UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Doctorado en Ciencias Biomédicas

Instituto de Ecología

Análisis integrativo de la regulación de la expresión genética en cáncer de mama

TESIS

QUE PARA OPTAR POR EL GRADO DE:

Doctora en Ciencias

PRESENTA:

María de la Soledad Ochoa Méndez

DIRECTOR DE TESIS

Dr. Enrique Hernández Lemus

Instituto Nacional de Medicina Genómica

COMITÉ TUTOR

Dra. Myrna G. Candelaria Hernández

Instituto Nacional de Cancerología

Dr. Alfredo Hidalgo Miranda

Instituto Nacional de Medicina Genómica

Ciudad de México 2023

Quiero empezar asentando los agradecimientos pertinentes. De entrada a las instituciones que me albergaron y sin cuya infraestructura y personal no habria tésis: el Programa de Doctorado en Ciencias Biomédicas, el Instituto Nacional de Medicina Genómica, el Instituto de Ecología y el Consejo Nacional de Ciencia y Tecnología. En concreto, quiero agradecer a quienes me orientaron en este proceso. A mi tutor principal, el Dr. Enrique Hernández Lemus, y los otros dos miembros de mi comité tutor, la Dra. Myrna G. Candelaria Hernández y el Dr. Alfredo Hidalgo Miranda; sin sus preguntas y recomendaciones, ni siquiera habría arrancado el doctorado. Realmente contar con un tutor como el Dr. Enrique hizo toda la diferencia. Aunque en otro rubro, también tengo que agradecerle a la Lic. Erika Rodríguez de la oficina de posgrado, sin su paciencia y disposición tampoco habría llegado a este punto.

La ciencia en este proyecto se nutrió enormemente de la interacción con mis compañeros, todos y cada uno, con sus intereses y estilos particulares, incluso los que dudaban de estar en el lugar correcto. Espero que mis silencios no crearan un desequilibrio insalvable en este intercambio. Quiero agradecerles particularmente a los doctores Guillermo de Anda-Jáuregui y Santiago Sandoval, tan accesibles, siempre. A Helena, Erandi, Diana Tapia, Diana Garcia, Fernanda, Cristobal, Yuriko, Tadeo, Daniel, Chucho, Chema, Hugo, Karol, Kahori,...

Ya que afuera del doctorado estaba la ciudad insondable, necesito mencionar aqui a Elsa y a Alsino, gracias por darme la mano cada que tropecé. Gracias a German por recordarme comer, a mi amplia familia, a los amigos. Gracias a lo que nunca notaron lo importantes que fueron para este proceso.

Le dedico este trabajo
a Francisco por el refugio pandemico,
pero mas por las contradicciones;
a Emilia por el valor;
a Simao por el amor.

Índice general

Siglas y abreviaturas

5mC 5-metilcitosina

5hmC 5-hidroximetilcitosina

5fC 5-formilcitosina

5caC 5-carboxilcitosina

AML Leucemia mieloide aguda

AR Receptor de endrogeno

ARSyN *ASCA Removal of Systematic Noise on Seq data*. Herramienta para corregir el efecto de lote

AVE Varianza explicada promedio

BER Reparación por escisión de bases

bHLH Dominio de unión a DNA *basic helix-loop-helix*

bZIP Dominio de unión a DNA *basic leucine zipper*

BMPs Morfógenos de hueso

C2H2-ZF Dominio de unión a DNA Cys(2)His(2) *Zinc finger*

CAF Fibroblastos asociado al cáncer

CCA Análisis de correlación canónica

CCLE Enciclopedia de Líneas Celulares del Cáncer

CGI Isla de CpG

CNA Alteración del número de copias

CoCA *Cluster-of-clusters*

CV Validación cruzada

DBD Dominio de unión al DNA

DNMT DNA metiltransferasa

DRAGON *Determining Regulatory Associations using Graphical models on multi-Omic Networks*

EMT Transformación epitelio-mesenquimal

ER Receptor de estrógeno

ERE Elemento de respuesta a estrógeno

EWAS Estudio de asociación del epigenoma completo

FDR *False discovery rate*. Método de ajuste del valor p por contrastes múltiples

FGF Factores de crecimiento fibroblásticos

GDC *Genomica Data Commons*

GSEA *Gene Set Enrichment Analysis*

GWAS Estudio de asociación del genoma completo

HER2E Subtipo enriquecido de HER2

HM450 *HumanMethylation450K BeadChip*

HR Receptor Hormonal

hTR Componente de RNA de la telomerasa

IHQ Inmunohistoquímica

JIVE *Joint and Individual Variation Explained*

KRAB

LASSO *Least Absolute Shrinkage and Selection Operator*

MBD Dominio de unión a CpG metilado

MCIA Análisis de co-inercia múltiple

MFA Análisis de Factores Múltiples

MI Información mutua

MOGSA *Multi-omics Gene Set Analysis*

MOTA *Multi-Omic inTegrative Analysis*

MRE Elemento de respuesta a miRNAs

NHR Receptor nuclear de hormonas

PAM50 *Prediction Analysis of Microarray 50*

PANDA *Passing Attributes between Networks for Data Assimilation*

PCA Análisis de componentes principales

PLS Mínimos cuadrados parciales

PPI Interacciones proteína-proteína

PR Receptor de progesterona

RGCCA Análisis generalizado de correlación canónica

RISC Complejo de silenciamiento iducido por RNA

RMSE Error de predicción. Error de raíz cuadrada media

SGCCA Análisis generalizado y escueto de correlación canónica

TCGA *The Cancer Genome Atlas*

TF Factor transcripcional

TMM *trimmed mean of M-values*

TN Triple negativo

TSG Gen supresor de tumores

TWAS Estudio de asociación del transcriptoma completo

Resumen

El cáncer de mama es el tipo de cáncer más diagnosticado y es la principal causa de muerte por cáncer entre las mujeres; a pesar de contar con subtipos con características moleculares y clínicas bien definidas. La expresión genética, empleada para identificar los subtipos, está sometida a mecanismos de regulación con patrones divergentes entre los subtipos y dentro de los mismos, dejando la interrogante de qué determinan que se exprese una firma transcripcional u otra. La identificación de los mecanismos que regulan la expresión en cada subtipo, permitiría comprender mejor las diferencias entre subtipos. En este contexto, se plantea la construcción de un modelo regulatorio por subtipo del cáncer de mama, que integre la metilación del DNA, la expresión de factores transcripcionales y de miRNAs.

A pesar de ser relativamente reciente, la integración multi-ómica cuenta con distintas aproximaciones con ventajas y desventajas particulares, de entre las que se eligieron dos herramientas complementarias: los modelos escuetos multivariados y las redes probabilísticas. En conjunto, estas herramientas permiten encontrar la relación global entre la expresión genética y las tres capas de regulación, señalar asociaciones específicas de los subtipos y las funciones afectadas, así como comparar entre modelos para buscar diferencias y similitudes entre los subtipos del cáncer de mama.

La implementación de dichas herramientas generó los resultados reportados en tres artículos de investigación diferentes, que se exponen de manera cronológica. En un primer acercamiento se ajustaron modelos de red elástica de cada uno de los genes en la firma PAM50, para cada subtipo y el tejido normal. Al tratarse de un modelo multivariado escueto, se seleccionan de manera automática y a partir de ómicas completas, los mejores predictores de la expresión genética. En su mayoría, los predictores seleccionados no se comparten entre subtipos ni entre genes; pero destacan los miRNAs miR-10a/b y miR-21, que son seleccionados recurrentemente. Estos miRNAs conectan de manera excluyente a los genes PAM50 en los distintos subtipos y el tejido normal, sugiriendo que la coordinación de la expresión de la firma PAM50 está alterada

en el cáncer de mama.

En un segundo artículo, se construyeron redes de información mutua para cada uno de los subtipos y el tejido normal. En este caso los resultados consisten en observaciones generales sobre las capas de regulación, así como una falta de conservación de circuitos asociados con una misma función en las distintas redes. En el aspecto general, hay que mencionar una diferencia en la conectividad de los nodos con potencial regulador, que repercute en la comunicación en red, e insinúa un aplicación de los CpGs como marcadores específicos de la funciones alteradas en los subtipos y un aprovechamiento de miRNAs y transcritos codificantes de TFs, como puntos de intersección entre funciones.

Finalmente se acoplaron las dos aproximaciones, comenzando con un análisis de correlación canónica generalizado y escueto, que encuentra funciones multi-ómicas, seguido de la construcción de redes de cada función, que señalan las interacciones puntuales involucradas. En este último artículo se logran concretar todos los objetivos del proyecto y se producen, de manera semi-automática, modelos multi-ómicos de las funciones afectadas por metilación del DNA, expresión de transcritos y de miRNAs, en los subtipos del cáncer de mama.

Introducción

El cáncer de mama es un problema de salud mundial, cuya heterogeneidad ha sido clasificada en subtipos con diferencias moleculares y clínicas. Las diferencias moleculares abarcan todos los niveles funcionales caracterizados hasta ahora [1–3], incluyendo el epigenético, el transcripcional y el post-transcripcional. El nivel transcripcional es particularmente importante, porque refleja el estado funcional de las células [4] y permite la identificación del subtipo al que pertenecen los tumores [5]. Sin embargo, hay una interrelación de los niveles funcionales, que sólo se ha comenzado a caracterizar recientemente [6]. En ese contexto, se plantea el estudio de manera integrada del papel de la metilación del DNA, la expresión de factores transcripcionales y de miRNAs en los cambios transcripcionales asociados al cáncer de mama.

Esta introducción justifica el objetivo por partes, empezando por exponer la importancia del cáncer de mama y sus subtipos en la sección 1. Hecho esto, se pasa a exhibir la importancia de una interpretación conjunta de los datos de metilación del DNA, la expresión de transcritos y de miRNAs en la sección 2. Aunque sólo son tres ómicas diferentes, considerando las plataformas de medición, se toman como cuatro niveles funcionales, pues los transcritos que codifican para factores de la transcripción representan un mecanismo de regulación distinto. Finalmente, se abordan antecedentes de la integración multi-ómica en la sección 3, concentrándose en las técnicas utilizadas en este proyecto: modelos de red elástica, redes de información mutua y análisis de correlación canónica escueto y generalizado.

1. El cáncer de mama

El cáncer de mama es el tipo de cáncer más diagnosticado y es la principal causa de muerte por cáncer entre las mujeres, tanto en el mundo, como en México [7]. En el país fallecieron por esta causa 6252 mujeres durante el 2015 y 7257 en el 2018 [8]. Si la tendencia de aumento en la

incidencia y mortalidad se mantiene como lo ha hecho en las últimas tres décadas [9], se espera que esta cifra siga creciendo; por lo cual es un problema que demanda estudio y atención.

En México el cáncer de mama se suele presentar en mujeres mayores a 40 años, a edades más tempranas que en otros países, y diagnosticarse en etapas localmente avanzadas [9]. El principal factor de riesgo, aunque de distintas maneras, es la exposición a estrógenos. Otros factores ampliamente reconocidos incluyen ser mujer, edad avanzada, tener antecedentes personales o familiares y aspectos del estilo de vida como obesidad y consumo de alcohol y tabaco. Aun así, se han encontrado tumores de mama en hombres, que representan menos del 1 % de los casos y solo el 10 % de los mismos se asocian a mutaciones hereditarias, siendo las más conocidas las que afectan a los genes *BRCA1* y *BRCA2* [10]. Por su parte, el estilo de vida occidental marca una clara división en la incidencia en el norte y centro del país, donde es más diagnosticado, respecto al sureste, donde en cambio es mayor la mortalidad [9].

El problema con los estrógenos es que, por una parte, promueven la proliferación celular y por otra, son oxidados a productos reactivos que dañan el DNA [11]. La exposición a estrógenos está ligada al tiempo de vida menstrual, siendo mayor el riesgo de las mujeres con menarquía temprana y menopausia tardía, pero también al consumo prolongado de anticonceptivos y a la obesidad. Antes de la menopausia la mayor parte del estrógeno en el cuerpo proviene de los ovarios y un pequeño porcentaje del tejido graso; pero después de la menopausia, la principal fuente de estrógeno es el tejido graso, y mientras más haya, mayor es el riesgo de cáncer de mama. De manera adicional, el sobrepeso ocasiona un mayor nivel de insulina en sangre, lo que también se ha asociado con cáncer de mama [10].

Dos factores relacionados que disminuyen el riesgo, son la edad del primer parto y la lactancia. Experimentos en ratones muestran que el embarazo ocasiona la diferenciación de los lóbulos mamarios en unidades secretoras, con menor actividad proliferativa, lo que disminuiría el subset de células susceptibles a la carcinogénesis. La reducción del riesgo que da la lactancia es independiente del parto y el estatus menopáusico, sin que haya una explicación funcional fuerte sino varias hipótesis, que abarcan: la interrupción de los ciclos ovulatorios, menor producción de estrógeno y la diferenciación terminal del tejido [11].

A pesar de la variación entre países y etapas, el cáncer de mama tiene una buena tasa de recuperación respecto a otros tipos de cáncer. Se estima que hasta 15 % de los pacientes desarrollan metástasis distantes, que en su mayoría se detectan en huesos, hígado, pulmón y cerebro, con una asociación entre el sitio de metástasis y el subtipo de cáncer de mama [10, 12]. Al respecto, se ha logrado identificar patrones que permiten agrupar a los tumores de distintas maneras [1, 3, 5],

que repercuten en el pronóstico y tratamiento de la enfermedad, como se detalla a continuación.

1.1. Los subtipos de cáncer de mama

La mayoría de los tumores de mama afectan el epitelio de las glándulas mamarias, esta malla de ductos ramificados, que se extienden de manera radial a partir del pezón y terminan en lóbulos [13]. Por lo cual, histológicamente, se trata de carcinomas, que pueden clasificarse como ductales o lobulares y ser invasivos o presentarse in situ. La conservación de los patrones de expresión genética indica que los carcinomas invasivos provienen de las lesiones in situ [14]. Un porcentaje menor al 1 % de tumores son sarcomas, que se desarrollan del estroma de las glándulas, incluyendo vasos sanguíneos y miofibroblastos [10].

También se ha usado a los receptores de estrógeno (ER), progesterona (PR) y del factor 2 de crecimiento epitelial humano (HER2), como marcadores inmunohistoquímicos para la clasificación clínica [1]. La presencia del receptor de estrógeno en hasta 1 % de las células del tumor, indica un tumor susceptible a terapia endocrina [15], bien diferenciado y menos agresivo. Los tumores positivos para HER2 pueden responder al tratamiento con anticuerpos monoclonales y a inhibidores de cinasa, pero la prognosis depende de los otros receptores. Los triple positivos tienen buen pronóstico, mientras que aquellos con el fenotipo ER-PR-HER2+, son más agresivos y poco diferenciados. Así, los tumores sin estos receptores no tienen terapias dirigidas [16].

La relevancia de los receptores tiene una razón biológica, puesto que el estrógeno estimula la proliferación de las células con el receptor e induce al receptor de progesterona -una hormona mitogénica-, haciendo que los tumores PR+, sean comúnmente también ER+ [16]. Por su parte, la unión del factor de crecimiento, ocasiona la heterodimerización de HER2 y la activación de su dominio intracelular, que entonces, participa en múltiples vías de transducción, como MAPK y PI3K [11]. A las subdivisiones anteriores se suma la clasificación por expresión genética o subtipos moleculares. La clasificación por expresión genética proviene de los patrones transcripcionales compartidos entre muestras distintas de un mismo tumor, que identifican a los subtipos intrínsecos: luminal A, luminal B, enriquecido de HER2 y basal [5]. Originalmente también se identificó un subtipo similar al tejido normal, sin embargo, la posibilidad de que se tratara de contaminación por tejido normal adyacente, mantiene la existencia de este subtipo en controversia [17].

Aunque se han empleado distintos clasificadores moleculares, como *Mammaprint* y *BluePrint* , e incluso se ha aproximado la subtipificación con marcadores inmunohistoquímicos de prolifera-

ción y de los receptores mencionados [9], en las bases de datos predomina el uso del clasificador PAM50, *Prediction Analysis of Microarray 50*. Se trata de un arreglo que mide la expresión de los 50 genes que mejor separan los subtipos intrínsecos [18] y que aporta información altamente predictiva sobre recurrencia y respuesta neoadjuvante [17].

El perfeccionamiento de las técnicas de alto rendimiento enriqueció la descripción de los subtipos de cáncer de mama, permitiendo el paso de un agrupamiento de firmas transcripcionales a subtipos con características multi-ómicas propias, como se resume en la tabla 1. De esta manera se puede separar claramente a los subtipos luminales, pues aunque ambos suelen ser positivos para receptores hormonales, y negativos para HER2; los tumores luminales B tienen mayor expresión de genes asociados a la proliferación celular y menor expresión de los genes ligados al tejido luminal, como PR. Un subconjunto de los tumores luminales B, son característicamente afectados por la hipermetilación de la vía de Wnt [1]. Los tumores luminales A, tienen la menor cantidad de mutaciones, pero un incremento de aquellas que afectan los genes de *PIK3CA* y *MAP3K1*, respecto al subtipo luminal B. De manera interesante, ambos subtipos tienen buen pronóstico y frecuencias elevadas de alrededor de 30 % de los casos -cada uno-, pero los luminales B exhiben mayor quimiosensibilidad y el mayor riesgo de recurrencia en 10 años independientemente de la terapia. Por lo que se ha propuesto a éste como el subtipo a estudiar, por encima de otros con peor pronóstico, para reducir la mortalidad por cáncer de mama [18].

Los tumores del subtipo enriquecido de HER2 (HER2E) se caracterizan por la sobre-expresión de *HER2* y genes cercanos, como *GRB7*, tanto a nivel transcripcional como a nivel de proteínas y por presentar la mayor cantidad de mutaciones en general y sobre el gen de la deaminasa de citidinas *APOBEC3B* [18]. La sobre-expresion de HER2 se asocia a la amplificación del brazo largo del cromosoma 17, que contiene siempre al receptor, pero cuya extensión varía. Sin embargo, este subtipo mantiene cierta controversia, pues cerca de la mitad de los tumores con la amplificación son clasificados como luminales, en su mayoría, o basales [1]. Aún más, los estudios de expresión diferencial entre tumores con la amplificación y sin ella, identifican pocos genes fuera del cromosoma 17, con cambios modestos. En cambio, al comparar tumores HER2E contra no HER2E, resaltan el receptor de andrógeno (AR) y distintas dianas de ER, que podrían explicarse por la redundancia entre ER y AR. Sumando la cooperación entre HER2 y AR, además de la relación inversa de la expresión de HER2 con ER/PR [16], se especula que la amplificación podría ser un evento driver que enmascara la naturaleza hormonal del subtipo, como mayormente apocrino (ER-PR-AR+) [19].

Los tumores basales sobre-expresan genes asociados a la proliferación celular y al tejido basal

de la mama, están característicamente hipometilados, tienen la mayor frecuencia de alteraciones sobre *TP53* y se asocian a la inactivación de *BRCA1*. Al compararlo con distintos tipos de cáncer, este subtipo resulta ser molecularmente más similar al cáncer de células escamosas de pulmón que a los subtipos luminales de mama, mientras que su patrón de mutaciones lo acerca a los tumores serosos de ovario. Aunque los tumores basales corresponderían al fenotipo triple negativo (TN) de marcadores inmunohistoquímicos, sólo el 75 % de los tumores TN tiene el patrón de expresión del subtipo basal [1]. Este patrón de expresión se asocia a tumores agresivos, que se presentan a edades tempranas, con mayor susceptibilidad en poblaciones de ancestría africana y el peor pronóstico a 5 años [18]. La correspondencia con el fenotipo TN implica que no hay tratamientos dirigidos, sin embargo, recientemente se ha aprobado el uso de inhibidores de PARP en los tumores con mutaciones de *BRCA1*.

Molecularmente el subtipo basal puede dividirse aún más, aunque aún no hay un consenso de cuántos y cuáles serían esos sub-subtipos, se han mencionado los subtipos bajo en claudina, metaplásico y rico en interferon [16] y, de manera independiente, los grupos similar a basal, mesenquimal y ligado al receptor luminal de andrógeno. Cada categoría tiene sus propias mutaciones y características clínicas, de entre las que cabe destacar una edad de diagnóstico superior para los tumores ligados al receptor luminal de andrógeno y la activación, sin amplificación, de HER2. Por su parte, los tumores similares a basal se separan en dos grupos, BL1 y BL2, por el riesgo de progresión. Aquellos clasificados como BL2 se asocian a un *checkpoint* G1/S intacto, mientras que los identificados como BL1 pierden copias de *RB*, lo que repercute en la expresión de la proteína. Finalmente, los tumores mesenquimales se caracterizan por un alto porcentaje de mutaciones sobre modificadores epigenéticos y genes de reparación del DNA, además de la deleción frecuente de la beta-2-microglobulina, que sugiere una presentación de antígenos disminuida. Los tumores mesenquimales también exhiben hipometilación del DNA, lo que coincide con mayor accesibilidad de la cromatina sobre diversos *enhancers* [20, 21].

La subdivisión de los tumores basales es particularmente interesante porque recientemente se han observado diferencias en la respuesta inmune de cada subtipo. Por mucho tiempo se consideró al cáncer de mama como poco inmunogénico dada su relativamente baja carga mutacional. Sin embargo, se ha observado una tasa de supervivencia elevada entre pacientes del subtipo basal con sobre-expresión de PDL2, lo que sugiere que un subconjunto de pacientes con cáncer de mama podría beneficiarse de terapia inmune [22]. Al caracterizar el microambiente de subtipo basal se identificaron tres grupos definidos por 1) la incapacidad de atraer células del sistema inmune innato, 2) quimiotaxis seguida de inactivación de la inmunidad innata y 3) aumento de

factores inmuno inhibidores. El fenotipo del primer grupo se ha explicado con la amplificación de *MYC*, que induce la expresión de distintas quimiocinas y de PDL1, además de la inactivación de celulas dentriticas y macrofagos, limitando el reclutamiento de celulas adaptativas. El segundo fenotipo estaría justificado por la gran infiltración de fibroblastos asociados a cáncer (CAF) [23], que correlaciona negativamente con la infiltracion de celulas T [22] y que depende del inmunomodulador TGFB; además del efecto inmuno inhibidor que la frecuente mutación de la vía de PI3K-AKT estaría permitiendo [23]. Por las características del tercer fenotipo, este sería el subconjunto de pacientes que podría beneficiarse más directamente de la terapia inmune.

Vistas las enormes diferencias entre los subtipos y los sub-subtipos, se han postulado células de origen diferentes [21]. En principio, la división luminal-basal refleja al epitelio normal de la glándula mamaria, formado por una bicapa de células luminales, que producen leche, y células basales, que expulsan la leche [14]. Así, la capa basal o mioepitelial está formada por células contráctiles, que expresan KRT14, TP63, ACTA2/SMA, MME/CD10 y THY1/CD90; mientras que la capa luminal está formado por células capaces de responder a hormonas, que además de los receptores expresan EpCAM, KRT8, KRT18 y MUC1. Sin embargo, la capa luminal puede separarse en células luminales como tal y progenitores luminales. Mientras las células luminales, se distinguen por los receptores ER y PR; las células progenitoras luminales carecen casi totalmente de estos receptores y en cambio expresa KRT5/6, un marcador de la capa basal en muchos tipos de epitelio. Distintas características de expresión genética y estructura de la cromatina, sugieren a las progenitoras luminales, como células intermedia a las basales y luminales [13].

Al examinar las capacidades de crecimiento de cada tipo de célula, se observó que los tres tipos pueden llegar a generar colonias, pero sólo cerca del 0.1 % de la fracción basal puede producir bicapas que forman estructuras similares a la glándula mamaria cuando son inyectadas en ratones y que además, producen leche si se estimulan apropiadamente. Los progenitores luminales sólo producen células con características luminales, con telómeros muy cortos incluso cuando se emplean muestras de mujeres jóvenes, y niveles elevados de especies reactivas de oxígeno [13]. De este modo, de acuerdo al modelo de carcinogénesis por células troncales, los tumores poco diferenciados, ER-, surgirían de las células más primitivas -de la fracción basal-; los tumores enriquecidos de HER2 y los luminales B, que se han descrito como basoluminales, provendrían de una célula troncal intermedia -los progenitores luminales-; finalmente, se predice que los tumores luminales A se originarían de la transformación de células troncales ER+ [14]. Considerando la escasa división de las células luminales [13], el modelo de carcinogénesis por evolución clonal podría ser más apropiado para abordar el origen de los tumores luminales A, pues

plantea una población de células genéticamente inestables que ganan aptitud por acumulaciòn mutaciones y selección [14, 24]. La célula de origen de los tumores es relevante, porque los tratamientos normalmente eliminan células proliferantes, eliminando la mayor parte del tumor pero ignorando células quiescentes como serían las células troncales [14]

Más allá del origen de cada subtipo, es claro que se trata de entidades molecularmente distintas y que estás diferencias pueden incidir en su comportamiento clínico. Si bien en este trabajo se ahondó en la descripción transcripcional, pueden observarse diferencias entre subtipos en muchos otros niveles como puede ser la tasa de interaciones cis y trans de la red de co-expresión [25] y la activación de vías metabólicas [26]. Aunque no se espera que los subtipos intrínsecos reemplacen a las pruebas inmunohistoquímicas, dada la dependencia en los receptores para la asignación de tratamientos, ni se debe sobresimplificar la heterogeneidad tumoral en estos grupos tan amplios [19]; la clasificación molecular se han establecido como la unidad de descripción del cáncer de mama y será utilizada en este trabajo.

Tabla 1: Características generales de los subtipos del cáncer de mama. HR:receptor hormonal.

PAM50	Basal	HER2E	LumA	LumB
IHQ	75 % TN	66.1 % HER2+/HR-	HR+/HER2- Ki-67 bajo	HR+/HER2- Ki-67 alto
posible origen	fracción basal	progenitores luminales	fracción luminal	progenitores luminales
frecuencia aproximada	20 %	<20 %	30 %	30 %
tratamiento	-	trastuzumab	terapia endocrina	
metilación	baja	-	-	alta (subset)
mutaciones	*TP53* (84 %); *BRCA* (20 %)	*HER2* (80 %)	*PIK3CA* (49 %); *MAP3K1* (14 %)	*PIK3CA* (32 %); *MAP3K1* (5 %)

2. Regulación transcripcional

Conforme se han ido recabando datos, el estudio del cáncer ha sobrepasado el enfoque reduccionista que lo consideraba una enfermedad de genes [27]. Así, ha pasado a considerarse una enfermedad de la desregulación de genes [28, 29], una enfermedad de procesos celulares [30] y vías [31] y, cuando se considera el origen de la desregulación, una enfermedad multiescala, donde las alteraciones subcelulares repercuten en el tejido, al mismo tiempo que las propiedades del tejido -léase irrigación-, repercuten en el fenotipo y eventualmente en el genotipo celular [32]. En otras palabras, se ha ido adoptando un enfoque de biología de sistemas, donde las interacciones

importan, ocurran entre genes o entre escalas. Después de todo, no son los genes aislados los que ejecutan las funciones, sino conjuntos de proteínas que han pasado por procesos regulados de transcripción y traducción y que necesitan señales para entrar en acción o dejar de hacerlo.

La cuestión es que la regulación o desregulación de los genes ya es un problema multiescala, que al menos involucra secuencias regulatorias, factores transcripcionales (TFs), histonas, metilación del DNA, RNAs no-codificantes y la conformación de la cromatina [33]; y esto aceptando el nivel de ruido que implica medir la cantidad efectiva de un gen en una célula cuantificando transcritos en lugar de proteínas [4, 34]. Los mecanismo regulatorios mencionados pueden organizarse en categorías diferentes, como serían epigenética, transcripcional y post-transcripcional, pero en realidad son interdependientes y su presencia simultánea puede ser identificada en una misma muestra, lo que permite los análisis integrativos que serán tratados en la tercer sección de este texto.

