UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

FACULTAD DE MEDICINA

BIOMEDICINA

IMPLEMENTACIÓN DE MÉTODOS COMPUTACIONALES DE BÚSQUEDA SISTEMÁTICA DE REGULADORES TRANSCRIPCIONALES MAESTROS EN CÁNCER

TESIS

QUE PARA OPTAR POR EL GRADO DE:

DOCTOR EN CIENCIAS

PRESENTA:

TOVAR ROMERO HUGO ANTONIO

TUTOR PRINCIPAL DE TESIS: DR. ENRIQUE HERNÁNDEZ LEMUS
INSTITUTO NACIONAL DE MEDICINA GENÓMICA

COMITÉ TUTOR: DR. LUIS ANTONIO MENDOZA SIERRA
INSTITUTO DE INVESTIGACIONES BIOMÉDICAS, UNAM

DR. LEÓN PATRICIO MARTÍNEZ CASTILLA
FACULTAD DE QUÍMICA, UNAM

MÉXICO, CD. MX. **ABRIL, 2016**

UNAM
POSGRADO
Ciencias Biológicas

COORDINACIÓN

Isidro Ávila Martínez
Director General de Administración Escolar, UNAM
P r e s e n t e

Me permito informar a usted que el Subcomité de Biología Evolutiva y Sistemática del Posgrado en Ciencias Biológicas, en su sesión ordinaria del día 19 de octubre de 2015, aprobó el jurado para la presentación del examen para obtener el grado de **DOCTOR EN CIENCIAS** del alumno **TOVAR ROMERO HUGO ANTONIO** con número de cuenta **93107084** con la tesis titulada **"IMPLEMENTACIÓN DE MÉTODOS COMPUTACIONALES DE BÚSQUEDA SISTEMÁTICA DE REGULADORES TRANSCRIPCIONALES MAESTROS EN CÁNCER"**, realizada bajo la dirección del **DR. JOAQUIN ALEJANDRO ZÚÑIGA RENRIQUE HERNÁNDEZ LEMUS:**

Presidente:	DRA. MARTHA ROBLES FLORES
Vocal:	DR. JESÚS ESPINAL ENRÍQUEZ
Secretario:	DR. LEÓN PATRICIO MARTÍNEZ CASTILLA
Suplente:	DR. GUSTAVO MARTÍNEZ MEKLER
Suplente:	DR. LUIS ANTONIO MENDOZA SIERRA

Sin otro particular, me es grato enviarle un cordial saludo.

A T E N T A M E N T E
"POR MI RAZA HABLARÁ EL ESPÍRITU"
Cd. Universitaria, Cd. Mx., a 31 de marzo de 2016

DRA. MARÍA DEL CORO ARIZMENDI ARRIAGA
COORDINADORA DEL PROGRAMA

POSGRADO EN CIENCIAS BIOLOGICAS
COORDINACIÓN

Agradecimientos

Agradezco al Posgrado en Ciencias Biológicas de la Universidad Nacional Autónoma de México (UNAM). También agradezco a la UNAM por la beca que me otorgó el primer año de mi doctorado así como al Consejo Nacional de Ciencia y Tecnología por la beca de los siguientes tres años (No de becario 202668). Quiero agradecer a mi tutor principal el Dr. Enrique Hernandez Lemus por todo su apoyo, guía y todos los consejos (personales y académicos) que siempre me ha dado. Así como a los miembros de mi comité tutor: el Dr. Luis Antonio Mendoza Sierra y el Dr. León Patricio Martínez Castilla por todo el apoyo y las ideas que siempre enriquecieron este trabajo. Igualmente agradezco a los miembros del jurado por todas sus enriquecedoras ideas y propuestas.

Agradecimientos

Quiero agradecer a título personal a mi compañera de vida y paternidad Paty que ha sabido comprender y apoyar este importante esfuerzo que ahora concluimos juntos. A mi maestro, gurú y amigo Jorge Meave. Siempre a mi lado, me ha aconsejado en tantos caminos y en tantos retos. Muchas gracias Gor por nunca escatimar tu amistad a mi y a mi familia. A Rodrigo por lo que he aprendido con el y sus siempre gratas discusiones. También agradezco a mis compañeros de periplo: Aldo y Ana, Alejandro, Chucho, Daniel, Diana, Jesus, Joshua, Kahorik, Karol, Luis Felipe, Memo, Mike, Mireya, Ollin, Raúl, Roselyn, Serch y Tadeo. Muchas gracias por su ñoñez. A mis amigos, que siempre han creído en mi: Agus, Álvaro, Ani, Auroridae, Daniel, Edgar, Jessi, Naye, Nuria y Violeta.

A Nuria,
amor de mi vida.

Índice general

Implementación de métodos computacionales de búsqueda sistemática de reguladores transcripcionales maestros en cáncer

por

Hugo Antonio Tovar Romero

Resumen

Las redes de regulación genética son responsables de los delicados mecanismos que controlan la expresión genética. Bajo ciertas circunstancias, los programas de regulación genética pueden dar lugar a cascadas de transcripción. Tales cascadas son eventos en los que la activación de factores de transcripción clave, que son llamados *reguladores maestros* desencadenan una serie de eventos de expresión génica. La acción de los reguladores maestros de la transcripción es entonces importante para el establecimiento de ciertos programas como el de desarrollo y de diferenciación celular. Sin embargo, tales cascadas también se han relacionado con la aparición y establecimiento de fenotipos cancerosos.

Aquí se presenta una implementación sistemática de una serie de algoritmos destinados a la inferencia y análisis de reguladores transcripcionales maestros en el contexto de células de cáncer primario de mama. Dichos estudios se llevaron a cabo con una base de datos de 880 experimentos de microarreglos de expresión genética de tejido de cáncer de mama primario y de controles sanos provenientes de biopsias. También se llevaron a cabo análisis de funciones bioquímicas y de enriquecimiento de vías bioquímicas para estudiar el papel de los procesos que controlan, a nivel transcripcional, dichos reguladores maestros y su relación con el cáncer primario de mama. Se encontró que moléculas tales como AGTR2, ZNF132, TFDP3 y otras son reguladores maestros en esta red de regulación genética. Algunos conjuntos de genes controlados por estos reguladores están involucrados en procesos que se sabe bien que son característicos del cáncer. Mientras que los estudios clásicos abordan el problema molécula por molécula, los trabajos como éste ofrecen una enfoque deductivo que enriquece el conocimiento del desarrollo de fenotipos, en particular, los relativos a la biología del cáncer.

Implementación de métodos computacionales de búsqueda sistemática de reguladores transcripcionales maestros en cáncer

by

Hugo Antonio Tovar Romero

Abstract

Gene regulatory networks account for the delicate mechanisms that control gene expression. Under certain circumstances, gene regulatory programs may give rise to transcriptional cascades. Such cascades are events in which activation of key-responsive transcription factors called *master regulators* trigger a series of gene expression events. The action of transcriptional master regulators is then important for the establishment of certain programs like cell development and differentiation. However, such cascades have also been related with the onset and maintenance of cancer phenotypes.

Here we present a systematic implementation of a series of algorithms aimed at the inference and analysis of transcriptional master regulators in the context of primary breast cancer cells. Such studies were performed in a database of 880 microarray gene expression experiments on biopsy-captured tissue corresponding to primary breast cancer and healthy controls. Biological function and biochemical pathway enrichment analyses were also performed to study the role that the processes controlled –at the transcriptional level– by such master regulators may have in relation to primary breast cancer. We have found that molecules such as AGTR2, ZNF132, TFDP3 and others are master regulators in this gene regulatory network. Sets of genes controlled by these regulators are involved in processes that are well-known hallmarks of cancer. While classical studies address the problem molecule by molecule, works like this offer a deductive approach that enriches the knowledge of the development of phenotypes in particular those relating to the development of phenotypes, in particular, those regarding cancer biology.

Capítulo 1

Introducción

...Dios mueve al jugador, y éste, la pieza.
¿Qué Dios detrás de Dios la trama empieza
de polvo y tiempo y sueño y agonía?
— Jorge Luis Borges, *El hacedor* (1960)

Se ha sugerido que el cáncer es una enfermedad de vías bioquímicas [Hanahan y Weinberg, 2000]. Se considera que los principales *hallmarks* (sellos característicos) del cáncer están relacionadas con la proliferación, apoptosis, diferenciación celular y en general a la desregulación del ciclo celular y la alteración de los procesos de reparación del DNA [Hanahan y Weinberg, 2000 y 2011]. El fenotipo de una célula, sana o enferma, parece estar determinado por la actividad de cientos de genes y sus productos [Basso *et al.*, 2005]. Esta actividad está coordinada por intrincadas relaciones de regulación transcripcional y de una o varías vías de señalización. Por desgracia, actualmente, las interacciones moleculares que provocan estos procesos no se entienden bien. Una razón para esta falta de conocimiento es que la mayoría de estas vías no tienen una estructura de cadena sencilla, son entramados complejos de muchas redes regulatorias que determinan los procesos y respuestas celulares [Emmert-Streib *et al.*, 2014a].

Se ha observado que algunas de estas enredadas cascadas de genes están comúnmente controladas por unos cuantos genes conocidos como *Reguladores Transcripcionales Maestros* (RTM) [Han *et al.*, 2004; Sun-Kin Chan y Kyba, 2013; Mullen *et al.*, 2011]. En estos casos, estos pocos genes son responsables del control de todo el programa de regulación de la célula que define un tipo celular [Han *et al.*, 2004; Basso *et al.*, 2005; Affara *et al.*, 2013]. Los RTM pueden actuar

sobre procesos generales de la células [Hosking, 2012], pero también en fenotipos celulares específicos [Hinnebusch y Natarajan, 2002; Medvedovic *et al.*, 2011; Affara *et al.*, 2013]. Entender esta organización y encontrar estos genes reguladores es crucial para elucidar tanto la fisiología celular normal como los fenotipos patológicos complejos [Basso *et al.*, 2005].

La identificación de estos RTM se basa en la relación (inferida o empírica) entre ellos y sus blancos en las redes de regulación genética [Hernández-Lemus y Siqueiros-García, 2013]. El implementar un método computacional sistemático para identificar y analizar reguladores transcripcionales maestros que resulten relevantes para el cáncer de mama, particularmente en sus etapas más tempranas, puede ser de gran ayuda para entender tanto los mecanismos de control de la transcripción de los seres vivos, como la biología de la enfermedad.

1.1. Biología de sistemas

Uno de los objetivos de la biología de sistemas es tratar de entender la maquinaria celular de los sistemas biológicos para resolver el problema de la interrelación del genotipo con el fenotipo [Jansen, 2003; Rockman, 2008; Liu *et al.*, 2010]. Los estudios convencionales de biología molecular permiten la identificación de grupos de genes que afectan a cierto fenotipo con un enfoque inductivo. El procedimiento más característico de este enfoque es la inducción de función de una molécula a partir de observar el fenotipo de un organismo sin ella. La disponibilidad de mediciones de grandes cantidades de fenotipos moleculares permite el uso de algoritmos que tratan de dilucidar redes de regulación que tracen las relaciones entre el genotipo y el fenotipo [Jansen, 2003; Rockman, 2008; Liu *et al.*, 2010]. Desde esta perspectiva, se trata de deducir la función e importancia de una molécula a partir de la relación que tiene con otras en un contexto dado. Para tener una visión de las relaciones complejas que se pueden encontrar con la ayuda del enfoque de la biología de sistemas la Figura 1-1 muestra una serie de temas característicos (a menudo interdependientes) de los sistemas biológicos complejos [Hernández-Lemus, 2014].

Con la intención de abarcar estos aspectos, se han combinado las tecnologías ómicas (estudio de grandes conjuntos de elementos celulares como la genómica, proteómica o la epigenómica.)

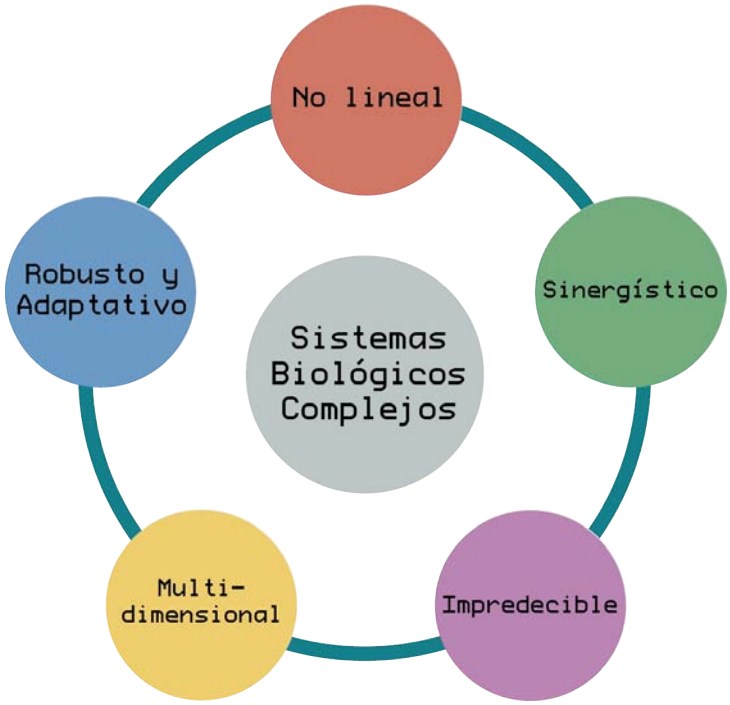


Figura 1-1: **Algunos temas característicos de los sistemas biológicos complejos.** Podemos notar que éstos son fenómenos dependientes que no pueden ser tratados separadamente, de ahí la necesidad de metodologías integrativas [Hernández-Lemus, 2014].

con modelos estadísticos [Visvanathan *et al.*, 2010; Jin *et al.*, 2007] y con técnicas computacionales diseñadas *ad hoc* [You, 2004; Kitano, 2002; Hernández Patiño *et al.*, 2013] para la integración de datos. Una tarea nada trivial ya que, incluso la gestión de la gran cantidad de datos representa ya un reto en el paradigma de "los grandes conjuntos de datos" (*big data*) [Tretyakov *et al.*, 2013; Schouten, 2013].

Éstas tecnologías nos proporcionan herramientas para medir cada uno de estos tipos de datos, de forma multiplexada (de muchos aspectos a las vez). Por ejemplo podemos tener experimentos para medir todo el transcriptoma o el proteoma, o incluso el interactoma. Por esta razón son necesarios los marcos de integración [Mosca *et al.*, 2010; Szabo *et al.*, 2000] para organizar e interpretar los datos experimentales [Kanehisa *et al.*, 2006; Emmert-Streib *et al.*, 2014b]. Además de la integración de datos, las bases de datos pueden proporcionar un sistema de consultas basadas en la ontología y de herramientas de análisis relacionadas con vías metabólicas, interacciones proteína-proteína, estructura de proteínas y modelado de sistemas de cáncer de mama. La razón detrás de este esfuerzo es que el cáncer necesita ser estudiado

también desde un punto de vista integrador tal como el que ofrece la biología de sistemas.

1.1.1. El acercamiento de la Biología de sistemas al cáncer de mama

Como ya se mencionó, uno de los temas primordiales dentro del enfoque de la biología de sistemas es la integración de datos [Baca-López *et al.*, 2014]. Dado que el cáncer es una patología que implica (al menos) la desregulación metabólica y hormonal, la inestabilidad genómica, una respuesta inmune anormal, anomalías en expresión génica, entrecruce de señalización, mutaciones, inflamación y plegamiento anómalo de proteínas, es un candidato natural para ser estudiado a través de la biología de sistemas [Hernández-Lemus, 2014].

Una de las primeras aplicaciones de las tecnologías ómicas –en particular del análisis de expresión génica basado en microarreglos– fue la búsqueda de firmas moleculares. Esto es, un conjunto específico de moléculas que, en teoría, definirían un fenotipo determinado. Éstas pueden utilizarse en la elaboración de perfiles de expresión génica de los tumores de cáncer de mama dirigido a la predicción de los resultados clínicos [van 't Veer *et al.*, 2002; Farmer *et al.*, 2005; Pau Ni *et al.*, 2010; Ruckhaeberle *et al.*, 2008; Tripathi *et al.*, 2008; Pollack *et al.*, 2002; Sotiriou *et al.*, 2006], o de los agentes de respuesta a las quimioterapias [Satih *et al.*, 2010]. Los perfiles de expresión génica han sido ampliamente usados también para mejorar el pronóstico [Sotiriou *et al.*, 2006; Liu *et al.*, 2007] y la subclasificación de tumores [Perou *et al.*, 2000; Sorlie *et al.*, 2003]; así como, para determinar la correlación de ciertas firmas moleculares con la metástasis del cáncer de mama [Wang *et al.*, 2005].

Los enfoques genómicos multinivel aplicados al estudio de la metástasis del cáncer de mama nos han llevado al desarrollo de las firmas moleculares, tanto a nivel de expresión génica como de interacción de proteínas. Se han diseñado conjuntos altamente reproducibles [Yao *et al.*, 2010] teniendo en cuenta que *diferentes* firmas de nodos pueden estar alteradas en *diferentes* conjuntos de pacientes lo cual pueden estar afectando la dinámica de *las mismas* vías asociadas con la metástasis del cáncer a través de la interacción con sus vecinos. En otras palabras, la metástasis del cáncer (algunos afirman que el cáncer en general) no es una condición centrada

en genes sino, más bien, una condición centrada en vías (o redes) [Baca-López *et al.*, 2012]. Yao *et al.* [2010] encuentran que la señalización en vías como la del ciclo celular, apoptosis, Jak-STAT, MAPK, ErbB, Wnt, y P53 se encontraban entre las más propensas a ser desreguladas de forma multi-gen/multi-objetivo; es decir, las vías asociadas a la enfermedad pueden depender de los cambios de co-expresión de *diferentes nodos de la firma* con el *mismo conjunto de vecinos* enriquecidos en esta vía metabólica.

En resumen, el enfoque de la biología de sistemas se basa en herramientas destinadas a un nivel de estudio global de los distintos conjuntos de datos ómicos; tales herramientas son métodos para la clasificación, cuantificación, cómputo, visualización, almacenamiento y recuperación de la información que proviene de los procesos biológicos. Un hito fundamental en biología de sistemas es que las redes de regulación bioquímica independientes de la escala proporcionan el marco integrador por excelencia, para explorar la compleja actividad de regulación en el contexto de fenotipos celulares específicos [Jeong *et al.*, 2000].

1.2. Redes de regulación bioquímica

Desde el punto de vista de la teoría de redes, los sistemas biológicos consisten en un conjunto (por lo general muy grande) de componentes en forma de moléculas biológicas: segmentos de DNA, transcritos de RNA, enzimas y otras proteínas, complejos de biomoléculas, entre otros. Estos componentes interactúan a través de una variedad de mecanismos: regulación génica, interacciones proteína-DNA, interacciones proteína-RNA, metabolismo, interacción proteína-proteína, vías bioquímicas por citar algunos ejemplos. En estas redes biológicas los componentes son los nodos, mientras que las interacciones son los enlaces.

Las redes bioquímicas se pueden construir a muchos niveles y pueden representar diferentes tipos de interacción. Comúnmente nos referimos a rutas bioquímicas, más que a redes, cuando estamos interesados en una serie muy particular de interacciones. Es importante recordar que las rutas bioquímicas nunca existen de manera aislada y éstas son, de hecho, parte de una

compleja red de interacción. Las redes bioquímicas que tradicionalmente se han considerado son:

1. Las redes metabólicas que representan las transformación química de los metabolitos.

2. Las redes de proteínas que representan interacciones proteína-proteína, como las de modificaciones de proteínas por señalización enzimática (también conocidas como redes de señalización) o la formación de complejos proteicos.

3. Las redes genéticas que representan las relaciones que se pueden establecer entre genes, cuando se observa cómo el nivel de expresión de cada uno afecta el nivel de expresión de otro [Brazhnik *et al.*, 2002].

Cada uno de estos tipos de red es una simplificación del sistema celular completo. La adopción de estas simplificaciones para la descripción de fenómenos específicos depende, con mucho, en cuál de los componentes celulares se va hacer la observación experimental. Esto es, por ejemplo, cuando se monitorea exclusivamente la expresión genética para estudiar cierto fenómeno, se está limitado a construir una red genética para explicar los datos.

La biología molecular, tradicionalmente ha propagado la idea de que los genes dictan todas las reglas en la célula [Brazhnik *et al.*, 2002]. Esto se materializa en el dogma central de de la biología molecular, el cual enfatiza que las proteínas, y sus metabolitos resultantes, solo son sintetizados cuando sus genes son activados. Este dogma no reconoce que la expresión de los genes también está influenciada por los niveles de metabolitos y proteínas [Brazhnik *et al.*, 2002]. ter Kuile y Westerhoff [2001] cuantificaron cómo la expresión genética y el metabolismo controlaban el flujo glicolítico en tres especies de protistas parásitos. Concluyeron que el flujo raramente es regulado por la simple expresión de genes. En algunos casos éste era regulado 30 % por expresión genética y 70 % por el metabolismo [ter Kuile y Westerhoff, 2001]. Si bien este tipo de hallazgos nos indica que en el futuro se necesita hacer más esfuerzo para monitorear los tres niveles de regulación, sigue siendo util estudiar la red genética por separado [Brazhnik *et al.*, 2002].

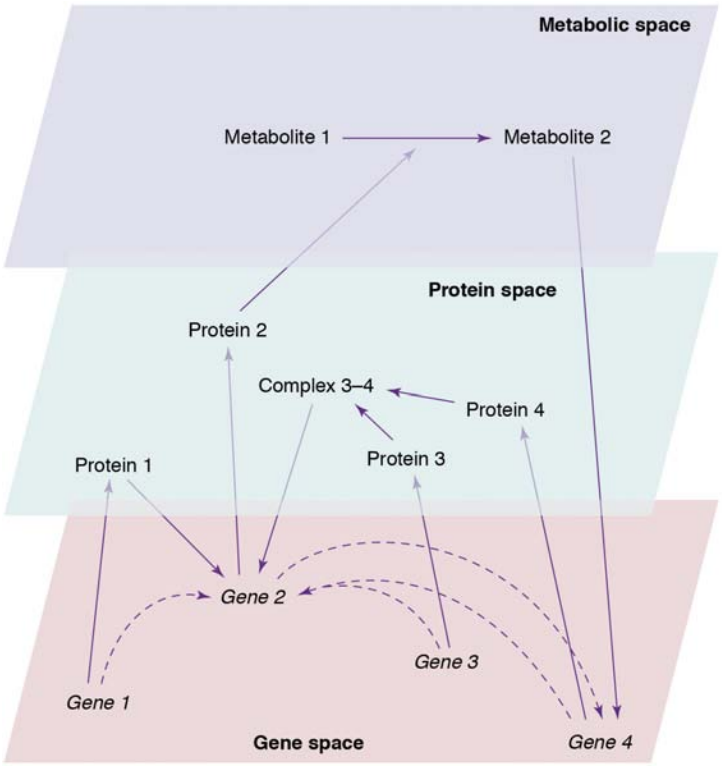


Figura 1-2: **Ejemplo de una red bioquímica**. Las moléculas están organizadas en tres niveles (espacios): mRNAs, proteínas y metabolitos. Las flechas continuas indican interacciones. Se muestran tres mecanismos diferentes de interacción gen-gen: regulación del gen 2 por el producto de la proteína del gen 1; regulación del gen 2 por el complejo 3-4 formado por el producto de los genes 3 y 4; y la regulación del gen 4 por el metabolito 2 el cual es producido por la proteína 2. Las líneas punteadas muestran la proyección de estas interacciones en el espacio del gen que constituiría su correspondiente red genética (figura obtenida de Brazhnik *et al.* [2002]).

En la figura 1-2 se representa un modelo de la red bioquímica global en la cual los tres niveles son representados como planos. En cada red bioquímica global, los genes no interactúan directamente con otros genes (ni tampoco sus correspondientes mRNAs); sin embargo, la inducción o represión de genes es consecuencia de la actividad de determinadas proteínas, las cuales, a su vez, son el producto de ciertos genes. Sin embargo, es útil abstraer estos estados entre proteínas y metabolitos, y representar la acción de los genes en otros genes en una red genética (también llamada red regulatoria genética o red de regulación o transcripción) [Brazhnik *et al.*, 2002].

La figura 1-3 es una representación gráfica de la red genética correspondiente de la red bioquímica de la figura 1-2. La mayoría de los genes en la redes genéticas presentan un efecto negativo en sus propias concentraciones ya que las tasas de degradación de su mRNA es proporcional a su concentración [Brazhnik *et al.*, 2002]. Sin embargo, estudios recientes señalan que la medición de solamente mRNA para inferir la presencia de su proteína puede ser una buena aproximación [Hernández-Lemus *et al.*, 2014].

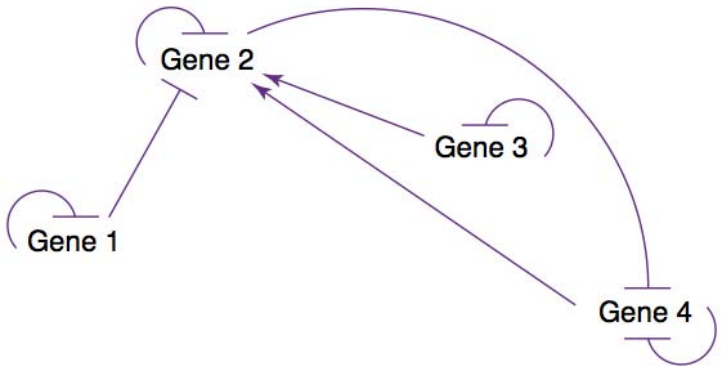


Figura 1-3: **Representación gráfica de la red bioquímica de la figura 1-2.** Las redes de genes son modelos que representan la relación causal entre las acciones de los genes, comúnmente a nivel de mRNA [Brazhnik *et al.*, 2002]. Las lineas muestran los efectos directos con flechas para señalar activación y barras para inhibición. Los ejes incluyen implícitamente los efectos del proteoma y del metaboloma como se representa en la figura 1-2.

1.2.1. Las redes de regulación genética y su inferencia

Las redes de regulación genética son modelos provenientes de la teoría de grafos que describen esta expresión genética en una población de células de manera unificada, bajo ciertas condiciones biológicas en un momento determinado. Estas redes se emplean a menudo para identificar interacciones genéticas a partir de datos experimentales por medio de modelos teóricos y computacionales. La biología de sistemas se basa en dos formas diferentes pero complementarias para la construcción de este tipo de redes, a menudo referidas como los enfoques "de abajo a arriba" (*bottom-up*) y "de arriba hacia abajo" (*top-down*) [Hernández Patiño *et al.*, 2013]. En el punto de vista bottom-up (a veces llamado cinético), se desarrollan modelos basados en la integración de la información ya disponible en las bases de datos, y luego esos modelos se prueban bajo una variedad de condiciones experimentales. En el enfoque top-down (o probabilístico), se usa un proceso guiado por los datos (*data-driven*) para inferir la correlación (o interacción) de los componentes de las redes a partir de datos masivos de experimentos genómicos; entonces se construyen modelos que puedan ser probados. Ambos puntos de vista son complementarios entre sí. El primero sirve para evaluar el modelo de ajuste, mientras que el segundo está dirigido a nuevos descubrimientos. El enfoque top-down se puede ejemplificar con la inferencia probabilística de redes genéticas de regulación [Margolin *et al.*, 2006a; Hernández-Lemus *et al.*, 2009] y su caracterización teórica fundamentalmente termodinámica.

En estas redes de regulación genética, los nodos representan genes y las aristas entre los nodos A y B representan que los genes A y B participan conjuntamente en alguna actividad reguladora. Por ejemplo, A puede ser el factor de transcripción de B, o A puede ser un miRNA que puede disminuir la expresión de B. Trabajos anteriores de regulación celular [Basso *et al.*, 2005; Sun-Kin Chan y Kyba, 2013; Mullen *et al.*, 2011] han mostrado que estas redes pueden estar dominadas por unos cuantos nodos que representarían a los reguladores transcripcionales maestros.

La inferencia de la topología de redes frecuentemente involucra la deconvolución de las interacciones a partir de las propiedades y la dinámica de, por ejemplo, niveles de mRNA en el perfil de un fenotipo celular específico [Hernández-Lemus, 2013]. En la última década, un gran número de estos métodos de deconvolución se han aplicado a estos sistemas. Entre los más utilizados están los conocidos como redes de co-expresión, los de algoritmo de agrupamiento, los métodos bayesianos, de ecuaciones diferenciales ordinarias y los métodos basados en teoría de la información [Bansal *et al.*, 2007]. Todos estos métodos se enfrentan, en diferentes grados, a diferentes problemas como son sobre-ajustes, elevada complejidad computacional, dependencia de modelos de redes no realistas o mucha dependencia de información suplementaria no disponible para todos los modelos biológicos [Margolin *et al.*, 2006a]. Este último aspecto está fuertemente relacionado con la complejidad de los modelos empleados para la búsqueda de dichas redes. Los modelos biológicos de mayor complejidad presentan especialmente dos problemas mayores: 1) la gran cantidad de variables con mucho ruido que resulta en la necesidad de poder computacional mayor y 2) la no-linealidad de las dependencias estadísticas que tiene como consecuencia que muchos de los métodos lineales carezcan de la capacidad de dilucidar interacciones entre elementos de la red [Hernández-Lemus y Rangel-Escareño, 2011].

1.2.2. Teoría de la información

Por lo mencionado en la sección anterior, se puede comprender por qué la inferencia, a partir de tecnologías genómicas, de redes de regulación genética, como la del modelo de cáncer de mama, presenta la problemática de lidiar con sistemas con miles de variables, altos niveles de ruido y un conjunto muy reducido de muestreo y, sobre todo, con el carácter fuertemente no-

lineal que subyace la dinámica bioquímica. Es en este contexto en el cual el papel de los métodos basados en la teoría de la información ha tenido éxito sobre otros enfoques [Basso *et al.*, 2005; Margolin *et al.*, 2006a; Hernández-Lemus y Rangel-Escareño, 2011]. El enfoque basado en la teoría de la información usa generalizaciones de los coeficientes de correlación pareada, llamada Información Mutua (MI), en un análisis de agrupamiento para comparar perfiles de expresión para un conjunto de microarreglos. Para cada par de genes i y j se calcula su información mutua (MI_{ij}) y considera a ésta como la relación entre estos dos genes (a_{ij}). Calculada de esta forma, la dirección de la relación resultante no puede ser calculada ($a_{ij} = a_{ji}$). De esta forma se puede usar MI para medir el grado de independencia entre dos genes [Bansal *et al.*, 2007; Hernández-Lemus y Rangel-Escareño, 2011].

La información mutua MI_{ij} entre el gen i y el gen j se puede calcular como:

$$MI_{ij} = H_i + H_j - H_{ij} \tag{1-1}$$

Donde H es la Entropía en el contexto de la teoría de la información. Para definir ésta vamos a suponer que X y Y son un par de variables aleatorias que tienen las siguientes características:

- Un alfabeto finito $\mathcal{X}$ y $\mathcal{Y}$ respectivamente
- Distribución de probabilidad conjunta de $p(X, Y)$
- Distribución de probabilidades marginales $p(X)$ y $p(Y)$

Dejemos también que $\hat{X}$ y $\hat{Y}$ expresan dos variables aleatorias discretas definidas en $\mathcal{X}$ y $\mathcal{Y}$ respectivamente. La distribución de probabilidades de la asociación sería $\hat{p}(X)$ y $\hat{p}(Y)$, su distribución de probabilidad conjunta sería $\hat{p}(X, Y)$ y estaría definida en $\mathcal{J}$, el espacio de muestreo de la probabilidad conjunta sería $\mathcal{J} = \mathcal{X} \times \mathcal{Y}$. Para casos particulares tendríamos $p(x) = P(X = X)$ y $\hat{p}(y) = P(\hat{Y} = y)$. De esta forma, para cada distribución de probabilidad discreta de X de un gen i, es posible definir la entropía de información H de esa distribución de la siguiente forma:

$$H_i = -K_s \sum_v p_v(X) log p_v(X) \tag{1-2}$$

aquí H_i es la entropía de Shannon-Waver [1949], K_s es una constante útil para determinar las unidades en las cuales la entropía es medida y $p_v(X)$ es la densidad de probabilidad del estado v de la variable aleatoria dada por $X = x$.

La entropía H_i tiene muchas propiedades interesantes. Específicamente ésta alcanza un máximo para variables distribuidas uniformemente, esto es, entre más alta la entropía, más azarosamente se encuentran distribuidos los niveles de expresión de los genes en el experimento. De la definición se puede ver que la MI se convierte en cero si ambas variables x_i y x_j son estadísticamente independientes ($P(x_i x_j) = P(x_i)P(x_j)$) pues su entropía conjunta es $H_{ij} = H_i + H_j$. Una alta MI indica que dos genes están estadísticamente asociados uno con el otro. Ya que la MI es simétrica entre ambos genes, se obtienen relaciones sin dirección [Bansal *et al.*, 2007; Hernández-Lemus y Rangel-Escareño, 2011].

La teoría de la información se ha convertido en una herramienta teórica fundamental para desarrollar algoritmos y técnicas computacionales para enfrentar el problema de la selección de características y de la deconvolución de redes de regulación genética aplicadas a datos reales. Sin embargo, hay metas y retos que involucran la aplicación de la teoría de la información al análisis genómico. Los algoritmos aplicados deben devolver modelos inteligibles, tienen que depender de muy poco conocimiento previo, lidiar con miles de variables, detectar dependencias no-lineales y todo esto a partir de decenas (o al menos pocos cientos) de muestras ruidosas [Hernández-Lemus *et al.*, 2009].

1.3. Factores de transcripción y reguladores transcripcionales maestros

1.3.1. Factores de transcripción

El fenotipo de una célula está determinado por la actividad de cientos de genes y sus productos [Basso *et al.*, 2005]. Esta actividad está coordinada por intrincadas relaciones de regulación transcripcionales y la existencia de una o varias vías de señalización. Actualmente, el modelo de

regulación transcripcional en Eucariontes más aceptado dice que, usualmente, unas proteínas regulatorias, llamadas Factores de Transcripción (FT), que se caracterizan por tener una secuencia y estructura que se une al DNA, actúan uniéndose a una secuencia específica en los módulos reguladores o *enhancers* de sus genes blanco [Lelli *et al.*, 2012]. Los FT pueden actuar de dos formas opuestas: pueden activar o inhibir la actividad transcripcional de sus blancos. Si bien las redes genéticas obtenidas con la deconvolución basada en teoría de la información son no dirigidas, dadas estas características intrínsecas de los FT, es aceptable pensar que su interacción en la red puede tener una dirección de éstos a sus genes blanco. Son componentes celulares clave que controlan la expresión genética: su actividad puede determinar cómo actúa y responde la célula al ambiente [Vaquerizas *et al.*, 2009].

En un principio se estimó, basados en una exploración inicial de todo el genoma humano [Lander *et al.*, 2001; Venter *et al.*, 2001], que la maquinaria transcripcional pudiera estar controlada por entre 200 y 300 genes y que estos pudieran tener entre 2,000 y 3,000 sitios de unión específico para factores de transcripción. Vaquerizas *et al.* [2009] mencionan que, en la base de datos Gene Ontology, se definen 1,052 factores de transcripción y que solo el 6 % (62 casos) de ellos fueron corroborados experimentalmente. Seis años después, la misma base de datos reconoce 1,846 FT de los cuales 14 % (260) de ellos cuentan con evidencia experimental. Este puede ser un indicativo del rápido progreso de la documentación de los mecanismos de transcripción, pero también de la apabullante complejidad del control genético.

1.3.2. Reguladores transcripcionales maestros

Un Regulador Transcripcional Maestro es un Factor de Transcripción que se expresa en el comienzo del desarrollo de un fenotipo o tipo celular, participa en las especificaciones de dicho fenotipo regulando múltiples genes que están corriente abajo, por interacción directa o a través de cascadas génicas [Sun-Kin Chan y Kyba, 2013].

Los Reguladores Transcripcionales Maestros son responsables del control de todo el programa de regulación genética que define un tipo celular [Han *et al.*, 2004; Basso *et al.*, 2005; Affara

et al., 2013]. Los RTM pueden actuar sobre procesos generales de la células [Hosking, 2012], pero también en fenotipos celulares específicos [Hinnebusch y Natarajan, 2002; Medvedovic *et al.*, 2011; Affara *et al.*, 2013]. Entender esta organización y encontrar estos genes reguladores es crucial para elucidar tanto la fisiología celular normal como los fenotipos patológicos [Basso *et al.*, 2005] como el del cáncer de mama.

El gen mTOR es un ejemplo clásico de RTM. Se sabe que es importante en la regulación del control de crecimiento y longevidad. Un mal funcionamiento en el complejo que acompaña a mTOR (mTORC1 o mTORC2 compuestos por cinco o seis moléculas respectivamente incluyendo a mTOR mismo) se ha asociado con anormalidades durante el desarrollo, enfermedades autoinmunes y cáncer [Hosking, 2012]. Parece que el papel principal de mTOR es la regulación de la síntesis de proteínas. Debido a la gran cantidad de interacciones de señalización que tiene mTOR, este factor de transcripción actúa como un regulador maestro en varias condiciones fenotípicas.

Otro buen ejemplo de regulador transcripcional maestro es Smad3 en el factor de crecimiento transformante beta (TGF-β). Smad3, en combinación con otros factores de transcripción (Figura 1-4) determina la diferenciación celular hacia células Pro-B, miotubos y células madre embrionarias [Mullen *et al.*, 2011].

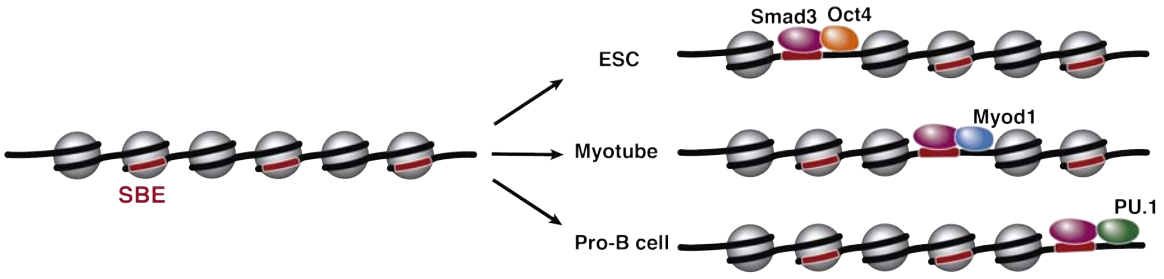


Figura 1-4: **Regulador Transcripcional Maestro Smad3**. Mullen *et al.* [2011] estudiaron la señalización del factor de crecimiento transformante beta (TGF-β) mediada por el factor de transcripción Smad3. Smad3 se pega al DNA acompañando a otros reguladores transcripcionales los cuales dirigen distintas respuestas hacia tipos celulares específicos. Smad3 junto con Oct4 dan origen a células madres embrionarias (ESCs), junto con Myod1 dan origen a miotubos, y con PU.1 a células Pro-B. Las cajas rojas indican los elementos de union a Smad (SBE) y las esferas grises representan nucleosomas. (Modificada de Mullen *et al.* [2011])

Como último ejemplo, podemos mencionar a los reguladores transcripcionales maestros TCF7 y RUNX1 que, en conjunto, regulan una red de factores de transcripción definiendo el estado de diferenciación de las células madre hematopoyéticas (Figura 1-5). Son estos dos genes la punta de una serie de cascadas génicas que determina la diferenciación actuando sinérgicamente [Wu *et al.*, 2012].

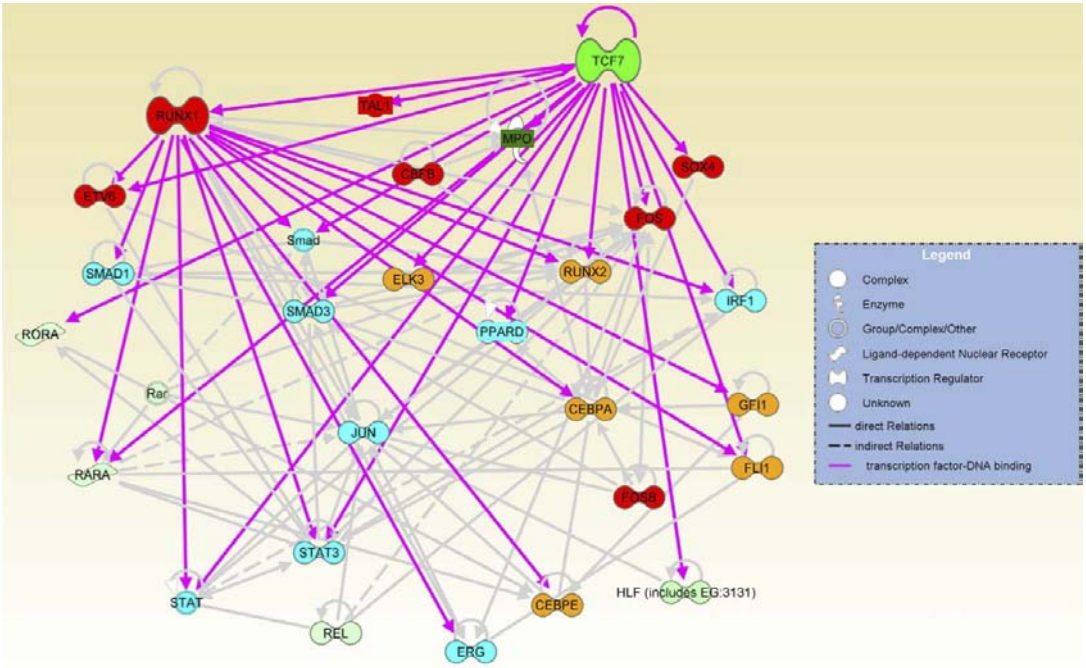


Figura 1-5: **Reguladores Transcripcionales Maestros TCF7 y RUNX1**. La red referente al establecimiento y desarrollo de las células madre hematopoyéticas (nodos rojos), control del crecimiento celular (nodos azules) y multipotencialidad (nodos naranja) fueron identificados entre los genes sobre regulados en células CD34+ (2 veces el valor del control) y mostrado por el software de *Ingenuity Pathway Analysis* (IPA). Las lineas grises corresponden a las relaciones registradas basadas en la literatura de IPA. Las lineas rosas indican los blancos de unión entre TCF7 o RUNX1 identificadas con experimentos de ChIP-Seq. Los tonos de color verde de los nodos de la red indican el nivel de regulación en las células CD34+. Sox4, Mpo, Tal1 y Ppard blancos de TCF7 que fueron anidados manualmente a la red por sus interesantes funciones en la hematopoiesis y la auto-renovación (fuente Wu *et al.* [2012]).

Lo anterior nos permite vislumbrar la importancia y las potencialidades de identificar cuáles son los Reguladores Transcripcionales Maestros tanto en los fenotipos sanos como, en los de las enfermedades como el cáncer de mama. Tanto para la comprensión de la regulación de la transcripción en Eucariontes como para el entendimiento y, posiblemente, tratamiento de dichas enfermedades.

