UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
POSGRADO DE CIENCIAS BIOQUÍMICAS
INSTITUTO NACIONAL DE MEDICINA GENÓMICA

ANÁLISIS PROTEÓMICO COMPUTACIONAL DE REDES REGULATORIAS EN CÁNCER DE MAMA

TESIS

QUE PARA OPTAR POR EL GRADO DE

MAESTRO EN CIENCIAS BIOQUÍMICAS

PRESENTA:

MARTÍN RÜHLE BOGGI

TUTOR

DR. ENRIQUE HERNÁNDEZ LEMUS
INSTITUTO NACIONAL DE MEDICINA GENÓMICA

MIEMBROS DEL COMITÉ TUTOR

DR. MAXIMINIO ALDANA
INSTITUTO DE CIENCIAS FÍSICAS, UNAM
DR. SERGIO ENCARNACIÓN GUEVARA
CENTRO DE CIENCIAS GENÓMICAS, UNAM

CIUDAD UNIVERSITARIA, CD. MX., SEPTIEMBRE, 2020

Agradecimientos

La curiosidad y la búsqueda del conocimiento son una parte importante en las cosas que le dan sentido a mi vida. Por ello, quiero agradecerle a todas las personas que me han ayudado, y lo siguen haciendo, para seguir el camino de la ciencia. Entre ellos y ellas quiero destacar a mis padres, Silvina, Fernando y Eduardo que no solo me incentivaron a seguir este camino tan hermoso, sino que también me han apoyado incondicionalmente en todo y se han asegurado que nunca me faltara nada para ser feliz, y a Cande, mi compañera, que me enseña y me ayuda ser mejor todos los días.

Los incentivos que recibí e hicieron posible mis estudios también vinieron por parte de la educación pública, gratuita y de calidad argentina durante mi carrera de grado, y mexicana en los dos últimos años de posgrado. Esto abre un abanico interminable de agradecimientos a todas las personas que hicieron posible mis estudios en México que voy a intentar mencionar. Van desde la ayuda inagotable que me otorgaron desde posgrados de la UNAM, para poder inscribirme y obtener la beca que hizo esto posible, mis compañeros y tutores del grupo de investigación del INMEGEN, que me apoyaron y me guiaron en mi proyecto, hasta todo el pueblo mexicano, que no solo luchó por la educación pública, gratuita, y de calidad de la que tuve suerte de formar parte, sino que siempre me trataron como un hermano. En esto último reside mi esperanza para un futuro más justo, donde no importe de donde venimos ni quienes somos, podamos tener las oportunidades para aspirar a ser quienes querramos, y así podamos tener todas y todos una vida digna.

Resumen

Las proteínas regulan y forman parte de gran cantidad de procesos biológicos, pero hasta la fecha, los perfiles tumorales se han centrado en la información genética. Aquí intentamos expandir este perfil a través del análisis de datos públicos de proteomas de muestras de cáncer de mama, los cuales incluyen el análisis de redes de co-expresión de información mutua. Pudimos observar que existen procesos biológicos particulares asociados a las comunidades de estas redes y cómo se pierden algunos fenómenos de co-expresión transcripcional a nivel de proteína. Este tipo de análisis de datos y redes son un recurso amplio y muy apropiado para explorar el comportamiento celular en la investigación del cáncer.

Abstract

Proteins regulate and are part of a large number of biological processes. However, to date, tumor profiles have focused on genetic information. Here we tried to expand this profiling through analysis of open proteome data of and mutual information co-expression networks analysis. We could see that there are well disctint biological processes associated with the communities of these networks and how some transcriptional co-expression phenomena is lost at the protein level. These kind of data and networks analysis are a broad appropriate resource to explore cellular behavior and cancer research.

Índice general

Índice de figuras

Índice de tablas

Siglas y Acrónimos

BH Prueba de ajuste: Benjamini & Hochberg

CPTAC Clinical Proteomics Tumor Analysis Consorptium

CAMs Cell Adhesion Molecules

FDR False Discovery Rate

GDC Genomic Data Commons

GO Gene Ontology

HCD Higher-energy Collisional Dissociation

IP-P Interacciones Proteína-Proteína

ISB Institute for Systems Biology

KEGG Kyotos Encyclopedia of Genes and Genomes

MHC Mayor Histocompatibility Complex

PCA Principal Component Analysis

pGenes Genes inferidos a partir de la expresión protéica

RPLC Reversed Phase Liquid Chromatography

STRING Search Tool for the Retrieval of Interacting Genes/Proteins

TCGA The Cancer Genome Atlas

WebGestalt WEB-based GEne SeT AnaLysis Toolkit

Capítulo 1

Introducción

1.1. Cáncer de mama

El cáncer de mama es una enfermedad que se caracteriza por el crecimiento descontrolado de las células del tejido mamario. Las células cencerígenas en general tiene muchas características que han sido estudiadas y las diferencian de las células normales, estás características se han denominado "sellos distintivos del cáncer" o "hallmarks of cancer" en inglés [27], y se muestran en la Figura 1.1.

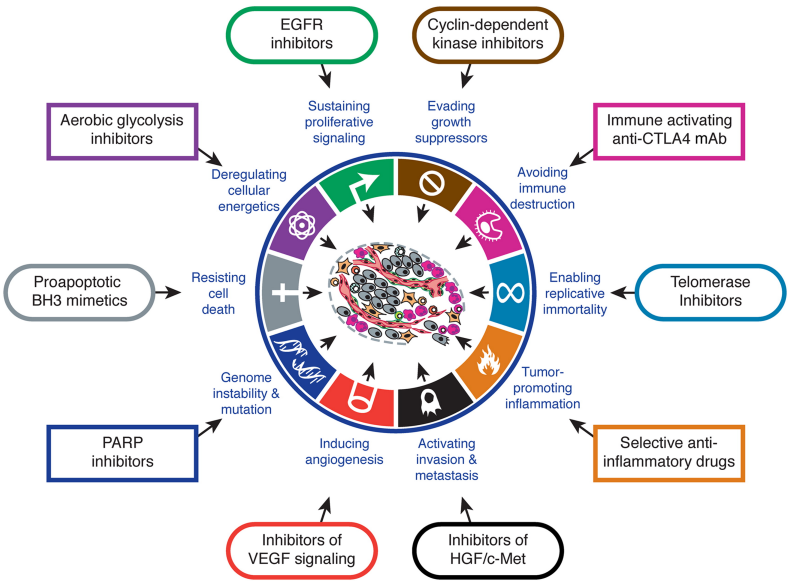


Figura 1.1: *Hallmarks of Cáncer* [27]: *Los sellos distintivos del cáncer*, son el conjunto de fenómenos o características biológicas que definen a esta enfermedad.

De acuerdo a la Organización Mundial de la Salud el cáncer de mama es el cáncer más frecuente en las mujeres a nivel mundial [44], y en México es la primera causa de muerte por cáncer en la mujer [10]. Sumado a estos importantes aspectos epidemiológicos, está el hecho de que ésta es una enfermedad altamente compleja y heterogénea, tanto en sus orígenes moleculares como en sus manifestaciones clínicas; debido a estos aspectos es fundamental mejorar la comprensión de los mecanismos moleculares detrás del desarrollo del cáncer de mama [24].

1.2. Subtipos de cáncer de mama

Como se mencionó anteriormente el cáncer de mama es una enfermedad heterogénea, con diferentes características clínicas, biológicas y moleculares [47]. La subtipificación o clasificación molecular basada en patrones de expresión genética ha demostrado que al implementarla provee un diagnóstico más preciso, así como un pronóstico más certero [2]. Motivos por los cuales ha sido incluida en las guías de práctica clínica internacionales para esta enfermedad y es una pieza central en la toma de decisiones terapéuticas [18]. El algoritmo con mayor aceptación hasta el momento fue desarrollado por Parker et al. [36], denominado PAM50 (del inglés Prediction Analysis of Microarray 50), que toma en cuenta la expresión de un conjunto de 50 genes para clasificar esta enfermedad en 4 subtipos intrínsecos diferentes: Luminal A, Luminal B, HER2-enriquecido y Basal [13]. Estos subtipos se mantuvieron significativos en los análisis multivariables que incorporaron parámetros estándar (estado del receptor de estrógeno, grado histológico, tamaño del tumor y estado de los ganglios) [36].

Aunque en este trabajo no vamos a analizar cada subtipo por separado por cuestiones del alcance del proyecto, nos parece relevante describir brevemente los subtipos tumorales.

- **Luminal A.** Es el subtipo de cáncer de mama más común, constituye alrededor del 40 % de todos los subtipos. Presenta una sobreexpresión del receptor de estrógenos y de genes regulados por dicho receptor. Por lo general no presenta una sobreexpresión de HER2, ni de genes que se relacionen con proliferación.

Este subtipo presenta el mejor pronóstico y las menores tasas de recurrencia de los 4 que se describen [70].

- **Luminal B.** Aproximadamente el 20 % de los tumores de cáncer de mama pertenecen a este subtipo. Tiene una variabilidad mayor en la expresión de receptores de estrógeno y genes relacionados al receptor HER2 en relación con el subtipo luminal A. Tiene peor pronostico y un mayor riesgo de recaídas en comparacion al luminal A [56].

- **HER2 enriquecido (HER2+).** Como su nombre lo indica se define por la sobreexpresión del receptor HER2 (derivado de su nombre en inglés, human epidermal growth factor receptor 2), involucrado en la regulación de la fisiología, crecimiento y proliferación celular . Estos tumores suelen ser negativos a receptores de estrógeno y progesterona, y tiene una peor prognosis en comparación con los subtipos luminales [28]. La sobreexpresión del oncogen HER2 ocurre entre un 25 a 30 % de los casos de cáncer de mama y se ha asociado con una mala respuesta a tratamientos hormonales [39].

- **Basal.** Tiene patrones de expresión similares a los del epitelio mamario basal, de ahí su nombre. Este subtipo representa un 20 % de los tumores mamarios aproximadamente. El 80 % de los tumores *triple negativo* (que presentan subexpresión de los receptores a estrógeno, progesterona y HER2) corresponden al subtipo basal [18]. Este subtipo se ha asociado a inestabilidad genómica, alta expresión de genes proliferativos, grados histológicos más altos, y son los más agresivos (presentan el peor pronostico) [8].

1.3. Técnologías de secuenciación de alto rendimiento

Gracias a la secuenciación de nueva generación, la gran abundancia de información de expresión génica concibe el desarrollo del enfoque teórico para el modelado de

procesos de regulación génica [53] y de esta manera establece una estimación inicial de la complejidad asociada a la biología molecular humana.

El enfoque genómico o transcriptómico nos ha provisto de información de gran utilidad, aunque se suele suponer, por cuestiones de practicidad y alcance tecnológico, que hay una relación directa entre la información de la expresión genética y la expresión protéica. Aunque este enfoque se use normalmente como estimador de la expresión de proteínas, en promedio, los datos provenientes de la expresión genética tienen baja precisión [38][22]. Esta falencia en la interpretación de los datos de expresión genética como predictores de la expresión protéica, se debe a que no toma en cuenta los diversos mecanismos de regulación intermedios que ocurren entre la transcripción y el producto final protéico, como las modificaciones post-transcripcionales del ARN mensajero o las post-traduccionales que afectan la función y la estabilidad de las proteínas [15].

La importancia del estudio de la expresión protéica reside en que son las últimas moléculas efectoras de las funciones celulares y ejecutoras de las características fenotípicas [48]. Además, el poder limitado de los biomarcadores de un solo gen o proteína, para predecir el desarrollo celular, ha aumentado la necesidad de análisis de datos masivos integrados a gran escala y con un enfoque de interacciones más globales [66].

Desafortunadamente, la generación de datos proteómicos de alta calidad está rezagada con respecto a los datos de expresión de ARN. Igualmente, los avances en la espectrometría de masas de los últimos años, como la tecnología dominante de la secuenciación protéica, nos han provisto de gran cantidad de información y particularmente ha abordado el estudio del cáncer como una de las prioridades [15], por este motivo el análisis proteómico computacional toma una relevancia preponderante a la hora de interpretar los datos a gran escala que se están generando.

1.4. Redes de coexpresión con información mutua

La magnitud de los datos ómicos en la actualidad proporciona la oportunidad de decodificar de manera alternativa el rol de las moléculas biológicas y los procesos de

los que forman parte, que caracterizan los fenotipos emergentes. En este escenario, un procedimiento común para evaluar los perfiles de expresión genética se basa en estadísticas que miden la dependencia entre variables, y las redes de coexpresión resultantes se utilizan para identificar genes funcionalmente relacionados o controlados por el mismo programa regulador [64].

En el ámbito de la genómica se suele utilizar distintos coficientes de correlación, como el de Spearman, Kendall o Pearson, para medir la dependencia entre variables, este ultimo es el más utilizado generalmente [46]. El coeficiente de Pearson es una medida de interdependencia lineal entre las variables x e y. Las variables deben de ser aleatorias y distribuirse idénticamente, además de seguir distribuciones normales.

La formula para calcular el coeficiente de correlación de Pearson es la siguiente:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Que es una covarianza estandarizada para las n observaciones emparejadas de las variables x e y, por lo que carece de unidades de medida. Toma los valores -1 y 1 cuando hay una correlación total (negativa y positiva respectivamente) entre las variables, y 0 cuando no existe relación lineal entre ellas (son independientes) [51].

El problema con el coeficiente de Pearson es que solo detecta las relaciones lineales entre variables, y la expresión de componentes biológicos, como genes o proteínas, es común que no siga una dependencia lineal [9] [19].

Luego, se podrían utilizar los coeficientes de Spearman o Kendall que no tienen restricción lineal para poder capturar las dependencias no lineales entre las variables (en nuestro caso las proteínas y genes). El problema con estos coeficientes es que están limitados a la medición de solamente la relación monótona de rankings entre variables ordinales [58].

Para capturar todas las relaciones entre la expresión de los componentes biológicos necesitamos de la medida más general de dependencia estadística entre variables, por este motivo utilizamos la información mutua.

La información mutua de dos variables aleatorias es una cantidad que mide la

reducción de la incertidumbre de una variable x debido al conocimiento del valor de otra variable y. El valor de información mutua dadas dos variables aleatorias continuas x e y se define como:

$$MI(X,Y) = \int\int fX,Y(x,y) \log\left(\frac{fX,Y(x,y)}{fX(x)fY(y)}\right) dxdy$$

El valor de información mutua siempre es positivo, un valor muy alto, indica una reducción mayor en la incertidumbre debido a la alta dependencia estadística, mientras que un valor bajo indica mayor incertidumbre entre variables. Para el caso en que MI sea 0, el dato nos indica que las dos variables son independientes.