El proyecto aquí tratado se acota a 3 mecanismos de regulación de la expresión genética que representan las 3 categorías: en el nivel epigenético se incluye la metilación del DNA; los representantes únicos de la regulación transcripcional, que son los TFs y los microRNAs del nivel post-transcripcional. De esta manera, el proyecto no intenta abarcar todas las interacciones regulatorias de la expresión genética en el cáncer de mama, sino que trata la relación de estos tres mecanismos con los subtipos de este cáncer. La idea es explorar el papel conjunto que podrían tener la metilación del DNA, los factores transcripcionales y los microRNAs, al ser mecanismos extensamente caracterizados, con mediciones robustas. Como antecedente, se describe a continuación cada mecanismo y su relación con el cáncer de mama.

2.1. Nivel epigenético: la metilación del DNA

La regulación epigenética incluye distintas modificaciones de la cromatina que afectan la unión de la maquinaria de transcripción; siendo la metilación del DNA, el mecanismo más estudiado. Se trata de la adición de un grupo metilo al carbono en la quinta posición del anillo de citosina (5mC) [28], que cambia la configuración espacial del surco mayor del DNA, de modo que se crean y pierden sitios de unión para proteínas específicas [35]. Aunque las adeninas también pueden ser metiladas [36], esta modificación ocurre mayormente en los dinucleótidos CG, identificados como CpGs, que se encuentran subrepresentados en el genoma, pero enriquecidos en regiones de 0.5 a 4 Kb de largo conocidas como islas CpG (CGI) [35, 37]. La coincidencia de casi 60 % de los promotores del genoma humano con tales islas, ilustra el potencial de este mecanismo de

regulación transcripcional [35].

El método de detección más simple implica el tratamiento con bisulfito de sodio, que convierte las citosinas sin metilar en uracilo, mientras las citosinas metiladas se mantienen. Después, el estado de metilación puede distinguirse mediante secuenciación o microarreglos. Sin embargo, también se han usado enzimas de restricción sensibles a la metilación e inmunoprecipitación con anticuerpos contra 5mC. Estos métodos pueden acoplarse con microarreglos para lograr un examen de alto rendimiento, pero suelen usarse para la detección de regiones delimitadas [38]. En este sentido, la PCR específica para metilación y el ensayo *MethyLight*, son las herramientas por excelencia para evaluar loci puntuales [39]. De este modo, los esfuerzos para caracterizar el metiloma se basan mayoritariamente en microarreglos, debido al bajo costo, alto rendimiento y buena precisión. Los microarreglos de Illumina han pasado por varias generaciones, cubriendo cada vez mayor parte del genoma, con el llamado *HumanMethylation450K BeadChip* (HM450) cubriendo 485,577 sitios de CpG. Esta plataforma fue empleada para generar epigenomas de referencia por el *International Cancer Genome Consortium*, el *International Human Epigenome Consortium* y por el TCGA para caracterizar más de 7500 muestras de 200 tipos de cáncer diferentes. A pesar de una carencia de pruebas sobre regiones regulatorias distales, el microarreglo HM450 permitió conocer el estatus de metilación del cuerpo de los genes RefSeq y su vecindad [40]

En general, la metilación es proporcional a la frecuencia de di-nucleótidos de CpG en las secuencias, siendo los transposones, las regiones intergénicas no repetitivas y los exones, ejemplos de secuencias altamente metiladas. Las islas CpG no siguen está tendencia, sino que parecen estar protegidas contra la metilación y en cambio, tienen a la RNA polimerasa II unida constitutivamente y una zona río arriba libre de nucleosomas [41]. Cuando la inhibición de los promotores CGI hace falta, se unen de manera transiente complejos represores [37]. De este modo, alrededor del 80 % de los CpGs en el genoma, descontando CGIs, se encuentran normalmente metiladas [28], lo que vuelve citosinas metiladas al 1 % del total de pares de bases [42]. Aunque no se ha podido establecer como causa o efecto, los promotores CGI tienen diferencias esenciales respecto el resto de los promotores, que incluyen regiones de inicio de la transcripción más largas, mayor frecuencia de transcritos en ambos sentidos y menos sitios de unión de TFs [41]. Además de las islas, se pueden distinguir por su densidad de CpG los márgenes de las mismas, con las primeros 2 kb identificadas como costas, y las regiones entre 2 y 4 kb, como plataformas [35].

Aunque la metilación del DNA sufre modificaciones con el paso del tiempo, se mantiene una firma de identidad celular, por lo que representa un mecanismo de programación de la

expresión genética a largo plazo [43,44]. Así, después de la fertilización ocurre una de-metilación generalizada del genoma y se establecen patrones permanentes durante la embriogénesis. La metilación de novo depende de las enzimas Dnmt3a y Dnmt3b y está confinada a las células pluripotentes de los embriones tempranos. Más tarde, durante la división celular, sólo tiene lugar una metilación de mantenimiento, en que DNMT1 se une al DNA hemimetilado y copia las marcas de la hebra parental a la hija, en una forma de memoria celular [35,43]. Además de las DNMTs, la metilación del DNA requiere de S-adenosil L-metionina como co-sustrato, estableciendo un vínculo entre la regulación de la expresión genética y el metabolismo [45].

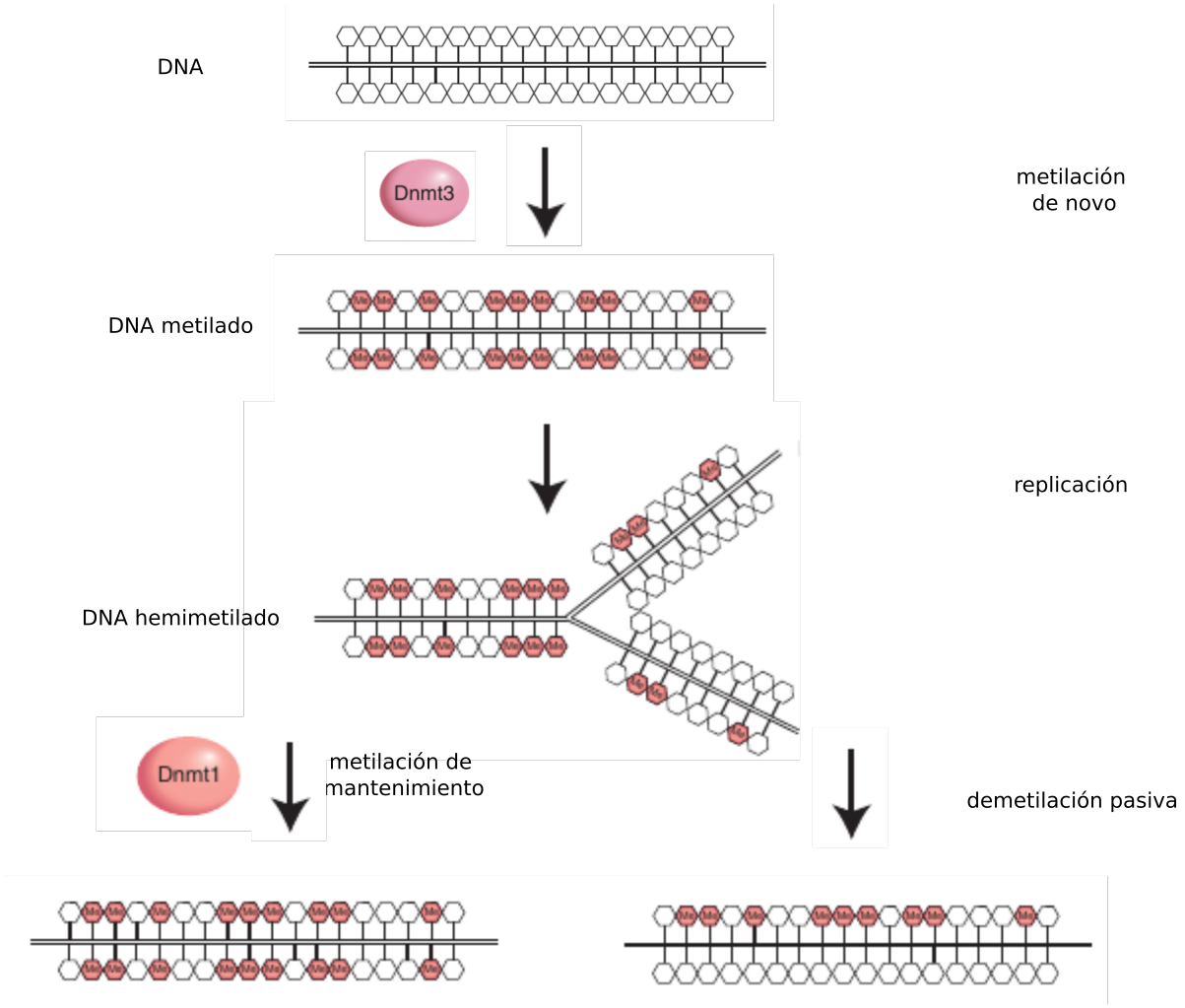


Figura 1: Proceso de metilación del DNA. Figura retomada de [35].

En el sentido contrario, para remover la metilación, se ha planteado un mecanismo pasivo, en que la metilación se va perdiendo conforme pasan las divisiones celulares, además de un mecanismo activo, asociado a las enzimas TET [35]. En la figura 1 se representan los dos tipos metilación del dinucléotido CpG. Adicionalmente, la figura 2 expande el proceso de demetilación

activa. Ambos mecanismos, la demetilación activa y pasiva, se han ligado a los cambios previos a la implantación, con la metilación del genoma materno diluyéndose de manera pasiva y el genoma paterno sufriendo la acción de Tet3. El mecanismo pasivo podría ser el responsable de la pérdida gradual de la metilación del DNA, que se observa mientras los individuos envejecen, más patente en gemelos monocigóticos [46].

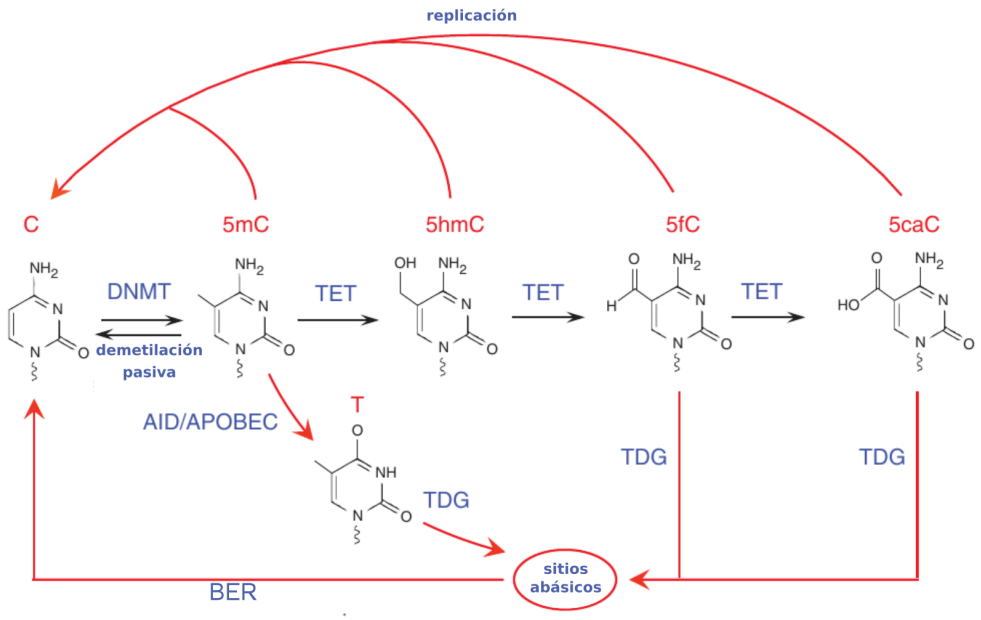


Figura 2: Proceso de demetilación del DNA. TDG: timina DNA glicosilasa; AID/APOBEC: deaminasa de citidinas. Figura retomada de [35].

La familia de proteínas TET (*Ten eleven translocation*) es un grupo de hidroxilasas de DNA encargadas de oxidar el metilo de la citosina y sus derivados de manera sucesiva. La acción de TET1, TET2 y TET3 cataliza el paso de 5-metilcitosina a 5-hidroximetilcitosina (5hmC), que se convierte en 5-formilcitosina (5fC). que a su vez se oxida a 5-carboxilcitosina (5caC). Las formas 5fC y 5caC pueden ser reemplazadas por citocinas mediante la acción de la glicosilasa de DNA y la reparación por escisión de bases (BER). Los tres derivados se encuentran simultáneamente sobre el DNA, pero no puede precisarse por secuenciación de bisulfito, ya que 5hmC se lee como 5mC, mientras que 5fC y 5caC como citosina. La identificación de cada forma es relevante, porque, al contrario de 5mC, los derivados no permiten la unión eficiente de reguladores transcripcionales; sino que 5fC y 5caC favorecen la unión de proteìnas involucradas en las reparación del DNA [47].

La unión de los reguladores transcripcionales sobre las citosinas metiladas depende de proteínas con dominio MBD (*methyl-CpG binding domain*), como MeCP2, que además reclutan deacetila-

sas y metiltransferas de histonas y entonces reconfiguran la cromatina a su forma inactiva [43]. Muchos TFs pueden unirse tanto a DNA metilado como DNA no metilado, pero con distintas afinidades [48], tal es el caso de MYC, que se une al motivo CACGTG a menos que el CpG central haya sido metilado. A diferencia de MYC, la metilación mejora la unión de otros factores transcripcionales como CEBPA y CEBPB [49].

El modelo general dicta que la metilación de los promotores impide la transcripción, al estorbar la unión los factores transcripciones; mientras que la metilación del cuerpo de los genes favorece la expresión genética al facilitar la elongación transcripcional [28, 35, 37, 43]. Sin embargo, hay ejemplos de todo tipo de interacciones entre transcripción y metilación del DNA, pasando por la protección contra la metilación, la promoción de la misma y la de-metilación [49]. La protección contra la metilación atañe a ese 60 % de promotores CGI que normalmente se encuentran de-metilados. Tal protección se asocia con proteínas con dedos de zinc CXXC, que se unen a sitios CpG no-metilados de manera inespecífica a la secuencia. El activador transcripcional CFP1, *CXXC Finger Protein 1*, protege la región, al unirse a CpGs sin metilar y reclutar a la metiltransferasa encargada de colocar la marca de transcripción activa H3K4me. Otro ejemplo es el de las proteínas TET1 y TET3, que también tienen el dominio CXXC y cuya localización sobre promotores no metilados, podría indicar que revierten la metilación de novo errónea. Por otra parte, se ha sugerido que los loops de DNA-RNA resultado de la transcripción activa, protegen a los promotores próximos de la acción de las DNMTs [49].

Por el contrario, la promoción de la metilación del DNA mediante proteínas involucradas en la transcripción, implica el reclutamiento de DNMTs, como se ha documentado que pasa con DNMT3B y los factores transcripcionales MYC y E2F6. Se espera que la mejor caracterización de la familia de TFs KRAB-ZNF provea más ejemplos de esta metilación dirigida, puesto que se unen a complejos represores con DNMT y, 224 de los 330 miembros presentan un motivo RH -histidina unida a zinc, precedida por arginina- que interactúa con el grupo metilo de la 5mC. De manera equivalente, los ejemplos de de-metilación implican el reclutamiento de proteínas TET. SPI1 es un factor de transcripción que interactúa tanto con DNMT3B, como con TET2, la unión con está última tiene como consecuencia la pérdida de la metilación sobre los promotores. Un ejemplo adicional está en la interacción de las proteínas TET y el co-activador PPARG unido a DNA, que desencadena la conversión región-específica de 5mC a 5hmC [49].

2.1.1. Metilación y cáncer

Viendo la importancia que la metilación de DNA tiene sobre la definición del tipo celular a través de la regulación transcripcional, es comprensible su alteración en síndromes y enfermedades. Los síndromes de Prader–Willi, Angelman, Beckwith–Wiedemann y Silver–Russell se han mapeado a aberraciones cromosomales, pero también a defectos en el *imprinting* por la metilación alterada de los genes involucrados: *UPD*, *ICR2* e *CR1* [46]. En cáncer, se han reportado niveles de expresión de la DNMTs similares a los observados en embriones, mientras que la mutación de las enzimas TET se ha identificado de manera recurrente en distintos tumores liquidos [50]. Las alteraciones en la metilación del DNA descritas en cáncer, no se limitan mutaciones o epimutaciones [1] puntuales, sino que incluyen tanto hipermetilación, como hipometilación simultanea de multiples regiones del genoma [43].

La hipermetilación del DNA en cáncer afecta del 5 al 10 % de los promotores CGI -que normalmente no están metilados-, y se ha asociado al silenciamiento de genes supresores de tumores (TSGs) [28], encargados de, por ejemplo, inducir la apoptosis y el arresto celular. Además de silenciamiento epigenético, los supresores de tumores suelen sufrir mutaciones disruptivas como indeles y sustituciones por codones de paro en ambos alelos, pues de acuerdo a la hipótesis de los dos impactos de Knudson, las dos copias del gen debe quedar inoperantes para la inactivación del TSG [27]. La hipermetilación del promotor suele ser el segundo impacto de estos genes y se cree que avanza de manera gradual, desde la heterocromatina circundante, hasta el sitio de inicio de lo transcripción, reduciendo de manera sutil y heterogénea la expresión genética y favoreciendo la plasticidad del tumor [37]. De este modo, incluso las costas de metilación están diferencialmente metiladas en cáncer. La cantidad de CGIs afectados también se incrementa gradualmente, conforme la diferenciación celular disminuye [43].

Alrededor de la mitad de los genes que causan formas familiares de cáncer, se pueden encontrar hipermetilados en tumores esporádicos. En el caso del cáncer de mama, del 10 al 15 % de las mujeres con tumores esporádicos exhiben hipermetilación del TSG *BRCA1*, acompañada de un patrón de expresión consistente con los tumores hereditarios [37]. A parte de los supresores de tumores como tal, la hipermetilación provoca un silenciamiento perjudicial de miRNAs y desregulaciones más complejas, como la interferencia con la unión ER-ERE [51] y la pérdida del imprinting de IGF2. La expresión de IGF2, involucrada en los síndromes de Beckwith–Wiedemann y Silver–Russell, normalmente es inhibida por el insulator H19, que impide la acción de un enhan-

[1] Las epimutaciones son cambios en el epigenoma respecto al consenso, equivalentes a las mutaciones [46], pero reversibles y más frecuentes [38].

cer distal sobre el promotor de IGF2; sin embargo, en distintos tipos de cáncer se ha encontrado a H19 hipermetilado, lo que permite la expresión de la copia materna de *IGF2* y ocasiona un exceso del factor de crecimiento. Como este, abundan los ejemplos de hipermetilación, al punto que hacen falta estrategias de filtrado para identificar sus consecuencias funcionales [28].

De manera equivalente, la hipometilación hace que el porcentaje de sitios CpG metilados en el genoma, pase del 80 al 60 o hasta el 40 % y avanza de manera que las metástasis tienen niveles menores de metilación que los tumores primarios [52]. El fenotipo metilador identificado en un subgrupo de tumores, se caracteriza por la metilación coordinada de gran cantidad de CGIs, similar a lo reportado previamente en cáncer colorectal,y tiene bajo riesgo de metástasis y mejores tasas de supervivencia. Aprovechando estas observaciones, se han encontrando agentes que revierten la de-metilación, inhibiendo la invasividad y metástasis de líneas celulares de cáncer de mama e hígado [43].

Al contrario de la hipermetilación, la hipometilación no ocurre de manera focalizada sobre promotores CGI, sino a gran escala, afectando elementos repetitivos que incluyen transposones y oncogenes y que mapean a regiones de replicación tardía asociadas a la lámina nuclear. La activación transcripcional de las repeticiones predispone el genoma a la recombinación, como evidencia el aumento en la frecuencia de alteraciones cromosomales en cáncer. Los transposones se mantienen bajo control en los tumores basales, debido a una compensación por la pérdida de la metilación mediante la tri-metilación de la lisina 27 de la histona 3 [51]. Mientras la hipometilación propicia indeles y translocaciones, la sóla metilación aumenta la susceptibilidad de las citocinas a la mutagénesis, pues aumenta la tasa de deaminación hidrolítica que, debido al grupo metilo, convierte la base en timina en vez de uracilo, como corresponde a las citosinas, impidiendo la reparación eficiente del daño. Este efecto es tan común, que la mitad de las mutaciones de p53 en cáncer colorrectal pueden ser atribuidas a este fenómeno. Por si fuera poco, la metilación también cambia el espectro de absorción de las citosinas al rango de la luz solar, lo que provoca mayor formación de dímeros de pirimidina en la piel [28].

A pesar de que consistentemente se ha encontrado en cáncer un exceso de variabilidad en los niveles de metilación respecto al tejido normal [52], se conocen patrones específicos, al menos, para los subtipos basal, luminal B y HER2E. El subtipo basal es el más hipometilado y, de acuerdo con lo esperado, también tiene una inestabilidad genómica elevada. Entre las muestras luminales B se ha reconocido un subgrupo hipermetilado, donde los CpGs afectados están ligados a la vía de Wnt [1]. Sobre el subtipo enriquecido de HER2 se ha reportado un sesgo hacia la hipermetilación -sobre la hipometilación- respecto al tejido normal, que se asocia con la amplificación de HER2

y afecta especialmente a los genes HOX [53].

Aunque la regulación por metilación actúa de forma local sobre los genes, la metilación coordinada entre loci lejanos puede reflejar un mismo programa transcripcional. En ese sentido, se ha reportado que más de la mitad de los pares de genes altamente co-metilados -con coeficientes de correlación de Pearson por encima de 0.75- en cáncer de mama, están en cromosomas diferentes y tienden a participar en funciones similares, con un enriquecimiento en las vías de diabetes del adulto de inicio juvenil, linaje hematopoyético, depresión a largo plazo e interacción entre receptores y la matriz extracelular [42]. Otros estudios con distintos tejidos, disputan la distancia a la que se observa co-metilación, pero en cáncer colorrectal se ha reportado correlación positiva a corta distancia y negativa en trans, con pocos genes co-metilados con muchos otros y un enriquecimiento de marcas asociadas al complejo polycomb [54]. Salvando las diferencias entre estudios, un análisis pan-cáncer, que incluye al cáncer de mama, reporta variabilidad por tejido, pero identifica 4 grupos de genes consistentemente co-metilados, dos de los cuales permiten discriminar entre muestras de cáncer y tejido normal, a pesar de contener sólo seis genes asociados a cáncer: *CSF2*, *GALR1*, *IRF4*, *PTPRT*, *SOX11* y *NRG1* [55].

Sin embargo, los niveles de metilación y co-metilación no necesariamente implican un cambio funcional en la célula, hay más mecanismos de regulación en juego y se estima que sólo 15 % de los genes diferencialmente metilados, exhiben también un cambio en la expresión [53]. La búsqueda específica de genes cuya expresión es dictada por la metilación del DNA, encontró en los datos de cáncer de próstata del TCGA, un enriquecimiento de procesos de oxidorreducción, exosoma celular y transporte de electrones; a la par de un enriquecimiento de las vías de metabolismo de drogas, tirosina, histidina, fenilalanina y glutatión. Revisando otros conjuntos de datos, se identificaron procesos asociados a la matriz extracelular y a la diferenciación del sistema nervioso, además de vías ligadas a distintas adicciones [56].

Además de aportar información sobre el origen de los tumores y los genes potencialmente activos, la metilación del DNA ha ganado interés clínico como marcador pronóstico. El DNA es un material relativamente resistente, que puede manipularse con más facilidad que el RNA necesario para medir expresión genética [43] y que puede recuperarse de distintos fluidos corporales dependiendo del tipo de cáncer. Conforme las células del tumor mueren, se libera DNA libre al flujo sanguíneo, donde pueden ser detectados con alta sensibilidad [37]. Por ejemplo, a partir de los niveles de metilación en suero de mujeres con cáncer de mama metastásico, se pudo distinguir un subgrupo con mayor supervivencia libre de progresión, ahora reconocible por la metilación de *SFN*, *HMLH1*, *HOXD13*, *PCDHGB7*, *RASSF1* y *P16* [39]. La hipermetilación de los

elementos de respuesta a estrógeno sirve para predecir la respuesta reducida a terapia endocrina, con la metilación de *PSAT1* como indicador específico de la respuesta a tamoxifeno. También hay numerosos estudios explorando la detección temprana de cáncer a partir de pruebas que miden la metilación del DNA. Tanto su sensibilidad como su especificidad, superan con creces las reportadas para el cribado mamográfico y son aún superiores para estadios avanzados [51].

La otra utilidad potencial de la metilación del DNA, es en el tratamiento del cáncer. El uso de los inhibidores de DNMT como estrategia de sensibilización a otros tratamientos es prometedor para el cáncer de mama, aunque aún no ha sido aprobado para el uso clínico rutinario. Los inhibidores de DNMT, decitabina y 5-azacitidina, son usados en el manejo de malignidades hematológicas y pueden inhibir el crecimiento tumoral de modelos de cáncer de mama ER+ en combinación con quimioterapia o inmunoterapia. Se cree que estos inhibidores activan la respuesta inmune al detener el silenciamiento de los antígenos tumorales. Lo que se ha demostrado es un aumento en la expresión del inmunomodulador PD-L1 en líneas celulares y xenoinjertos tratados con decitabina, que mejora el reclutamiento de células CD8+ y la eficacia de la inmunoterapia. Adicionalmente se ha reportado un beneficio en los pacientes con metilación de *BRCA1* ante el uso de inhibidores de PARP y se cree que la caracterización epigenética de la respuesta a inhibidores de CDK4/6 podría mejorar el manejo de los pacientes con cáncer de mama ER+ y metastásico, que reciben este medicamento en primera línea, pero no siempre responden al tratamiento [51].

2.2. Nivel transcripcional: los factores transcripcionales

Los factores transcripcionales conforman la familia de proteínas más grande, abarcando cerca del 8 % de los genes humanos [57]. Históricamente el término ha descrito cualquier proteína involucrada en la transcripción y su modulación [58], incluyendo tanto a los factores generales que participan en la transcripción de la mayoría de los genes, como a los factores secuencia-específicos, que dirigen los distintos patrones de expresión espaciotemporales de los organismos [57]. De este modo, la base de datos de ENCODE Factorbook resguarda los perfiles de unión de casi 700 proteínas relacionadas a la transcripción, incluyendo factores específicos, co-factores y miembros del complejo de RNA polimerasa II [59].

Entonces, los factores transcripcionales pueden reclutar directamente a la RNA polimerasa o depender de factores accesorios que ejerzan la función. La mayor parte de los TFs eucariotas necesitan complejos co-activadores o co-represores, con múltiples subunidades o dominios,

involucrados en la remodelación de la cromatina [60]. Aunque también hay TFs cuya función simplemente consiste en la interferencia con la unión de otras proteínas [58].

La base de datos *HumanTFs* acota los factores transcripcionales a proteínas que se unen al DNA, normalmente a través de un dominio de unión a DNA (DBD), y regulan la transcripción. Con esta definición recuperan casi en su totalidad los TFs de Fulton et al. y Vaquerizas et al. [61,62], además de nuevas adiciones. El conjunto de 1639 probables factores transcripcionales humanos está enriquecido de los DBD C2H2-ZF y homeodominio, que aparecen en el 54 % de los TFs. La mayoría de los factores tienen varias copias de un sólo tipo de DBD y una combinación de dominios efectores de entre 391 dominios diferentes. Los patrones de expresión de los TFs dependen en gran medida del DBD, con los homeodominios siendo tejido-específicos y el 88 % de los C2H2-ZFs con el dominio KRAB expresados sin especificidad [58].

