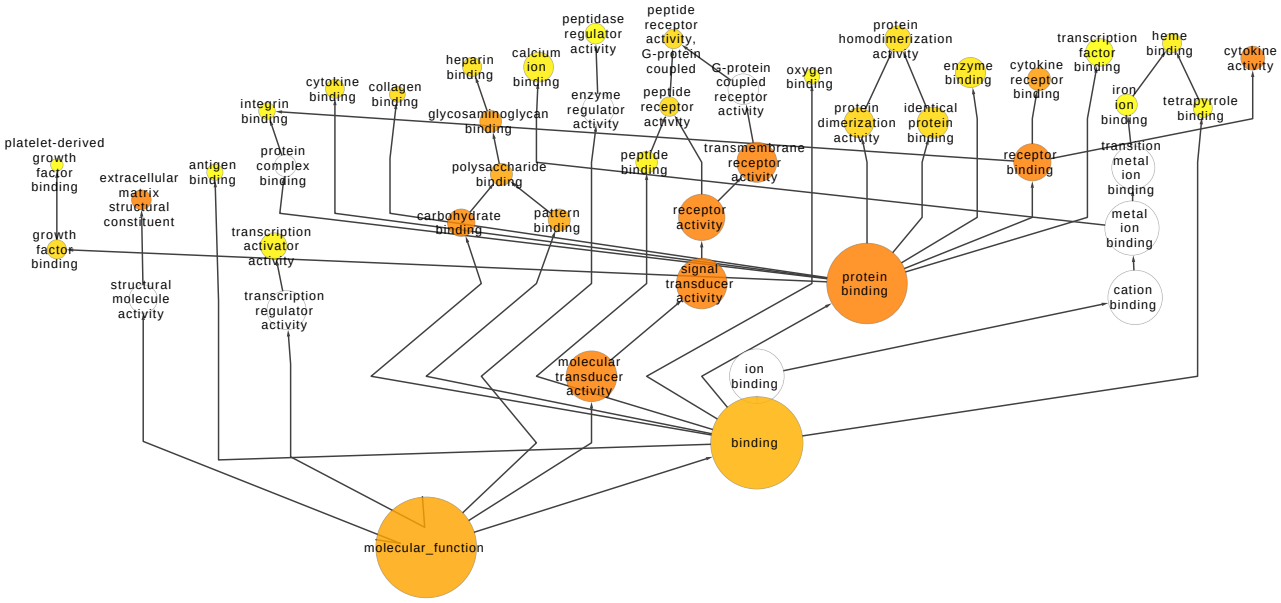


Figura 5.4: Resultados para el análisis de enriquecimiento estadístico para todos los genes de la segunda red en la categoría *molecular function*. Las funciones enriquecidas muestran una participación importante en la regulación de la comunicación celular por medio de citocinas y sus receptores. También se muestra la participación de componentes estructurales de la matriz extracelular. Color:blanco, menos significativo; rojo, más significativo. Tamaño de los nodos proporcional al número de genes que contiene.

Enriquecimientos en la categoría *componente celular*

La categoría *cellular component* (componente celular)(Figura 5.5) mostró la participación de componentes integrales de la membrana plasmática, componentes de la cara externa de la membrana, componentes del espacio extracelular, incluyendo la matriz extracelular de colágena, además de componentes específicos del complejo mayor de histocompatibilidad (MHC).

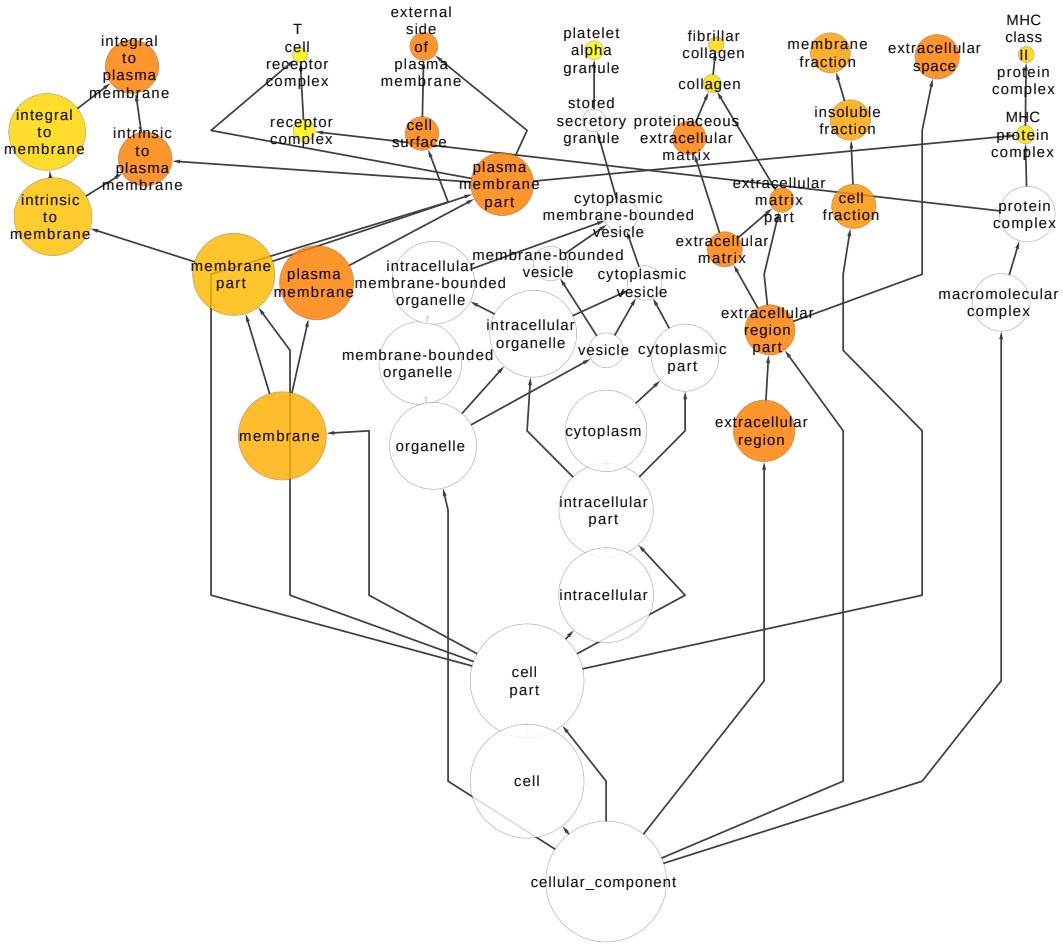


Figura 5.5: Resultados para el análisis de enriquecimiento estadístico para todos los genes de la segunda red en la categoría *cellular component*. Se observa la participación predominante de componentes membranales y del espacio extracelular, lo que sugiere que existe una modulación sobre los componentes de superficie celular que puede estar involucrada directamente en la capacidad de sensar estímulos. Color:blanco, menos significativo; rojo, más significativo. Tamaño de los nodos proporcional al número de genes que contiene.

5.2.2. Análisis de enriquecimiento estadístico para las comunidades de la red de segunda iteración

El análisis de enriquecimiento por comunidades (Cuadro 5.6) se llevó a cabo para investigar la siguiente relación entre los grupos de genes altamente conectados entre sí y procesos biológicos determinados por éstos.

La primera comunidad no mostró un enriquecimiento significativo para ningún proceso biológico, pese a ser la de mayor tamaño. Esta comunidad presenta genes de componentes del Citocromo P450, glicoproteínas específicas del embarazo (PSGs), transportadores de solutos iónicos y numerosas proteńas con patrones de dedos de zinc que son asociados con la unión al DNA.

La segunda comunidad se encuentró enriquecida en *respuesta a virus*. Los genes asociados incluyen receptores de siete dominios transmembranales, interferones de tipo I, canales iónicos, receptores inmunológicos de tipo KIR (*killer cell immunoglobulin-like receptor*) que están asociados en inmunomodulación y factores de transcripción como CIITA que controla la activación transcripcional de numerosos miembros del complejo mayor de histocompatibilidad de clase II.