De esta manera utilizamos la información mutua entre los datos de expresión protéica y genética para inferir redes de coexpresión y de esta manera poder analizar la relación que hay entre los distintos componentes biológicos. Aunque la cantidad de datos ha crecido en gran manera en los ultimos años gracias al desarrollo de las tecnologías de secuenciación masiva, esta forma de analizar e interpretar set de datos de expresión se ha utilizado ya en muchos estudios y se han obtenido resultados valiosos [29] [63], así como también se han capturados eventos o procesos previamente estudiados [1].

Estos análisis computacionales, junto con el desarrollo de redes de interacción, están permitiendo trasladar perfiles de expresión a modelos matemáticos para analizar los patrones de correlación fenotípica, que aparte de ayudar a entender mejor el desarrollo del cáncer pueden llegar a descubrir biomarcadores [11] o blancos farmacológicos [57].

A pesar de las altas expectativas, la investigación en proteómica relacionada con la de redes ha experimentado solamente un crecimiento moderado [61] [66]. Por lo tanto, integrar y complementar el análisis genómico con el enfoque proteómico y la aplicación de redes, es de gran relevancia para mejorar la comprensión, clasificación y diagnóstico del cáncer de mama [62].

Capítulo 2

Objetivos

2.1. Objetivo General

A través del análisis computacional genómico y proteómico mejorar el entendimiento en la perdida de regulación en cáncer de mama mediante la identificación y comparación de los procesos celulares más desregulados.

2.2. Objetivos Particulares

- Descargar y curar datos proteómicos para realizar el análisis.
- Analizar la expresión diferencial protéica entre muestras sanas y tumorales.
- Analizar la coexpresión de proteínas y transcriptos en muestras de cáncer de mama.
- Generar redes de regulación protéica y de interacción proteína-proteína.
- Mediante el análisis de la expresión protéica analizar las vías y procesos celulares más desregulados.

Capítulo 3

Metodología

En esta sección se van a presentar los distintos puntos o etapas del flujo de trabajo seguidos en este proyecto, resumidos de forma esquemática en la figura 3.1

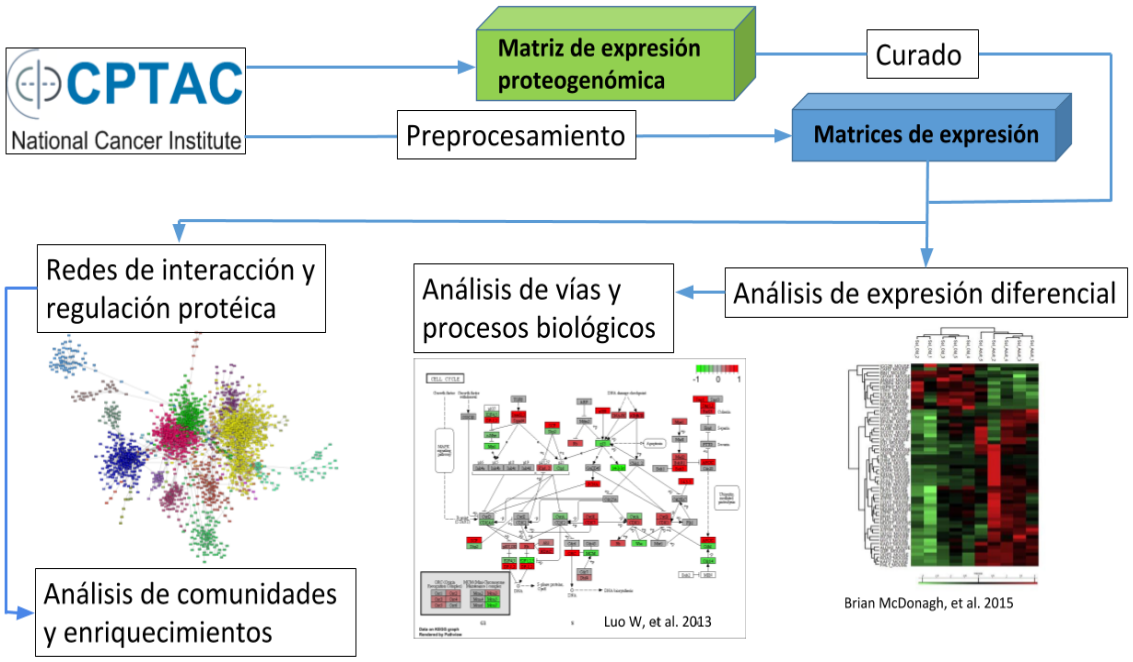


Figura 3.1: Flujo de trabajo empleado en este trabajo.

3.1. Datos

Todas las muestras tuorales, de las cuales fueron extraidos los espectros de masas utilizados en este trabajo, fueron adquiridas de pacientes recién diagnosticados con adenocarcinoma invasivo de mama que pasaron por una extracción quirúrgica y que no habían tenido tratamientos previos contra la enfermedad. Las muestras seleccionadas debian tener al menos 130 mg de material residual de peso húmedo, la cantidad objetivo para el procesamiento de proteómica entre los equipos de investigación colaboradores [31], esto también se cumplió con las 3 muestras de tejido sano, provenientes de biopsias de pacientes con tejido mamario saludable. Se consideraron 126 muestras tumorales, de las cuales 105 produjeron al menos el mínimo preespecificado de 0,7 mg de proteína total después de la extracción de proteínas que será expliacado en detalle más adelante. Fueron adquiridos los espectros de masas de los proteomas de 105 muestras de cáncer de mama, junto con 3 muestras de tejido sano y 3 réplicas de muestras tumorales, desde el portal de Clinical Proteomics Tumor Analysis Consorptium (CPTAC)[31], que a su vez adquirieron las muestras del The Cancer Genome Atlas (TCGA), ahora llamado Genomic Data Commons (GDC)[32]. Hay una representación de los 4 subtipos del cáncer de mama (Luminal A, Luminal B, Basal-like y HER2-enriched). La metodología usada para la generación de los espectros de masas realizada por el CPTAC fue:

- Preparación analítica de la muestra. Digestión tríptica y alquilación con iodoacetamida.
- Cromatografía. Cromatografía líquida de fase reversa (RPLC) a pH 10.
- Espectrometría de masas. Fragmentación por disociación por colisión inducida (HCD) y doble cuadrupolo Orbitrap.
- Estrategia para la cuantificación. Etiquetado isobárico para la cuantificación relativa y absoluta (4-plex iTRAQ).

Paralelamente se descargó una matriz de expresión proteogenómica, generada por el CPTAC a partir de los mismos datos proteómicos utilizados en este trabajo. Esta

matriz consiste en una tabla de expresión de genes inferidos a partir de la expresión protéica (de ahora en más los llamaremos pGenes). Cada muestra tiene 2 columnas, una que muestra la expresión de péptidos que son únicos para cada gen (ningún otro gen puede expresar ese péptido), denominados Unshared peptides, y la expresión de todos los posibles péptidos pertenecientes a ese gen. El requisito para que un gen sea apreciado es que tiene que estar representado por al menos dos Unshared peptides. El protocolo para la generación de la matriz de expresión proteogenómica se encuentra detallado en la publicación [48].

Para asegurarnos de la integridad de la matriz proteogenómica descargada se realizó un análisis de calidad. Para ello se utilizó la colección de paquetes de tidyverse en el entorno de trabajo de R [26].

Por último, utilizamos redes de coexpresión generadas por nuestro grupo a partir de datos de RNA-Seq de las mismas muestras con las que se generó la matriz proteogenómica. La generación de estas redes será explicada en detalla más adelante.

3.2. Análisis de datos

Para preprocesar y analizar los espectros de masas crudos descargados de CPTAC se utilizó la plataforma Trans Proteomic Pipeline (TPP), desarrollada por el Institute for Systems Biology [59], que posee las herramientas bioinformáticas necesarias para la generación de matrices de expresión a partir de los espectros de masas, y tiene la gran ventaja de que todos los programas que utiliza son de código abierto. Para cada etapa de preprocesamiento de los datos crudos, necesaria para llegar a la matriz de expresión, se utilizaron diferentes herramientas o programas que están compiladas en la plataforma del TPP. A continuación se explican, brevemente, los procesos llevados a cabo en cada etapa, y que programas fueron utilizados para llevarlos a cabo:

- Identificación de péptidos. Se utilizó X!Tandem, un open source software que mediante un algoritmo genera coincidencias de los espectros con secuencias peptídicas cargadas en motores de búsqueda, y de esta manera se generó una lista de posibles péptidos presentes en la muestra [43].

- Validación de péptidos. Mediante la utilización de PeptideProphet e Iprophet, se validaron estadísticamente los péptidos encontrados por X!Tandem mediante la utilización de motores de búsqueda más precisos [6], [16].

- Cuantificación. Ya que las muestras se encuentran etiquetadas isobáricamente, es posible determinar la concentración relativa de cada péptido en la muestra. Para ello se utilizó el software Libra, encargado de cuantificar los péptidos validados en la etapa anterior [45].

- Asignación de proteínas. Finalmente, con los péptidos ya validados y cuantificados, estos fueron asignados a las proteínas de las que forman parte con ProteinProphet, para finalmente tener la lista de proteínas que fueron capturadas mediante espectrometría de masas [5].

3.3. Análisis de expresión diferencial

Con la matriz de expresión proteogenómica se realizó el análisis de la expresión diferencial. Este consiste en determinar, mediante el paquete de R-Bioconductor, limma, Linear models for microarrays and RNA-seq [55] [41], si existe diferencia significativa de la expresión de las proteínas de las muestras de tejido tumoral en comparación con las muestras de tejido sano. Este análisis se basa en un ajuste de modelos lineales, los pGenes que se consideran diferencialmente expresados, son aquellos que presentan lo siguiente:

- Un log fold change mayor a 1. El log fold change es la proporción de cambio en el nivel de expresión entre condiciones experimentales, expresado en logaritmo base 2.

- Un estadístico B mayor a 5. El estadístico B es un indicador de consistencia estadística en los contrastes entre grupos de niveles de expresión.

Luego se utilizó Python para la visualización de los datos de expresión diferencial (Figura 4.1) y el análisis de la similitud entre los pGenes inferidos por los péptidos

shared y unshared.

3.4. Análisis de desregulación de vías y procesos celulares

Mediante gene set enrichment analysis (también llamado Over-Representation Analysis) es posible determinar cuáles son las vías o procesos celulares que están significativamente desregulados. Para ello se utilizó una herramienta web denominada WebGestalt (WEB-based GEne SeT AnaLysis Toolkit)[69] [68] que mediante pruebas hipergeométricas determina la probabilidad de que el grupo de genes incógnita del que nosotros queremos averiguar su función, esté relacionado a una vía o proceso celular particular. Esta herramienta utiliza las bases de datos, ampliamente utilizadas para este tipo de análisis, de GeneOntology (GO) y Kyotos Encyclopedia of Genes and Genomes (KEGG) [67] [35] [65], y se utilizó para enriquecer los pGenes diferencialmente expresados (con distinto punto de corte) y para los pGenes pertenecientes a las distintas comunidades de las redes generadas (detallado más adelante).

Hay diversos parámetros a tener en cuenta para realizar un ORA con GO, como el tipo de proceso dentro de la base de datos, la prueba de ajuste, en número de genes por categoría y el nivel de significancia, luego de realizar diversas pruebas se llegó a la combinación más rápida y con menor False Discovery Rate (FDR) para los distintos procesos, que se mantuvo en todos los análisis en los que se utilizó GO:

- Tipo de proceso: Biological Process no Redundant
- Número de genes por categoría: 5-2000
- Prueba de ajuste: Benjamini & Hochberg (BH)
- Nivel de significancia: TOP 20 o TOP 3
- Número de categorías esperadas por el “ set cover ”: 10
- Número de categorías visualizadas en el reporte: 40

3.5. Generación de redes de coexpresión (con información mutua)

Para la generación de redes de coexpresión con información mutua se utilizaron dos herramientas. La primera, ARACNe-AP (Algorithm for the Reconstruction of Accurate Cellular Networks, Adaptive Partitioning strategy), es el algoritmo encargado de calcular los valores de información mutua entre las distintas parejas de pGenes, a partir de los valores de expresión de la matriz proteogenómica. La segunda, el software Cytoscape, fue utilizado para la visualización de la red y los análisis subsiguientes. Las redes generadas fueron siempre a partir de los datos de los tumores, ya que el número de muestras control no es suficiente para esta tarea. De la misma manera se generaron las redes de coexpresión transcriptómica provistas por nuestro grupo de trabajo [34]. Detallaremos a continuación los distintos parámetros utilizados en cada herramienta para llegar al resultado final de la red.

ARACNe-AP: Este algoritmo tiene la capacidad de capturar dependencias estadísticas generales (información mutua) entre la expresión de los distintos genes o proteínas, como ya se explicó en la introducción. Es importante destacar que este algoritmo es muy utilizado en nuestro laboratorio ya que tiene posee gran velocidad y precisión [3]. Esta característica es ventajosa ya que hay muchas relaciones entre genes o proteínas que no siguen una dependencia estadística lineal que pueda ser capturada por coeficientes de correlación de Pearson o Spearman. Esto se podría ejemplificar en la relación que tienen las distintas subunidades proteicas de un complejo que responda a señales ambientales que determinen su estructura y estequiometría, como sucede, por ejemplo, con las proteínas *Non-Exponentially Degraded* [19]. Para realizar la red se utilizó la guía del software ARACNe (disponible en https://sourceforge.net/projects/aracne-ap/files/). El valor de p elegido para generar el punto de corte de las interacciones fue 10^{-8}.

Cytoscape: Este software libre provee una plataforma para la integración, análisis y visualización de redes de interacción. Luego de generar la tabla de información mutua con ARACNe, esta se importó en el entorno de Cytoscape, aclarando cuál era

el nodo de salida y el nodo blanco. Con esto se logró visualizar la primera versión de la red, para poder empezar con los análisis tuvimos que eliminar los enlaces repetidos (ya que durante la generación de la red no se distingue direccionalidad y se generan 2 enlaces, uno por cada dirección). Posteriormente, a partir de bases de datos públicos, se descargaron tablas que contienen la información de a qué cromosoma pertenece cada gen (del genoma humano GRCh38.p13), y de esta manera poder analizar si hay un agrupamiento de expresión correlacionado a la ubicación cromosomal.