En el mundo, el cáncer de mama es un grave problema de salud pública. En el 2012, a nivel mundial, se estimó que es el tipo de cáncer en mujeres con el mayor número de nuevos casos (1,676,600 casos, el 25 % de todos los casos de cáncer) y de muertes (521,900 de muertes, el 15 % de las muertes de cáncer) [American Cancer Society, 2015]. Estas cifras se agravan en los países pobres. No cabe duda que los estudios dirigidos a la compresión de los mecanismos con que actúa son fundamentales.

1.4. Objetivo

Hasta ahora, el grueso de la información acerca del cáncer se sustenta en investigaciones que se centran en estudiar moléculas individuales o en pequeños conjuntos de éstas. Puede ser muy útil abordar este tema con una perspectiva que no considere molécula por molécula sino que trate de obtener información de la medición de grandes conjuntos de moléculas y su interacción. En años recientes han surgido avances que nos dan mayor posibilidad de hacer esto: las tecnologías ómicas de alto rendimiento y el aumento de poder de cómputo con el cual ejecutar algoritmos que nos permitan enfocar la atención sobre los sistemas que subyacen detrás del ruido de los datos crudos.

Esta tesis trata de contribuir con este esfuerzo basándose en algunos hitos ampliamente aceptados de los mecanismos moleculares. Concretamente de la actividad regulatoria de los Factores de Transcripción a sus blancos, la aparente organización jerárquica de las redes moleculares y del supuesto de que el fenotipo enfermo debería ser, en buena parte, resultado de la actividad de los genes diferencialmente expresados (firma molecular).

Explícitamente, el objetivo de este trabajo es inferir sistemáticamente reguladores transcripcionales maestros usando un conjunto relativamente grande de datos de expresión de mRNA.

Para ello se tuvo que: 1) Obtener un conjunto de datos de expresión de genoma completo suficientemente grande para asegurar poder estadístico y, al mismo tiempo, homogéneo para hacerlos comparables. Este conjunto de datos debe contener un número importante de microarreglos de tejido sano para hacer contrastes. 2) Encontrar formas de preprocesar los datos de tal manera que los experimentos de diferente origen fuera comparables. Esto es, que la variación fuera, en parte importante al menos, de origen biológico y no artefactos de los experimentos. 3) Inferir una firma molecular a partir de la comparación entre casos y controles (microarreglos de tejido enfermo contra microarreglos de tejido sano). 4) Inferir redes de factores de transcripción que nos permitan explorar en su topología los alcances de los FT en la firma molecular y así reconocer su posible papel como Reguladores Transcripcionales Maestros. 5) Estimar y comparar el alcance de los FT sobre la firma molecular.

Capítulo 2

Materiales y métodos

Como ya mencionamos, el estudio de sistemas complejos como las redes de regulación genética requiere de una combinación de enfoques y técnicas de análisis para su estudio [Hernández-Lemus, 2013]. Este trabajo procura explorar los datos para obtener información del sistema estudiado (*data-driven*) haciendo uso de las herramientas matemáticas y computacionales que se adecuen al sistema. El flujo de trabajo general de este proyecto puede verse en la figura 2-1. Básicamente se puede resumir en:

1. Obtención de datos.
2. Eliminar, en lo posible, artefactos (Preproceso)
3. Inferencia de la Red de Regulación de Transcripcional
4. Inferencia de Reguladores Transcripcionales Maestros
    a) Inferencia de la dirección de la regulación de los FT
    b) Cálculo de la Firma Molecular
    c) Generación del Modelo nulo
    d) Inferencia de Reguladores Maestros
5. Análisis de los resultados con Redes Causales

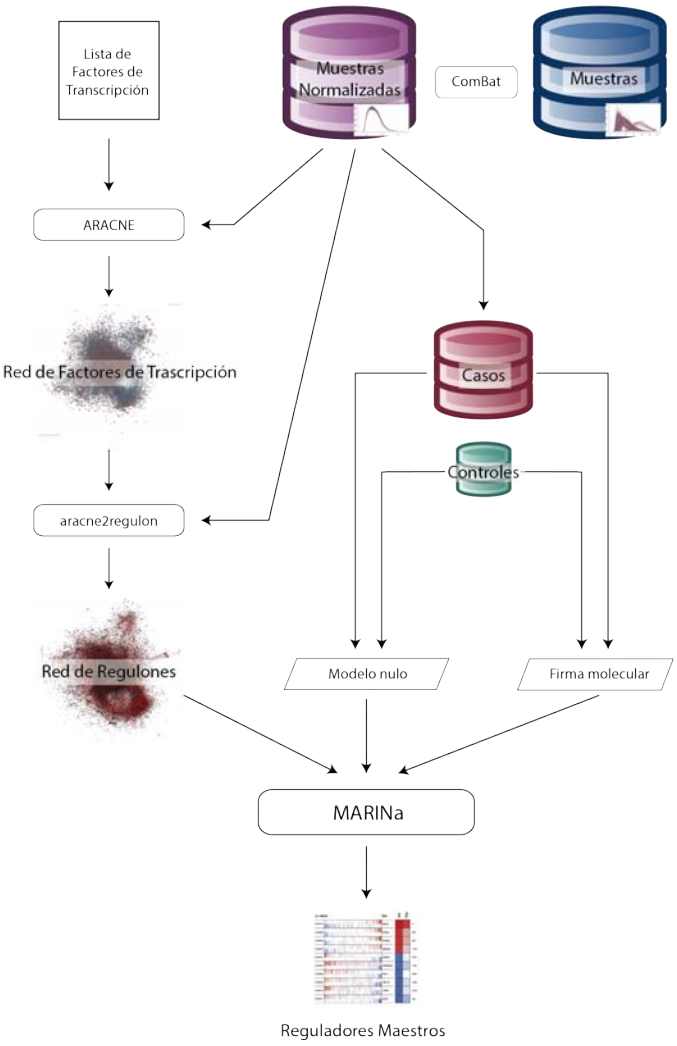


Figura 2-1: **Flujo de trabajo.** Este diagrama muestra los pasos a seguir para inferir Reguladores Transcripcionales Maestros con MARINa (*Master Regulator Inference Algorithm*) [Lefebvre *et al.*, 2010]. Después de obtener las muestras se lleva a cabo el preproceso de éstas con el fin de eliminar efectos de lote y hacer comparables los experimentos. La matriz de expresión resultado de este paso es la entrada principal de ARACNe (*Algorithm for the Reconstruction of Accurate Cellular Networks*) [Margolin *et al.*, 2006a] que, dada la lista de Factores de Transcripción humanos, infiere la relación de éstos con todos los genes del microarreglo. A continuación se determinó la dirección de la regulación transcripcional calculando la correlación de Spearman de los FT con cada uno de sus blancos. Por otro lado se separan las muestras en casos y controles y se estima cuán diferencialmente expresados están cada uno de los genes para generar la Firma Molecular. Además, a partir de esta firma, se genera un modelo nulo permutándola 1,000 veces con la finalidad de estimar valores de confianza (ver sección 2.5). Estos tres elementos, el conjunto de todos los regulones, la Firma Molecular y el modelo nulo, son las entradas de MARINa. Éste elige aquellos FT en la Firma Molecular con más de N blancos (valor por defecto: $N = 20$) y, para cada uno de ellos ejecuta un análisis de enriquecimiento con la finalidad de encontrar cuál de ellos tiene más blancos en su regulón que contenga más genes de la Firma Molecular.

2.1. Repositorio en línea

Con la finalidad de hacer reproducible y colaborativo este trabajo, se ha publicado el código generado (que también se puede consultar en el Apéndice A) a disposición de quien quiera consultarlo en el la dirección https://github.com/CSB-IG/tmr_search en *Github* que es una plataforma de desarrollo colaborativo en la internet.

2.2. Obtención de datos

2.2.1. Lista de factores de transcripción humanos

La lista definitiva de los Factores de Transcripción Humanos aún está por escribirse [Vaquerizas *et al.*, 2009]. Su contenido dependerá mucho de la definición de Factor de Transcripción y de la exploración experimental de las proteínas candidatas para serlo, como se menciona en la introducción. Se precisaba tener una lista fiable de factores de transcripción que sirviera de base para analizar el control de la transcripción en nuestro conjunto de datos. Se analizó el conjunto de datos con cuatro listas de FT: La lista sugerida en el trabajo de Vaquerizas *et al.* [2009], la lista disponible de manera abierta en http://www.bioguo.org/AnimalTFDB/, la lista ofrecida por Shimoni y Alvarez [2013] disponible en linea en http://dx.doi.org/10.6084/m9.figshare.871524 y una lista generada a partir del archivo de anotación del microarreglo de Affymetrix Human Genome U133A 2.0 Array (HGU133A) buscando en las columnas correspondientes a "Gene Ontology / Molecular Function" y "InterPro" por el término *transcription factor* (Disponible en https://github.com/CSB-IG/tmr_search/tree/master/TF_lists). Dado que los resultados para las cuatro listas eran muy similares, la última lista generada en este proyecto es de la que presentamos resultados.

2.2.2. Microarreglos de expresión

Para este trabajo se obtuvo un conjunto de datos de 880 microarreglos con perfiles de expresión de mRNA de diez distintos experimentos (Tabla 2-1) los cuales están disponibles en el sitio de *Gene Expression Omnibus* (http://www.ncbi.nlm.nih.gov/geo/) . Todos los

experimentos fueron hechos usando extracciones de mRNA total bajo el protocolo GPL96 el cual se basa en la plataforma de microarreglos de Affymetrix HGU133A que consiste de 12,500 genes (que incluye solo aquellos genes mejor caracterizados del genoma humano). 819 muestras de los 880 totales corresponden a tejido de cáncer de mama primario no tratado, mientras que las otras 61 muestras corresponden a tejido mamario sano (ver Tabla 2-1). La lista completa de los 880 archivos CEL descargados de GEO y usados en este trabajo se encuentra en https://github.com/CSB-IG/tmr_search/blob/master/CEL_files_list.txt.

Tabla 2-1: Identificadores GEO y su referencias a los experimentos de microarreglos usados en este trabajo. La primera columna contiene la clave de identificación del GEO, la segunda y la tercera columna corresponden al número de muestras casos/controles respectivamente, la cuarta columna incluye una breve descripción de las muestras y la quinta la referencia asociada. (La lista de los 880 archivos CEL puede ser consultada en GitHub).

GEO Series ID	Tumores	Controles	Descripción	Referencia
GSE1456	159		Pacientes de cáncer de mama que recibieron cirugía	Pawitan *et al.* [2005]
GSE1561	49		Se tomaron biopsias de pacientes cáncer de mama localizado o inflamatorio.	Farmer *et al.* [2005]
GSE2603	99		Tejidos de cáncer de mama primarios que se obtuvieron de procedimiento terapéutico.	Minn *et al.* [2005]
GSE2990	61		Experimentos de microarreglos de tumores primarios de mama.	Sotiriou *et al.* [2006]
GSE3494	4		Tejidos de cáncer de mama congelados en fresco.	Miller *et al.* [2005]
GSE4922	249		Tumores de mama primarios invasivos.	Ivshina *et al.* [2006]
GSE7390	198		Experimentos de microarreglos de tumores primarios de mama.	Desmedt *et al.* [2007]
GSE6883		3	Las muestras se procesaron dentro de la hora después de la cirugía de reducción de senos.	Liu *et al.* [2007]
GSE9574		15	Las muestras se obtuvieron de pacientes de mamoplastia de reducción sin cáncer de mama aparente	Tripathi *et al.* [2008]
GSE15852		43	Tejido normal pareado	Pau Ni *et al.* [2010]
Total	819	61		

Es importante mencionar que, para tener resultados sólidos, es necesario un número importante de muestras. La mayoría de los tejidos sanos proviene de tejido adyacente sin rastros de la enfermedad de pacientes de cáncer de mama, pero también se incluyeron muestras no pareadas, con el fin de tener un mayor número de muestras y para hacer más robusta nuestra estadística.

2.2.3. Conjunto de datos del Atlas Genómico del Cáncer

Como ejercicio de comparación, se llevaron a cabo casi todos estos pasos con un segundo conjunto de datos que fueron obtenidos de la base de datos *The Cancer Genome Atlas* (Atlas Genómico del Cáncer, TCGA, http://cancergenome.nih.gov/). Se obtuvieron 597 muestras de mRNA de los cuales, 534 correspondían a muestras de tejido de cáncer de mama invasivo y otras 63 eran no tumorales. Todos los datos colectados de esta base de datos corresponden al nivel 3, es decir que ya se encontraban preprocesados y normalizados.

2.3. Preproceso

El efecto de lote, esto es, variación técnica añadida durante el procesamiento de las muestras, es uno de los factores de confusión más comunes durante el análisis de datos de microarreglos [Grass, 2009]. Chen *et al.* [2011] probaron seis algoritmos distintos para el control del efecto de lote y encontraron que los mejores resultados se obtuvieron usando el método empírico bayesiano conocido como `ComBat` (Combating Batch Effects When Combining Batches of Gene Expression Microarray Data) [Johnson *et al.*, 2007]. Sin embargo, dado que siete de los diez conjuntos de datos corresponden a tejido tumoral exclusivamente (esto es, que no contienen muestras control), y los tres restantes conjuntos de datos solo tienen muestras de tejido sano, no hay intersección entre estos conjuntos de datos. De acuerdo con Leek *et al.* [2010], los tratamientos y los lotes están completamente confundidos. Puesto que no hay un método para estimar el efecto de lote bajo estas condiciones [Leek *et al.*, 2010], `ComBat` no puede llevar a cabo la normalización de todo el conjunto de datos. Considerando que `ComBat` no elimina el efecto de lote con las condiciones de nuestro conjunto de datos, se decidió resolver este problema de la siguiente forma: Después de procesar los arreglos con el paquete de Bioconductor `frma` McCall *et al.* [2010], y usando sumarización con promedio ponderado robusta (*robust weighted average*) sin corrección de fondo. Se dividió el conjunto de datos en casos y controles, entonces se aplicó el algoritmo de `ComBat` a los dos conjuntos por separado. Hecho esto, se reunieron los dos conjuntos resultantes y se renormalizaron con el algoritmo de *Regresión Local Cíclica* [Bolstad *et al.*, 2003; Ballman *et al.*, 2004], de tal forma que ambos conjuntos estuvieran ahora en el *mismo rango dinámico*.

Se necesitaba tener una medida del efecto de lote entre las muestras, para tener una estimación de cuánto efecto de lote se está eliminando. Para este fin usamos el el Análisis de Componentes Principales de la Varianza (*Principal Variance Component Analysis*, PVCA). PVCA es un algoritmo que conjunta las ventajas del Análisis de Componentes Principales (reducción de la dimensionalidad) así como del Análisis de Componentes de la Varianza ajustando un modelo mixto usando los factores de interés así como el efecto aleatorio para estimar y separar la variabilidad total [Grass, 2009]. Esta herramienta está disponible en http://www.niehs.nih.gov/research/resources/software/biostatistics/pvca/ (pero también disponible en un paquete para R en Bioconductor).

Una vez que el efecto de lote se redujo por separado, un análisis de PVCA corroboró que dicho efecto casi desapareció y el efecto del tratamiento aún estaba presente (Figura 2-2). Debido a las condiciones del diseño experimental, no fue posible eliminar el efecto de lote permanentemente. Como el efecto de lote mezclado con el diseño experimental es un tema importante de discusión en la investigación en genómica computacional, se puede visualizar un escenario en el cual el presente trabajo pueda ser revisado así como algunas de sus conclusiones reinterpretadas. Mientras tanto, el método descrito anteriormente destinado a reducir el efecto de lote se puede considerar una primera aproximación para los propósitos de esta tesis.

Dada la complejidad de la regulación transcripcional [Lim *et al.*, 2009], la búsqueda de Reguladores Transcripcionales Maestros parece extremadamente sensible al efecto de lote. Puesto que MARINa detecta detalles finos de la expresión transcripcional, incluso las más pequeñas modificaciones pueden acarrear cambios importantes en el resultado final. Después de preprocesadas las 880 muestras de esta forma se está en condiciones de trabajar con una mínima duda razonable de que el efecto de lote no afectará de forma fundamental en los resultados del análisis principal. Estos análisis son la inferencia de redes de regulación genética y la inferencia de Reguladores Transcripcionales Maestros. El código utilizado en el preproceso de muestras puede ser consultado en el Apéndice A.1 (Github: https://github.com/CSB-IG/tmr_search/blob/master/Normalization_precollapsing.R).

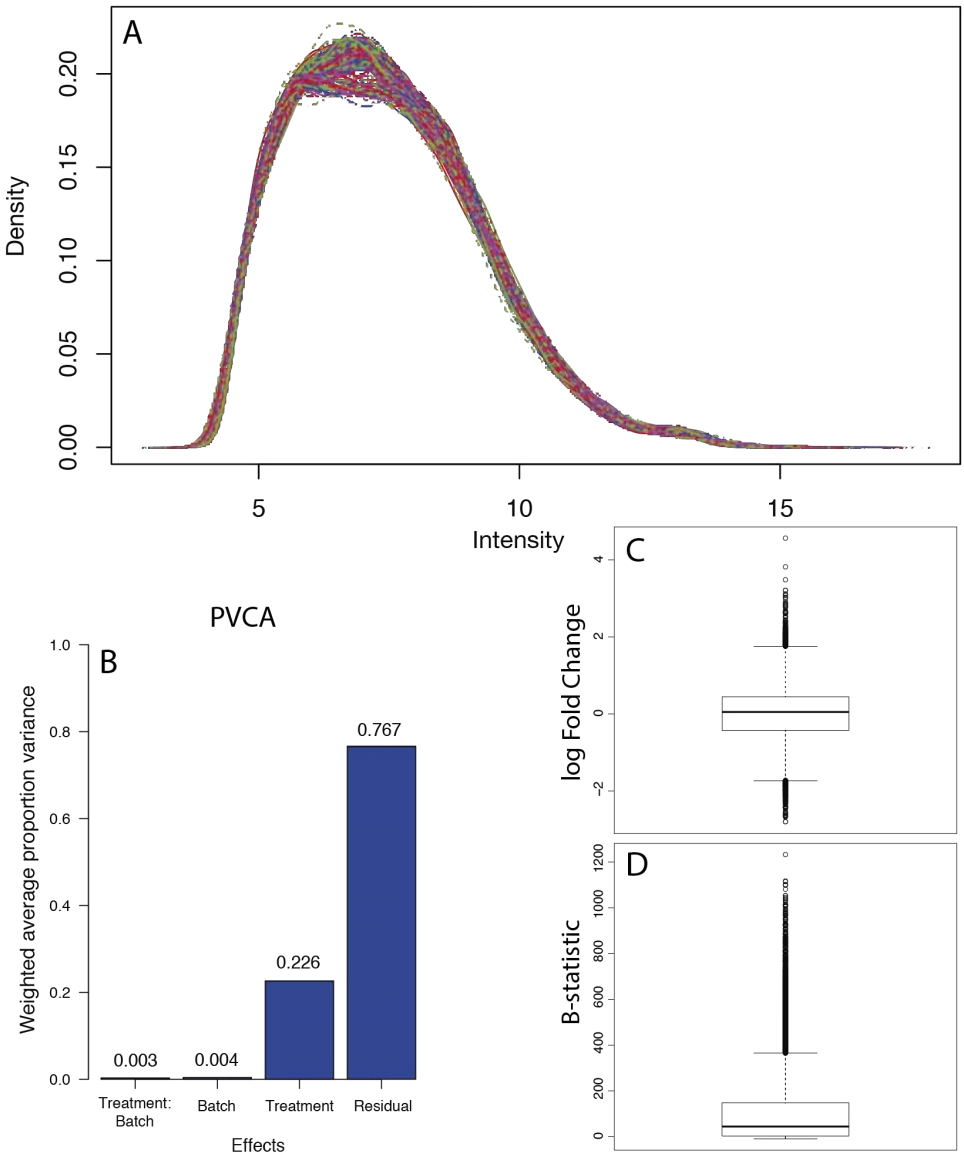


Figura 2-2: **Normalización por tratamiento**. A: Gráfico de densidad asociado con la matriz de expresión usada para este trabajo después de aplicar a casos y controles por separado y luego renormalizar ambos conjuntos con la normalización de Regresión Local Cíclica. B: Resultado del PVCA. Aquí se puede observar un mínimo efecto de lote así como una importante contribución de varianza debido al tratamiento. C: log fold-change y D: estadística B con comportamiento aproximado al esperados en conjuntos de datos de este tipo.

2.3.1. Preproceso del conjunto de datos de TCGA

Como ya se mencionó, los datos obtenidos del TCGA son del nivel 3, esto es, que ya pasaron por un preproceso que normaliza y unifica los datos. Este preproceso está determinado por la plataforma de donde fueron obtenidos los niveles de expresión. Estas son Agilent 244K

y Affymetrix HT-HG-U133a básicamente [Cancer Genome Atlas Research Network, 2008]. En particular, los datos provenientes de los arreglos de Agilent fueron normalizados usando el algoritmo de regresiones lineales ponderadas localmente (*locally weighted linear regression* o lowess [Cleveland, 1979]) y la relación entre el canal de Cy5 (rojo, de muestra) y el canal de Cy3 (verde, de referencia) se transformó usando log_2 para crear los niveles de expresión de 18.624 genes [Cancer Genome Atlas Research Network, 2008]. Por otro lado los microarreglos de Affymetrix que pasaban el control de calidad fueron normalizados con el algoritmo de Promedio de multiarreglos robusto (*Robust Multiarray Average*, RMA [Irizarry *et al.*, 2003]) en combinación con el de affymetrix.aroma para obtener la matriz de expresión de de 12,042 genes [Cancer Genome Atlas Research Network, 2008].

A continuación, ambos conjuntos de datos fueron unificados en un solo conjunto de datos de expresión [Cancer Genome Atlas Research Network, 2008]. Cada conjunto de datos de expresión de Affymetrix fue transformado logarítmicamente y se restó la media. Como la plataforma Agilent genera datos de tasa logarítmica, solamente se restó la media. Las matrices de datos resultantes se fusionaron con el valor de la mediana. Siguiendo este método, se generó una matriz de expresión con 19,692 genes [Cancer Genome Atlas Research Network, 2008].

2.4. Inferencia de redes de regulación transcripcional

Las herramientas computacionales necesarias para hacer biología de sistemas pueden ser clasificadas, a grandes rasgos, en identificación de sistemas y de análisis de comportamiento [Kitano, 2001]. En biología molecular, la deconvolución de sistemas se encarga de identificar las relaciones de regulación entre genes, proteínas y otras moléculas, así como sus dinámicas inherentes. Podría argumentarse que la deconvolución de sistemas es uno de los problemas más complicados de la ciencia. Aunque el análisis de esta dinámica se lleva a cabo únicamente con modelos, la construcción de estos modelos es un proceso estrechamente conectado a la realidad, parte de un proceso iterativo entre el análisis de datos, simulación y validación experimental [Kriete y Eils, 2014].

2.4.1. ARACNe

ARACNe (*Algorithm for the Reconstruction of Accurate Cellular Networks*) es un algoritmo usado para identificar interacción entre genes, calculando información mutua a partir de datos derivados de microarreglos expresión de mRNA [Basso *et al.*, 2005; Margolin *et al.*, 2006b]. ARACNe calcula la información mutua (MI ecuación 1-1) para cada par de genes i y j (MI_{ij}) a partir de una matriz de niveles de expresión de genes (renglones) en un conjunto de muestras (columnas).

Para hacer más eficiente computacionalmente la estimación de la MI, ARACNe usa un estimador del núcleo Gaussiano [Steuer *et al.*, 2002] en datos transformados en cópulas [Margolin *et al.*, 2006a]. Los núcleos Gaussianos son una forma no paramétrica de estimar la función densidad de probabilidad entre puntos en variables aleatorias. Si dos puntos en el espacio de datos son vecinos cercanos, entonces el ángulo entre los vectores que los representan en el espacio del núcleo será pequeño. Si los puntos están lejos, entonces el vector correspondiente estará cerca a la "perpendicular". La estimación de densidad de núcleos es un problema importante de suavizado de datos cuando se hacen inferencias sobre poblaciones basándose en un número finito de muestras.

Dadas dos variables, X y Y, sus marginales y la unión de sus densidades de probabilidad pueden ser estimadas usando núcleos Gaussianos:

$$\hat{f}(X) = \frac{1}{N}\frac{1}{\sqrt{2\pi h}}\sum_i exp\left(\frac{(x-x_i)^2}{2h^2}\right) \tag{2-1}$$

y

$$\hat{f}(X,Y) = \frac{1}{N}\frac{1}{\sqrt{2\pi h}}\sum_i exp\left(\frac{(x-x_i)^2+(y-y_i)^2}{2h^2}\right) \tag{2-2}$$

Donde N es el tamaño de muestra y h la anchura del núcleo. La MI se puede computar como:

$$\hat{MI}(X,Y) = \frac{1}{N}\sum_i log\frac{\hat{f}(x_i,y_i)}{\hat{f}(x_i)\hat{f}(y_i)} \tag{2-3}$$

Para una h espaciada uniformemente, el sesgo del estimador del núcleo Gaussiano de MI se pierde asintóticamente cuando $M \to \infty$, en tanto que $h(M) \to 0$ y $[h(M)]^2 M \to \infty$. Sin embargo, para M finita el sesgo depende fuertemente de $h(M)$ y la elección correcta de un núcleo Gaussiano no es universal.

Una vez que es calculada MI_{ij} para todos los pares de genes, ARACNe excluye a todos lo pares de genes en los cuales la hipótesis nula de independencia mutua de genes no puede ser desechada ($Ho : MI_{ij} = 0$). El valor de P de la hipótesis nula, calculada usando simulaciones Montecarlo, es asociado a cada valor de información mutua.

En este trabajo se usó ARACNe en linea de comandos compilado a partir de la fuente en C++ disponible en http://wiki.c2b2.columbia.edu/califanolab/index.php/Software/ARACNE. ARACNe recibe la matriz de expresión en formato de texto separado por tabuladores. Para este trabajo únicamente se calculó la interacción de todos los factores de transcripción (en https://github.com/CSB-IG/tmr_search/tree/master/TF_lists) con el resto de los genes. De tal forma que se obtuvo una red de todos los factores de transcripción y sus blancos. Este paquete, lo que devuelve de salida es un archivo adj que es una lista de valores de adyacencia que representan una matriz de adyacencia en la cual solo son representadas las interacciones con MI mayores al umbral configurado. Este procedimiento fue idéntico tanto para los datos de GEO como de TCGA.

Paralelización

Ya que todos estos cálculos requieren un cantidad de tiempo de ejecución considerable, vale la pena mencionar las implementaciones que se llevaron a cabo para optimizar el tiempo de cómputo. Dado que las estimaciones que hace ARACNe se ejecutan gen por gen, los cálculos por gen pueden ser ejecutados de forma independiente y paralela en procesadores por separado y no uno a uno en fila esperando el uso de un procesador como sucede con el

software tal como éste se distribuye. Ya que contábamos con una computadora de 32 procesadores y 378 GB de RAM, era muy atractivo poder usar todos estos recursos de forma simultánea. HTCondor (*High Throughput Computing*) es un sistema de gestión de carga de trabajo especializada para tareas de cálculo intensivo que se distribuye libre y gratuitamente en http://research.cs.wisc.edu/htcondor/.

HTCondor provee un sistema de filas de trabajos, políticas de programación de tareas, esquema de prioridades, monitoreo y manejo de recursos, de tal forma que fue un candidato excelente para que pudiera administrar y monitorear el trabajo de ARANCE gen por gen. Dado que una salida generada con un límite de tolerancia de $P = 1$ contiene todas las interacciones posibles, se generó la red de FT con una tolerancia de este valor y, posteriormente, se eliminaron aquellas interacciones por debajo del umbral deseado. En específico y con la intención de explorar la estructura de la red regulatoria de FT, se generaron redes con un umbral de P de 1×10^{-30}, 1×10^{-40}, 1×10^{-50} y 1×10^{-100} para ambos conjuntos de datos (GEO y TCGA).

El código correspondiente a la paralelización con HTCondor puede ser consultado en el Apéndice A.2 (Github: https://github.com/CSB-IG/tmr_search/tree/master/parallel_aracne) y requiere tener ya instalado HTCondor.

2.5. Algoritmo de inferencia de reguladores maestros

La búsqueda sistemática de Reguladores Transcripcionales Maestros actualmente está implementada sólo con un par de métodos. Uno es a través de MARINa [Lefebvre *et al.*, 2010] y otro es Biobase ExPlain® (Qiagen) [Kel *et al.*, 2010]. Usando Transfac®, ExPlain® dice ser capaz de identificar cómo los factores de transcripción afectan la expresión genética. Esto se ha hecho para predecir cómo los FT inducen patrones de expresión tanto en microarreglos como en experimentos de RNA-seq. Su sitio de internet está en: http://www.biobase-international.com/transfac-upgrade.

Mientras que ExPlain® es un algoritmo privativo y de paga, MARINa (*Master Regulator Inference algorithm*) se encuentra libre y disponible para cualquiera. Éste es un algoritmo

diseñado para inferir aquellos factores de transcripción que controlan la transición entre dos fenotipos, A y B y el mantenimiento del segundo fenotipo. Los niveles de expresión del mRNA son un predictor pobre de la actividad regulatoria de los factores de transcripción y un peor predictor de su relevancia en la regulación de fenotipos específicos [Lefebvre *et al.*, 2010]. Para abordar este problema, MARINa infiere los factores transcripcionales maestros de la actividad transcripcional global del conjunto de regulones (esto es, de la activación o represión de todos los blancos de los FT) y de su relevancia biológica a través de la superposición de esta actividad en los programas específicos del fenotipo (Firma Molecular) [Lefebvre *et al.*, 2010] evaluada con una puntuación de enriquecimiento genético (*gene enrichment score* GES) [Subramanian *et al.*, 2005].

En este trabajo usamos el paquete `ssmarina` [Alvarez, 2013], una implementación de MARINa para R, el cual no solo ejecuta MARINa sino otros algoritmos necesarios para este análisis. Por ejemplo la inferencia del efecto regulatorio de los FT sobre sus blancos (conjunto de regulones), crea el modelo nulo, además de los análisis posteriores como el análisis de sombras y de sinergia [Lefebvre *et al.*, 2010]. El paquete `ssmarina` no está en el repositorio oficial de R pero se distribuye de forma libre y gratuita en `http://figshare.com/articles/ssmarina_R_system_package/785718`.

El código correspondiente al análisis con el paquete `ssmarina` (correspondiente a la la sección 2.4) puede consultarse en el Apéndice A.4 y en el documento en línea

2.5.1. Firma molecular

La firma molecular (FM) en este caso, son aquellos genes diferencialmente expresados entre casos y controles y se obtiene de comparar la expresión de los genes de los microarreglos de muestras de tejido sano contra la del tejido tumoral. Para esto se ejecuta una prueba t de Student comparando casos y controles renglón por renglón. Este análisis se lleva a cabo por la función optimizada para el caso llamada `rowTtest` que se encuentra en la biblioteca `ssmarina` y recibe como argumentos primero el conjunto de muestras problema y después el conjunto de muestras control. Regresa un objeto en forma de lista que contiene el estadístico t y el valor de P de la prueba estimada para dos colas. Por último, para ser consistentes con la unidades

del modelo nulo, se calcula el z-score de cada una de las comparaciones. Dado que la FM se calcula para toda la matriz de expresión normalizada, no se requiere más que generarla una vez para todo el conjunto de datos. Al final, la firma molecular es la lista de todos los genes considerados, ordenados por su nivel de expresión diferencial en ambos tratamientos.

2.5.2. Generación del conjunto de regulones

MARINa necesita, además de la Firma Molecular, una red regulatoria específica para cada tipo celular. Esta es justo la generada con ARACNe. Sin embargo esta red carece de información acerca del tipo de interacción (activación o represión) que tienen los FT sobre sus blancos. La función `aracne2regulon` del paquete `ssmarina`, toma el archivo .adj proveniente de ARACNe así como la matriz de expresión normalizada. Con ellas determina la dirección de la regulación calculando la correlación de Spearman entre los FT con más de 20 interacciones y todos sus blancos. El resultado es un objeto de R tipo regulón que contiene una lista de listas de todos los FT con más de 20 interacciones que contiene la lista de todos sus blancos y el tipo de la interacción con cada uno de ellos.

Como contamos con versiones de la red con diferente magnitud de corte, lo que tenemos que hacer es generar un conjunto de regulones para cada P. Por tanto tenemos conjuntos de regulones con valores de corte de P de 1×10^{-30}, 1×10^{-40}, 1×10^{-50} y 1×10^{-100}.

2.5.3. Modelo nulo

Dado el alto grado de co-regulación en redes transcripcionales, el supuesto de independencia estadística de la expresión genética es poco realista y potencialmente se pueden estar subestimando los valores de P. Para tener en cuenta la estructura de correlación entre los genes, se define un modelo nulo para MARINa utilizando un conjunto de firmas moleculares obtenidas permutando las muestras, tratamientos y controles, al azar. En `ssmarina` la función `ttestNull` lleva a cabo este proceso y, como salida, produce una matriz numérica de z-scores, con genes (o pruebas) en los renglones e iteraciones de permutaciones en las columnas. Este objeto puede ser ya usado como modelo nulo para el MARINa.

Al igual que con la Firma Molecular, el mismo modelo nulo generado con la matriz de

expresión normalizada puede ser usado con cualquier conjunto de regulones para calcular los Reguladores Transcripcionales Maestros con MARINa.

2.5.4. MARINa

Una vez que se tiene la firma molecular, el conjunto de regulones y el modelo nulo, MARINa busca aquellos regulones con mayor efecto en la firma molecular. Esto lo hace ejecutando múltiples análisis de enriquecimiento para calcular cuál de todos los regulones (conjunto de blancos de un factor de transcripción) está enriquecido con más genes con mejor z-score. A esta evaluación se le llama puntuación de enriquecimiento (*Enrichment Score*, ES). Se usa el modelo nulo para estimar el valor de P por permutación al comparar el z-score de cada regulón con los generados aleatoriamente. El resultado final es una lista de genes con el ES normalizado (NES) lo que permite compararlo, además se puede evaluar su efecto (positivo o negativo) sobre la firma molecular y un valor de P, que es el valor de confiabilidad contra el modelo nulo (Figura 2-3).

Una vez obtenida esta lista de candidatos, hay algunos análisis que se pueden hacer con estos los regulones mejor calificados. Por ejemplo, se puede estimar qué factor de transcripción controla los blancos de otro más colocándose "sombreándolo". O bien qué par de factores de transcripción podrían estar actuando sobre el mismo conjunto de blancos de manera conjunta ("sinérgicamente"). Estos dos análisis serán explicados en las siguientes secciones.

2.5.5. Análisis de sombras

Si los regulones R_1 y R_2 del FT_1 y FT_2 se traslapan significativamente y sólo FT_1 está enriquecido, tal vez FT_2 aparece también enriquecido por el enriquecimiento de sus blancos comunes. En este caso la actividad del FT_2 es la sombra de la del FT_1 y podemos llamar al FT_2 un regulador sombra (*Shadow Regulator*, Figura 2-4). Esto puede ser fácilmente detectado ya que el análisis de enriquecimiento de los blancos que no son comunes ($R_2 \backslash R_1 \cap R_2$) tendría que ser significativamente menor que todo el regulón R_2. La función `shadow` de `ssmarina` toma el puntaje de enriquecimiento (ES) de los blancos del FT_2 (ES_2) que no son blancos del FT_1 ($R_2 \backslash R_1 \cap R_2$). Entonces calcula el ES de 1,000 subgrupos aleatorios del FT_2 del mismo tamaño que $R_2 \backslash R_1 \cap R_2$. Si el regulón restante es mayor de 20 blancos, se computa entonces el

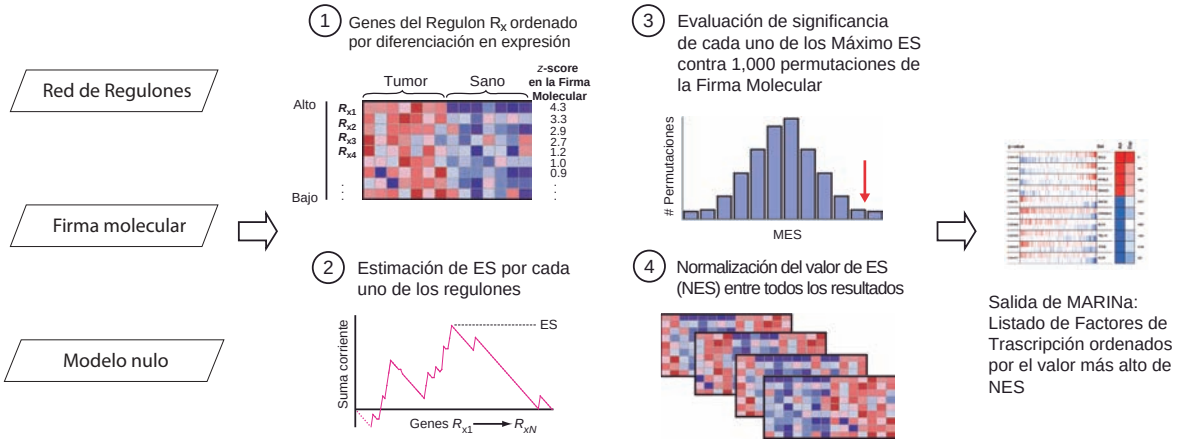


Figura 2-3: **Estimación de los Reguladores Transcripcionales Maestros por MARINa**. El algoritmo MARINa toma como entradas la Red de Regulones, la Firma Molecular y el Modelo Nulo ya descritos. ① Se ordenan los genes de cada uno de los regulones pertenecientes a un FT con más de 20 blancos. Toma los diferencialmente expresados entre casos y controles ordenados según el nivel de expresión en los casos. Por último, cuenta cuántos de éstos están presentes en la Firma Molecular. ② Se hace una suma corriente a lo largo de la cuenta del paso uno. El punto más alto es el puntaje de enriquecimiento (ES). ③ Se compara el máximo ES con los generados con permutaciones del modelo nulo para estimar su valor de significancia (P). ④ Se estandarizan los valores de ES para hacerlos comparables. El resultado es una lista de los FT con valor absoluto más alto de ES, la dirección de su efecto sobre la FM y un valor de significancia.

valor empírico de P de observar un ES menor que ES_2 si $\mathrm{ES}_2 > 0$ y mayor que ES_2 si $\mathrm{ES}_2 < 0$. Un FT_2 es sombra de FT_1 si $P < 0.01$ y si el FT_1 no es sombra del FT_2. Nótese que se prueba cada par en ambas direcciones $\mathrm{FT}_1 \rightarrow \mathrm{FT}_2$ y $\mathrm{FT}_2 \rightarrow \mathrm{FT}_1$, y se define al FT_2 como sombra del FT_1 sólo cuando el FT_1 no es sombra de FT_2. Se ignoran algunos casos ambiguos cuando no se puede probar una dirección por limitaciones de tamaño.

2.5.6. Análisis de sinergia

Alternativamente, dos FT pueden tener un efecto sinérgico en sus blancos comunes ($R_1 \cap R_2$). En este caso, el enriquecimiento de $R_1 \cap R_2$ debería ser significativamente mayor que ambos R_1 y R_2 (Figura 2-5). Se seleccionan los pares de FT que tengan una superposición significativa en sus genes blanco calculados con una prueba exacta de fisher. Los pares de regulones se ensamblan como sigue: si ambos FT se correlacionan positivamente, se define un regulón positivo o negativo intersectando los regulones positivos y negativos de los FT ($R^+_{FT_1 FT_2} = R^+_{FT_1} \cap R^+_{FT_2}$ y $R^-_{FT_1 FT_2} = R^-_{FT_1} \cap R^-_{FT_2}$). Si los FT están anti-correlacionados, se intersectan los regulones positivos de un FT con los negativos del segundo FT y vice versa ($R^+_{FT_1 FT_2} = R^+_{FT_1} \cap R^-_{FT_2}$ y

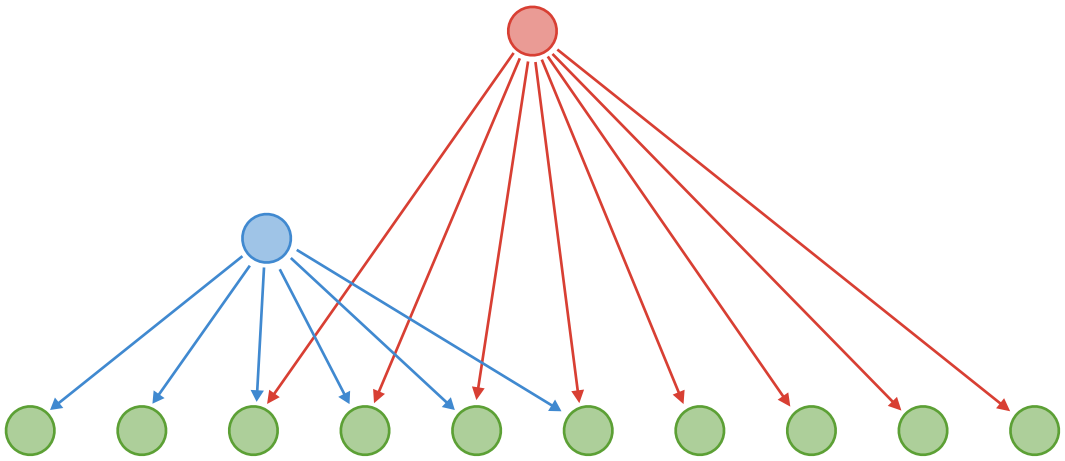


Figura 2-4: **Ensombrecimiento entre Factores Maestros de Transcripción**. Esquema de dos FT que se hacen sombra. El FT azul tiene como blancos algunos genes blanco del FT rojo. Sin embargo el FT rojo contiene estos blancos y más. Por tanto el FT azul es sombra del FT rojo.

$R^{-}_{FT_1FT_2} = R^{-}_{FT_1} \cap R^{+}_{FT_2}$ los símbolos positivo y negativo no tienen significado en los pares). Entonces se corre el análisis de enriquecimiento con los regulones de los pares FT_1/FT_2 como conjunto de genes y la unión de los regulones de FT_1 y FT_2 como referencia. Los pares de FT probados incluyen todos los pares de dos Reguladores Transcripcionales Maestros y los pares Regulador Transcripcional Maestro-Regulador Sombra de los casos ambiguos definidos en la sección anterior. Un par sinérgico se define como aquel en el que el análisis de enriquecimiento es significativo a un valor de corte de P de 0.01 y la lista final de los RTMs contiene todos los FT participantes en los pares sinérgicos.