Dada la importancia del DBD, los factores transcripcionales pueden agruparse de acuerdo a la familia del dominio, que equivale a un agrupamiento por las secuencias que reconocen [59]. De hecho, muchos TFs fueron propuestos originalmente por homología, haciendo que las familias más grandes C2H2-ZF, homeodominio, bHLH -*basic helix-loop-helix*-, bZIP -*basic leucine zipper*- y NHR -*nuclear hormone receptor*-, sean también las descritas con mayor antigüedad [63]. Esto podría haber restringido la identificación de nuevos factores, pero está justificado por la historia evolutiva de los DBD, que provendrían de un set pequeño de ancestros comunes que experimentaron duplicación y divergencia [58].

Aunque no todas las instancias de un dominio tienen realmente la misma especificidad de secuencia, ortólogos de TFs tan lejanos como la relación entre humano y *Drosophila* se unen a las mismas secuencias [64]. Los sitios de unión de los factores transcripcionales son secuencias de entre 6 y 20 pares de bases, en la región regulatoria de los genes diana, denominadas motivos [59]. Los motivos se determinan mediante la identificación de tantas secuencias de unión como sea posible, por métodos como ChIP-seq y HT-SELEX [59]. Posteriormente, las discrepancias aceptadas por el TF son caracterizadas por matrices de peso o modelos ocultos de Markov, en colecciones de motivos tales como JASPAR [65].

Un mismo factor puede regular genes diferentes y un gen puede ser regulado por distintos TFs, sin que la presencia de un motivo baste para saber si el factor transcripcional realmente regula al gen. Únicamente CTCF, un factor estructural -y general-, se une a las casi 14000 instancias de su motivo en el genoma [66], pues la unión de los factores transcripcionales depende de la accesibilidad de la cromatina, que a su vez depende del estado de metilación del DNA, la posición de los nucleosomas y la unión de otros TFs. Para unirse a sus motivos, los factores

transcripcionales tienen que competir o interactuar con los nucleosomas [58]. La unión de los TFs asociados con el re-posicionamiento de nucleosomas anti-correlaciona con el nivel de metilación del DNA [3], mientras que la ausencia de nucleosomas en una región indicaría un nivel elevado y sin fluctuaciones del transcrito correspondiente. Por su parte, el seguimiento de molécula única ha revelado que los TFs se unen al DNA de manera transiente y cualquier interacción, como las que se darían entre TFs, puede retrasar la difusión [67].

Aunque los factores transcripcionales han sido divididos en activadores y represores, muchos TFs pueden reclutar múltiples cofactores con efectos opuestos, haciendo que lo más adecuado sea incluir la diana y la condición bajo la cual está funcionando un factor. Los factores KRAB C2H2-ZF son represores de los elementos transponibles, al promover su silenciamiento [58]; mientras que HOXA5 funciona como activador de P53 en células de cáncer de mama [68, 69]. Entonces, la unión al motivo puede ser insuficiente para determinar el efecto del TF sobre el locus, y simplemente reflejar la accesibilidad de la cromatina [57].

Por otra parte, la unión de un TF cerca del sitio de inicio de la transcripción puede servir para predecir el nivel de expresión genética. La unión de E2F4 sobre los promotores CGI, que no suelen estar metilados, explica 47 % de la varianza en la expresión; pero este porcentaje baja hasta 14 % en los promotores con baja densidad de CpGs, donde la metilación jugaría un papel regulatorio. Mientras tanto, la unión de los factores generales explica hasta 73 % de la varianza en la expresión genética y este porcentaje crece menos del 15 % cuando se incluyen los factores secuencia-específicos y las modificaciones de histonas en el modelo, sugiriendo que los distintos mecanismos regulatorios son fuertemente redundantes. Los genes más difíciles de predecir son aquellos regulados de manera post-transcripcional, involucrados en el control del ciclo celular y con diferencias de expresión significativas entre tejidos [57]

Al estudiar 2073 pares de genes y sus factores transcripcionales en 135 condiciones, Inoue et al. identificaron cuatro patrones de expresión: 1) sin cambio, 2) expresión correlacionada, 3) expresión no correlacionada por que el gen se mantiene constante a pesar de la variación del factor transcripcional y 4) falta de correlación debido a niveles constantes del TF y variables del gen. Al clasificar los pares de acuerdo al patrón mostrado en la mayoría de las condiciones, encontraron que la expresión correlacionada se asocia con genes de ciclo celular y la replicación del DNA, mientras que diversas enfermedades humanas están ligadas al tercer patrón y el metabolismo y la transducción de señales al cuarto. Aunque se asume que la expresión correlacionada es la regla, menos de 20 % de los pares entran en esta categoría y en cambio resalta la falta de correlación, que acepta mecanismos de regulación adicionales. En particular, el tercer patrón se asocia con

genes cuya expresión es determinada por la degradación del transcrito, y no por la síntesis, debido a la unión constante del TF sobre su motivo; haciendo que la alteración de la degradación sea perjudicial, como sucede con la acumulación del oncogen β-catenina [70].

2.2.1. TFs y cáncer

Los factores transcripcionales regulan una gran cantidad de procesos biológicos y son esenciales para el mantenimiento de la homeostasis, por lo cual no es sorprendente que su alteración esté asociada con distintas enfermedades. En particular, los TFs representan casi el 20 % de los oncogenes identificados [50]. Sin embargo, los factores transcripcionales no sólo son afectados por mutaciones directas, sino que la mutación y metilación de las regiones regulatorias puede perturbar su unión y funcionamiento, como sucede en cáncer de colon con la unión de TCF4 a un *enhancer* que promueve la expresión de MYC y que al estar mutado, altera los niveles del oncogen [58].

Además, hay una gran cantidad de cascadas transcripcionales disparadas por la acción de unos cuantos factores, que funcionan como reguladores transcripcionales maestros. Los reguladores maestros son los genes que controlan la especificación de un linaje ya sea por regulación directa o indirecta, cuya expresión alterada puede cambiar el destino celular [71]. En cáncer de mama se ha identificado a AGTR2, ZNF132 y TFDP3 como reguladores maestros ligados a los rasgos distintivos del cáncer. Centrando el análisis en las vías de transducción de señales también se identificó a TSHZ2, HOXA2, MEIS2, HOXA3, HAND2, HOXA5, TBX18, PEG3, GLI2 y CLOCK, siendo este último el único regulador positivo. Los reguladores de ambos conjuntos muestran cierta redundancia en sus dianas, lo que sugiere robustez en la regulación. En el caso de la transducción de señales resalta la vía de *Hedgehog*, por su relación con la morfogénesis y la auto-renovación de las células troncales [69, 72].

Esta relación del cáncer con la morfogénesis y la diferenciación celular, encaja con la teoría oncogerminativa del cáncer, según la cual, la expresión aberrante de los genes del desarrollo permite la reprogramación de las células somáticas a un linaje inmortal de células troncales del cáncer y después de eso, hacia una nueva identidad celular [73]. La transformación epitelio-mesenquimal es un buen ejemplo de esta teoría, pues depende de los mismos factores transcripcionales -SNAIL, SLUG, TWIST y FOXD3- tanto durante el desarrollo, como durante la progresión del cáncer. Eventualmente también la metástasis se asemeja al desarrollo embrionario de distintas estructuras, al depender de los mismos morfógenos: los ligandos Wnt y *Hedgehog*, las proteína

morfogenéticas de hueso (BMPs), y los factores de crecimiento fibroblásticos (FGF) [50].

Por otro lado, mientras la alteración de los factores transcripcionales o su expresión modifican procesos completos, la alteración de los motivos de unión también tiene un efecto, quizá más acotado, al afectar únicamente la relación entre el TF y un gen diana, pero igualmente problemático. Al analizar la accesibilidad del DNA en 23 tipos de cáncer diferentes, se encontraron cientos de mutaciones no codificantes y somáticas, que afectan la unión de los factores transcripcionales, sugiriendo un mecanismo ubicuo de manipulación de la expresión genética. El agrupamiento de los tipos de cáncer por accesibilidad del DNA concuerda con el agrupamiento por multi-omica -expresión de transcritos, microRNAs y proteínas, además de la metilación del DNA y el número de copias-, sugiriendo relevancia funcional. Las regiones accesibles y específicas de un grupo están hipometilados respecto a los otros *clusters*, al mismo tiempo que exhiben un enriquecimiento de SNPs y motivos para TFs asociados al cáncer mejor representado en el *cluster*. Alrededor del 65 % de estos SNPs no tienen como diana putativa al gen más cercano. Al enfocarse en el cáncer de mama, se observó un 36 % de regiones accesibles que también lo son en el resto de los tipos de cáncer, una división entre los tumores basales y los no basales, además de una diferencia en la supervivencia que depende de la accesibilidad de los motivos de *ESR1* [3].

De entre los 294 TFs oncogènicos [74], resaltan los receptores de andrógeno y estrógeno, los genes *BRCA1* y *BRCA2*, *MYC* y *GATA3* por su asociación con los subtipos del cáncer de mama. El receptor de andrógeno ha sido asociado con el subtipo HER2E [19], aunque también tiene relevancia clínica, y de hecho es más común en los tumores ER+ [75]. Por su parte, el receptor de estrógeno es el marcador por excelencia de los subtipos luminales; mientras que las mutaciones -germinales o no- de los genes de susceptibilidad al cáncer de mama y la activación de MYC son frecuentes en el subtipo basal. Finalmente, el factor transcripcional GATA3 está particularmente mutado en los tumores luminales, donde además suele estar sobre-expresado [1].

Dada la relevancia del receptor de estrógeno en la clasificación del cáncer de mama, vale la pena ahondar en su funcionamiento. Además de su rol como factor transcripcional, ER es un miembro de la superfamilia de receptores nucleares de hormonas, codificado por los parálogos *ESR1*, en 6q25.1 y *ESR2*, en 14q22-24. Los receptores que resultan de cada gen, ERα y ERβ, respectivamente, tienen expresión tejido específica y diferencias en cuanto a la estructura y la unión a DNA, que sin embargo, permiten la formación homodímeros y heterodímeros con una afinidad similar por el DNA. Las hormonas esteroideas difunden a través de la membrana plasmática y una vez que el dominio de unión a ligando del receptor recibe estrógeno, se forma un dímero estable, capaz de interactuar con secuencias específicas a través del dominio de

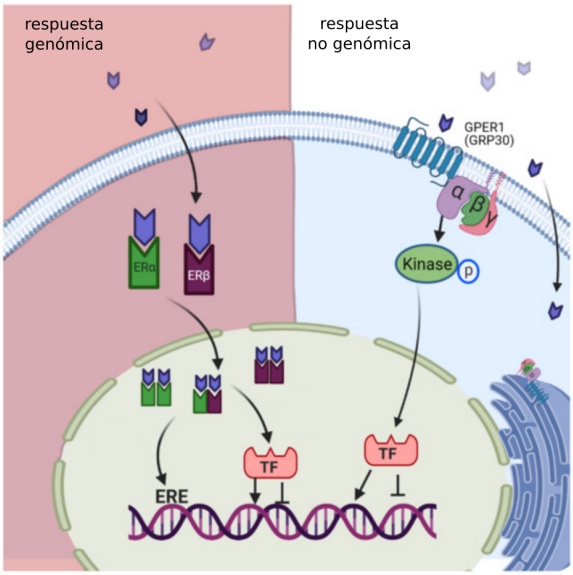


Figura 3: Mecanismos del receptor de estrógeno. Retomado de [76]

unión a DNA. Los elementos de respuesta a estrógeno (EREs) son palíndromos de 5 pares de bases separados por 3 bps, cuya secuencia consenso es GGTCAnnnTGACC. Cuando el receptor activado se une al ERE, se cree que se forma un complejo de preiniciación para la RNA polimerasa, mediante la inactivación o disociación de co-represores y el reclutamiento de co-activadores, que favorece la proliferación celular [11].

De manera adicional al ER nuclear, hay receptores en la membrana plasmática y en la mitocondria. En la membrana, el ER se asocia con las balsas de lípidos, interactúa con receptores de factores de crecimiento como EGFR y HER2 y participa en respuestas no genómicas al estrógeno,que abarcan desde la activación de cinasas, hasta la modulación de la migración, la supervivencia y la proliferación celular. En la mitocondria, la presencia de ERβ incide sobre el metabolismo y las señales anti-apoptóticas. La figura 3 muestra los dos tipos respuesta que involucran al receptor, genómica y no genómica.

La actividad del receptor cambia con la naturaleza del ligando, la fosforilación y la interacción con otros TFs. El ER puede propiciar la transcripción sin necesidad de hormona, ya sea mediante la interacción con el factor transcripcional SP1 y sus elementos de respuesta o porque los factores de crecimiento extracelulares provocan la fosforilación y activación del ER, entrecruzando las vías de transducción de señales y receptores esteroideos. La interacción con otros TFs explica la activación de genes sin ERE, mientras que la interacción de ER con las ciclina D1 permite la unión del receptor a los EREs, también sin necesidad de estrógeno y de manera aditiva cuando hay hormona.

Además, la función de ER depende de la expresión del receptor, que está sometida a regulación en múltiples niveles. El promotor del receptor contiene los motivos de distintos factores transcripcionales como SP1, FOXA1 y EZH2; a parte de varios EREs incompletos. Por su parte, las seis isoformas conocidas del mensajero codifican para la misma proteína, pero exhiben patrones de expresión tejido específicos y comprenden 5'UTR diferentes, que parecen plegarse con más o menos estabilidad y podrían alterar la eficiencia de la traducción. Del lado opuesto, el 3'UTR contiene las semillas de 72 microRNAs, incluyendo a miR-22, miR-206, miR-221 y miR-222, que están sobre-expresados en tumores ER-, respecto a los ER+ y; al cluster miR-17-92 -miR-18a, miR-19b y miR-20b-, cuya expresión depende de ERα y MYC, formando un bucle de retroalimentación negativo. Normalmente los 29 CpGs sobre *ESR1* carecen de metilación, sin embargo se ha documentado una extensa metilación en líneas celulares ER- [11].

La principal alteración del ER durante la progresión del cáncer de mama es en cuanto a su expresión genética. Aunque el tejido normal sólo presenta ERα, los tumores ductales tempranos tienen niveles elevados de ERα y bajos de ERβ, mientras que en los estadios más avanzados se pierden ambos receptores. Por el contrario, los tumores lobulares empiezan con niveles elevados de los dos receptores y terminan perdiendo a ERβ [77]. Las grandes disrupciones y la pérdida de la heterocigosidad raramente afectan al receptor, por lo que no pueden usarse para explicar el estatus ER-. En otras palabras, hay pocas mutaciones documentadas en tumores primarios, que en cambio se vuelven frecuentes en las lesiones metastásicas. Por ejemplo la mutante Y537N, que ha sido ligada a metástasis en hueso y permite la activación constitutiva del TF, al abolir el sitio de fosforilación. Además, cerca del 7 % de los tumores tienen mutaciones en los *enhancers* ligados a *ESR1* [13]. Entonces, la alteración en cáncer de mama del ER es más bien a nivel de la expresión y tiene efectos transcripcionales.

Estudios de precipitación de la cromatina señalan entre 5000 y 1000 EREs, que se reducen a cerca de 1500 genes de respuesta a estrógeno [11]. Sin embargo, el efecto del TF no es únicamente local. Inicialmente se describió que ERα, FOXA1 y AP-2γ mediaban la interacción de larga distancia entre GREB1 y TFF1, pero gracias a estudios de ChIA-PET, ahora se conocen 689 bucles de cromatina formados por la interacción entre EREs distales y proximales. Los bucles se forman tanto de manera intracromosomal como intercromosomalmente y se cree que forman subcompartimentos en el espacio nuclear [11, 78].

2.3. Nivel postranscripcional: los microRNAs

Los microRNAs, miRNAs o miRs son reguladores post-transcripcionales de la expresión genética [38], que inhiben la traducción por complementariedad de bases, ya sea induciendo la degradación de los mensajeros si la coincidencia entre secuencias es perfecta, o secuestrando los transcritos diana cuando no lo es [79]. También hay evidencia de que los miRNAs podrían incidir positivamente en la traducción, al aumentar la biogénesis de ribosomas o reclutando complejos a la región rica en AU de los mensajeros [80]. Se trata de RNAs no codificantes, de cadena única, de alrededor de 22 nucleótidos, cuya secuencia está evolutivamente conservada [81] y cuya expresión depende del tejido y el contexto celular [82].

Se estima que un tercio de los genes codificantes son susceptibles de regulación por miRNAs, por lo cual, este tipo de regulación influye casi todos los procesos celulares [80], incluyendo la proliferación celular, la diferenciación y la apoptosis [83]. Además, un microRNA puede regular simultáneamente múltiples procesos, haciendo que la de-regulación de pocos miRNAs, cambie en gran medida el perfil de expresión de una célula [80]. En consistencia con esto, los miRNAs son los RNAs pequeños más comunes en los tejidos somáticos [80] y son esenciales para mantener la cohesión de la red transcripcional [84].

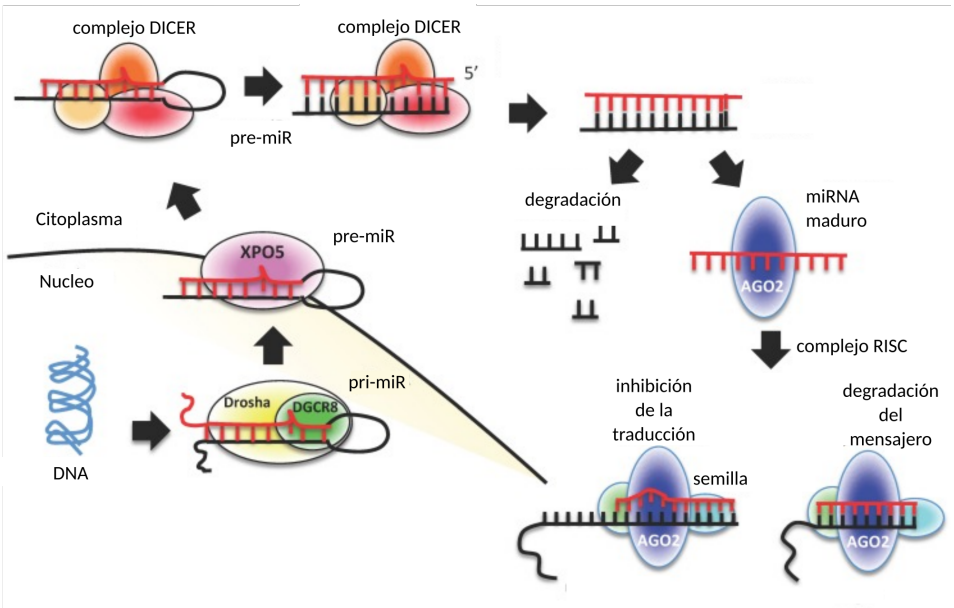


Figura 4: Biogénesis de los miRNAs. Retomado de [80]

La producción de microRNAs sigue una serie de pasos, representados en la figura 4, que inician con la transcripción de *primary miRNAs* o pri-miRNAs por la RNA-polimerasa II [85]. Los pri-

miRNAs son RNAs no-codificantes de más de 1 kb, con la poliadenilación y el casquete 5'de todos los mensajeros, pero también con estructura secundaria [86]. Los pri-miRNAs son reconocidos por el complejo micro-procesador de DROSHA, una RNAse III endonucleasa, y DGCR8, una proteína de unión a RNA, que producen otra forma precursora, los pre-miRNA. Los pre-miRNAs comprenden entre 60 y 110 nucleótidos que forman un asa-bucle y son exportados del núcleo por el complejo XPO5/RANGTP. Ya en el citoplasma sufren un nuevo corte, esta vez por DICER, una RNAsa III que lleva los pre-mIRNAs a una forma transiente, de doble cadena y con la extensión correcta de entre 21 y 23 bases, más dos nucleótidos en cada extremo 3' [87]. En asociación con los cofactores TRBP y PACT, DICER transfiere el miRNA al complejo RISC, donde la hebra guía se mantendrá, mientras la hebra pasajera es degradada [80,81]. Aunque durante el desarrollo se transcriben pri-miRNAs que no llegan a la forma madura, en los tejidos adultos y sanos el nivel de miRNAs maduros es similar al nivel de pri-miRNAs, sugiriendo una producción eficiente [88].

En la mitad de los casos la transcripción de los pri-miRNAs parte de promotores propios y en la otra mitad de promotores de genes codificantes, que contienen pri-miRNAs en los intrones o incluso en los exones [86]. Estos precursores pueden ser mono o policistrónicos [79], con miR-200 siendo un ejemplo del segundo caso. Los microRNAs se anotan en familias por su similitud de secuencia, que refleja una ancestría en común y una funcionalidad compartida. La familia de miR-200 está formada por 5 miRNAs que se transcriben desde dos loci distintos: miR-200b, miR-200a y miR-429 están codificados en la región intergénica del cromosoma 1, mientras que miR-200c y miR-141 están en el cromosoma 12 [89]. Otras familias se transcriben desde loci distintos, pero comparten la misma secuencia madura, como sucede con miR-9-1 (cromosoma 1), miR-9-2 (cromosoma 5) y miR-9-3 (cromosoma 15), dificultando el mapeo entre miRNAs maduros y promotores específicos con modificaciones epigenéticas específicas [85].

Ya dentro del complejo RISC (*RNA-induced silencing complex*), los microRNAs sirven como guía hacia los mensajeros complementarios. La complementariedad del mensajero se da específicamente en la región 3'UTR, donde se suelen localizar los elementos de respuesta a miRNA, MRE por sus siglas en inglés [81]. En el caso del microRNA, hay una semilla de 6 a 8 nucleótidos en el extremo 5' que determina la unión con el mensajero. Una vez unidos, dependiendo del grado de homología, el mensajero puede ser degradado por las proteínas argonauta AGO1-4 [80].

A pesar de la aparente simplicidad de la unión de dos secuencias complementarias, el tamaño y la baja especificidad de los miRNAs dificulta la predicción de mensajeros diana. Por ello, a parte de secuencias en sí, deben considerarse la conservación y la estabilidad termodinámica [80]. No sólo se han construido algoritmos basados en la secuencia, sino también algoritmos basados en

la expresión génica; estos últimos pueden separarse en los que explotan la correlación negativa entre un miRNAs y sus dianas potenciales, y aquellos que siguen enfoques más complejos y computacionalmente costosos, como puede ser la predicción causal invariante [90]. Dada la dificultad para predecir mensajeros diana, las bases de datos como miRanda, TargetScan o miRTarBase, que guardan tanto predicciones como casos validados, son de gran utilidad [91].

2.3.1. miRNAs y cáncer

Contraintuitivamente a su rol pleiotrópico, muchos microRNAs se encuentran en regiones frágiles del genoma y sufren de alteraciones en el número de copias [92], como pasa con miR-125b, let-7g, miR-21 y 72.8 % de los miRNAs asociados con cáncer de mama [29]. Aunque las mutaciones sobre microRNAs específicos no tienen un efecto acotado, las alteraciones sobre el proceso de producción de los miRNAs afectan la célula de manera aún más amplia, pues alteran de manera simultánea a múltiples reguladores pleiotrópicos. De este modo, las mutaciones de *DROSHA* y *DICER* están ligadas con baja supervivencia en pacientes con cáncer de ovario, pulmón y mama. La alteración en la expresión genética se ha atribuido a los reguladores, MYC y ADARB1 en el caso de *DROSHA*, y miR-103/107 y let-7 en el caso de *DICER*. La sub-expresión de *DICER* está asociada con el subtipo basal de cáncer de mama [80]. Curiosamente, hay miRNAs que se sobre-expresan cuando *DROSHA* o *DICER* están sub-expresadas, sugiriendo un mecanismo alternativo. La unión de KSRP al complejo RISC junto con algunos pre-miRNAs, como miR-21, postula a esta proteína de *splicing*, que es inducida ante el daño a DNA, como posible parte de ese mecanismo [86].

Otros componentes del complejo microprocesador que se encuentran alterados en cáncer son DGCR8 y las helicasas p68 y p72, que conectan al complejo microprocesador con p53. En el siguiente paso en la producción de miRNAs, se han identificado mutaciones inactivantes de *XPO5* en tumores con inestabilidad de microsatélites de colon, gástricos y endometriales. La mutación de *XPO5* aumenta el riesgo de cáncer de mama. La fosforilación de XPO5 por MAPK/ERK en cáncer de hígado, tiene el mismo resultado que las mutaciones inactivantes, al impedir la exportación de pre-mIRNAs al citoplasma. Fuera del núcleo, los factores asociados con DICER, como TARBP2 y AGO2, también exhiben alteraciones. La mutaciones de *TARBP2* identificadas en carcinomas con inestabilidad de microsatélites, cambian el marco de lectura del gen; mientras que su sub-expresión se asocia con melanomas y tumores metastásicos de mama y próstata. La sobre-expresión de *AGO2* se ha reportado en tumores de mama, gástricos y de cabeza y cuello [93].

Al depender de la transcripción, los miRNAs también son modulados por la metilación del DNA y la unión de factores transcripcionales. Se estima que cerca de un 33 % de los miRNAs de-regulados en cáncer tienen alteraciones en la metilación del DNA [88]. En líneas celulares sin DNTM1 ni DNTM3B, se observa expresión de miRNAs placentarios, normalmente silenciados. Al respecto, se ha reportado un sobrelape importante entre los microRNAs marcados por el complejo de silenciamiento Polycomb en células troncales embrionarias y aquellos con metilación CGI en células tumorales [94]. Para mencionar un ejemplo puntual, se puede hablar de miR-205, cuya sub-expresión se asocia con la metilación de su promotor y con la resistencia a tratamiento y la transformación epitelio-mesenquimal (EMT) [79].

Ejemplos de regulación transcripcional de los microRNAs incluyen la regulación de MYC sobre miR23a y de NFkB sobre miR-29b [82]. El caso de miR-29 es interesante, porque se conocen tanto reguladores, como efectores del miRNA. La unión de MYC parece ser el paso inicial del silenciamiento, y se sigue del reclutamiento de modificadores de histonas. Al ser parte de los llamados "epi-miRNAs", la familia miR-29 inhibe a DNMT3A, DNMT3B y a DNMT1 y con ello, la metilación del DNA en cáncer de pulmón y leucemia mieloide aguda [79, 82].

Finalmente, el microambiente tumoral también puede alterar los niveles de miRNAs, como se ha observado en tumores hipóxicos de mama, donde la hipoxia inhibe a las de-metilasas de histonas dependientes de oxígeno KDM6A y KDM6B. Como resultado, la metilación -a nivel de histonas- del promotor de DICER aumenta y su expresión disminuye y con ello también disminuye el procesamiento de miRNAs. La familia de miR-200 es de las principales afectadas por la sub-expresión de DICER [86]. Al regular la expresión de los factores transcripcionales ZEB1 y ZEB2, que inhiben la transcripción de genes epiteliales como la E-caderina; la pérdida de miR-200 favorece la transformación epitelio-mesenquimal y se asocia con tumores de mama metaplásicos y agresivos [83]. De manera paralela a la EMT, la pérdida de miR-200 libera al factor transcripcional ETS1 de la represión del miRNA. ETS1 regula la expresión de factores angiogénicos y junto con ELK1, dispara la metilación -a nivel DNA- del promotor de DROSHA, reduciendo aún más los niveles de miRNAs, lo que se ha asociado con tumores poco diferenciados [83, 86].