Glay: Otra de las herramientas que usamos dentro de la plataforma de Cytoscape fue el community cluster Glay, una aplicación incorporada a Cytoscape que combina distintos algoritmos de detección de comunidades [21].

3.6. Generación y análisis de redes de interacción proteína-proteína

Para generar las redes de interacción se utilizó la base de datos y herramienta web STRING (Search Tool for the Retrieval of Interacting Genes/Proteins). Se seleccionaron los sets de pGenes pertenecientes a las comunidades, generadas por Glay, de las redes de información mutua para saber si la coexpresion está relacionada a interacciones proteína-proteína (IP-P) reportadas en experimentos validados.

Debido a la gran base de datos que posee y a la cual accede STRING, fue necesario configurar ciertos parametros en la generación de las redes para que representara la información que a nosotros nos interesa. Se seleccionó que los enlaces representaran los experimentos (ya que hay interacciones de textmining, co-occurrence, gene fusion, database, co-expression, y Neighborhood). Para el peso de los enlaces se seleccionó la confianza, en donde los niveles de confianza de interacción, que van de 0 a 1, están representados por la evidencia disponible (el número de estudios en donde fueron reportadas, y las diferentes técnicas con las que se reportaron). La otra opción disponible es la acción molecular entre las proteínas (inhibición, activación, binding, etc).

Luego de generada la red es posible visualizar, en la sección analysis, el p-value del enriquecimiento de IP-P.

Para lograr una mejor caracterización de la relación entre las redes de coexpresión y las de IP-P, en la plataforma STRING se generaron 3 redes para cada comunidad de la red de coexpresión variando los niveles de confianza de las IP-P. Los niveles utilizados en este trabajo fueron bajo (0.15), medio (0.4) y alto (0.7). Luego de esto, se extrajeron las redes resultantes de la interacción entre las redes de coexpresión y las de IP-P, mediante Cytoscape, para analizar la relación entre coexpresión e interacción.

Capítulo 4

Resultados

4.1. Análisis de datos

Debido a limitantes computacionales del servidor del TPP no fue posible seguir el protocolo de preprocesamiento de los espectros de masas para todas las muestras (cada archivo de espectros tiene un tamaño mayor a 60 gb y el tiempo requerido para el paso inicial de "Identificación de péptidos" en una muestra rondó en las 100 horas). Puesto que la matriz no pudo completarse en los primeros dos semestres del proyecto, continuamos trabajando con la matriz de expresión proteogenómica.

4.2. Curado de los datos

Cuando realizamos un análisis de calidad de la matriz proteómica descargada del CPTAC, Figura adicional 6.1, pudimos observar las siguientes características:

- Ya se encontraba normalizada, se corroboró buscando en los protocolos de la generación de la matriz en donde aclaraban que se había extraído la mediana de todos los datos.
- La matriz se encontraba incompleta. Muchos de los valores de los pGenes dentro de la matriz no se encontraban en todas las muestras.

- Muchas de las muestras presentaban una desviación considerable, haciendo difícil la distinción de los grupos de muestras controles y tumores (PCA Figura adicional 2).

Por estos motivos fue necesario realizar varias modificaciones a la matriz. Lo primero que se realizó fue extraer los pGenes que no tenían información en mas de 60 muestras, que representa menos del 10 % de los pGenes, como se muestra en la Figura adicional 6.1.A, ya que podrían traer ruido a los análisis. Luego, realizando una búsqueda en la literatura del CPTAC averiguamos que 28 de las muestras de tumor habían pasado por un proceso de degradación protéica, lo que generaba datos poco fiables. De esta manera excluimos de la matriz a dichas muestras, lo que mejoró considerablemente la separación de los grupos, así como la desviación de los datos de las muestras de tumor (PCA de la Figura adicional 6.1.B). Para poder realizar el análisis de expresión diferencial, por una cuestión estadística del método, fue necesario eliminar los pGenes que no se encontraran en las 3 muestras de tejido sano.

4.3. Análisis de expresión diferencial

El análisis de expresión diferencial aplicado a la matriz de expresión proteogenómica curada tuvo como objetivo observar el cambio en los niveles de expresión de los pGenes en las muestras de tumores en comparación a las muestras sanas (a partir de los valores de log fold change). En la Figura 4.1 se muestra un mapa de calor, resultado del análisis de expresión diferencial de los 9124 pGenes de la matriz curada.

Se muestra la expresión protéica de dos maneras: A partir de péptidos únicos de cada gen, que no se encuentran en proteínas que no sean expresadas por ese gen ("unshared"); y a partir de todos los péptidos que pueden conformar a las proteínas expresadas por el gen en cuestión, sin importar que también puedan pertenecer a proteínas expresadas por otros genes ("shared").

En el mapa de calor se muestran, de forma esquemática, solo algunos de los nombres de los pGenes, ya que la tabla posee 9124 pGenes y no sería posible poner el nombre de todos en la figura.

A simple vista se observa en el mapa de calor que el perfil de expresión para ambas aproximaciones, péptidos “unshared” y "shared", no muestra gran diferencia en la expresión diferencial. Esto muestra que la inclusión del resto de los péptidos, que no son únicos, al análisis de expresión diferencial no generaría gran cambio en el perfil de expresión, pero para tener una medida cuantitativa de esta comparación analizamos la similitud entre ambos perfiles de expresión.

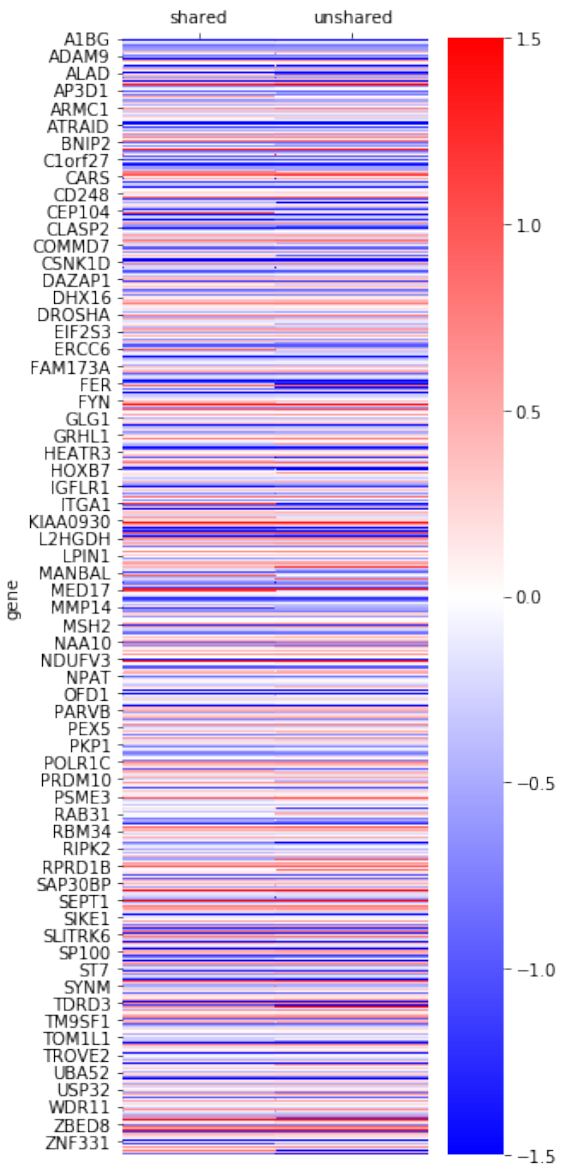


Figura 4.1: Mapa de calor de expresión diferencial. Se representan los pGenes inferidos a partir de la expresión protéica de muestras tumorales en comparación a muestras de tejido sana. La sobreexpresión está expresada en rojo y la subexpresión en azul.

Para los pGenes inferidos de los péptidos únicos (“unshared”) se contaron 4608 pGenes sub-expresados (con un valor de log fold change menor a cero) y 4545 sobre-expresados (con un valor de log fold change mayor a cero). Para los de péptidos totales (“shared”) fueron 4772 sub-expresados (de los cuales 4369 fueron compartidos con los péptidos únicos, lo que representa el 95 %) y 4654 sobre-expresados (de los cuales 4248 están compartidos con los únicos, que representa el 93 %).

Puesto que hay una alta similitud en la expresión diferencial (alrededor del 95 %) para los péptidos “shared” y “unshared” decidimos usar los datos de la expresión de los péptidos “unshared” para análisis posteriores, y de esta manera asegurarnos que los pGenes con los que trabajamos sólo provengan de una sola proteína medida en las

muestras.

4.4. Análisis de desregulación de vías y procesos celulares

Para poder determinar qué procesos biológicos están siendo afectados por la expresión diferencial de los distintos pGenes, seleccionamos distintos subsets de pGenes variando el punto de corte de la expresión diferencial, los puntos de corte fueron 0.5, 1.0 y 1.5.

Cuando se hace un ORA para procesos biológicos mediante la base de datos de GO no importa si los pGenes diferencialmente expresados están sobre o sub-expresados, ya que no sabemos qué función cumplen en ese proceso, lo que podría estar inhibiéndolo o activándolo. Por este motivo el ORA con GO se realizó con una línea de corte de expresión diferencial incluyendo pGenes sobre y sub-expresados.

El subconjunto de pGenes diferencialmente expresados con una línea de corte menos restringente, 0.5 y -0.5, está conformado por 3838 pGenes, de los cuales 3161 están anotados a una categoría funcional de GO. Cuando se realiza un ORA de este set de pGenes mediante WebGestalt (ver métodos), se obtiene una lista de procesos como se muestra en la Tabla 4.1. En este tipo de tablas, que seguiremos analizando más adelante, mostramos los nombres de los 20 procesos más enriquecidos (como están anotados en la base de datos de GO), su proporción de enriquecimiento y FDR. La proporción de enriquecimiento representa el número de genes de nuestro set, que forman parte del conjunto de genes del proceso biológico, sobre el número de genes teóricos que formarían parte del conjunto de genes del proceso biológico si el set de genes fuera aleatorio.

Cuando el punto de corte se define en 1 y -1, el número de pGenes desciende a 1381, de los cuales 1116 están anotados en GO. Este subconjunto se enriqueció dando como resultado los procesos que se observan en la Tabla 4.2.

Cuando el punto de corte se define en 1.5 y -1.5, el número de pGenes desciende a

Tabla 4.1: Procesos biológicos enriquecidos a traves de GO para pGenes diferencialmente expresados con un corte de 0.5 y -0.5.

Procesos biológicos	Proporción de Enriquecimiento	FDR
protein activation cascade	3.41	<1E-16
cytolysis	2.81	6.66E-06
platelet degranulation	2.69	<1E-16
acute inflammatory response	2.16	7.65E-10
vitamin metabolic process	1.89	2.26E-05
antibiotic metabolic process	1.87	1.25E-05
humoral immune response	1.84	1.29E-08
small molecule catabolic process	1.72	2.57E-11
negative regulation of proteolysis	1.70	1.29E-08
protein maturation	1.69	4.61E-08
fatty acid metabolic process	1.68	1.29E-08
extracellular structure organization	1.66	3.77E-09
granulocyte activation	1.64	1.32E-10
neutrophil mediated immunity	1.63	2.02E-10
sulfur compound metabolic process	1.61	3.61E-07
coenzyme metabolic process	1.60	6.70E-07
negative regulation of hydrolase activity	1.55	3.45E-07
mRNA processing	1.52	3.45E-07
RNA splicing	1.48	1.93E-05
generation of precursor metabolites and energy	1.46	1.54E-05

Tabla 4.2: Procesos biológicos enriquecidos a traves de GO para pGenes diferencialmente expresados con un corte de 1 y -1.

Procesos biológicos	Proporción de Enriquecimiento	FDR
protein activation cascade	7.21	<1E-16
protein-containing complex remodeling	5.95	3.63E-06
cytolysis	5.31	1.03E-06
acute inflammatory response	4.22	<1E-16
platelet degranulation	3.97	1.25E-11
cofactor catabolic process	3.79	2.27E-05
interaction with symbiont	3.59	1.62E-05
humoral immune response	3.29	0
neutral lipid metabolic process	3.21	1.14E-06
regulation of response to wounding	2.88	1.84E-06
antibiotic metabolic process	2.84	5.55E-06
regulation of inflammatory response	2.54	6.85E-11
fatty acid metabolic process	2.49	3.04E-10
protein maturation	2.36	9.42E-08
extracellular structure organization	2.36	8.94E-10
coagulation	2.12	1.06E-05
sulfur compound metabolic process	2.06	1.08E-05
negative regulation of proteolysis	2.05	2.21E-05
small molecule catabolic process	1.94	1.18E-05
organic hydroxy compound metabolic process	1.91	4.28E-06

528, de los cuales 420 están anotados en GO. Este subconjunto se enriqueció dando como resultado los procesos que se observan en la tabla 4.3.

Tabla 4.3: Procesos biológicos enriquecidos a traves de GO para pGenes diferencialmente expresados con un corte de 1.5 y -1.5.

Procesos biológicos	Proporción de Enriquecimiento	FDR
protein-containing complex remodeling	12.15	4.55E-07
protein activation cascade	8.81	2.83E-13
protein-lipid complex subunit organization	8.07	3.69E-06
cofactor catabolic process	7.83	2.62E-07
platelet degranulation	7.77	<1E-16
regulation plasma lipoprotein particle levels	5.39	3.14E-05
neutral lipid metabolic process	5.29	6.38E-07
acute inflammatory response	5.04	8.84E-08
antibiotic metabolic process	4.86	5.20E-07
drug catabolic process	4.67	3.15E-06
humoral immune response	3.79	5.46E-07
negative regulation of proteolysis	3.57	3.54E-08
coagulation	3.24	1.04E-06
steroid metabolic process	3.22	3.15E-06
fatty acid metabolic process	3.16	7.01E-07
receptor-mediated endocytosis	3.07	3.04E-05
regulation of inflammatory response	3.03	2.76E-06
organic hydroxy compound metabolic process	2.89	1.69E-07
extracellular structure organization	2.82	6.22E-06
negative regulation of hydrolase activity	2.78	3.69E-06

Luego de revisar los procesos que comparten los 3 análisis con los distintos puntos de corte, pudimos observar que hay 8 de 20 que están compartidos por los 3 análisis, y 14 que se comparten entre los dos análisis con los puntos de corte más restringentes (1.5 y 1). Partiendo de esta similitud, y las proporciones de enriquecimiento crecientes, nos centraremos en el análisis de los procesos para el punto de corte más restringente de 1.5.