2.6. Análisis de redes causales

El análisis de redes causales (*Causal Networks Analysis*. CNA) se llevó a cabo a través de la metodología del *Ingenuity Pathway Analysis* (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity). IPA genera redes causales basada en su base de datos altamente curada (la *Ingenuity Knowledge Base* (IKB)). Esta IKB reporta una serie de relaciones causa y efecto basadas en *observaciones experimentales* relacionadas con la transcripción, expresión, activación, modificación molecular, efectos de unión y procesos de transporte. Dado que estas interacciones han sido medidas experimentalmente, pueden ser asociadas con una dirección definida de efecto

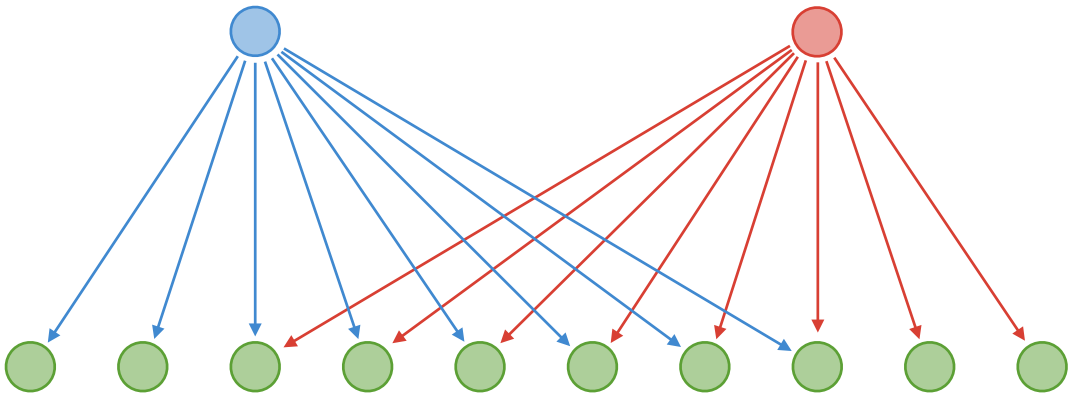


Figura 2-5: **Sinergia entre Reguladores Transcripcionales Maestros**. El conjunto de blancos de los FT azul y rojo sufren un efecto sinérgico dado por el efecto de cinética multiplicativa ya que están enriquecidos significativamente uno con el otro.

causal, ya sea activación o inhibición de los procesos mencionados arriba a nivel de redes en todo el genoma. Para más detalles metodológicos, por favor referirse a las referencias Krämer *et al.* [2014] y Espinal-Enríquez *et al.* [2015].

Este análisis se aplicó a los primeros 100 (top 100) RTMs y todos sus blancos, sólo para los Reguladores Transcripcionales Maestros del conjunto de regulones de la red con valores de $P > 1 \times 10^{-40}$. Al elegir los primeros cien RTMs se trató de colectar un número de genes $> 1,000$ en todos los casos. Con este número aseguramos un número de genes $> 4,000$ en la mayoría de los análisis. La información de las redes generadas es complementada con análisis de expresión diferencial entre muestras de tumores y muestras sanas y éstos constituyen la entrada para el estudio de CNA. Sólo se consideran los genes con *log fold-change* ($lfc = \log_2 \frac{e^{caso}}{e^{control}}$) en sus niveles de expresión > 1 y valores de $P < 0.0001$. De esta forma se trató de asegurar que fueran representadas diferencias realmente significativas entre ambos tratamientos. Es de hacer notar que éstos umbrales son, en realidad, el filtro selectivo de genes del top 100 de los cuatro conjuntos de regulones. Los genes diferencialmente expresados (GDE) fueron calculados por medio del paquete `limma` [Ritchie *et al.*, 2015]. Todo el código usado durante el procesamiento de las muestras puede ser consultado en el apéndice A.1. Para este análisis solo se trabajó con los resultados de la red de $P < 1 \times 10^{-40}$.

Capítulo 3

Resultados y discusión

Usando algoritmos de estimación de información mutua, en este trabajo se infirieron dos redes de regulación transcripcional a partir de dos conjuntos de microarreglos de expresión de mRNA que corresponden a muestras de tejido de biopsias de pacientes con cáncer primario de mama sin tratar y de tejido sano adyacente. Un conjunto compuesto por 880 muestras provenientes de la base de datos GEO y otro compuesto de 597 proveniente de la base de datos TCGA. A partir de estas redes, se eliminaron conexiones de las redes conservando aquellos con valores de P mayores a cuatro diferentes niveles de corte: 1×10^{-30}, 1×10^{-40}, 1×10^{-50} y 1×10^{-100}. Después se infirió la dirección de la regulación de los Factores de Transcripción basados en los niveles de expresión de sus genes blancos para las ocho subredes y se estimaron, con el algoritmo MARINa, aquellos FT que resultaron con un mayor número de blancos en la Firma Molecular de dicho fenotipo, esto es, los Reguladores Transcripcionales Maestros. Se obtuvieron ocho listas de RTM, una para cada valor de P en ambas redes. También se ejecutaron análisis de sombra y análisis de sinergia con el objetivo de encontrar aquellos genes que pudieran co-regular un subconjunto de genes de la FM. Finalmente, con los resultados obtenidos, se analizaron los genes en términos de las rutas canónicas de la IKB y las redes relacionadas con cáncer.

Por ser muy homogéneos los datos obtenidos en todos los análisis, a continuación se presentrán y discutirán los resultados de la red generada con un valor de restricción de $P = 1 \times 10^{-40}$. Primero para la red del conjunto de datos de GEO y luego para la de TCGA. Los resultados del resto de las redes se muestran en la sección 3.6.

3.1. Inferencia de redes de regulación transcripcional

Con ARACNe se obtuvo una red de regulación transcripcional para cada conjunto de datos correspondientes a cuatro valores de corte de P (1×10^{-30}, 1×10^{-40}, 1×10^{-50} y 1×10^{-100}). Como se puede ver en la Tabla 3-1 para los datos de GEO el número de elementos en las redes va de 11,026 en la menos restrictiva hasta 547 en la de mayor restricción. Otro elemento que varía mucho es el número de vecinos promedio que va de 727.48 hasta 8.76 así como el número de componentes en los que se separan de 2 hasta 28. Un elemento importante de la topología de la red para el algoritmo MARINa, es el nivel de interconectividad de la red. En una red con mucha interconectividad entre FT y blancos hay un mayor número de regulones que deben ser computados para su estimación como RTMs.

Tabla 3-1: Comparación de parámetros de las redes generadas con el conjunto de los datos de GEO.

Parámetro	P de 1×10^{-30}	P de 1×10^{-40}	P de 1×10^{-50}	P de 1×10^{-100}
Coeficiente de agrupamiento	0.52	0.47	0.422	0.191
Componentes conectados	2	10	28	26
Diámetro de la red	8	9	11	11
Radio de la red	1	1	1	1
Centralización de la Red	0.326	0.237	0.197	0.124
Rutas más cortas	121,517,554 (99 %)	60,427,348 (99 %)	24,586,898 (97 %)	53,944 (18 %)
Tamaño de eje promedio	2.656	3.301	4.036	3.954
Número de vecinos promedio	93.940	31.702	16.563	3.283
Número de nodos	11,026	7,797	5,033	547
Densidad de la red	0.009	0.004	0.003	0.006
Heterogeneidad de la Red	2.893	3.195	3.160	1.934
Nodos aislados	0	0	0	0
Número de autoconexiones	0	0	0	0
Pares de nodos multi-eje	21,898	5,456	1,960	77

3.2. Reguladores transcripcionales maestros

Los 10 reguladores transcripcionales maestros mejor calificados de entre los inferidos con el algoritmo MARINa para la red de $P = 1 \times 10^{-40}$ se muestran en la Figura 3-1.

El segundo RTM fue la proteína dedo de zinc 132, ZNF132. La subexpresión de este gen se asocia con una promoción aberrante de la hipermetilación y una mala prognosis en cáncer de próstata [Abildgaard *et al.*, 2012]. En las muestras este gen también está subexpresado, en concordancia con este último trabajo.

Otros RTMs obtenidos en esta lista tienen diferentes funciones relacionadas con el cáncer.

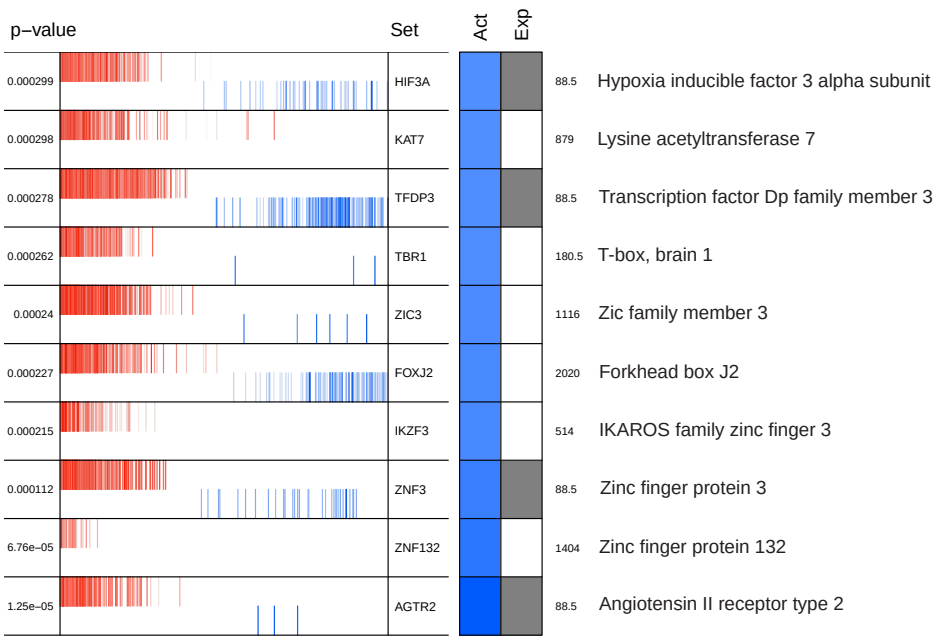


Figura 3-1: **Top 10 Reguladores Transcripcionales Maestros inferidos de los microarreglos de las muestras de tejido de cáncer de mama para GEO.** Este gráfico muestra la lista de los primeros diez mejores candidatos a RTMs (columna Set) en el conjunto de datos. El valor de P asociado a cada candidato a regulador transcripcional maestro se muestra a la izquierda. Se muestran a la derecha la actividad diferencial (Act) y la expresión diferencial (Exp) predicha por MARINa. Los tonos rojo de Exp o Act muestran sobreexpresión y los azules subexpresión. El puntaje de Exp de cada candidato regulador transcripcional maestro también se muestra en el lado derecho de la gráfica. Las líneas rojas en el medio de la gráfica representan objetivos de cada regulador transcripcional maestro que predijo ARACNe. La posición de cada línea en el eje horizontal corresponde a su z-score en la lista de genes de la Firma Molecular. Los genes con mayor sobrexpresión diferencial se muestran a la izquierda en rojo y los más diferencialmente subexpresados se muestran a la derecha en azul.

Por ejemplo, HIF3A es bien conocido como un regulador negativo de la tumorogénesis [Hara y Kondo, 2011; Heikkila *et al.*, 2011; Ando *et al.*, 2013]. Este comportamiento parece estar acorde con nuestras muestras de cáncer de mama, dado que HIF3A está también subexpresado. FOXJ2 es otro RTM que en nuestras muestras está subexpresado. La sobre expresión de este gen disminuye la migración de las células con cáncer de mama [Wang *et al.*, 2012]. Otros RTMs importantes están también involucrados en la regulación de otros tipos de cáncer. Por ejemplo el gen IKZF3 (AIOLOS), es un gen cuya sobreexpresión inhibe la proliferación celular en las células Nalm-6 [Zhuang *et al.*, 2014]. En resumen, podemos decir que algunos de los RTMs encontrados aquí son ampliamente conocidos por ser claves en el desarrollo de fenotipos cancerosos. Sin embargo, de otro conjunto de RTMs se desconoce la función que juegan en el desarrollo de la enfermedad.

El regulador transcripcional maestro con la mejor calificación fue el receptor de angiotensina 2, AGTR2. El hallasgo de esta proteína de membrana puso en evidencia que la notación de la lista de factores de transcripción no fue la mejor. Sin embargo este gen se ha mostrado que media el programa de muerte celular en células de *leiomiosarcoma* [Zhao *et al.*, 2015]. Además, la inhibición de AGTR2 está relacionada con el crecimiento celular y la evasión de la apoptosis en cáncer de mama [De Paepe *et al.*, 2002]. Este gen está subexpresado en nuestras muestras de cáncer, por tanto, esto puede ser indicativo de la disminución de los procesos apoptóticos en las muestras, uno de los *hallmarks* del cáncer.

Gráfico de colmena

En la Figura 3-2 podemos observar una visualización novedosa de la red del conjunto de regulones, llamada gráfico de colmena (*hiveplot* [Krzywinski *et al.*, 2011]). Con esta visualización se pueden apreciar algunas características interesantes de la red del conjunto de regulones. Se puede observar que un número pequeño de factores de transcripción controlan la mayoría de los genes diferencialmente expresados (firma molecular). Dado que esta visualización es una red no dirigida, es posible que aquellos factores de transcripción que aparecen en la firma molecular (eje de la derecha) pueden estar regulando los factores de transcripción de los eje verticales.

Es importante recalcar que este es otro análisis independiente del algoritmo de MARINa. Considerando esto, es notable que nueve reguladores transcripcionales maestros del top 10 inferidos por MARINa sean de los nodos más conectados dentro de la firma molecular. Además, de que parecen actuar sobre otro subgrupo de factores de transcripción en el eje vertical, el cual, a su vez, controla una gran parte del resto de la firma molecular. Estas conclusiones independientes hacen pensar en la robustez del algoritmo empleado para la búsqueda de reguladores transcripcionales maestros. A su vez, nos puede dar algunas pistas de la relevancia de los factores de transcripción diferencialmente expresados como inductores de fenotipo. Se necesita investigaciones adicionales para este último punto, pero es importante destacar que con esta metodología podemos argumentar características relevantes que corresponden a la regulación de la transcripción en células eucariotas, particularmente el carácter jerárquico que tienen estas redes regulatorias.

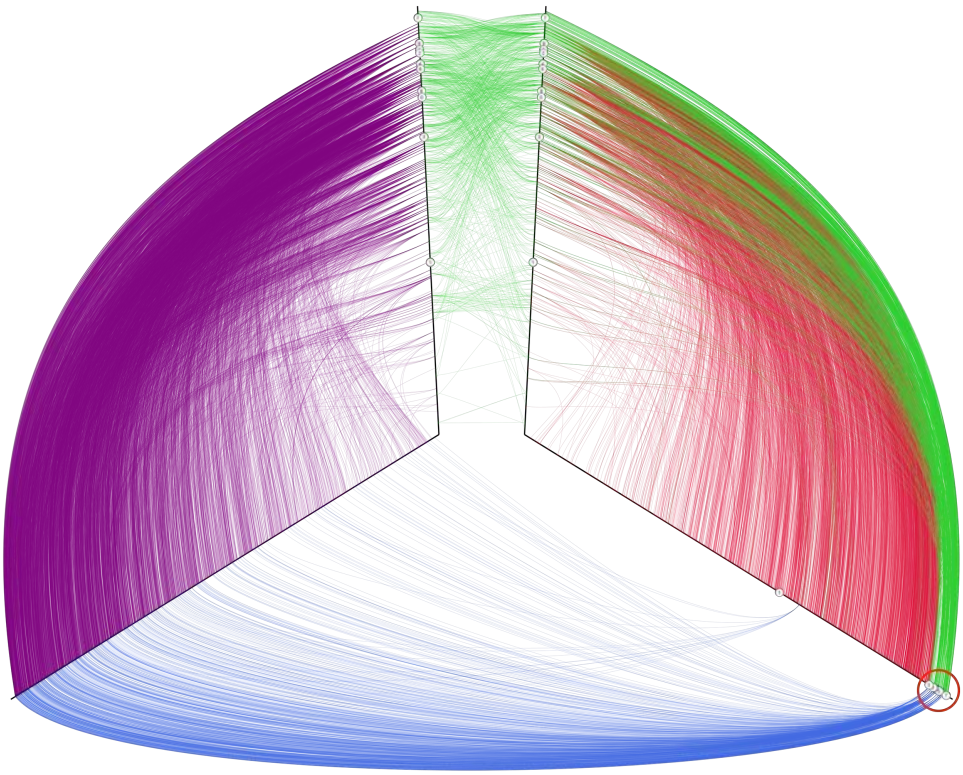


Figura 3-2: **Hiveplot de los Conjuntos de Regulones para los datos de GEO** En las lineas rectas en negro están representados los nodos que, en este caso, son genes y las lineas curvas representan interacciones (dadas por su valor de información mutua). Estos están ordenados de tal manera que los nodos más conectados están en la parte más distal de las lineas y al centro los nodos menos conectados. Ambos ejes verticales contienen la lista de los Factores de transcripción mientras que el eje derecho representa genes diferencialmente expresados entre sanos y enfermos. El eje izquierdo contiene los genes no diferenciados. La interacción entre dos Factores de transcripción se representa en verde. La interacción entre FT y un gen diferencialmente expresado en rojo, mientras que las curvas moradas representan interacciones entre FT y genes no diferencialmente expresados. Por último, las lineas azules representan las interacciones de aquellos FT que están diferencialmente expresados y los genes no diferencialmente expresados. Dentro del círculo rojo se encuentran nueve de los 10 RTMs mejor calificados por MARINa: (0) AGTR2, (1) ZNF132, (2) ZNF3, (3) IKZF3, (4)FOXJ2, (5) ZIC3, (6) TBR1, (7) TFDP3 (8) HIF3A y (9) KAT7. Para una mejor visualización, las interacciones con una información mutua menor de 0.85 fueron eliminados. Una copia en linea con máxima resolución puede ser descargada de https://goo.gl/9b1FJI

3.2.1. Análisis de sombras

Un par de factores de transcripción $\mathrm{FT_n}$ y $\mathrm{FT_m}$ pueden compartir una proporción importante de blancos. Sin embargo, los blancos no compartidos nos pueden ayudar a diferenciar cuál de ambos factores de transcripción tiene un influencia mayor sobre el fenotipo que el otro. Si el enriquecimiento de los blancos exclusivos de $\mathrm{FT_n}$ afecta más significativamente la firma molecular que de los blancos exclusivos de $\mathrm{FT_m}$ y no al contrario, decimos que el $\mathrm{FT_m}$ es *sombra* del $\mathrm{FT_n}$. Los resultados de análisis de sombras para la red de $P < 1 \times 10^{-40}$ se pueden observar

en la figura 3-3. En esta red dirigida (dado que el análisis de sombra genera una relación FT $\rightarrow$ $\text{FT}_{\text{sombra}}$) el conjunto de blancos únicos del factor de transcripción blanco enriquecería menos la firma molecular que los bláncos exclusivos del factor de transcripción origen, aunque sus blancos comunes la enriquezcan igual. Para ver fácilmente estas relaciones, se recurrió a la visualización conocida como BioFabric [Longabaugh, 2012], la cual facilita observar las relaciones jerárquicas entre nodos en redes de pequeñas a medianas (hasta alrededor de 10,000 nodos). Su principal característica es que los nodos están representados por líneas horizontales y sus relaciones como líneas verticales. Esto es, los ejes de la red son identificables porque corren de forma transversal hasta alcanzar su nodo vecino. Si se ordenan los nodos por grado se puede representar un arreglo que permite contemplar las jerarquías entre nodos. En este caso los nodos principales contienen los blancos de sus nodos subsecuentes. En la figura se puede observar que algunos blancos del RTMs del top 10 son sombra de algunos otros RTMs, por ejemplo el caso de ZHX3, ZNF132 y otros (Figura 3-3).

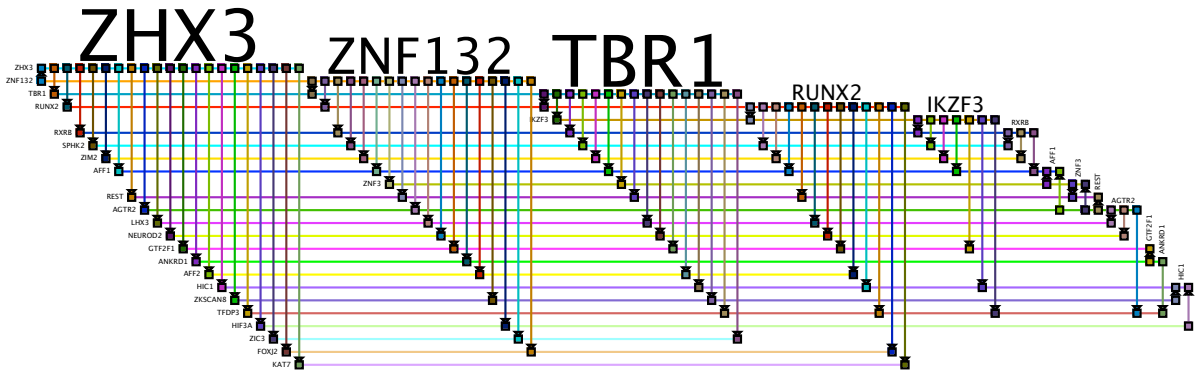


Figura 3-3: **Red generada con los pares de factores de transcripción resultado del análisis de sombra para el conjunto de datos de GEO.** En esta visualización (hecha con el software BioFabric [Longabaugh, 2012]) los nodos se representan como líneas horizontales, una por renglón, y los ejes son lineas verticales, una por columna. En estas redes, la dirección de las interacciones $\text{FT}_{\text{origen}} \rightarrow \text{FT}_{\text{blanco}}$ significa que el enriquecimiento de los genes no compartidos entre $\text{FT}_{\text{origen}}$ y $\text{FT}_{\text{blanco}}$ no es significativamente distinto. Dado que la interacción es directa, se puede observar, por ejemplo, que el gen que ensombrece más FT es ZHX3, seguido por ZNF132 etcétera. Una copia en linea con máxima resolución puede ser consultada en https://goo.gl/hJPiIv

3.2.2. Análisis de sinergia

El principal objetivo del análisis de sinergia es detectar aquellos reguladores transcripcionales maestros que tienen un efecto en conjunto sobre sus blancos. Esto es, que el resultado de la actividad de un par (o grupo) de reguladores transcripcionales maestros es más que la suma de

sus actividades por separado [Carro *et al.*, 2010; Aytes *et al.*, 2014]. En este trabajo se ejecutó el análisis de sinergia para los primeros 25 reguladores transcripcionales maestros obtenidos en MARINa. Los resultados se pueden ver en la Figura 3-4 para el conjunto de GEO y en la Figura 3-11 para las redes del conjunto de datos de TCGA. Éstas muestran los 10 primeros conjuntos de RTMs que actúan sinérgicamente sobre sus blancos.

Se puede observar que TBR1 y RUNX2 se encuentran presentes en los 10 grupos de genes reguladores maestros (GGRMs). En cambio, AFF2 y TFDP3 se encuentra en 7 de 10 GGRMs. Otros RTMs están presentes en el núcleo central de regulación de diversos GGRMs. Un GGRM está compuesto de un conjunto de reguladores maestros que regulan sinérgicamente la transcripción de *un mismo conjunto de genes blanco*. Este hecho hace al control transcripcional por grupos sinérgicos un fenómeno muy robusto. Las implicaciones biológicas de esto podrían ser de gran importancia en contextos de las enfermedades de vías metabólicas como la aquí estudiada. Por ejemplo de Anda-Jáuregui *et al.* [2015] llaman a esta conjunción de blancos "*Crosstalk*" y discuten la posibilidad de que este tipo de entrecruzamiento entre diferentes vías pueda ser la fuente de la resistencia a fármacos de algunos subtipos de cáncer de mama. Es muy probable que los RTMs que actúan sinérgicamente serían nodos clave en este fenómeno de *Crosstalk*.

3.3. Análisis de redes causales

Este análisis tuvo como objetivo explorar qué procesos son enriquecidos por los primeros reguladores transcripcionales maestros y todos sus blancos usando una base de datos altamente curada, la *Ingenuity Knowledge Base*. Para este análisis se tomaron los top 100 RTMs obtenidos por MARINa junto con todos sus blancos inferidos por ARACNe. A continuación fue sometida esta lista junto con su valores de log_2 *fold change* y su valor de P a la aplicación *Ingenuity Pathway Analysis* http://www.ingenuity.com/products/ipa.

En esta exploración obtuvimos que la tercera via canónica más importante en el grupo de regulones fue la de "*Mecanismos moleculares del cáncer*" ($P = 4.78\times10^{-8}$). Además, con los resultados del Análisis de Redes Causales (*Causal Network Analysis*, CNA) fuimos capaces de

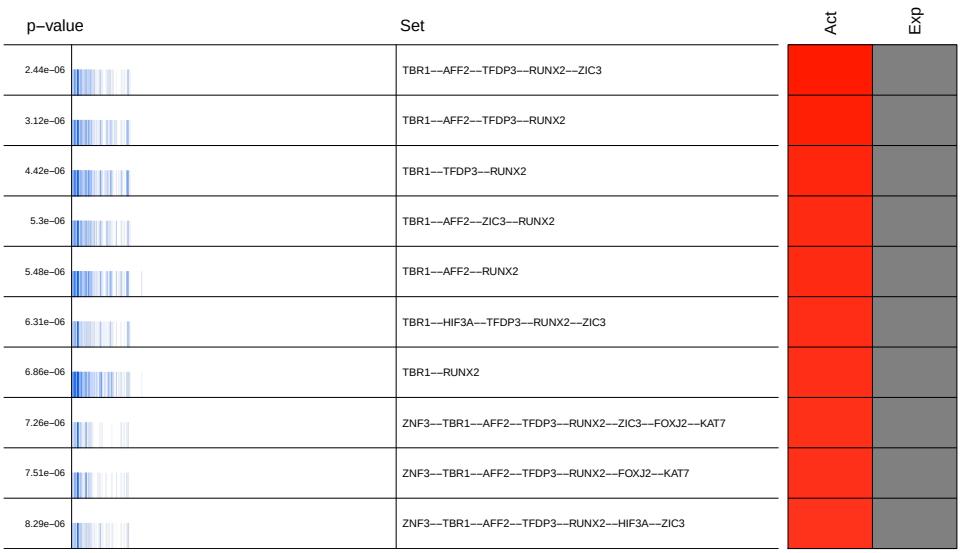


Figura 3-4: **Gráficos del Análisis de Sinergia para el conjunto de GEO.** Se muestran los primeros 10 conjuntos de genes reguladores maestros que regulan sinérgicamente conjuntos de genes de la firma molecular en el conjunto de datos. La columna "Set" muestra aquellos factores de transcripción que corregulan sinérgicamente el mismo conjunto de blancos. El conjunto restante de elementos de la gráfica corresponden con los descritos en la Figura 3-1. Los ocho genes involucrados en el fenómeno de sinergia fueron *T-box, brain 1* (TBR1), *AF4/FMR2 family member 2* (AFF2), *transcription factor Dp family member 3* (TFDP3), *runt related transcription factor 2* (RUNX2), *Zic family member 3* (ZIC3), *hypoxia inducible factor 3 alpha subunit* (HIF3A), *forkhead box J2* (FOXJ2) y *lysine acetyltransferase 7* (KAT7).

encontrar módulos funcionales relacionados con algunos conocidos *hallmarks* de cáncer [Hanahan y Weinberg, 2000, 2011].

Es de destacar que unos de los primero 10 RTM *IKAROS family zinc finger 3* (IKZF3) el cual está involucrado en la remodelación de la cromatina está presente el la red relacionada con control del ciclo celular (Figura 3-5). Esta red tiene como protagonistas destacados al gen *cyclin D1* (CCND1) y a *E2F transcription factor 4* (E2F4). El primero pertenece a la familia muy conservada de las ciclinas y que se sabe que tienen una fuerte relación con la regulación del ciclo celular. El segundo pertenece a la familia de las proteinas E2F también con una importancia crucial en el control del ciclo celular.

Además, en el caso de la red de apoptosis (Figura 3-6), podemos destacar la actividad del regulador maestro *transcription factor Dp family member 3* (TFDP3), el cual heterodimeriza con la proteína E2F para fortalecer su unión con el DNA y promover la transcripción de los ge-

nes blanco de E2F. En esta red relacionada con muerte celular, si bien se presentan importantes moléculas como *tumor necrosis factor* (TNF) y *caspase 8* (CASP8) hay que aclarar que éstas no forman parte del conjunto de moleculas blanco de los 100 regulones. Sin embargo *caspase 8 associated protein 2* (CASP8AP2) sí está presente así como *PDLIM2 PDZ and LIM domain 2* (PDLIM2), un promotor de la migración y de la adhesión celular.

Con respecto a la red relacionada con cáncer (Figura 3-7), se puede ver al RTM *lysine acetyltransferase 7* (KAT7) el cual participa en la regulación de algunas oncoproteínas. Esta red está centrada en *tumor protein p63* (TP63) el cual forma parte de la familia de p53 muy conocida por su imporatancia en el proceso del control de la apoptosis.

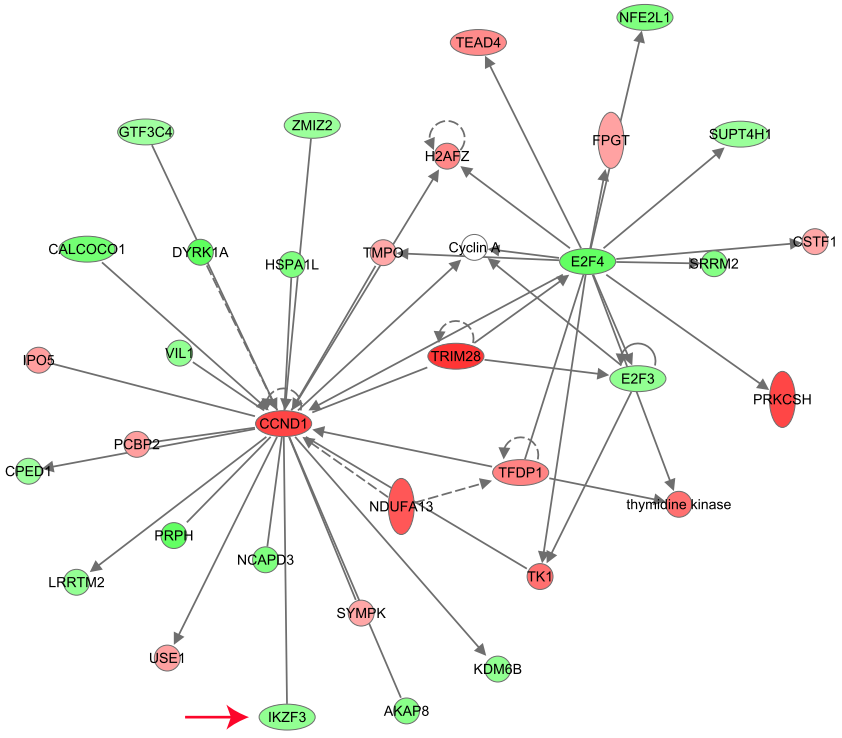


Figura 3-5: **Red causal obtenida con la IPA-KB relacionada con Ciclo Celular.** En esta red es de destacar que unos de los primero 10 RTM *IKAROS family zinc finger 3* (IKZF3) el cual está involucrado en la remodelación de la cromatina. Todas las moléculas en color están presentes entre algunos de los regulones de los RTMs del top 100. Las moléculas en rojo están sobre-expresadas y las moléculas verdes muestran subexpresión. La intensidad de color representa la diferencia entre las muestras de tejido con cáncer de mama comparadas con las normales.

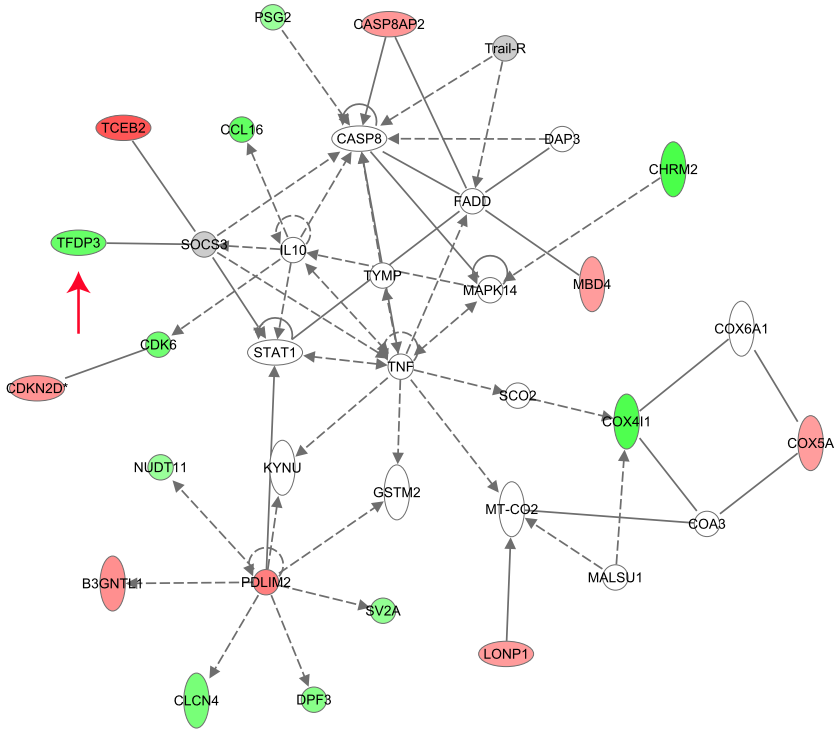


Figura 3-6: **Red causal obtenida con la IPA-KB relacionada con Muerte Celular.** En esta red podemos mencionar la actividad del regulador maestro *transcription factor Dp family member 3* (TFDP3), el cual heterodimeriza con la proteína E2F para fortalecer su unión con el DNA y promover la transcripción de los genes blanco de E2F. Todas las moléculas en color están presentes entre algunos de los regulones de los RTMs del top 100. Las moléculas en rojo están sobre-expresadas y las moléculas verdes muestran subexpresión. La intensidad de color representa la diferencia entre las muestras de tejido con cáncer de mama comparadas con las normales.

3.3.1. El análisis de sinergia resaltó un conjunto de RTMs subregulados

Como se puede observar en la Figura 3-4, los RMTs TBR1, RUNX2 y TFDP3 están presentes en todo el top 10 de grupos de genes reguladores maestros. Así, en las demás redes se puede ver un fenómeno similar (H2AFX en la de con $P< 1\times10^{-30}$ y TP63 y LHX3 en la de $P< 1\times10^{-50}$ Apéndice 3.6.2). Este hecho puede ser atribuido a la regulación concomitante de éstos en un conjunto particular de blancos. El hecho de que los mismos, (o casi los mismos) conjuntos de RTMs son capaces de regular un número diferente de GGRMs puede indicar un programa regulatorio robusto capaz de *definir fenotipos*.

Si consideramos algunos de los *hallmarks* del cáncer [Hanahan y Weinberg, 2011], por ejemplo desregulación del ciclo celular, inhibición de la apoptosis, migración, angiogénesis y prolife-

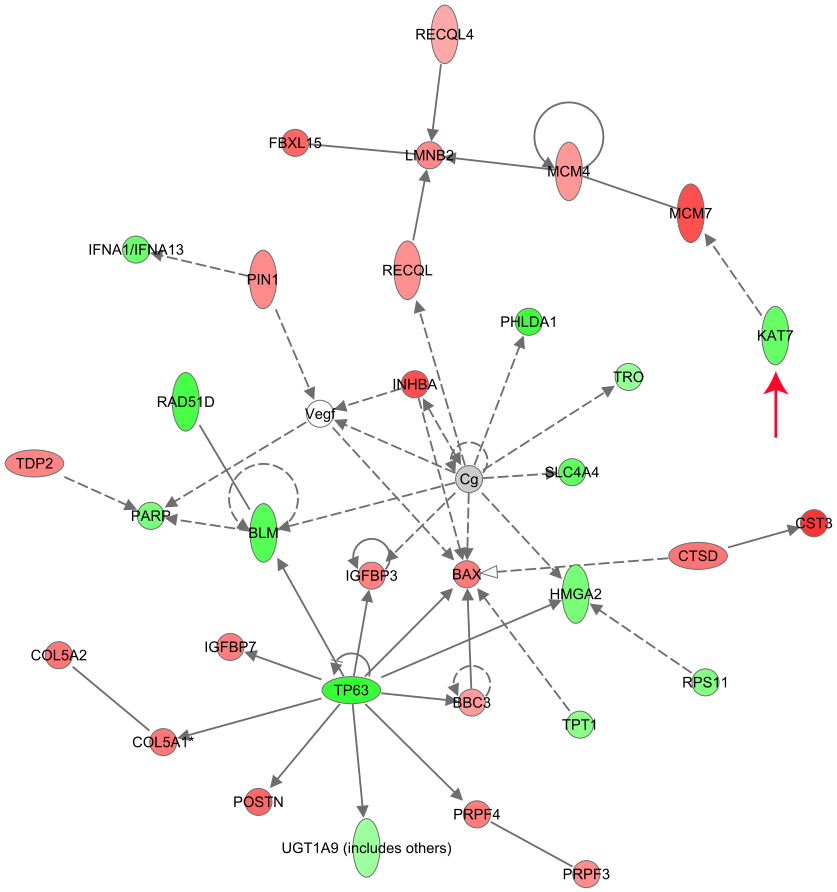


Figura 3-7: **Red Causal obtenida con la IPA-KB relacionada con cáncer.** En esta red se puede ver al RTM *lysine acetyltransferase 7* (KAT7). Todas las moléculas en color están presentes entre algunos de los regulones de los RTMs del top 100. Las moléculas en rojo están sobre-expresadas y las moléculas verdes muestran subexpresión. La intensidad de color representa la diferencia entre las muestras de tejido con cáncer de mama comparadas con las normales.

ración; podemos inferir algunos efectos que tienen los grupos de RTMs sobre éstos. Por ejemplo, el gen KAT7 (Figura 3-7), el cual se presenta en dos grupos sinérgicos, se ha relacionado con desregulación del ciclo celular [Siriwardana *et al.*, 2014]. TFDP3 el cual se comentó previamente, se presenta en siete de los 10 grupos y su desregulación se ha asociado a la evasión de la apoptosis [Tian *et al.*, 2007]. RUNX2 así como FOXJ2 están relacionados en procesos de migración [Boregowda *et al.*, 2014; Wang *et al.*, 2014, 2012]. HIF3 está implicado en el proceso de angiogénesis [Ando *et al.*, 2013]. Finalmente, la proteína de dedo de zinc ZNF3, está relacionada con proliferación [Gao *et al.*, 2008]. También se encontró dos RTMs que no están caracterizados en la literatura como relacionados con cáncer. Este es el caso de TBR1 y ZIC3, los cuales se ha reportado que están relacionados con el desarrollo del cerebro y corazón [Bulfone *et al.*, 1995; Cowan *et al.*, 2014]. Estos hallazgos previos deben ser analizados en estudios posteriores para tratar de entender las bases de la regulación transcripcional en Eucariontes, en particular, en relación con el cáncer.

3.4. Resultados y discusión del conjunto de TCGA

En esta sección se describen los resultados de este flujo de trabajo con los datos del TCGA con la red recortada a un valor de $P< 1\times10^{-40}$. Es importante señalar algunos aspectos fundamentales para tratar de hacer una comparación entre estos resultados y los del conjunto de datos de GEO: El número de muestra entre ambos conjuntos es un factor importante de considerar (880 para el conjunto de GEO y 597 de TCGA). Esto podemos notarlo al comparar los parámetros de las redes generadas que, a pesar de tener los mismos valor de corte de P (equivalente a MI), su topología es visiblemente diferente (Tabla 3-2 y figuras 3-14 y 3-15). Otro aspecto importante a considerar es que los datos del TCGA incluyen sólo muestras de tejido de cáncer invasivo que debemos ponderar al comparar los resultados de ambos conjuntos de datos. Sin embargo, a pesar de todas estas diferencias, se pudo capturar cómo los patrones de control de la transcripción parecen similares en ambos conjuntos como se podría interpretar de los hiveplots para ambos conjuntos (figuras 3-2 y 3-9).

Tabla 3-2: Comparación de parámetros de las redes inferidas por ARACNe con un valor de corte de $P< 1\times10^{-40}$ tanto para el conjunto de GEO como de TCGA.

Parámetro	Red de GEO	Red de TCGA
Coeficiente de agrupamiento	0.47	0.474
Componentes conectados	10	87
Diámetro de la red	9	17
Radio de la red	1	1
Centralización de la Red	0.237	0.123
Rutas más cortas	60,427,348 (99 %)	15,169,902 (85 %)
Tamaño de eje promedio	3.301	4.583
Número de vecinos promedio	31.702	14.350
Número de nodos	7,797	4,215
Densidad de la red	0.004	0.003
Heterogeneidad de la Red	3.195	2.725
Pares de nodos multi-eje	5,456	1,655

3.4.1. Inferencia de redes de regulación transcripcional

La Tabla 3-3 nos muestra estos parámetros para las redes construidas con el conjunto de datos de TCGA. Como se mencionó arriba, hay que considerar que esta red fue construida con un número menor de experimentos y, por tanto, los valores de *MI* son menores. En la tabla podemos observar claramente como hay mayor número de componentes conectados (58, 87, 86 y 31) en comparación con las redes de GEO (2, 10, 28 y 26). Éste es un dato fundamental para la inferencia de RTMs dado que MARINa considera el número de conexiones que tiene cada FT y deja de lado aquellos que tienen menos de 20 blancos. Esto es, en las redes de TCGA, con redes más ralas y un alto número de elementos disgregados, se estimó un número menor de factores de transcripción como RTMs.

Tabla 3-3: Comparación de parámetros de las redes generadas con el conjunto de datos de TCGA.

Parámetro	P de 1×10^{-30}	P de 1×10^{-40}	P de 1×10^{-50}	P de 1×10^{-100}
Coeficiente de agrupamiento	0.461	0.474	0.483	0.406
Componentes conectados	58	87	86	31
Diámetro de la red	14	17	16	10
Radio de la red	1	1	1	1
Centralización de la Red	0.147	0.123	0.104	0.207
Rutas más cortas	51,129,972 (95 %)	15,169,902 (85 %)	4,970,594 (15 %)	25,726 (15 %)
Tamaño de eje promedio	3.967	4.583	4.983	2.478
Número de vecinos promedio	21.503	14.350	10.974	7.348
Número de nodos	7,301	4,215	2,597	411
Densidad de la red	0.003	0.003	0.004	0.018
Heterogeneidad de la Red	3.040	2.725	2.494	1.934
Nodos aislados	0	0	0	0
Número de autoconexiones	0	0	0	0
Pares de nodos multi-eje	3,730	1,655	891	180

3.4.2. Reguladores transcripcionales maestros

Los 10 reguladores transcripcionales maestros mejor calificados inferidos con el algoritmo MARINa para las cuatro redes de TCGA se muestran en la Figura 3-8. Mientras que en la figura 3-9 se muestra el hiveplot para la red de $P< 1\times10^{-40}$. En ésta última alcanzamos a ver el mismo carácter jerárquico de las interacciones entre FT y firma molecular así como de los principales reguladores transcripcionales maestros hallados con el algoritmo MARINa.