Por otro lado, es común encontrar miRNAs circulantes en fluidos como plasma y saliva. Los microRNAs en el suero sanguíneo incluso pueden usarse como biomarcadores con capacidad pronóstica en el cáncer de mama, próstata, colón, ovario y pulmón. Específicamente la detección de mir-21, miR-92a, miR-10b, miR-125b, miR-155, miR-191, miR-382 y miR-30a permitiría identificar de manera temprana al cáncer de mama [38]. Estos miRNAs están protegidos de la acción de la RNAsas gracias a la unión con lipo y ribonucleoproteínas o por el empaquetado

en microvesículas [80]. Una vez que son endocitados, la regulación de la traducción en las células receptoras se ve alterada, implicando a los microRNAs como moléculas señalizadoras. En este sentido, se ha demostrado que los fibroblastos asociados a cáncer secretan un espectro de miRNAs diferente respecto a los fibroblastos normales, y estos no son los únicos componentes del microambiente liberando microRNAs [81].

Aún descartando los miRNAs circulantes, hay una diferencia clara entre los perfiles del tejido normal de la mama y los tumores, con miR-10b, miR-125b, miR-145, miR-21 y miR- 155 mostrando las diferencias más significativas [92]. Además, los perfiles de expresión de miRNAs permiten distinguir entre subtipos de cáncer de mama [1] y entre subpoblaciones celulares, siendo los progenitores luminales las células más similares a los tumores basales y las células luminales maduras las más cercanas a los tumores del subtipo luminal B. Los microRNAs luminales regulan la diferenciación y el desarrollo celular; mientras que los basales regulan la localización intracelular, el transporte y la biosíntesis de organelos, la secreción y la interacción entre células [95]. A pesar de que la correspondencia entre los subtipos intrínsecos y los perfiles de miRNAs es ruidosa [1], la sobre-expresión de miR-206 se ha asociado con los tumores ER- y la sub-expresión de miR-125a/b con aquellos enriquecidos de HER2 [83].

MiR-206 inhibe la expresión de *ESR1*; mientras que su expresión es favorecida por ERα y no por ERβ ni progesterona, sugiriendo un ciclo de retroalimentación negativa. Otros miRNAs que regulan a *ESR1* son miR-18a/b, miR-193b y miR- 302c, cuya expresión, junto con la de miR-206, induce el arresto del ciclo celular e inhibe la proliferación ligada a estrógeno. Además, miR-17-5p tiene el mismo efecto, debido a una regulación indirecta de ERα, a través de AIB1.

El perfíl de microRNAs de las células troncales del cáncer de mama también es diferente, al estar enriquecido de miRNAs asociados a la autorenovación, tales como let-7 y miR-34. Let-7 regula a oncogenes como HRAS, HMGA2, MYC y caspasa-3. MiR-34 regula el mantenimiento del fenotipo en cáncer de colon, a través de la señalización Notch; mientras que en cáncer de mama, la sobre-expresión de miR-34 ocasiona arresto en el ciclo celular y su sub-expresión aumenta la capacidad invasiva [80].

Los miRNAs con un rol en el cáncer pueden funcionar como oncogenes o como supresores de tumores, dependiendo de sus dianas. Los microRNAs oncosupresores inhiben la expresión de genes que favorecen el desarrollo del tumor, por lo que su sub-expresón es perjudicial, como es el caso de miR-200. Los oncomiRs por el contrario, regulan a supresores de tumores y es su sobre-expresión lo que resulta perjudicial, como sucede con miR-21, que regula promotores de la apoptosis y de la migración celular [80]. Además, podría definirse una subcategoría de oncomiRs

con los miRNAs exclusivamente pro-metastaticos, como son miR10b y la familia miR-373/520c. Se ha reportado que miR-10b está sobre-expresado sólo en las células metastásicas de cáncer de mama y no en el tumor primario; miR-10b inhibe al factor transcripcional HOXD10 y con ello, provoca una cascada de alteraciones que terminan con la expresión pro-metastásica de RHOC, la migración celular y la invasión [83].

Sin embargo, el rol de un miRNA podría depender del contexto celular, ya que las interacciones regulatorias miRNA-mRNA no necesariamente existen en todos los tipos de cáncer [90]. En un estudio pan-cáncer, computacional, de miRNAs que dirigen la expresión genética, se observó que las interacciones miRNA-gen no se conservan, a pesar de que hay 22 miRNAs que sí fungen como *drivers* en distintos tipos de cáncer. Exceptuando a miR-5001 en cáncer colorrectal y a miR-2276 en cáncer endometrial, en este estudio todos los miRNAs son calificados como supresores de tumores y la familia let-7 funciona como TSG y como oncomiR al mismo tiempo [96].

A pesar de que cada miRNA puede regular cientos de genes, se ha planteado a los miRNAs como posibles medios para regular genes del cáncer, ya sea introduciendo oligonucleótidos similares a miRNA para restaurar la expresión del miRNA y suprimir oncogenes o introduciendo antagonistas para inhibir al miRNA de interés. Un ejemplo de antagonistas o antagomiRs, son las esponjas de miRNA, mensajeros sintéticos con múltiples sitios de unión para un miRNA específico, que entonces lo captan, impidiéndole inhibir TSGs. Hay formulaciones de oligonucleótidos similares a miRNA, de esponjas de miRNAs, de oligonucleótidos anti-miRNA y de pequeñas moléculas que están siendo estudiadas en modelos de cáncer. Para cáncer de mama al menos se han probado un antagomiR-10b y un oligo similar a miR-195. El antagonista inhibe la metástasis a pulmón, no así el crecimiento del tumor primario en ratón; mientras que el oligo similar a miR-195 aumenta la sensibilidad al tratamiento e inhibe la traducción de RAF1 y BCL2 en líneas celulares [80, 81].

3. Integración multi-ómica

Ya que se han descrito los mecanismos regulatorios a tratar, es posible hablar de su interacción. Y es que el nivel de transcrito de un gen no necesariamente depende sólo de la metilación o de la presencia de un factor transcripcional, sino que depende, en algunos casos y probablemente no en toda situación, de una combinatoria de los distintos mecanismos. El asunto es que no sabemos cuáles son esos casos ni situaciones o si esto afecta el desenlace del desarrollo tumoral. Por esta razón, se plantea el análisis integrado de la metilación del DNA, de la expresión de factores transcripcionales y de miRNAs, de los subtipos de cáncer de mama.

La integración multi-ómica se refiere a la interpretación conjunta de distintas ómicas y es posible gracias a la acumulación de mediciones obtenidas con distintas tecnologías de alto rendimiento y su depósito en bases de datos, en particular bases de datos públicas. En los estudios más ambiciosos también se habla de pan-ómica. Al implicar grandes cantidades de mediciones, en distintas unidades, no necesariamente sincrónicas, es válido hablar de una perspectiva compleja que demanda herramientas estadísticas propias. De esta manera, se trata de una aproximación con un periodo de gestación largo, en que hizo falta estandarizar las tecnologías de alto rendimiento, alcanzar un mínimo de muestras para satisfacer los requisitos estadísticos y ajustar las herramientas computacionales a los nuevos objetivos. Esta sección abordará justamente las herramientas computacionales que se han desarrollado, al constituir tanto antecedentes como métodos de interés.

3.1. Generalidades de la integración computacional

La promesa de la integración multi-ómica es aportar una perspectiva más completa del cáncer al considerar los distintos niveles funcionales, en lugar de concentrarse en un sólo aspecto de este heterogéneo fenómeno. Específicamente se han mencionado tres objetivos:

- descubrir mecanismos moleculares, así como su asociación con los fenotipos;
- agrupar muestras o mejorar la caracterización de los grupos conocidos y;
- predecir fenotipos [6, 97].

Los dos últimos objetivos pueden ordenarse de manera sucesiva, primero encuentro muestras que se agrupan, luego predigo lo que pasará con las nuevas muestras que se integren a los

grupos. Un subproducto de tal sucesión sería la identificación de biomarcadores, que permitan reconocer la pertenencia de una muestra a un grupo. Estos dos objetivos además empatan con problemas de aprendizaje estadístico conocidos, como *clustering*, clasificación y regresión; es decir que se pueden abordar tanto con perspectivas supervisadas -si hay una respuesta a predecir-, como no supervisadas -si en vez de centrarse en una respuesta se necesita analizar las relaciones entre variables u observaciones-. Por su parte, el descubrimiento de mecanismos moleculares, que idealmente seguiría a la identificación de marcadores, descansa en la inferencia de redes, y requiere, en gran medida, la generación de datos de validación.

Usando distintos nombres, la integración multi-ómica se ha dividido de acuerdo el momento de integración y al objeto a integrar. Se habla de integración vertical o integración-N cuando se incorporan ómicas diferentes referidas a las mismas muestras, es decir, al empleo de observaciones co-ocurrentes de niveles funcionales diferentes. Este es el tipo de integración al que aspira este trabajo y en el que se concentrará la presente sección. Por el contrario, la integración horizontal o integración-P agrega estudios de un mismo nivel molecular, hechos sobre sujetos diferentes, para aumentar el tamaño de muestra [98, 99].

Además, se habla de integración temprana y tardía de acuerdo al momento de ejecución. La integración temprana se refiere a la concatenación de mediciones obtenidas con ómicas diferentes desde el principio, antes de cualquier análisis de clasificación o regresión, lo que desdeña la heterogeneidad entre plataformas. Por otro lado, la integración tardía combina múltiples modelos predictivos, obtenidos por separado para cada ómica, ignorando las interacciones entre niveles y la posibilidad de sinergia o antagonismo [99]; este es el primer tipo de integración multi-ómica que se produjo y a pesar de entregar resultados sumamente útiles, se ha abandonado por otras aproximaciones. Aunque menos discutido, también se ha propuesto un abordaje intermedio, en que se modela un sólo conjunto de datos, después de la transformación de las ómicas a través de análisis por separado, lo que respeta la diversidad de plataformas, sin necesariamente capturar las interacciones entre niveles funcionales [100]. La tabla 2 resume distintas herramientas de integración multiómica en términos de los criterios de clasificación mencionados.

Aparte del problema de compatibilidad entre plataformas, la integración multi-ómica afronta retos en cuanto al ruido, que aumenta con la cantidad de variables aportadas por cada ómica; a la dimensionalidad, puesto que la cantidad de variables siempre sobrepasa el tamaño de la muestra y la interpretabilidad del modelo final, que se vuelve más difícil mientras más variables se tengan [6, 101]. Para resolver la cuestión de la compatibilidad se emplean distintas normalizaciones, posteriormente al pre-procesamiento independiente y de acuerdo a los requerimientos de cada

plataforma. El método de normalización requerido por las mayoría de las herramientas es la estandarización de los datos concatenados, esto es, llevar todos los valores a media de cero y varianza uno, sin importar la ómica de procedencia. Cuando el número de variables y ruido difiere entre plataformas, se recomienda la normalización del análisis de factorización de matrices (MFA), que divide el bloque de datos de cada ómica entre la raíz cuadrada del primer valor propio, de esta manera, todas las plataformas tienen el mismo peso en el análisis. Para evitar que el bloque más grande domine el análisis, también se ha empleado la división de cada bloque por la raíz cuadrada del número de variables o la varianza total [102, 103] e incluso se ha planteado un algoritmo para detectar el método de normalización óptimo [104].

Los problemas de dimensionalidad e interpretabilidad pueden afrontarse de una sola vez con la aplicación de los métodos multivariados escuetos. Los métodos multivariados estudian de manera simultánea múltiples variables, formando parte de la integración temprana. Para volverlos escuetos se agrega una penalización a la función de ajuste, que contrae los coeficientes de modo que algunas variables terminan con coeficiente cero y salen del modelo, mejorando así su interpretabilidad, al mismo tiempo que permite el ajuste a pesar del exceso de dimensiones. Estos métodos además emplean la descomposición de las matrices de datos, en particular la descomposición en valores singulares [105], produciendo herramientas estadísticas bien fundamentadas, veloces y, como se describe a continuación, listas para su aplicación sobre las interrogantes del cáncer.

A su vez, la construcción de redes exhibe las interacciones entre pares de entidades, normalmente sin restricción en cuanto al origen de las mismas, permitiendo la integración de cualquier conjunto de ómicas. El foco sobre un subconjunto de interacciones interpretable se pone, ya sea haciendo uso de redes funcionales conocidas a priori como las vías metabólicas o de señalización o, mediante umbrales de significancia o conectividad. En este caso, el problema del exceso de variables respecto al tamaño de la muestra depende en gran medida del método de inferencia, aunque se han propuesto herramientas para identificar el tamaño de muestra óptimo de acuerdo al error de clasificación aceptable [106].

3.2. Integración tardía

El análisis de *cluster-of-clusters*, CoCA, es quizá el método de integración tardía que más impacto ha tenido, al ser la herramienta base del TCGA. Se trata de un algoritmo de *clustering* consenso a partir de los grupos identificados por separado en cada ómica. Aunque fue introducido precisamente con los datos de cáncer de mama [1], ya en el 2018 se optó por un árbol de decisión

para agrupar los tumores ginecológicos [2], pues CoCA los agrupaba por tipo de cáncer. A pesar de desestimar la relación entre ómicas, esta herramienta identificó los patrones que permiten elucubrar respecto al efecto que tiene una ómica sobre la otra. Por ejemplo, fue en el artículo del 2018 que se describió la coincidencia entre el subtipo basal, la mayor hipometilación del DNA y una inestabilidad genómica elevada, que muy bien podría ser consecuencia de la metilación alterada de los transposones, pero demostrarlo requiere más datos.

Una mejora que podría calificar como integración intermedia, es el *clustering* omica-específico, pero compartiendo las etiquetas de grupo entre ómicas, a través de un modelo bayesiano [107, 108]. Este enfoque es interesante porque reporta baja correlación entre los *clusters* encontrados por metilación, expresión y número de copias (CNA), concluyendo que un agrupamiento único de todos los pacientes no da cuenta de la heterogeneidad y en cambio puede ser engañoso. Resalta la correlación relativamente alta -y esperada- entre CNA y expresión para el subtipo HER2E y, en el sentido contrario, entre metilación y CNA para el basal. Este último resultado se explica con la dificultad de agrupar las muestras del subtipo basal, debido a su inestabilidad genómica [108].

Dejando los algoritmos de agrupamiento, *ActivePathways* agrega la significancia por gen de distintos análisis, a partir de ómicas diferentes y hace un análisis de enriquecimiento funcional, tanto de la lista integrada, como de las significancias separadas, determinando así qué enriquecimiento depende de qué evidencia. Al aplicar esta herramienta a los datos de CNA y expresión genética de METABRIC, los autores son capaces de identificar vías cuyo enriquecimiento depende de la integración de datos, como la regulación negativa de procesos apoptóticos en el subtipo HER2E, y sugerir marcadores [109]

Entonces, como puede verse, las herramientas de integración tardía retoman resultados independientes y no consideran las relaciones entre ómicas, sin que esto impida de alguna manera la satisfacción de los objetivos de clasificación o enriquecimiento funcional. Estrictamente, el único objetivo de la integración multi-ómica que demanda integración temprana es el de descubrir mecanismos moleculares multi-ómicos. Aún así, existen ejemplos de integración temprana con todo tipo de fines.

3.3. Integración temprana

En oposición a la adición de resultados independientes, la concatenación de los datos desde el inicio permite observar efectos conjuntos. Sin embargo, la concatenación también produce matrices de datos muy grandes, que por regla general, requieren de estrategias de reducción de

la dimensionalidad para su análisis. La reducción de la dimensionalidad consiste en construir un set reducido, q, de variables nuevas, mediante combinación lineal de las variables originales, p. Las nuevas variables son denominadas ejes principales, vectores propios o variables, componentes o factores latentes y están formadas por p coeficientes (*loadings*), con al menos un valor distinto de cero. No hay una manera estándar de elegir el tamaño de q, sino que se suele elegir de acuerdo al punto donde la varianza explicada por cada factor latente se estabiliza. Para encontrar estos componentes latentes se busca maximizar la varianza que cada uno representa, pero manteniendo ortogonalidad entre ellos, es decir, que capturen la mayor cantidad de información diferente [110].

La estrategía de reducción de la dimensionalidad más común es el análisis de componentes principales (PCA), para representar datos complejos en el plano. Aunque existen tantos sabores como cualidades de los datos, como el análisis de correspondencia, la factorización no negativa de matrices (NMF), el análisis de correlación canónica (CCA), la regresión de mínimos cuadrados parciales (PLS) o el análisis de co-inercia (CIA) [110]. Sobre el PCA se construyó el análisis de factores múltiples (MFA), especialmente útil para la integración multi-ómica, al considerar la estructura de los datos, normalizando cada bloque de variables (ómica) con su primer valor propio y obteniendo entonces los componentes principales de la matriz concatenada completa. La normalización del MFA intenta que todas las ómicas tengan un peso en los resultados relativo a su varianza -y no a su tamaño - y solventa el problema de la integración temprana con la heterogeneidad entre plataformas [111].

Multi-omics Gene Set Analysis (MOGSA) es una herramienta de enriquecimiento funcional especialmente pensada para datos de célula única, pero que ejemplifica muy bien lo que se persigue con la integración temprana. Para estimar el enriquecimiento, MOGSA proyecta los conjuntos diana sobre los ejes del MFA, generando un valor de enriquecimiento para cada uno. Un valor de enriquecimiento alto implica variables que explican una gran proporción de la información global, ya sea de uno o varios bloques de datos, pero siempre descartando efectos exclusivos de un bloque, como podría ser el efecto de lote asociado a una plataforma. De esta manera es posible encontrar el enriquecimiento funcional sobre ómicas dispares, sin mapear las distintas variables a los mismos genes, como sería preciso para la mencionada *ActivePathways*. Tampoco se necesitan valores de significancia por ómica, eliminando la comparación entre grupos que requeriría un análisis de expresión diferencial. En este caso sólo se necesita conocer las membresía de cada variable, de cada bloque, en los conjuntos diana. Además, la descomposición permite excluir factores que no sean de interés, mejorando la interpretabilidad [112].

Otra herramienta que parte del PCA es *Joint and Individual Variation Explained* (JIVE), que

descompone la matriz concatenada en submatrices -de menor rango- de varianza compartida, varianza individual y ruido. Además de reducir la dimensionalidad, JIVE permite la exploración visual de la matriz de varianza compartida y a partir de eso, la identificación de potenciales biomarcadores [113]. Al analizar datos de expresión de miRNAs y transcritos de glioblastoma, JIVE encuentra mayor varianza individual que compartida entre las ómicas y mayor en los datos de transcritos que de miRNAs. A su vez, la estructura compartida recupera más varianza de los miRNAs, que de los transcritos. Puesto que las estructuras son ortogonales, la información de la matriz de varianza compartida, donde pueden apreciarse los subtipos de cáncer, no está relacionada con las matrices individuales. Puesto que originalmente los subtipos se identificaron agrupando expresión genética, el peso de los miRNAs fue un resultado inesperado. Al examinar los coeficientes sobre la estructura compartida, los autores pueden identificar tanto transcritos como miRNAs con roles en la enfermedad [103].

Además del enriquecimiento funcional y la búsqueda de estructuras compartidas, las descomposición de matrices se puede usar para agrupar tumores, como hicieron Cantini y compañía con distintas herramientas, con supuestos diferentes. Trabajando con datos simulados, intNMF e iCluster identificaron los mejores agrupamientos, siendo iCluster peor, pero permitiendo todo tipo de datos y distribuciones, sin la restricción de valores negativos del NMF y encontrando un componente latente común entre ómicas [**?**]. Trabajando con datos del TCGA, MCIA, RGCCA y JIVE son los mejores para encontrar factores asociados a características clínicas o a funciones biológicas. El análisis regularizado de correlación canónica generalizada (RGCCA), que considera factores latentes distintos por ómica y no compartidos como iCluster, tiene el mejor desempeño en los datos de cáncer de mama. Se sugiere que esta identificación de factores latentes por ómica no sólo permite encontrar procesos biológicos compartidos, sino procesos que son complementarios entre los bloques de datos [114].

El RGCCA también ha sido usado para identificar metabolitos con un potencial como marcadores de cáncer hepático o tejidos cirróticos. *Multi-Omic inTegrative Analysisf* (MOTA) es una herramienta que utiliza el RGCCA para estimar eficientemente la correlación entre elementos de ómicas diferentes. Bajo la premisa de que los fenotipos no sólo divergen en la abundancia de las moléculas, sino también en la manera en que éstas se conectan, MOTA estima correlaciones diferenciales con los resultados del RGCCA y sólo conecta los pares de elementos con valores sobre un umbral pre-especificado. De está manera, MOTA produce una red escueta y un puntaje por nodo. Este puntaje refleja tanto la conectividad del nodo, como su expresión diferencial y facilita la selección de nodos de interés de entre aquellos con mayor puntaje. En su

comparación de tumores hepáticos y tejidos cirróticos, los autores encuentran que el top 30 de transcritos con mejor puntaje está enriquecido de genes del cáncer y procesos relacionados con el cáncer hepático, lo que no sucede con otras herramientas equivalentes [115]. Así, se suma a la lista de objetivos alcanzados con técnicas de integración temprana, la identificación de posibles marcadores.

El RGCCA es generalizado por la capacidad de analizar más de dos bloques de variables simultáneamente y regularizado, por incluir una penalización sobre los coeficientes [116]. Este no es un método escueto, porque la penalización que emplea, conocida como *ridge*, no lleva a cero ningún coeficiente, sino que simplemente los acerca a este valor. Esta penalización permite reducir la varianza del ajuste y arroja coeficientes cuyo valor absoluto sirve para discernir las variables más relevantes en los componentes. Sin embargo, el punto de corte sobre "lo más relevante" siempre es arbitrario. En cambio, las penalizaciones LASSO y ENET superan esta dificultad contrayendo algunos coeficientes hasta cero, en métodos escuetos que tienen como producto la selección automática de variables [105].

3.3.1. LASSO

La penalización *Least Absolute Shrinkage and Selection Operator*, LASSO, ha sido añadida a distintas metodologías estadísticas, generando la versión escueta de análisis conocidos como PCA, JIVE, PLS y RGCCA. Su capacidad para la selección de variables ha sido explotada para sugerir posibles reguladores de la expresión genética, biomarcadores y, como se desea hacer aquí, plantear relaciones entre distintos niveles funcionales.

La diferencia entre las penalizaciones LASSO y *ridge*, es el objeto sobre el cual trabajan. Mientras la penalización *ridge* escala la norma dos (l2) del vector de coeficientes; LASSO transforma la norma uno (l1), usando el mismo parámetro lambda con valores entre 0 y 1 que se elige por validación cruzada (CV). Por su parte, la diferencia entre dichas normas es una potencia, ya que la norma uno de un vector se define como la suma de los valores absolutos; mientras que la norma dos involucra la suma de los cuadrados [105]. Fuera de está pequeña diferencia en la definición, los métodos de análisis estadístico mencionados se mantienen en cuanto a objetivos y supuestos.

En el caso del SGCCA, donde la 'S' significa sparse o escueto, se conserva el objetivo del RGCCA de extraer la información compartida entre bloques de datos, para cada uno de los mismos, es decir, no se obtiene una sola estructura compartida como con el JIVE, sino que

se genera un componente latente para cada bloque. Este componente de bloque resume la varianza de sus propios datos, al mismo tiempo que se correlaciona con otros bloques. Los bloques correlacionados dependen de un parámetro C. Además del cambio de regularizado por escueto, el SGCCA tiene como principal característica la maximización de la covarianza. Mientras el RGCCA puede maximizar covarianza, correlación o un compromiso entre ambas, el SGCCA sólo optimiza la covarianza, teniendo como primer prioridad encontrar componentes de bloque que contengan la mayor varianza posible y, como segunda prioridad, que recuperen la correlación con los componentes vecinos. Así, el SGCCA está definido con el problema de optimización:

$$max_{w_1,...,w_j} \sum_{j,k=1}^{J} c_{jk} g(cov(X_j w_j, X_k w_k)) \text{ s.t. } \begin{cases} ||w_j||_2 = 1 \\ ||w_j||_1 \leq s_j \end{cases} , \quad j = 1, ..., J$$

Donde:

$X_1, ..., X_j =$ bloques de datos
$w_j =$ vectores de coeficientes
$X_j w_j =$ componentes de bloque
$g =$ función convexa continua, que permite optimizar criterios diferentes, como:
- la función identidad, que simplemente optimiza la covarianza
- la función de *Horst*, que penaliza la correlación negativa entre bloques
- la función centroide que permite correlaciones negativas

$C =$ matriz cuadrada del tamaño del número de bloques, con valores 1 y 0, dependiendo si los bloques están conectados o no
$s_j =$ constante positiva que determina el grado de penalización, es decir, qué tan escueto es w [102, 116]

La matriz C permite examinar distintas relaciones entre niveles funcionales. En la publicación original del SGCCA se estudiaron tres diseños de C respecto a la relación entre expresión genética, imbalance cromosómico y la localización o el subtipo de los gliomas pediátricos. El diseño con los tres bloques conectados busca alteraciones simultáneas de los dos niveles funcionales respecto a los subtipos. Cuando el punto de unión es la localización, el objetivo son las alteraciones asociadas a los subtipos, independientemente de la relación entre expresión e imbalance. Finalmente, poner la expresión como el puente entre los otros dos bloques habla de imbalances que afectan la expresión y que a su vez repercuten en el subtipo. El segundo diseño tuvo los mejores valores predictivos, confirmando que en ese caso el fin era determinar los subtipos, aunque el bloque de expresión arroja más información discriminativa que el de imbalances. También se reporta la selección de una cantidad menor de variables que métodos equiparables y baja sensibilidad respecto a la función g [116].

Así como PCA y PLS pueden derivarse como casos especiales del RGCCA [102], la versión escueta de PLS, sPLS, descansa sobre el SGCCA. En particular, en el paquete mixOmics de R se han implementado instancias del SGCCA de Tenenhaus, específicamente pensadas tanto para la integración vertical, como para la integración horizontal y, de manera especial, extensiones supervisadas dedicadas a la clasificación y predicción. Por usabilidad, mixOmics reemplaza el parámetro de penalización por el número de elementos a recuperar de cada dimensión y facilita su ajuste con funciones sacadas directamente de la investigación estadística. Al mismo tiempo, el paquete captura medidas de error y de la estabilidad de las variables seleccionadas, permitiendo su visualización eficiente. Al examinar sus propias herramientas, los autores reportan mayor capacidad discriminatoria de los subtipos del cáncer de mama en la integración de transcriptoma y proteoma, que empleando los datos de expresión de miRNAs, que sin embargo, están fuertemente correlacionados con el transcriptoma [99, 117, 118].

Las ventajas del sPLS incluyen mayor estabilidad ante datos colineales -lo que es frecuente en las ómicas- que otros tipos de regresión o que el mismo CCA; varianza explicada similar entre distintos niveles de penalización; ortogonalidad de los factores latentes del mismo bloque, pero no de los factores del bloques diferentes -que surgen de la maximización de la covarianza- [99]; desempeño superior a herramientas de clasificación como bosques aleatorios y centroides más cercanos y; 4 modos diferentes de descomponer la matriz de datos, con finalidades diferentes. El modo regresión, o PLS2, tiene como objetivo explicar Y a partir de X, por lo que la descomposición es asimétrica y los factores latentes obtenidos no serán los mismos que al predecir X con Y. El modo clásico es idéntico al modo de regresión y también se identifica como PLS2, pero usa una normalización diferente, lo que produce coeficientes de Y diferentes. El modo canónico en cambio es simétrico, pues su objetivo es modelar las relaciones entre bloques, sin asumir un sentido en las mismas. Al ser un modo exploratorio, no puede someterse a los mismos criterios de ajuste que las alternativas supervisadas. Finalmente, el modo invariante también sería asimétrico pues no descompone la respuesta sino que hace una análisis de redundancia de X respecto a Y [119, 120].