El proceso más enriquecido es protein-containing complex remodeling (remodelación de complejos que contienen proteínas). Este proceso tiene una proporción de enriquecimiento de 12.16, este es el resultado de una prueba hipergeométrica, como ya se explicó anteriormente, esto significa que el número de pGenes de nuestro set de pGenes diferencialmente expresado, que forman parte de este proceso, son 12.16 veces más que los que serían si nuestro set de genes fuera aleatorio. La definición general de este proceso en GO es “la adquisición, pérdida o modificación de macromoleculas dentro de un complejo, resultando en la alteración de un complejo existente”. Debido a que el tipo de proceso con el cual se hizo el ORA es noRedundant (procesos no redundantes), aglomera 4 procesos que son: protein-DNA complex

remodeling, protein-RNA complex remodeling, protein-DNA-RNA complex remodeling y protein-lipid complex remodeling. Haciendo un análisis más fino de los pGenes que se encuentran diferencialmente expresados para este proceso, se puede observar que la mayoría están relacionados con el metabolismo y transporte de lípidos, particularmente varias de estos pGenes se encuentran en un cluster de apolipoproteínas (transportadoras de lípidos) del cromosoma 11 y están sub-expresadas, lo que podría indicar que este proceso enriquecido se trata de algún fenómeno estructural. Las apolipoproteínas A1/C3/A5, que pertenecen a este cluster y regulan la formación de HDL y la actividad de la lipoproteína lipasa [30], han sido objeto de varios estudios. Desde deleciones o translocaciones asociadas con un mayor riesgo de cáncer de mama [4] a la asociación entre los niveles de algunas de estas apolipoproteínas a la incidencia de cáncer [54][7]. Tambien se ha informado la posible existencia de un loci supresor de tumores, donde se detectaron pérdidas alélicas frecuentes en muchos tipos de tumores, incluido cáncer de mama [40], en particular el gen APOC3, que se encuentra dentro de los pGenes regulados negativamente en la región mencionada [37].

Muchos de los procesos enriquecidos están vinculados con el transporte/ metabolismo/ organización de lípidos, ya que las apolipoproteínas, previamente mencionadas, forman parte y son un factor importante en el enriquecimiento de estos procesos. De esta manera nos enfocaremos en los demás procesos que no involucren lípidos (debido a la posible inferencia del fenómeno estructural en los procesos asociados a lípidos).

El segundo proceso más enriquecido es protein activation cascade, el cual está definido por GO como "Una respuesta a un estímulo que consiste en una serie secuencial de modificaciones a un conjunto de proteínas donde el producto de una reacción actúa catalíticamente en la siguiente reacción. La magnitud de la respuesta se amplifica típicamente en cada paso sucesivo en la cascada. Las modificaciones típicamente incluyen proteólisis o modificaciones covalentes, y también pueden incluir eventos de unión."; al ser éste un proceso muy general, se realizó un análisis más fino de los pGenes que están enriqueciendo este proceso y se pudo observar que hay un grupo de 17 pGenes (de los 23 que enriquecen al proceso) que pertenecen a la cascada de complemento y coagulación, de la cual se hablará en más detalle más adelante, pero que tienen un

rol importante en la tumorigénesis [50].

Dentro del resto de los procesos se destacan los relacionados a la inflamación y respuesta inmune ya que representan 4 de los 20. Entre ellos se encuentran platelet degranulation, acute inflammatory response, humoral immune response y regulation of inflammatary response. El número de pGenes diferencialmente expresados de estos procesos equivale al 13 % de todos los pGenes. Como veremos más adelante en el trabajo, estos procesos parecen ser claves en el funcionamiento tumoral debido a la persistencia en la que aparecen en los distintos análisis.

Todavía es necesario seguir haciendo un análisis fino de los pGenes pertenecientes a los distintos procesos enriquecidos para tener una mejor caracterización del perfil tumoral, pero debido a limitantes de tiempo hasta este nivel de profundidad llega el alcance de este trabajo.

Como se mencionó antes, se realizó el mismo análisis utilizando la base de datos de KEGG para distintos puntos de corte, a excepción de que se analizó a los pGenes sobre-expresados y sub-expresados por separado. Esta clasificación se debe a que mediante el análisis por KEGG es posible determinar si la función del gen es activadora o represora de la vía. Debido a que los pGenes pertenecientes al sub set de pGenes sobre-expresados para las líneas de corte más restringentes (1 y 1.5) son muy pocos, el FDR es para todas las vías es mayor a 0.05, por lo que los resultados para estos puntos de corte no son mostrados. Los resultados restantes pueden verse en las tablas 4.4, 4.5. De este análisis de enriquecimiento realizado mediante KEGG, se pueden extraer varios resultados.

Para los subsets de pGenes sobre-expresados, aunque la línea de corte no sea tan alta, se puede observar en la tabla 4.4 que la vía más enriquecida es la de “protein export”, que se refiere a la exportación de proteínas al medio extracelular. Esta vía tiene 23 pGenes involucrados, de los cuales 12 están sobre-expresados, y uno en particular se ha estudiado en profundidad, el gen SEC11A ha sido asociado a migración, invasión y metástasis en nodos linfáticos [42] y tmabién se encuentra sobre-expresado en la línea de corte de 1.

Se sabe que la secreción de proteínas al medio extracelular es fundamental para

Tabla 4.4: Vías celulares más enriquecidas a través de KEGG para pGenes sobre-expresados con un corte de 0.5.

Vías celulares	Proporción de Enriquecimiento	FDR
Protein export	5.95	4.97E-05
Allograft rejection	4.51	7.21E-04
Graft-versus-host disease	4.02	0.02
Type I diabetes mellitus	3.58	3.41E-03
Spliceosome	3.52	2.24E-05
Viral myocarditis	3.47	0.02
Ribosome	3.29	2.58E-05
Autoimmune thyroid disease	3.28	0.01
Base excision repair	3.12	4.48E-04
Steroid biosynthesis	3.01	0.07
Asthma	2.98	0.09
RNA transport	2.78	2.44E-03
Antigen processing and presentation	2.69	0.04
ABC transporters	2.37	0.06
Cell cycle	2.02	0.03
Herpes simplex infection	1.98	0.01
Phagosome	1.62	0.15
Cell adhesion molecules (CAMs)	1.59	0.08
Protein processing in endoplasmic reticulum	1.58	0.08
Epstein-Barr virus infection	1.58	0.21

la supervivencia de células tumorales [52], debido a esto se ha desarrollado una rama propia del estudio del cáncer que es el secretoma (proteínas secretadas por las células tumorales). Por lo que vemos en el enriquecimiento este proceso está teniendo un rol activo en nuestro grupo de tumores de estudio. Un análisis interesante que podría enriquecer este resultado sería la captura de cuales son las moléculas secretadas y en que concentración. No hay estudios que le den mayor profundidad a la vía de "protein export", por lo que podría ser un foco de atención en el futuro.

Otro resultado interesante que puede mencionarse de este enriquecimiento, viene de la segunda vía más enriquecida, “Allograft rejection”, vía que contiene 38 componentes, de los cuales están sobre-expresados 15, y 12 de estos son distintas subunidades de los complejos mayores de histocompatibilidad (MHC), complejos fundamentales en la presentación antigénica y la respuesta inmunológica. Cuando hay inflamación, uno de los hallmarks del cáncer [27], se suelen sobreexpresar los complejos MHC-1 en las distintas células del organismo y el complejo MHC-2 en las células presentadoras de antígeno [60], de esta manera lo que podríamos estar capturando aquí son los complejos provenientes de las células presentadoras de antígeno que se encuentran dentro

de la heterogeneidad tumoral.

Finalmente pasamos al análisis de las vias que contienen a los pGenes sub-expresados. En la tabla 4.5 podemos observar las vías más enriquecidas con los pGenes más subexpresados.

Tabla 4.5: Vías celulares más enriquecidas a través de KEGG para pGenes sub-expresados con un corte de -1.5.

Vías celulares	Proporción de Enriquecimiento	FDR
Complement and coagulation cascades	10.12	1.45E-13
Cholesterol metabolism	6.85	5.29E-04
Phenylalanine metabolism	6.72	0.18
Nitrogen metabolism	6.72	0.18
Histidine metabolism	6.62	0.09
Tyrosine metabolism	6.34	0.02
PPAR signaling pathway	6.17	6.58E-05
Fatty acid degradation	6.06	0.01
Folate biosynthesis	5.86	0.11
Ferroptosis	4.76	0.10
African trypanosomiasis	4.35	0.22
Chemical carcinogenesis	4.18	0.02
Glycolysis / Gluconeogenesis	3.92	0.08
ECM-receptor interaction	3.71	0.06
Drug metabolism	3.70	0.09
Regulation of lipolysis in adipocytes	3.52	0.22
Metabolism of xenobiotics by cytochrome P450	3.51	0.10
Drug metabolism	3.37	0.11
Adipocytokine signaling pathway	3.31	0.18
Metabolic pathways	1.37	0.21

La via principal, que se destaca por su alto valor de enriquecimiento y bajo FDR en todos los puntos de corte, es Complement and coagulation cascades. Esta vía, como su nombre lo indica tiene dos componentes, la cascada de complemento y la de coagulación.

En la cascada de coagulación observamos que un 33 % de los genes dentro de la vía están subexpresados en gran medida (con línea de corte en -1.5), y este número de genes abarca más del 63 % de la vía cuando el corte está en -1. Sin embargo, muchos de los genes subexpresados son inhibidores de la formación del coágulo (producto final de la vía), dificultando el análisis. Varios de los factores determinantes de la vía (como la trombina, o el factor tisular) se encuentran subexpresados, lo que no coincide con estudios anteriores que determinan la posible importancia de estas proteínas en procesos de progresión tumoral como la angiogénesis y metástasis [50][14]. Igualmente,

como se mencionó anteriormente, muchos de los inhibidores de estas proteínas también se encuentran subexpresados, lo que dificulta determinar si el proceso de coagulación está activo o reprimido.

En la cascada de complemento sucede algo similar a lo que sucede en la cascada de coagulación. El 34 % de las proteínas están subexpresadas para un corte de -1.5, y el 69 % con corte en -1. Pero al igual que en la cascada de coagulación observamos que es difícil discernir si la vía está inactiva, ya que muchos de las proteínas subexpresadas son inhibidoras de la vía.

Igualmente los resultados obtenidos para los pGenes subexpresados no fueron los esperados, ya que estudios recientes muestran que la vía de complemento se encuentra activa en los tumores, ya que les otorga algunos beneficios debido a que es causante de inflamación, y sus componentes se encuentran sobreexpresados [50]. Debido a que estos estudios son recientes, todavía no hay un consenso pleno sobre el rol que juegan las vías de complemento y coagulación en el desarrollo y prognosis tumoral.

Este resultado contradictorio a los estudios recientes podría indicar que estas vias no son esenciales para los tumores en general, y que su estado de activación pueda variar en cada caso, así como tambien que la expresión y el estado de los componentes de cada vía puedan variar con el tiempo dentro de los mismos tumores. De una u otra manera, este resultado podría aportar nueva información para que el estudio de estas vías se mantenga activo y pueda dilucidarse el verdadero rol que cumplen en el cáncer de mama.

Al igual que en los procesos biológicos, en el análisis de vías todavía hay muchos resultados por profundizar.

4.5. Análisis de redes de coexpresión

Utilizando ARACNe y Cytoscape generamos, visualizamos y analizamos una red de interacción mutua (Figura 4.2) en donde los enlaces representados tienen un p-value de al menos 10^{-8}. Los nodos representan los pGenes (inferidos a partir de la expresión protéica) y los enlaces la información mutua que estos comparten. En esta

red también están representados los cromosomas a los cuales pertenecen los pGenes, el color de los nodos representa el cromosoma al cual pertenece cada pGen.

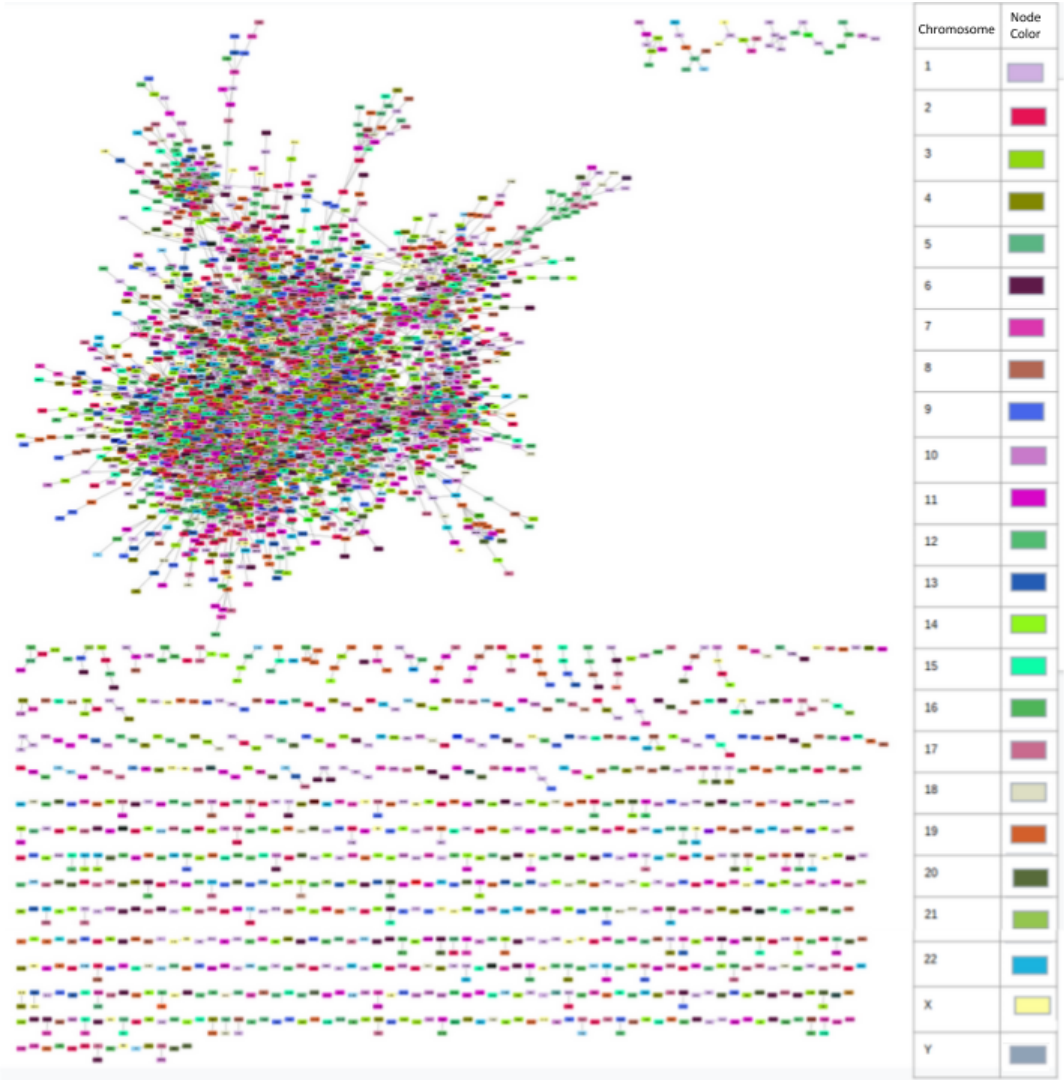


Figura 4.2: Red de coexpresión de los pGenes inferidos a partir de la expresión protéica. El color de los pGenes representa su localización cromosomal, presentado en la tabla del lado derecho.