La tercera comunidad presentó un enriquecimiento en procesos de señalización celular, diferenciación celular y morfogénesis. El número de genes en cada categoría enriquecida es pequeño, menor a 5 genes. Entre los genes asociados tenemos al receptor de DAMPs AGER y a BCL2.

La cuarta comunidad se caracterizó por tener una predominancia de genes sobreexpresados. Presentó un enriquecimiento en procesos de morfogénesis que incluyen: *desarrollo del sistema esquelético*, *desarrollo de la piel*, así como otros de caracter genérico como *desarrollo organísmico multicelular*. Los componentes celulares están localizados en la region extracelular. Los genes asociados a esta comunidad son sobre todo de proteínas de matriz extracelular: versicano, colágenas tipo I, III, V y VI, trombospondina 2 y cadherina 11 que se asocian con adhesión

celular.

Cuadro 5.6: Análisis de enriquecimiento estadístico de procesos biológicos por comunidad. Se muestran los 5 procesos de enriquecimiento más significativo con un *p-value* ≤ 0.05.

Descripción del proceso	**p-value corregido**	**Genes comunidad**	**Genes proceso**	**Nombres de los genes en la comunidad**
Segunda comunidad (Figura 5.3(b))				
respuesta a virus	6.0171E-3	7	142	IFNA21 IFNA7 CHRM2 IFNA5 IFNA10 IFNA4 IFNA14
Tercera comunidad (Figura 5.3(c))				
regulación de la actividad de heterodimerización de proteínas	7.3748E-3	2	4	BCL2 NRG1
regulación positiva de la diferenciación celular	7.3748E-3	5	255	FGF18 PDLIM7 BCL2 GNAS NRG1
regulación positiva de la diferenciación celular del músculo estriado	7.3748E-3	2	6	BCL2 NRG1
regulación de la actividad de homodimerización de proteínas	1.0307E-2	2	8	BCL2 NRG1
regulación positiva de procesos del desarrollo	1.4931E-2	5	347	FGF18 PDLIM7 BCL2 GNAS NRG1
osificación endocondral	1.6143E-2	2	12	FGF18 GNAS
Cuarta comunidad (Figura 5.3(d))				
desarrollo del sistema esquelético	8.5803E-7	8	332	ASPN AEBP1 FBN1 COL3A1 COL1A2 SPARC COL5A2 CDH11
desarrollo organísmico multicelular	3.1259E-6	15	2972	ASPN AEBP1 COL3A1 FBN1 DPYSL3 SPARC COL5A2 PCOLCE COL5A1 DACT1 FAP COL6A3 COL1A2 VCAN CDH11
organización de fibrillas de colágeno	3.1259E-6	4	27	COL3A1 COL1A2 COL5A2 COL5A1
desarrollo de la piel	5.0182E-6	4	34	COL3A1 COL1A2 COL5A2 COL5A1
proceso del desarrollo	5.0182E-6	15	3235	ASPN AEBP1 COL3A1 FBN1 DPYSL3 SPARC COL5A2 PCOLCE COL5A1 DACT1 FAP COL6A3 COL1A2 VCAN CDH11
desarrollo de sistema	1.7332E-5	13	2422	ASPN AEBP1 FAP COL6A3 FBN1 COL3A1 COL1A2 VCAN DPYSL3 SPARC COL5A2 COL5A1 CDH11

5.2.3. Localización de procesos enriquecidos en la red completa de segunda iteración

En la sección anterior se muestran los resultados de buscar procesos biológicos conocidos en grupos de genes altamente conectados entre sí. En esta sección presentamos el resultado de una pregunta complementaria: ¿Como se encuentran distribuidos los procesos biológicos conocidos en la red?

Se exploraron aquellos procesos que tuvieran 1) un buen nivel de enriquecimiento (entre los primeros 20), 2) un número grande de genes involucrados y 3) que se relacionaran de manera directa con la inflamación. Finalmente fueron analizados: *Proceso del sistema inmune*, *señalización* y *respuesta inflamatoria* (Figuras

5.6, 5.7 y 5.8 respectivamente y señalados en el cuadro 4.1 con negritas).

Las subredes resultantes formaron redes de componentes separados y numerosos nodos sin conectar. Las tres subredes mostraron grupos de nodos sobreexpresados asociados a nodos subexpresados y comparten grupos de genes entre sí. La categoría de *señalización*, que es la más general, comparte 84 genes con *proceso del sistema inmune* y 34 genes con *respuesta inflamatoria*. *Respuesta inflamatoria* y *proceso del sistema inmune* comparten 41 genes. Los genes compartidos se distribuyeron a lo largo de todos los componentes de estas subredes.

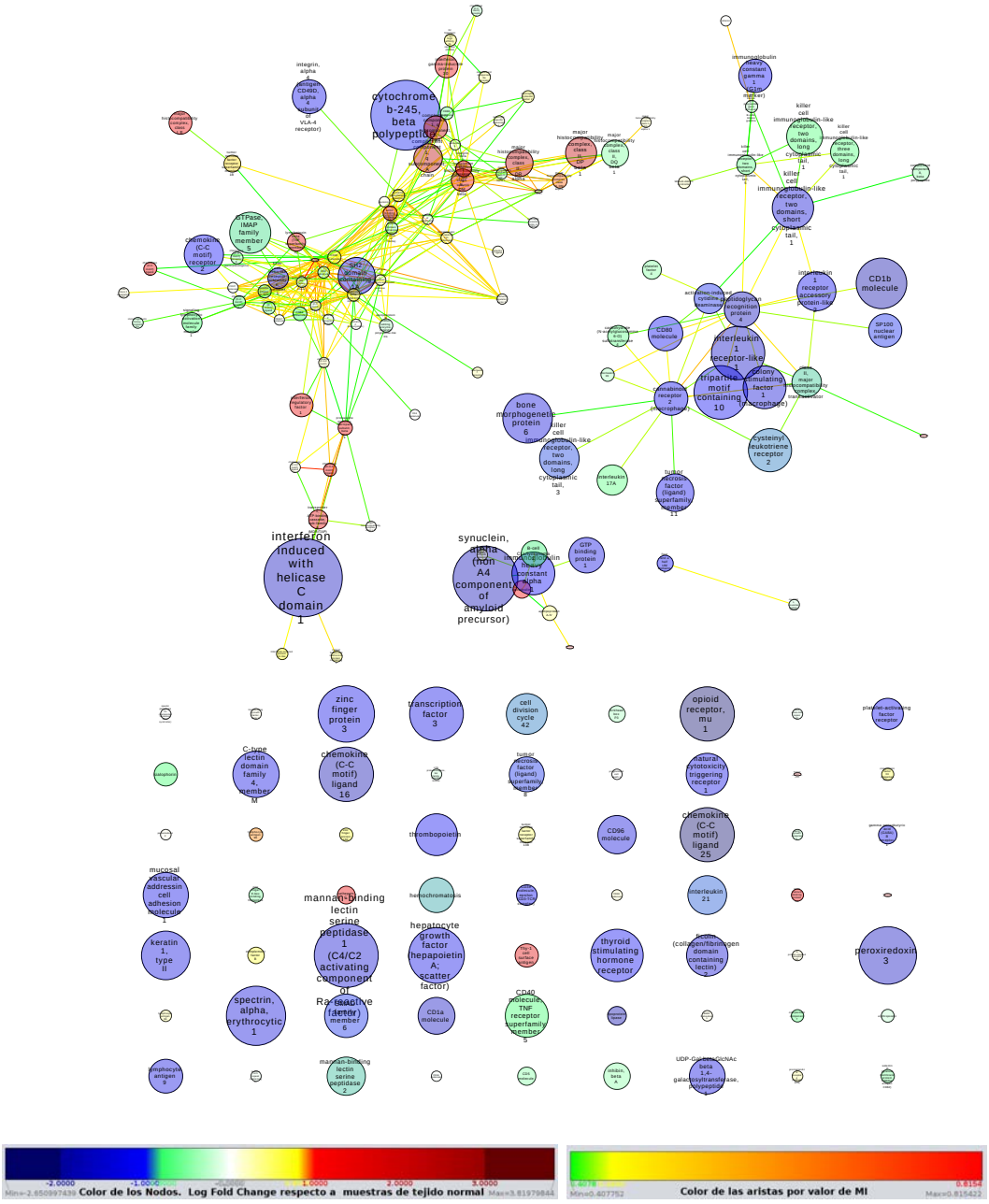


Figura 5.6: Subred correspondiente a *Proceso del sistema inmune*. Los genes se encuentran formando tres componentes y numerosos genes sin interacciones entre sí. El componente más grande consta de 70 nodos, 327 de las 388 interacciones totales y los niveles de expresión incluyen genes sub y sobreexpresados. El tamaño de los nodos es proporcional al estadístico B. Mayor tamaño indica consistencia en los niveles de expresión.