Sin necesariamente usar el ajuste por mínimos cuadrados parciales, los análisis basados en regresión con penalización LASSO han alcanzado el objetivo hasta ahora ignorado de la identificación de elementos regulatorios potenciales. Un ejemplo es el uso que hace miRDriver para sugerir miRNAs que extienden el efecto de una alteración en el número de copias a sus dianas en trans. Utilizando los datos de cáncer de mama del TCGA, puede verse que miRDriver tiende a seleccionar miRNAs relacionados con cáncer y con valor prognótico, como miR- 1224, miR-31,

let-7g y let-7b [29]. Otro ejemplo es la integración asimétrica de metilación del DNA, CNA, miRNAs y sitios de unión de TFs, con los genes diferencialmente expresados de los subtipos del cáncer de mama. Comparando la capacidad de predicción de las ómicas por separado y en conjunto, los modelos integrados tuvieron mejor desempeño, con un aumento significativo al agregar los datos de metilación. Aunque hay reguladores compartidos entre los cuatro subtipos, que incluyen a E2F1 y CITED2, no se desarrolla sobre el posible mecanismo de regulación, sino que se comprueba la capacidad de estratificación de una firma transcripcional con los reguladores del subtipo basal [121].

Yendo más lejos, se han planteado posibles mecanismos regulatorios a partir de este tipo de regresiones. Setty et al. modelaron la expresión diferencial entre glioblastomas y tejido normal , con una regresión LASSO que integra la metilación del promotor y el número de copias promedio, con la cuenta de TFs y miRNAs que pueden unirse a la región regulatoria. El número de copias fue seleccionado por el LASSO en todos los casos, mientras que los coeficientes de metilación se mantuvieron siempre como grandes valores negativos. La variabilidad de los coeficientes que corresponden a miRNAs puede explicarse por una baja correlación con la expresión del gen diana, debido a la acción simultánea de otros reguladores o, como se ha sugerido antes, a que los miRNAs sólo afectan la expresión génica de manera modesta. Cuando se buscaron los predictores clave por subtipo y gen, mediante un análisis de dependencia, se observó que REST, un represor de genes neuronales en células no neuronales, es clave en los distintos subtipos. Aunque las interacciones que se muestran en la figura 5a provienen de la literatura, el modelo regulatorio depende de la selección de REST y miR-124 en las regresiones de todos subtipos, y la selección de YY1 y miR-132 en uno de ellos [122].

De manera similar, Li et al. integraron CNA, metilación y microRNAs, para predecir los datos de expresión de cáncer de ovario del TCGA. La descomposición de las matrices arroja factores de, en promedio, 45 CNA loci, 43 sitios CpG, 5 miRNAs y 44 genes. Mientras la correlación entre los eventos de CNA y los niveles de expresión se mantiene positiva, la correlación con la metilación varía. Al derivar redes de correlación de los elementos co-seleccionados, los autores recalcan que los elementos se mantendrían aislados si no fuera por la integración. Aunque estas redes no necesariamente reflejan relaciones causales, si pueden servir como punto de partida para estudiar los mecanismos subyacentes. La figura 5b muestra la red del EGR1. En este caso tanto nodos como interacciones surgen del análisis, y la literatura sólo verifica las conexiones EGR1-PITX2 y WT1-AMH [123]

Acoplando el sPLS de mixOmics y MOGSA, Chapell y compañía no sugieren un posible me-

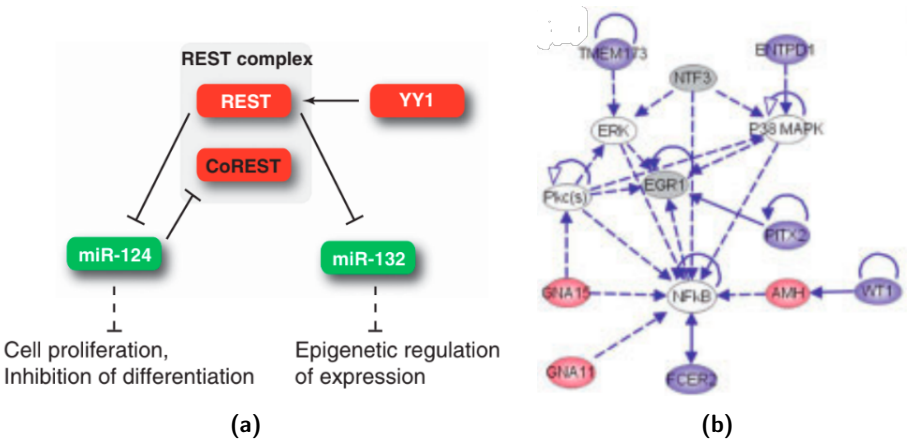


Figura 5: Mecanismos regulatorios propuestos a partir de aproximaciones escuetas a la integración multi-ómicas. (a) Modelo regulatorio del represor REST, propuesto por Setty et al. [122]. (b) Modelo regulatorio de EGR1, propuesto por Li et al. [123]

canismo multi-ómico, sino un eje regulatorio que sería deseable intervenir. Integrando datos de líneas celulares de cáncer de mama basal, que abarcan metilación del DNA, expresión genética, expresión de proteínas, fosfoproteómica y modificación de histonas, se recuperó a TGFB1, TGFBR2, KLF6, KLF12, PIK3R3, VIM, NES, RASL11B, HOXC9, LAMB3, PRKCD, PRKCE y MELK en los dos primeros componentes latentes. Al revisar más a fondo, los autores encuentran que la señalización por TGFb efectivamente difiere entre líneas con y sin mutaciones de *BRCA1* y que su enriquecimiento depende de la fosforilación de SMAD5. Puesto que en células tumorales el receptor estimulado de TGFb fosforila a SMAD1/5 y promueve la migración celular, se concluye que el eje TGFBR1 – SMAD1/5 podría ser de interés clínico [124].

Sin enfocarse en algún mecanismo, también se ha aprovechado la integración para construir redes multi-ómicas, sobre las que es posible extraer conclusiones generales respecto a la importancia de los distintos niveles funcionales en el cáncer. Sohn et al construyeron redes de CpGs, miRNAs y CNA con los coeficientes de un modelo predictivo de la expresión genética en cáncer de ovario. El modelo integrado predice la expresión mejor que los modelos de una sola ómica. Los patrones de los genes altamente expresados tienen como predictor predilecto la alteración del número de copias; mientras que los genes con mayor variabilidad se explican mejor con la metilación, lo que indicaría un efecto dinámico de la regulación por metilación del DNA, que cuenta cada vez con más evidencia [125]. En cuanto a las redes, la red del modelo integrado tiene mayor modularidad y enriquece para funciones más específicas que las redes de una sola ómica. Las aristas con mayor peso (coeficiente) tienden a involucrar sitios de metilación y conforme baja el peso aumenta la proporción de CNA, mientras que los miRNAs se mantienen estables [126].

Aunque sólo integran datos de metilación del DNA y expresión de cáncer de mama, Lee et al. construyeron redes con los coeficientes de regresión LASSO; implementado a la par un *kernel* pesado para compartir información entre todas las muestras, sin dejar de obtener coeficientes específicos de cada subtipo. El uso del *kernel* mejora las predicciones de los genes ligados a algún subtipo y especialmente al enriquecido de HER2, que sólo cuenta con 16 muestras. Para cada gen diana sólo consideran los sitios CpG de las vías en las que el gen participa y el modelo selecciona entre 200 y 300 sitios, que forman redes con 88.82 % de las aristas en común entre subtipos. Los CpGs más conectados participan en la progresión del cáncer, como sucede con LEP y FGFR3 en el subtipo luminal B. Los genes mejor predichos codifican GTPasas, factores de la transcripción y proteínas de unión al DNA. Sumando las tasas de error de los genes en una vía, es posible estimar el impacto de la metilación en la vía, lo que deja en el top de predicción de los cuatro subtipos vías del metabolismo de carbohidratos como glicólisis/gluconeogénesis, la ruta de las pentosas fosfato y el metabolismo de fructosa y manosa [127].

Finalmente, la capacidad de selección del LASSO también se ha usado para filtrar dianas terapéuticas potenciales de entre múltiples elementos genómicos. Esto fue posible ajustando un modelo Cox multivariado y escueto de los datos de expresión genética, miRNAs, metilación del DNA y alteración del número de copias de cáncer de ovario. La firma resultante contiene 156 elementos y predice mejor la supervivencia libre de progresión que firmas más grandes pero basadas en una sola ómica. La firma integrada está enriquecida de genes ligados a la respuesta inmune y el metabolismo. Al ordenar los elementos de la firma por su capacidad de estratificar pacientes es posible filtrar aún más la lista de biomarcadores potenciales [128].

Aunque la sola reducción de la dimensionalidad puede bastar para predecir fenotipos y agrupar muestras, la penalización LASSO facilita la identificación de reguladores potenciales y la construcción de redes escuetas y con ello, abre la posibilidad de cumplir la promesa de la integración temprana sobre el descubrimiento de mecanismos multi-ómicos. El gran problema de la penalización LASSO es la inestabilidad de las variables seleccionadas. Éste es un problema inherente y cuenta con un paliativo bien establecido, basado en la frecuencia con que las variables son seleccionadas en subconjuntos aleatorios de los datos. Las frecuencias pueden resumirse con el puntaje de Fleiss, que refleja la concordancia entre subconjuntos y toma valores más grandes mientras mayor sea la estabilidad [116]. También se puede simplemente elegir un punto de corte sobre la frecuencia de selección [29, 129], lo que sin embargo suma arbitrariedad. Sea cual sea la estrategia elegida, las variables altamente correlacionadas tienen mayor probabilidad de ser seleccionadas en cada subconjunto [120].

De manera adicional, la penalización LASSO ha mostrado limitaciones cuando hay grupos de variables fuertemente correlacionadas dentro de un mismo bloque, caso en el cual, LASSO tiende a elegir una sola variable. En las situaciones donde hay más muestras que variables, la penalización *ridge* tiene mejor desempeño. En las situaciones donde hay más variables que muestras, como sucede con las ómicas, LASSO sólo puede seleccionar tantas variables como muestras haya. Un método ideal debería eliminar las variables triviales e incluir grupos completos ya que una de las variables sea seleccionada. Con esto en mente se planteó una penalización intermedia a ridge y LASSO, que es la red elástica [130].

3.3.2. ENET

La red elástica, ENET por sus siglas en inglés, combina las dos penalizaciones, *ridge* y LASSO, para obtener un método que contrae los coeficientes de regresión hasta cero; pero también filtra grupos completos de variables correlacionadas. Debido a la selección de grupos, la red elástica selecciona más variables que el LASSO y frecuentemente alcanza predicciones más precisas, especialmente en presencia de colinealidad.

En su derivación más simple la penalización de red elástica consiste en la suma de las penalizaciones *ridge* y LASSO, por lo que implica dos parámetros, lambda 2 que actúa sobre l2 y lambda 1 para escalar l1. Para simplificar, los parámetros lambda son sustituidos por una valor alpha, que convierte la penalización en:

$$(1-\alpha)|\beta_1| + \alpha|\beta|^2 \leq t; \qquad \alpha = \frac{\lambda_2}{\lambda_2 + \lambda_1}$$

De esta manera, cuando el parámetro de mezcla, α, vale 1, la red elástica se convierte en la penalización *ridge* y cuando $\alpha = 0$ en el LASSO. Entonces la red elástica existe entre 0 y 1, aunque la evidencia indica que el mejor desempeño se logra con valores muy cercanos a cualquiera de los dos extremos. Al tratarse de una función estrictamente convexa para cualquier valor mayor de 0, los predictores idénticos reciben el mismo coeficiente, lo que garantiza el agrupamiento de las variables fuertemente correlacionadas. La similitud con el LASSO, permite un cómputo eficiente y su aplicación para resolver tanto problemas de regresión, como de clasificación [130].

Así, el parámetro de mezcla controla qué tan escueto es el modelo y se elige por validación cruzada. Tal y como pasa con las otras penalizaciones, la CV basada en conjuntos pequeños de datos, menores a 100 muestras, produce un exceso de varianza que puede ser problemático. De manera similar, los valores de los coeficientes no tienen relación con las medidas originales, lo que

dificulta su interpretación [131]. Finalmente, en ambas penalizaciones las variables carecen de una medida de significancia que soporte su selección [132]. La diferencia entre LASSO y ENET radica en una ventaja de las redes elásticas cuando hay un exceso de predictores relevantes o cuando hay grupos de variables correlacionadas. De otra manera, cuando los modelos son altamente escuetos, el desempeño de ambas penalizaciones es equivalente [133].

La superioridad de la red elástica aumenta con el número de muestras y el nivel de correlación; pero decae después de un lapso cuando lo que incrementa es el número de predictores [133]. Contraintuitivamente, los falsos positivos también aumentan mientras más muestras, menos predictores y mayor correlación, lo que se controla elevando el valor de alpha. Es con base en esto que se recomienda el valor de 0.5 para el parámetro de mezcla, pues controla el error de tipo I y tiende a filtrar grupos completos de variables correlacionadas [131].

En lugar de elegir una penalización sobre la otra, se ha propuesto su uso conjunto para la identificación robusta de marcadores. A partir de tres cohortes con datos de lipidómica, la integración horizontal con las dos penalizaciones arroja un marcador y una clasificación de los sujetos a partir de ese marcador, que supera las diferencias entre cohortes. Probando el mismo esquema conjunto con datos de expresión de adenocarcinoma pancreático y, por separado, de leucemia mieloide, se repite la recuperación de genes altamente discriminativos, que, al conectarse en una red de interacciones, modelan las diferencias entre cáncer y tejido normal [129].

Además del uso conjunto, Pineda y compañía propusieron una estrategia de permutación con el fin de evaluar la significancia de la selección. Las dos penalizaciones se usaron para predecir la expresión del cáncer de vejiga a partir de variantes genéticas, metilación o una combinación de ambas ómicas. El resultado del LASSO es una selección de 9 genes significativamente explicados por variantes genéticas, 19 por CpGs y 23 por el modelo combinado. Contra lo esperado, la selección de la red elástica es menor, con 11, 6 y 4 genes, respectivamente. Aunque la intersección entre penalizaciones es pequeña, se toma como evidencia adicional del modelo. Llamativamente, los genes seleccionados tanto por LASSO, como por ENET, alcanzan valores p similares [132]

Otra innovación es el ajuste de penalizaciones diferentes por ómica, que implementan Liu et al. para clasificar muestras de leucemia mieloide aguda y, de manera independiente, de adenocarcinoma prostático, con base en la integración de datos de expresión genética y metilación del DNA. La idea es que forzar una contracción uniforme sobre ómicas con tamaños diferentes y efectos de magnitud distinta, puede castigar injustamente variables importantes pero más sutiles que todo el bloque de la otra ómica. Entonces, el ajuste de penalizaciones diferentes aborda el problema de la integración temprana con la compatibilidad entre plataformas. Aunque esto suma

un parámetro a ajustar por ómica, lo que exige mayor tiempo de cómputo, no cambia el método de fondo, pues los paquetes de programación que contienen la red elástica, como glmnet y caret, normalmente ya incluyen esta opción. Así, los autores fijan alpha en 0.5 y optimizan un valor de contracción lambda y un radio de contracción kappa, que finalmente mejoran la predictibilidad del modelo. Los datos simulados indican que mientras mayor sea el contraste de número de variables entre los bloques, mayor debe ser la penalización diferencial, cargando la contracción de coeficientes sobre la ómica con más ruido. La clasificación de AML mejora cuando la penalización afecta menos los datos de metilación, pues los subtipos caracterizados por aberraciones cromosomales no pueden identificarse sólo con la expresión. Por el contrario, la clasificación de los tumores de próstata no mejora con la penalización diferencial, sino que el kappa óptimo está ligeramente por encima de uno y selecciona un exceso de transcritos, en su mayoría ya ligados al cáncer [134].

Debido a su origen, la red elástica comparte dificultades con el LASSO. A saber, la parametrización por validación cruzada, la interpretabilidad de los coeficientes y su falta de significancia. Afortunadamente cada cuestión tiene soluciones, sino totales, bien establecidas, como el uso alternativo de *k-folds*, *leave-one-out* o *bootstrapping* en el caso de la elección de parámetros o la obtención de valores empíricos, tal y como hicieron Pineda et al. [132], para afrontar la carencia de valores de significancia. Aunque los coeficientes no son comparables con los valores originales de las ómicas, Huang et al. [121] muestran como los coeficientes reflejan la fuerza de la asociación entre predictores e incluso le dan un signo. Como herramientas de integración temprana, ambos métodos escuetos multivariados se enfrentan a las diferencias entre plataformas, ya sea con normalizaciones que controlan el peso que tiene cada ómica [111] o variando la penalización a la que son sometidas, como demostraron Liu y compañía [134].

Finalmente, al depender de los métodos de reducción de dimensiones, heredan las cuestiones de la descomposición de matrices. Puesto que los componentes latentes son combinaciones lineales de las variables originales, sólo se pueden capturar relaciones con un efecto lineal extendido sobre distintos predictores, cuando probablemente el efecto de las alteraciones no se concentre en un sólo factor [99, 135]. Adicionalmente, el número de componentes limita la información captada. Aunque los primeros componentes explican la mayor parte de la varianza y los últimos estarían ligados con ruido; los datos de cáncer comprenden señales del tejido, la exposición a mutágenos, el tratamiento y el infiltrado inmune, entre otros, lo que pueden elevar la cantidad de componentes necesaria. Aún más, los componentes de los modelos escuetos, con la mayoría de los coeficientes en cero, explican menor porcentaje de varianza [119]. Cuando se trata de

clasificar muestras, la convención es recuperar K-1 componentes, donde K representa el número de clases [120]. Sin embargo, al buscar mecanismos multi-ómicos, no hay clases para guiarse.

Se ha sugerido que la división de los bloques de datos en vías o sub-bloques funcionales, puede simplificar la elección del número de componentes, al dotarlos de una interpretación más directa [102]. Puesto que ambos métodos escuetos llegan a desembocar en redes de correlación, también es pertinente resaltar que estás redes fallan al distinguir entre relaciones directas e indirectas, lo que se podría resolver estimando la matriz de precisión [115]. Aunque la matriz de precisión no es otra cosa que la inversa de la matriz de covarianza, su obtención no es un proceso sencillo cuando hay bloques co-lineales o más predictores que observaciones [102]. Entonces, al implementar un modelo multivariado escueto es necesario sopesar las desventajas enunciadas, contra la capacidad de integrar grandes cantidades de variables sin la necesidad estricta de un filtrado previo [132].

3.4. Integración con redes

En su revisión del 2016 Bersanelli dividió los métodos integrativos de manera jerárquica, comenzando por distinguir los métodos basados en redes, de aquellos independientes de las mismas [97]. Como se ha visto, los métodos multivariados pueden generar redes multi-ómicas, pero no las necesitan para concretar la integración, sino que funcionan al nivel de la matriz de datos. Por el contrario, hay herramientas que explotan la capacidad de encontrar relaciones de dependencia entre cualquier par de variables aleatorias, convertidas en nodos [136], como la expresión de un miRNA y la metilación de un CpG.

Al avanzar por pares, las relaciones son ciegas al efecto de terceros y terminan en redes densas de interpretabilidad limitada [136]. Así, hace falta un paso extra, que ponga el foco sobre la interacciones que satisfacen cierto criterio, como hacen los algoritmos de la sección 3.4.1 o, como se describe en 3.4.2, restringirse a las interacciones con sustento en lo ya conocido.

También es posible estimar dependencias condicionales (correlaciones parciales) a través de modelos gráficos gaussianos, que obtienen la red a partir de la matriz de precisión [136]. Sin embargo, este enfoque aún está siendo explorado y sólo se tocará en el contexto de DRAGON. Puesto que las ómicas producen matrices de datos con más predictores que observaciones, la matriz de covarianza no es invertible y no sirve para calcular correlaciones parciales. DRAGON (*Determining Regulatory Associations using Graphical models on multi-Omic Networks*) introduce una penalización que contrae la matriz de covarianza y permite la obtención de una matriz

de precisión también contraída. De está manera se consigue una herramienta en proceso de evaluación, pero con resultados prometedores, como la identificación de ELF4 y ZBTB33, dos TFs co-expresados debido a su co-metilación, en los datos de cáncer de mama del TCGA [137].

Mientras la conexión de cualquier par de variables genera redes multipartitas, la integración multi-ómica también es posible mediante redes multicapa. Las redes multipartitas tienen tantos tipos de nodos como ómicas integradas y las aristas representan tanto relaciones intra como inter-ómicas. En cambio, las redes multicapa representan relaciones del mismo conjunto de nodos a través de distintos niveles funcionales, siendo entonces un compendio de las redes derivadas de cada ómica [97]. Los múltiples enlaces conectando un par de nodos en este formalismo son la base de los algoritmos de integración de la sección 3.4.3.

Además hay ejemplos de integración que no encajan en ninguna de las secciones. Tal es el caso de la división de los tipos de cáncer en aquellos caracterizados por mutaciones y los ligados al número de copias. Esta división surge del análisis de la modularidad de un grafo bipartita, de pacientes y alteraciones, y ubica al cáncer de mama en el grupo asociado con CNA. Aunque la red involucra CNA, mutaciones puntuales y metilación del DNA, efectivamente integrando ómicas diferentes; los nodos no son variables aleatorias, sino entidades aisladas que se conectan a la red de acuerdo a la presencia o ausencia. La red tampoco coincide con el formalismo multicapa, pues los nodos sólo existen en su propio nivel funcional, que sería el de alteraciones en cáncer. Sin embargo, al ignorar las relaciones entre alteraciones, que sólo se conectan a través de la co-ocurrencia en un paciente e implicar un filtrado de ómicas completas a eventos frecuentes, este trabajo califica como integración intermedia. A pesar de la estrategia relativamente simple, el enfoque de redes permite una conclusión tan potente como sugerir mecanismo oncogénicos distintos para los tumores, de tejidos diferentes, que se agrupan en torno a combinaciones específicas de eventos funcionales [138].

Otro ejemplo es la identificación de factores transcripcionales que son sensibles a la metilación del DNA en la mayoría de los tipos de cáncer. En este caso todo comienza con una red de regresión de la expresión genética, que une TFs y genes, y la integración es más una incorporación de los patrones de metilación a los atributos de los genes, para evaluar el efecto de la epigenética sobre la relación TF-gene. Comparando los distintos tipos de cáncer encontraron que solo el 0.28 % de las relaciones regulatorias mediadas por metilación aparecen en más de 4 tipos de cáncer y sin embargo, hay TFs que consistentemente son sensibles a la metilación. Estos TFs regulan más dianas, exhiben expresión diferencial y están enriquecidos en las vías de transducción de señales, en la adhesión celular y en la familia ETS [139]. Aquí los nodos sí representan variables

aleatorias, pero de la misma ómica y, la red es esencial para la interpretación conjunta de los datos de expresión y metilación, que es el objetivo final de la integración. Adicionalmente, este ejemplo tan debatible de integración multi-ómica por redes, si considera el efecto de un nivel funcional sobre el otro y se concentra en las interacciones, primero para reconocer las que dependen de la metilación del DNA y luego para contrastar los distintos subtipos.

Tan dispares como resulten, estos ejemplos resaltan la importancia de las conexiones, ya sea entre TFs y genes o entre tumores a través de alteraciones y por ello empatan con la biología de redes, esta aproximación de la biología de sistemas concentrada en la inferencia de modelos de redes sobre fenómenos biológicos y su análisis con teoría de grafos. En particular, los ejemplos aquí presentados, pertenecen al enfoque probabilístico de la biología de redes. Éste es un enfoque arriba-abajo, que construye el modelo a probar desde de datos masivos y no sólo partiendo de la información (limitada), ya depositada en las bases de datos [140]. A continuación se ahonda en este enfoque probabilístico y sus implicaciones para la integración multi-ómica.

3.4.1. Redes probabilísticas

La mayoría de los métodos de reconstrucción de redes de genes, empecemos con una sola ómica, sugieren mecanismos a partir de las relaciones entre genes. Pensemos en los culpables por asociación en las redes de interacciones [141]. La diferencia con las redes de regulación genética es que las aristas reflejan dependencias estadísticas entre los patrones de expresión. Dos genes unidos por sus patrones de expresión pueden sostener una relación funcional o simplemente de co-ocurrencia [142]. La diferencia con las redes multi-ómicas es que las variables existen en escalas diferentes y hace falta normalizarlas. Si la medida de dependencia estadística es la correlación, no basta con tener valores comparables, las distribuciones de varianza también deben ser similares [106]. Aunque la correlación es quizá la medida más accesible, hay múltiples herramientas que usan la información mutua (MI). La información mutua viene de la teoría de la información y, al tomar la información como la reducción de la incertidumbre, mide la reducción en la incertidumbre de una variable respecto a la información sobre otra variable [143].

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

$$I(X:Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = I(Y:X)$$

La información mutua aventaja a la correlación al capturar relaciones no lineales, ser insensible a la parametrización y poder estimarse velozmente, todas características deseables para la integración multi-ómica, con relaciones no lineales, diferentes escalas y gran cantidad de variables. Por otro lado, la información mutua siempre es positiva, incluso cuando se evalúa sobre patrones aleatorios, por lo que es necesario establecer umbrales empíricos [142]. Además, se trata de una medida simétrica, que no permite darle dirección a las interacciones [144].

En las llamadas redes de relevancia, se calcula la información mutua de manera pareada y se desechan todos los pares por debajo de un umbral. Aunque esta estrategía resalta los nodos funcionales, es incapaz de distinguir relaciones directas e indirectas. Así, sobre las redes de relevancia se construyó ARACNE (*Algorithm for the Reconstruction of Accurate Cellular Networks*), que encuentra dependencia estadísticas irreductibles, probablemente regulatorias, al deshacerse de las interacciones indirectas. El DPI (*Data Processing Inequality*) de ARACNE examina cada triángulo en la red de relevancia y descarta la arista con menos peso, considerando que la información mutua decrece rápidamente conforme aumenta la distancia entre nodos [142].

ARACNE ha sido explotada para la integración de datos de expresión de miRNAs y transcritos en cáncer de mama. Comparando las redes de cáncer y tejido normal adyacente, Drago-García et al. observaron una reducción de nexos entre miRNAs y transcritos en cáncer, a pesar de que hay más miRNAs, que se conectan entre ellos y son esenciales para mantener la cohesión de la red. Igualmente reportan un enriquecimiento de procesos relacionados con el sistema inmune y la adhesión celular, además de una asociación muy clara entre la familia de miR-200 y el *cluster* DKL1-DIO3, que participa en la transformación epitelio-mesenquimal. Respecto a la técnica, los autores encuentran diferencias cuantitativas en la distribución de MI entre miRNAs y transcritos, lo que puede reflejar las diferencias entre las moléculas y augura contratiempos para discernir las interacciones directas en el enfoque multi-ómico [84].

La respuesta para discriminar entre relaciones directas e indirectas puede ser la información mutua condicional. Si la información condicional entre un par de nodos, respecto a un tercero, es menor a un umbral, hay que quitar la arista que une al par [143]. De este modo, Liu y compañía construyeron redes para distintos tipos de cáncer, de acuerdo a la información mutua entre transcritos y TFs, dado el patrón de metilación del promotor y su número de copias. Luego, para examinar los efectos regulatorios, ajustaron regresiones lineales que muestran mejor predictibilidad integrando las ómicas que sólo usando los datos de expresión, genes que dependen fuertemente del número de copias y genes que dependen mayormente de la metilación. Estos últimos sobrelapan entre tipos de cáncer y se asocian con la tumorigénesis. Finalmente usaron los

promotores cuya metilación afecta la transcripción para agrupar los tipos de cáncer y examinar las curvas de supervivencia de los grupos, identificando 10 tipos de cáncer, incluyendo el cáncer de mama, donde la metilación es un determinante importante de la agresividad de los tumores. A pesar de que la aproximación arroja resultados sumamente interesantes, los autores advierten que demanda tamaños de muestra grandes y datos con suficiente varianza para encontrar redes confiables [145].