La red está conformada por 4318 nodos y 18735 enlaces. Tiene un componente gigante (que contiene más del 50 % de los nodos) de 3276 pGenes (76 % del total), y el resto de los pGenes se reparten en pequeños componentes de no más de 7 pGenes.

El motivo por el cual se pintaron los pGenes de acuerdo a su cromosoma fue para poder comparar esta red con la de transcritos y analizar la inferencia de los procesos post-transcripcionales en la coexpresión a nivel de ARN y proteico [34].

4.6. Comparación topológica entre redes de transcritos y proteínas

Un fenómeno característico que se da en cáncer es que la red de coexpresión del ARN mensajero está totalmente agrupada en los cromosomas [23][17]. En la Figura 4.3 se muestra la diferencia entre las redes de coexpresión de transcriptos (generadas a partir de datos provenientes de la técnica de RNA-Seq) para muestras tumorales y sanas a partir de datos de nuestro laboratorio [34]. Estas redes se generaron con las mismas muestras utilizadas para el análisis proteómico, ya que tambien fueron provistas por The Cancer Genome Atlas.

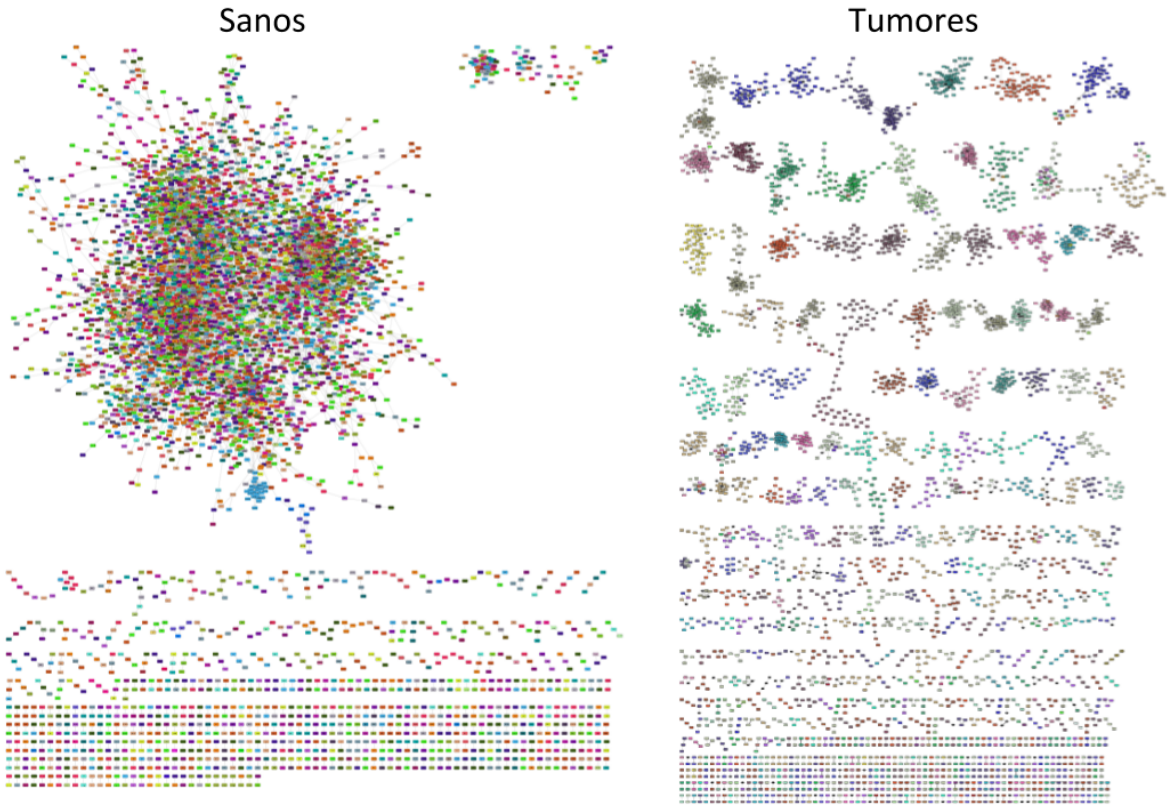


Figura 4.3: Redes de coexpresión a partir de datos de RNA-Seq. La red de la izquierda está inferida a partir de muestras de tejido sano, la red de la derecha de muestras tumorales. Los colores refieren al cromosoma de cada gen.

No fue posible hacer el mismo análisis mostrado en la Figura 4.3 para proteínas ya que se ha buscado en distintas bases de datos proteicas y hasta el momento no hay suficientes proteomas de muestras de tejido sano para generar una red de coexpresión

protéica con significancia estadística. Por este motivo lo que hicimos fue contrastar solamente las redes de coexpresión que se generaron a partir de muestras tumorales.

Para poder comparar topológicamente las redes generadas por ARACNe es deseable que se tenga el mismo número de enlaces para cada red. Por ellos se seleccionaron los 11651 enlaces con mayor información mutua para cada red, este número viene de los enlaces que poseía la red original del transcriptoma.

En la Figura 4.4 podemos observar las redes de coexpresión del transcriptoma y del proteoma.

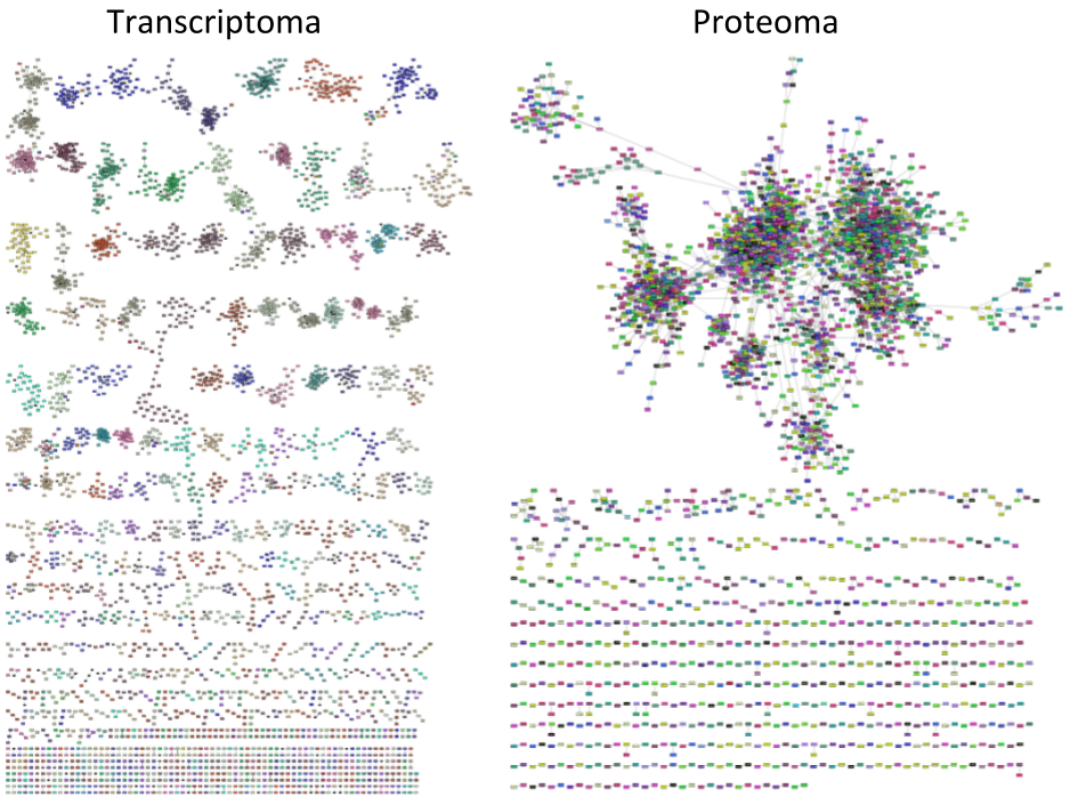


Figura 4.4: Redes de coexpresión del transcriptoma y el proteoma. La red de la izquierda está inferida a partir de datos de RNA-Seq, la red de la derecha de la matriz de epxresión protegenómica. Los colores refieren al cromosoma de cada pGen.

Como se puede observar, en la red del proteoma, hay una gran coexpresión entre pGenes de los distintos cromosomas (interacciones trans), a diferencia de la red del transcriptoma. Esto indicaría que la coexpresión de genes que se encuentran en los mismos cromosomas es un fenómeno a nivel transcripcional, pero que a nivel traduccional no se mantiene [22]. Entonces, podría suponerse que la correlación que se

observa en la transcripción de genes de los mismos cromosomas no responde necesariamente a una necesidad de las células tumorales, ya que la traducción protéica es la que representa en gran medida los requerimientos celulares, y en esta etapa no se sostiene el fenómeno visto para los datos provenientes de RNA Seq [23][17]. Como se expone en distintos artículos, la pérdida de las interacciones trans podría deberse a un fenómeno estructural de la organización genómica más que alguna dependencia entre las necesidades celulares [22]. Algunas teorías que podrían explicar estos fenómenos están siendo formuladas en nuestro laboratorio, como la falta de control de la ARN-polimerasa, lo que llevaría a la transcripción en tandem de los genes en un mismo cromosoma [17], y que la variación en el número de copias podría tener alguna incidencia en el fenómeno estructural transcripcional [49]. Otros estudios recientes, que podrían ser complementarios a las teorías del laboratorio, apuntan a que aberraciones epigenéticas podrían estar alterando la integridad topológica del genoma [12].

Puesto que la red del proteoma representa, en cierta medida, los requisitos celulares para la supervivencia y desarrollo tumoral, lo siguiente que hicimos fue buscar cuales son las comunidades o módulos de esta red y cuales pueden ser las funciones o fenómenos asociados por las cuales se estén formando.

4.7. Detección de comunidades

Partimos de la red con las 11651 interacciones más fuertes para detectar las comunidades que tiene esta red, con el objetivo de analizar qué relación tienen las proteínas contenidas dentro de cada comunidad.

Utilizamos la aplicación que se encuentra dentro de Cytoscape llamada clusterMaker, que posee muchos programas y algoritmos para detectar comunidades. El programa utilizado fue community cluster (Glay), que utiliza distintos algoritmos dependiendo de las características de la red (ver métodos) [21]. A los pGenes de la red se les asignaron colores según el cluster al que pertenecen como se observa en la Figura 4.5.

Con Glay se generaron 359 clusters, de los cuales 266 son pares de pGenes, con más

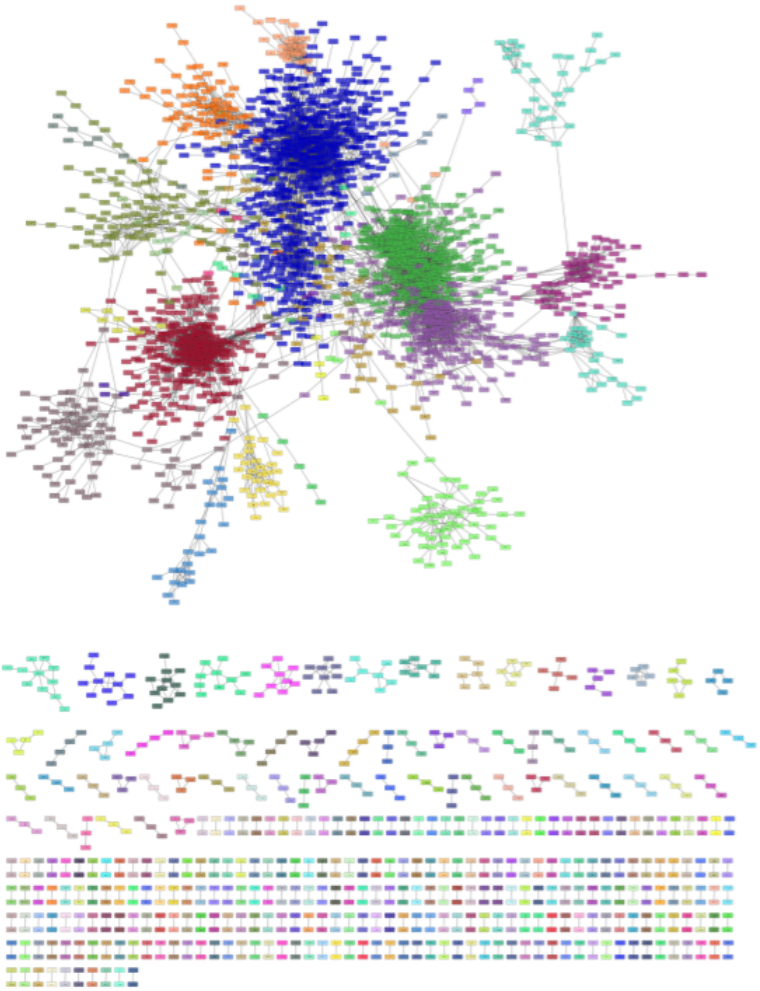


Figura 4.5: Redes de coexpresión del proteoma agrupada mediante el algoritmo Glay. Los colores representan a qué comunidad pertenece cada pGen

de 10 pGenes hay únicamente 19 módulos. Por lo tanto analizamos las 5 comunidades que contienen mayor número de pGenes y que representan más del 50 % de los pGenes de la red. A los pGenes pertenecientes a esas comunidades se les hizo un análisis de enriquecimiento para determinar si hay procesos biológicos asociados con esos grupos de pGenes.