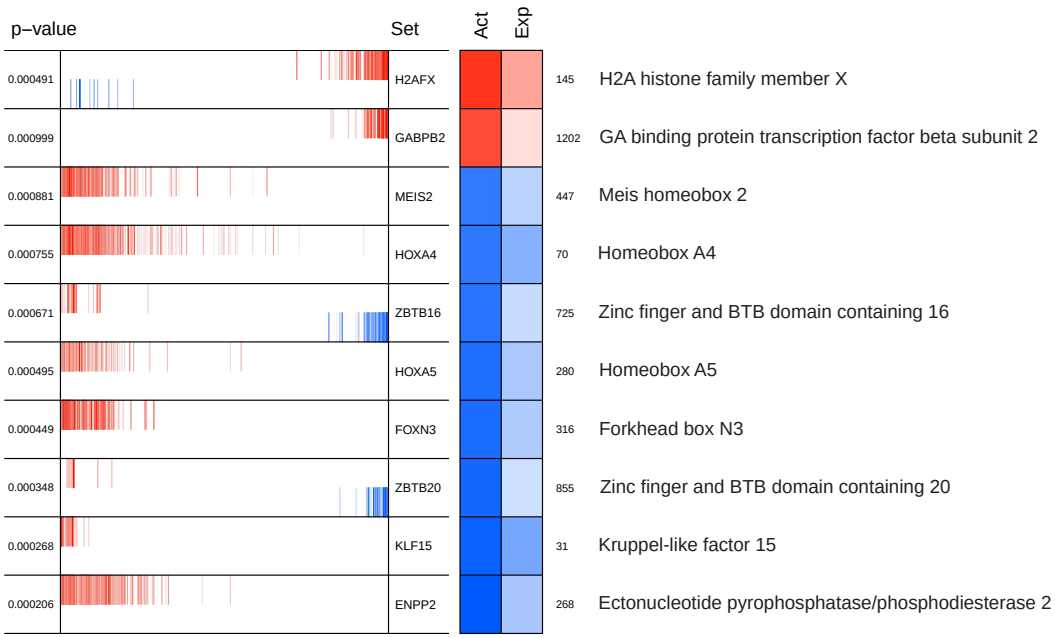


Figura 3-8: **Top 10 Reguladores Transcripcionales Maestros inferidos de los microarreglos de las muestras de tejido de cáncer de mama para TCGA.** Este gráfico muestra la lista de los primeros diez mejores candidatos a RTMs (columna Set) en el conjunto de datos. El valor de P asociado a cada candidato a regulador transcripcional maestro se muestra a la izquierda. Se muestran a la derecha la actividad diferencial (Act) y la expresión diferencial (Exp) predicha por MARINa. Los tonos rojo de Exp o Act muestran sobre expresión y los azules subexpresión. El puntaje de Exp de cada candidato regulador transcripcional maestro también se muestra en el lado derecho de la gráfica. Las líneas rojas en el medio de la gráfica representan objetivos de cada regulador transcripcional maestro que predijo ARACNe. La posición de cada línea en el eje horizontal corresponde a su z-score en la lista de genes de la firma molecular. Los genes con mayor sobrexpresión diferencial se muestran a la izquierda en rojo y los más diferencialmente subexpresados se muestran a la derecha en azul.

Análisis de sombras

En la figura 3-10 se puede observar que algunos candidatos del top 10 a reguladores transcripcionales maestros son sombra de otros. Por ejemplo PPARG, un factor de transcripción conocido por su relación con el cáncer de mama [Abduljabbar *et al.*, 2015; Pon *et al.*, 2015;

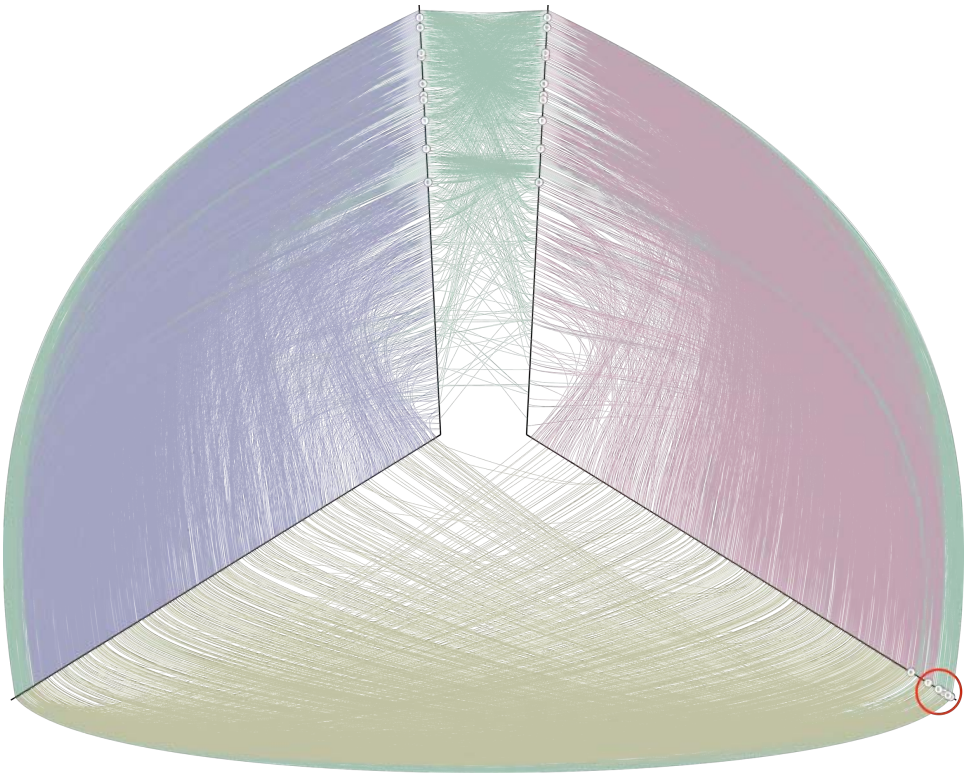


Figura 3-9: **Hiveplot de los Conjuntos de Regulones para los datos de TCGA** En las líneas rectas en negro están representados los nodos que, en este caso, son genes y las líneas curvas representan interacciones (dadas por su valor de información mutua). Ambos ejes verticales contienen la lista de los Factores de transcripción mientras que el eje derecho representa genes diferencialmente expresados entre sanos y enfermos. El eje izquierdo contiene los genes no diferenciados. La interacción entre dos Factores de transcripción se representa en verde. La interacción entre FT y un gen diferencialmente expresado en rojo, mientras que las curvas moradas representan interacciones entre FT y genes no diferencialmente expresados. Por último, las lineas azules representan las interacciones de aquellos FT que están diferencialmente expresados y los genes no diferencialmente expresados. Dentro del círculo rojo se encuentran nueve de los 10 RTMs mejor calificados por MARINa: (0) ENPP2, (1) KLF15, (2) ZBTB20, (3) FOXN3, (4) HOXA5, (5) ZBTB16, (6) HOXA4, (7) MEIS2, (8) H2AFX y (9) GABPB2. Para una mejor visualización, las interacciones con una información mutua menor de 0.85 fueron eliminados. Una imagen con gran resolución puede ser descargada en linea de https://goo.gl/beH1hB

Park *et al.*, 2014; Zhang *et al.*, 2015], se encuentra sombreando a KLF15, HOXA5, ENPP2, HOXA4 y MEIS2. Esto es, los blancos de PPARG no compartidos con estos RTMs enriquecen más la firma molecular. Por otro lado KLF15, un gen poco anotado, se encuentra sombreando a HOXA5, FOXN3, ENPP2, HOXA4, MEIS2 y al mismo PPARG (Figura 3-10). Según estos resultados, KLF15 es un interesante candidato para ser investigado a profundidad en el contexto de cáncer de mama.

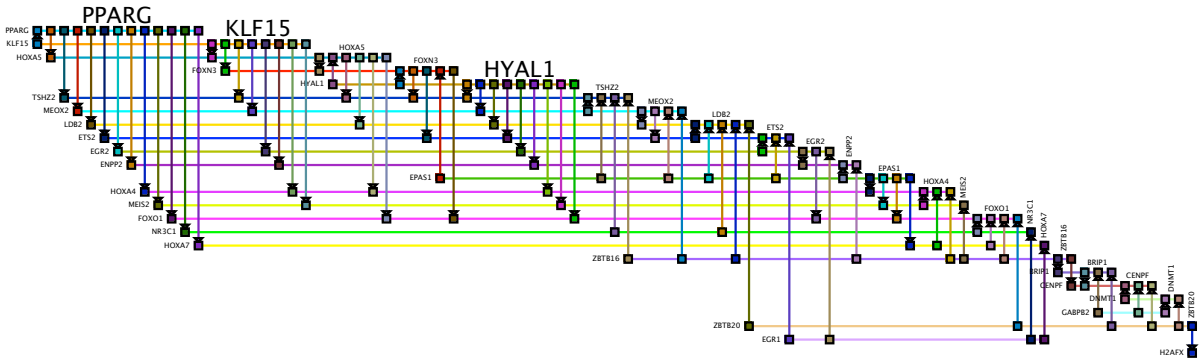


Figura 3-10: **Red generada con los pares de FT resultado del análisis de sombra para el conjunto de datos de TCGA.** En esta visualización (hecha con el software BioFabric [Longabaugh, 2012]) los nodos se representan como líneas horizontales, una por renglón, y los ejes son lineas verticales, una por columna. En estas redes, la dirección de las interacciones $\mathrm{FT}_{\mathrm{origen}} \rightarrow \mathrm{FT}_{\mathrm{blanco}}$ significa que el enriquecimiento de los genes no compartidos entre $\mathrm{FT}_{\mathrm{origen}}$ y $\mathrm{FT}_{\mathrm{blanco}}$ no es significativamente distinto. Una copia de esta figura en alta resolución puede ser descargada de https://goo.gl/tYe98i.

Análisis de sinergia

Los resultados del análisis de sinergia para las redes del conjunto de datos de TCGA se pueden ver en la Figura 3-11. En éstos también se puede apreciar el comportamiento de grupos de genes reguladores maestros como se discutió en el conjunto de GEO. Una serie de reguladores transcripcionales maestros se repite a lo largo de los conjuntos de reguladores en sinergia y solo unos cuantos por grupo no se repiten o apenas se repiten.

3.5. Análisis de redes causales

Con el mismo procedimiento, los cien primeros RTMs obtenidos con MARINa para la red de $P< 1\times10^{-40}$ y todos sus blancos, se analizó con la *IPA-KB*. Dos redes obtenidas son interesantes por contener genes del top 10 además de ser redes con relación a *hallmarks* del cáncer.

El RTM obtenido entre los primeros diez *ectonucleotide pyrophosphatase/phosphodiesterase 2* (ENPP2) la cual se sabe que está involucrada con la proliferación celular y la quimiotaxis, se encuentra interactuando en la red relacionada con cáncer (Figura 3-12). Ésta es otra red centrada en el gen *cyclin D1* (CCND1, esta vez no presente en el conjunto de moléculas analizadas) y que sabe que tienen una fuerte relación con la regulación del ciclo celular. Sí está presemte *FBJ murine osteosarcoma viral oncogene homolog* (FOS) la cual se sospecha implicada en la

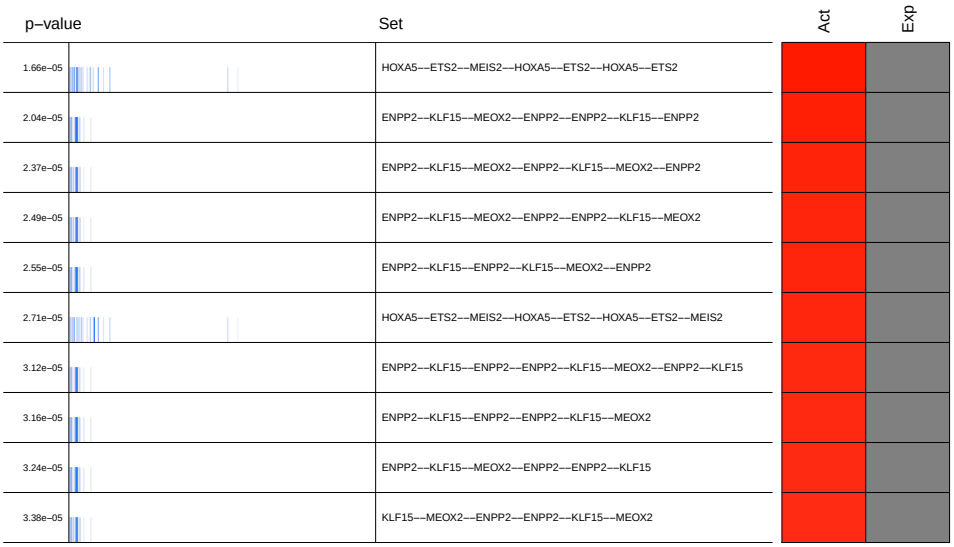


Figura 3-11: **Gráficos del Análisis de Sinergia para el conjunto de TCGA.** Se muestran los primeros 10 conjuntos de genes reguladores maestros que regulan sinérgicamente conjuntos de genes de la FM en nuestro conjunto de datos para las cuatro redes. La columna *Set* muestra aquellos genes que corregulan sinérgicamente el mismo conjunto de blancos. Los seis genes involucrados en el fenómeno de sinergia fueron *homeobox A5* (HOXA5), *ETS proto-oncogene 2, transcription factor* (ETS2), *Meis homeobox 2* (MEIS2), *ectonucleotide pyrophosphatase/phosphodiesterase 2* (ENPP2), *Kruppel-like factor 15* (KLF15) y *mesenchyme homeobox 2* (MEOX2)

regulación de la proliferación, diferenciación y transformación celular.

En la red relacionada con desarrollo y función del sistema cardiovascular (Figura 3-13) podemos ver al RTM *zinc finger and BTB domain containing 16* (ZBTB16) una porteína de dedos de zinc la cual está involucrada la progresión del ciclo celular. Es uno de los primeros genes mejor punteados por MARINa en el análisis de esta red particular. En esta red se pueden destacar moléculas como CD40 y CD80 miembros de una familia de receptores involucrados en la respuesta inmune y el reconocimiento celular.

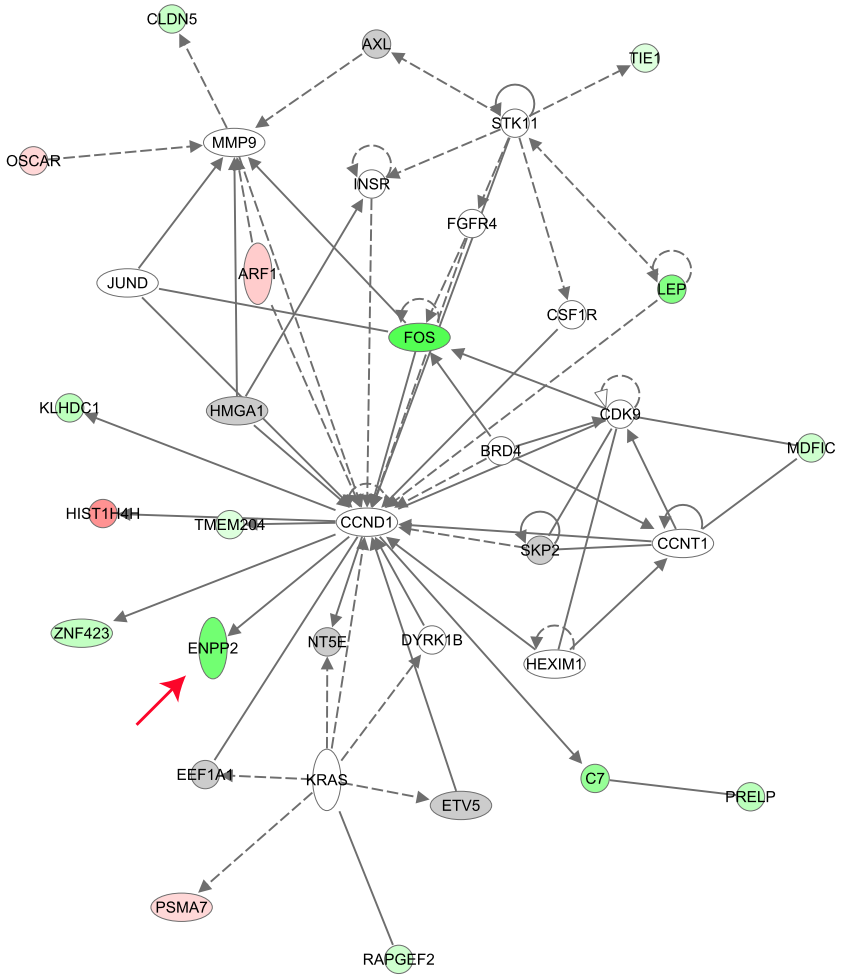


Figura 3-12: **Red causal obtenida con la IPA-KB para el conjunto de TCGA relacionada con cáncer.** El RTM obtenido entre los primeros diez *ectonucleotide pyrophosphatase/phosphodiesterase 2* (ENPP2) la cual se sabe que está involucrada con la proliferación celular y la quimiotaxis, está presente en esta red. Las moléculas en rojo están sobre-expresadas y las moléculas verdes muestran subexpresión. La intensidad de color representa la diferencia entre las muestras de tejido con cáncer de mama comparadas con las normales.

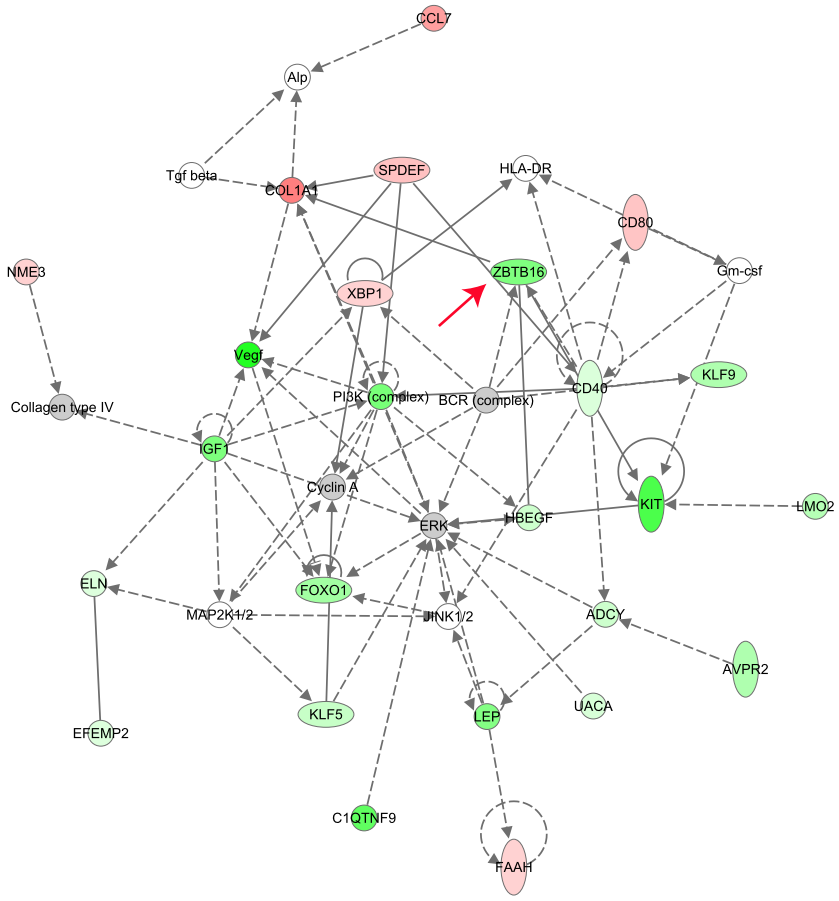


Figura 3-13: **Red causal obtenida con la IPA-KB para el conjunto de TCGA relacionada con desarrollo y función del sistema vascular.** En la red podemos ver al RTM *zinc finger and BTB domain containing 16* (ZBTB16) una porteína de dedos de zinc la cual está involucrada la progresión del ciclo celular. Las moléculas en rojo están sobre-expresadas y las moléculas verdes muestran subexpresión. La intensidad de color representa la diferencia entre las muestras de tejido con cáncer de mama comparadas con las normales.

3.6. Resultados de las redes con distintos niveles de corte

Además de los resultados presentados hasta ahora, se llevó a cabo todos estos análisis para el resto de las subredes en ambos conjuntos de datos. Para hacer más clara y fluida la presentación de los datos, el resto de los resultados se presenta en esta sección. Presenteamos las visualizaciones generadas con Cytoscape de las ocho redes (cuatro de cada conjunto de datos) que corresponden a los cuatro niveles de corte de P ya mencionados (1×10^{-30}, 1×10^{-40}, 1×10^{-50} y 1×10^{-100}). El resto de los resultados se presentan en dos subsecciones, una para cada conjunto de datos. Se muestran los diez primeros reguladores maestros mejor calificados para las otras tres redes (1×10^{-30}, 1×10^{-50} y 1×10^{-100}), los seis hiveplots restantes (tres y tres), las seis redes restantes generadas con BioFabric que muestran los resultados del análisis de sombras con sus ligas respectivas a las figuras en alta resolución disponibles en la Internet y las gráficas de `ssmarina` para los análisis de sinergia.

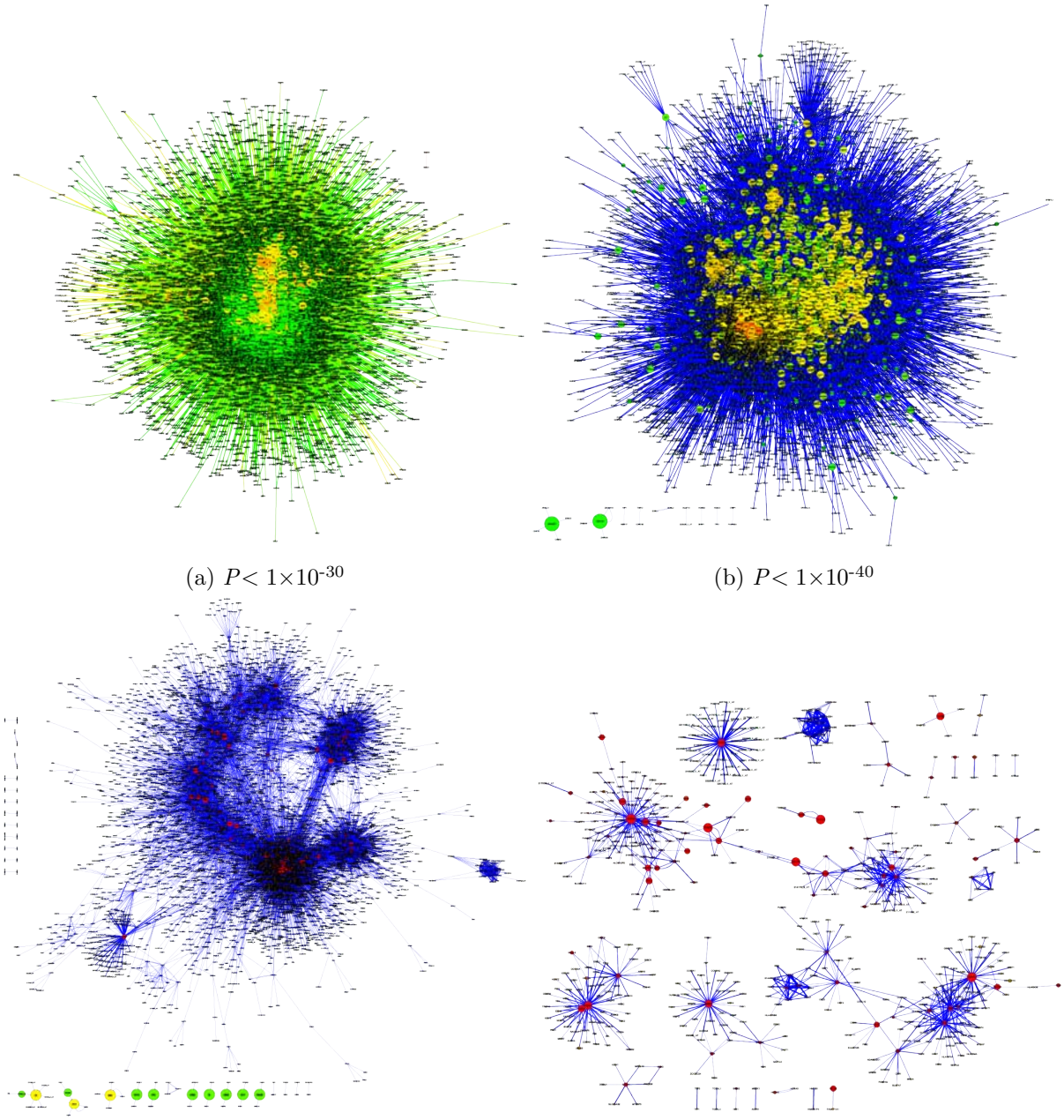
3.6.1. Redes inferidas por ARACNe

Conjunto de GEO

La Figura 3-14 muestra las topologías de la red generada con el conjunto de GEO recortado con diferentes valores de P.

Conjunto de TCGA

La Figura 3-15 muestra la topología de las redes inferidas por ARACNe para el conjunto de datos de TCGA.

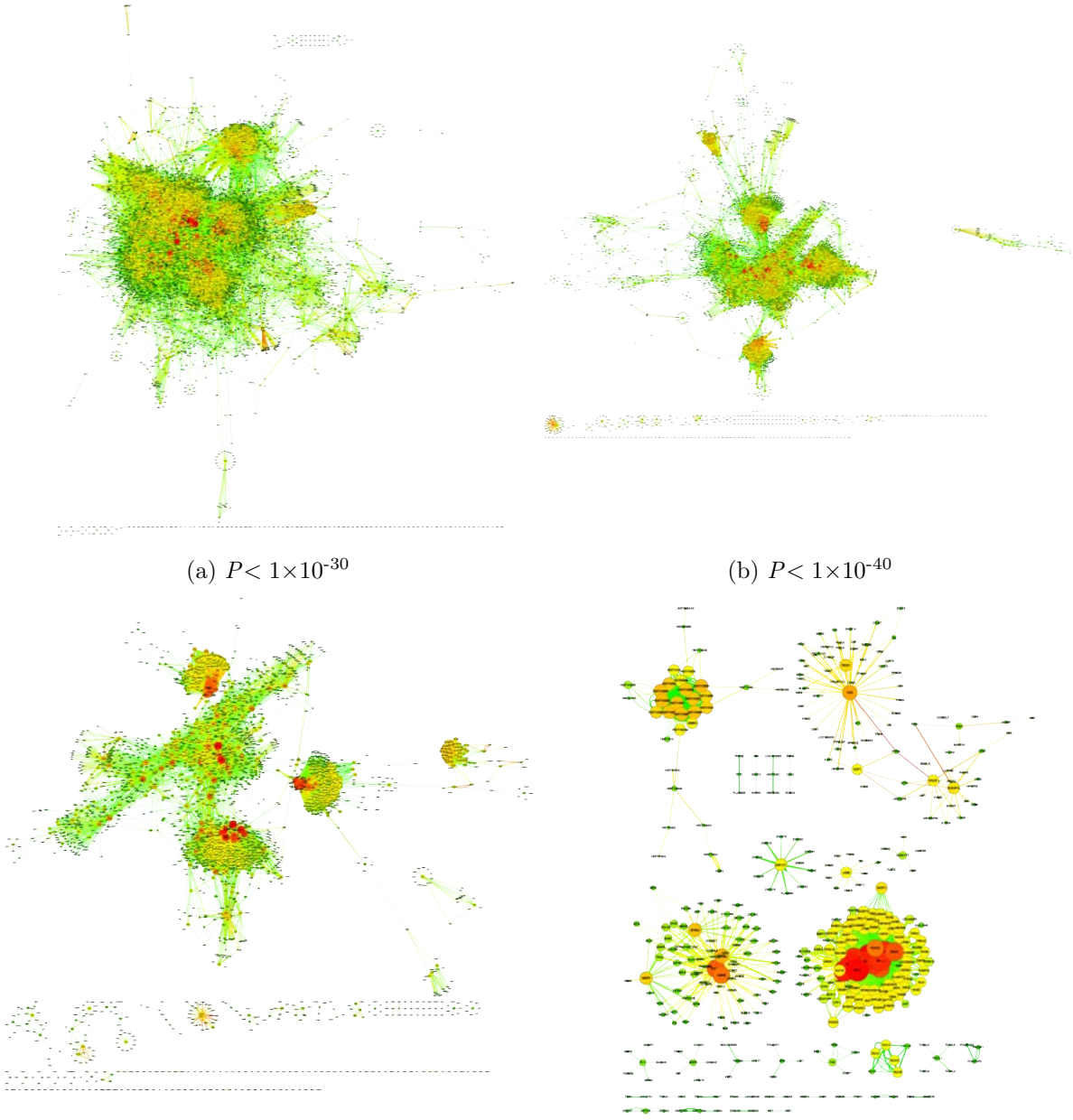


(a) $P< 1\times10^{-30}$ (b) $P< 1\times10^{-40}$

(c) $P< 1\times10^{-50}$ (d) $P< 1\times10^{-100}$

Figura 3-14: **Topología de las redes inferidas con ARACNe para el conjunto de datos de GEO.** El color de los nodos representa el grado siendo las de mayor grado de color rojo y las de menor verde. Todas las redes tienen disposición orgánica y fueron dibujadas con el software *Cytoscape*. Las redes en formato .adj fueron transformadas a un archivo .sif legible para *Cytoscape* con el script reproducido en el apéndice A.3

.



(a) $P < 1 \times 10^{-30}$

(b) $P < 1 \times 10^{-40}$

(c) $P < 1 \times 10^{-50}$

(d) $P < 1 \times 10^{-100}$

Figura 3-15: **Topología de las redes inferidas con ARACNe para el conjunto de datos de TCGA.** El color de los nodos representa el grado siendo las de mayor grado de color rojo y las de menor verde. Todas las redes tienen disposición orgánica y fueron dibujadas con el software *Cytoscape*. Las redes en formato .adj fueron transformadas a un archivo .sif legible para *Cytoscape* con el script reproducido en el apéndice A.3

.

3.6.2. Conjunto de muestras del GEO

Reguladores transcripcionales maestros inferidos con MARINa

Las siguientes figuras 3-16 3-17 y 3-18 muestran los primeros 10 mejores candidatos obtenidos con el algoritmo MARINa usando las subredes restantes obtenidas cortando los ejes con valores de P menores de 1×10^{-30}, 1×10^{-50} y 1×10^{-100} respectivamente. La Figura 3-19 muestra los *hiveplots* de estas subredes.

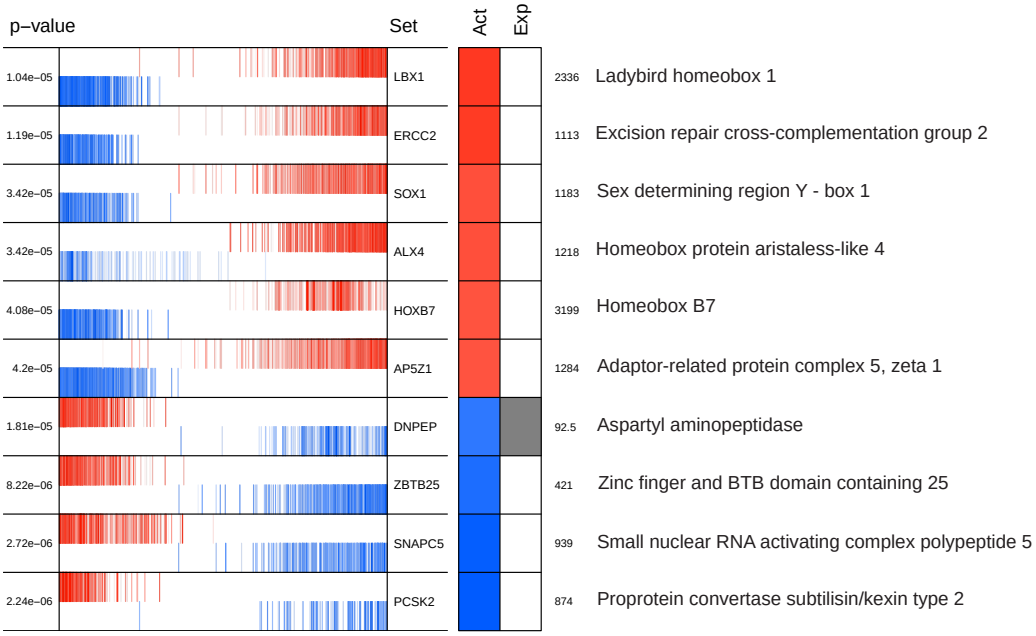


Figura 3-16: **Top 10 Reguladores Transcripcionales Maestros inferidos del conjunto de datos del GEO para la red de $P = 1 \times 10^{-30}$.** Este gráfico muestra la lista de los primeros diez mejores candidatos a RTMs (columna Set) en el conjunto de datos. El valor de P asociado a cada candidato a regulador transcripcional maestro se muestra a la izquierda. La actividad diferencial (Act) y la expresión diferencial (Exp) predicha por MARINa se muestran a la derecha. Los tonos rojo de Exp o Act muestran sobreexpresión y los azules subexpresión. El puntaje de Exp de cada candidato regulador transcripcional maestro también se muestra en el lado derecho de la gráfica. Las líneas rojas en el medio de la gráfica indican el número de objetivos de los RTM que predijo ARACNe. La posición de cada línea en el eje horizontal corresponde a su puntaje en la lista de genes que se determina por su expresión diferencial. Los genes con mayor sobrexpresión diferencial se muestran a la izquierda en rojo y los más diferencialmente subexpresados se muestran a la derecha en azul.

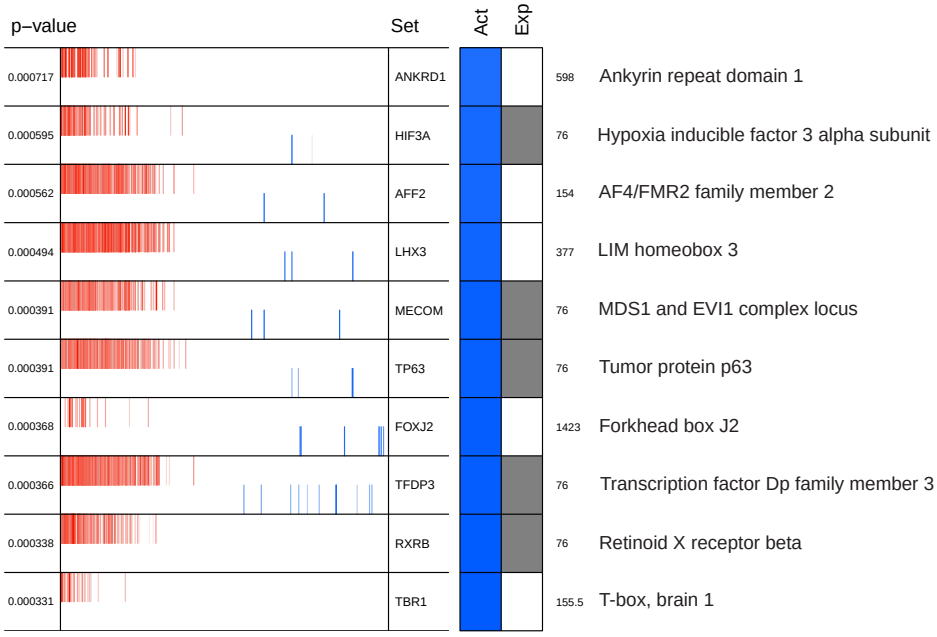


Figura 3-17: **Top 10 Reguladores Transcripcionales Maestros inferidos del conjunto de datos del GEO para la red de $P = 1 \times 10^{-50}$.** Este gráfico muestra la lista de los primeros diez mejores candidatos a RTMs (columna Set) en el conjunto de datos. El valor de P asociado a cada candidato a regulador transcripcional maestro se muestra a la izquierda. La actividad diferencial (Act) y la expresión diferencial (Exp) predicha por MARINa se muestran a la derecha. Los tonos rojo de Exp o Act muestran sobreexpresión y los azules subexpresión. El puntaje de Exp de cada candidato regulador transcripcional maestro también se muestra en el lado derecho de la gráfica. Las líneas rojas en el medio de la gráfica indican el número de objetivos de los RTM que predijo ARACNe. La posición de cada línea en el eje horizontal corresponde a su puntaje en la lista de genes que se determina por su expresión diferencial. Los genes con mayor sobrexpresión diferencial se muestran a la izquierda en rojo y los más diferencialmente subexpresados se muestran a la derecha en azul.

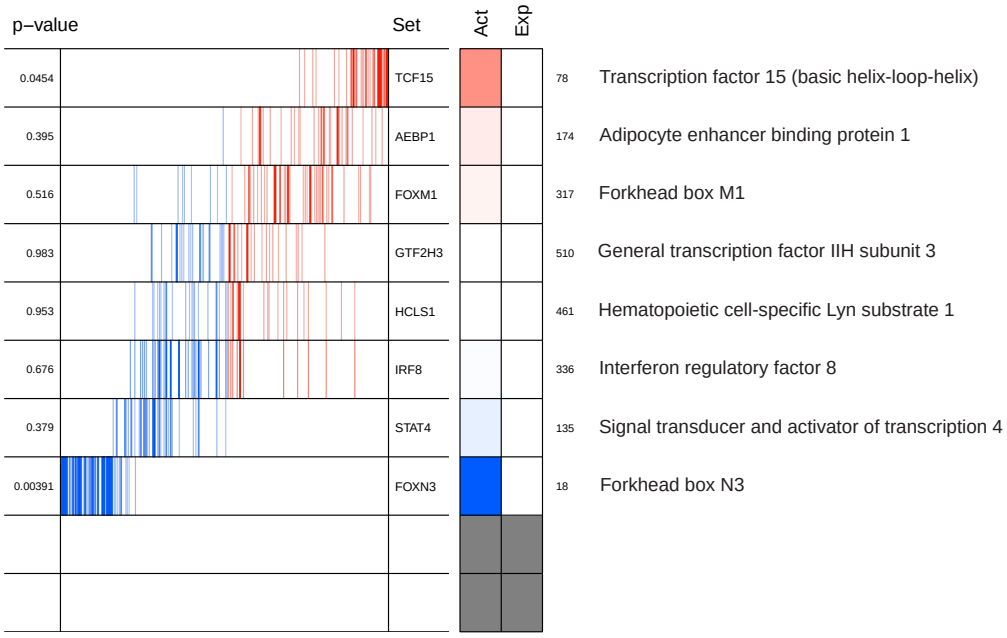
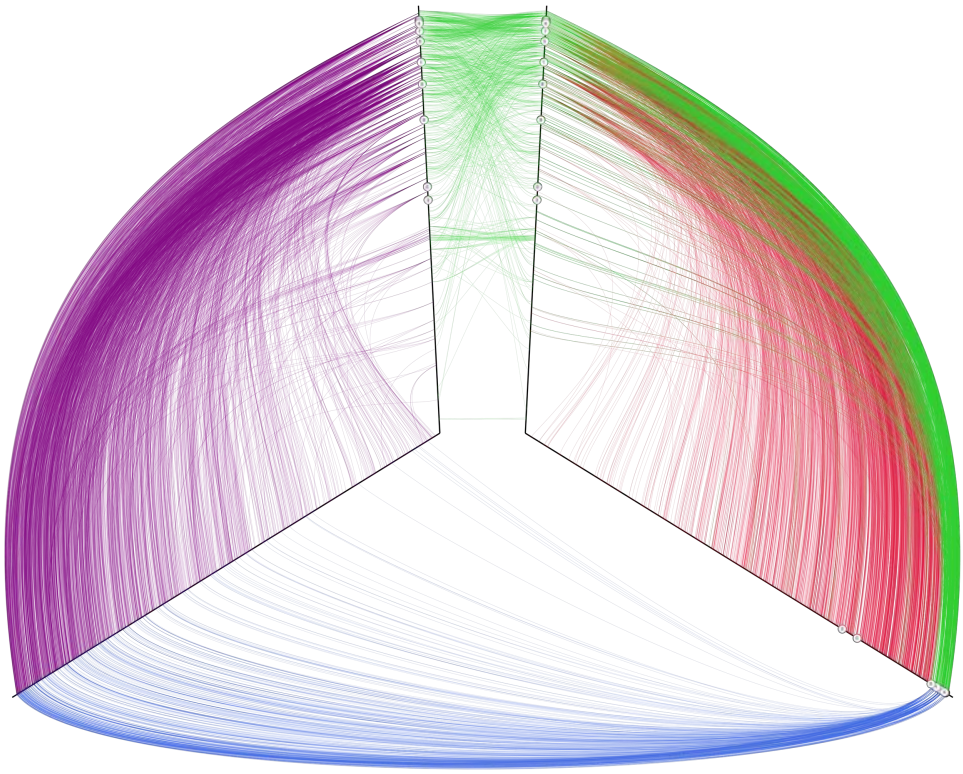
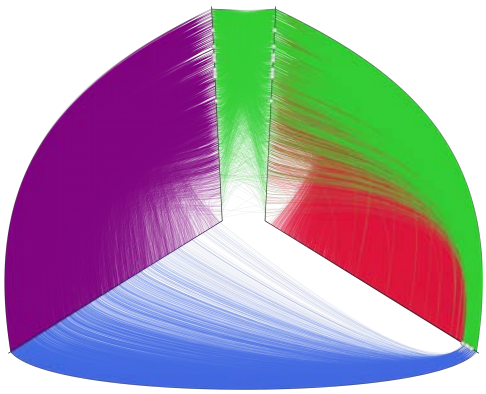


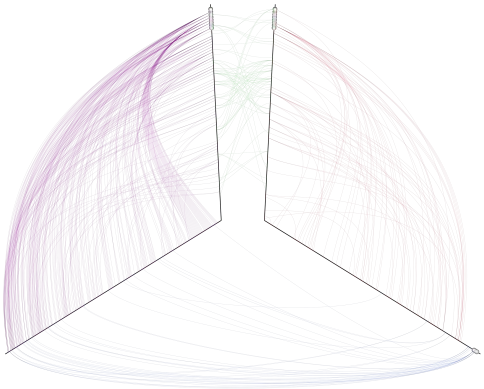
Figura 3-18: **Top 10 Reguladores Transcripcionales Maestros inferidos del conjunto de datos del GEO para la red de $P = 1 \times 10^{-100}$.** Este gráfico muestra la lista de los primero diez mejores candidatos a RTMs (columna Set) en el conjunto de datos. El valor de P asociado a cada candidato a regulador transcripcional maestro se muestra a la izquierda. La actividad diferencial (Act) y la expresión diferencial (Exp) predicha por MARINa se muestran a la derecha. Los tonos rojo de Exp o Act muestran sobreexpresión y los azules subexpresión. El puntaje de Exp de cada candidato regulador transcripcional maestro también se muestra en el lado derecho de la gráfica. Las líneas rojas en el medio de la gráfica indican el número de objetivos de los RTM que predijo ARACNe. La posición de cada línea en el eje horizontal corresponde a su puntaje en la lista de genes que se determina por su expresión diferencial. Los genes con mayor sobrexpresión diferencial se muestran a la izquierda en rojo y los más diferencialmente subexpresados se muestran a la derecha en azul.