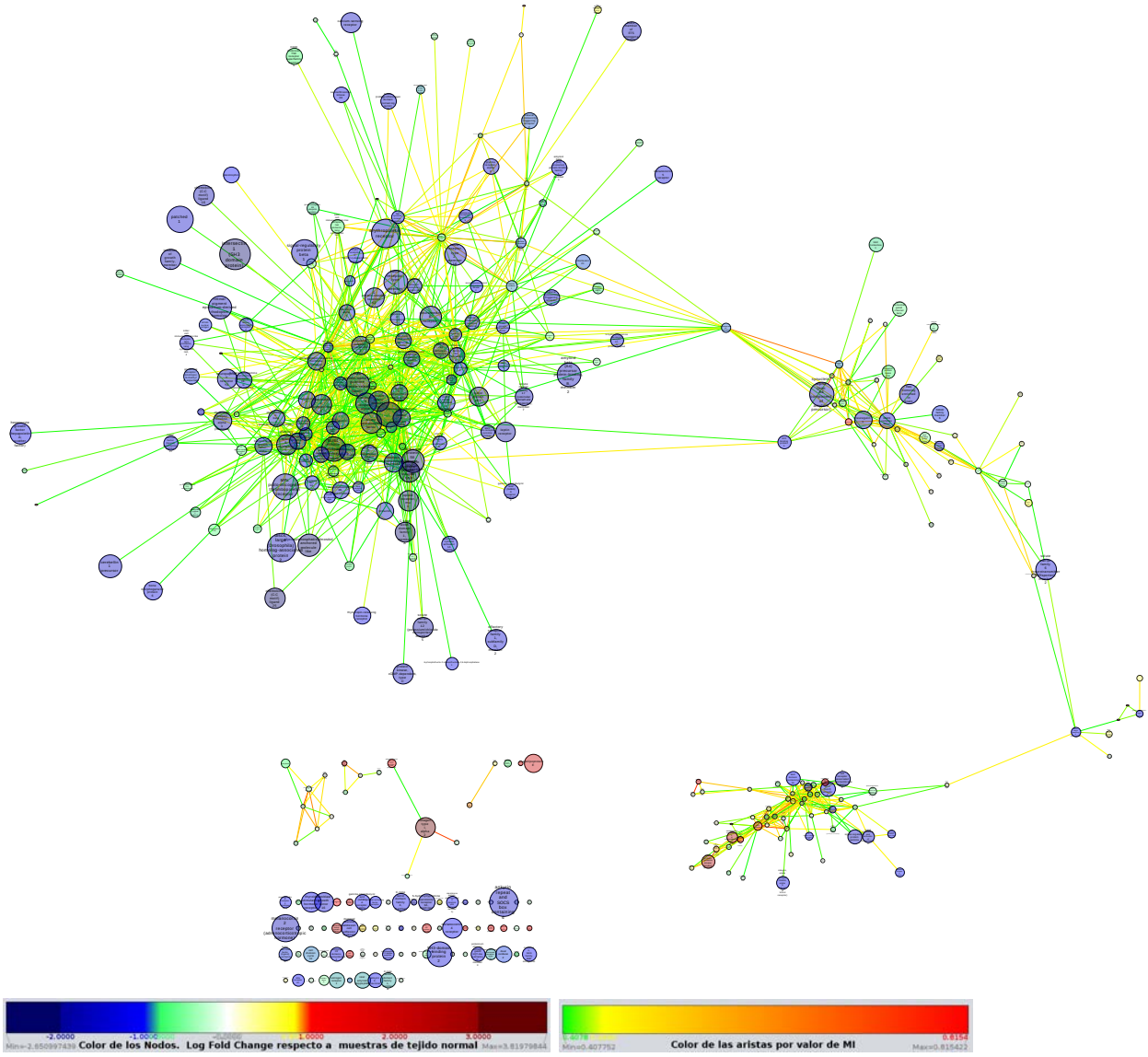


Figura 5.7: Subred correspondiente al proceso *señalización*. Se distingue un componente principal y cuatro componentes pequeños, así como numerosos genes aislados. En el componente principal se distinguen claramente tres segmentos, uno de los cuáles (parte inferior derecha) presenta genes asociados con valores de expresión diferencial opuestos. Se aprecia la predominancia de genes subexpresados lo que sugiere una fuerte regulación negativa de la señalización celular, lo que se corrobora con el análisis de vías de Pathway Commons (Cuadro 4.2). El tamaño de los nodos es proporcional al estadístico B. Mayor tamaño indica consistencia en los niveles de expresión.