4.8. Análisis de enriquecimiento

Se enriquecieron los pGenes pertenecientes a las distintas comunidades mediante el software online WebGestalt, utilizando la base de datos de Gene Ontology para

procesos biológicos no redundantes. Para no perdernos entre tanta información, en la Tabla 4.6 solo se presentan los 3 procesos biológicos más enriquecidos para cada comunidad (detallando el numero de pGenes que contiene) y se muestra la confianza del enriquecimiento, en donde todos los procesos tienen un FDR menor al 0.05.

Tabla 4.6: Análisis de enriquecimiento para las 5 comunidades más grandes.

Procesos Biológicos	Proporción de Enriquecimiento	FDR
Comunidad de 578 pGenes		
RNA splicing	3.9897	<1E-16
mRNA processing	3.6212	9.4369E-14
regulation of mRNA metabolic process	3.8778	2.603E-08
Comunidad de 373 pGenes		
humoral immune response	8.3981	<1E-16
acute inflammatory response	13.825	<1E-16
platelet degranulation	15.241	<1E-16
Comunidad de 275 pGenes		
neutrophil mediated immunity	4.8739	<1E-16
regulation of leukocyte activation	6.1568	<1E-16
adaptive immune response	6.012	<1E-16
Comunidad de 228 pGenes		
extracellular structure organization	8.2849	<1E-16
cell-substrate adhesion	6.7961	<1E-16
collagen metabolic process	13.811	3.1456E-14
Comunidad de 91 pGenes		
mitotic cell cycle phase transition	8.8822	<1E-16
chromosome segregation	13.331	<1E-16
DNA replication	14.899	<1E-16

En resumidas cuentas podemos observar lo siguiente sobre los procesos enriquecidos de cada comunidad:

- Para el módulo 1, que contiene 578 pGenes, los 3 procesos más enriquecidos corresponden al procesamiento del ARN mensajero.
- Los 3 procesos del módulo 2 (con 373 pGenes) pertenecen a la respuesta inmune e inflamatoria. La desgranulación de plaquetas está asociada a la presencia de histamina y serotonina que son reguladores en la vasodilatación y permeabilidad de los vasos sanguíneos [33]. La respuesta inmune humoral incluye a los anticuerpos así como, las proteínas pertenecientes al sistema de complemento.
- Los procesos asociados al tercer módulo, que contiene 275 pGenes, también están asociados a la respuesta inmune, pero más focalizados en la respuesta

adaptativa.

- Los procesos enriquecidos para la cuarta comunidad, de 228 pGenes, pertenecen a la organización de la estructura extracelular, proceso fundamental para el desarrollo tumoral y la invasión a otros tejidos mediante metástasis.

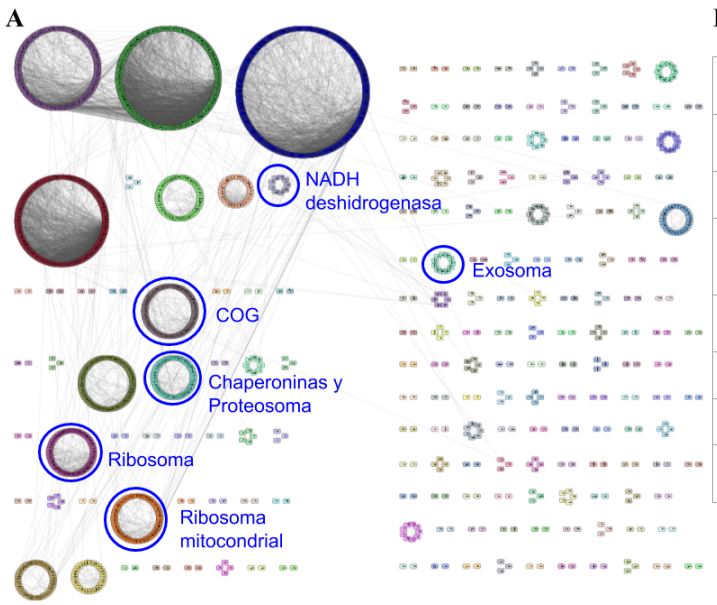
- En la quinta comunidad, con 91 pGenes, los procesos pueden agruparse al procesamiento del ADN durante la mitosis, otro proceso esencial para el crecimiento desregulado de las células cancerígenas.

Podemos observar que los procesos enriquecidos para las distintas comunidades pertenecen a distintas categorías bien definidas, y que cada una de ellas corresponden a diversos hallmarks del cáncer [27], así como a procesos fisiológicos bien marcados. El procesamiento del ARN mensajero y del ADN durante la mitosis puede asociarse a la inestabilidad genómica y la inmortalidad replicativa. Los procesos asociados a la respuesta inmunitaria pueden estar asociados a la evasión de la destrucción inmunológica y la inflamación promovida por el tumor. Finalmente la comunidad con procesos relacionados a la organización de la estructura extracelular está asociada a la invasión metastásica y la angiogénesis [20].

Podemos ver que la red captura un comportamiento funcional de los tumores, y para darle mayor validez a la relevancia de la red de coexpresión de información mutua y su consiguiente análisis, lo que hicimos a continuación fue buscar si la red podía capturar complejos proteicos, ya que las proteínas que los componen sabemos que se encuentran coexpresadas [22].

4.9. Captura de complejos proteicos en la red de interacción mutua

Cuando continuamos con el análisis de las comunidades más pequeñas observamos que distintos complejos proteicos se veían reflejadas en estas, como se observa en la Figura 4.6.



B

Complejo protéico	Subunidades del complejo	Porcentaje de la comunidad
NADH deshidrogenasa	9/45	100%
Exosoma	8/10	73%
Complejo Oligomérico del Golgi (COG)	9/9	10%
Chaperoninas (TRiC)	7/18	13%
Proteosoma	10/28	18%
Ribosoma	57/79	83%
Ribosoma mitocondrial	50/82	68%

Figura 4.6: Captura de complejos protéicos. A. Red de información mutua con una visualización jerárquica circular en torno a las comunidades. Las comunidades donde se encuentran los complejos están señaladas en azul. B. Tabla detallando el número de subunidades de los complejos dentro de la comunidad y cuánto representa de ella.

Como esperabamos, se pudo observar que muchas de las comunidades están enriquecidas con pGenes que codifican para proteínas que pertenecen al mismo complejo proteico. Este resultado le da mayor respaldo a la idea de que el agrupamiento de los pGenes en las distintas comunidades responde a la función biológica que cumplen.

Ya que se encuentran disponibles los datos de las interacciones directas entre proteínas, lo siguiente que se realizó fue analizar la relación de las interacciones inferidas en las redes de coexpresión y las interacciones proteína-proteína validadas experimentalmente.

4.10. Generación y análisis de redes de interacción proteína-proteína

Cuando cargamos el set de pGenes de cada comunidad en la plataforma de STRING, y configuramos que la red generada represente Interacciones Proteína-Proteína (IPP) validadas experimentalmente, con una confianza de interacción media (0.4), se

puede observar una red como la de la Figura 4.7, que representa las IP-P que mantienen las proteínas sintetizadas a partir de los pGenes que se encuentran en la quinta comunidad más grande, que contiene 91 pGenes.

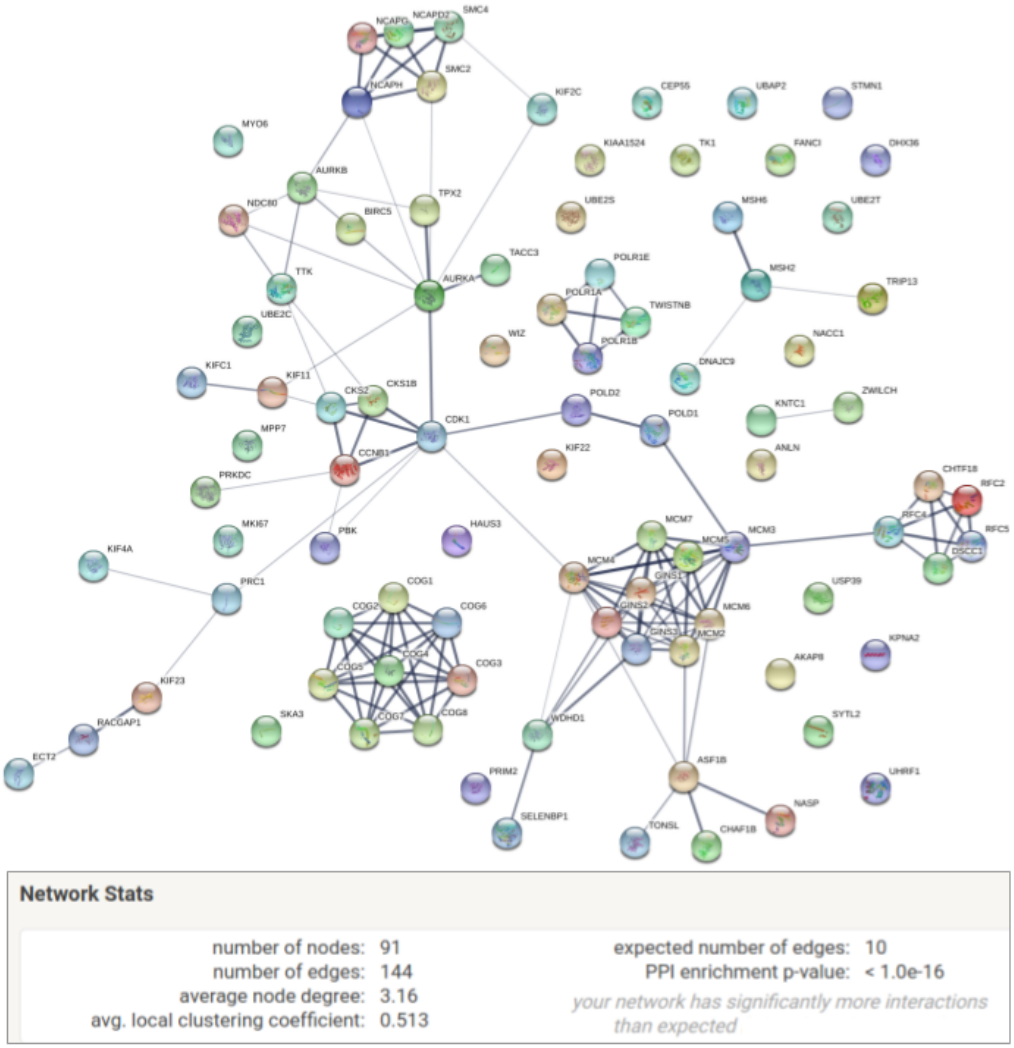


Figura 4.7: Red de IP-P con las estadísticas provistas por STRING. El grosor de los enlaces representa la confianza con la que se han reportado.

Las redes generadas en la plataforma de STRING poseen mucha información disponible, desde la información más elemental de cada proteína analizada, como su función o sus distintos identificadores, hasta el modelado tridimensional de su estructura. Los enlaces también poseen información de cómo fueron reportados, el ensayo utilizado y el correspondiente artículo donde fue publicado, con base en esta información se le asigna el peso del enlace que representa la confianza.

En el caso de la red particular de la Figura 4.7, podemos observar en las estadísticas que los enlaces reportados para ese set de proteínas es mucho más alto que el esperado para un set aleatorio. Este resultado se da en parte por la presencia del complejo protéico de COG, que ya había sido capturado mediante la red de coexpresión.

Luego de generar 3 redes por cada comunidad, variando los niveles de confianza de las IP-P, pudimos observar que los valores de p del enriquecimiento de IP-P para cada una de las redes fueron menores a 10^{-16}, el cual es el límite de cálculo del servidor de STRING. Como ya habíamos sugerido con la red de coexpresión, este resultado indica que las comunidades están conectadas biológicamente de manera significativa.

Finalmente, lo que buscamos fue analizar los enlaces compartidos entre la red de coexpresión y las de IP-P, de esta manera se podría entender mejor la relación subyacente que existe entre la coexpresión y las interacciones directas, y cómo la confianza de las IP-P validadas experimentalmente influye en esta relación. Para ello, lo que se realizó fue generar una red de intersección entre las redes de IP-P, con sus distintas confianzas de interacción, y la red de coexpresión. Los resultados se ven en la Tabla 4.7, donde se muestran los enlaces que comparten ambas redes (en la tabla se muestra como la columna de "Intersecciones") y la relación entre los enlaces que comparten ambas redes y los esperados de IP-P si el set de genes hubiera sido aleatorio (en la tabla se representa como "I/E").

Tabla 4.7: Análisis de los enlaces compartidos entre la red de coexpresión y la red de IP-P.

Confianza	Baja		Media		Alta	
	Intersecciones	I/E	Intersecciones	I/E	Intersecciones	I/E
Módulo 1	136	0.09	68	0.19	49	0.21
Módulo 2	172	0.54	56	3.73	39	7.80
Módulo 3	205	0.44	70	2.59	42	6.00
Módulo 4	85	0.22	38	2.38	20	5.00
Módulo 5	73	0.76	56	5.60	54	10.80

Podemos observar, para todos los módulos, que a medida que la confianza aumenta la relación de enlaces compartidos entre ambas redes y enlaces de IP-P esperados para un set de proteínas aleatorias tambien aumenta, y en 4 de los 5 módulos la relación es al menos 5 veces más grande cuando la confianza de IP-P es alta. Esto nos ayuda, una vez más, a reafirmar la idea de que la coexpresión está asociada a la funcionalidad.

Capítulo 5

Discusión y conclusiones

Debido a los altos índices epidemiológicos y a la dificultad de encontrar tratamientos efectivos contra el cáncer, es de primordial importancia continuar con el estudio de esta enfermedad tan heterogénea.

Los resultados obtenidos en este trabajo amplían la visión panorámica del perfil de expresión proteómico del cáncer de mama para este set de muestras tumorales.