(a) Hive Plot del regulón $P< 1\times10^{-50}$



(b) Hive Plot del regulón $P< 1\times10^{-30}$



(c) Hive Plot del regulón $P< 1\times10^{-100}$

Figura 3-19: **Hiveplots de los Conjuntos de Regulones para los datos de GEO** En las lineas rectas en negro están representados los nodos que, en este caso, son genes y las lineas curvas representan interacciones (dadas por su valor de información mutua). Ambos ejes verticales contienen la lista de los Factores de transcripción mientras que el eje derecho representa genes diferencialmente expresados entre sanos y enfermos. El eje izquierdo contiene los genes no diferenciados. La interacción entre dos Factores de transcripción se representa en verde. La interacción entre FT y un gen diferencialmente expresado en rojo, mientras que las curvas moradas representan interacciones entre FT y genes no diferencialmente expresados. Por último, las lineas azules representan las interacciones de aquellos FT que están diferencialmente expresados y los genes no diferencialmente expresados. Los números indican los genes RTMs del top 10 estimados con MARINa. Para una mejor visualización, las interacciones con una información mutua menor de 0.85 fueron eliminados excepto en la red de $P = 1 \times 10^{-100}$.

Resultados del análisis de sombreado

Las siguientes figuras 3-20, 3-21 y 3-22 presentan los resultados del análisis de sombras para las subredes restantes obtenidas cortando los ejes con valores de P menores de 1×10^{-30}, 1×10^{-50} y 1×10^{-100} respectivamente.

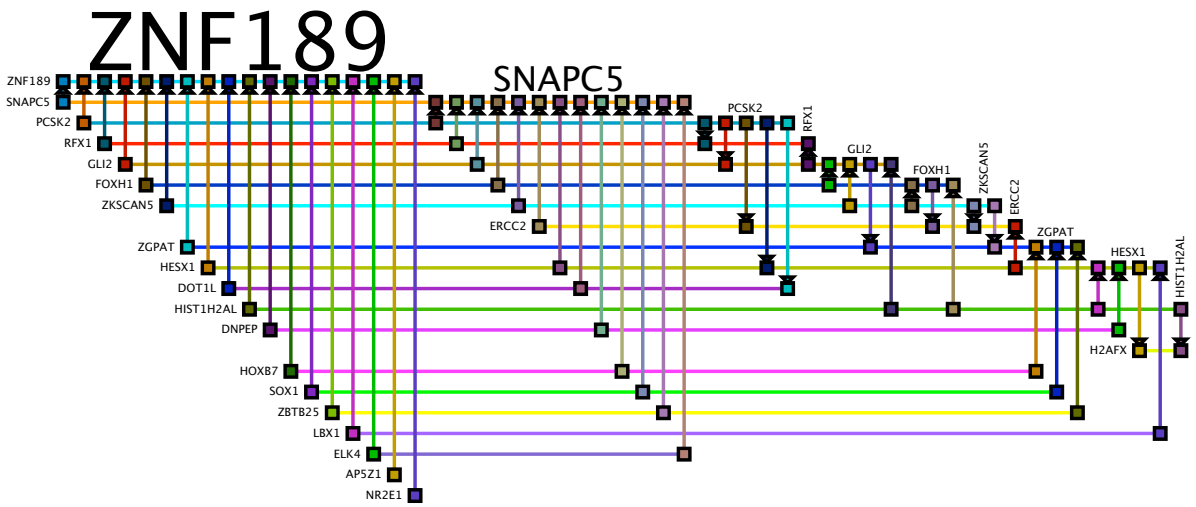


Figura 3-20: **Red generada con los pares de FT resultado del análisis de sombra para el conjunto de datos de GEO con $P< 1\times10^{-30}$.** En esta visualización (hecha con el software BioFabric [Longabaugh, 2012]) los nodos se representan como líneas horizontales, una por renglón, y los ejes son lineas verticales, una por columna. En estas redes, la dirección de las interacciones $\mathrm{FT}_{\mathrm{origen}} \rightarrow \mathrm{FT}_{\mathrm{blanco}}$ significa que el enriquecimiento de los genes no compartidos entre $\mathrm{FT}_{\mathrm{origen}}$ y $\mathrm{FT}_{\mathrm{blanco}}$ no es significativamente distinto. Una copia de esta figura en alta resolución puede ser descargada de https://goo.gl/FsEsxO.

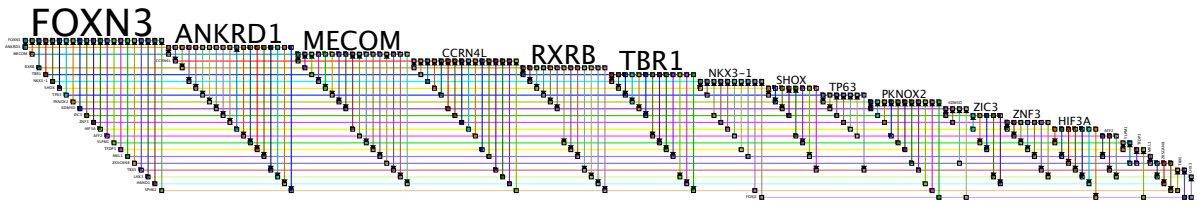


Figura 3-21: **Red generada con los pares de FT resultado del análisis de sombra para el conjunto de datos de GEO con $P< 1\times10^{-50}$.** En esta visualización (hecha con el software BioFabric [Longabaugh, 2012]) los nodos se representan como líneas horizontales, una por renglón, y los ejes son lineas verticales, una por columna. En estas redes, la dirección de las interacciones $\mathrm{FT}_{\mathrm{origen}} \rightarrow \mathrm{FT}_{\mathrm{blanco}}$ significa que el enriquecimiento de los genes no compartidos entre $\mathrm{FT}_{\mathrm{origen}}$ y $\mathrm{FT}_{\mathrm{blanco}}$ no es significativamente distinto. Una copia de esta figura en alta resolución puede ser descargada de https://goo.gl/53EXqk.

Figura 3-22: **Red generada con los pares de FT resultado del análisis de sombra para el conjunto de datos de GEO con $P < 1 \times 10^{-100}$.** En esta visualización (hecha con el software BioFabric [Longabaugh, 2012]) los nodos se representan como líneas horizontales, una por renglón, y los ejes son lineas verticales, una por columna. En estas redes, la dirección de las interacciones $\mathrm{FT}_{\mathrm{origen}} \rightarrow \mathrm{FT}_{\mathrm{blanco}}$ significa que el enriquecimiento de los genes no compartidos entre $\mathrm{FT}_{\mathrm{origen}}$ y $\mathrm{FT}_{\mathrm{blanco}}$ no es significativamente distinto. Una copia de esta figura en alta resolución puede ser descargada de https://goo.gl/7DAi4a.

Resultados del análisis de sinergia

Las siguientes figuras 3-23, 3-24 y 3-25 muestra los grupos de factores de transcripción que actúan sinergicamente en las subredes restantes obtenidas cortando los ejes con valores de P menores de 1×10^{-30}, 1×10^{-50} y 1×10^{-100} respectivamente.

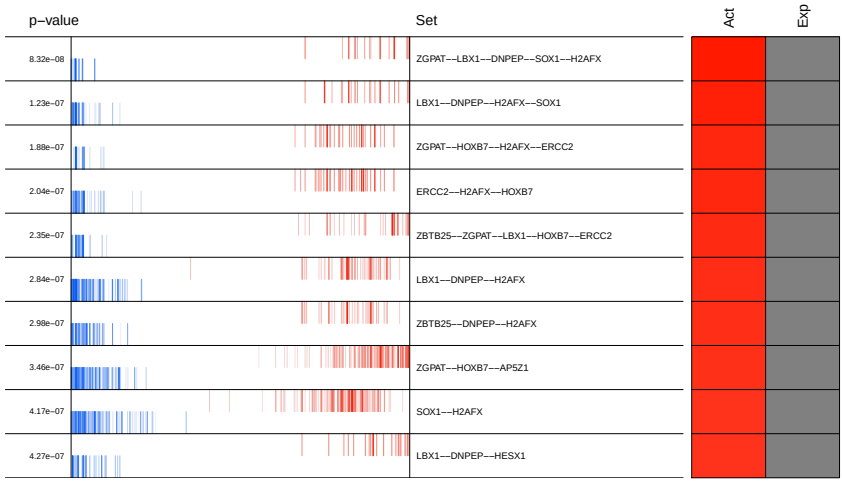


Figura 3-23: **Gráficos del Análisis de Sinergia para el conjunto de GEO con $P< 1\times10^{-30}$.** Se muestran los primeros 10 conjuntos de genes reguladores maestros que regulan sinérgicamente conjuntos de genes de la FM en nuestro conjunto de datos para las cuatro redes. La columna *Set* muestra aquellos genes que corregulan sinérgicamente el mismo conjunto de blancos. Los 10 genes involucrados en el fenómeno de sinergia fueron *zinc finger CCCH-type and G-patch domain containing* (ZGPAT), *ladybird homeobox 1* (LBX1), *aspartyl aminopeptidase* (DNPEP), *sex determining region Y-box 1* (SOX1), *H2A histone family member X* (H2AFX), *homeobox B7* (HOXB7), *excision repair cross-complementation group 2* (ERCC2), *zinc finger and BTB domain containing 25* (ZBTB25), *adaptor related protein complex 5 zeta 1 subunit* (AP5Z1) y *HESX homeobox 1* (HESX1).

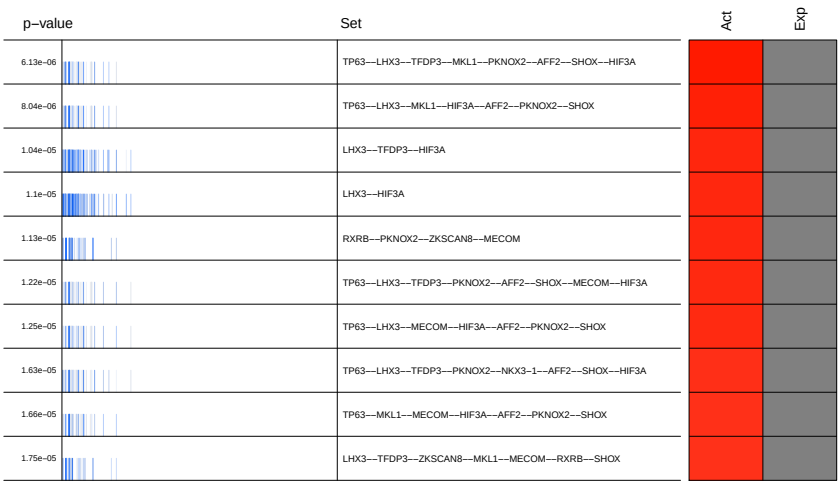


Figura 3-24: **Gráficos del Análisis de Sinergia para el conjunto de GEO con $P < 1 \times 10^{-50}$.** Se muestran los primeros 10 conjuntos de genes reguladores maestros que regulan sinérgicamente conjuntos de genes de la FM en nuestro conjunto de datos para las cuatro redes. La columna *Set* muestra aquellos genes que corregulan sinérgicamente el mismo conjunto de blancos. Los 12 genes involucrados en el fenómeno de sinergia fueron *tumor protein p63* (TP63), *LIM homeobox 3* (LHX3), *transcription factor Dp family member 3* (TFDP3), *megakaryoblastic leukemia (translocation) 1* (MKL1), *PBX/knotted 1 homeobox 2* (PKNOX2), *AF4/FMR2 family member 2* (AFF2), *short stature homeobox* (SHOX), *hypoxia inducible factor 3 alpha subunit* (HIF3A), r*etinoid X receptor beta* (RXRB), *zinc finger with KRAB and SCAN domains 8* (ZKSCAN8), *MDS1 and EVI1 complex locus* (MECOM) y *NK3 homeobox 1* (NKX3-1).

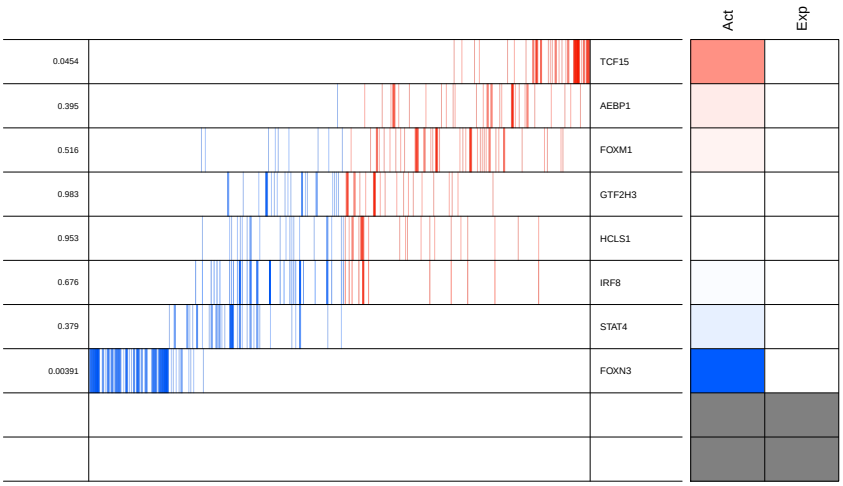


Figura 3-25: **Gráficos del Análisis de Sinergia para el conjunto de GEO con $P < 1 \times 10^{-100}$.** Se muestran los primeros 10 conjuntos de genes reguladores maestros que regulan sinérgicamente conjuntos de genes de la FM en nuestro conjunto de datos para las cuatro redes. Los ocho genes involucrados en el fenómeno de sinergia fueron *transcription factor 15 (basic helix-loop-helix)* (TCF15), *AE binding protein 1* (AEBP1), *forkhead box M1* (FOXM1), *general transcription factor IIH subunit 3* (GTF2H3), *hematopoietic cell-specific Lyn substrate 1* (HCLS1), *interferon regulatory factor 8* (IRF8), *signal transducer and activator of transcription 4* (STAT4) y *forkhead box N3* (FOXN3).

3.6.3. Conjunto de muestras del TCGA

Reguladores transcripcionales maestros inferidos con MARINa

Las siguientes figuras 3-26, 3-27 y 3-28 muestran los primeros 10 mejores candidatos obtenidos con el algoritmo MARINa usando las subredes restantes obtenidas cortando los ejes con valores de P menores de 1×10^{-30}, 1×10^{-50} y 1×10^{-100} respectivamente. La Figura 3-29 muestra los *hiveplots* de estas subredes.

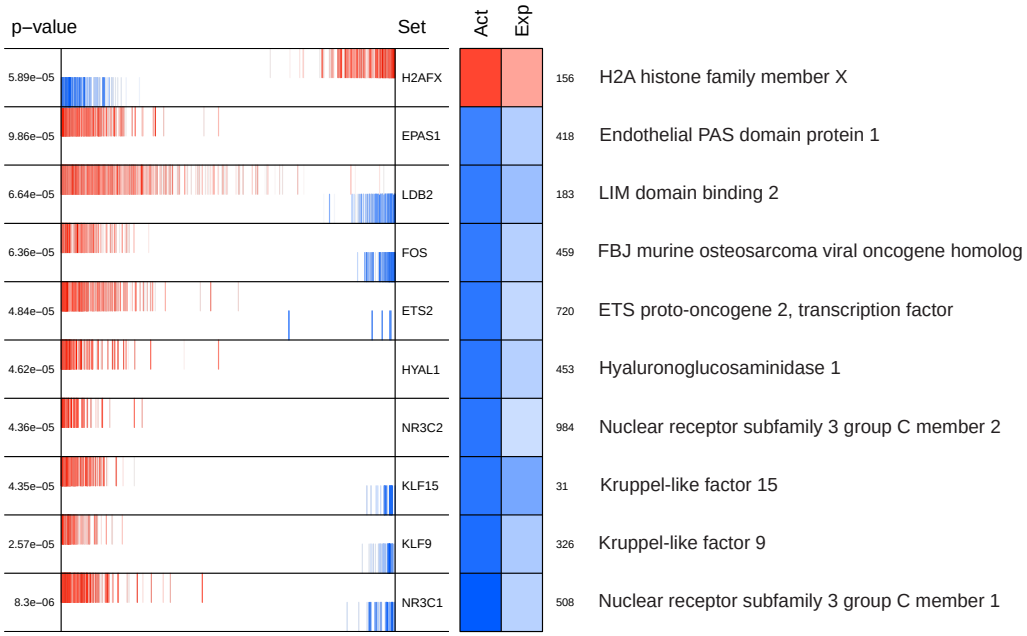


Figura 3-26: **Top 10 Reguladores Transcripcionales Maestros de las muestras de tejido de cáncer de mama para TCGA con $P< 1\times10^{-30}$.** Este gráfico muestra la lista de los primero diez mejores candidatos a RTMs (columna Set) en el conjunto de datos. El valor de P asociado a cada candidato a regulador transcripcional maestro se muestra a la izquierda. La actividad diferencial (Act) y la expresión diferencial (Exp) predicha por MARINa se muestran a la derecha. Los tonos rojo de Exp o Act muestran sobreexpresión y los azules subexpresión. El puntaje de Exp de cada candidato regulador transcripcional maestro también se muestra en el lado derecho de la gráfica. Las líneas rojas en el medio de la gráfica indican el número de objetivos de los RTM que predijo ARACNe. La posición de cada línea en el eje horizontal corresponde a su puntaje en la lista de genes que se determina por su expresión diferencial. Los genes con mayor sobrexpresión diferencial se muestran a la izquierda en rojo y los más diferencialmente subexpresados se muestran a la derecha en azul.

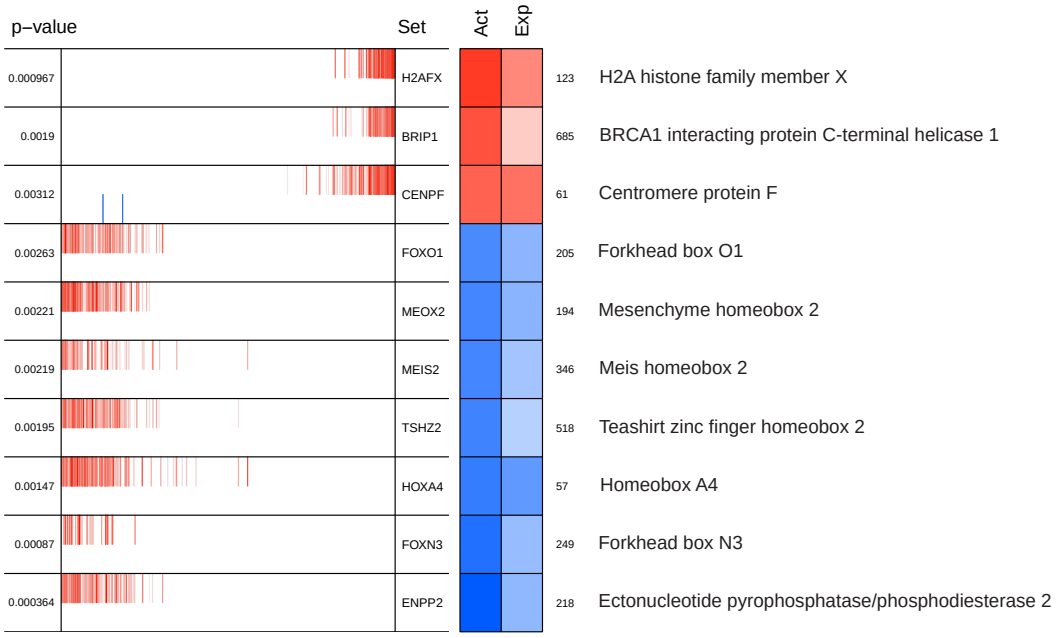


Figura 3-27: **Top 10 Reguladores Transcripcionales Maestros de las muestras de tejido de cáncer de mama para TCGA con $P < 1 \times 10^{-50}$.** Este gráfico muestra la lista de los primero diez mejores candidatos a RTMs (columna Set) en el conjunto de datos. El valor de P asociado a cada candidato a regulador transcripcional maestro se muestra a la izquierda. La actividad diferencial (Act) y la expresión diferencial (Exp) predicha por MARINa se muestran a la derecha. Los tonos rojo de Exp o Act muestran sobreexpresión y los azules subexpresión. El puntaje de Exp de cada candidato regulador transcripcional maestro también se muestra en el lado derecho de la gráfica. Las líneas rojas en el medio de la gráfica indican el número de objetivos de los RTM que predijo ARACNe. La posición de cada línea en el eje horizontal corresponde a su puntaje en la lista de genes que se determina por su expresión diferencial. Los genes con mayor sobrexpresión diferencial se muestran a la izquierda en rojo y los más diferencialmente subexpresados se muestran a la derecha en azul.

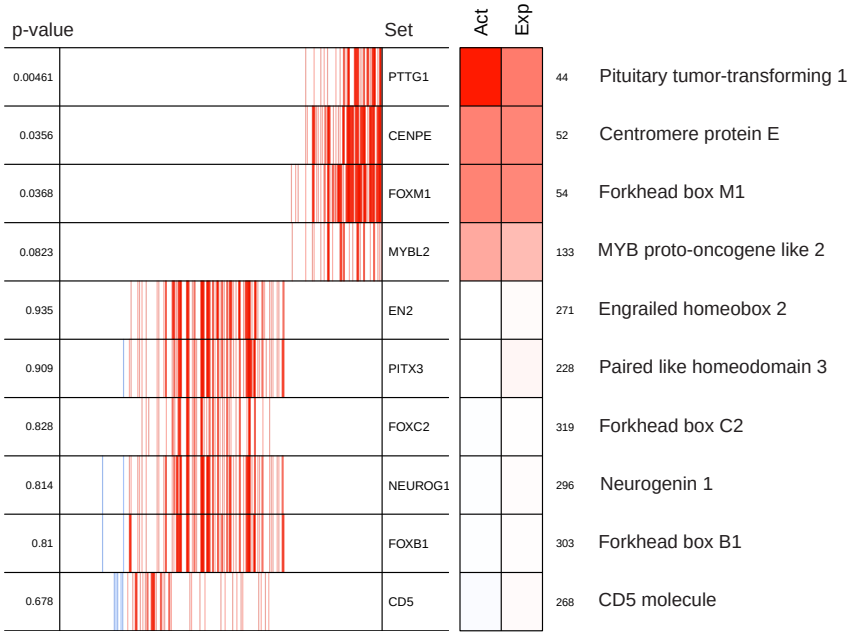
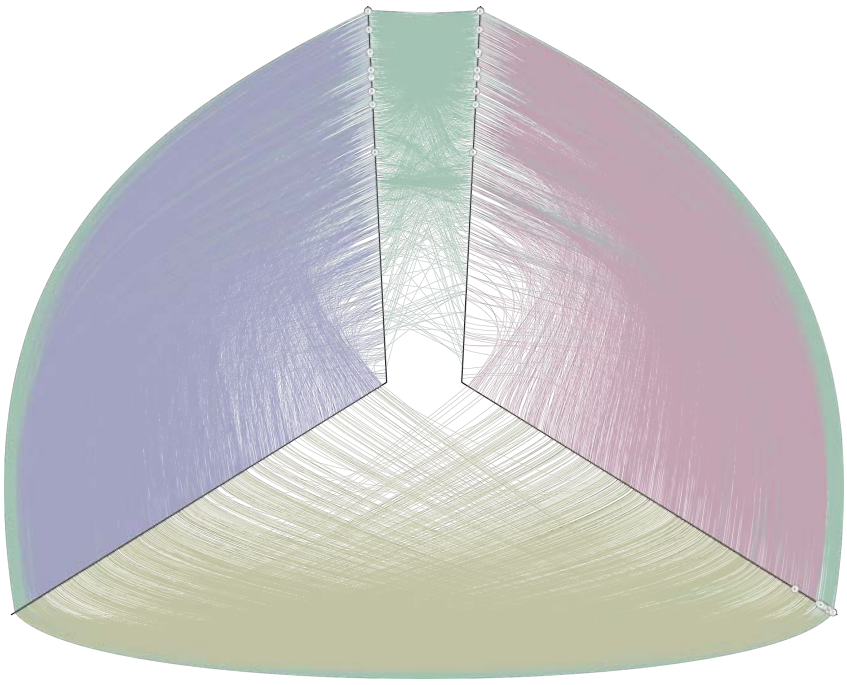
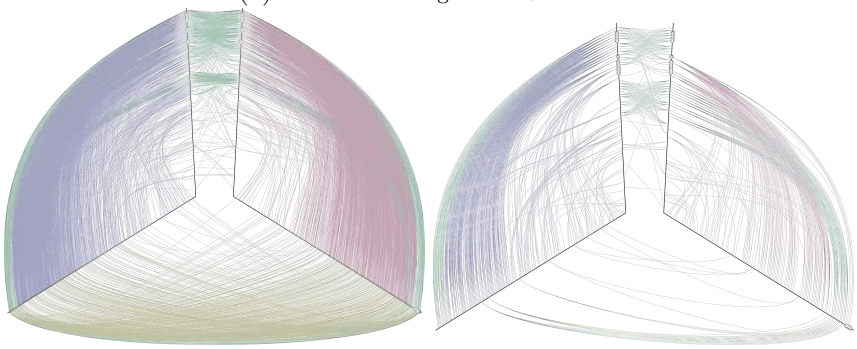


Figura 3-28: **Top 10 Reguladores Transcripcionales Maestros de las muestras de tejido de cáncer de mama para TCGA con $P < 1 \times 10^{-30}$.** Este gráfico muestra la lista de los primero diez mejores candidatos a RTMs (columna Set) en el conjunto de datos. El valor de P asociado a cada candidato a regulador transcripcional maestro se muestra a la izquierda. La actividad diferencial (Act) y la expresión diferencial (Exp) predicha por MARINa se muestran a la derecha. Los tonos rojo de Exp o Act muestran sobreexpresión y los azules subexpresión. El puntaje de Exp de cada candidato regulador transcripcional maestro también se muestra en el lado derecho de la gráfica. Las líneas rojas en el medio de la gráfica indican el número de objetivos de los RTM que predijo ARACNe. La posición de cada línea en el eje horizontal corresponde a su puntaje en la lista de genes que se determina por su expresión diferencial. Los genes con mayor sobrexpresión diferencial se muestran a la izquierda en rojo y los más diferencialmente subexpresados se muestran a la derecha en azul.



(a) Hive Plot del regulón $P< 1\times10^{-30}$



(b) Hive Plot del regulón $P< 1\times10^{-50}$ (c) Hive Plot del regulón $P< 1\times10^{-100}$

Figura 3-29: **Hiveplots de los Conjuntos de Regulones para los datos de TCGA** En las lineas rectas en negro están representados los nodos que, en este caso, son genes y las lineas curvas representan interacciones (dada por su valor de información mutua). Ambos ejes verticales contienen la lista de los Factores de transcripción mientras que el eje derecho representa genes diferencialmente expresados entre sanos y enfermos. El eje izquierdo contiene los genes no diferenciados. La interacción entre dos Factores de transcripción se representa en verde. La interacción entre FT y un gen diferencialmente expresado en rojo, mientras que las curvas moradas representan interacciones entre FT y genes no diferencialmente expresados. Por último, las lineas azules representan las interacciones de aquellos FT que están diferencialmente expresados y los genes no diferencialmente expresados. Los números indican los genes RTMs del top 10 estimados con MARINa. Para una mejor visualización, las interacciones con una información mutua menor de 0.85 fueron eliminados excepto en la red de $P< 1\times10^{-100}$.

Resultados del análisis de sombreado

Las siguientes figuras 3-30, 3-31 y 3-32 presentan los resultados del análisis de sombras para las subredes restantes obtenidas cortando los ejes con valores de P menores de 1×10^{-30}, 1×10^{-50} y 1×10^{-100} respectivamente.

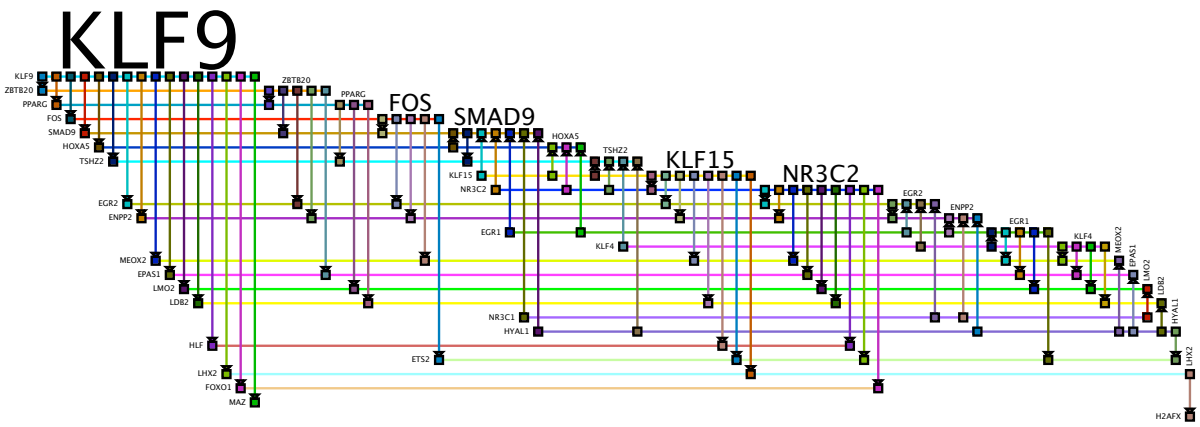


Figura 3-30: **Red generada con los pares de FT resultado del análisis de sombra para el conjunto de datos de TCGA con $P< 1\times10^{-30}$.** En esta visualización (hecha con el software BioFabric [Longabaugh, 2012]) los nodos se representan como líneas horizontales, una por renglón, y los ejes son lineas verticales, una por columna. En estas redes, la dirección de las interacciones $\mathrm{FT}_{\mathrm{origen}} \rightarrow \mathrm{FT}_{\mathrm{blanco}}$ significa que el enriquecimiento de los genes no compartidos entre $\mathrm{FT}_{\mathrm{origen}}$ y $\mathrm{FT}_{\mathrm{blanco}}$ no es significativamente distinto. Una copia de esta figura en alta resolución puede ser descargada de https://goo.gl/51TDFn

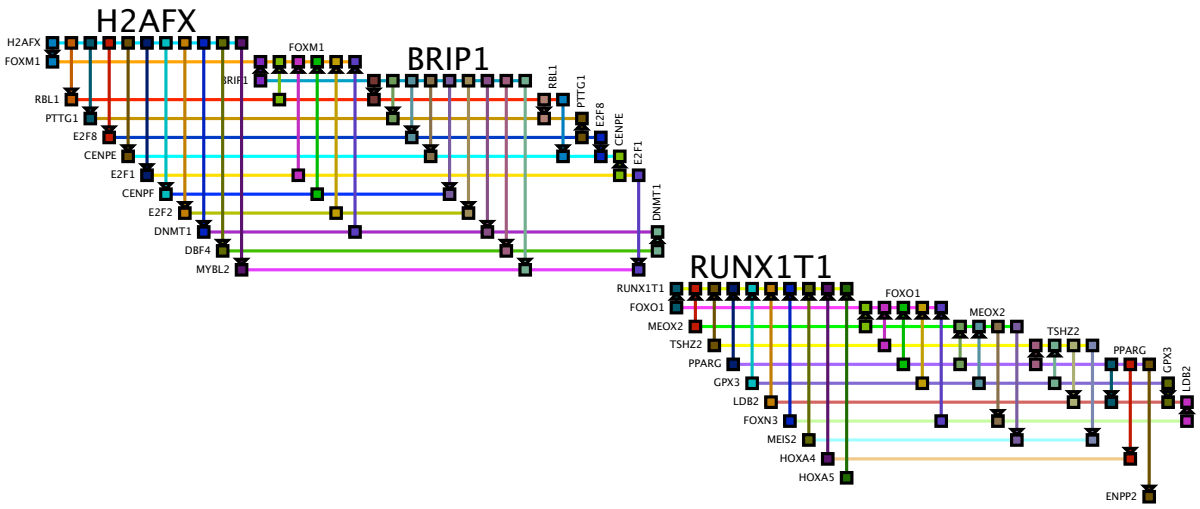


Figura 3-31: **Red generada con los pares de FT resultado del análisis de sombra para el conjunto de datos de TCGA con $P < 1 \times 10^{-50}$.** En esta visualización (hecha con el software BioFabric [Longabaugh, 2012]) los nodos se representan como líneas horizontales, una por renglón, y los ejes son lineas verticales, una por columna. En estas redes, la dirección de las interacciones $\text{FT}_{\text{origen}} \rightarrow \text{FT}_{\text{blanco}}$ significa que el enriquecimiento de los genes no compartidos entre $\text{FT}_{\text{origen}}$ y $\text{FT}_{\text{blanco}}$ no es significativamente distinto. Una copia de esta figura en alta resolución puede ser descargada de https://goo.gl/EmjNCt

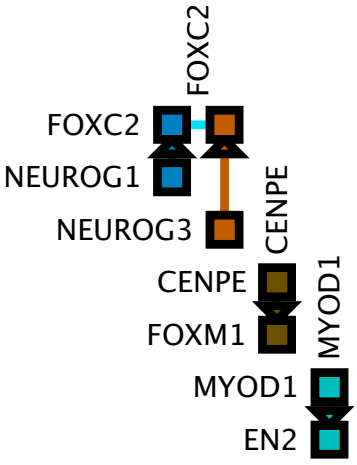


Figura 3-32: **Red generada con los pares de FT resultado del análisis de sombra para el conjunto de datos de TCGA con $P < 1 \times 10^{-100}$.** En esta visualización (hecha con el software BioFabric [Longabaugh, 2012]) los nodos se representan como líneas horizontales, una por renglón, y los ejes son lineas verticales, una por columna. En estas redes, la dirección de las interacciones $\text{FT}_{\text{origen}} \rightarrow \text{FT}_{\text{blanco}}$ significa que el enriquecimiento de los genes no compartidos entre $\text{FT}_{\text{origen}}$ y $\text{FT}_{\text{blanco}}$ no es significativamente distinto. Una copia de esta figura en alta resolución puede ser descargada de https://goo.gl/85Rg48

Resultados del análisis de sinergia

Las siguientes figuras 3-33, 3-34 y 3-35 muestra los grupos de factores de transcripción que actúan sinergicamente en las subredes restantes obtenidas cortando los ejes con valores de P menores de 1×10^{-30}, 1×10^{-50} y 1×10^{-100} respectivamente.

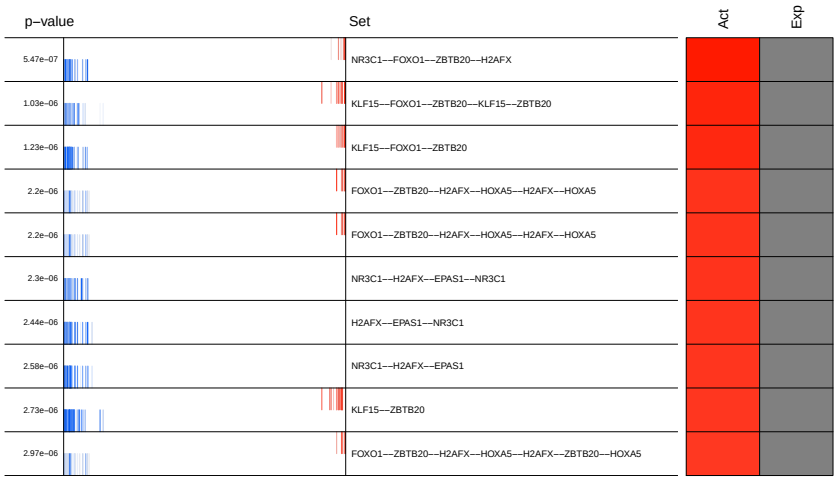


Figura 3-33: **Gráficos del Análisis de Sinergia para el conjunto de TCGA con una P de corte $< 1\times\mathbf{10^{-30}}$.** Se muestran los primeros 10 conjuntos de genes reguladores maestros que regulan sinérgicamente conjuntos de genes de la FM en nuestro conjunto de datos para las cuatro redes. La columna *Set* muestra aquellos genes que corregulan sinérgicamente el mismo conjunto de blancos. Los seis genes involucrados en el fenómeno de sinergia fueron *nuclear receptor subfamily 3 group C member 1* (NR3C1), *forkhead box O1* (FOXO1), *zinc finger and BTB domain containing 20* (ZBTB20), *H2A histone family member X* (H2AFX), *Kruppel-like factor 15* (KLF15) y *endothelial PAS domain protein 1* (EPAS1).

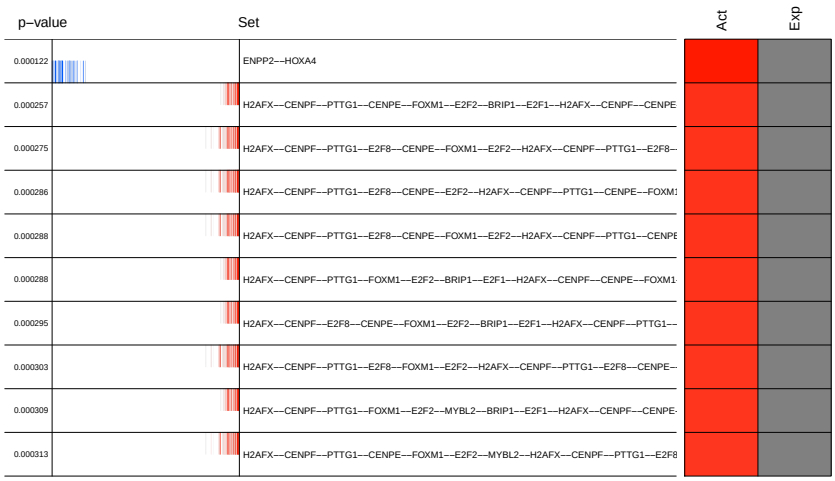


Figura 3-34: **Gráficos del Análisis de Sinergia para el conjunto de TCGA para una red con una *P* de corte $< 1\times\mathbf{10^{-50}}$.** Se muestran los primeros 10 conjuntos de genes reguladores maestros que regulan sinérgicamente conjuntos de genes de la FM en nuestro conjunto de datos para las cuatro redes. La columna *Set* muestra aquellos genes que corregulan sinérgicamente el mismo conjunto de blancos. Los 13 genes involucrados en el fenómeno de sinergia fueron *ectonucleotide pyrophosphatase/phosphodiesterase 2* (ENPP2), *homeobox A5* (HOXA5), *H2A histone family member X* (H2AFX), *centromere protein F* (CENPF), *centromere protein E* (CEMPE), *pituitary tumor-transforming 1* (PTTG1), *forkhead box M1* (FOXM1), *E2F transcription factor 2* (E2F2), *BRCA1 interacting protein C-terminal helicase 1* (BRIP1), *E2F transcription factor 1* (E2F1), *H2A histone family member X* (H2AFX), *E2F transcription factor 8* (E2F8) y *MYB proto-oncogene like 2* (MYBL2).

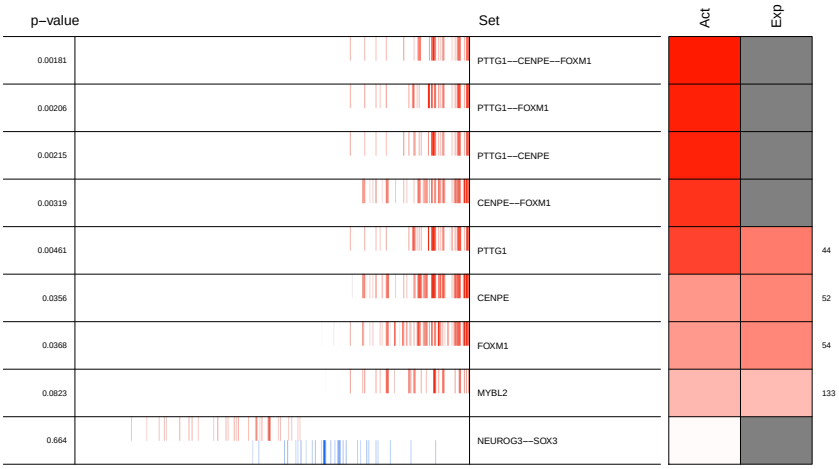


Figura 3-35: **Gráficos del Análisis de Sinergia para el conjunto de TCGA con una *P* de corte $< 1\times\mathbf{10^{-100}}$.** Se muestran los primeros 10 conjuntos de genes reguladores maestros que regulan sinérgicamente conjuntos de genes de la FM en nuestro conjunto de datos para las cuatro redes. En esta red el algoritmo solo pudo generar 9 conjuntos. La columna *Set* muestra aquellos genes que corregulan sinérgicamente el mismo conjunto de blancos. Los seis genes involucrados en el fenómeno de sinergia fueron *pituitary tumor-transforming 1* (PTTG1), *centromere protein E* (CEMPE), *forkhead box M1* (FOXM1), *MYB proto-oncogene like 2* (MYBL2), *neurogenin 3* (NEUROG3) y *sex determining region Y-box 3* (SOX3).

Capítulo 4

Conclusiones

En este estudio se implementó un método basado en la combinación de los algoritmos de la inferencia de redes de regulación génica y el análisis de enriquecimiento genético y otros más como firmas moleculares. Todo esto usando conjuntos de experimentos de expresión génica que cuentan con casos y controles con los cuales se pueden inferir firmas moleculares. Los casos fueron muestras pertenecientes a tejido primario de cáncer de mama obtenido de biopsias mientras que los controles fueron de tejido de mama sano. Este flujo de trabajo nos permitió develar una serie de fenómenos de la regulación transcripcional que probablemente pueden estar detrás del establecimiento del fenotipo tumoral. Por ejemplo, fue posible recobrar vías enriquecidas funcionalmente relacionadas con cáncer de la red conformadas por un grupo de reguladores transcripcionales maestros y sus blancos directos. Es de destacar que, incluso en los casos en los cuales la anotación de los factores de transcripción fue incorrecta, los genes resultantes fueron de moléculas importantes en los *hallmarks* del cáncer.

Otro elemento importante que hay que resaltar es el hecho de que en este trabajo se desarrolló toda la metodología en muestras de cáncer de mama no clasificado por subtipos moleculares. El enfoque de esta metodología radica en la búsqueda de diferencias entre los casos y controles. Ya que creemos que una descripción más precisa de los fenómenos se mejorará con muestras subtipificadas, el siguiente paso en esta línea de investigación es, precisamente, separar las muestras por subtipo molecular. Por ejemplo, en el análisis de redes causales, se encontró que lo que ellos definen como la vía canónica *"Mecanismos moleculares del cáncer"*

(https://goo.gl/fOYDVg) está activa significativamente.