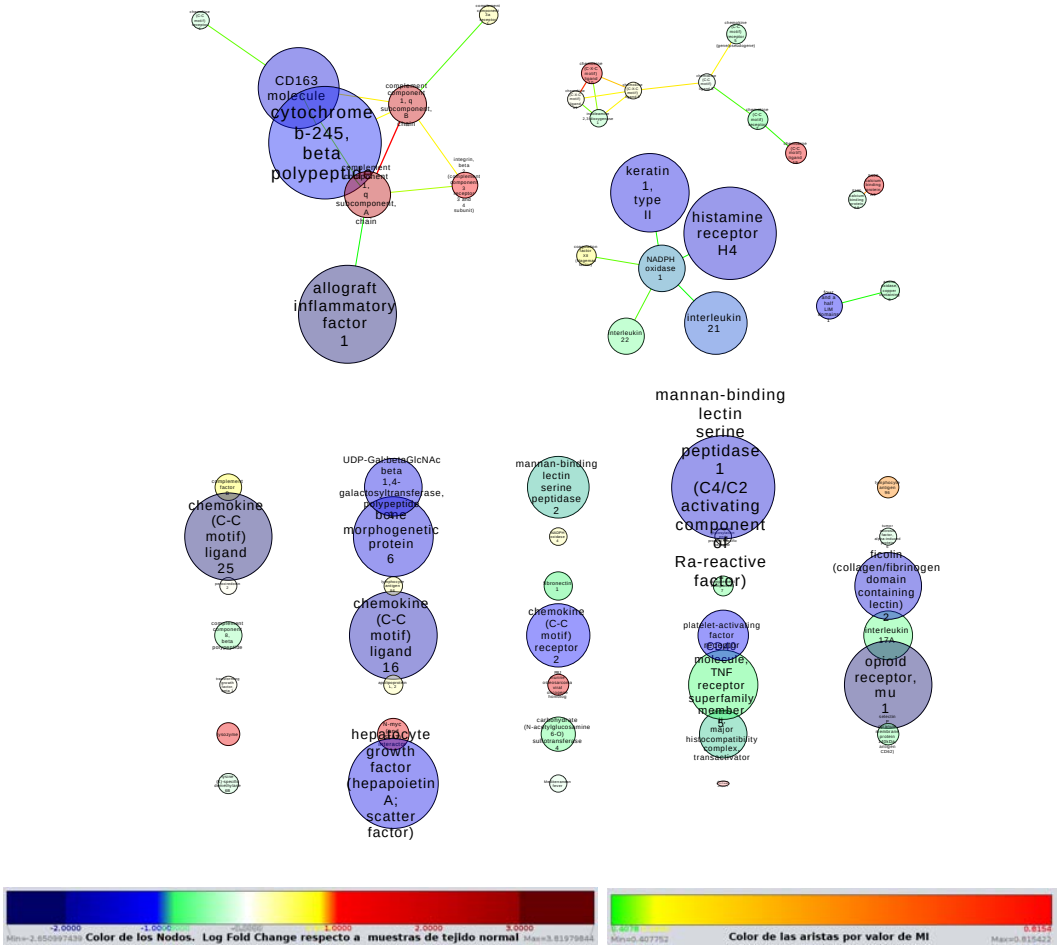


Figura 5.8: Subred correspondiente al proceso *respuesta inflamatoria*. En esta subred se observan componentes de pequeño tamaño y numerosos genes sin interacciones directas. Aunque la mayoría de los genes están subexpresados es interesante notar que los componentes iniciadores de la cascada de complemento están sobreexpresados y con uno de los niveles de correlacion más altos de toda la red. Esta subred comparte genes con la de *proceso del sistema inmune* lo que es razonable si tomamos en cuenta la relación existente entre la inflamación como un intermediario entre la detección de la alteración en la homeostasis y la eliminación de posibles amenazas. El tamaño de los nodos es proporcional al estadístico B. Mayor tamaño indica consistencia en los niveles de expresión.

Capítulo 6

Discusión

La inferencia de redes a partir de datos de microarreglos de expresión de genoma completo, permite encontrar las relación de cada gen con el resto de los genes en la célula. La estructura de la red refleja de este modo los patrones globales de expresión genética, lo que no es posible realizar con un enfoque de genes individuales. Así mismo las redes combinadas con análisis de expresión diferencial puede revelar patrones de regulación complejos con genes que presentan una expresión diferencial dispar, lo que pasaría desapercibido en un análisis únicamente de expresión diferencial.

Se ha mostrado que las redes que modelan sistemas reales (por ejemplo, el internet o las redes sociales) tienen características topológicas particulares, como una distribución del grado de conectividad que se ajusta a una ley de potencia, coeficiente de agrupamiento alto, longitud típica de caminos con valor pequeño y la presencia de comunidades de nodos con alta densidad de conecciones [62]. Se ha sugerido que las comunidades tienen una importancia funcional [77] y se han reportado en redes biológicas que modelan la interacción de proteínas [74]. La búsqueda de comunidades en redes de regulación transcripcional puede revelar módulos de genes coexpresados [31]. Aunque parece natural que estos módulos transcripcionales correspondan a vías celulares, los análisis de enriquecimiento estadístico mostraron que las comunidades de la red de inflamación están relacionadas con

más de un proceso. Por otro lado, la ubicación de los procesos enriquecidos a nivel de la red completa revela que éstos no se encuentran formando un solo módulo conectado, mas bien distribuyendose en varias partes de la red. La relevancia de estos resultados será discutida mas adelante.

6.1. Topología de las redes

El análisis topológico de las redes (Cuadro 5.1) muestra que con el cálculo de la red de segunda iteración aumentó el número de nodos, de 689 a 1280 (1.87 veces), además de incrementar considerablemente el número de aristas de 1632 a 21332 (13 veces). Los parámetros calculados para ambas redes indican un aumento en el coeficiente de agrupamiento y de la densidad de la red, así como en el número de caminos de longitud mínima (aproximadamente cuatro veces mayor), con una longitud típica de caminos de 3.72. El número promedio de vecinos aumentó en un factor de 7. Este cambio en los parámetros indica que la ganancia de aristas entre genes no es uniforme. Por otro lado, la distribución del grado de conectividad en la red, que a primera vista se muestra como una distribución de *cola pesada* (*heavy-tailed*) no se ajustó satisfactoriamente a una ley de potencia siguiendo el procedimiento descrito en [81]. Estos resultados sugieren características comunes con una *red compleja* de la manera discutida en [73], aunque el no ajustarse a una ley de potencia, mostrando una distribución de cola pesada, parece ser común para redes metabólicas y de regulación transcripcional (ibídem).

6.2. Expresión diferencial en el contexto de la red

El patrón de expresión diferencial de los genes de la red muestra el *apagado* de una gran cantidad de genes en una proporción de alrededor de 6 genes subexpresados por cada uno sobreexpresado. La visualización de los valores de expresión diferencial (Figura 5.2) revela el agrupamiento de genes con patrones de expresión diferencial similar. Esta tendencia se mantiene en las comunidades para las

cuales se hace evidente que incluso los genes que no tienen un valor de expresión diferencial $-1 \leq logFCh \geq 1$ siguen tendencias similares, como en el caso de la comunidad 2 (Figura 5.3(b)), donde los nodos marcados en verde se pueden considerar como *ligeramente subexpresados* en un contexto de genes francamente subexpresados (color azul).

La expresión de genes está regulada en varios niveles, con mecanismos de control a nivel de la transcripción, procesamiento de RNA mensajeros, traducción y modificaciones posttraduccionales [56]. La regulación a nivel transcripcional puede llevarse a cabo con diferentes grados de especificidad, desde la regulación de genes individuales, de conjuntos de genes que forman módulos funcionales o incluso de regiones enteras del genoma que pueden activarse o ser silenciadas [31]. En algunas partes de la red se encuentran grupos de nodos con patrones de expresión dispar, esto es, genes subexpresados asociados a genes sobreexpresados. Aunque los genes de expresión diferencial similar tienden a agruparse, no se encuentran aislados del resto de la red. Estos grupos de genes participan en procesos biológicos que serán detallados mas adelante (Figura 5.6). El significado biológico de estos patrones de expresión podría relacionarse con la regulación a nivel de genes individuales (ibídem).

Los resultados obtenidos en las redes aquí mostradas bien pueden estar reflejando todos estos mecanismos de regulación, sin embargo, debido a que sólo podemos observar el efecto de la regulación, y no sus causas, es necesario complementar nuestro enfoque con experimentos que sean capaces de medir variables adicionales, como los patrones de metilación de genes y compactación de la cromatina (que revelen la activación o silenciamiento de regiones del genoma), así como integrar la información de bases de datos sobre los blancos confirmados de factores de transcripción.

6.3. Análisis de enriquecimiento estadístico de procesos biológicos

El análisis en las categorías de Gene Ontology (Cuadro 5.3) arroja procesos relacionados con el sistema inmunológico, señalización y respuesta inflamatoria, con una ubicación celular (Figura 5.5) principalmente en la membrana plasmática y el espacio extracelular, así como funciones moleculares (Figura 5.4) centradas en las actividades de comunicación celular, incluidas la señalización por receptores acoplados a proteínas G, la unión de antígenos y actividad estructural relacionada con la matriz extracelular. Esto resulta interesante si se toma en cuenta que, durante la respuesta inflamatoria, la producción de mediadores como quimiocinas y citocinas, hace uso de los mecanismos de señalización celular para coordinar el comportamiento de las células presentes en el tejido afectado [21].

En el análisis de Pathway Commons (Figura 5.4), donde fueron probados los genes con expresión diferencial en dos grupos: sobreexpresados y subexpresados con el fin de conocer el prendido y apagado de vías específicas, se obtuvieron resultados similares a los enriquecimientos basados en *Gene Ontology*. Entre las vías enriquecidas para los genes sobreexpresados destacan aquellas relacionadas a la regulación del sistema inmune, con la presentación de antígenos, interacciones con integrinas y señalización por citocinas en el sistema inmune. En conjunto, estas vías sugieren la regulación positiva de los procesos del sistema inmunológico, posiblemente relacionado con el reclutamiento y maduración de linfocitos [20].

El papel de los GPCRs está asociado directamente a la actividad de quimiocinas [82] de las que son receptores y que tienen una función quimioatrayente sobre las células inmunes en circulación. Estos datos sugieren que en el caso de los tumores de cáncer primario de mama existe una regulación de la señalización celular, que podría tener un papel en el reclutamiento y maduración de las células inmunes presentes en el microambiente tumoral. Esto último es apoyado por el enriquecimiento de categorías como *regulación de la activación de células T*, que

tienen un papel importante en la inmunoregulación y que pueden coordinar según su fenotipo la función antitumoral o bien la inmunosupresión local por medio de la inhibición de células efectoras[48].