Los resultados de enriquecimiento de los pGenes diferencialmente expresados arrojaron diversas características. Primero, mediante este tipo se análisis, fue posible capturar un posible fenómeno estructural habitual, que podría ir desde una translocación a una deleción del cluster de apolipoproteínas del cromosoma 11. También vimos que los procesos relacionados con el sistema inmune y la inflamación toman un rol preponderante cuando se analizan los pGenes diferencialmente expresados, ya que son la rama de procesos que más pGenes abarca, representando el 13 % de los pGenes más diferencialmente expresados. Por otro lado, pudimos observar que los resultados de los últimos estudios realizados con objetivo de determinar el rol del sistema de complemento en cáncer de mama, no se vieron reflejados en nuestros resultados. Esta variabilidad que observamos con resultados anteriores podría aportar una visión novedosa en cuanto al papel que puede tener el complemento, así como la vía de coagulación, en el desarrollo tumoral. Igualmente todavía queda mucho trabajo por hacer con estos resultados, ya que hay mucha información que puede analizarse a partir de los datos de expresión diferencial.

Cuando generamos la red de coexpresión a partir de la información mutua que comparten los pGenes observamos que el fenómeno de pérdida de regulación trans observado a nivel transcriptómico se pierde. Aunque se ha estudiado que la correlación entre el proteoma y el transcriptoma no es alta, resulta interesante que la coexpresión genética tan ligada a la ubicación cromosomal, que se observa en la red inferida a partir de los datos de RNASeq, se perdiera casi totalmente a nivel protéico. Esto indica que hay, de hecho, un programa regulatorio de la traducción muy distinto al de la transcripción, ya que la coexpresión de genes no representa la cotraducción de proteínas.

Durante el análisis de enriquecimiento de los procesos asociados a las comunidades generadas volvimos a observar procesos relacionados a inflamación y sistema inmune, lo que sugiere que las proteínas están siendo sintetizadas de manera tal que este tipo de procesos se encuentren funcionalmente activos y coordinados. También pudimos advertir que la organización de las comunidades responde a procesos claramente determinados y que son parte de los hallmarks del cáncer. Esto nos da la idea de que el análisis de redes de coexpresión de información mutua, es muy valioso a la hora de caracterizar y representar los procesos biológicos que se están dando en muestras complejas como lo son los tumores. Esta noción viene aparejada de la capacidad de capturar varios complejos proteicos de relevancia funcional para las células.

La sección de la generación y análisis de redes de interacción proteína-proteína nos ha provisto de resultados que sugieren, nuevamente, que la funcionalidad celular está correlacionada con la coexpresión. Se pudo observar que la funcionalidad asociada a las comunidades de la red de coexpresión sigue estando vigente a la hora de calcular el valor de p del enriquecimiento de IP-P. Observar que el aumento de confianza de las IP-P aumenta el numero de interacciones que comparten las redes de coexpresión e IP-P respecto a las que se esperarían, nos lleva a pensar que la coexpresión captura las interacciones más directas. Una comparación de estos resultados con los de muestras sanas nos podría otorgar una visión más clara de los rearreglos en la regulación de la síntesis protéica que sufren las células cancerígenas. Por este motivo esperamos que los estudios proteómicos sigan en su curso de avance para poder disponer de mayor

cantidad de datos de calidad con los cuales seguir trabajando.

Aunque la literatura todavía no presenta tanta evidencia, las redes de coexpresión de proteínas representan un enfoque válido para obtener una nueva visión general de los datos proteómicos y proporcionar nuevas hipótesis sobre los procesos clave que actúan en estados patológicos. Por supuesto, su valor debe seguir siendo evaluado por más estudios, pero los resultados que se obtienen con este enfoque los hacen prometedores para convertirse en una herramienta estándar para el análisis de los perfiles de expresión.

Como sabemos, los tumores son heterogéneos, pero la estructura de los tumores se define por la presencia de células cancerosas y no cancerosas en distintos arreglos espaciales [25]. Podemos observar la probable presencia de células no cancerosas, por ejemplo, en el análisis de enriquecimiento, donde las proteínas MHCII, expresadas por las células presentadoras de antígeno, están sobreexpresadas. Debido a esto, un enfoque más preciso para analizar los procesos biológicos tumorales sería determinar el proteoma de las diferentes secciones del tumor. Este tipo de análisis multidimensional (single cell proteomics), en donde también se incorporan datos genómicos, está comenzando [25], y podría ser un paso realmente importante hacia la comprensión de los ecosistemas tumorales y su evolución, para finalmente desarrollar nuevas terapias.

Bibliografía

[1] A J Butte, I S Kohane. "Mutual Information Relevance Networks: Functional Genomic Clustering Using Pairwise Entropy Measurements". En: *Pac Symp Biocomput.* 29.418 (2000). DOI: 10.1142/9789814447331_0040.

[2] A. Prat, L. A. Carey, B. Adamo, M. Vidal, J. Tabernero, J. Cortés, J. S. Parker, C. M. Perou, and J. Baselga. "Molecular features and survival outcomes of the intrinsic subtypes within her2-positive breast cancer". En: *Journal of the National Cancer Institute* 106.8 (2014), dju152. DOI: 10.1093/jnci/dju152..

[3] Alexander Lachmann, Federico M. Giorgi, Gonzalo Lopez, Andrea Califano. "ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information". En: *Bioinformatics* 32.14 (2016), págs. 2233-2235. DOI: https://doi.org/10.1093/bioinformatics/btw216.

[4] Alexander S. Hill, Nicola J. Foot, Tracy L. Chaplin, Bryan D. Young. "The most frequent constitutional translocation in humans, the t(11;22)(q23;q11) is due to a highly specific Alu-mediated recombination". En: *Human Molecular Genetics* 9 (2000), págs. 1525-1532. DOI: https://doi.org/10.1093/hmg/9.10.1525.

[5] Alexey I Nesvizhskii, Andrew Keller, Eugene Kolker, Ruedi Aebersold. "A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry". En: *Anal Chem.* 75.17 (2003), págs. 4646-4658. DOI: 10.1021/ac0341261..

[6] Robert Hubley the ISB Andrew Keller Jimmy K. Eng. *PeptideProphet.* URL: http://peptideprophet.sourceforge.net/. (accessed: 04.07.2020).

[7] Atsushi Kasamatsuab, Katsuhiro Uzawaab, Katsuya Usukuraa, Kazuyuki Koikea, Dai Nakashimaa, Takashi Ishigamia, Kazuaki Fushimia, Katsunori Ogawarab, Masashi Shiibab, Hideki Tanzawa. "Loss of heterozygosity in oral cancer". En: *Oral Science International* 8.2 (2011), págs. 37-43. DOI: https://doi.org/10.1016/S1348-8643(11)00027-9.

[8] Bayraktar, S. y Glück, S. "Molecularly targeted therapies for metastatic triple-negative breast cancer". En: *Breast Cancer Research and Treatment* 138 (2013), págs. 21-35. DOI: 10.1007/s10549-013-2421-5.

[9] Brunel H, Gallardo-Chacón J, Buil A, Vallverdú M, Soria J, et al. "Miss: a non-linear methodology based on mutual information for genetic association studies in both population and sib-pairs analysis". En: *Bioinformatics* 26 (2010), págs. 1811-1818. DOI: https://doi.org/10.1093/bioinformatics/btq273.

[10] Centro Nacional de Equidad de Género y Salud Reproductiva. "Información Estadística Cáncer de Máma". En: *Secretaría de Salud, Gobierno de México* (2016). DOI: https://www.gob.mx/salud/acciones-y-programas/informacion-estadistica.

[11] Chaowang Lan, Hui Peng, Eileen M. McGowan, Gyorgy Hutvagner Jinyan Li. "An isomiR expression panel based novel breast cancer classification approach using improved mutual information". En: *BMC Medical Genomics* 11.118 (2018), págs. 67-77. DOI: https://doi.org/10.1186/s12920-018-0434-y.

[12] Charalampos Lazaris, Iannis Aifantis, Aristotelis Tsirigos. "On Epigenetic Plasticity and Genome Topology". En: *Trends in Cancer* 6.3 (2020), págs. 177-180. DOI: https://doi.org/10.1016/j.trecan.2020.01.006.

[13] Charles M. Perou, et al. "Molecular portraits of human breast tumours". En: *J Clin Oncol.* 406 (2000), págs. 747-752. DOI: https://doi.org/10.1038/35021093.

[14] Daniel Ajona, Sergio Ortiz-Espinosa, Ruben Pio and Fernando Lecanda. "Complement in Metastasis: A Comp in the Camp". En: *Front Immunol* 10 (2019), pág. 669. DOI: 10.3389/fimmu.2019.00669.

[15] David P. Nusinow, John Szpyt, Mahmoud Ghandi, ..., Levi A. Garraway, William R. Sellers, Steven P. Gygi. "Quantitative Proteomics of the Cancer Cell Line Encyclopedia". En: *Cell* 180 (2020), págs. 387-402. DOI: https://doi.org/10.1016/j.cell.2019.12.023.

[16] David Shteynberg, Eric W Deutsch, Henry Lam, Jimmy K Eng, Zhi Sun, Natalie Tasman, Luis Mendoza, Robert L Moritz, Ruedi Aebersold, Alexey I Nesvizhskii. "iProphet: Multi-Level Integrative Analysis of Shotgun Proteomic Data Improves Peptide and Protein Identification Rates and Error Estimates". En: *Mol Cell Proteomics.* 10.12 (2011). DOI: 10.1074/mcp.M111.007690.

[17] Diana García-Cortés, Guillermo de Anda-Jáuregui, Cristobal Fresno, Enrique Hernández-Lemus, Jesús Espinal-Enriquez. "Loss of trans regulation in breast cancer molecular subtypes". En: *bioRxiv* (2018). DOI: doi:https://doi.org/10.1101/399253.

[18] E. Senkus et al. "Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up". En: *Annals of Oncology* 26.5 (2015), págs. 8-30. DOI: https://doi.org/10.1093/annonc/mdv298.

[19] Erik McShane,Celine Sin,Henrik Zauber,Jonathan N. Wells,Neysan Donnelly,Xi Wang,Jingyi Hou,Wei Chen,Zuzana Storchova,Joseph A. Marsh,Angelo Valleriani,Matthias Selbach. "Kinetic Analysis of Protein Stability Reveals Age-Dependent Degradation". En: *Cell* 167.3 (2016), págs. 803-815. DOI: https://doi.org/10.1016/j.cell.2016.09.015.

[20] Fui Boon Kai, Allison P.Drain, Valerie M.Weaver. "The Extracellular Matrix Modulates the Metastatic Journey". En: *Developmental Cell* 49.3 (2019), págs. 332-346. DOI: https://doi.org/10.1016/j.devcel.2019.03.026.

[21] Gang Su, Allan Kuchinsky, John H. Morris, David J. States, Fan Meng. " GLay: community structure analysis of biological networks ". En: *Bioinformatics* 26.24 (2010), págs. 3135-3137. DOI: https://doi.org/10.1093/bioinformatics/btq596.

[22] Georg Kustatscher, Piotr Grabowski, and Juri Rappsilber. "Pervasive coexpression of spatially proximal genes is buffered at the protein level". En: *Molecular Systems Biology* 8 (2017), pág. 937. DOI: 10.15252/msb.20177548.

[23] Guillermo de Anda-Jáuregui, Cristobal Fresno, Diana García-Cortés, Jesús Espinal Enríquez and Enrique Hernández-Lemus. "Intrachromosomal regulation decay in breast cancer". En: *Applied Mathematics and Nonlinear Sciences* 4.1 (2014), págs. 223-230. DOI: https://doi.org/10.2478/AMNS.2019.1.00020.

[24] Guillermo de Anda-Jáuregui, Jesús Espinal-Enríquez , Diana Drago-García, and Enrique Hernández-Lemus. "Nonredundant, Highly Connected MicroRNAs Control Functionality in Breast Cancer Networks". En: *Int J Genomics* 2018 (2018). DOI: 10.1155/2018/9585383.

[25] H. Raza Ali, Hartland W. Jackson, Vito R. T. Zanotelli, Esther Danenberg, Jana R. Fischer, Helen Bardwell, Elena Provenzano, CRUK IMAXT Grand Challenge Team, Oscar M. Rueda, Suet-Feung Chin, Samuel Aparicio, Carlos Caldas Bernd Bodenmiller. "Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer". En: *Nature Cancer* 1 (2020), págs. 163-175. DOI: https://doi.org/10.1038/s43018-020-0026-6.

[26] Hadley Wickham, et al. "Welcome to the Tidyverse". En: *Journal of Open Source Software* 4.(43) (2019), pág. 1686. DOI: https://doi.org/10.21105/joss.01686.

[27] Hanahan D, Weinberg RA. "Hallmarks of Cancer: The Next Generation". En: *Cell* 144.5 (2011), págs. 646-674. DOI: https://doi.org/10.1016/j.cell.2011.02.013.

[28] Harold J Burstein. "The Distinctive Nature of HER2-positive Breast Cancers". En: *N Engl J Med* 353.(16) (2005), págs. 1652-4. DOI: 10.1056/NEJMp058197.

[29] Hugo Tovar, Rodrigo García-Herrera, Jesús Espinal-Enríquez, Enrique Hernández-Lemus. "Transcriptional Master Regulator Analysis in Breast Cancer Genetic Networks". En: *Comput Biol Chem.* 59 (2015), págs. 67-77. DOI: 10.1016/j.compbiolchem.2015.08.007..

[30] Hung Chun Chen Jinn, Yu Gu Jer, Min Chan Min, Chi Hsieh Shyi, Jang Shin Yung, Hsiun Lai. "Role of lipid control in diabetic nephropathy". En: *Kidney International* 67 (2005), S60-S62. DOI: https://doi.org/10.1111/j.1523-1755.2005.09415.x.

[31] National Cancer Institute. *Clinical Proteomic Tumor Analysis Consortium.* URL: https://cptac-data-portal.georgetown.edu/. (accessed: 11.03.2020).

[32] National Cancer Institute. *Genomic Data Commons.* URL: https://gdc.cancer.gov/. (accessed: 11.03.2020).

[33] Jennifer A. Frontera, J. Javier Provencio, Fatima A. Sehba, Thomas M. McIntyre, Amy S. Nowacki, Errol Gordon, Jonathan M. Weimer Louis Aledort. "The Role of Platelet Activation and Inflammation in Early Brain Injury Following Subarachnoid Hemorrhage". En: *Neurocrit Care* 26 (2017), págs. 48-57. DOI: https://doi.org/10.1007/s12028-016-0292-4.

[34] Jesús Espinal-Enríquez, Cristóbal Fresno, Guillermo Anda-Jáuregui Enrique Hernández-Lemus. "RNA-Seq based genome-wide analysis reveals loss of inter-chromosomal regulation in breast cancer". En: *Scientific Reports volume* 7.1760 (2017). DOI: https://doi.org/10.1038/s41598-017-01314-1.