A su vez, la falta de dirección de las aristas puede resolverse pasando de buscar reguladores para cada gen, a buscar reguladores para grupos de genes co-expresados a través de condiciones diferentes. LeMoNe, más tarde LemonTree, empieza con un agrupamiento de dos vías, que une los genes co-expresados con alta probabilidad en *clusters* enriquecidos de categorías funcionales específicas [146]. Como segundo paso el algoritmo ordena una lista de posibles reguladores, de acuerdo a su capacidad para predecir la expresión del *cluster* en distintas condiciones. Aunque sólo se han probado como reguladores TFs, miRNAs y CNA por separado, no hay restricciones sobre el tipo de variables que se pueden integrar. En el caso de los miRNAs, la mayoría se asignan a un sólo grupo, pero hay miRNAs repetidos entre grupos. En el caso de CNA, LemonTree produce módulos con enriquecimientos más significativos que la herramienta dedicada CONEXIC [147, 148]. Este ejemplo propone una solución al problema de la simetría en las medidas de dependencia estadística, como MI y correlación, y logra señalar reguladores potenciales para grupos de genes asociados con funciones, pero en el proceso borra las fronteras entre reguladores e ignora sus posibles interacciones.

Los algoritmos aquí descritos infieren redes potencialmente regulatorias a partir de ómicas completas. Los grafos resultantes no son completos, porque algunas aristas no pasan los umbrales, si se habla de redes de MI o de MI condicional, o porque se restringe de inicio quién se une a quién en LeMoNe/LemonTree. Ya con los grafos, el foco del análisis puede cerrarse aún más en torno a los nodos con roles importantes en la red. Como resultado de esto hay filtro indirecto sobre los nodos, sólo los conectados aparecen, de manera que la perspectiva del análisis es guiada por los propios datos. Por el contrario, hay todo un conjunto de herramientas guiadas por las interacciones ya depositadas en las bases de datos, como se detalla a continuación.

3.4.2. Redes conocidas a priori

En lugar de depender completamente de la estadística, los datos multi-ómicos pueden organizarse sobre las interacciones ya conocidas, conservando así aristas demostradas, que de otra

manera podrían diluirse en el ruido acumulado [149]. Al simplemente retomar lo conocido, esta aproximación no sirve para proponer mecanismos multi-ómicos, porque las relaciones entre distintos niveles funcionales no son conocidas, al menos no al grado de especificidad requerido. Por otro lado, al ceñirse a la información ya asentada, lleva la discusión directo a funciones y genes, sin perderse en el mar de datos nuevos. En otras palabras, la integración guiada por redes conocidas no es la aproximación para encontrar novedades sino para darle un uso a la información ya recabada, siendo ese uso la propuesta de marcadores o firmas y blancos terapéuticos. Además, ya que está aproximación no depende necesariamente de la estadística, el tamaño de muestra pasa a segundo plano e incluso se puede hablar de una red por cada paciente.

Las aplicaciones más simples de esta aproximación sólo mapean los genes alterados a la red y registran los nodos topológicamente importantes, como hubs y cuellos de botella. APODHIN (*Analysis of Pan-omics Data in Human Interactome Network*) suma los valores de logFC de las distintas ómicas, si se trata de integración vertical, o cohortes, para la integración horizontal, y reporta las vías asociadas, además de los genes topológicamente importantes [150]. En vez de los genes alterados, Cava et al. mapean los genes con capacidad pronóstica del cáncer de mama basal. La innovación aquí es doble, pues primero extraen una subred por paciente, utilizando los genes pronósticos que además tienen una alteración del número de copias en ese paciente y luego buscan fármacos ligados a los cinco genes con mayor grado, convirtiendo un análisis sencillo en una herramienta para la medicina personalizada. En los tumores basales, la alteración simultánea de BRCA1 y TP53 se observa en la mitad de los casos, siendo ambos nodos centrales en la red, potencialmente susceptibles a varios fármacos y entonces, blancos terapéuticos deseables [151].

Las aplicaciones más sofisticadas calculan algún tipo de puntaje aprovechando la red. Tal es el caso de NetICS (*Network-based Integration of Multi-omics Data*), que busca genes mediadores, que conectan las alteraciones con los efectores, a partir de las interacciones funcionales depositadas en bases de datos, como KEGG y miRTarBase. Ya que las redes a priori tienen dirección, basta con difundir río abajo el puntaje de perturbación y río arriba el puntaje de expresión diferencial de los efectores, con un algoritmo de difusión de calor, para obtener un puntaje de mediación. Después de repetir el proceso para cada muestra, la combinación de los puntajes genera un valor global. En el cáncer de mama los genes mediadores están enriquecidos en las vías de señalización. Resaltan EP300 y TP53, cada uno río abajo de 5 genes diferentes, alterados en 50 % de las muestras y conectados a través de los efectores. La gran desventaja de NetICS es un sesgo hacia los genes altamente conectados, que por mero azar, tienen mayor probabilidad de tener genes alterados como vecinos [152].

De manera similar se han integrado datos de GWAS, eQTLs, mQTLs, ATAC-seq y la anotación del proyecto Roadmap, para proponer genes clave. El proceso empieza conectando genes con los SNPs que los afectan, directa o indirectamente, y ubicando esos genes en dos redes diferentes de interacciones entre proteínas (PPI), una general y una tejido-específica. Por cada uno de los genes afectados hay que verificar el enriquecimiento funcional y de firmas pronósticas, como MammaPrint y PAM50, entre los primeros vecinos y obtener un puntaje combinado del enriquecimiento pronóstico en las dos redes. Calculando dicho puntaje para el cáncer de mama se puede concluir que los SNPs inciden sobre vías de cáncer y de señalización, como las MAP cinasas, TGF-beta y WNT. Los 20 genes con mayor enriquecimiento pronóstico incluyen genes del cáncer conocidos y novedades como RNASEH2A, que exhibe una expresión anormal en los tumores y se asocia con alteraciones del número de copias, menor supervivencia y tumores ER- [153].

Mergeomics promete integrar cualquier conjunto de ómicas, siguiendo en gran medida los mismos pasos que en el ejemplo anterior. Primero se evalúa el enriquecimiento de cualquier conjunto de genes predefinido, en cualquier tipo de datos de asociación con enfermedad (GWAS, TWAS, EWAS) y si hay distintos tipos se calcula un meta valor de significancia. Posteriormente los conjuntos enriquecidos se proyectan sobre una red de regulación conocida, buscando subgrafos asociados a la enfermedad y específicamente *hubs*. Después de darle a los nodos un peso que refleja la confianza de las aristas adyacentes, se compara la contribución al peso del hub del conjunto enriquecido respecto al azar. Si la contribución es significativa se postula al hub como clave para la enfermedad [154].

Mientras los ejemplos descritos hasta ahora se enfocan en la identificación de genes de interés, iOmicsPASS usa las redes para descubrir firmas que distinguen entre subgrupos de muestras y que, al mismo tiempo, forman un grafo de nodos funcionales, que puede usarse para aprender sobre el fenotipo asociado. Esta implementación se enfoca en integrar transcriptoma y proteoma, normalizando la expresión genética con el número de copias. Avanzando sobre redes validadas de regulación transcripcional y PPI, la abundancia de un par de nodos conectados se convierte en el peso de la arista que los une. Si la proteína de un TF está elevada al mismo tiempo que su diana, se considera que la interacción es probable y se le da mayor peso. Luego, un algoritmo de centroides más cercanos busca subgrafos que predigan el fenotipo, que constituyen las firmas de clasificación. Las firmas obtenidas para el cáncer de mama no separan completamente al subtipo HER2E de los tumores luminales, sino que dividen al grupo HER2+ de acuerdo a un sub-grafo ligado a la replicación y reparación del DNA. Otros subgrafos interesantes son el enriquecimiento de la señalización por el receptor de estrógeno y la regulación por FOXA1 y AP1 en el subtipo

luminal B, respecto al luminal A, que en cambio exhibe sub-expresión de vías relacionadas con el ciclo celular. El subtipo basal no se define claramente al analizar las ómicas por separado, pero sí al integrarlas, resaltado la importancia de la regulación transcripcional [149].

Siguiendo con la idea de adaptar la red a priori al fenotipo estudiado, Glass et al. plantean una herramienta que combina las tres aproximaciones a la integración con redes. PANDA (*Passing Attributes between Networks for Data Assimilation*) parte de un red probabilística y dos redes a priori y encuentra una red que incorpora ambas fuentes, mediante el intercambio de aristas [155]. En realidad PANDA no es una herramienta de integración multi-ómica, sino de inferencia de redes transcripcionales basada en el formalismo multicapa. Sin embargo, se han construido sobre PANDA algoritmos para incorporar a la red transcripcional datos de miRNAs, la accesibilidad de la cromatina y ómicas en general, todo a través de la fusión de redes probabilísticas y a priori.

3.4.3. Redes multicapa

La integración por redes multicapa consiste en el colapso de las distintas capas en una sola red que agrega las distintas fuentes de información. Las capas son las redes derivadas de cada ómica bajo análisis, con la peculiaridad de incluir siempre el mismo grupo de nodos. Esto permite pesar las aristas por el nivel de evidencia y darle un soporte fuerte a cualquier posible descubrimiento. Por otro lado, la restricción en cuanto a los nodos exige una cantidad de información no siempre disponible. Revisando el algoritmo de una sóla ómica PANDA, es posible ver estas ventajas y desventajas.

PANDA busca la concordancia entre capas, refinando cada red con la información de la otra, al resaltar los aspectos de los datos que atañen a las redes a priori y los aspectos de las redes a priori que mejor reflejan los datos. Con este fin, PANDA parte de: 1) una red inferida de los datos, que normalmente es de correlación y aporta la probabilidad inicial de co-regulación; 2) una red de regulación, conocida a priori y formada exclusivamente por nodos incluidos en los datos y; 3) una red de interacciones entre factores transcripcionales, con las mismas restricciones que la red de regulación. A grandes rasgos, la idea es combinar las redes de manera iterativa, cada vez con cambios pequeños regidos con un parámetro de actualización y calculando un puntaje de concordancia, cuya convergencia da la señal de paro. La concordancia considera la evidencia acumulada de que el gen j sea regulado por el TF i, dado la cooperación (interacción) entre i y otros TFs que regulan a j, y la co-expresión de j con otros genes regulados por i. En un inicio el peso de todas las aristas se normaliza a z-scores y se va actualizando con la información

compartida entre redes en cada iteración, para terminar con un valor que refleja la confianza en la interacción, siendo negativo cuando la evidencia dice que no hay conexión entre los nodos y positiva en el caso opuesto [155].

Como puede verse por las premisas de PANDA, el algoritmo es bastante general y podría servir para buscar otros reguladores de la transcripción, como de hecho pasa. PUMA (*PANDA Using MicroRNA Associations*) es una adaptación que encuentra miRNAs y TFs, siguiendo los mismo pasos, pero comenzando con una red de co-expresión transcritos-miRNAs, la red de PPI, una red regulatoria que incluye vínculos validados con miRNAs y una lista de reguladores (miRNAs) que no cooperan entre ellos y que entonces no deben aparecer en la red de PPI [156]. Aunque la cooperatividad entre reguladores distintos a los TFs no ha sido descartada, tampoco cuenta con redes establecidas que puedan sumarse a la red de PPI. Por ahora, podrían integrarse con PUMA tantos predictores de la expresión como se desee, simplemente anotándolos en la lista de no cooperadores; alternativamente, SPIDER permite considerar la accesibilidad de la cromatina propia de un fenotipo. La particularidad de SPIDER (*Seeding PANDA Interactions to Derive Epigenetic Regulation*) es que la red de regulación que alimenta a PANDA incorpora datos de DNAse-seq, de modo que las aristas implican un motivo del TF que sobrelapa con una región abierta de la cromatina y con la región regulatoria del gen diana, incluyendo sitios fuera del promotor proximal. Añadir la información sobre la accesibilidad de la cromatina predice redes similares a los observado con ChIP-seq y funciona aún mejor cuando el foco se pone sobre regiones regulatorias distales [157].

Así, el llamado *message passing* permite descubrir interacciones nuevas, específicas de un fenotipo, que no podrían encontrarse con la integración guiada por redes a priori y que sería difícil identificar entre las aristas de las redes probabilísticas. Sin embargo, en el caso de PANDA, las interacciones nuevas accesibles están acotadas al mecanismo donde un regulador actúa sobre dianas co-expresadas o coopera con otros reguladores de la misma diana [155]. Otras herramientas que implementan *message passing* no sufren de esta restricción, porque se dedican a la clasificación.

SNF (*Similarity Network Fusion*) es un ejemplo de *message passing* para la agrupación de pacientes. Empieza construyendo redes de similaridad entre muestras de acuerdo a las distintas ómicas, que luego se combinan de manera iterativa, hasta la convergencia de todos los grafos en un solo. La red final incorpora tanto información común como complementaria y contiene una cantidad reducida de ruido, al descartar las similitudes sin mucho peso. La integración con SNF de los datos de metilación del DNA, expresión de transcritos y de miRNAs de glioblastoma

multiforme, muestra que cada ómica produce topologías muy diferentes, pero en conjunto sostienen tres subtipos con diferencias significativas en cuanto a supervivencia. En la red integrada 49.5 % de las aristas provienen de dos ómicas y 17.2 % de las tres estudiadas, el resto sólo son soportados por una ómica, pero representan grandes similitudes. Además de la clasificación como tal, las redes ayudan entender la heterogeneidad entre pacientes, por ejemplo, la similitud intra-subtipo puede rastrearse a alteraciones particulares, como la sobre-expresión de *CTSD* en el segundo grupo, que afecta la respuesta al fármaco normalmente usado para tratar este tipo de cáncer [158].

MONET (*Multi Omic clustering by Non-Exhaustive Types*) sigue una estrategia similar, iniciando con una red de similaridad por ómica y después buscando sub-grafos pesados y recurrentes entre las ómicas, con un algoritmo codicioso. Cada subgrafo representa un módulo de pacientes similares, sin que necesariamente haya conexión entre los módulos ni que dos módulos surjan de las mismas ómicas. Esto posibilita la integración de ómicas incompletas y la identificación de pacientes atípicos que no encajan en ningún módulo. La evaluación de la herramienta con metilación, expresión de transcritos y expresión de miRNAs de distintos tipos de cáncer, arroja una mayoría de módulos basados en una sola ómica y agrupamientos consistentes entre reinicios [159].

SNF y MONET comparten hasta cierto punto las redes de similaridad por ómica, lo que las hace aptas para el análisis de pocas muestras e inmunes a las diferencias de escala y tamaño entre plataformas. Aunque difieren en cuanto a los criterios para integrar las redes de cada ómica, ambas herramientas van descartando las aristas con menos peso, dándoles robustez ante el ruido y la heterogeneidad de los datos [158]. En suma, las propiedades del *message passing* son deseables para la clasificación del cáncer. Si el objetivo es buscar mecanismos multi-ómicos, el *message passing* también es funcional, como los demuestran PANDA, PUMA y SPIDER, aunque existen alternativas.

La aproximación de Kim et al., comienza igual que SNF y MONET con redes de similitud por ómica, además de una red de relaciones entre ómicas conocida a priori. Pares de regulación conocidos si se integra la expresión de transcritos y miRNAs; loci afectados por CNA, cuando se incorpora esta ómica. Luego, en vez de fusionar las redes iterativamente, se ajustan los coeficientes para obtener una combinación lineal de las mismas, en lo que pareciera ser el equivalente análitico de los ejemplos anteriores. Para encontrar la mejor opción hace falta un estudio comparando tanto los resultados de cada herramienta, como las propiedades de los algoritmos, que, hasta donde se tiene conocimiento, actualmente no está disponible. Lo que se puede de-

cir al momento es que SNF/MONET tendrían una ventaja, al depender exclusivamente de las ómicas, sin necesitar de redes a priori. Tal y como se ha observado previamente, los resultados de Kim et al. basados en más información superan a los que dependen de información parcial. Coincidentemente, hay componentes aislados en la red integrada, pero la mayoría de los nodos se conectan, sugiriendo que los distintos niveles funcionales interactúan, más aún, los autores sospechan sinergia entre metilación y miRNAs en la regulación de la expresión [100].

Otro ejemplo sin *message passing*, sólo para integrar transcritos y proteínas, parte de dos redes independientes de co-expresión, cuyas matrices de adyacencia permiten computar el sobrelape entre redes y encontrar módulos conservados en los dos niveles. Posteriormente las aristas se filtran de acuerdo a su asociación con el fenotipo. El enriquecimiento funcional de los módulos destaca la coordinación entre niveles [160]. En este caso el requisito de las redes multicapa del mapeo entre niveles es muy claro, aunque la relación entre transcritos y proteínas no es 1:1, ni lo sería entre transcritos y metilación o miRNAs, exhibiendo otra ventaja de PANDA y sus variantes.

En conclusión, las redes permiten la integración multi-ómica sin demasiados problemas con la compatibilidad entre plataformas. Es posible distinguir, con algunas excepciones, tres aproximaciones diferentes a la integración con redes, cada una con sus ventajas y desventajas, pero claramente orientadas hacia objetivos específicos. La integración con redes probabilísticas es más versátil y permite observaciones generales sobre la manera en que se conectan los distintos niveles funcionales, pero los artículos discutidos tienen un sesgo claro hacia la búsqueda de reguladores potenciales y eventualmente hacia la propuesta de mecanismos multi-ómicos. Las redes a priori facilitan el hallazgo de nodos de interés, con la expectativa de que los predictores que explotan múltiples tipos de datos sean más robustos y reflejen la complejidad de la enfermedad [160]. Finalmente, las aproximaciones que fusionan redes plantean una clasificación más completa de los tumores, sin necesariamente asumir una estructura común a todos los niveles funcionales [159]. La gran excepción son los derivados de PANDA, más cercanos a una herramienta de fusión de redes que al resto de aproximaciones, pero orientados hacia la propuesta de reguladores.

Para simplificar la recuperación y comparación de las redes producidas por estas herramientas se han creado bases de datos dedicadas. iNetModels contiene redes multi-ómicas relacionadas con distintas condiciones metabólicas [161]; mientras que GRAND (*Gene Regulatory Network Database*) contiene redes transcripcionales y de la integración transcritos-miRNAs, obtenidas con los derivados de PANDA, a partir de los datos en la enciclopedia de líneas celulares del cáncer (CCLE) y el TCGA [162]. El propósito final de estos repositorios es facilitar la reutilización de

redes multi-ómicas ya publicadas y propiciar la experimentación dirigida. Sin embargo, aún sin las dificultades para distinguir interacciones directas e indirectas o para asignarle dirección a las aristas, es necesario recordar que la correlación no basta para inferir causalidad [143].

Tabla 2: Herramientas de integración multiómica mencionadas. El objetivo puede ser cualquier combinación de: sugerir mecanismos *m*, predicción *p*, agrupamiento de muestras *a*, enriquecimiento funcional *f* o identificación de marcadores *b*. Las temporalidad de la integración puede ser temprana *e*, intermedia *i* o tardía *l*.Se resaltan las herramientas utilizadas en este proyecto.

nombre	objetivo	técnica	temporalidad
MONET	ab	*message passing*	i
SNF	ab	*message passing*	i
PUMA/SPIDER	mb	*message passing*	i
iOmicsPASS	ap	centroides y redes a priori	i
Mergeomics	b	difusión de calor y redes a priori	t
NetICS	b	difusión de calor sobre redes a priori	i
APODHIN	fb	meta-LFC	t
LeMoNe/LemonTree	m	redes de coexpresión y regresión	e
ARACNE	mb	información mutua	e
DRAGON	b	matriz de precisión penalizada	e
Red elástica	pmb	descomposición y penalización ENET	e
miRDriver	b	LASSO	e
sPLS	pamb	descomposición y penalización LASSO	e
SGCCA	m	descomposición y LASSO	e
RGCCA	mb	descomposición de matrices	e
MOTA	b	RGCCA	e
iCluster	a	descomposición de matrices	e
JIVE	mb	descomposición de matrices	e
MOGSA	f	proyección a menor dimensión	e
ActivePathways	f	meta-significancia	t
CoCa	a	clustering consenso	t

Objetivo general

Analizar de manera integrada el papel de la metilación del DNA, de la expresión de factores transcripcionales y de miRNAs en los cambios transcripcionales asociados al cáncer de mama.

1. Objetivos específicos

1. Encontrar las relaciones entre la expresión genética y las tres capas de regulación -metilación del DNA, expresión de factores transcripcionales y expresión de miRNAs- mediante redes probabilisticas.
2. Asociar los patrones de expresión del cáncer de mama con la alteración de estas capas, integrando un analisis diferencial de la expresión y la metilación con las interacciones encontradas.
3. Identificar los procesos biológicos más afectados por dichas capas a través del ajuste de modelos escuetos.
4. Buscar diferencias en estas asociaciones entre los subtipos del cáncer de mama.

Materiales y métodos

Durante el desarrollo del proyecto se exploraron tres aproximaciones diferentes a la integración multi-ómica con los datos del cáncer de mama del TCGA. Específicamente se integraron datos de metilación del DNA, expresión de transcritos codificantes y expresión de miRNAs, mediante: un modelo de red elástica para cada gen en la firma prognóstica PAM50, un modelo de red probabilística y un modelo de SGCCA. En los tres casos se construyó un modelo por subtipo y uno para el tejido normal. El flujo de trabajo general incluye tres pasos secuenciales: pre-procesamiento de los datos, integración y análisis de resultados. Sin embargo, más allá del pre-procesamiento, cada modelo tiene sus propios métodos y alcances, por lo cual esta sección se divide en concordancia.

1. Pre-procesamiento

Los modelos de red elástica y probabilística surgen de una descarga de la base de datos *Genomic Data Commons* en mayo del 2019. Para el SGCCA se volvieron a descargar y pre-procesar los datos en diciembre del 2021, con el objetivo de aumentar el tamaño de la muestra, pues la intersección entre los datos de metilación y expresión aumentó entre las dos fechas. En ambos casos se trata de datos abiertos, del nivel 3, que consisten en cuentas de expresión ya mapeadas pero sin normalizar y valores beta de metilación. Así, los primeros dos modelos surgen de la caracterización por el TCGA de 45 tumores enriquecidos de HER2, 395 luminales A, 128 luminales B, 125 basales y 75 muestras de tejido normal adyacente. A su vez, el SGCCA tiene como entrada la información de 46 tumores enriquecidos de HER2, 416 luminales A, 140 luminales B, 128 basales y las mismas muestras de tejido normal. Los subtipos fueron asignados por el TCGA mediante PAM50. Respetando los estándares de cada ómica [163–165], en la segunda ocasión se ajustaron algunos parámetros, como se detalla a continuación. Para facilitar el contraste, además se enlistan las diferencias entre descargas en la tabla **3**.

Al tratarse del mismo tipo de plataforma, el procesamiento de transcritos y miRNAs es muy similar y está orientado hacia el análisis de expresión diferencial. La diferencia entre RNA-seq y miRNA-seq es un enriquecimiento de moléculas pequeñas [1]. Así, el pre-procesamiento de transcritos codificantes demanda el filtrado de transcritos con bajo número de cuentas, la normalización de sesgos entre transcritos y entre muestras y la corrección del efecto de lote. El enriquecimiento del miRNA-seq hace que no haya sesgos significativos -o conocidos- entre miRNAs, de modo que su pre-procesamiento no requiere la normalización entre miRNAs.

En la primera ocasión el umbral de bajo número de cuentas se puso en 10 cuentas por millón, se aplicó la normalización TMM (*trimmed mean of M-values*) entre muestras y se corrigieron los posibles efectos de lote con ARSyN, tanto en el caso de transcritos como en el de miRNAs. Además, con el objetivo de hacer posible la comparación entre transcritos, se aplicó la normalización FULL sobre los sesgos por el largo y el contenido de GC, antes de la normalización entre muestras.

Dado la baja representación de miRNAs en los modelos, en el 2021 el umbral de bajo número de cuentas se bajó hasta 0. Esto mantiene entre los datos a todos los transcritos y miRNAs que de hecho llegan a ser secuenciados, pasando de 16475 transcritos codificantes y 433 precursores de miRNA, a 17077 transcritos y 604 precursores. Como consecuencia se espera más ruido, pues un bajo número de cuentas puede ser el resultado de problemas en la secuenciación y no reflejar alguna señal biológica [164]. Sin embargo, la gran diferencia en la expresión que se ha documentado entre cáncer y tejido normal hace suponer que un filtro laxo no afectará el análisis de los transcritos [166]; mientras que el bajo número de cuentas de miRNAs que se espera por librería [165] justifica el umbral. Aqui hay que resaltar que los identificadores que acompañan los datos de miRNA-seq corresponden a precursores, con la finalidad de ligar CpGs y miRNAs. Aunque el uso de pre-miRNAs introduce una fuente de ambiguedad en cuanto al efecto real de los reguladores sobre las dianas, permite identificar la secuencia de origen del miRNA y entonces a los CpGs inmediatos. Mientas los miRNAs maduros son los que efectuan la regulación, sus secuencias no siempre puede ser mapeadas a un loci únicos sobre el DNA [167], impidiendo la identificación de CpGs reguladores, por lo que se prefirio el uso de precursores.

En cuanto al siguiente paso, también se cambió la normalización entre muestras de miRNAs al método de medianas, porque la prueba de valores M [2] arroja mejores resultados con esta

[2]Los valores M reflejan la diferencia en composición del RNA entre dos muestras:

$$M_{gk} = log_2 \frac{\frac{Y_{gk}}{N_k}}{\frac{Y^r_{gk}}{N_r}}$$

normalización que con la TMM.

Por su parte, la metilación del DNA proviene del microarreglo *Illumina Human Methylation 450*. Aunque la base de datos también incluye observaciones hechas con la versión *Illumina Human Methylation 27*, sólo se tomaron los datos del primer microarreglo, que cubre mayor porcentaje del genoma. El microarreglo elegido cubre el 99 % de los genes en RefSeq, tanto en la regiones regulatorias alrededor del promotor, como en el cuerpo de los genes, lo que equivale a 485,512 sitios de CpG [168]. Esta cantidad abarca sitios que pueden aportar más ruido que información, como son aquellos medidos con sondas que tienen mapeo ambiguo, sondas sobre cromosomas sexuales y sondas que coinciden con SNPs. Además de estos, se descartaron las sondas sin medición en más del 25 % de las muestras de los cuatro subtipos y el tejido normal. El resto de los valores faltantes fueron imputados con los 15 vecinos más cercanos. Finalmente, los valores beta originales fueron transformados en valores M, con distribuciones más cercanas a la normal preferida por los análisis de metilación diferencial [169].

A diferencia del 2019, en la segunda descarga de los datos no se encontraron muestras ligadas a pacientes masculinos, de manera que los cromosomas sexuales no son fuente de ruido y sus CpGs pueden conservarse. Sumando la diferencia en el número de muestras, que afecta el filtro por valores faltantes, el total de CpGs que terminan el pre-procesamiento en cada ocasión es de 384,575 y 393,132, respectivamente.

1.1. Análisis de expresión y metilación diferencial

Aunque no es precisamente pre-procesamiento, el análisis de expresión y metilación diferencial se describe en esta sección por estar ligado a las descargas de los datos.