Las proteínas de la vía clásica del complemento se encuentran en la red, comp parte de los procesos del sistema inmune y de respuesta inflamatoria. Se ha propuesto que las proteínas efectoras de la cascada de complemento pueden promover el desarrollo de tumores [49]. De acuerdo con este modelo, la deposición de C1q en las paredes de los capilares promueve una respuesta inflamatoria pro-tumorigénica por medio de la producción de anafilatoxinas C3y C5a y el reclutamiento de MDSCs (*myeloid-derived suppressor cells*). La producción de altos niveles de RNA mensajero de C1qa y C1qb en nuestro conjunto de datos parece consistente con este modelo y asocia el fenotipo canceroso con un ambiente pro-inflamatorio [23].

Si bien existen en la red grupos considerables de genes con un patrón de expresión uniforme (todos *apagados*), una revisión más atenta a los procesos de *señalización* (Figura 5.7) y *proceso del sistema inmune* (Figura 5.6 componente de la parte superior izquierda) resalta un grupo de genes integrados por miembros del MHC clase I y II, quimiocinas de tipo CC, receptores de quimiocinas (GPCRs), receptores de citocinas (IL-2Rγ, IL-7R), marcadores CD (usualmente receptores de membrana), Factores de transcripción de la familia IRF (IRF1, IRF8), componentes del complemento y receptores (C1qa, C1qb, C3Ar1) todos ellos conectados entre sí y con patrones de expresión diferencial dispares.

Debido a su interacción como vecinos cercanos en la red y a sus niveles de expresión no uniformes parece razonable pensar que se trata de un grupo de genes que pueden coexpresarse, pero que son objeto de una regulación a nivel individual [31], posiblemente importante dentro del fenotipo. La presencia de moléculas del del tipo de MHC I y II sugiere así mismo una regulación del sistema inmune adaptativo [28]. Una interpretación posible es que procesos como la presentación de antígenos, la cicatrización y la inflamación están estrechamente coordinados en

los tejidos. La inmunosupresión es importante como un mecanismo que impide la destrucción del tejido una vez que se ha iniciado con el proceso de reparación de daños en la fase de cicatrización [83] [46] [29].

Estos resultados tomados en conjunto apuntan a un fenotipo inflamado y son consistentes con varias de las características del cáncer mencionadas anteriormente [15] [16]: la *autosuficiencia de señales de crecimiento* y la *insensibilidad a las señales anticrecimiento* son en buena medida explicadas por la modulación de los procesos de comunicación celular mediados por los distintos tipos de receptores y mensajeros extracelulares: citocinias y quimiocinas. La *evasión de la apoptosis* y *angiogénesis sostenida* están asociadas al proceso de cicatrización, aunque su efecto debe ser temporal para no derivar en un proceso patológico [84].

El proceso *preñez* (female pregnancy) resulta un tanto inesperado en el contexto de tumorigénesis, sin embargo aparece enriquecido en la red con un p de 1.5979×10^{-8} y 33 genes que incluyen glicoproteínas específicas del embarazo (PSG1 a la 7 y PSG11), prolactina, receptor de eritropoietina, receptor de leptina y otras más todas las cuáles se encuentran subexpresadas. Es interesante la representación de un proceso relacionado con una inmunosupresión temporal que evita el rechazo del feto. Moléculas como PSG1 son las proteínas fetales más abundantes en la sangre materna durante el embarazo y contribuyen a la inmunotolerancia y desarrollo de la vasculatura placentaria [85], sin embargo no existe aún una anotacion satisfactoria para las demás PSGs. Se debe notar que el nivel de expresión de las PSGs durante el embarazo es alto, en contraste con los genes en este conjunto de muestras de cáncer de mama donde se encuentran subexpresados.

6.4. Distribución de los procesos en la red y procesos en las comunidades

La primera comunidad (Figura 5.3(a)), que cuenta con 105 genes, no mostró un enriquecimiento significativo para ningun proceso de GO. Esta comunnidad contiene genes subexpresados, muchos de los cuales corresponden a citocromos de la familia de P450, como CYP2C9, que es el gen con mayor número de interacciones de la red (408). Los miembros de la familia del citocromo P450 se asocian comunmente con el procesamiento de fármacos y la activación de carcinógenos químicos [86]. La actividad de algunos antiinflamatorios no esteroideos como el ácido acetilsalicílico depende de su procesamiento a otras formas biológicamente activas como el salicilato, el ácido gentisico y el salicil coenzima a que inhiben la producción de prostaglandinas acetilando a las ciclooxigenasas que las producen [87]. La presencia de variantes activas de CYP2C9 se ha asociado con una reducción en el riesgo de desarrollar adenoma del colon en pacientes que consumen aspirina (ibídem). El uso de antiinflamatorios no esteroides, incluida la aspirina, también se ha asociado con una reducción en el riesgo de desarrollar cáncer de mama [88].

En el caso de los genes en esta comunidad, así como de los demás genes subexpresados en la red, elque estén subexpresados y aparezcan con una correlación alta en los niveles de expresión con otros genes, nos muestra que éstos no se encuentran simplemente inactivados. Por el contrario, los niveles de expresión de estos genes varían dentro de rangos (Figura 4.1), que son diferentes cuando comparamos a los tejidos de tumores primarios de mama contra los tejidos mamarios normales.

La segunda comunidad (Figura 5.3(b), cuadro 5.6) se conforma por genes subexpresados, entre los que se encuentran numerosos interferones de clase I. Los interferones son citocinas que se producen típicamente en respuesta a la infección por virus y poseen funciones inmunomoduladoras [89]. Esta comunidad muestra un enriquecimiento estadísticamente significativo para el proceso *respuesta a virus*, sin embargo consta de genes subexpresados, lo que sugiere un proceso subregulado.

La tercera comunidad (Figura 5.3(c), cuadro 5.6) se encuentra enriquecida en procesos relacionados con la transducción de señales a través de la heterodimerización de proteínas, y en procesos relacionados con morfogénesis. Es notable el hecho de que el enriquecimiento estadístico de los procesos en esta comunidad se encuentre asociado a una cantidad pequeña de genes (de 2 a 5). BCL-2 es un gen que codifica para una proteína que regula la apoptosis asociada a estrés mitocondrial [90]; este gen se enuentra representado en cuatro de los cinco procesos enriquecidos para esta comunidad. El enriquecimiento de procesos con cinco genes, en una comunidad que se conforma por veintiseis genes hace pertinente la pregunta de si los demás genes están cumpliendo una funcion biológica determinada que no estamos detectando, basados en las funciones (procesos) que conocemos y cuyos genes tenemos identificados.

Entre los genes con sobreexpresión, los encontrados en la comunidad número 4 (Figura 5.3(d), cuadro 5.6) corresponden a procesos relacionados con la producción de matriz extracelular. La colágena III y colágena I son depositadas en el proceso de cicatrización [29]. La formación de vasos sanguíneos defectuosos se ha propuesto como un mecanismo para inducir la cicatrización en favor del crecimiento tumoral [13].

Una cuestión que resulta interesante es el de la distribución de los procesos en la red. El resultado del análisis de enriquecimiento por comunidades, muestra que para grupos de genes estrechamente relacionados en la red no existe una función única y bien definida en términos de los procesos biológicos *estándar* (Cuadro **??** y Figura 5.3). Por otro lado, al tomar la red completa y extraer de ésta los conjuntos de genes correspondientes a un solo proceso o vía conocidos, éstos genes no forman un solo grupo conectado (Figuras 5.6, 5.7 y 5.8).