[35] Jing Yang, Lei Chen, Xiangyin Kong, Tao Huang, Yu-Dong Cai. "Analysis of Tumor Suppressor Genes Based on Gene Ontology and the KEGG Pathway". En: *PLOS ONE* 9.(9) (2014), e107202. DOI: https://doi.org/10.1371/journal.pone.0107202.

[36] Joel S. Parker, et al. "Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtype". En: *J Clin Oncol.* 27.8 (2009), págs. 1160-1167. DOI: 10.1200/JCO.2008.18.1370.

[37] Katsuhiro Uzawa Hiroyoshi Suzuki Akira Komiya Hiroshi Nakanishi Katsunori Ogawara Hideki Tanzawa Kenichi Sato. "Evidence for two distinct tumor-suppressor gene loci on the long arm of chromosome 11 in human oral cancer". En: *International Journal of Cancer* 67.4 (1996), págs. 510-514. DOI: https:

//doi.org/10.1002/(SICI)1097-0215(19960807)67:4<510::AID-IJC8>3.0.CO;2-V.

[38] Liu Y, Beyer A, Aebersold R. "On the Dependency of Cellular Protein Levels on mRNA Abundance." En: *Cell* 165(3) (2016), págs. 535-50. DOI: 10.1016/j.cell.2016.03.014..

[39] Margaret Flowers 1, Patricia A Thompson. "t10c12 Conjugated Linoleic Acid Suppresses HER2 Protein and Enhances Apoptosis in SKBr3 Breast Cancer Cells: Possible Role of COX2". En: *PLoS One* 4.(4) (2009), e5342. DOI: 10.1371/journal.pone.0005342..

[40] Massimo Negrini, Debora Rasio, Garrett M. Hampton, Silvia Sabbioni, Shashi Rattan, Stephen L. Carter, Anne L. Rosenberg, Gordon F. Schwartz, Yosef Shiloh, Webster K. Cavenee and Carlo M. Croce. "Definition and Refinement of Chromosome 11 Regions of Loss of Heterozygosity in Breast Cancer: Identification of a New Region at 11q23.3". En: *Cnacer Research* 55.14 (1995), págs. 3003-3007.

[41] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, Gordon K. Smyth. "limma powers differential expression analyses for RNA-sequencing and microarray studies". En: *Nucleic Acid Research* 43.7 (2015), pág. 47. DOI: https://doi.org/10.1093/nar/gkv007.

[42] N Oue, Y Naito, T Hayashi, M Takigahira, A Kawano-Nagatsuma, K Sentani, N Sakamoto, H Zarni Oo, N Uraoka, K Yanagihara, A Ochiai, H Sasaki W Yasui. "Signal peptidase complex 18, encoded by SEC11A, contributes to progression via TGF- secretion in gastric cancer". En: *Oncogen* 33 (2014), págs. 3918-3926. DOI: https://doi.org/10.1038/onc.2013.364.

[43] The Global Proteome Machine Organization. *X! TANDEM Spectrum Modeler*. URL: https://www.thegpm.org/TANDEM/index.html. (accessed: 04.07.2020).

[44] World Health Organization. *Breast cancer*. URL: https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/. (accessed: 06.03.2020).

[45] Andrew Keller Nichole King Patrick Pedrioli. *Libra, protein quatification software*. URL: http://tools.proteomecenter.org/wiki/index.php?title=Software:Libra#More_detail. (accessed: 04.07.2020).

[46] Patrik Waldmann. "On the Use of the Pearson Correlation Coefficient for Model Evaluation in Genome-Wide Prediction". En: *Frontiers in Genetics* (2019). DOI: https://doi.org/10.3389/fgene.2019.00899.

[47] Peter T Simpson, Jorge S Reis-Filho, Theodora Gale Sunil R Lakhani. "Molecular evolution of breast cancer". En: *The journal of Pathology* 205.3 (2005), págs. 248-254. DOI: https://doi.org/10.1002/path.1691.

[48] Philipp Mertins, D. R. Mani1, Kelly V. Ruggles, Michael A. Gillette, Karl R. Clauser, Pei Wang, Xianlong Wang, Jana W. Qiao, Song Cao, Francesca Petralia, Emily Kawaler, ..., Steven A. Carr the NCI CPTAC. "Proteogenomics connects somatic mutations to signalling in breast cancer". En: *Nature* 534 (2016), págs. 55-62. DOI: 10.1038/nature18003.

[49] Phillippa C. Taberlay, et al. "Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations". En: *Genome Research* 26.(6) (2016), págs. 719-731. DOI: 10.1101/gr.201517.115.

[50] Reis, E., Mastellos, D., Ricklin, D. et al. "Complement in cancer: untangling an intricate relationship". En: *Nat Rev Immunol* 18 (2018), págs. 5-18. DOI: https://doi.org/10.1038/nri.2017.97.

[51] Ana Belén Pazos Ruiz. *Análisis de correlación moderno: ¿Qué alternativas existen para la correlación de Pearson?* España de Creative Commons, 2018. ISBN: http://openaccess.uoc.edu/webapps/o2/bitstream/10609/81845/7/apazosrTFM0618memoria.pd

[52] Sannino S, Brodsky JL. "Targeting protein quality control pathways in breast cancer". En: *BMC Biology* 15 (), pág. 109. DOI: 10.1186/s12915-017-0449-4..

[53] Sergio Antonio Alcalá-Corona, Jesús Espinal-Enríquez, Guillermo de Anda-Jáuregui, and Enrique Hernández-Lemus. "The Hierarchical Modular Structure of HER2+ Breast Cancer Network". En: *Front Physiol* 9 (2018), pág. 1423. DOI: 10.3389/fphys.2018.01423.

[54] Signe Borgquist Talha Butt Peter Almgren Dov Shiffman Tanja Stocks Marju Orho-Melander Jonas Manjer Olle Melander. "Apolipoproteins, lipids and risk of cancer". En: *Cancer Epidemiology* 138 (2016), págs. 2648-2656. DOI: https://doi.org/10.1002/ijc.30013.

[55] Smyth G.K. "limma: Linear Models for Microarray Data". En: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor.* (2005), págs. 397-420. DOI: https://doi.org/10.1007/0-387-29362-0_23.

[56] Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, et al. "Repeated observation of breast tumor subtypes in independent gene expression data sets". En: *Proc Natl Acad Sci USA* 100 (2003), págs. 8418-8423. DOI: 10.1186/1471-2164-7-96..

[57] Suhas Vasaikar, Chen Huang, Xiaojing Wang, ..., Tao Liu, Bing Zhang, Clinical Proteomic Tumor Analysis Consortium. "Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities". En: *Cell* 177 (2019), págs. 1035-1049. DOI: https://doi.org/10.1016/j.cell.2019.03.030.

[58] Suzana de Siqueira Santos, Daniel Yasumasa Takahashi, Asuka Nakata, André Fujita. "A comparative study of statistical methods used to identify dependencies between gene expression signals". En: *Briefings in Bioinformatics* 15.(6) (2014), págs. 906-918. DOI: https://doi.org/10.1093/bib/bbt051.

[59] Institute for Systems Biology (ISB). *Trans Proteomic Pipeline*. URL: Institute%20for%20Systems%20Biology%20(ISB). http://tools.proteomecenter.org/software.php(accessed: 30.06.2020).

[60] Tadeo Enrique Velazquez-Caldelas, Sergio Antonio Alcalá-Corona, Jesús Espinal-Enríquez and Enrique Hernandez-Lemus. "Unveiling the Link Between Inflam-

mation and Adaptive Immunity in Breast Cancer". En: *Frontiers in Immunology* (2019). DOI: https://doi.org/10.3389/fimmu.2019.00056.

[61] Wilson Wen Bin Goh, Limsoon Wong. "Integrating Networks and Proteomics: Moving Forward". En: *Cell* 34.12 (2016), págs. 951-959. DOI: https://doi.org/10.1016/j.tibtech.2016.05.015.

[62] Wu C., Zheng L. "Proteomics promises a new era of precision cancer medicine". En: *Springer Science* 46 ().

[63] Xiujun Zhang, Xing-Ming Zhao, Kun He, Le Lu, Yongwei Cao, Jingdong Liu, Jin-Kao Hao, Zhi-Ping Liu, Luonan Chen. "Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information". En: *Bioinformatics* 28.1 (2012), págs. 98-104. DOI: https://doi.org/10.1093/bioinformatics/btr626.

[64] Y Guo, Y Xing. "Weighted gene co-expression network analysis of pneumocytes under exposure to a carcinogenic dose of chloroprene". En: *Life Sci* (2016). DOI: 10.1016/j.lfs.2016.02.074.

[65] Yanshan Ge, Zhengxi He, Yanqi Xiang, Dawei Wang, Yuping Yang, Jian Qiu Yanhong Zhou. "The identification of key genes in nasopharyngeal carcinoma by bioinformatics analysis of high-throughput data". En: *Molecular Biology Reports* 46 (2019), págs. 2829-2840. DOI: https://doi.org/10.1007/s11033-019-04729-3.

[66] Young Kwang Chaea and Ana Maria Gonzalez-Angulo. "Implications of Functional Proteomics in Breast Cancer". En: *Oncologist* 19 (2014), págs. 328-335. DOI: 10.1634/theoncologist.2013-0437.

[67] Yuxiang Lin, Fangmeng Fu, Jinxing Lv, Mengchi Wang, Yan Li, Jie Zhang, and Chuan Wang. "Identification of potential key genes for HER-2 positive breast cancer based on bioinformatics analysis". En: *Medicine (Baltimore)* 99.(1) (2020), e18445. DOI: 10.1097/MD.0000000000018445.

[68] Yuxing Liao, Jing Wang, Eric J Jaehnig, Zhiao Shi, Bing Zhang. "WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs". En: *Nucleic Acids Research* 47.1 (2019), W199-W205. DOI: https://doi.org/10.1007/s11033-019-04729-3.

[69] Zhiao Shi Yuxing Liao y Bing Zhang. *WEB-based Gene SeT AnaLysis Toolkit*. URL: http://www.webgestalt.org/. (accessed: 11.03.2020).

[70] Zhiyuan Hu, et al. "The Molecular Portraits of Breast Tumors Are Conserved Across Microarray Platforms". En: *BMC Genomics* (2006), 7:96. DOI: 10.1186/1471-2164-7-96..

Capítulo 6

Figuras adicionales

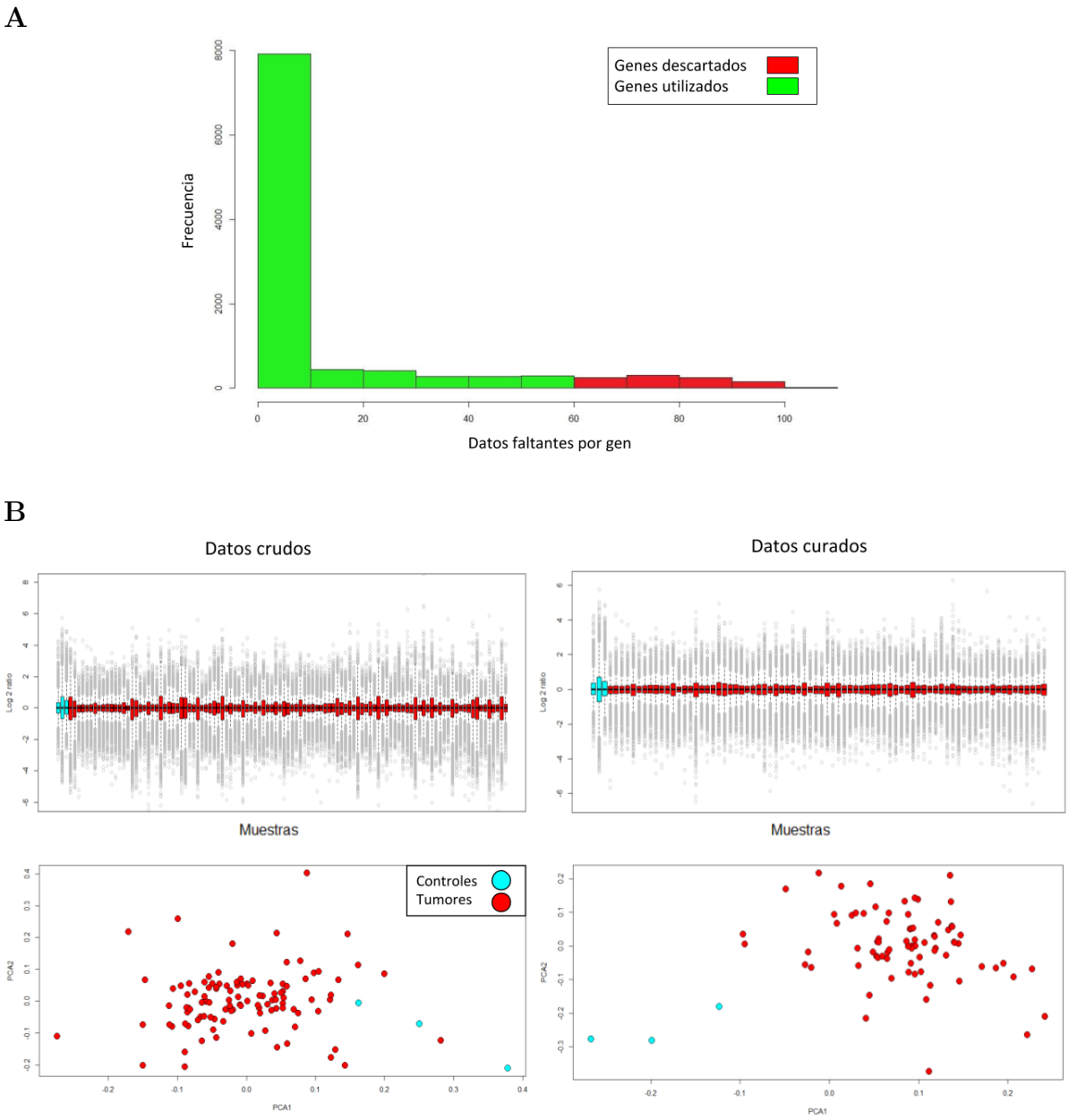


Figura 6.1: Curado de la matriz proteogenómica. A. Histograma de la frecuencia de datos faltantes por gen. B. Boxplot y PCA de los datos de la matriz protegenómica antes y después de su curado, respectivamente.