De nuevo, el análisis de expresión diferencial de miRNAs es equivalente al de transcritos, en el sentido de que en ambos casos hay que plantear un modelo base con las comparaciones de interés, ajustar el modelo y recuperar los elementos con expresión diferencial significativa [170]. En ambas ómicas, se ajustó un modelo con un coeficiente por subtipo y se evaluó la expresión diferencial con TREAT (*t-tests relative to a threshold*), que a diferencia del método de Bayes empírico usualmente empleado, considera un umbral de cambio en la hipótesis, disminuyendo los falsos positivos. La diferencia entre transcritos y miRNAs es justamente este umbral, que se toma

dónde M_gk son las cuentas del gen g en la muestra k y N_r es el total de cuentas en la muestra r. Se considera que las muestras son comparables si la mediana de los valores M es 0, pues tienen composición similar, de otra manera su comparación podría señalar expresión diferencial erróneamente

como 1.5 y 1.1, respectivamente. Aunque se recomienda usar un valor entre 1.2 y 1.5, enfocado a diferencias de dos órdenes en la expresión genética [171], se eligió un valor ligeramente menor para el análisis de miRNAs, dado su bajo número de cuentas.

La diferencia entre 2019 y 2021, es que en la primera ocasión se aplicó la normalización voom antes del ajuste del modelo, con la finalidad de disgregar la varianza del nivel de expresión [172]. En la segunda oportunidad, ARSyN regresa logaritmos base dos, haciendo inviable la aplicación de la normalización voom. Siguiendo con el espíritu del pre-procesamiento más laxo, también se cambió el método de identificación de miRNAs diferencialmente expresados al empírico de Bayes.

Respecto a la metilación del DNA se tomaron caminos completamente diferentes. En el 2019 se optó por la estrategía más simple de ajustar un modelo lineal, tal y como se hizo en el análisis de expresión diferencial, pero considerando las covariables: plato, muestra, vial, porción y género. Aunque esté método no está pensado para datos de metilación, cuya distribución bimodal claramente no coincide con la distribución normal asumida, se ha comprobado que las diferencias entre cáncer y tejido normal tienen tal magnitud, que todos los métodos arrojan resultados similares [173]. Por el contrario, en el 2021 se removió la varianza no deseada con una aproximación específica para este tipo de datos [174]. Al terminar el preprocesamiento de

Tabla 3: Diferencias entre los datos obtenidos en las dos ocasiones

	2019	2021
umbral de bajo número de cuentas	10	0
normalización entre muestras (miRNAs)	TMM	medianas
filtro cromosomas sexuales(HM450)	si	no
ARSyN regresa log2	no	si
expresión diferencial de miRNAs	TREAT	eBayes
metilación diferencial	limma ajustado	RUVfit
muestras		
HER2E	45	46
luminal A	395	416
luminal B	128	140
basal	125	128
tejido normal	75	75
predictores finales		
CpG	384,575	393,132
transcritos codificantes	16475	17077
miRNAs	433	604

las ómicas por separado, las matrices de datos fueron concatenadas en una matriz para cada subtipo y el tejido normal, de modo que los valores se alinean por paciente, independientemente

de la plataforma de origen. Dado que la base de datos tiene algunas mediciones repetidas de un mismo tumor y paciente, se calculó el promedio por elemento, obteniendo una matriz que tiene predictores (CpGs, transcritos y miRNAs) únicos en un sentido y pacientes únicos en el otro sentido. La separación de los datos por subtipo permite que los análisis subsecuentes capturen la heterogeneidad del cáncer de mama, al menos en lo que respecta a las diferencias entre subtipos.

2. Redes elásticas

Las redes elásticas son modelos de regresión lineal escuetos, que permiten la identificación de los predictores que mejor explican la variable dependiente. Con el objetivo de encontrar los CpGs, transcritos y miRNAs que predicen las expresión de los genes en la firma pronóstica PAM50, se ajustó un modelo de red elástica por gen y subtipo. La idea es filtrar, de entre ómicas completas, los predictores que podrían estar funcionando como reguladores de la firma que mejor distingue entre subtipos y así, observar similitudes y diferencias.

Como primer paso cada matriz concatenada fue normalizada, de manera que la media sea 0 y la desviación estándar 1, y dividida en datos de entrenamiento y datos de prueba. Siguiendo el ejemplo de Liu et al. [134], los datos de prueba representan el 20 % de las observaciones. El 80 % restante se usó para ajustar los modelos. Durante el ajuste se fijó el parámetro de mezcla en 0.5, ya que este valor tiende seleccionar grupos completos de variables correlacionadas [131], y se optimizó el parámetro de contracción usando valores entre 0.001 y 1000, mediante validación cruzada.

La validación cruzada se repitió 100 veces, fraccionando los datos del subtipo HER2E y el tejido normal en 3 partes y los subtipos luminal A, luminal B y basal en 5 partes, debido a las diferencias en cuanto al tamaño de muestra. La elección de parámetros se basa en el menor error RMSE (*root mean squared error*) obtenido en los datos de prueba. De manera alternativa se repitió el ajuste, pero usando sólo 40 muestras por subtipo, para verificar el efecto que tiene el tamaño de la muestra. Una vez teniendo los parámetros, se calculó el RMSE sobre los datos de prueba ya sea con todos los predictores seleccionados o manteniendo en el modelo sólo los predictores de cada nivel funcional, de manera similar al análisis de dependencia de Setty et al. [122]. Los valores de RMSE así obtenidos fueron comparados a nivel distribución por subtipo y ómica o exclusivamente por ómica, con una prueba de Kolmorogov-Smirnov y corrección FDR de la significancia.

2.1. Análisis de resultados

La selección de predictores de la red elástica implica una conexión entre el transcrito predicho y un conjunto de CpGs, transcritos y miRNAs. Agrupando los modelos por subtipo, se generan redes a analizar en cuanto al enriquecimiento funcional y de interacciones regulatorias. En el aspecto topológico, sólo se estimó el grado de entrada y salida de los nodos considerando la dirección del modelo, es decir, con los predictores apuntando (prediciendo) los genes del PAM50. Los coeficientes de los predictores con el mayor grado de salida se compararon con la distribución global de su ómica, mediante una prueba de Kolmorogov-Smirnov, para evaluar su predictibilidad.

Los transcritos seleccionados en cada modelo fueron sometidos a un análisis de enriquecimiento funcional contra la ontología de procesos biológicos. Posteriormente se evaluó si los procesos enriquecidos en más de una red, involucran a los mismo predictores. El enriquecimiento de interacciones regulatorias se evaluó con pruebas de Fisher y corrección FDR de la significancia. Las fuentes de interacciones regulatorias cambian dependiendo del nivel funcional. La metilación del DNA se tomó de la anotación del microarreglo, que mapea cada sonda con los genes afectados, de acuerdo a su posición sobre el genoma. Los pares TF-diana se descargaron de las bases de datos TRED, ITFP, ENCODE, TRRUST, Neph2012 [175] y Marbach2016 [176], mediante tftargets, incluyendo tanto predicciones como relaciones validadas. Finalmente, los pares miRNA-diana provienen de DIANA-microT-CDS, ElMMo, MicroCosm, miRanda, miRDB, PicTar, PITA, TargetScan, miRecords, miRTarBase y TarBase, también incluyendo predicciones y pares validados. Al tomar todas las relaciones regulatorias que involucran a los predictores seleccionados como universo de fondo, es posible probar si los modelos de red elástica tienen más interacciones regulatorias de lo que se esperaría por azar.

3. Redes probabilisticas

Con el objetivo de encontrar relaciones potencialmente regulatorias que involucren a los tres niveles funcionales, no sólo a los genes del PAM50, se construyeron redes MI. Puesto que la distribución del MI cambia con el tipo de moléculas que se conecten [84], en vez de fijar un umbral de información mutua único, se seleccionaron las 10000 interacciones con mayor valor de cada par de niveles, como se ha hecho con las redes transcripcionales [177]. Debido al espacio requerido para almacenar todos los pares posibles de CpGs, este tipo de interacción no fue estimado, de modo que las redes están formadas por pares CpG-transcrito, CpG-miRNA, transcrito-transcrito,

transcrito-miRNA y miRNA-miRNA.

Fijar un umbral de MI podría dejar fuera relaciones relevantes pero que tiene MI menor que el umbral sólo por el tipo de predictores que conectan. Por otro lado, un top arbitrario podría recuperar aristas sin señal biológica. Aprovechando la relación entre MI y p-value, donde las aristas con mayor valor son las que pueden atribuirse al azar con menor probabilidad, se registró el umbral de significancia que comprende las aristas en el top 10000. Así, el análisis impone una topología, comparable entre subtipos gracias al tamaño común y permite examinar la fuerza de las dependencias estadísticas, al acompañar cada arista de mínimo de significancia. Las distribuciones de MI fueron comparadas usando una prueba de Kolmogorov–Smirnov y la corrección FDR.

Finalmente se comprobó el efecto del tamaño de la muestra, evaluando la variabilidad en la información mutua cuando se toman subconjuntos de los datos, a través de z-scores. Con este motivo se repitió 100 veces la inferencia de las redes luminales, basal y del tejido normal, partiendo cada vez de 45 muestras, que es la cantidad de observaciones disponibles para el subtipo HER2E.

3.1. Análisis de resultados

Tal como en la aproximación anterior, se examinó el enriquecimiento funcional en las redes y la selección de interacciones regulatorias conocidas. En este caso, el análisis de la topología de las redes incluye la estimación de los caminos más cortos de cada nodo, además de la centralidad de grado. Al someter las distribuciones de los niveles funcionales a una prueba de Wilcoxon rank sum tests con corrección de FDR, es posible evaluar si las ómicas juegan roles distintos dentro de la red.

Dado que las redes exhiben a los transcritos que tienen relaciones de dependencia estadísticas con otros niveles funcionales, se abre la oportunidad de buscar funciones asociadas a varios niveles funcionales y funciones alteradas en el cáncer de mama, cuya de-regulación podría explicarse en esos niveles. Con esto en mente, se examinó la sobre-representación de los procesos biológicos con puntaje de GSEA significativo en las redes. Primero se hizo el análisis de GSEA sobre los resultados de expresión diferencial de cada subtipo y luego se evaluó la sobre-representación de los procesos con p-value menor que 0.01 en la red correspondiente. Para el análisis de sobre-representación se puso un umbral de significancia de 0.05, con el fin de conservar tantos procesos como sea posible.

Habiendo encontrado las funciones alteradas en cáncer y asociadas a múltiples niveles funcionales, se examinaron los CpGs, transcritos codificantes de TFs y miRNAs asociados, ante la posibilidad de que estén desempeñando un papel como reguladores. Como primer paso se revisó el enriquecimiento de cada tipo de reguladores potenciales respecto a la red de tejido normal, por proceso, con una prueba de Fisher, controlando las múltiples pruebas con la corrección FDR. Para hacer una comparación global, se calculó el índice de Jaccard, por proceso y entre subtipos y, para cada par de procesos del mismo subtipo. El índice de Jaccard divide el tamaño de la intersección de dos conjuntos, entre su unión. El contraste entre subtipos cuenta aristas en lugar de nodos, pues la pregunta es si las interacciones potencialmente regulatorias se conservan. Las distribuciones así obtenidas fueron sometidas a pruebas de Kolmogorov–Smirnov tests con la corrección FDR, para comprobar la significancia de las diferencias observadas. Al calcular el índice entre procesos del mismo subtipo, también se clasifican los reguladores potenciales en exclusivos y compartidos, para después verificar el enriquecimiento de cada clase en los distintos subtipos con una prueba de Fisher.

4. SGCCA

La última aproximación a la integración multi-ómica que se implementó es el análisis de correlación canónica escueto y generalizado. Esto se hizo con los datos descargados en el 2021, por lo que no es directamente comparable con los análisis descritos anteriormente, que usan sólo el subconjunto de datos disponibles en el 2019. La normalización aplicada sobre la matriz concatenada también es diferente, pues en este caso se divide cada bloque de datos por la raíz cuadrada de su primer valor propio. Esto equivale a pesar cada ómica por su varianza e impide que la metilación del DNA domine los resultados sólo por ser de mayor tamaño [111]. Por otro lado, el análisis de resultados se mantiene muy similar, concentrándose en el enriquecimiento funcional y usando las interacciones regulatorias conocidas ya mencionadas.

Para correr el SGCCA hace falta elegir el número de componentes latentes a recuperar (ncomp), qué tan escuetos serán los componentes, una matriz de diseño (C), y una función para maximizar la covarianza (g). Los parámetros entre paréntesis se eligieron de manera arbitraria, pero justificadas. Un PCA de los datos muestra que cada componente latente explica menos del 5 % la varianza en los datos, por lo que se decidió que ncomp sea igual al número de muestras en cada matriz menos uno, con la intención de explicar la mayor parte posible de los datos. Para la matriz C se eligió el diseño más simple, que dicta la existencia de vínculos entre las tres ómicas;

y para maximizar la covarianza se eligió la función centroide, que admite tanto correlaciones positivas, como negativas.

Finalmente, se ajustó un parámetro de contracción de los componentes latentes por ómica, de la secuencia [0.01,0.02,...,0.09,0.1,..,0.9], mediante validación cruzada. Puesto que se desea producir resultados comparables entre subtipos, se usó el mismo parámetro de contracción. Para elegir parámetros universales, el ajuste se hizo sobre sets de datos balanceados, formados por 10 muestras tomadas al azar de cada subtipo y el tejido normal. Se corrieron 10 instancias del SGCCA por cada combinación de parámetros -uno por ómica-, con datos balanceados diferentes, recuperando sólo un componente latente y registrando el número de predictores seleccionados y la varianza explicada promedio (AVE). En total, cada valor se probó 11340 veces por ómica. Los parámetros de contracción elegidos optimizan simultáneamente la varianza explicada promedio y el número de predictores seleccionados, **figura 6**. Los modelos definitivos usan el total de muestras de cada subtipo y el tejido normal, y los parámetros elegidos.

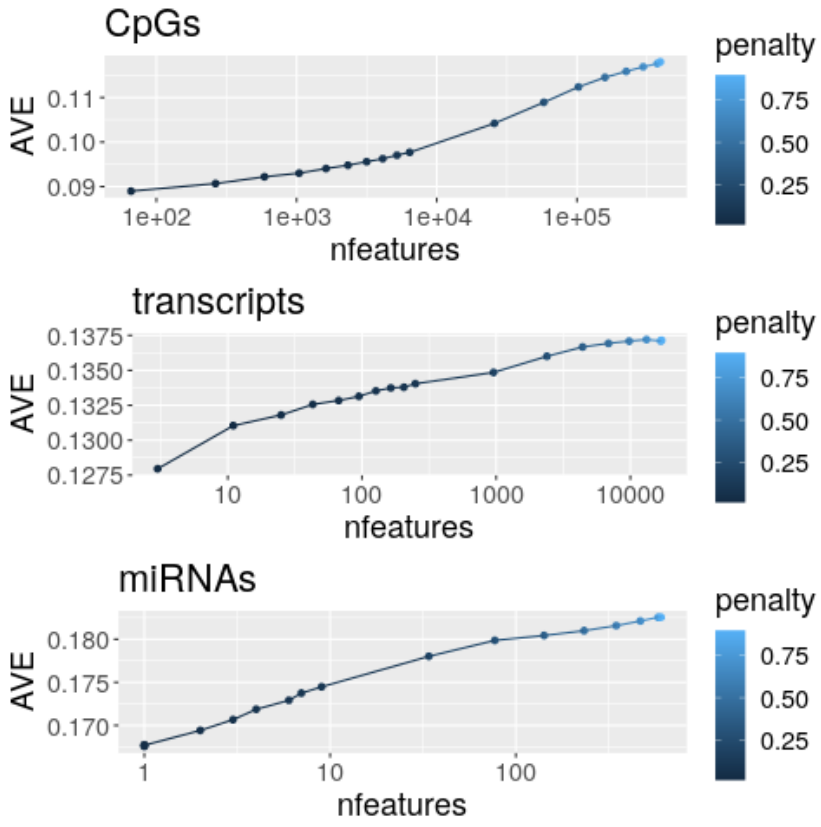


Figura 6: Elección de los parámetros de contracción, de acuerdo a la mediana de predictores seleccionados (nfeatures) y la varianza explicada promedio (AVE). Los parámetros elegidos son los puntos previos al mayor cambio en la pendiente, es decir, 0.02 para sitios CpG y transcritos y 0.05 para microRNAs.

4.1. Análisis de resultados

Los resultados del SGCCA incluyen una matriz escueta de predictores por componentes latentes, donde las celdas contienen los coeficientes contraídos. Los predictores con coeficientes distintos a cero maximizan la covarianza entre ómicas [99] y permiten encontrar las funciones asociadas a la covariación entre niveles funcionales. Con el objetivo de explotar todos los predictores co-seleccionados y no solamente los transcritos, todos los identificadores se tradujeron a genes de Entrez. Mientras los transcritos y miRNAs cuentan con anotación directa en Entrez; los CpGs se mapearon a través de sus dianas, lo que resulta en una amplificación de su representación, ya que un sitio puede afectar todo un *cluster* de genes. Una vez terminado el mapeo, se corrió un análisis de enriquecimiento funcional contra la ontología de procesos biológicos y las vías de KEGG, de cada grupo de predictores co-seleccionados. Siguiendo las jerarquías de ambas bases de datos y concentrándose en las funciones encontradas exclusivamente en algún subtipo o el tejido normal, se evaluó además el enriquecimiento de categorías funcionales, con una prueba de Fisher, aplicando la corrección de Holm sobre los valores de significancia.

De manera independiente se corrió un análisis GSEA, con los valores de expresión diferencial de cada subtipo respecto al tejido normal y sin aplicar ningún umbral de significancia. De esta manera se recuperan todas las funciones examinadas y es posible asignarle un puntaje GSEA a cada proceso y vía identificadas con el SGCCA. La pregunta es si las funciones ligadas a los predictores que maximizan la covarianza entre ómicas son afectadas por la expresión diferencial.

Aunque no todos los componentes están enriquecidos funcionalmente, hay funciones repetidas en varios componentes, por lo que se agruparon todos los predictores co-seleccionados que comparten algún enriquecimiento y construyeron redes de MI. Para elegir un punto de corte de MI que respete las diferencias entre niveles funcionales, se calculó también la información mutua entre pares regulatorios conocidos. Esto es, de las bases de datos ya mencionadas, se extrajeron interacciones regulatorias que involucran a los CpGs, transcritos codificantes de TFs y miRNAs seleccionados, se estimó su MI y se tomaron las medianas como umbral específico de cada mecanismo de regulación. Dado que la idea es basar los umbrales en relaciones regulatorias, en esta ocasión las bases de datos se limitaron a interacciones validadas, descartando predicciones y concentrándose en lo demostrado, que en el caso de los TFs incluye los resultados de ChIP-seq, huella de DNasa y experimentos de pequeña escala. La elección de la mediana sobre la media evita que el umbral sea dominado por valores extremos. Además, con la finalidad de admitir más aristas en la red final, cuando las distribuciones de MI no son significativamente diferentes de

acuerdo a una prueba de Kolmorogov-Smirnov, se toma como umbral único el valor más bajo de entre las medianas calculadas, independientemente del mecanismo regulatorio.

Las redes obtenidas permiten visualizar reguladores potenciales de las funciones ligadas a la covarianza entre niveles funcionales. Estas redes se usaron como guía para una revisión de la literatura, que verifica si hay evidencia externa apoyando a las relaciones regulatorias sugeridas. De esta manera, se construyó un flujo de trabajo semi-automatizado que produce modelos regulatorios escuetos y verificables.

Todos los procesos aquí descritos están registrados con paquetes, funciones y parámetros específicos en el repositorio de github mSolEdadO/TCGAmiRmethyRNAIntegration.

Resultados

Como se expuso en la introducción, hay distintas aproximaciones a la integración multi-ómica, cada una con sus ventajas y desventajas. Cada aproximación también cuenta ya con herramientas de *software* definidas para lograr objetivos de clasificación, identificación de funciones, marcadores y de relaciones entre niveles funcionales. Con esto en mente, se planteó el flujo de trabajo de la **figura 7**, que aprovecha dos aproximaciones complementarias para satisfacer los objetivos específicos del proyecto. No solo se trata de integrar los datos de metilación del DNA, expresión de transcritos y de miRNAs; sino de entender, o aportar al entendimiento, del efecto que tienen las ómicas en cada uno de los subtipos del cáncer de mama.

Partiendo de datos comparables entre muestras y entre niveles funcionales, es decir, preprocesados, se proyecta la construcción de dos modelos multi-ómicos: una red probabilística y un modelo escueto multivariado. Mientras la red expone las relaciones entre niveles funcionales, resolviendo el primer objetivo específico; el modelo escueto multivariado, selecciona los predictores asociados a la expresión genética, cumpliendo con el segundo objetivo. Posteriormente, un análisis de enriquecimiento funcional sobre el modelo escueto solventa el tercer objetivo. Finalmente, al incorporar los modelos y los análisis de expresión y metilación diferencial, la idea es producir redes fácilmente comparables, cuyo contraste entre subtipos satisfaga el último objetivo.

Dada la estrecha relación entre cada paso y los objetivos del proyecto, el flujo de trabajo original se mantuvo casi intacto durante el desarrollo del proyecto; cambiando únicamente el orden en que se incorporan los modelos y explorando distintas herramientas. La opción definitiva, que pasa por cada proceso de la **figura 7**, consiste en el acoplamiento del SGCCA con redes funcionales de MI. Sin embargo, antes de llegar a esta solución, se obtuvieron resultados igualmente relevantes y válidos, publicados en revistas indexadas bajo los títulos *Multi-omic regulation of the PAM50 gene signature in breast cancer molecular subtypes* [178] y *An information theoretical multilayer network approach to breast cancer transcriptional regulation* [179]. Tanto estos, como el artículo final, se encuentran anexos. Para explicar los resultados se seguirá un orden

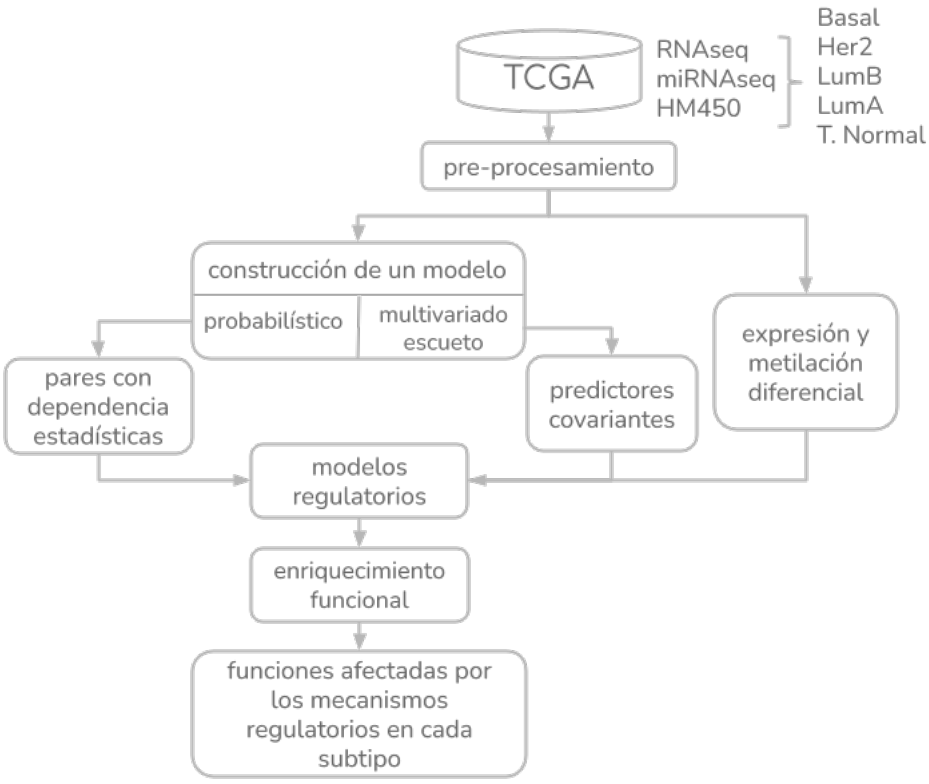


Figura 7: Flujo de trabajo propuesto

cronológico, abordando cada artículo por separado, tal y como se dividió la sección de métodos.

1. Regulación multi-ómica de la firma PAM50 en los subtipos del cáncer de mama

Empezando con el segundo objetivo, de asociar los patrones de expresión del cáncer de mama con la alteración de la metilación, la expresión de TFs y de miRNAs, se propuso la predicción de la expresión de los 50 genes del PAM50, como un primer acercamiento al análisis completo. De este modo, se ajustaron modelos de red elástica por cada uno de los genes y cada uno de los subtipos, tomando como predictores todos los datos disponibles, excepto el patrón de expresión del gen correspondiente. A pesar de necesitar el ajuste de más parámetros y carecer de paquetes exclusivamente dedicados a ello, se prefirió la red elástica sobre la penalización LASSO, por la capacidad de seleccionar grupos completos de variables correlacionadas. Esta capacidad justifica el uso de ómicas completas, sin filtrar los transcritos codificantes para TFs ni los CpGs ubicados en las cercanías del gen de interés, pues el modelo recuperaría el módulo completo de genes

covariantes, incluidos los reguladores potenciales.

1.1. La contribución de las ómicas a la expresión del PAM50 cambia entre el tejido normal y los cuatro subtipos del cáncer de mama

La salida del modelo es una lista de predictores que explican la expresión del gen. A cada predictor lo acompaña un coeficiente de regresión, cuyo valor refleja el peso de la asociación con el gen y puede ser negativo o positivo. La **tabla 4** muestra el total de predictores seleccionados en los 50 modelos de cada uno de los subtipos. Ahí se sugiere una correlación entre el tamaño de la ómica y la cantidad de predictores seleccionados, siendo los sitios CpGs los más seleccionados. Sin embargo, al evaluar esta hipótesis con una prueba de χ^2 resulta que el tamaño de las ómicas no explican la proporción en que son seleccionadas, sino que hay una sobrerrepresentación de transcritos y miRNAs, visible en la **figura 8a**, que podría deberse a que ambos pertenecen al mismo nivel molecular y están sometidos a las mismas presiones. Al mismo tiempo, considerando relaciones regulatorias conocidas, hay un enriquecimiento de TFs y miRNAs que regulan a los genes del PAM50; pero una sub-representación de CpGs cercanos, siendo la excepción el subtipo luminal B, que tiene más CpGs regulatorios de lo esperado.

Tabla 4: Tamaño de la entrada y la salida de los modelos. TF se refiere a transcritos que codifican factores transcripcionales

	Basal	HER2E	LumA	LumB	Normal
muestras	125	45	395	128	75
CpGs seleccionados	3090	2514	7173	1485	5373
CpGs reguladores seleccionados	9	0	21	12	0
transcritos seleccionados	1525	591	3115	888	2340
TFs seleccionados	207	91	465	133	327
TFs reguladores seleccionados	4	3	25	7	9
miRNAs	101	85	174	116	123
miRNAs reguladores seleccionados	8	5	8	12	5

Tomando la sub-selección de sitios CpG como indicador de que esta ómica tiene menor capacidad de predicción, se hizo un análisis de dependencia sobre las ómicas. Este análisis consiste en evaluar el error del modelo, a través del RMSE, cuando se eliminan predictores mediante la sustitución de sus coeficientes por cero. Al eliminar todos los predictores de una ómica, se puede medir la contribución de la misma sobre la predicción. Aunque el resultado de la comparación de distribuciones no siempre es significativo, el desplazamiento de las distribuciones en la **figura 8b** indica que la metilación del DNA aporta menos a los modelos de todos los subtipos, mientras

que los transcritos y los miRNAS contribuyen por igual. Esto contrasta con reportes previos de una mejora significativa cuando se incluyen datos de metilación [121]. También en oposición a lo esperado, los modelos multi-omicos tienen el mismo error que los modelos que sólo incluyen transcritos, indicando una contribución baja de CpGs y miRNAs a la predicción de la firma PAM50 y apoyando a la expresión genética como el mejor clasificador de los subtipos. Lo que concuerda con el origen de esta clasificación en los niveles de expresión de los transcritos.