Estos resultados toman sentido cuando consideramos la forma en que se desarrolla la respuesta inflamatoria en el tiempo, en el cual la integración de señales

ambientales conduce a la expresión de grupos de genes en etapas sucesivas [91]. Éstos grupos son capaces de coordinar los pasos siguientes de la respuesta [24]. La integración de señales es crucial para el proceso, ya que existe un compromiso entre los mecanismos destructivos que aseguran la eliminación de patógenos y los mecanismos que permiten la reparación de los tejidos dañados [23][24][92]. La fragmentación de los procesos representados por los grupos de genes en la red podría estar relacionada con diferentes momentos en la respuesta; así mismo, las comunidades que representan a varios procesos podrían estar representando los ajustes que realiza la célula, para los diferentes procesos que actuarán en conjunto, en una etapa particular de la respuesta inflamatoria.

Para comprobar esta hipótesis es necesario realizar análisis detallados de vías, complementados con redes metabólicas, de regulación traduccional y redes de interacción de proteínas de tal modo que se integre la información de la estructura y dinámica de estos sistemas [3]. Desde un punto de vista topológico, las vías se pueden considerar como mapas de caminos, que pueden estar abiertos o cerrados. Los niveles de expresión observados en los grupos de genes coregulados pueden proporcionar indicios sobre la transitabilidad de las vías en las que participan [93].

Los resultados obtenidos apuntan a que la inflamación es un factor importante para el desarrollo del fenotipo del cáncer primario de mama, y a que ésta se encuentra estrechamente relacionada con procesos de comunicación celular, presentación de antígenos y modificaciónes en la matriz extracelular. Dichos procesos han sido propuestos en la literatura como mecanismos importantes del desarrollo tumoral. Ésto constituye un primer nivel de validación sobre la utilidad del análisis de redes complejas, ya que estamos recuperando información biológica que se ha obtenido por medio de análisis más tradicionales, como es el caso de experimentos con modelos animales (Sección 2.2). Así mismo, la asociación de genes basada en los datos permite recuperar ótros procesos corregulados de los que no tenemos una sospecha previa, como es el caso de los genes del proceso de preñez.

Al tratarse de una metodología nueva, que integra el uso de herramientas de varios campos que incluyen a la biología, estadística y ciencias de la computación, es esperable que surjan problemas por resolver, como es el caso de la calidad de las anotaciones, de la validez de los conjuntos de datos experimentales y de lo adecuado de los análisis estadísticos realizados.

La calidad de las anotaciones resulta relevante para obtener interpretaciones que resulten claras, lo que representa un problema si muchas de las moléculas son aún poco conocidas. Un solo gen puede servir como molde para numerosas proteínas, que a su vez pueden ser modificadas de diversas formas y jugar papeles biológicos diferentes dependiendo del contexto, de su localización en la célula, etc. Es posible que moléculas bien conocidas tengan funciones que aún no hemos reconocido; tambien es posible que funciones biológicas conocidas se lleven a cabo por otras moléculas no identificadas aún.

La validez de los conjuntos experimentales, debido a que éstos consisten en varios cientos, o incluso miles de muestras, serían difíciles de producir por medio de un solo experimento controlado, especialmente en el caso de datos provenientes de pacientes humanos. La selección de los conjuntos de datos depende, en este caso de un curado meticuloso de muestras provenientes de distintos conjuntos de datos que cuenten con información detallada sobre el tipo de estudio para el que se realizaron, los antecedentes clínicos del paciente, el historial de tratamiento recibido, etc. que constituyen las variables experimentales que controlamos por medio de la selección de muestras.

Finalmente, respecto al uso de métodos estadísticos para el análisis de datos, es importante considerar que muchas de estas herramientas fueron adaptadas en un principio para el manejo de conjuntos de datos relativamente pequeños. El uso de conjuntos de datos de gran volumen trae consigo la generación de valores de significancia estadística en un rango de varios ordenes de magnitud, en algunos casos extremadamente pequeños. Como ejemplo, el valor de corte utilizado para

decidir si un gen está diferencialmente expresado en un conjunto del orden de una decena de muestras, es de *log-dds* ≥ 6, mientras que para nuestro conjunto de datos con alrededor de 800 muestras obtenemos valores que puede ir hasta *log-odds* $=$ 250 o mayores. El mismo caso se presenta con la significancia estadística o *p-value* de las interacciones entre genes de la red en donde, para esta red particular se utilizaron valores de *p-value*$\leq 1 \times 10^{-100}$.

Capítulo 7

Conclusiones

La inferencia y el análisis de redes de regulación transcripcional son herramientas que permiten la integración de datos experimentales de gran volumen, proporcionando una perspectiva global de la estructura biológica. Si bien los estudios a nivel de genes individuales son fundamentales en la búsqueda de las diferencias fenotípicas que expliquen el cáncer de mama, es dificil interpretar resultados altamente heterogéneos [2]. Así mismo, debemos recordar que el fenotipo resulta de la integración funcional de todas las moléculas que conforman al sistema biológico, ya sea una célula, un tejido, un órgano o el organismo completo. La exploración desde una perspectiva de sistemas es un paso natural en el proceso de integración de los datos de alto volumen con los que contamos actualmente [3].

A través de los análisis de de enriquecimiento estadístico de procesos biológicos, encontramos evidencia de procesos inflamatorios activos en cáncer primario de mama, estrechamente relacionados a nivel transcripcional con procesos del sistema inmune adaptativo, particularmente con la presentación de antígenos, así como la modificación de la matriz extracelular característica del proceso de cicatrización.

El análisis pone en evidencia que los grupos de genes agrupados en comunidades no forman parte de procesos únicos. Los genes pueden codificar proteínas que toman parte en una gran cantidad de procesos biológicos; Ésto puede estar relacio-

nado con el enriquecimiento de múltiples procesos dentro de la misma comunidad, todos los cuales involucran al mismo gen (o genes). Los genes coexpresados establecen un contexto funcional una vez que todos estos se hayan traducido, lo que no podemos observar aún con éste tipo de análisis debido a que estamos limitados a los niveles de expresión de RNA mensajero.

La estructura de la red muestra que la coexpresión de genes que participan en procesos biológicos conocidos no se da en forma de bloques únicos; por el contrario, en procesos biológicos en los que están involucrados un gran número de proteínas, la red muestra grupos de genes coexpresados, en muchos casos separados, lo que sugiere una regulación de los procesos por bloques, posiblemente separados en el tiempo.

7.1. Perspectivas

Como un siguiente paso consideraremos la exploración de las redes de regulación transcripcional por subtipos de cáncer de mama, especialmente debido a que éstos subtipos moleculares presentan diferencias en los patrones de expresión genética, así como en la respuesta al tratamiento, lo que tiene un impacto terapéutico. Entender las diferencias en los patrones de regulación genética en los distintos subtipos podrá, potencialmente, revelar los mecanismos que llevan al efecto diferencial de los fármacos, lo que repercutirá en la prescripción de tratamientos a los pacientes.

Creemos recomendable explorar la ubicación de los grupos de genes en los cromosomas como una forma de investigar si estamos viendo una regulación a nivel de regiones completas del genoma, o a nivel de genes individuales; particularmente con el grupo de genes que son vecinos de CYP2C9, que es el gen con más conexiones en la red.

Para entender más detalladamente el papel que están jugando los procesos biológicos encontrados, es necesario el análisis de expresión diferencial de sus regu-

ladores positivos y negativos, con la finalidad de entender si realmente los procesos son activados o apagados. Esta cuestión nos remite de nuevo al problema de la anotación de los genes y las proteínas que se traducen a partir de ellos, así como de otros productos que no se traducen como miRNAs, lncRNAs, etc.

Para explorar los patrones de regulación de los procesos en el tiempo, es necesario el ordenamiento de las muestras con respecto a algún indicador cronológico o de progresión de la enfermedad. Esto supone el buscar si existe dicho indicador para el tipo de tejidos que estamos estudiando y el desarrollo de técnicas para clasificar las muestras.