En cuanto a la estructura de la red de regulación génica inducida por la acción de reguladores transcripcionales maestros, se observaron dos fenómenos interesantes y relacionados. El primero, denominado sombreado, se refiere al hecho de que hay algunos reguladores maestros que son capaces de controlar la transcripción de los objetivos de otros Factores de Transcripción. Y el segundo la acción sinérgica de algunos reguladores transcripcionales maestros los cuales pusieron de manifiesto módulos funcionales, indicativos de algunos de los *hallmarks* del cáncer, como son la proliferación, la evasión de la apoptosis y la invasividad. Esta podría ser una razón por la cual parece tan robusto el programa de regulación de la enfermedad.

De todo lo anterior podemos destacar estas conclusiones:

1. Se infirieron reguladores transcripcionales maestros usando ambos conjuntos de datos de expresión génica (Tabla 4-1). Todos ellos o con una sabida importancia en los procesos cancerígenos o con muy poca anotación (lo que nos hace interesantes blancos de estudio).

Tabla 4-1: Primeros diez mejores candidatos a reguladores transcripcionales maestros inferidos con las redes con valores de corte a $P < 1 \times 10^{-40}$ tanto para el conjunto de GEO como de TCGA.

Lugar	GEO	TCGA
1	*Angiotensin II receptor type 2*	*Ectonucleotide pyrophosphatase/phosphodiesterase 2*
2	*Zinc finger protein 132*	*Kruppel-like factor 15*
3	*Zinc finger protein 3*	*Zinc finger and BTB domain containing 20*
4	*IKAROS family zinc finger 3*	*Forkhead box N3*
5	*Forkhead box J2*	*Homeobox A5*
6	*Zic family member 3*	*Zinc finger and BTB domain containing 16*
7	*T-box, brain 1*	*Homeobox A4*
8	*Transcription factor Dp family member 3*	*Meis homeobox 2*
9	*Hypoxia inducible factor 3 alpha subunit*	*H2A histone family member X*
10	*Lysine acetyltransferase 7*	*GA binding protein transcription factor beta subunit 2*

2. Se evidenció el carácter jerárquico de la regulación transcripcional en el control de la expresión. Dicho carácter se observó en los resultados de ambos conjuntos de datos.

3. El análisis de sombras puso en evidencia lo imbricado de la interacción transcripcional.

4. Se observó la actividad sinérgica sobre algunos conjuntos de genes denominados Grupos de Genes Reguladores Maestros que muestran la acción simultánea de conjuntos de

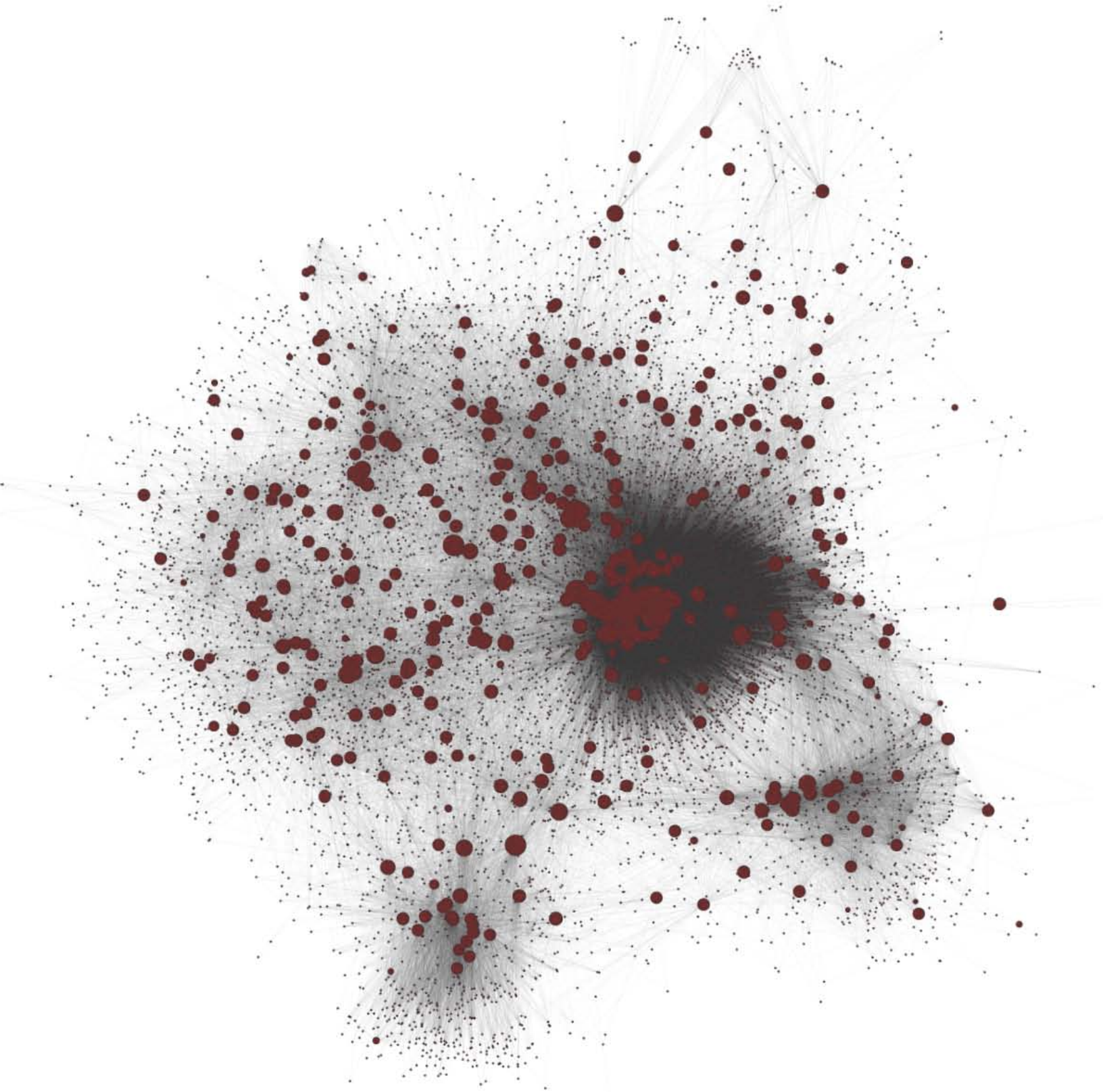
Reguladores Maestros tope sobre grandes conjuntos de blancos que son colectivamente regulados por éstos.

5. Junto con el sombreado, este último fenómeno puede estar contribuyendo a la creación y mantenimiento de condiciones robustas en el contexto de los fenotipos asociados a tumores como ya se discutió.

De algunos años a la fecha, una serie de adelantos tecnológicos ha sorprendido a la comunidad cientifica. A través de éstos desarrollos ahora somos capaces de generar una cantidades nunca antes vista de datos extraídos del análisis de los fenómenos naturales. La biología molecular es un testigo de primera línea de esta oleada de información. Este hecho ofrece la posibilidad de explorar los fenómenos de interés desde nuevos ángulos de acercamiento. Juntos, estas nuevas herramientas y los enfoque clásicos, nos permiten fijar nuestra atención en fenómenos muy específicos a un alto nivel de detalle. Básicamente las nuevas maneras de ver a un fenómeno, mediante la consideración de grandes conjuntos de datos generan hipótesis que pueden ser corroboradas de forma clásica, esto es en el laboratorio básico, en modelos animlaes y posteriormente en la clínica.

Sin embargo, los grandes conjuntos de datos conllevan la necesidad de resolver el problema de reducir su complejidad para generar dichas hipótesis. Explorar las opciones para hacer esta reducción de complejidad es pues fundamental para poder hacer esta integración. En este trabajo recuperamos la información de los niveles de expresión de un gran número de genes provenientes de experimentos de microarreglos. Se generaron redes de regulación génica a partir de éstos y se infirieron reguladores transcripcionales maestros basados en la importancia de los blancos de cada factor de transcripción en la firma molecular. Por ello, esta exploración sugiere que el Algoritmo de Inferencia de Reguladores Maestros (MARINa) puede convertirse en una herramienta metodológica muy importante en el estudio molecular de fenotipos celulares complejos, particularmente aquellos relacionados con enfermedades complejas como el cáncer. Por supuesto es mucho el camino que ha de recorrerse para que estos novedosos enfoques nos brinden una comprensión global de fenómenos biológicos tan complejos; sin embargo, los trabajos recientes, entre los que se encuentra este, son suficientemente promisorios para considerarlos

como parte de las herramientas que la genómica moderna requiere.



Literatura citada

Abduljabbar, R., Al-Kaabi, M.M., Negm, O.H., Jerjees, D., Muftah, A.A., Mukherjee, A., Lai, C.F., Buluwela, L., Ali, S., Tighe, P.J., Green, A., Ellis, I., y Rakha, E. Prognostic and biological significance of peroxisome proliferator-activated receptor-gamma in luminal breast cancer. *Breast Cancer Res. Treat.* **150**(3):511–522 (2015)

Abildgaard, M.O., Borre, M., Mortensen, M.M., Ulhøi, B.P., Tørring, N., Wild, P., Kristensen, H., Mansilla, F., Ottosen, P.D., Dyrskjøt, L., Ørntoft, T.F., y Sørensen, K.D. Downregulation of zinc finger protein 132 in prostate cancer is associated with aberrant promoter hypermethylation and poor prognosis. *Int. J. Cancer* **130**(4):885–895 (2012)

Affara, M., Sanders, D., Araki, H., Tamada, Y., Dunmore, B.J., Humphreys, S., Imoto, S., Savoie, C., Miyano, S., Kuhara, S., Jeffries, D., Print, C., y Charnock-Jones, D.S. Vasohibin-1 is identified as a master-regulator of endothelial cell apoptosis using gene network analysis. *BMC Genomics* **14**(1):1–1 (2013)

Alvarez, M.J. *ssmarina: Single sample-optimized Master Regulator Analysis* (2013). R package version 1.01

American Cancer Society. Global Cancer Facts & figures (2015)

Ando, H., Natsume, A., Iwami, K., Ohka, F., Kuchimaru, T., Kizaka-Kondoh, S., Ito, K., Saito, K., Sugita, S., Hoshino, T., y Wakabayashi, T. A hypoxia-inducible factor (HIF)-3 splicing variant, HIF-34 impairs angiogenesis in hypervascular malignant meningiomas with epigenetically silenced HIF-34. *Biochem. Biophys. Res. Commun.* **433**(1):139–144 (2013)

Aytes, A., Mitrofanova, A., Lefebvre, C., Alvarez, M.J., Castillo-Martin, M., Zheng, T., Eastham, J.A., Gopalan, A., Pienta, K.J., Shen, M.M., Califano, A., y Abate-Shen, C. Cross-species regulatory network analysis identifies a synergistic interaction between FOXM1 and CENPF that drives prostate cancer malignancy. *Cancer Cell* **25**(5):638–651 (2014)

Baca-López, K., Mayorga, M., Hidalgo-Miranda, A., Gutiérrez-Nájera, N., y Hernández-Lemus, E. The Role of Master Regulators in the Metabolic/Transcriptional Coupling in Breast Carcinomas. *PLoS ONE* **7**(8):e42678 (2012)

Baca-López, K., Correa-Rodriguez, M.D., Flores-Espinosa, R., García-Herrera, R., Hernández-Armenta, C.I., Hidalgo-Miranda, A., Huerta-Verde, A.J., Imaz-Rosshandler, I., Martinez-Rubio, A.V., Medina-Escareno, A., Mendoza-Smith, R., Rodriguez-Dorantes, M., Salido-Guadarrama, I., Hernández-Lemus, E., y Rangel-Escareno, C. A three-state model for multidimensional genomic data integration. *Systems Biomedicine* **1**(2):122–129 (2014)

Ballman, K.V., Grill, D.E., Oberg, A.L., y Therneau, T.M. Faster cyclic loess: normalizing RNA arrays via linear models. *Bioinformatics* **20**(16):2778–2786 (2004)

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., y di Bernardo, D. How to infer gene networks from expression profiles. *Molecular Systems Biology* **3** (2007)

Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R., y Califano, A. Reverse engineering of regulatory networks in human B cells. *Nature Genetics* **37**(4):382–390 (2005)

Bolstad, B.M., Irizarry, R.A., Astrand, M., y Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**(2):185–193 (2003)

Boregowda, R.K., Olabisi, O.O., Abushahba, W., Jeong, B.S., Haenssen, K.K., Chen, W., Chekmareva, M., Lasfar, A., Foran, D.J., Goydos, J.S., y Cohen-Solal, K.A. RUNX2 is overexpressed in melanoma cells and mediates their migration and invasion. *Cancer Lett.* **348**(1-2):61–70 (2014)

Brazhnik, P., de la Fuente, A., y Mendes, P. Gene networks: how to put the function in genomics. *Trends in biotechnology* **20**(11):467–472 (2002)

Bulfone, A., Smiga, S.M., Shimamura, K., Peterson, A., Puelles, L., y Rubenstein, J.L. T-brain-1: a homolog of Brachyury whose expression defines molecularly distinct domains within the cerebral cortex. *Neuron* **15**(1):63–78 (1995)

Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**(7216):1061–1068 (2008)

Carro, M.S., Lim, W.K., Alvarez, M.J., Bollo, R.J., Zhao, X., Snyder, E.Y., Sulman, E.P., Anne, S.L., Doetsch, F., Colman, H., Lasorella, A., Aldape, K., Califano, A., y Iavarone, A. The transcriptional network for mesenchymal transformation of brain tumours. *Nature* **463**(7279):318–325 (2010)

Chen, Y., Fujita, T., Zhang, D., Doan, H., Pinkaew, D., Liu, Z., Wu, J., Koide, Y., Chiu, A., Lin, C.C., Chang, J.Y., Ruan, K.H., y Fujise, K. Physical and functional antagonism between tumor suppressor protein p53 and fortilin, an anti-apoptotic protein. *J. Biol. Chem.* **286**(37):32575–32585 (2011)

Cleveland, W.S. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association* **74**(368):829–836 (1979)

Cowan, J., Tariq, M., y Ware, S.M. Genetic and functional analyses of ZIC3 variants in congenital heart disease. *Hum. Mutat.* **35**(1):66–75 (2014)

de Anda-Jáuregui, G., Mejía-Pedroza, R.A., Espinal-Enríquez, J., y Hernández-Lemus, E. Crosstalk events in the estrogen signaling pathway may affect tamoxifen efficacy in breast cancer molecular subtypes. *Computational Biology and Chemistry* (2015)

De Paepe, B., Verstraeten, V.M., De Potter, C.R., y Bullock, G.R. Increased angiotensin II type-2 receptor density in hyperplasia, DCIS and invasive carcinoma of the breast is paralleled with increased iNOS expression. *Histochem. Cell Biol.* **117**(1):13–19 (2002)

Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., Viale, G., Delorenzi, M., Zhang, Y., d'Assignies, M.S., Bergh, J., Lidereau, R., Ellis, P., Harris, A.L., Klijn, J.G.M., Foekens, J.A., Cardoso, F., Piccart, M.J., Buyse, M., Sotiriou, C., y TRANSBIG Consortium. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clinical cancer research : an official journal of the American Association for Cancer Research* **13**(11):3207–3214 (2007)

Emmert-Streib, F., De Matos Simoes, R., Mullan, P., Haibe-Kains, B., y Dehmer, M. The gene regulatory network for breast cancer: integrated regulatory landscape of cancer hallmarks. *Statistical Genetics and Methodology* **5** (2014a)

Emmert-Streib, F., Tripathi, S., Simoes, R.d.M., Hawwa, A.F., y Dehmer, M. The human disease network. *Systems Biomedicine* **1**(1):20–28 (2014b)

Espinal-Enríquez, J., Muñoz-Montero, S., Imaz-Rosshandler, I., Huerta-Verde, A., Mejía, C., y Hernández-Lemus, E. Genome-wide expression analysis suggests a crucial role of dysregulation of matrix metalloproteinases pathway in undifferentiated thyroid carcinoma. *BMC Genomics* **16**(1):207 (2015)

Farmer, P., Bonnefoi, H., Becette, V., Tubiana-Hulin, M., Fumoleau, P., Larsimont, D., Macgrogan, G., Bergh, J., Cameron, D., Goldstein, D., Duss, S., Nicoulaz, A.L., Brisken, C., Fiche, M., Delorenzi, M., y Iggo, R. Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene* **24**(29):4660–4671 (2005)

Gao, J., Li, W.X., Feng, S.Q., Yuan, Y.S., Wan, D.F., Han, W., y Yu, Y. A protein-protein interaction network of transcription factors acting during liver cell proliferation. *Genomics* **91**(4):347–355 (2008)

Grass, P. Experimental Design. En A. Scherer (editor), *Batch Effects and Noise in Microarray Experiments*, capítulo 3, págs. 19–31. John Wiley & Sons, Ltd (2009)

Han, J.D.J., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy,

D., Walhout, A.J., Cusick, M.E., y Roth, F.P. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* **430**(6995):88–93 (2004)

Hanahan, D. y Weinberg, R.A. The hallmarks of cancer. *Cell* **100**(1):57–70 (2000)

Hanahan, D. y Weinberg, R.A. Hallmarks of Cancer: The Next Generation. *Cell* **144**(5):646–674 (2011)

Hara, S. y Kondo, Y. [Hypoxia-inducible factor-3alpha as a negative regulator of tumorigenesis]. *Seikagaku* **83**(1):50–55 (2011)

Heikkila, M., Pasanen, A., Kivirikko, K.I., y Myllyharju, J. Roles of the human hypoxia-inducible factor (HIF)-3 variants in the hypoxia response. *Cell. Mol. Life Sci.* **68**(23):3885–3901 (2011)

Hernández-Lemus, E. Systems Biology and Integrative Omics in Breast Cancer. En D. Barh (editor), *Omics Approaches in Breast Cancer*, págs. 333–352. Springer India, New Delhi (2014)

Hernández-Lemus, E. y Rangel-Escareño, C. The role of information theory in gene regulatory network inference. En P. Deloumeaux (editor), *Information Theory New Research*, págs. 137–184. Information Theory: New Research (2011)

Hernández-Lemus, E. y Siqueiros-García, J.M. Information theoretical methods for complex network structure reconstruction. *Complex Adaptive Systems Modeling* **1**(1):8 (2013)

Hernández-Lemus, E. Further steps toward functional systems biology of cancer. *Frontiers in physiology* **4**:256 (2013)

Hernández-Lemus, E., Velázquez-Fernández, D., Estrada-Gil, J.K., Silva-Zolezzi, I., Herrera-Hernández, M.F., y Jiménez-Sánchez, G. Information theoretical methods to deconvolute genetic regulatory networks applied to thyroid neoplasms. *Physica A* **388**(24):5057–5069 (2009)

Hernández-Lemus, E., Tovar, H., y Mejía, C. Non-equilibrium thermodynamics analysis of transcriptional regulation kinetics. *Journal of Non-Equilibrium Thermodynamics* **39**(4):205–218 (2014)

Hernández Patiño, C.E., Jaime-Muñoz, G., y Resendis-Antonio, O. Systems biology of cancer: moving toward the integrative study of the metabolic alterations in cancer cells. *Frontiers in physiology* **3** (2013)

Hinnebusch, A.G. y Natarajan, K. Gcn4p, a Master Regulator of Gene Expression, Is Controlled at Multiple Levels by Diverse Signals of Starvation and Stress. *Eukaryotic Cell* **1**(1):22–32 (2002)

Hosking, R. mTOR: The Master Regulator. *Cell* **149**(5):955–957 (2012)

Irizarry, R.A., Hobbs, B., Collin, F., Barclay, Y.D.B., Antonellis, K.J., Scherf, U., y Speed, T.P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**(2):249–264 (2003)

Ivshina, A.V., George, J., Senko, O., Mow, B., Putti, T.C., Smeds, J., Lindahl, T., Pawitan, Y., Hall, P., Nordgren, H., Wong, J.E.L., Liu, E.T., Bergh, J., Kuznetsov, V.A., y Miller, L.D. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer research* **66**(21):10292–10301 (2006)

Jansen, R.C. Studying complex biological systems using multifactorial perturbation. *Nature Reviews Genetics* **4**(2):145–151 (2003)

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., y Barabási, A.L. The large-scale organization of metabolic networks. *Nature* **407**(6804):651–654 (2000)

Jin, V.X., O'Geen, H., Iyengar, S., Green, R., y Farnham, P.J. Identification of an OCT4 and SRY regulatory module using integrated computational and experimental genomics approaches. *Genome Research* **17**(6):807–817 (2007)

Johnson, W.E., Li, C., y Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**(1):118–127 (2007)

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., y Hirakawa, M. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **34**(Database issue):D354–357 (2006)

Kel, A., Voss, N., Valeev, T., Stegmaier, P., Kel-Margoulis, O., y Wingender, E. ExPlain^{TM}: finding upstream drug targets in disease gene regulatory networks. *SAR and QSAR in Environmental Research* **19**(5-6):481–494 (2010)

Kitano, H. *Foundations of Systems Biology*. MIT Press (MA) (2001)

Kitano, H. Computational systems biology. *Nature* **420**(6912):206–210 (2002)

Krämer, A., Green, J., Pollard, J., y Tugendreich, S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics (Oxford, England)* **30**(4):523–530 (2014)

Kriete, A. y Eils, R. Computational systems biology (2014)

Krzywinski, M., Birol, I., Jones, S.J., y Marra, M.A. Hive plots—rational approach to visualizing networks. *Briefings in Bioinformatics* **13**(5):bbr069–644 (2011)

Lander, E.S., Linton, L.M., y et. al. Initial sequencing and analysis of the human genome. *Nature* **409**(6822):860–921 (2001)

Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., y Irizarry, R.A. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics* **11**(10):733–739 (2010)

Lefebvre, C., Rajbhandari, P., Alvarez, M.J., Bandaru, P., Lim, W.K., Sato, M., Wang, K., Sumazin, P., Kustagi, M., Bisikirska, B.C., Basso, K., Beltrao, P., Krogan, N., Gautier, J., Dalla-Favera, R., y Califano, A. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Molecular Systems Biology* **6**:1–10 (2010)

Lelli, K.M., Slattery, M., y Mann, R.S. Disentangling the many layers of eukaryotic transcriptional regulation. *Annual review of genetics* **46**(1):43–68 (2012)

Lim, W.K.W., Lyashenko, E.E., y Califano, A.A. Master regulators used as breast cancer metastasis classifier. *Audio and Electroacoustics Newsletter, IEEE* págs. 504–515 (2009)

Liu, B., Hoeschele, I., y de la Fuente, A. Handbook of Research on Computational Methodologies in Gene Regulatory Networks. En S. Das, D. Caragea, S.M. Welch, y W.H.

Hsu (editores), *Handbook of Research on Computational Methodologies in Gene Regulatory Networks*, págs. 79–107. IGI Global Snippet (2010)

Liu, R., Wang, X., Chen, G.Y., Dalerba, P., Gurney, A., Hoey, T., Sherlock, G., Lewicki, J., Shedden, K., y Clarke, M.F. The prognostic role of a gene signature from tumorigenic breast-cancer cells. *The New England journal of medicine* **356**(3):217–226 (2007)

Longabaugh, W.J. Combing the hairball with BioFabric: a new approach for visualization of large networks. *BMC Bioinformatics* **13**(1):275 (2012)

Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R., y Califano, A. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics* **7**(Suppl 1):S7 (2006a)

Margolin, A.A., Wang, K., Lim, W.K., Kustagi, M., Nemenman, I., y Califano, A. Reverse engineering cellular networks. *Nature Protocols* **1**(2):662–671 (2006b)

McCall, M.N., Bolstad, B.M., y Irizarry, R.A. Frozen robust multiarray analysis (fRMA). *Biostatistics* **11**(2):242–253 (2010)

Medvedovic, J.J., Ebert, A.A., Tagoh, H.H., y Busslinger, M.M. Pax5: a master regulator of B cell development and leukemogenesis. *Advances in Immunology* **111**:179–206 (2011)

Miller, L.D., Smeds, J., George, J., Vega, V.B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E.T., y Bergh, J. From the Cover: An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences of the United States of America* **102**(38):13550–13555 (2005)

Minn, A.J., Gupta, G.P., Siegel, P.M., Bos, P.D., Shu, W., Giri, D.D., Viale, A., Olshen, A.B., Gerald, W.L., y Massagué, J. Genes that mediate breast cancer metastasis to lung. *Nature* **436**(7050):518–524 (2005)

Mosca, E., Alfieri, R., Merelli, I., Viti, F., Calabria, A., y Milanesi, L. A multilevel data integration resource for breast cancer study. *BMC Systems Biology* **4**(1):76 (2010)

Mullen, A.C., Orlando, D.A., Newman, J.J., Lovén, J., Kumar, R.M., Bilodeau, S., Reddy, J., Guenther, M.G., DeKoter, R.P., y Young, R.A. Master Transcription Factors Determine Cell-Type-Specific Responses to TGF-β Signaling. *Cell* **147**(3):565–576 (2011)

Park, B., Shin, A., Kim, K.Z., Lee, Y.S., Hwang, J.A., Kim, Y., Sung, J., Yoo, K.Y., y Lee, E.S. Lack of effects of peroxisome proliferator-activated receptor gamma genetic polymorphisms on breast cancer risk: a case-control study and pooled analysis. *Asian Pac. J. Cancer Prev.* **15**(21):9093–9099 (2014)

Pau Ni, I.B., Zakaria, Z., Muhammad, R., Abdullah, N., Ibrahim, N., Aina Emran, N., Hisham Abdullah, N., y Syed Hussain, S.N.A. Gene expression patterns distinguish breast carcinomas from normal breast tissues: the Malaysian context. *Pathology, research and practice* **206**(4):223–228 (2010)

Pawitan, Y., Bjöhle, J., Amler, L., Borg, A.L., Egyhazi, S., Hall, P., Han, X., Holmberg, L., Huang, F., Klaar, S., Liu, E.T., Miller, L., Nordgren, H., Ploner, A., Sandelin, K., Shaw, P.M., Smeds, J., Skoog, L., Wedrén, S., y Bergh, J. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast cancer research : BCR* **7**(6):R953–64 (2005)

Perou, C.M., Sørlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., Fluge, Ø., Pergamenschikov, A., Williams, C., Zhu, S.X., Lønning, P.E., Børresen-Dale, A.L., Brown, P.O., y Botstein, D. Molecular portraits of human breast tumours : Article : Nature. *Nature* **406**(6797):747–752 (2000)

Pollack, J.R., Sorlie, T., Perou, C.M., Rees, C.A., Jeffrey, S.S., Lonning, P.E., Tibshirani, R., Botstein, D., Borresen-Dale, A.L., y Brown, P.O. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences* **99**(20):12963–12968 (2002)

Pon, C.K., Firth, S.M., y Baxter, R.C. Involvement of insulin-like growth factor binding

protein-3 in peroxisome proliferator-activated receptor gamma-mediated inhibition of breast cancer cell growth. *Mol. Cell. Endocrinol.* **399**:354–361 (2015)

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., y Smyth, G.K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43** (2015)

Rockman, M.V. Reverse engineering the genotype–phenotype map with natural genetic variation. *Nature* **456**(7223):738–744 (2008)

Ruckhaeberle, E., Rody, A., Engels, K., Gaetje, R., von Minckwitz, G., Schiffmann, S., Groesch, S., Geisslinger, G., Holtrich, U., Karn, T., y Kaufmann, M. Microarray analysis of altered sphingolipid metabolism reveals prognostic significance of sphingosine kinase 1 in breast cancer. *Breast Cancer Research and Treatment* **112**(1):41–52 (2008)

Satih, S., Chalabi, N., Rabiau, N., Bosviel, R., Fontana, L., Bignon, Y.J., y Bernard-Gallon, D.J. Gene Expression Profiling of Breast Cancer Cell Lines in Response to Soy Isoflavones Using a Pangenomic Microarray Approach. *Omics : a journal of integrative biology* **14**(3):231–238 (2010)

Schouten, P. Big data in health care. *Healthcare financial management : journal of the Healthcare Financial Management Association* **67**(2):40–42 (2013)

Shannon, C.E. y Weaver, W. *The mathematical theory of communication.* University of Illinois, Urbana and Chicago (1949)

Shimoni, Y. y Alvarez, M. *TF list* (2013)

Siriwardana, N.S., Meyer, R., Havasi, A., Dominguez, I., y Panchenko, M.V. Cell cycle-dependent chromatin shuttling of HBO1-JADE1 histone acetyl transferase (HAT) complex. *Cell Cycle* **13**(12):1885–1901 (2014)

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J.S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C.M., Lonning, P.E.,

Brown, P.O., Borresen-Dale, A.L., y Botstein, D. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences* **100**(14):8418–8423 (2003)

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., Desmedt, C., Larsimont, D., Cardoso, F., Peterse, H., Nuyten, D., Buyse, M., Van de Vijver, M.J., Bergh, J., Piccart, M., y Delorenzi, M. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute* **98**(4):262–272 (2006)

Steuer, R., Kurths, J., Daub, C.O., Weise, J., y Selbig, J. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics (Oxford, England)* **18 Suppl 2**:S231–40 (2002)

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., y Lander, E.S. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* **102**(43):15545–15550 (2005)

Sun-Kin Chan, S. y Kyba, M. What is a Master Regulator? *Journal of Stem Cell Research & Therapy* **03**(02) (2013)

Szabo, C., Masiello, A., Ryan, J.F., y Brody, L.C. The breast cancer information core: database design, structure, and scope. *Hum. Mutat.* **16**(2):123–131 (2000)

ter Kuile, B.H. y Westerhoff, H.V. Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS letters* **500**(3):169–171 (2001)

Tian, C., Lv, D., Qiao, H., Zhang, J., Yin, Y.h., Qian, X.p., Wang, Y.p., Zhang, Y., y Chen, W.f. TFDP3 inhibits E2F1-induced, p53-mediated apoptosis. *Biochemical and Biophysical Research Communications* **361**(1):20–25 (2007)

Tretyakov, K., Laur, S., Smant, G., Vilo, J., y Prins, P. Fast probabilistic file fingerprinting for big data. *BMC Genomics* **14**(Suppl 2):S8 (2013)

Tripathi, A., King, C., de la Morenas, A., Perry, V.K., Burke, B., Antoine, G.A., Hirsch, E.F., Kavanah, M., Mendez, J., Stone, M., Gerry, N.P., Lenburg, M.E., y Rosenberg, C.L. Gene expression abnormalities in histologically normal breast epithelium of breast cancer patients. *International journal of cancer. Journal international du cancer* **122**(7):1557–1566 (2008)

van 't Veer, L.J., Dai, H., Van de Vijver, M.J., He, Y.D., Hart, A.A.M., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., y Friend, S.H. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**(6871):530–536 (2002)

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., y Luscombe, N.M. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics* **10**(4):252–263 (2009)

Venter, J.C., Adams, M.D., y et. al. The sequence of the human genome. *Science* **291**(5507):1304–1351 (2001)

Visvanathan, M., Baumgartner, C., Tilg, B., y Lushington, G.H. Systems Biology Approach for Mapping TNFα-NFκB Mathematical Model to a Protein Interaction Map. *The Open Systems Biology Journal* **3**(1):1–8 (2010)

Wang, W., Chen, B., Zou, R., Tu, X., Tan, S., Lu, H., Liu, Z., y Fu, J. Codonolactone, a sesquiterpene lactone isolated from Chloranthus henryi Hemsl, inhibits breast cancer cell invasion, migration and metastasis by downregulating the transcriptional activity of Runx2. *Int. J. Oncol.* **45**(5):1891–1900 (2014)

Wang, Y., Yang, S., Ni, Q., He, S., Zhao, Y., Yuan, Q., Li, C., Chen, H., Zhang, L., Zou, L., Shen, A., y Cheng, C. Overexpression of forkhead box J2 can decrease the migration of breast cancer cells. *J. Cell. Biochem.* **113**(8):2729–2737 (2012)

Wang, Y., Klijn, J.G.M., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M.E., Yu, J., Jatkoe, T.,

Berns, E.M.J.J., Atkins, D., y Foekens, J.A. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet (London, England)* **365**(9460):671–679 (2005)

Wu, J.Q., Seay, M., Schulz, V.P., Hariharan, M., Tuck, D., Lian, J., Du, J., Shi, M., Ye, Z., Gerstein, M., Snyder, M.P., y Weissman, S. Tcf7 Is an Important Regulator of the Switch of Self-Renewal and Differentiation in a Multipotential Hematopoietic Cell Line. *PLOS Genetics* **8**(3):e1002565 (2012)

Yao, C., Li, H., Zhou, C., Zhang, L., Zou, J., y Guo, Z. Multi-level reproducibility of signature hubs in human interactome for breast cancer metastasis. *BMC Systems Biology* **4**(1):151 (2010)

You, L. Toward computational systems biology. *Cell Biochemistry and Biophysics* **40**(2):167–184 (2004)

Zhang, Y., Luo, H.Y., Liu, G.L., Wang, D.S., Wang, Z.Q., Zeng, Z.L., y Xu, R.H. Prognostic significance and therapeutic implications of peroxisome proliferator-activated receptor overexpression in human pancreatic carcinoma. *Int. J. Oncol.* **46**(1):175–184 (2015)

Zhao, Y., Lutzen, U., Fritsch, J., Zuhayra, M., Schutze, S., Steckelings, U.M., Recanti, C., Namsoleck, P., Unger, T., y Culman, J. Activation of intracellular angiotensin AT receptors induces rapid cell death in human uterine leiomyosarcoma cells. *Clin. Sci.* **128**(9):567–578 (2015)

Zhuang, Y., Li, D., Fu, J., Shi, Q., Lu, Y., y Ju, X. Overexpression of AIOLOS inhibits cell proliferation and suppresses apoptosis in Nalm-6 cells. *Oncol. Rep.* **31**(3):1183–1190 (2014)

Apéndice A

Código

A.1. Código de R usado para el preproceso

A.1.1. Normalización y colapso de la matriz de expresión

```
# Required packages
library(affy)
library(frma)
library(sva)
library(annotate)
library(hgu133a.db)
library(limma)


# Read CEL files from the Working Directory
Data = ReadAffy()

# Parameters for graphics

N = length(Data@phenoData@data$sample)
pm.mm = 0
for (i in 1:N) {
    pm.mm[i] = mean(mm(Data[, i]) > pm(Data[, i]))
  }
```

```
mycolors = rep(c("blue", "red", "green", "magenta"), each = 2)

# Control plot for raw data

pdf("DataGraphs.pdf", width = 7, height = 5)
  hist(Data, col = mycolors, main = "Raw data distribution")
  boxplot(Data, col = mycolors, main = "Raw data distribution")
  plot(100 * pm.mm, type = 'h', main = 'Percent of MMs > PMs', ylab = "%", xlab = "
      Microarrays", ylim = c(0, 50), col = "red", lwd = 5)
  grid(nx = NULL, ny = 6, col = "blue", lty = "dotted", lwd = par("lwd"), equilogs
      = TRUE)
dev.off()

# Frozen Robust Multi-Array summarization/normalization
frmaData <- frma(Data, summarize = "robust_weighted_average")

# Extracting the expresion matrix
edata <- exprs(frmaData)

# Control plot for summarized data

pdf("frmaNormalized.pdf", width = 7, height = 5)
  mycolors = rep(c("blue", "red","green", "magenta"), each = 2)
  plotDensity(edata, col = mycolors, main = "frma normalization")
  boxplot(edata, col = mycolors, main = "Normalized data distribution")
dev.off()

################ BATCH FILE ########################


GSMnames <- colnames(edata)

GSMnames <- data.frame(lapply(GSMnames, function(v) {
    if (is.character(v))
    return(toupper(v))
    else return(v)
  }))
```

```
GSMnames <- as.character(GSMnames)

colnames(edata) <- GSMnames

batch <- c((rep(0, N)))

# Treatments

# This .txt files contain the list of GSM ID for any GSE
GSE1456<-read.table("/lists_of_names/all_GSE1456.txt", colClasses = "character")
GSE1561<-read.table("/lists_of_names/all_GSE1561.txt", colClasses = "character")
GSE2603<-read.table("/lists_of_names/all_GSE2603.txt", colClasses = "character")
GSE2990<-read.table("/lists_of_names/all_GSE2990.txt", colClasses = "character")
GSE3494<-read.table("/lists_of_names/all_GSE3494.txt", colClasses = "character")
GSE4922<-read.table("/lists_of_names/all_GSE4922.txt", colClasses = "character")
GSE7390<-read.table("/lists_of_names/all_GSE7390.txt", colClasses = "character")


GSE1456 <- GSE1456[[1]]
GSE1561 <- GSE1561[[1]]
GSE2603 <- GSE2603[[1]]
GSE2990 <- GSE2990[[1]]
GSE3494 <- GSE3494[[1]]
GSE4922 <- GSE4922[[1]]
GSE7390 <- GSE7390[[1]]

n1456 <- which(GSMnames %in% GSE1456)
n1561 <- which(GSMnames %in% GSE1561)
n2603 <- which(GSMnames %in% GSE2603)
n2990 <- which(GSMnames %in% GSE2990)
n3494 <- which(GSMnames %in% GSE3494)
n4922 <- which(GSMnames %in% GSE4922)
n7390 <- which(GSMnames %in% GSE7390)

batch[n1456] = 1
batch[n1561] = 2
```

```
batch[n2603] = 3
batch[n2990] = 4
batch[n3494] = 5
batch[n4922] = 6
batch[n7390] = 7

case <- c(n1456, n1561, n2603, n2990, n3494, n4922, n7390)

# Controls

# This .txt files contain the list of GSM ID for any GSE
GSE15852 <-read.table("/lists_of_names/all_GSE15852.txt", colClasses = "character")
GSE6883 <-read.table("/lists_of_names/all_GSE6883.txt", colClasses = "character")
GSE9574 <-read.table("/lists_of_names/all_GSE9574.txt", colClasses = "character")



GSE15852 <- GSE15852[[1]]
GSE6883 <- GSE6883 [[1]]
GSE9574 <- GSE9574 [[1]]

n15852 <- which(GSMnames %in% GSE15852)
n6883  <- which(GSMnames %in% GSE6883)
n9574  <- which(GSMnames %in% GSE9574)

batch[n15852] = 8
batch[n6883] = 9
batch[n9574] = 10

control <- c(n15852, n6883, n9574)

##  ComBat for separated treatments ##

# split of the expression matrix
caseExp <- edata[, -control]
healtExp <- edata[, -case]

# split of the batch file
```

```
caseBatch <- c(batch[-control])
healtBatch <- c(batch[-case])

# ComBat for bouth groups
case_combat = ComBat(dat=caseExp, batch=caseBatch, mod=NULL)
healt_combat = ComBat(dat=healtExp, batch=healtBatch, mod=NULL)

# Join of the tu matrix
combat_2ways <- matrix(rep(0, (22283*N)), ncol=N)
colnames(combat_2ways) <- colnames(edata)
rownames(combat_2ways) <- rownames(edata)
combat_2ways[, control] <- healt_combat
combat_2ways[, case] <- case_combat

# Control plot for jioned matrix

pdf("pre2waysCombat_robust_weighted_average.pdf", width = 7, height = 5)
  mycolors = rep(c("blue", "red", "green", "magenta"), each = 2)
  plotDensity(combat_2ways, col = mycolors, main = "Pair␣combat␣normalization", sub
       = "background␣=␣none␣summarize␣=␣random_effect")
  boxplot(combat_2ways, col = mycolors, main = "Normalized␣data␣distribution")
dev.off()

# Normalization of both Combat normalized subsets
n_combat_2ways <- normalizeBetweenArrays(combat_2ways, method = "cyclicloess")

# Control plot for jioned cyclic loess normalized matrix

pdf("2waysCombat_robust_weighted_average_cyclicloess.pdf", width = 7, height = 5)
  mycolors = rep(c("blue", "red", "green", "magenta"), each = 2)
  plotDensity(n_combat_2ways, col = mycolors, main = "Pair␣combat␣normalization",
      sub = "␣background=␣none␣summarize␣=␣random_effect")
  boxplot(n_combat_2ways, col = mycolors, main = "Normalized␣data␣distribution")
dev.off()

### Pre-collapse preparation  ###
```

```
# Design and contingence matrix
design = matrix(rep(0,(N*2)), nrow=N)
colnames(design) = c('case','healt')
rownames(design) = colnames(combat_edata)
design[n1456,1]=1
design[n1561,1]=1
design[n2603,1]=1
design[n2990,1]=1
design[n3494,1]=1
design[n4922,1]=1
design[n7390,1]=1
design[n15852,2]=1
design[n6883,2]=1
design[n9574,2]=1
cont.matrix = makeContrasts('case - healt', levels=design)

# Treatment vector (for PVCA analysis)
treatment<-c(rep(0, N))
treatment[n1456] = 1
treatment[n1561] = 1
treatment[n2603] = 1
treatment[n2990] = 1
treatment[n3494] = 1
treatment[n4922] = 1
treatment[n7390] = 1
treatment[n15852] = 0
treatment[n6883] = 0
treatment[n9574] = 0

# limma for Differential Expression's analysis
fit = lmFit(n_combat_2ways, design)
fit2 = contrasts.fit(fit, cont.matrix)
fit2 = eBayes(fit2)

n_combat_2ways_statistics <- topTable(fit2n, coef = 1, adjust = 'fdr', n = length(
    row.names(n_combat_2ways)),  sort = "none")
```

```
# Control plot for log Fold Change
pdf(file="2ways_combat_cyclicloess_logFC_statisitics.pdf")
  boxplot(n_combat_2ways_statistics[, 1])
dev.off()

# Control plot for B statistic
pdf(file="2ways_combat_cyclicloess_B_statisitics.pdf")
  boxplot(n_combat_2ways_statistics[, 6])
dev.off()

  # Read my oun annotation of the chip
  my_annotation <- read.table(file = "my_annotation_hgu133A.txt", header = TRUE,
      row.names = 1,
    colClasses = "character")

  # Expresion matrix whith B statistic for collapse
  precolaps_2ways_combat <- cbind(my_annotation,  n_combat_2ways_statistics[, 6], n
      _combat_2ways)
  colnames(precolaps_2ways_combat)[2] <- c("b")
  write.table(precolaps_2ways_combat, file = "exp_matrix_normalized_precolaps.txt",
       quote = FALSE, sep = "\t", row.names = FALSE)
```

Procesamiento fuera de R del archivo de la matriz de expresión normalizada:

Script de Python: "microarray_colapser.py"

input: exp_matrix_normalized_precolaps.txt

output: exp_matrix_normalized_precolaps_collapsed.txt

```
  # This script helps to collapse the archives of microarray expression
  # It is required that the first and second column are the Gene Symbol and the
      statistical b consecutively ("loads" in our case)
  # It runs like any Python script:
  # $ python microarray_colapser.py path/to/the/exp_matrix_file.txt
  # output in this example would exp_matrix_file_colapsed.txt
  # The output file still has the original first two columns

  import sys
```

```
exp_file = open(sys.argv[1], 'rb')

new_lines = {}
dict1 = {}
names = []

for line in exp_file :
  p = line.split("\t")
  if p[0] in dict1:
    if p[1] > dict1[p[0]]:
      dict1 [p[0]] = p[1]
      new_lines [p[0]] = line.strip()
  else:
    dict1 [p[0]] = p[1]
    new_lines [p[0]] = line.strip()
    names.append(p[0])

exp_file.close()

output = open(sys.argv[1].replace(".txt", "", 1)+"_colapsed.txt", "w")
for n in names:
  output.write(new_lines[n]+"\n")
output.close()
```

Comando de shell para finalizar el archivo de expresión:

```
cut --complement -f2 \
exp_matrix_normalized_precolaps_colapsed.txt > \
exp_collapsed.txt
```

Lectura de la matriz de expresión colapsada y guardado del objeto de R final con los objetos útiles para el posterior análisis:

```
eset <- read.table("exp_collapsed.txt",header=TRUE, sep="\t", row.names=1)
save(eset, case, control, batch, treatment, file="exp_set.RData")
```

A.1.2. Análisis de PVCA

Con el fin de medir el efecto de lote se usó el siguiente código de R.

```
#Programmer:  Pierre R. Bushel
#Location: NIEHS
#email  Bushel@niehs.nih.gov
#Code: R
#Program name: pvca.R
#Date: May 26, 2009

# Modified for two treatments without time series
# and lme4 v 1.1-8
# Hugo Tovar
# https://github.com/hachepunto

########## load libraries ##########
library(lme4)
####### Edit these variables according to user defined parameters and the path to
     your data and data files names ##########

# myPath <- "../../Data/"
# theGene_expression_file <- "ge_data_transformed_mvi_tab_delimited.TXT"
# theExperiment_data_file <- "expinfo_tab_delimited2.TXT"

load("exp_set.RData")
setwd( "~/working/directory/")

pct_threshold = 0.8 # Amount of variability desired to be explained by the
    principal components.  Set to match the results in book chapter and SAS code.
      User can adjust this to a higher (>= 0.8) number but < 1.0

### In addition, be sure to modify the mixed linear model by adding the
    appropriate random effects terms in the model

n_sample = 880

###############################################
```

```
# theGEDFilePath = paste(myPath,theGene_expression_file, sep="")
# theExpDataFilePath = paste(myPath,theExperiment_data_file, sep="")

########## Load data ##########

theDataMatrix <- eset
dataRowN <- nrow(theDataMatrix)
dataColN <- ncol(theDataMatrix)

########## Center the data (center rows) ##########
theDataMatrixCentered <- matrix(data = 0, nrow = dataRowN, ncol = dataColN)
theDataMatrixCentered_transposed = apply(theDataMatrix, 1, scale, center = TRUE,
    scale = FALSE)
theDataMatrixCentered = t(theDataMatrixCentered_transposed)


exp_design = as.data.frame(matrix(rep(0,(n_sample*4),nrow=n_sample))
exp_design[,1]<-c(1:n_sample)
rownames(exp_design)<-exp_design[,1]
colnames(exp_design) <-c("sample","Treatment","Batch","columnname")
exp_design[,2]<-tratamiento
exp_design[,3]<-batch
exp_design[,4]<-GSMnames

expDesignRowN <- nrow(exp_design)
expDesignColN <- ncol(exp_design)
myColNames <- names(exp_design)


########## Compute correlation matrix ##########

theDataCor <- cor(theDataMatrixCentered)

########## Obtain eigenvalues ##########

eigenData <- eigen(theDataCor)
```

```
eigenValues = eigenData$values
ev_n <- length(eigenValues)
eigenVectorsMatrix = eigenData$vectors
eigenValuesSum = sum(eigenValues)
percents_PCs = eigenValues /eigenValuesSum

########## Merge experimental file and eigenvectors for n components ##########

my_counter_2 = 0
my_sum_2 = 1
for (i in ev_n:1){
my_sum_2  = my_sum_2 - percents_PCs[i]
  if ((my_sum_2) <= pct_threshold ){
    my_counter_2 = my_counter_2 + 1
  }

}
if (my_counter_2 < 3){
  pc_n  = 3

}else {
  pc_n = my_counter_2
}

# pc_n is the number of principal components to model

pc_data_matrix <- matrix(data = 0, nrow = (expDesignRowN*pc_n), ncol = 1)
mycounter = 0
for (i in 1:pc_n){
  for (j in 1:expDesignRowN){
  mycounter <- mycounter + 1
    pc_data_matrix[mycounter,1] = eigenVectorsMatrix[j,i]

  }
}

AAA <- exp_design[rep(1:expDesignRowN,pc_n),]
```

```
Data <- cbind(AAA,pc_data_matrix)

####### Edit these variables according to your factors #######

Data$Treatment <- as.factor(Data$Treatment)
Data$Batch <- as.factor(Data$Batch)

########## Mixed linear model ##########
op <- options(warn = (-1))
effects_n = (expDesignColN - 2) + ((expDesignColN - 2)*(((expDesignColN - 2)-1)))
    /2 + 1
randomEffectsMatrix <- matrix(data = 0, nrow = pc_n, ncol = effects_n)


for (i in 1:pc_n){
  y = (((i-1)*expDesignRowN)+1)
  #randomEffects <- (summary(Rm1ML <- lmer(pc_data_matrix ~ (1|Time) + (1|
      Treatment) + (1|Batch) + (1|Time:Treatment) + (1|Time:Batch) + (1|Treatment
      :Batch), Data[y:(((i-1)*expDesignRowN)+expDesignRowN),], REML = TRUE,
      control=lmerControl(maxIter = 1000000, msMaxIter=1000000, singular.ok=TRUE,
      tolerance=1e-4, returnObject=TRUE),verbose = FALSE, na.action = na.omit))
      @REmat)
  randomEffects <- as.data.frame(summary(Rm1ML <- lmer(pc_data_matrix ~ (1|
      Treatment) + (1|Batch) + (1|Treatment:Batch), Data[y:(((i-1)*expDesignRowN)
      +expDesignRowN),], REML = TRUE, verbose = FALSE, na.action = na.omit))$
      varcor)
  randomEffectsMatrix[i,] = randomEffects[,4]

}

effectsNames <- randomEffects[,1]

########## Standardize Variance ##########

randomEffectsMatrixStdze <- matrix(data = 0, nrow = pc_n, ncol = effects_n)
for (i in 1:pc_n){
```

```
  mySum = sum(randomEffectsMatrix[i,])
  for (j in 1:effects_n){
    randomEffectsMatrixStdze[i,j] = randomEffectsMatrix[i,j]/mySum
  }
}

########## Compute Weighted Proportions ##########

randomEffectsMatrixWtProp <- matrix(data = 0, nrow = pc_n, ncol = effects_n)
for (i in 1:pc_n){
  weight = eigenValues[i]/eigenValuesSum
  for (j in 1:effects_n){
    randomEffectsMatrixWtProp[i,j] = randomEffectsMatrixStdze[i,j]*weight
  }
}

########## Compute Weighted Ave Proportions ##########

randomEffectsSums <- matrix(data = 0, nrow = 1, ncol = effects_n)
randomEffectsSums <-colSums(randomEffectsMatrixWtProp)
totalSum = sum(randomEffectsSums)
randomEffectsMatrixWtAveProp <- matrix(data = 0, nrow = 1, ncol = effects_n)

for (j in 1:effects_n){
  randomEffectsMatrixWtAveProp[j] = randomEffectsSums[j]/totalSum

}

pdf("pvca.pdf",width=5,height=7)
bp <- barplot(randomEffectsMatrixWtAveProp,  xlab = "Effects", ylab = "Weighted␣
    average␣proportion␣variance", ylim= c(0,1.1),col = c("blue"), las=2)

axis(1, at = bp, labels = effectsNames, xlab = "Effects", cex.axis = 0.5, las=2)
values = randomEffectsMatrixWtAveProp
new_values = round(values , 3)
text(bp,randomEffectsMatrixWtAveProp,labels = new_values, pos=3, cex = 0.8) #
    place numbers on top of bars
```

```
dev.off()
```

A.2. Paralelización con HTCondor

A.2.1. Paralelización de ARACNe

Para generar los archivos .condor que correrán ARACNe se usa el siguiente script de Python llamado "genera_condor.py"

Todos estos scripts de Python requieren de $Jinja2 == 2.7.3$.

```
import argparse
import os, stat
from jinja2 import Environment, FileSystemLoader

parser = argparse.ArgumentParser(description='Generates condor submit file for
    aracne runs.')
parser.add_argument('--path_to_aracne2',  type=argparse.FileType('r'), required=
    True, help='path to aracne2 binary')
parser.add_argument('--expfile',  type=argparse.FileType('r'), required=True,
    help='expression file')
parser.add_argument('--probes',  type=argparse.FileType('r'), required=True, help
    ='probes, one in every line' )
parser.add_argument('--run_id',  required=True, help="name of condor run" )
parser.add_argument('--outdir',  required=True, help="outdir for adj matrices" )
parser.add_argument('--p',  required=True, help="P-value: e.g. 1e-7" )

args    = parser.parse_args()

expfile = args.expfile.name
p       = args.p
outdir  = args.outdir

# make sane affy ids
probes = []
for id in args.probes.readlines():
    probes.append( id.strip() )
```

```
# create exec dir
if not os.path.exists(outdir):
    os.makedirs(outdir)

# create aracne.sh
aracne_path = os.path.dirname(os.path.realpath(args.path_to_aracne2.name))
aracne_sh = """#!/bin/bash
cd {aracne_path}
./aracne2 $@"""
with open(os.path.join(outdir,'aracne.sh'), 'w') as f:
    f.write(aracne_sh.format(aracne_path=aracne_path))
    os.chmod(f.name, stat.S_IREAD | stat.S_IWRITE | stat.S_IEXEC | stat.S_IXGRP |
         stat.S_IXOTH )
#
# create condor script
#
scriptname = "%s/%s.condor" % (outdir, args.run_id)

# use same dir as this file's as environment
env = Environment(loader=FileSystemLoader(os.path.dirname(os.path.realpath(
    __file__))))
template = env.get_template('condor_aracne.tt')

with open(scriptname, 'w') as f:
    f.write( template.render( expfile = expfile,
                              probes = probes,
                              p       = p,
                              outdir  = outdir,
                              run_id  = args.run_id ) )
```

Este script requiere del templete de texto contenido en el archivo .tt llamado "condor_aracne.tt":

```
executable = aracne.sh
error      = {{ run_id }}.error
universe   = vanilla
log        = {{ run_id }}.log
```

```
  {% for id in probes %}

  Arguments = -i {{ expfile }} \
              -h {{ id }} \
              -p {{ p }} \
              -o {{ outdir }}/{{ id }}_{{ p }}.adj
  Output       = {{ id }}_{{ p }}.log
  Error        = {{ id }}_{{ p }}.err
  Queue

  {% endfor %}
```

Un ejemplo de comando utilizado para generar los .condor es:

```
python ~/breast_cancer_networks/parallel_aracne/genera_condor.py \
    --path_to_aracne2 /ARACNE/aracne2 \
    --expfile /exp_collapsed.txt \
    --probes /human_TF_list.txt \
    --run_id breast_cancer_TF_network \
    --outdir /breast_cancer_TF_network_p1 \
    --p 1
```

Para echar a andar condor:

```
cd /breast_cancer_TF_network_p1
condor_submit breast_cancer_TF_network.condor
```

Para unir los .adj resultantes:

```
cd breast_cancer_TF_network/
cat *.adj | grep -v ">" > ../breast_cancer_TF_network_p1.adj
```

A.2.2. Paralelización del cortador de redes

Para generar los archivos .condor se utiliza el script: “genera_prune_condor.py”

```
  import argparse
  import os, stat
  from math import log
  from jinja2 import Environment, FileSystemLoader
```

```
parser = argparse.ArgumentParser(description='Prune interaction below given
    threshold.')
parser.add_argument('--adj',  type=argparse.FileType('r'), required=True, help='
    one or more adjacency files')
parser.add_argument('--outdir',  required=True, help="directory to place condor
    scripts" )
parser.add_argument('--p',  required=True, help="P-value: e.g. 1e-7" )
parser.add_argument('--n',  required=True, help="sample size" )

args    = parser.parse_args()

# compute mi value, from bootstrap Aldo Huerta 2014
alfa  = 1.062
beta  = -48.7
gamma = -0.634
p     = float(args.p)
n     = int(args.n)
mi = (alfa - log(p)) / ((-beta) + (-gamma)*n)
print "will generate prune scripts for mi= %f"  % mi


# use same dir as this file's as environment, load template
env = Environment(loader=FileSystemLoader(os.path.dirname(os.path.realpath(
    __file__))))
template = env.get_template('prune_adj.tt')

# create output dir
if not os.path.exists(args.outdir):
    os.makedirs(args.outdir)
os.chdir(args.outdir)

# create one prune script per line
lineas = args.adj.readlines()
matrix_name = os.path.basename(args.adj.name)

scripts = []
```

```
for linea in lineas:
    if not linea.startswith('>'):
        gene_line = linea.strip()
        gene_list = gene_line.split()
        gene_id = gene_list[0]
        prune_script = "condor_%s_prune.py" % gene_id
        with open(prune_script, 'w') as f:
            f.write( template.render( gene_line   = gene_line,
                                      matrix_name = matrix_name,
                                      p           = args.p,
                                      mi          = "%f" % mi ) )
        scripts.append(prune_script)

stanza = """
Arguments = {s}
Error = {s}.err
Queue
"""

with open('%s.condor' % matrix_name,'w') as condor_file:
    condor_file.write("""executable = /usr/bin/python
universe   = vanilla
""")
    for s in scripts:
        condor_file.write(stanza.format(s=s))
```

Con el templete "prune_adj.tt"

```
import os

mi = {{ mi }}
p  = {{ p }}
matrix_name = '{{ matrix_name }}'
gene_line = '{{ gene_line }}'

gene_mi_list = gene_line.split()

main_gene = gene_mi_list[0]
```

```
  filtered = [main_gene, ]
  for n in range(1,len(gene_mi_list),2):
      if float(gene_mi_list[n+1]) >= mi:
          filtered.append(gene_mi_list[n])
          filtered.append(gene_mi_list[n+1])

  adj_filename = "%s_%s_%s.adj" % (main_gene, matrix_name, p)

  with open(adj_filename, 'w') as f:
      f.write("\t".join(filtered))
      f.write("\n")

  os.remove(os.path.realpath(__file__))
```

Un ejemplo de comando utilizado para generar los .condor es:

```
python ~/breast_cancer_networks/parallel_aracne/genera_prune_condor.py \
    --adj /breast_cancer_TF_network_p1.adj \
    --outdir BC_network_prunned/p1exp-100/ \
    --p 1e-100 \
    --n 880
```

Para echar a andar condor:

```
cd BC_network_prunned/p1exp-100/
condor_submit breast_cancer_TF_network_p1.condor
```

Para unir los .adj resultantes:

```
cd BC_network_prunned/p1exp-100/
cat *adj > ../breast_cancer_TF_network_p-100.adj
```

A.3. Script de conversión .adj a .sif listo para Cytoscape

```
  # Convert aracne2 .adj files to tab separated values suitable .sif files for
      cytoscape
  # Usage: $ python adj2cytoscape.py file.adj > file.sif
```

```
import argparse

parser = argparse.ArgumentParser(description='Convert␣aracne2␣adjacency␣files␣to␣
    tab␣separated␣values␣suitable␣for␣cytoscape.')
parser.add_argument('adj', type=argparse.FileType('r'), nargs="+", help="One␣or␣
    more␣ADJ␣files." )
args    = parser.parse_args()

for f in args.adj:
    for l in f.readlines():
        if not l.startswith('>'):
            cols   = iter(l.strip().split('\t'))
            source = cols.next()
            for target, mi in ((item, cols.next()) for item in cols):
                print "%s␣%s␣%s" % (source, mi, target)
```

A.4. Código para **ssmarina**

A.4.1. Instalación de **ssmarina** y dependencias

```
install.packages("http://files.figshare.com/1217799/ssmarina_1.01.tar.gz", repos =
    NULL)
install.packages("mixtools")
```

A.4.2. Inferencia de reguladores transcripcionales maestros con **ssmarina**

```
###     M A R I N a     ###

# Required packages
library(mixtools)
library(ssmarina)

# Load expresion matrix data
load(exp_set.RData)
```

```
## Regulon generation (P = 1exp-100 breast cancer Transcription Factor network in
    this example) ##
regulon <- aracne2regulon(breast_cancer_TF_network_p-100.adj, eset)

# Create a working directory
dir.create("p1exp-100/")
setwd("p1exp-100/")

save(regulon, file = "./regulon.RData")

##  Signature generation  ##

# Row by row T test comparation of the two matrix
signature <- rowTtest(eset[, case], eset[, control])
# z-score values for the GES
signature <- (qnorm(signature$p.value/2, lower.tail = FALSE) * sign(signature$
    statistic))[, 1]

# Null model by sample permutations
nullmodel <- ttestNull(eset[, case], eset[, control], per = 1000, repos = T)

save(signature, nullmodel, file = "./sign_null.RData")
# This molecular signature and null model can be used for analysis of other P
    values networks

########  M A R I N A ########

mrs <- marina(signature, regulon, nullmodel)
mrs_noledges <- mrs # Backup results

# Leading-edge analysis
mrs <- ledge(mrs)

# Plotting masters regulators
pdf("Top10mrs.pdf", width = 6, height = 7)
  plot(mrs_noledges, cex = 0.7)
dev.off()
```

```
# Saving Top 100 in a file
top100 <- summary(mrs, mrs = 100)
write.table(top100, file = "./top100mrs.txt", quote = FALSE, sep = "\t")

#######  S H A D O W  and  S Y N E R G Y  A N A L Y S I S #########

# Shadow analysis
mrshadow <- shadow(mrs, pval = 25)

# Writing the SIF file of the shadow interactions
write.table(summary(mrshadow)$Shadow.pairs, file = "shadow_pairs.sif", quote =
    FALSE, sep = "\t", row.names = FALSE, col.names = FALSE)

# Synergy analysis
mrs <- marinaCombinatorial(mrs, regulators = 15)

mrs <- marinaSynergy(mrs)

# plotting synergy regulators
pdf("Synergy.pdf", width = 11, height = 7)
  plot(mrs, mrs = 10, cex= 0.7)
dev.off()

##  F I N A L  S A V E  ##
save(mrs, mrs_noledges, mrshadow, file = "MARINa.RData")
```

Computational Biology and Chemistry 59 (2015) 67–77

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Transcriptional master regulator analysis in breast cancer genetic networks

Hugo Tovar[a], Rodrigo García-Herrera[a], Jesús Espinal-Enríquez[a,b], Enrique Hernández-Lemus[a,b,*]

[a] Computational Genomics Department, National Institute of Genomic Medicine (INMEGEN), Mexico
[b] Center for Complexity Sciences, National Autonomous University of Mexico (UNAM), Mexico

ARTICLE INFO

Article history:
Received 12 March 2015
Received in revised form 17 August 2015
Accepted 17 August 2015
Available online 22 August 2015

Keywords:
Transcriptional master regulators
Breast cancer
Gene regulatory networks
Systems biology

ABSTRACT

Gene regulatory networks account for the delicate mechanisms that control gene expression. Under certain circumstances, gene regulatory programs may give rise to amplification cascades. Such transcriptional cascades are events in which activation of key-responsive transcription factors called *master regulators* trigger a series of gene expression events. The action of transcriptional master regulators is then important for the establishment of certain programs like cell development and differentiation. However, such cascades have also been related with the onset and maintenance of cancer phenotypes. Here we present a systematic implementation of a series of algorithms aimed at the inference of a gene regulatory network and analysis of transcriptional master regulators in the context of primary breast cancer cells. Such studies were performed in a highly curated database of 880 microarray gene expression experiments on biopsy-captured tissue corresponding to primary breast cancer and healthy controls. Biological function and biochemical pathway enrichment analyses were also performed to study the role that the processes controlled – at the transcriptional level – by such master regulators may have in relation to primary breast cancer. We found that transcription factors such as AGTR2, ZNF132, TFDP3 and others are master regulators in this gene regulatory network. Sets of genes controlled by these regulators are involved in processes that are well-known hallmarks of cancer. This kind of analyses may help to understand the most upstream events in the development of phenotypes, in particular, those regarding cancer biology.

1. Introduction

Cancer is a pathway-disease (Hanahan and Weinberg, 2000). The main hallmarks of cancer are associated to the action of pathways related to cell proliferation, apoptosis evasion, cell differentiation and in general, to the dysregulation of cell cycle and the alteration of DNA-repairing processes (Hanahan and Weinberg, 2000, 2011). The phenotype of a cell is determined by the activity of a large number of genes and proteins (Basso et al., 2005). Hence, transcriptional regulation lies at the heart of many of the intricate molecular relationships driving the activity of biological pathways (Emmert-Streib et al., 2014).

It has been observed that a number of large scale transcriptional cascades behind such complex cellular processes are actually triggered by the action of a relatively small number of transcription factor molecules that have been called Transcriptional Master Regulators (TMRs) (Han et al., 2004; Sun-Kin Chan and Kyba, 2013; Mullen et al., 2011). It has been argued that these genes control the entire transcriptional regulatory program for specific cellular phenotypes (in eukaryotic cells; Han et al., 2004; Basso et al., 2005; Affara et al., 2013). However, TMRs are also able to act on general cellular processes at the same time (Hinnebusch and Natarajan, 2002; Medvedovic et al., 2011; Affara et al., 2013). A proper understanding of the organization of these TMR-mediated highly-regulated events is thus crucial to elucidate normal cell physiology as well as complex pathological phenotypes (Basso et al., 2005).

Given the complex mechanisms underlying transcriptional regulations on eukaryotes, the identification of TMRs is often based on the (inferred or observed) relationship among them and their cascade of RNA targets in gene regulatory networks (Hernández-Lemus and Siqueiros-García, 2013). Being a primary upstream event in the cell regulatory program, dysregulation of TMRs may have a high impact on cancer-related phenotypes, since under genetic instability conditions, uncontrolled synthesis of these

* Corresponding author at: Computational Genomics Department, National Institute of Genomic Medicine (INMEGEN), Mexico.
E-mail address: ehernandez@inmegen.gob.mx (E. Hernández-Lemus).

http://dx.doi.org/10.1016/j.compbiolchem.2015.08.007
1476-9271/

molecules could originate the activation/amplification of several transcriptional cascades (Basso et al., 2005; Baca-Lopez et al., 2014; Baca-López et al., 2012).

A TMR is a transcription factor (TF) that is expressed at the early onset of the development of a particular phenotype or cell type (Sun-Kin Chan and Kyba, 2013). It also participates in the specifications of such a phenotype by regulating multiple downstream genes, either directly or by means of genetic cascades. Transcription factors are hence key cellular components that control gene expression: their activities may determine how cells function and respond to the environment (Vaquerizas et al., 2009).

Transcription factors may act in two opposite directions: either activating or repressing transcriptional activity of their targets. Based on the initial estimations of the whole human genome sequence, it was calculated that the transcriptional machinery could be composed of 200 to 300 genes and there could exist between 2000 to 3000 specific union sites for transcription factors (Lander et al., 2001; Venter et al., 2001). In Vaquerizas et al. (2009) it is stated that in the http://amigo.geneontology.orgGene Ontology database 1052 TFs were defined and just 6% (62 cases) of them had experimental corroboration. Six years later, the same database recognized 1846 TFs and 14% (260) of them had experimental evidence. This is indicative of the fast progress on documenting the transcription mechanisms, but this also points to the overwhelming complexity of the mechanisms of genomic control.

Implementation of computational methods to identify and analyze TMRs is relevant in the context of breast cancer, particularly at its earliest stages. We have probabilistically inferred the gene regulatory network associated with this phenotype, then a computational analysis has uncovered its active TMRs in the context of primary breast cancer. In our study we have considered such an analysis, as well as the resulting TMR-related phenomena in the context of transcriptional regulatory programs. We also discuss here some of the implications of our results in breast cancer biology. The article is structured as follows: Section 2 presents an overview of the materials and methods used in this work. This includes both the experimental datasets used, the network inference strategy and the molecular signature analysis, as well as the algorithm for the discovery of transcriptional master regulators. Section 3 presents some of the main results of the application of this pipeline in primary breast cancer microarray gene expression data. Finally, Section 4 presents some conclusions mainly related with the advantages of implementing a method such as MARINa (Lefebvre et al., 2010) in order to unveil some aspects of regulatory control that may lie behind the establishment of tumor phenotypes.

2. Materials and methods

2.1. Experimental datasets

For the analysis presented here, we obtained 880 microarray expression profiles from several experimental datasets that are available on the Gene Expression Omnibus site (http://www.ncbi.nlm.nih.gov/geo/GEO) (Edgar et al., 2002). All experiments were performed by using total mRNA on the microarray platform Affymetrix HGU133A (GPL96), which consists of 18,400 transcripts and variants, including 14,500 well-characterized human genes (Liu et al., 2003). From the total 880 samples, 819 correspond to primary breast cancer tissue, whereas the remaining 61 samples correspond to healthy breast tissue. In the case of experiments that included any kind of treatment or cell modification, we only used the unaltered samples (see Table 1).

A second dataset for comparing the results was obtained from The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/). We used 597 mRNA samples of invasive breast cancer, of which 534 correspond to tumor samples and the other 63 were non-tumor. All data used for this analysis correspond to level 3, which means they are already normalized.

2.2. Batch effect control

Batch effect is one of the most recurrent factors of error during data analysis from microarrays (Grass, 2009). Chen et al. (2011) tested six different algorithms to eliminate batch effect and found that the best results were obtained by using the empirical bayesian method known as ComBat (Combating Batch Effects When Combining Batches of Gene Expression Microarray Data) (Johnson et al., 2007). However, since seven out of the ten datasets corresponded to tumor tissue exclusively (i.e. there are no control samples), and the three remaining datasets had only healthy tissues, there is no intersection between those datasets. According to Leek et al. (2010), treatments and batches are completely confounded. Since currently there is no method to estimate the batch effect under these conditions (Leek et al., 2010), ComBat (Johnson et al., 2007) cannot perform the normalization of the whole dataset. Taking into account that ComBat does not eliminate batch effect with the conditions of our dataset, we decided to partially solve this issue as follows: After preprocessing all arrays with frma (McCall et al., 2010), and using summarization with robust weighted average with no background correction, we split the datasets into cases/controls, and then applied ComBat to both datasets separately. After that, we re-joined the two resulting datasets and re-normalized them together with the cyclic loess algorithm (Ballman et al., 2004), in such way that both conditions belong now to the *same dynamic range*

We needed to have a measure of the batch effect within the samples so that we could remove the corresponding bias as accurately as possible. To this end we resort to Principal Variance Component Analysis (PVCA) that is an algorithm that combines the advantages of the principal component analysis (reduction of dimensionality) with the components of the analysis of variance (Grass, 2009). Once the batch effect is reduced separately, a PVCA analysis corroborated that such a batch effect almost disappeared and the treatment effect was important enough. (Fig. 1).

Given our design conditions, it was not possible to eliminate batch effect completely. Since batch effect in such mixed experimental designs is an important topic of current research in computational genomics, we can envisage a scenario in which the present work may be revisited and some of its conclusions may need to be revised. In the meantime, the method described above aimed at reducing and estimating batch effects may be considered a first approximation for the purposes of the work presented here.

For the TCGA dataset, since we analyzed data level 3 samples, normalization had already been performed by the TCGA site. For batch effect correction, the data were computed using ComBat (Johnson et al., 2007), Median Polish and ANOVA.

2.3. Network inference

Gene regulatory networks (GRN) are models that describe the relationship between genes under certain given conditions. Network inference can be defined as the process of identifying gene interactions from experimental data by performing a computational analysis (Bansal et al., 2007). To infer the breast cancer transcription factor regulatory network (interactome), we proceeded as follows. First, we generated a network for every known human TF in the primary breast cancer gene expression dataset by using the Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) (Basso et al., 2005; Margolin et al., 2006). ARACNE is a computational algorithm widely used to identify statistical relationships among genes, by calculating the mutual information (*MI*) between gene pairs from microarray expression data (Basso et al., 2005; Margolin et al., 2004).

Table 1
GEO identifier and references for the microarray experiments used here, the first column is GEO key ID, the second and third columns are the corresponding number of samples cases/controls, respectively. The fourth column is a brief description of the samples, and the fifth one presents the associated reference.

GEO ID Series	Tumors	Controls	Description	Reference
GSE1456	159		Breast cancer patients receiving surgery	Pawitan et al. (2005)
GSE1561	49		Biopsies were taken from patients with large operable, locally advanced or inflammatory breast cancer.	Farmer et al. (2005)
GSE2603	99		Tissues from primary breast cancers were obtained from therapeutic procedures performed as part of routine clinical management.	Minn et al. (2005)
GSE2990	61		Microarray experiments from primary breast tumors	Sotiriou et al. (2006)
GSE3494	4		Freshly frozen breast tumors tissues	Miller et al. (2005)
GSE4922	249		Primary invasive breast tumors	Ivshina et al. (2006)
GSE7390	198		Microarray experiments from primary breast tumors	Desmedt et al. (2007)
GSE6883		3	Samples were processed within an hour after breast reduction surgery	Liu et al. (2007)
GSE9574		15	Samples were obtained from patients undergoing reduction mammoplasty without apparent breast cancer	Tripathi et al. (2008)
GSE15852		43	Paired normal tissues	Pau Ni et al. (2010)
Total	819	61		

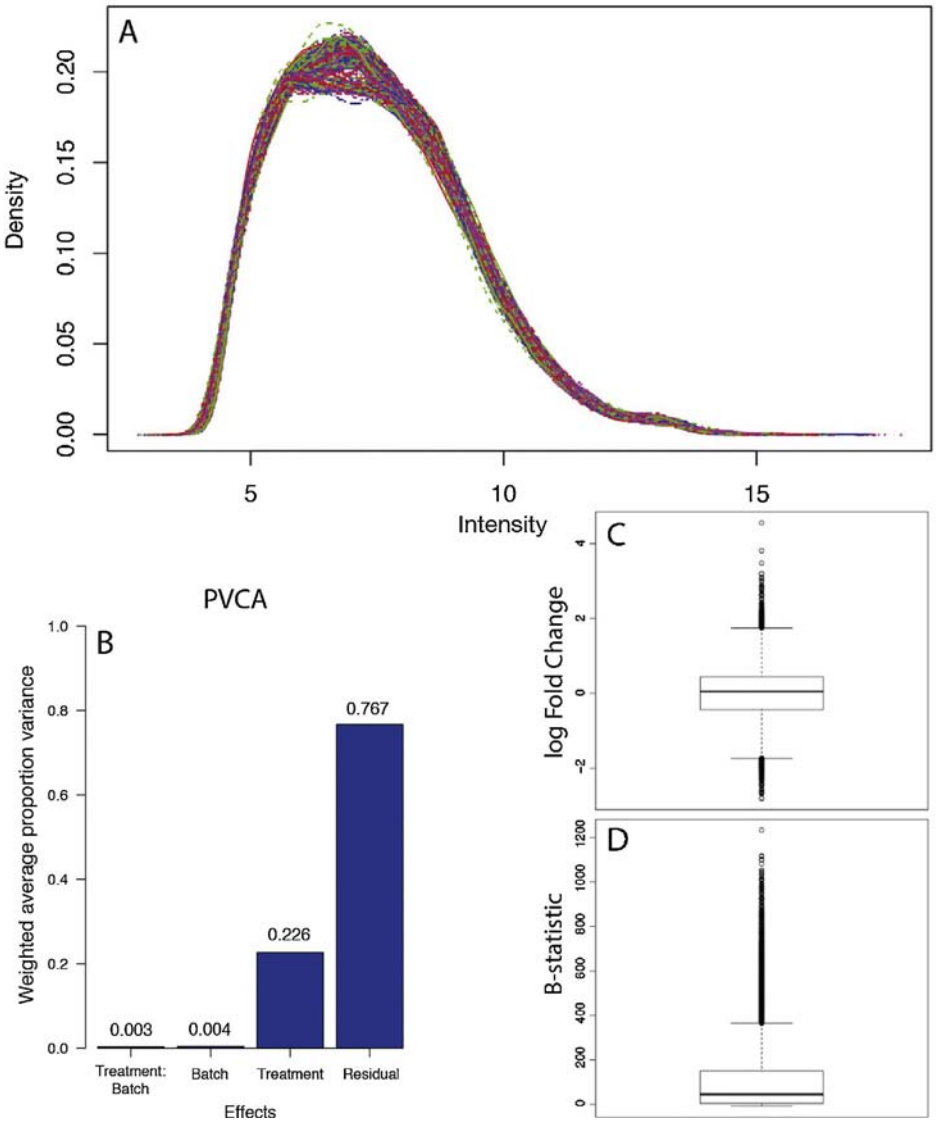


Fig. 1. Batch effect control. A. Density plot of the GEO expression matrix used for this analysis after application of ComBat for cases and controls separately. Thereafter, samples were joined with a cyclic loess normalization. B. PVCA analysis shows a minimum batch effect as well as an important contribution of treatment to variance. C. log fold-change and D. B-statistics.

A critical factor in the analysis of genetic regulation is the selection of variables (also called feature selection), which leads to the best predictive model. The methods of feature selection applied to genomic data could enhance the diagnosis for complex diseases, such as cancer, through the identification of a small subset of features or variables that represent the phenotype in a more accurate way. Thus we selected features by generating ARACNE networks with *p*-values of 10^{-30}, 10^{-40}, 10^{-50} and 10^{-100} in order to explore the structure of the TF's regulatory network at different threshold values of *MI*. Each network was inferred using the default DPI value (1) and also using the full list of transcription factors (Supplementary Material 1).

2.4. Molecular signature

The Molecular Signature (MS) was obtained by comparing gene expression of the microarrays of healthy tissue samples with those of tumor tissue. To this end a Student t-test was performed to compare both matrices row by row. This analysis was performed with an optimized function *rowTtest* from the package http://figshare.com/articles_ssmarina_R_system_package/785718*ssmarina*(Alvarez, 2013); inputs are the sets of both problem and control samples. It returns a list objects with their *t* statistic and *p*-value. To be consistent with the null model, the *z*-score of each comparison is calculated.

2.5. Master regulator inference algorithm

Only a few methods have been developed so far for the identification of TMRs. One of them is Biobase ExPlain® (Qiagen). ExPlain usesTransfac® and claims to be able to identify how the TFs affect gene expression both in microarrays and RNA-Seq experiments. This is done to predict how the TFs induce apattern of gene expression. Their site is hosted on http://www.biobase-international.com/transfac-upgrade. Another method is MARINa (MAster Regulator INference algorithm) which identifies TFs whose ARACNE-inferred targets (their regulon) have increased or decreased their expression levels in the context of a particular genetic signature. This aids to elucidate the way in which phenotypic differentiation is carried out (Lefebvre et al., 2010)

MARINa is designed to infer transcription factors that control the transition between two phenotypes A and B, as well as the maintenance of the latter phenotype. If the A → B transition is supported by the activation or repression of specific TFs then their targets should be among the most differentially expressed genes between the two cellular phenotypes, with activated and repressed targets at opposite ends of the expression range. MARINa estimates the importance and biological relevance of a TF on a given phenotype by computing the statistical significance of the overlap between its regulon and the gene expression signature using sample permutation to estimate the distribution of the enrichment score (ES) (Subramanian et al., 2005) in the null condition (Lefebvre et al., 2010).

Fig. 2 shows an overview of the MARINa pipeline. Brief descriptions of ARACNE and MARINa algorithms are provided in Supplementary Material 2. In this work we used the [R] (R Development Core Team, 2014) MARINa implementation *ssmarina* (Alvarez, 2013) which considers all aforementioned features. The networks obtained by ARACNE, as well as our calculated regulon-set network, were analyzed with the *Cytoscape* (Shannon et al., 2003) plugin *NetworkAnalyzer* (v1.0) (Assenov et al., 2008).

An important step for this algorithm is the selection of the Transcription Factors, since they will determine the rest of the calculation. A proper annotation of transcription factors is crucial for an accurate description of the process under investigation. Here, we used the HGU133A annotation file, in which we found 1142 TFs (Supplementary Material 1). This list was compared with other three lists. Those lists are available in Shimoni and Alvarez (2013), Vaquerizas et al. (2009) and http://www.bioguo.org/AnimalTFDB/ Animal Transcription Factor DataBase, respectively. We want to stress that all four lists show consistency among them.

2.5.1. Network structural features induced by transcriptional master regulators

Master regulator analysis (as outlined in the MARINa algorithm) may also point out to emerging phenomena related with the collective, cooperative action of groups of TMRs. Two quite interesting related emerging structural properties of TMR-driven GRNs are *shadowing* and *synergy* of TMRs over their target gene sets (Lefebvre et al., 2010). These phenomena are defined as follows:

Definition 1. A **gene regulatory network** can be defined as a graph $\mathcal{G}(V, E)$ over a duplex formed by two sets, a set V of nodes or vertices ($v_i \in V$) given by **genes**, and a set E of edges connecting such vertices ($e_i \in E$) representing **transcriptional regulatory** interactions among such genes. The connectivity rule is represented by the so-called **adjacency matrix** $\mathcal{A} = A_{i,j}$, where $A_{i,j} \neq 0$ implies a non-null interaction among gene v_i which regulates gene v_j or vice versa.
Let M_i and M_j be two genes, acting as **potential transcriptional master regulators** in $\mathcal{G}(V, E)$.
Let $\Omega(M_i)_\gamma$ be the set of gene targets for M_i enriched in a given signature γ that is $\forall v_k \in \Omega(M_i)_\gamma \exists A_{i,k} \neq 0$.
In the case of **indirect** transcriptional interactions, adjacency relations may be written as: $A_{i,k} = A_{i,m} \circ A_{m,n} \circ A_{n,o} \ldots \circ A_{o,k}$. With $\circ$ the composition function implying sequential interaction and genes v_m, v_n, v_o being intermediary regulators. Similarly defined is $\Omega(M_j)_\gamma$ the set of gene targets for M_j enriched in the signature γ.
We are then able to define **shadowing** of transcriptional master regulators as follows:
If $\Omega(M_j)_\gamma \subseteq \Omega(M_i)_\gamma$ then we say that M_i is **shadowing** M_j for the signature γ. It may happen that a TF gene may *appear* to be a TMR due to shadowing phenomena, i.e. it has common target enrichment with a real TMR. In such cases we call that TF a **Shadow Regulator** and do not consider it a TMR.

In the case of the cooperative phenomena of synergy in the regulatory programs of several TMRs, definition is as follows:

Definition 2. Following the tenets of **Definition 1** we can define **synergistic regulatory activity** (or simply **synergy**) between two transcriptional master regulators M_i and M_j as follows:
If, for a given signature γ there exists a set $\Omega^{i,j}_{syn} = \Omega(M_i)_\gamma \bigcap \Omega(M_j)_\gamma \neq \emptyset$ then there is a **synergy** between M_i and M_j for that signature.
To account for the transcriptional effect of synergistic interactions it is useful to define two scenarios:

1. If the two TMR are **positively correlated** in their activity over their target genes (either both activate or both inhibit expression of their targets) it is possible to define a **positive synergistic regulon-set** $\Omega^{i,j,+}_{syn} = \Omega(M_i)^+_\gamma \bigcap \Omega(M_j)^+_\gamma$, as well as a **negative synergistic regulon-set** $\Omega^{i,j,-}_{syn} = \Omega(M_i)^-_\gamma \bigcap \Omega(M_j)^-_\gamma$
2. If the two TMR are **anti-correlated** in their activity over their target genes (one activates and the other one inhibits expression of their targets) then we may define **positive synergistic regulon-set** $\Omega^{i,j,+}_{syn} = \Omega(M_i)^+_\gamma \bigcap \Omega(M_j)^-_\gamma$, as well as a **negative synergistic regulon-set** $\Omega^{i,j,-}_{syn} = \Omega(M_i)^-_\gamma \bigcap \Omega(M_j)^+_\gamma$. Here, by convention, the sign of the synergistic interaction has been established based on the action of M_i. In all the cases synergistic interactions are referred to the molecular signature γ

Synergistic transcriptional regulation activity involving more than two master regulators can also be defined straightforwardly as the **multiple set intersection**.

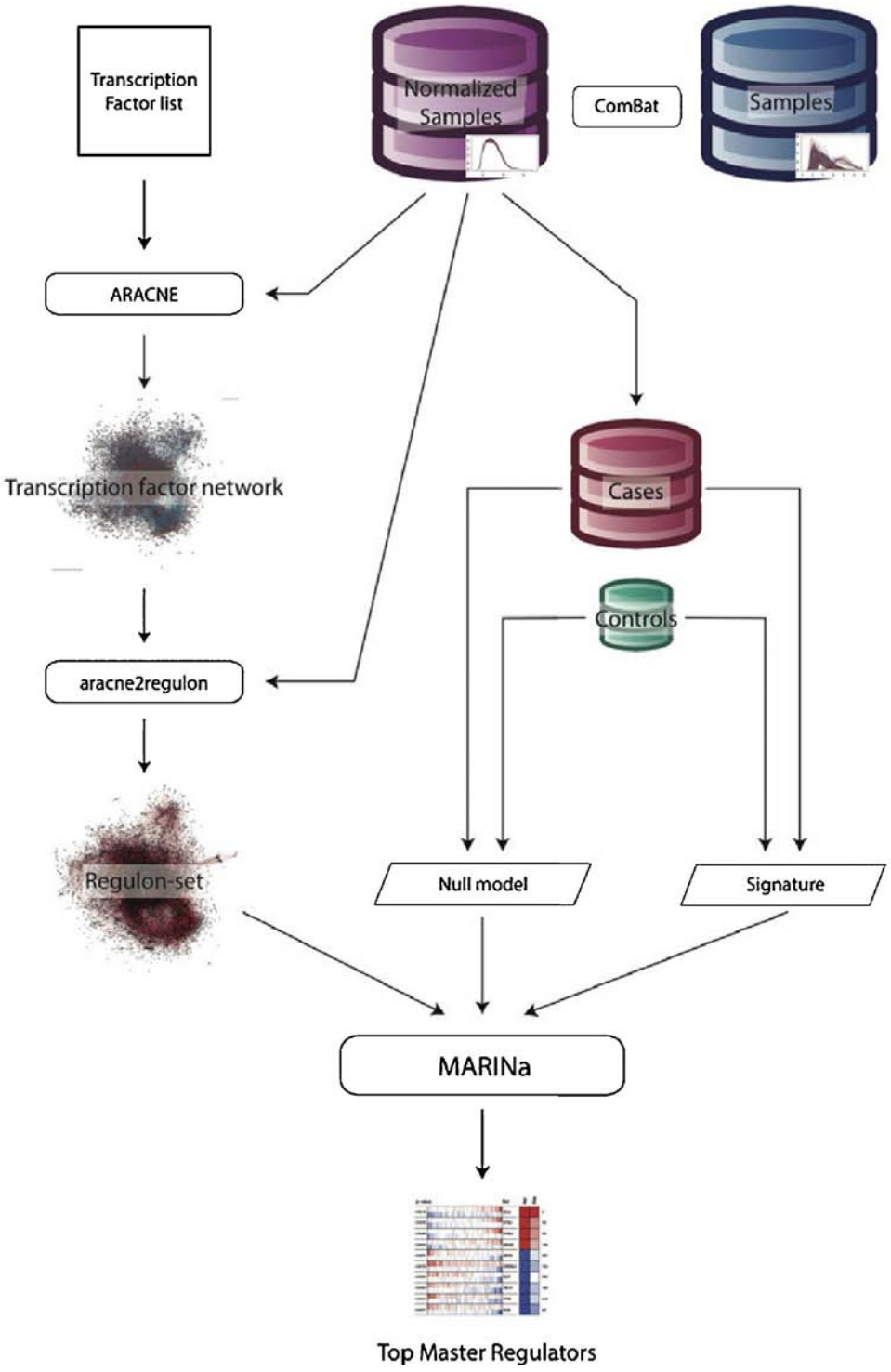


Fig. 2. MARINa pipeline. After data normalization, statistical dependency relations for all TFs versus the whole-genome breast cancer samples were inferred by using *ARACNE*. Then we determined the direction of the regulation of these TF calculating Spearman's correlations between each TF with its targets using *aracne2regulon* function from the *ssmarina* package. (2). Separately, in the set of samples divided by cases and controls, differentially expressed genes (DEGs) were chosen as the molecular signature. *ssmarina* generates a null-model by using a set of randomly chosen molecular signatures and estimates the corresponding set of *p*-values. These three elements (regulon-set, molecular signature and null model) are the inputs of *MARINa*. The algorithm computes the enrichment of each regulon, given they include at least 20 target genes, on the tails of the genome-wide expression signature (Lefebvre et al., 2010).