Bibliografía

[1] J Ferlay, I Soerjomataram, M Ervik, R Dikshit, S Eser, C Mathers, M Rebelo, DM Parkin, D Forman, and F Bray. Globocan 2012 v1. 0, cancer incidence and mortality worldwide: Iarc cancerbase no. 11 [internet]. lyon, france: International agency for research on cancer. c2013 [cited 2013 oct 17]. *globocan. iarc. fr/Default. aspx. Accessed*, 2014.

[2] Cancer Genome Atlas Network et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[3] Hiroaki Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–1664, 2002.

[4] Hernández-Lemus. M Mayorga. K, Baca-López. E. Information-theoretical analysis of gene expression data to infer transcriptional interactions. *REVISTA MEXICANA DE FÍSICA*, 55(6):456–466, Diciembre 2009.

[5] Mukesh Bansal, Vincenzo Belcastro, Alberto Ambesi-Impiombato, and Diego Di Bernardo. How to infer gene networks from expression profiles. *Molecular systems biology*, 3(1), 2007.

[6] Julian Peto. Cancer epidemiology in the last century and the next decade. *Nature*, 411(6835):390–395, 2001.

[7] Cristian Tomasetti and Bert Vogelstein. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, 347(6217):78–81, 2015.

[8] Janet S Butel. Viral carcinogenesis: revelation of molecular mechanisms and etiology of human disease. *Carcinogenesis*, 21(3):405–426, 2000.

[9] Fran Balkwill and Alberto Mantovani. Inflammation and cancer: back to virchow? *The lancet*, 357(9255):539–545, 2001.

[10] Karin E de Visser, Lidiya V Korets, and Lisa M Coussens. De novo carcinogenesis promoted by chronic inflammation is b lymphocyte dependent. *Cancer cell*, 7(5):411–423, 2005.

[11] Alberto Mantovani, Tiziana Schioppa, Chiara Porta, Paola Allavena, and Antonio Sica. Role of tumor-associated macrophages in tumor progression and invasion. *Cancer and Metastasis Reviews*, 25(3):315–322, 2006.

[12] Sergei I Grivennikov, Florian R Greten, and Michael Karin. Immunity, inflammation, and cancer. *Cell*, 140(6):883–899, 2010.

[13] Harold F Dvorak. How tumors make bad blood vessels and stroma. *The American journal of pathology*, 162(6):1747–1757, 2003.

[14] Howard Y Chang, Julie B Sneddon, Ash A Alizadeh, Ruchira Sood, Rob B West, Kelli Montgomery, Jen-Tsan Chi, Matt Van De Rijn, David Botstein, and Patrick O Brown. Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. *PLoS biology*, 2(2):e7, 2004.

[15] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *cell*, 100(1):57–70, 2000.

[16] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

[17] Alberto Mantovani, Paola Allavena, Antonio Sica, and Frances Balkwill. Cancer-related inflammation. *Nature*, 454(7203):436–444, 2008.

[18] Hans Clevers. At the crossroads of inflammation and cancer. *Cell*, 118(6):671–674, 2004.

[19] Jeffrey R Jackson, MP Seed, CH Kircher, DA Willoughby, and JD Winkler. The codependence of angiogenesis and chronic inflammation. *The FASEB Journal*, 11(6):457–465, 1997.

[20] David G DeNardo and Lisa M Coussens. Balancing immune response: crosstalk between adaptive and innate immune cells during breast cancer progression. *Breast Cancer Res*, 9(4):212, 2007.

[21] Lisa M Coussens and Zena Werb. Inflammation and cancer. *Nature*, 420(6917):860–867, 2002.

[22] Alberto Mantovani. Cancer: inflaming metastasis. *Nature*, 457(7225):36–37, 2009.

[23] Ruslan Medzhitov. Origin and physiological roles of inflammation. *Nature*, 454(7203):428–435, 2008.

[24] Carl Nathan. Points of control in inflammation. *Nature*, 420(6917):846–852, 2002.

[25] Grace Y Chen and Gabriel Nuñez. Sterile inflammation: sensing and reacting to damage. *Nature Reviews Immunology*, 10(12):826–837, 2010.

[26] Kezhong Zhang and Randal J Kaufman. From endoplasmic-reticulum stress to the inflammatory response. *Nature*, 454(7203):455–462, 2008.

[27] Peter J. Murray and Thomas A. Wynn. Protective and pathogenic functions of macrophage subsets. *Nat Rev Immunol*, 11(11):723737, Oct 2011.

[28] Thomas F Gajewski, Hans Schreiber, and Yang-Xin Fu. Innate and adaptive immune cells in the tumor microenvironment. *Nature immunology*, 14(10):1014–1022, 2013.

[29] Matthias Schäfer and Sabine Werner. Cancer as an overhealing wound: an old hypothesis revisited. *Nature reviews Molecular cell biology*, 9(8):628–638, 2008.

[30] Osamu Takeuchi and Shizuo Akira. Pattern recognition receptors and inflammation. *Cell*, 140(6):805–820, 2010.

[31] Ruslan Medzhitov and Tiffany Horng. Transcriptional control of the inflammatory response. *Nat Rev Immunol*, 9(10):692–703, Oct 2009.

[32] Khyati Bhatelia, Kritarth Singh, and Rajesh Singh. Tlrs: Linking inflammation and breast cancer. *Cellular signalling*, 2014.

[33] Shizuo Akira and Kiyoshi Takeda. Toll-like receptor signalling. *Nature reviews immunology*, 4(7):499–511, 2004.

[34] Ute Oltmanns, Razao Issa, Maria B Sukkar, Matthias John, and K Fan Chung. Role of c-jun n-terminal kinase in the induced release of gm-csf, rantes and il-8 from human airway smooth muscle cells. *British journal of pharmacology*, 139(6):1228–1234, 2003.

[35] Kate Schroder and Jurg Tschopp. The inflammasomes. *Cell*, 140(6):821–832, 2010.

[36] Mansi Saxena and Garabet Yeretssian. Nod-like receptors: master regulators of inflammation and cancer. *Frontiers in immunology*, 5, 2014.

[37] Subhashini Sadasivam, Sanjeev Gupta, Vegesna Radha, Kiran Batta, Tapas K Kundu, and Ghanshyam Swarup. Caspase-1 activator ipaf is a p53-inducible gene involved in apoptosis. *Oncogene*, 24(4):627–636, 2005.

[38] Peyman Nakhaei, Thibault Mesplede, Mayra Solis, Qiang Sun, Tiejun Zhao, Long Yang, Tsung-Hsien Chuang, Carl F Ware, Rongtuan Lin, and John Hiscott. The e3 ubiquitin ligase triad3a negatively regulates the rig-i/mavs signaling pathway by targeting traf3 for degradation. *PLoS pathogens*, 5(11):e1000650, 2009.

[39] Pierre Guermonprez, Jenny Valladeau, Laurence Zitvogel, Clotilde Théry, and Sebastian Amigorena. Antigen presentation and t cell stimulation by dendritic cells. *Annual review of immunology*, 20(1):621–667, 2002.

[40] Facundo D Batista and Naomi E Harwood. The who, how and where of antigen presentation to b cells. *Nature Reviews Immunology*, 9(1):15–27, 2009.

[41] Teunis BH Geijtenbeek and Sonja I Gringhuis. Signalling through c-type lectin receptors: shaping immune responses. *Nature Reviews Immunology*, 9(7):465–479, 2009.

[42] Toby Lawrence. The nuclear factor nf-κb pathway in inflammation. *Cold Spring Harbor perspectives in biology*, 1(6):a001651, 2009.

[43] Jordan S Pober and William C Sessa. Evolving functions of endothelial cells in inflammation. *Nature Reviews Immunology*, 7(10):803–815, 2007.

[44] Lionel B Ivashkiv and Laura T Donlin. Regulation of type i interferon responses. *Nature Reviews Immunology*, 14(1):36–49, 2014.

[45] Nancy Hogg, Irene Patzak, and Frances Willenbrock. The insider's guide to leukocyte integrin signalling and function. *Nature Reviews Immunology*, 11(6):416–426, 2011.

[46] Dmitry I Gabrilovich and Srinivas Nagaraj. Myeloid-derived suppressor cells as regulators of the immune system. *Nature Reviews Immunology*, 9(3):162–174, 2009.

[47] David M Mosser and Justin P Edwards. Exploring the full spectrum of macrophage activation. *Nature Reviews Immunology*, 8(12):958–969, 2008.

[48] Dmitry I Gabrilovich, Suzanne Ostrand-Rosenberg, and Vincenzo Bronte. Coordinated regulation of myeloid cells by tumours. *Nature Reviews Immunology*, 12(4):253–268, 2012.

[49] Maciej M Markiewski, Robert A DeAngelis, Fabian Benencia, Salome K Ricklin-Lichtsteiner, Anna Koutoulaki, Craig Gerard, George Coukos, and John D Lambris. Modulation of the antitumor immune response by complement. *Nature immunology*, 9(11):1225–1235, 2008.

[50] Charles A Janeway, Paul Travers, MJ Walport, Mark J Shlomchik, et al. *Immunobiology: the immune system in health and disease*, volume 2. Churchill Livingstone, 2001.

[51] Hydar Ali and Reynold A Panettieri Jr. Anaphylatoxin c3a receptors in asthma. *Respir Res*, 6:19, 2005.

[52] Michael J Heller. Dna microarray technology: devices, systems, and applications. *Annual review of biomedical engineering*, 4(1):129–153, 2002.

[53] Jörg D Hoheisel. Microarray technology: beyond transcript profiling and genotype analysis. *Nature reviews genetics*, 7(3):200–210, 2006.

[54] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

[55] Karol Baca-López, Miguel Mayorga, Alfredo Hidalgo-Miranda, Nora Gutiérrez-Nájera, and Enrique Hernández-Lemus. The role of master regulators in the metabolic/transcriptional coupling in breast carcinomas. *PloS one*, 7(8):e42678, 2012.

[56] Paul Anderson. Post-transcriptional regulons coordinate the initiation and resolution of inflammation. *Nat Rev Immunol*, 10(1):2435, Jan 2010.

[57] C Rangel-Escareño. *Herramientas computacionales para el análisis de datos de microarreglos de expresión*. INMEGEN, Octubre 2010.

[58] Ronald Aylmer Fisher. The design of experiments. 1935.

[59] Raymond Hubbard and María Jesús Bayarri. Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *The American Statistician*, 57(3):171–178, 2003.

[60] Ronald Christensen. Testing fisher, neyman, pearson, and bayes. *The American Statistician*, 59(2):121–126, 2005.

[61] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[62] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[63] Enrique Hernández-Lemus, David Velázquez-Fernández, Jesús K Estrada-Gil, Irma Silva-Zolezzi, Miguel F Herrera-Hernández, and Gerardo Jiménez-Sánchez. Information theoretical methods to deconvolute genetic regulatory networks applied to thyroid neoplasms. *Physica A: Statistical Mechanics and its Applications*, 388(24):5057–5069, 2009.

[64] E Hernández-Lemus and C Rangel-Escareño. The role of information theory in gene regulatory network inference. *Information Theory: New Research*, pages 109–144, 2011.

[65] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo D Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7(Suppl 1):S7, 2006.

[66] James Frey, Todd Tannenbaum, Miron Livny, Ian Foster, and Steven Tuecke. Condor-g: A computation management agent for multi-institutional grids. *Cluster Computing*, 5(3):237–246, 2002.

[67] Melissa S Cline, Michael Smoot, Ethan Cerami, Allan Kuchinsky, Nerius Landys, Chris Workman, Rowan Christmas, Iliana Avila-Campilo, Michael Creech, Benjamin Gross, et al. Integration of biological networks and gene expression data using cytoscape. *Nature protocols*, 2(10):2366–2382, 2007.

[68] Kristian A Gray, Bethan Yates, Ruth L Seal, Mathew W Wright, and Elspeth A Bruford. Genenames. org: the hgnc resources in 2015. *Nucleic acids research*, page gku1071, 2014.

[69] Jing Wang, Dexter Duncan, Zhiao Shi, and Bing Zhang. Web-based gene set analysis toolkit (webgestalt): update 2013. *Nucleic acids research*, page gkt439, 2013.

[70] Tomihisa Kamada and Satoru Kawai. An algorithm for drawing general undirected graphs. *Information processing letters*, 31(1):7–15, 1989.

[71] Nadezhda T Doncheva, Yassen Assenov, Francisco S Domingues, and Mario Albrecht. Topological analysis and interactive visualization of biological networks and protein structures. *Nature protocols*, 7(4):670–685, 2012.

[72] Jun Dong and Steve Horvath. Understanding network concepts in modules. *BMC systems biology*, 1(1):24, 2007.

[73] Reka Albert. Scale-free networks in cell biology. *Journal of cell science*, 118(21):4947–4957, 2005.

[74] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[75] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-worldnetworks. *nature*, 393(6684):440–442, 1998.

[76] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[77] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[78] Jongkwang Kim and Thomas Wilhelm. What is a complex graph? *Physica A: Statistical Mechanics and its Applications*, 387(11):2637–2652, 2008.

[79] Gary D Bader and Christopher WV Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics*, 4(1):2, 2003.

[80] Steven Maere, Karel Heymans, and Martin Kuiper. Bingo: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16):3448–3449, 2005.

[81] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661703, Nov 2009.

[82] Israel F Charo and Richard M Ransohoff. The many roles of chemokines and chemokine receptors in inflammation. *New England Journal of Medicine*, 354(6):610–621, 2006.

[83] Gregory T Motz and George Coukos. The parallel lives of angiogenesis and immunosuppression: cancer and other tales. *Nature Reviews Immunology*, 11(10):702–711, 2011.

[84] David G Greenhalgh. The role of apoptosis in wound healing. *The international journal of biochemistry & cell biology*, 30(9):1019–1030, 1998.

[85] Felipe A Lisboa, James Warren, Gisela Sulkowski, Marta Aparicio, Guido David, Enrique Zudaire, and Gabriela S Dveksler. Pregnancy-specific glycoprotein 1 induces endothelial tubulogenesis through interaction with cell surface proteoglycans. *Journal of Biological Chemistry*, 286(9):7577–7586, 2011.

[86] C Rodriguez-Antona and M Ingelman-Sundberg. Cytochrome p450 pharmacogenetics and cancer. *Oncogene*, 25(11):16791691, Mar 2006.

[87] Jeannette Bigler, John Whitton, Johanna W Lampe, Lisa Fosdick, Roberd M Bostick, and John D Potter. Cyp2c9 and ugt1a6 genotypes modulate the protective effect of aspirin on colon adenoma risk. *Cancer research*, 61(9):3566–3569, 2001.

[88] LA Garcia Rodriguez and A Gonzalez-Perez. Risk of breast cancer among users of aspirin and other anti-inflammatory drugs. *British journal of cancer*, 91(3):525–529, 2004.

[89] Jos M. Gonzlez-Navajas, Jongdae Lee, Michael David, and Eyal Raz. Immunomodulatory functions of type i interferons. *Nat Rev Immunol*, Jan 2012.

[90] Eicke Latz, T. Sam Xiao, and Andrea Stutz. Activation and regulation of the inflammasomes. *Nat Rev Immunol*, 13(6):397411, May 2013.

[91] Karuppiah Kannan, Ninette Amariglio, Gideon Rechavi, Jasmine Jakob-Hirsch, Itai Kela, Naftali Kaminski, Gad Getz, Eytan Domany, and David Givol. Dna microarrays identification of primary and secondary target genes regulated by p53. *Oncogene*, 20(18):2225–2234, 2001.

[92] Charles N Serhan, Nan Chiang, and Thomas E Van Dyke. Resolving inflammation: dual anti-inflammatory and pro-resolution lipid mediators. *Nature Reviews Immunology*, 8(5):349–361, 2008.

[93] Adilson E Motter, Natali Gulbahce, Eivind Almaas, and Albert-László Barabási. Predicting synthetic rescues in metabolic networks. *Molecular Systems Biology*, 4(1), 2008.