2.6. Causal network analysis

Causal network (CN) analysis was performed with the *Ingenuity Pathway Analysis* method ((IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity). IPA generates CNs relying on a highly curated knowledge-based source (the *Ingenuity Knowledge Base* (IKB)). IKB reports a series of *experimentally observed* cause-effect relationships related to transcription, expression, activation, molecular modification, binding events and transport processes. Since these interactions have been experimentally measured they can be associated with a definite direction of the causal effect, either activation or inhibition of the above-mentioned processes at a

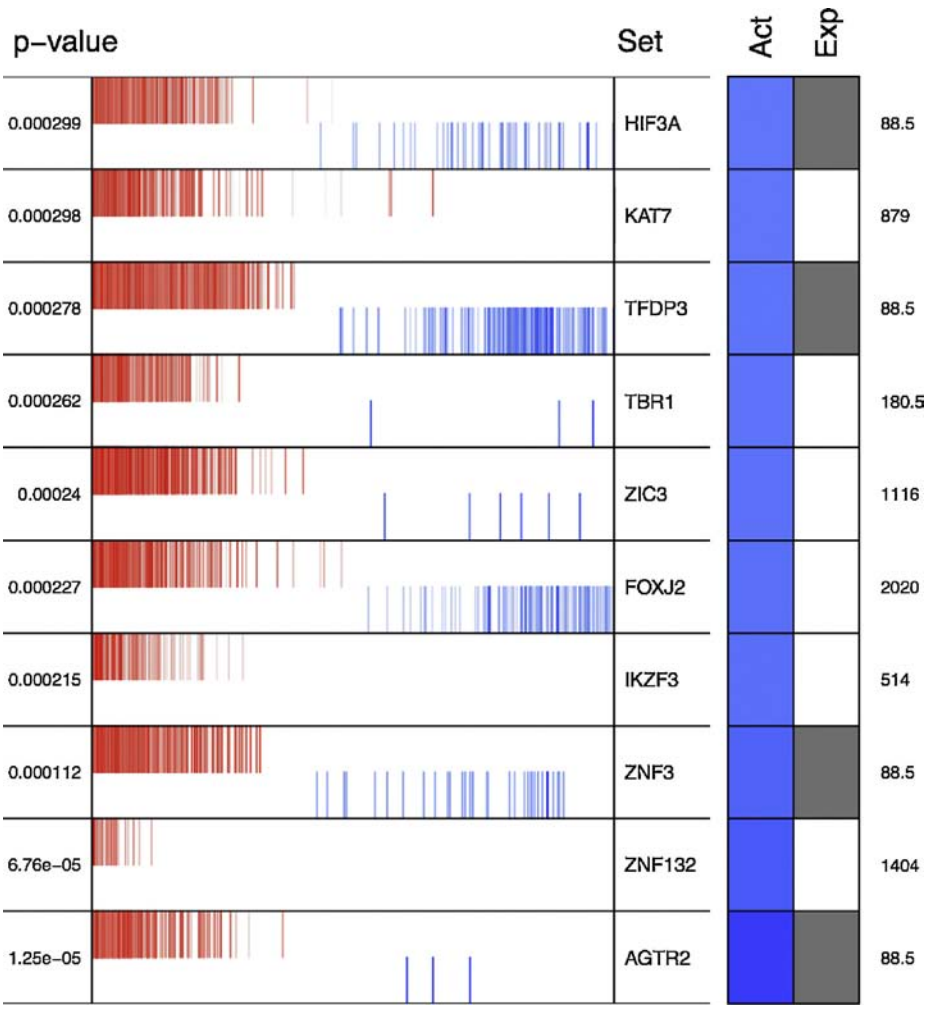


Fig. 3. Top-10 Master Regulators inferred from primary breast cancer tissue microarrays. MARINa plot of top-10 master regulators resulting from this analysis. This plot shows the projection of the repressed (blue color) and activated (red color) targets for each TF (vertical lines) on the GES (X-axis), where the genes in the GES were rank-sorted from the one most underexpressed to the one most overexpressed in the healthy vs. tumor conditions. ARACNE *p*-value for these results was 10^{-40}. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

whole genome network-wide level. Further methodological details of the CN analysis that we performed are described in (Krämer et al. (2014) and Espinal-Enríquez et al. (2015)).

This analysis was performed over the top-100 TMRs set and all their regulons. By choosing regulons for 100 TMRs we attempted to collect >1000 genes in all cases. With this number of TMRs we ensured having >4000 genes in most analyses. Network information was supplied with differential expression analysis between tumoral and healthy samples, which defined the input for the CN study. Differentially expressed genes (DEGs) were calculated by means of the *limma* algorithm (Ritchie et al., 2015). For this causal network analysis, only genes with expression log fold-change >1 and *p*-values <0.0001 were taken into account. By defining values of log fold-change >1 and a *p*-value <0.0001 a sufficiently significant difference between treatments was pursued. It must be noted that this is actually the true filter for the selected genes in the top 100 regulons.

3. Results and discussion

In this work we inferred a transcriptional regulatory network that is based on mutual information for 14,500 genes and 880 microarray gene expression samples corresponding to biopsy-captured tissue in breast cancer patients and controls (a comparative table of inferred networks with the four threshold values is presented in Supplementary Material 3). The direction of the regulation of the Transcription Factors (TFs) was inferred based on the expression of the target genes (a comparative table between regulon-sets is shown in Supplementary Material 3); TFs with the highest number of targets in the molecular signature of a given phenotype were chosen. We obtained a list of Transcriptional Master Regulators (TMRs). We also performed a shadow analysis and a synergy analysis in order to find those genes that could co-regulate subsets of molecular signature genes. Finally, with these results, we analyzed the genes in terms of the IPA-KB canonical pathways and cancer-related networks.

Some of the TMRs we found are broadly known to be key players in the development of cancerous phenotypes. Another subset of TMRs however, has unknown functions regarding the progression of the disease. The analysis presented here reveals some important issues that we discuss in following subsections.

3.1. MARINa and batch effect

As mentioned before, the batch effect control is a fundamental step in the management of expression microarrays (Grass,

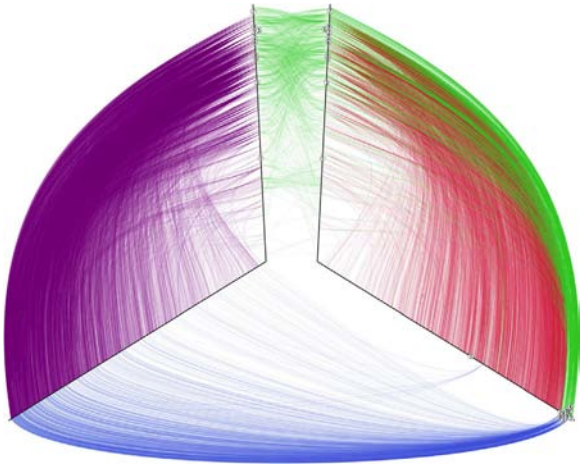


Fig. 4. Hive plot network visualization of the regulon-set with a ***p***-value of = 10^{-40}. Both vertical axes contain the full set of TFs. The bottom right axis contains the MS. The bottom left axis contains the non-differentiated genes. Green curves represent interactions among TFs. Red curves represent interactions among TFs and non TF targets in the MS. Blue curves represent interactions among TFs in the MS and targets in the non-differentiated set. Purple curves represent interactions among TFs and the non-differentiated set. Nodes in the axes are sorted by their degree, depicting the most connected genes in the outermost side of each axis and the less connected towards the center of the figure. Numbers indicate the genes of the Top10 TMRs calculated with MARINA: (0) AGTR2, (1) ZNF132, (2) ZNF3, (3) IKZF3, (4)FOXJ2, (5) ZIC3, (6) TBR1, (7) TFDP3 (8) HIF3A and (9) KAT7. For graphical purposes, axes with mutual information less than 0.85 were eliminated.

2009). Given the complexity of transcriptional regulation (Lim et al., 2009), the search for Transcriptional Master Regulators (TMRs) is extremely sensitive to batch effects. Since MARINa uses fine details of transcriptional expression, small modifications in the structure of the data may entail important changes in the final results. Expression microarrays used to perform this work were normalized with the ComBat algorithm. It is important to stress that the algorithm used by the TCGA consortium to correct batch effect is also ComBat, hence, the algorithm performed by us is reliable in order to obtain a dataset with a minimum batch effect.

3.2. Master regulators

The top-10 TMRs according to the MARINa algorithm with a *p*-value = 1^{-40} are shown in Fig. 3. TMRs for other *p*-values are shown in Supplementary Material 4.

The TMR with the best score is the angiotensin receptor 2, AGTR2. This gene has been shown to mediate programmed cell death in human leiomyosarcoma cells (Zhao et al., 2015). Furthermore, downregulation of AGTR2 is related to cell growth and evasion of apoptosis in breast cancer (De Paepe et al., 2002). This gene is underexpressed in our breast cancer samples; hence, this could be indicative of a diminished apoptotic process in the samples.

The second highest scored TMR was the zinc finger protein 132, ZNF132. Downregulation of this gene is associated with aberrant promoter hypermetylation and poor prognosis in prostate cancer (Abildgaard et al., 2012). In our samples this gene is also underexpressed.

Other TMRs obtained by this list have different cancer-related functions. For instance, HIF3A is a well-known negative regulator of tumorigenesis (Hara and Kondo, 2011; Heikkila et al., 2011; Ando et al., 2013). In addition, TFDP3 inhibits p53-mediated apoptosis Tian et al. (2007), Ingram et al. (2011), Ma et al. (2014) and Qiao et al. (2007). This behavior is consistent in our breast cancer samples, since TFDP3 is also downregulated. FOXJ2 is another TMR which in our samples is underexpressed. Overexpression of this gene decreases breast cell cancer migration (Wang et al., 2012). Other top TMRs are also involved in regulation of other cancer types. For example, IKZF3 (AIOLOS) is a gene whose overexpression inhibits cell proliferation in Nalm-6 cells (Zhuang et al., 2014).

Fig. 4 shows a novel visualization of the regulon-set network, called a Hive Plot (Krzywinski et al., 2011). In a hive plot nodes are set along axes that represent different categories; we chose to place the nodes so that the outermost have the highest degree. Edges among the nodes are represented as curves that connect them.

Both vertical axes contain the full set of TFs. The bottom right axis contains the MS. The bottom left axis contains the non-differentiated genes. Green curves represent interactions among TFs. Red curves represent interactions among TFs and non TF targets in the MS. Blue curves represent interactions among TFs in the MS and targets in the non-differentiated set. Purple curves represent interactions among TFs and the non-differentiated set.

Some interesting features of the regulon-set network are clearly revealed. For instance, a small number of TFs controls the majority of the molecular signature genes. Since this visualization is a non-directed network, some TFs that appear in the MS axis also regulate TFs in the vertical axes. Some of the Top-10 TMRs that appear in the MS axis control a subset of TFs in the vertical axes which in turn control the rest of the molecular signature.

This figure may give some hints about the relevance of the differentially expressed TFs as phenotype inducers. Supplementary Material 4 includes a hive plot for the TCGA data which shows the same pattern. It is worth mentioning that TCGA data includes only invasive breast cancers. This should be one of the reasons for the differences observed between GEO and TCGA results, caused by sample heterogeneity of various stages and tumor grades between them (results for four *p*-values of TCGA MARINa analysis are shown in Supplementary Material 4). Further investigation regarding this last point is still necessary, but it is important to stress that with this method we can study relevant features corresponding to the regulation of the transcription in eukaryote cells.

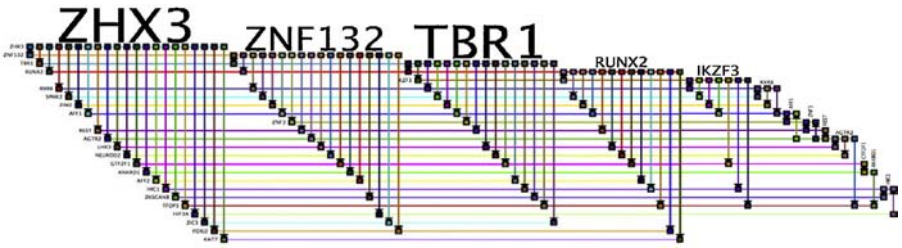


Fig. 5. Network generated with the gene pairs resulting from the shadow analysis. In this visualization (BioFabric (Longabaugh, 2012)) nodes are represented as one-dimensional horizontal lines, one per row, and edges are vertical lines, one per column. In this particular network, the direction of the interaction A → B signifies that gene A contains the targets of gene B. Larger gene tags indicate that such TF contains more of the targets of other master regulators. Since the interaction is directed, it can be seen that the gene that shadows more TFs is ZHX3, followed by ZNF132 and so on.

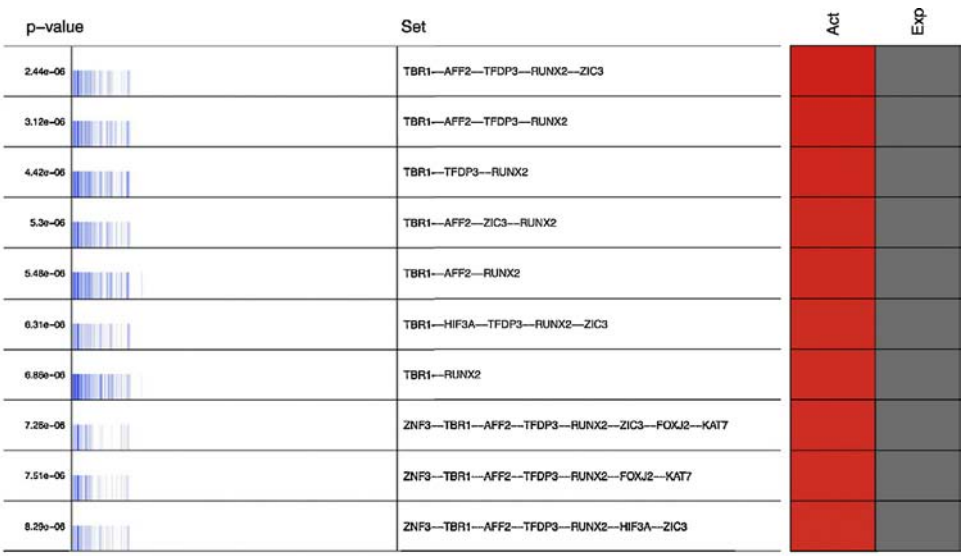


Fig. 6. Synergy analysis graph. The first 10 subsets of TFs that regulate synergistically the molecular signature of our dataset are shown. The column *Set* shows genes that co-regulate the same set of targets.

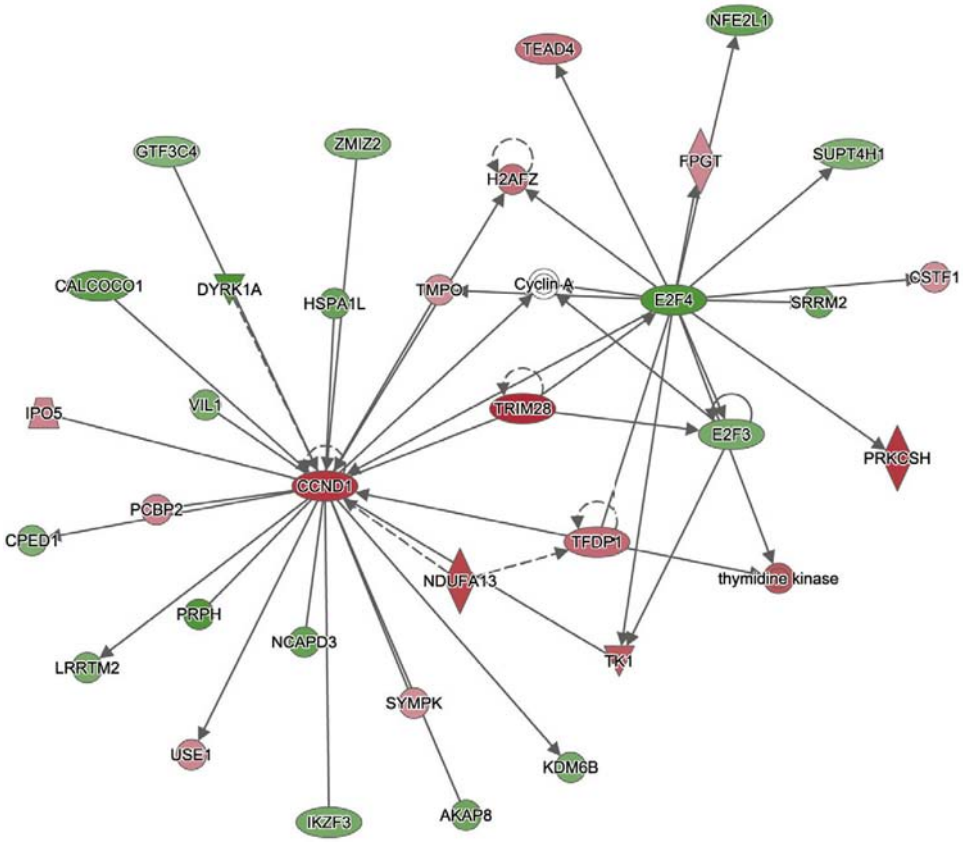


Fig. 7. IPA-KB's cancer-related network. Red molecules are overexpressed, green molecules are underexpressed. Color intensity indicates the difference between breast cancer tumor and normal samples.

3.3. Shadow analysis

A transcription factor TF_n with a large number of targets could appear to be more important than another TF_m. But it could be the case that TF_n is inferred as differentially active, while it is not, simply because it shares genes with a truly differentially active TF_m. In this case, TF_n is a "*shadow*" of TF_m. MARINa includes an algorithm which compares the targets of each pair of TFs with more than N (in this case 20) targets taking into account the *shadowing* of the TF in such a way that TF_n contains the targets of TF_m (Lefebvre et al., 2010)

The results of this analysis are presented in Fig. 5, which shows that the targets of some of the top-10 genes are already contained in the targets of some other genes not previously observed from the first analysis. Such is the case of ZHX3, ZNF132, and others.

3.4. Synergy analysis

The main objective of the synergy analysis is to detect those TFs which co-regulate a given subset of the molecular signature (Carro et al., 2010; Aytes et al., 2014). Such co-regulated sets are of a highly cooperative nature. We performed the synergy analysis for the top-25 TMRs as given by their enrichment score. The results are shown in Fig. 6, which displays the top-10 sets that are regulated by these top-25 TMRs.

We can see that TBR1 and RUNX2 are indeed present in all 10 master regulator gene sets (MRGSs). A MRGS is composed by a set of transcriptional master regulators that regulates the transcription of *the same set of gene targets*. This fact makes of the transcriptional control by synergistic groups an extremely robust phenomenon. In turn, AFF2 and TFDP3 are in 7 out of 10 MRGSs. Other TMRs are present in the central regulatory cores of diverse MRGSs.

3.5. Causal network and pathway analysis

For this analysis, the first 100 TMRs obtained by MARINa with all their ARACNE-inferred targets were used. We analyzed those genes with *IPA-KB*. As a result of Causal Network Analysis we were able to find a functional module related with some well-known hallmarks of cancer (Hanahan and Weinberg, 2000, 2011). It is quite remarkable that the Top-10 TMR IKZF3 participates in both, shadow and Cell-Cycle-related-Causal Network (Fig. 7). Furthermore, regarding the cancer-related network, we have KAT7, one of our TMRs, participating in the regulation of several oncoproteins (Fig. 1 in Supplementary Material 5). In turn, in the case of apoptosis we can highlight the action of the TFDP3 master regulator (Fig. 2 in Supplementary Material 5).

3.6. Synergy analysis highlights a core of downregulated TMRs

As observed in Fig. 6, TBR1, RUNX2 and TFDP3 are present in almost all of the top-10 MRGSs. This fact may be attributed to the concomitant regulation of them in a particular set of target genes. The fact that the same (or almost the same) set of TMRs is able to regulate a number of different MRGSs may be indicative of a strongly robust regulatory program, one that may be able to account for the *establishment of phenotypes*. A large number of genes involved on a wide variety of phenotype-defining processes are indeed regulated by a handful of coordinated TMRs.

The principal effects of the synergystic genes related to the hallmarks of cancer (Hanahan and Weinberg, 2011), namely, dysregulation of cell cycle, inhibition of apoptosis, migration, angiogenesis and proliferation, can be discussed now. For example, KAT7, which is present in two synergystic groups, is related with cell cycle dysregulation (Siriwardana et al., 2014). TFDP3, which was previously commented, is present in 7 out of 10 groups and its dysregulation is linked to evasion of apoptosis. RUNX2 and also FOXJ2 are involved in migration processes (Boregowda et al., 2014; Wang et al., 2014, 2012). HIF3 is involved in the process of angiogenesis (Ando et al., 2013). Finally, the zinc finger protein ZNF3 is related to proliferation (Gao et al., 2008). We also found two TMRs that are not annotated in cancer-related literature. This is the case of TBR1 and ZIC3, which are involved in brain and heart development (Bulfone et al., 1995; Cowan et al., 2014). We contend that these findings must be further studied in order to understand the basis of regulation of transcription in eukaryotes, specifically in the context of cancer.

4. Conclusions

In this study, we implemented a method based on the combination of gene regulatory network inference and gene set enrichment analysis algorithms. We did this across a set of gene expression experiments capable of inducing a molecular signature that distinguishes cases from controls. In this approach, cases were samples belonging to biopsy-captured primary breast cancer tissue while controls were healthy breast tissue. This algorithm called Master Regulator Inference Analysis (MARINa) has allowed us to unveil a series of transcriptional regulatory phenomena that may lie behind the establishment of the tumor phenotype. For instance, we were able to recover cancer-related enriched functional pathways for the networks conformed by the set of transcriptional master regulators and their direct targets. Causal network analysis also led us to the discovery of a quite active functional module involving known hallmarks of cancer such as proliferation, apoptosis evasion and invasiveness.

Another important issue that we want to stress is the fact that in this work we implemented the entire procedure in unclassified breast cancer samples. The approach of this method lies in finding generalities in comparing cases and controls. Since we strongly believe that a more accurate description of the phenomena will be improved with subtypified samples, the next step on this research avenue is precisely to separate the samples by subtype. Another issue worth mentioning is that in order to have robust results a large number of samples is necessary. Most healthy tissue samples came from breast cancer patients' adjacent tissue without traces of the disease, but we also included non-paired samples in order to have a greater number of them and make our statistics more robust.

Regarding the structure of the gene regulatory network induced by the action of transcriptional master regulators, we observed two interesting and related phenomena. The first one, termed *shadowing* refers to the fact that there are some master regulators that are able to control the transcription of the targets of others, thus adding to the robustness of the regulatory program. Also, we were able to observe synergistic activity over some gene sets termed master regulator gene sets that show the concomitant regulatory action of sets of top master regulators over very large sets of targets that are collectively regulated by them. Along with shadowing, this phenomenon may be contributing to the establishment and maintenance of robust conditions as was discussed in the context of tumor associated phenotypes.

For these reasons, we believe that master regulator inference analysis may become a very important methodological tool in the molecular study of complex cellular phenotypes, particularly those related with disease. As such, they may add to the tool set in computational cancer biology.

Acknowledgements

Funding statement: This work was supported by CONACYT (grant no. 179431/2012), as well as by federal funding from the National Institute of Genomic Medicine (Mexico). This work has been submitted to comply with the requirements of the *Ph.D. program in Biological Sciences at the Universidad Nacional Autónoma de México* of Hugo Antonio Tovar-Romero (HATR-Hugo Tovar). HATR is grateful to CONACYT for the financial support provided via a PhD Scholarship (grant no. 202668). Authors are grateful to the anonymous reviewers for their guiding comments.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.compbiolchem.2015.08.007.

References

Abildgaard, M.O., Borre, M., Mortensen, M.M., Ulhøi, B.P., Tørring, N., Wild, P., Kristensen, H., Mansilla, F., Ottosen, P.D., Dyrskjøt, L., Ørntoft, T.F., S?rensen, K.D., 2012. Downregulation of zinc finger protein 132 in prostate cancer is associated with aberrant promoter hypermethylation and poor prognosis. Int. J. Cancer 130 (February (4)), 885–895.

Affara, M., Sanders, D., Araki, H., Tamada, Y., Dunmore, B.J., Humphreys, S., Imoto, S., Savoie, C., Miyano, S., Kuhara, S., Jeffries, D., Print, C., Charnock-Jones, D.S., 2013. Vasohibin-1 is identified as a master-regulator of endothelial cell apoptosis using gene network analysis. BMC Genomics 14 (January (1)), 1.

Alvarez, M.J., 2013. ssmarina: Single sample-optimized Master Regulator Analysis. R package version 1.01., http://dx.doi.org/10.6084/m9.figshare.785718.

Ando, H., Natsume, A., Iwami, K., Ohka, F., Kuchimaru, T., Kizaka-Kondoh, S., Ito, K., Saito, K., Sugita, S., Hoshino, T., Wakabayashi, T., 2013. A hypoxia-inducible factor (HIF)-3 splicing variant, HIF-34 impairs angiogenesis in hypervascular malignant meningiomas with epigenetically silenced HIF-34. Biochem. Biophys. Res. Commun. 433 (March (1)), 139–144.

Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T., Albrecht, M., 2008. Computing topological parameters of biological networks. Bioinformatics 24 (2), 282–284.

Aytes, A., Mitrofanova, A., Lefebvre, C., Alvarez, M.J., Castillo-Martin, M., Zheng, T., Eastham, J.A., Gopalan, A., Pienta, K.J., Shen, M.M., Califano, A., Abate-Shen, C., 2014. Cross-species regulatory network analysis identifies a synergistic interaction between FOXM1 and CENPF that drives prostate cancer malignancy. Cancer Cell 25 (May (5)), 638–651.

Baca-López, K., Mayorga, M., Hidalgo-Miranda, A., Gutiérrez-Nájera, N., Hernández-Lemus, E., 2012. The role of master regulators in the metabolic/transcriptional coupling in breast carcinomas. PLoS ONE 7 (August (8)), e42678.

Baca-Lopez, K., Correa-Rodriguez, M.D., Flores-Espinosa, R., García-Herrera, R., Hernández-Armenta, C.I., Hidalgo-Miranda, A., Huerta-Verde, A.J., Imaz-Rosshandler, I., Martinez-Rubio, A.V., Medina-Escareno, A., Mendoza-Smith, R., Rodriguez-Dorantes, M., Salido-Guadarrama, I., Hernández-Lemus, E., Rangel-Escareno, C., 2014. A 3-state model for multidimensional genomic data integration. Syst. Biomed. 1 (October (2)), 122–129.

Ballman, K.V., Grill, D.E., Oberg, A.L., Therneau, T.M., 2004. Faster cyclic loess: normalizing RNA arrays via linear models. Bioinformatics 20 (November (16)), 2778–2786.

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., di Bernardo, D., 2007. How to infer gene networks from expression profiles. Mol. Syst. Biol. 3.

Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R., Califano, A., 2005. Reverse engineering of regulatory networks in human B cells. Nat. Genet. 37 (March (4)), 382–390.

Boregowda, R.K., Olabisi, O.O., Abushahba, W., Jeong, B.S., Haenssen, K.K., Chen, W., Chekmareva, M., Lasfar, A., Foran, D.J., Goydos, J.S., Cohen-Solal, K.A., 2014. RUNX2 is overexpressed in melanoma cells and mediates their migration and invasion. Cancer Lett. 348 (June (1–2)), 61–70.

Bulfone, A., Smiga, S.M., Shimamura, K., Peterson, A., Puelles, L., Rubenstein, J.L., 1995. T-brain-1: a homolog of Brachyury whose expression defines molecularly distinct domains within the cerebral cortex. Neuron 15 (July (1)), 63–78.

Carro, M.S., Lim, W.K., Alvarez, M.J., Bollo, R.J., Zhao, X., Snyder, E.Y., Sulman, E.P., Anne, S.L., Doetsch, F., Colman, H., Lasorella, A., Aldape, K., Califano, A., Iavarone, A., 2010. The transcriptional network for mesenchymal transformation of brain tumours. Nature 463 (January (7279)), 318–325.

Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., Liu, C., 2011. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS ONE 6 (2), e17238.

Cowan, J., Tariq, M., Ware, S.M., 2014. Genetic and functional analyses of ZIC3 variants in congenital heart disease. Hum. Mutat. 35 (January (1)), 66–75.

De Paepe, B., Verstraeten, V.M., De Potter, C.R., Bullock, G.R., 2002. Increased angiotensin II type-2 receptor density in hyperplasia, DCIS and invasive carcinoma of the breast is paralleled with increased iNOS expression. Histochem. Cell Biol. 117 (January (1)), 13–19.

Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., Viale, G., Delorenzi, M., Zhang, Y., d'Assignies, M.S., Bergh, J., Lidereau, R., Ellis, P., Harris, A.L., Klijn, J.G.M., Foekens, J.A., Cardoso, F., Piccart, M.J., Buyse, M., Sotiriou, C., TRANSBIG Consortium, 2007. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. Clin. Cancer Res. 13 (June (11)), 3207–3214.

Edgar, R., Domrachev, M., Lash, A.E., 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 30 (January (1)), 207–210.

Emmert-Streib, F., De Matos Simoes, R., Mullan, P., Haibe-Kains, B., Dehmer, M., 2014. The gene regulatory network for breast cancer: integrated regulatory landscape of cancer hallmarks. Stat. Genet. Methodol. 5.

Espinal-Enríquez, J., Mu noz-Montero, S., Imaz-Rosshandler, I., Huerta-Verde, A., Mejía, C., Hernández-Lemus, E., 2015. Genome-wide expression analysis suggests a crucial role of dysregulation of Matrix Metalloproteinases Pathway in Undifferentiated Thyroid Carcinoma. BMC Genomics 16, 207, http://dx.doi.org/10.1186/s12864-015-1372-0.

Farmer, P., Bonnefoi, H., Becette, V., Tubiana-Hulin, M., Fumoleau, P., Larsimont, D., Macgrogan, G., Bergh, J., Cameron, D., Goldstein, D., Duss, S., Nicoulaz, A.-L., Brisken, C., Fiche, M., Delorenzi, M., Iggo, R., 2005. Identification of molecular apocrine breast tumours by microarray analysis. Oncogene 24 (July (29)), 4660–4671.

Gao, J., Li, W.X., Feng, S.Q., Yuan, Y.S., Wan, D.F., Han, W., Yu, Y., 2008. A protein-protein interaction network of transcription factors acting during liver cell proliferation. Genomics 91 (April (4)), 347–355.

Grass, P., 2009. Experimental design. In: Scherer, A. (Ed.), Batch Effects and Noise in Microarray Experiments, chapter 3. John Wiley & Sons, Ltd, pp. 19–31, http://dx.doi.org/10.1002/9780470685983.ch3, ISBN 9780470685983.

Han, J.-D.J., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J.M., Cusick, M.E., Roth, F.P., Vidal, M., 2004. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. Nature 430 (June (6995)), 88–93.

Hanahan, D., Weinberg, R.A., 2000. The hallmarks of cancer. Cell 100 (1), 57–70.

Hanahan, D., Weinberg, R.A., 2011. Hallmarks of Cancer: The Next Generation. Cell 144 (5), 646–674.

Hara, S., Kondo, Y., 2011. [Hypoxia-inducible factor-3alpha as a negative regulator of tumorigenesis]. Seikagaku 83 (January (1)), 50–55.

Heikkila, M., Pasanen, A., Kivirikko, K.I., Myllyharju, J., 2011. Roles of the human hypoxia-inducible factor (HIF)-3 variants in the hypoxia response. Cell. Mol. Life Sci. 68 (December (23)), 3885–3901.

Hernández-Lemus, E., Siqueiros-García, J.M., 2013. Information theoretical methods for complex network structure reconstruction. Complex Adapt. Syst. Model. 1 (1), 8.

Hinnebusch, A.G., Natarajan, K., 2002. Gcn4p, a master regulator of gene expression, is controlled at multiple levels by diverse signals of starvation and stress. Eukaryotic Cell 1 (February (1)), 22–32.

Ingram, L., Munro, S., Coutts, A.S., La Thangue, N.B., 2011. E2F-1 regulation by an unusual DNA damage-responsive DP partner subunit. Cell Death Differ. 18 (January (1)), 122–132.

Ivshina, A.V., George, J., Senko, O., Mow, B., Putti, T.C., Smeds, J., Lindahl, T., Pawitan, Y., Hall, P., Nordgren, H., Wong, J.E.L., Liu, E.T., Bergh, J., Kuznetsov, V.A., Miller, L.D., 2006. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. Cancer Res. 66 (November (21)), 10292–10301.

Johnson, W.E., Li, C., Rabinovic, A., 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8 (January (1)), 118–127.

Krämer, A., Green, J., Pollard, J., Tugendreich, S., 2014. Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics (Oxford, England) 30 (February (4)), 523–530.

Krzywinski, M., Birol, I., Jones, S.J., Marra, M.A., 2011. Hive plots-rational approach to visualizing networks. Briefings Bioinform. 13 (December (5)), bbr069–bbr644.

Lander, E.S., Linton, L.M., et al., 2001. Initial sequencing and analysis of the human genome. Nature 409 (February (6822)), 860–921.

Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., Irizarry, R.A., 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet. 11 (October (10)), 733–739.

Lefebvre, C., Rajbhandari, P., Alvarez, M.J., Bandaru, P., Lim, W.K., Sato, M., Wang, K., Sumazin, P., Kustagi, M., Bisikirska, B.C., Basso, K., Beltrao, P., Krogan, N., Gautier, J., Dalla-Favera, R., Califano, A., 2010. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. Mol. Syst. Biol. 6, 1–10.

Lim, W.K.W., Lyashenko, E.E., Califano, A.A., 2009. Master regulators used as breast cancer metastasis classifier. IEEE Audio Electroacoust. Newsl., 504–515.

Liu, G., Loraine, A.E., Shigeta, R., Cline, M., Cheng, J., Valmeekam, V., Sun, S., Kulp, D., Siani-Rose, M.A., 2003. NetAffx: affymetrix probesets and annotations. Nucleic Acids Res. 31 (January (1)), 82–86 [PubMed Central:PMC165568] [PubMed:12519953].

Liu, R., Wang, X., Chen, G.Y., Dalerba, P., Gurney, A., Hoey, T., Sherlock, G., Lewicki, J., Shedden, K., Clarke, M.F., 2007. The prognostic role of a gene signature from tumorigenic breast-cancer cells. N. Engl. J. Med. 356 (January (3)), 217–226.

Longabaugh, W.J., 2012. Combing the hairball with BioFabric: a new approach for visualization of large networks. BMC Bioinform. 13 (October (1)), 275.

Ma, Y., Xin, Y., Li, R., Wang, Z., Yue, Q., Xiao, F., Hao, X., 2014. TFDP3 was expressed in coordination with E2F1 to inhibit E2F1-mediated apoptosis in prostate cancer. Gene 537 (March (2)), 253–259.

Margolin, A.A., Nemenman, I., Wiggins, C., Stolovitzky, G., Califano, A., 2004. On the reconstruction of interaction networks with applications to transcriptional regulation, arXiv.org.

Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R., Califano, A., 2006. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinform. 7 (Suppl. 1), S7.

McCall, M.N., Bolstad, B.M., Irizarry, R.A., 2010. Frozen robust multiarray analysis (fRMA). Biostatistics 11 (April (2)), 242–253.

Medvedovic, J.J., Ebert, A.A., Tagoh, H.H., Busslinger, M.M., 2011. Pax5: a master regulator of B cell development and leukemogenesis. Adv. Immunol. 111, 179–206.

Miller, L.D., Smeds, J., George, J., Vega, V.B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E.T., Bergh, J., 2005. From the Cover: an expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc. Natl. Acad. Sci. USA 102 (September (38)), 13550–13555.

Minn, A.J., Gupta, G.P., Siegel, P.M., Bos, P.D., Shu, W., Giri, D.D., Viale, A., Olshen, A.B., Gerald, W.L., Massagué, J., 2005. Genes that mediate breast cancer metastasis to lung. Nature 436 (July (7050)), 518–524.

Mullen, A.C., Orlando, D.A., Newman, J.J., Lovén, J., Kumar, R.M., Bilodeau, S., Reddy, J., Guenther, M.G., DeKoter, R.P., Young, R.A., 2011. Master transcription factors determine cell-type-specific responses to TGF-β signaling. Cell 147 (October (3)), 565–576.

Pau Ni, I.B., Zakaria, Z., Muhammad, R., Abdullah, N., Ibrahim, N., Aina Emran, N., Hisham Abdullah, N., Syed Hussain, S.N.A., 2010. Gene expression patterns distinguish breast carcinomas from normal breast tissues: the Malaysian context. Pathol. Res. Pract. 206 (April (4)), 223–228.

Pawitan, Y., Bjöhle, J., Amler, L., Borg, A.-L., Egyhazi, S., Hall, P., Han, X., Holmberg, L., Huang, F., Klaar, S., Liu, E.T., Miller, L., Nordgren, H., Ploner, A., Sandelin, K., Shaw, P.M., Smeds, J., Skoog, L., Wedrén, S., Bergh, J., 2005. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. Breast Cancer Res. 7 (6), R953–R964.

Qiao, H., Di Stefano, L., Tian, C., Li, Y.Y., Yin, Y.H., Qian, X.P., Pang, X.W., Li, Y., McNutt, M.A., Helin, K., Zhang, Y., Chen, W.F., 2007. Human TFDP3, a novel DP protein, inhibits DNA binding and transactivation by E2F. J. Biol. Chem. 282 (January (1)), 454–466.

R. Development Core Team, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.r-project.org/.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, URL http://www.statsci.org/smyth/pubs/limmaPreprint.pdf.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13 (11), 2498–2504.

Shimoni, Y., Alvarez, M., 2013. TF list., http://dx.doi.org/10.6084/m9.figshare.871524.

Siriwardana, N.S., Meyer, R., Havasi, A., Dominguez, I., Panchenko, M.V., 2014. Cell cycle-dependent chromatin shuttling of HBO1-JADE1 histone acetyl transferase (HAT) complex. Cell Cycle 13 (12), 1885–1901.

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., Desmedt, C., Larsimont, D., Cardoso, F., Peterse, H., Nuyten, D., Buyse, M., Van de Vijver, M.J., Bergh, J., Piccart, M., Delorenzi, M., 2006. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J. Natl. Cancer Inst. 98 (February (4)), 262–272.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. PNAS 102 (43), 15545–15550.

Sun-Kin Chan, S., Kyba, M., 2013. What is a master regulator? J. Stem Cell Res. Ther. 03 (02.).

Tian, C., Lv, D., Qiao, H., Zhang, J., Yin, Y.H., Qian, X.P., Wang, Y.P., Zhang, Y., Chen, W.F., 2007. TFDP3 inhibits E2F1-induced, p53-mediated apoptosis. Biochem. Biophys. Res. Commun. 361 (September (1)), 20–25.

Tripathi, A., King, C., de la Morenas, A., Perry, V.K., Burke, B., Antoine, G.A., Hirsch, E.F., Kavanah, M., Mendez, J., Stone, M., Gerry, N.P., Lenburg, M.E., Rosenberg, C.L., 2008. Gene expression abnormalities in histologically normal breast epithelium of breast cancer patients. Int. J. Cancer 122 (April (7)), 1557–1566.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., Luscombe, N.M., 2009. A census of human transcription factors: function, expression and evolution. Nat. Rev. Genet. 10 (April (4)), 252–263.

Venter, J.C., Adams, M.D., et al., 2001. The sequence of the human genome. Science 291 (February (5507)), 1304–1351.

Wang, Y., Yang, S., Ni, Q., He, S., Zhao, Y., Yuan, Q., Li, C., Chen, H., Zhang, L., Zou, L., Shen, A., Cheng, C., 2012. Overexpression of forkhead box J2 can decrease the migration of breast cancer cells. J. Cell. Biochem. 113 (August (8)), 2729–2737.

Wang, W., Chen, B., Zou, R., Tu, X., Tan, S., Lu, H., Liu, Z., Fu, J., 2014. Codonolactone, a sesquiterpene lactone isolated from Chloranthus henryi Hemsl, inhibits breast cancer cell invasion, migration and metastasis by downregulating the transcriptional activity of Runx2. Int. J. Oncol. 45 (November (5)), 1891–1900.

Zhao, Y., Lutzen, U., Fritsch, J., Zuhayra, M., Schutze, S., Steckelings, U.M., Recanti, C., Namsoleck, P., Unger, T., Culman, J., 2015. Activation of intracellular angiotensin AT, receptors induces rapid cell death in human uterine leiomyosarcoma cells. Clin. Sci. 128 (May (9)), 567–578.

Zhuang, Y., Li, D., Fu, J., Shi, Q., Lu, Y., Ju, X., 2014. Overexpression of AIOLOS inhibits cell proliferation and suppresses apoptosis in Nalm-6 cells. Oncol. Rep. 31 (March (3)), 1183–1190.