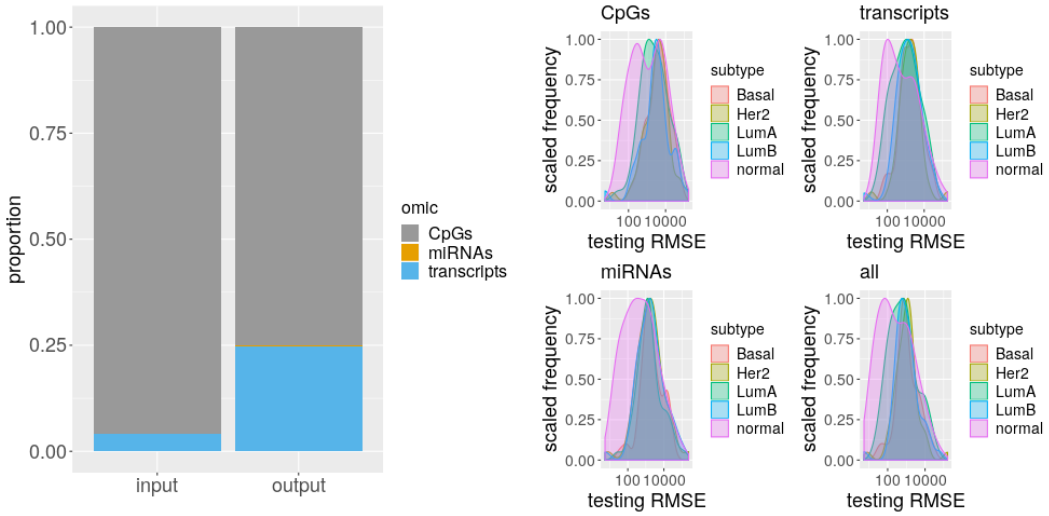


Figura 8: Diferencias entre las ómicas en cuanto a (a), la proporción de predictores seleccionados por los modelos y (b), el error de prediccion.

Las distribuciones de RMSE además permiten la comparación entre subtipos y con el tejido normal. De ahí puede advertirse una alteración de la relación entre las ómicas y la expresión del PAM50 en el cáncer de mama, ya que la capacidad de predicción de los modelos multi-ómicos, y particularmente de los modelos con CpGs, cambia en los subtipos respecto al tejido normal, lo que concide con la hipometilación generalizada que se espera en cáncer [52].

1.2. La fuerza con que las ómicas se asocian a la expresión del PAM50 cambia entre subtipos

Después de comparar las distribuciones de RMSE, había que comparar las distribuciones de los coeficientes, mostradas en la **figura 10**. Puesto que los coeficientes representan la correlación entre predictores y respuesta [121], las diferencias entre ómicas y subtipos son relevantes, pues, como el RMSE, potencialmente hablan de la alteración de los mecanismos de regulación.

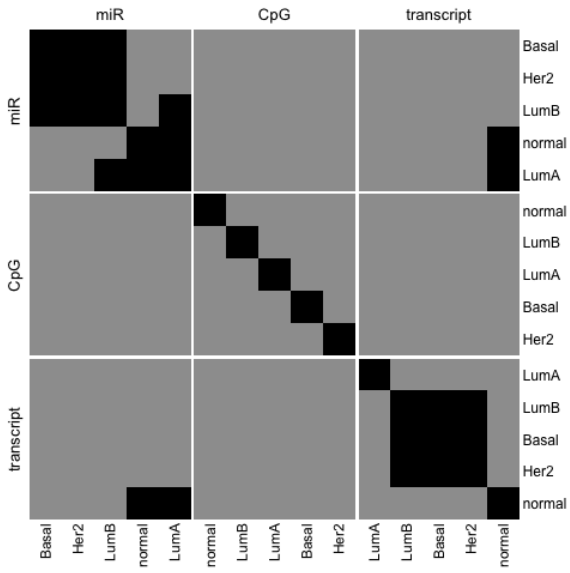


Figura 9: Significancia,ajustada por FDR, de los contrastes correspondientes. Gris representa valores por debajo de 0.05 y negro lo opuesto.

A pesar de las excepciones mostradas en la **figura 9**, en general, los contrastes entre distribuciones de coeficientes son significativos. En otras palabras, la distribución de los coeficientes ajustados para una misma ómica cambia entre subtipos y lo mismo pasa dentro de cada subtipo al contrastar ómicas.

De acuerdo a las distribuciones, la metilación del DNA tiene una asociación fuerte con los genes del PAM50; mientras que miRNAs y transcritos tienden a formar asociaciones positivas. El pico de la distribución sobre valores relativamente elevados, podría explicar la selección de tantos CpGs a pesar de su baja predictibilidad y coincide los resultados de Sohn et al., de mayor asociación entre metilación y transcripción que entre miRNAs y transcripción [126].

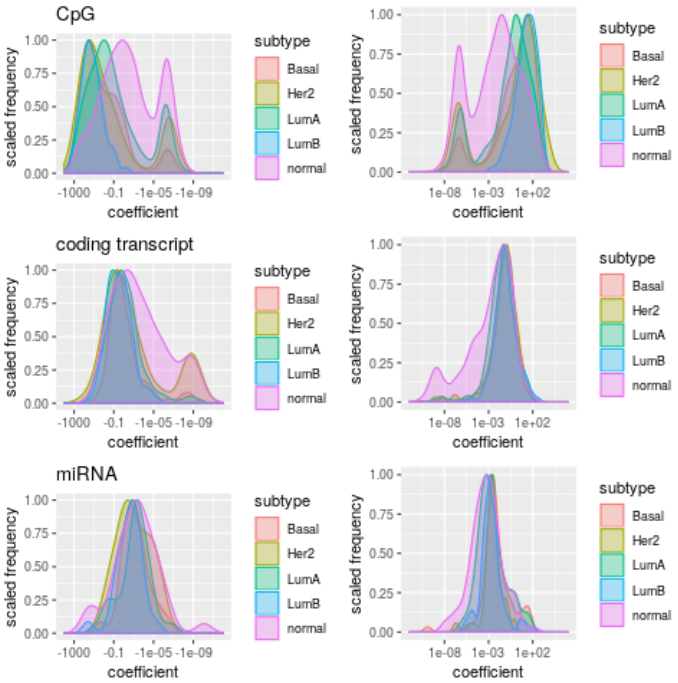


Figura 10: Distribución de los coeficientes de cada ómica. Cada fila del gráfico muestra una omica diferente, con la primera columna mostrando a los coeficientes negativos y la segunda columna a los positivos.

1.3. miR-10b y miR-21 son predictores universales de PAM50

Pasando de las observaciones generales a los predictores específicos, como siguiente paso se buscaron predictores selecionados para un mismo gen del PAM50, en los cuatro subtipos. Dicha intersección contiene relaciones con 13 transcritos y 2 miRNAs. Los CpGs sólo se conectan con el mismo gen en máximo dos subtipos. Hay 24 sitios en esta situación, de los cuales 15 son comunes al subtipo HER2E y algún otro subtipo; mientras que 9 sitios se conectan con HER2 sin estar en el mismo cromosoma, **tabla 5**. Por su parte, los transcritos involucrados están físicamente ligados al gen que predicen, insinuando que se trata de asociaciones triviales, **tabla 6**. Mientras ELP2 y SLC39A6 están codificados en cadenas opuestas del mismo locus, el resto de los pares son contiguos; cuando el 84.77 % del total de asociaciones con transcritos, considerando todos los modelos, conectan cromosomas diferentes.

Tabla 5: Sitios CpG seleccionados recurrentemente como predictores del mismo gen del PAM50.

data1	data2	PAM50	predictor	PAM50 chr	predictor chr
LumA	LumB	BCL2	cg25373630	18	18
Basal	LumA	CDC6	cg04243581	17	7
Basal	LumA	CXXC5	cg01008405	5	5
HER2E	LumB	ERBB2	cg03322619	17	2
HER2E	Basal	ERBB2	cg03414134	17	3
HER2E	LumB	ERBB2	cg03519711	17	1
HER2E	Basal	ERBB2	cg03903647	17	4
HER2E	Basal	ERBB2	cg04053045	17	2
HER2E	Basal	ERBB2	cg07239593	17	1
HER2E	LumB	ERBB2	cg12009778	17	6
HER2E	Basal	ERBB2	cg16021483	17	12
HER2E	LumA	ERBB2	cg22627876	17	1
HER2E	LumB	ESR1	cg26496205	6	16
HER2E	normal	FOXA1	cg00157855	14	7
LumB	normal	FOXA1	cg23768510	14	6
HER2E	normal	FOXA1	cg24665320	14	13
HER2E	normal	KRT17	cg05551922	17	3
LumB	normal	KRT17	cg16259464	17	2
LumA	LumB	MAPT	cg19108736	17	17
LumA	normal	PGR	cg10647644	11	5
LumA	LumB	SLC39A6	cg14058239	18	18
HER2E	LumA	SLC39A6	cg23152248	18	2
HER2E	LumA	SLC39A6	cg24830619	18	1
LumA	normal	UBE2C	cg09010017	20	8

Respecto a los miRNAs, las relaciones recurrentes entre subtipos únicamente implican a dos miRNAs: miR-10b y miR-21. Ambos tienen interacciones regulatorias conocidas con los genes del

PAM50, específicamente miR-21 actúa sobre *BCL2* [180], *MYC* [181], *EGFR* [182] y *HER2* [183], además de haber sido predicho como regulador de *ESR1* [184] y *FOXA1* [185]. Mientras tanto, miR-10b regula a *CDC6*, a *EGFR* y *SFRP1* [186,187] y su selección se extiende al tejido normal.

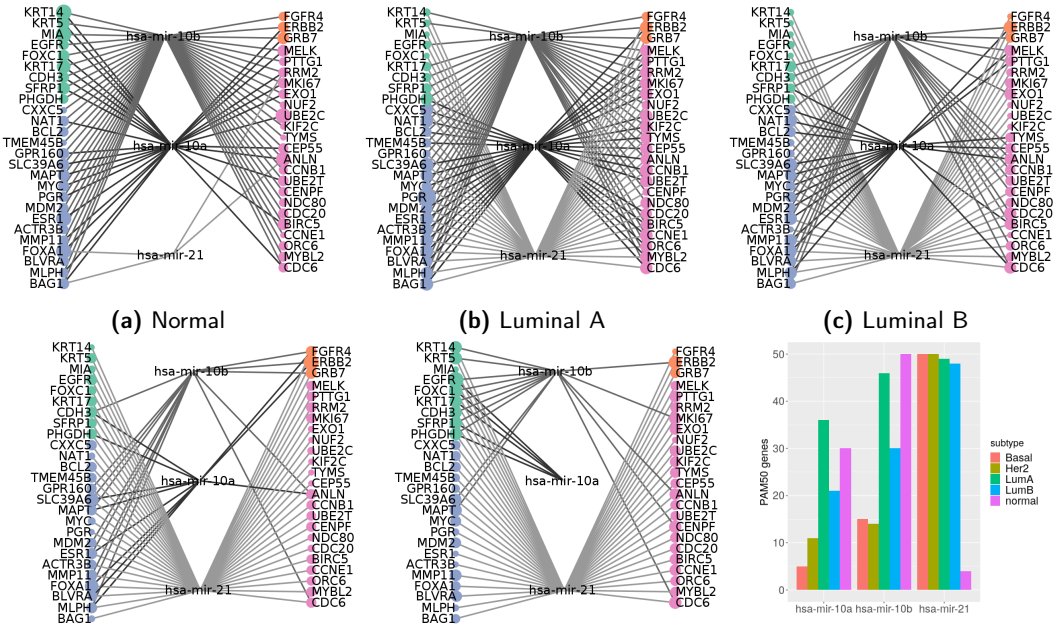
Tabla 6: Transcritos seleccionados como predictores del mismo gen del PAM50 en los 4 subtipos del cáncer de mama. Las columna inicio y fin indican las respectivas distancias que separan al predictor del gen del PAM50.

PAM50	predictor	cromosoma	inicio	fin
ANLN	KIAA0895	7	26060	23825
BIRC5	PGS1	17	-164463	-199478
ORC6	VPS35	16	33511	8876
CDC6	WIPF2	17	68575	20521
MYBL2	IFT52	20	76082	69196
BLVRA	COA1	7	150224	77625
CDC20	SZT2	1	-30901	-91044
CCNB1	CDK7	5	-67660	-99185
MDM2	CNOT2	12	-1434825	-1509449
MDM2	YEATS4	12	-551533	-545326
MDM2	SLC35E3	12	62052	51580
SLC39A6	ELP2	18	-20913	-51172
ERBB2	MIEN1	17	-40582	-106
GRB7	MIEN1	17	9431	16759

Adicionalmente, miR-10b y miR-21 fueron recurrentemente seleccionados en los modelos de distintos genes. Esto es llamativo porque, en general, un predictor sólo participa en un modelo por subtipo, 93.45 % de los CpGs, 74.24 % de los transcritos y 81.37 % de los miRNAs, no se comparte entre ningún par de modelos (o genes del PAM50). Aunque hay otros predictores compartidos entre genes, ninguno vincula a todos los genes de la firma, ni lo hace de manera recurrente. De hecho, hay una sub-representación de sitios CpGs entre los predictores compartidos y una sobre-representación de transcritos y miRNAs.

Entre los modelos del tejido normal, miR-10b fue seleccionado como predictor de los 50 genes, mientras que miR-21 sólo aparece en 4 modelos. Por el contrario, miR-21 está conectado con la mayoría de los genes en los cuatro subtipos, mientras que miR-10b es escasamente seleccionado, y se puede ver una pérdida de la conectividad conforme se avanza hacia los subtipos con peor prognosis, **figura 11**. Además, surge miR-10a, otro miembro de la familia miR-10, codificado río arriba de miR-10b, cuya similitud en secuencia [188] y ubicación justifica que se hable de miR-10a/b.

De manera coherente con el reemplazo de un miRNA por el otro, miR-21 está significativamente sobre-expresado en los cuatro subtipos, mientras que miR-10b está sub-expresado, tal y como



(a) Normal **(b)** Luminal A **(c)** Luminal B

(d) HER2E **(e)** Basal **(f)**

Figura 11: Los predictores que conectan a la mayoría de los genes en el PAM50 pasan de miR-10a/b en el tejido normal a miR-21 en los subtipos del cáncer de mama. (a–e) Los genes del PAM50 están a los lados, coloreados de acuerdo a su patrón de expresión: verde para basal, naranja para asociado a HER2, azul para luminal y rosa para pro-proliferativo. El tamaño de los nodos refleja el número de predictores selecccionados. Los predictores están en medio y las aristas los conectan con los genes que predicen. El gradiente de color de las aristas sólo facilita la visualización. (f) Genes del PAM50 conectados con cada miRNA por subtipo.

se había observado previamente [83]. A su vez, mir-10a está sub-expresado en los subtipos basal y HER2E, pero ligeramente sobre-expresado en los luminales, sin alcanzar un p-value significativo en luminal B. Así, un patrón normal de miR-10b explica -parcialmente- la expresión de todos los genes del PAM50 y cuando pierde expresión, miR-21 toma su lugar, sugiriendo un mecanismo tipo interruptor. En este sentido, la introducción de un antagonista de miR-21 en células MCF7 y MDA-MB-231 implantadas en ratones, inhibe el crecimiento del tumor y la migración celular [189]. Al mismo tiempo, la sub-expresión de miR-10b-3p, una de las formas maduras de miR-10b, está involucrada en la tumorigénesis, al permitir la sobre-expresión de los reguladores del ciclo celular BUB1, PLK1 y CCNA2 [190]. Adicionalmente, la sobre-expresión de miR-10b participa en la metástasis, al inhibir a HOXD10 y favorecer la invasión [83], lo que resalta la necesidad de construir modelos de las distintas etapas del cáncer.

El comportamiento tipo hub[3] de estos miRNAs coincide con observaciones previas de nuestro grupo, reportando a los miRNAs como nodos altamente conectados [191] y esenciales para la

[3]Los hubs son los nodos con mayor centralidad de grado en una red libre de escala. Se habla de un comportamiento tipo hub por tratarse de nodos muy conectados, pero sin una evaluacion estricta del tipo de red.

cohesión de las redes transcripcionales [84]. Cuando se representa la salida de los modelos de un subtipo como una red, donde cada predictor está unido al gen del PAM50 cuya expresión predice, puede verse que la desconexión de miR-10a/b y miR-21 no rompe la red, pero hace que sea necesario recorrer decenas de predictores para ir de un gen a otro. Esto significaría que la expresión de la firma PAM50 está coordinada en los cuatro subtipos y el tejido normal, con y sin miRNAs, pero miR-10a/b y miR-21 aceleran el flujo de información.

Considerando que cada miRNA puede regular cientos de genes [192], el comportamiento de miR-10a/b y miR-21 no es extraordinario. Sin embargo, como ya se mencionó, la acción de estos miRNAs sobre la firma PAM50 sólo ha sido descrita para algunos genes, apuntando a un posible efecto indirecto. Al respecto, los coeficientes asociados tienen valores dos órdenes de magnitud por debajo del rango general de los miRNAs y las distribuciones de coeficientes son significativamente diferentes, **figura 12**. La potencial regulación indirecta de los 50 genes en los cuatro subtipo y la transición entre tejido normal y cáncer ligada a la alteración de su expresión, perfila a mi10a/b y miR21 como posibles reguladores maestros del cáncer de mama [71]. Sin embargo, los modelos escuetos no seleccionan específicamente reguladores, sino que los recuperan de listas pre-filtradas [121–123], limitando lo que se puede aseverar sobre miR-10a/b y miR-21, a simples predictores universales de los genes en el PAM50.

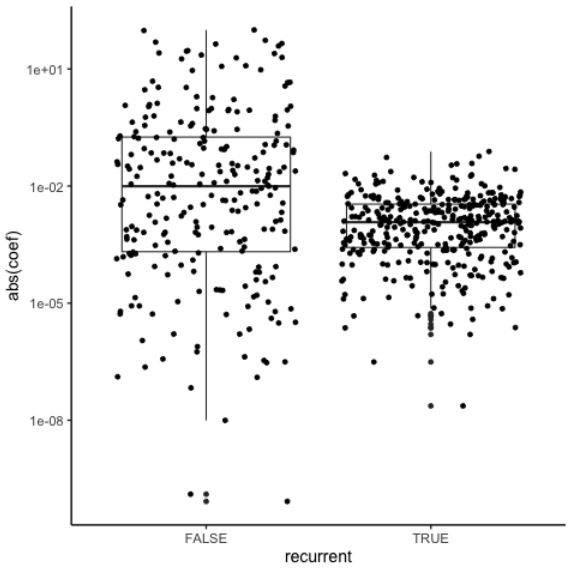


Figura 12: Contraste entre los coeficientes que acompañan a miR-10b y miR-21(*recurrent*=*TRUE*) y el resto los miRNAs.

1.4. El enriquecimiento funcional de los modelos cambia entre subtipos

El último análisis que se corrió sobre los predictores seleccionados por gen y subtipo, es un enriquecimiento funcional, que permite verificar si los grupos de predictores correlacionados funcionan en conjunto. Los genes del PAM50 tienen modelos con más predictores en los subtipos en los que están sobre-expresados. Los resultados del enriquecimiento contra la ontología de procesos biológicos están representados en la red de la **figura 13**.

En el subtipo basal, los modelos de *FOXC1* y *ANLN* están enriquecidos de transcritos ligados al estímulo por TGFβ y la protección del telómero, respectivamente. Ninguno de los dos genes del PAM50 está anotado dentro de estás funciones, aunque *FOXC1* y TGFβ comparten dianas [193]. En el caso de HER2E, sólo *ORC6* está enriquecido y el proceso es el ensamblado de sinapsis. El subtipo luminal B se conecta con la división celular a través de *MELK* y *CCNB1*, participantes conocidos de esta función y los modelos del tejido normal muestran enriquecimiento de distintos aspectos de la división celular. El subtipo luminal A es el que exhibe mayor enriquecimiento, lo que podría explicarse por la mayor selección de transcritos. Los genes cuyos modelos se asocian a funciones en este subtipo se conectan con su función anotada, que en la mayoría de los casos es la división celular.

CCNB1, *MKI67*, *UBE2C* y *MELK* están ligados a las mismas funciones en subtipos diferentes; pero tal enriquecimiento depende de predictores diferentes, sugiriendo que los procesos de división celular involucrados son robustos y pueden depender de genes diferentes entre los subtipos. *UBE2C* y *MELK* exhiben aristas comunes entre el subtipo luminal A y el tejido normal, cuyo enriquecimiento desaparece si se remueven los predictores comunes, indicando que el proceso depende de su mecanismo normal.

Por el contrario, *ANLN*, *CEP55*, *KRT17*, *MYBL2* y *ORC6* se conectan con procesos diferentes en los distintos subtipos o en el tejido normal. Específicamente, los predictores de *CEP55* en el luminal A son excluyentes de los predictores seleccionados en el tejido normal, sugiriendo que esta proteína centrosomal tiene un rol divergente en el cáncer, como se había reportado previamente [194]. El mismo razonamiento aplica para *KRT17* y *ORC6*, aunque se esperaría que la queratina 17 también se conecte con los receptores de tirosina en el subtipo luminal A [195]. Los predictores no son significativamente excluyente en el caso de *ANLN* y *MYBL2*, lo que implicaría divergencia funcional, al participar de procesos diferentes en compañía de los mismos genes.

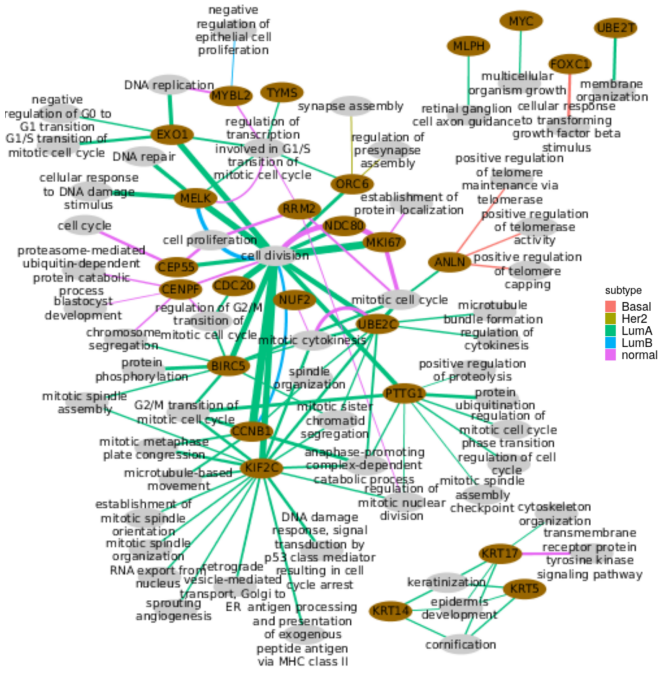


Figura 13: Enriquecimiento functional de los transcritos seleccionados como predictores del PAM50. Las elipses marrones son genes del PAM50 y las grises procesos biológicos enriquecidos entre sus predictores. El color de las aristas refleja subtipo donde se observó el enriquecimiento. El grosor de las arista indica la cantidad de predictores involucrados en el proceso.

1.5. Conclusión

La selección de predictores que ejercen los modelos de red elástica permite:

- examinar efectos globales de la metilación del DNA, la expresión de transcritos codificantes de TFs y de miRNAs sobre la expresión de los genes del PAM50
- encontrar reguladores potenciales, de entre los que destaca el interruptor formado por miR-10b y miR-21
- señalar posibles eventos de convergencia y divergencia funcional

Es notorio sin embargo, que luminal A, el subtipo con mayor número de muestras, también tiene la mayor cantidad de predictores seleccionados y que lo opuesto pasa con HER2E. Al verificar la comparabilidad entre subtipos, repitiendo el ajuste con subconjuntos aleatorios de 40 muestras, se mantienen las distribuciones de RMSE y los patrones generales de predictores seleccionados por subtipo. Sin embargo, no se puede descartar la sub-selección de predictores por la baja

representación del subtipo enriquecido de HER2, lo que afecta la capacidad del análisis para aseverar que hay divergencia funcional entre subtipos.

Además, este estudio hereda limitaciones del método en su forma más simple, como la carencia de valores de significancia ligados a los predictores y la penalización uniforme de ómicas inherentemente distintas. Ambos problemas tiene soluciones posibles, como la inferencia de p-values empíricos [132] y el ajuste de coeficientes de penalización diferentes para cada ómica [134], que también podría ayudar con el desequilibro en la representación de cada subtipo [127]. Finalmente, tendrían que probarse distintos valores del parámetro de mezcla, que en esta ocasión se fijó en 0.5.

2. Redes multi-omicas de la regulación transcripcional del cáncer de mama

La selección de predictores de los modelos de red elástica permite la construcción de redes que exponen las relaciones entre la expresión genética de los genes modelados y los niveles moleculares analizados. Sin embargo, estas redes no llegan a satisfacer el primer objetivo específico, que es encontrar las interacciones entre la expresión genética y las tres capas de regulación, porque no abarcan ómicas completas y para hacerlo se tendría que ajustar un modelo para cada uno de los 17077 genes codificantes o enfocarse en un subconjunto de interés, como los genes diferencialmente expresados o los del PAM50. Este problema se puede superar construyendo redes de información mutua con las matrices de cada subtipo. El modelo de redes sugerido sólo se infiere una vez por subtipo, sin la necesidad de elegir algún parámetro particular, lo que reduce el tiempo de cómputo. Además, la información mutua captura relaciones no lineales y no necesita de normalizaciones específicas.

Así, se construyeron redes de MI para cada subtipo del cáncer de mama y para el tejido normal, integrando los datos de metilación de sitios CpG, la expresión de los transcritos codificantes y de los miRNAs. Entre los transcritos codificantes está la tercer capa regulatoria de los transcritos codificantes de TFs. Tal como en la aproximación anterior, se incluyeron todos los predictores, sin restringir los CpGs que se pueden vincular a un transcrito por su distancia en el genoma ni los TFs por la presencia de su motivo de unión.

Las redes pasaron por dos procesos de poda. Primero se sometieron a un filtro por MI, respetando las diferencias entre los distintos tipos de arista, de modo que la significancia de todas

las aristas está determinada por un p-value que varía dependiendo del tipo de nodos conectados, **tabla 7**. Como segundo paso, se extrajeron los subgrafos asociados a funciones, reteniendo únicamente los nodos involucrados en procesos con puntaje de GSEA significativo y, que como conjunto funcional están sobre-representados en la red, además de sus primeros vecinos. La red del tejido normal, que no tiene transcritos diferencialmente expresados al tomarse como punto de referencia para cada subtipo, surge del mapeo de todos los procesos identificados en los subtipos. Las redes finales sólo incluyen interacciones CpG-transcrito, transcrito-transcrito y miRNA-transcrito, ligadas a funciones alteradas por la expresión particular del subtipo, de manera que inciden sobre el tercer objetivo específico, al exhibir funciones alteradas que se conectan con los tres niveles funcionales.

La hipótesis es que los nodos con capacidad regulatoria podrían estar influyendo en las funciones con las que se conectan a través de los transcritos. Para que esto pase, los reguladores tienen que covariar con sus dianas, con una señal estadística superior a la del resto de pares covariables, como podrían ser los genes co-regulados. Sin embargo, hay más mecanismos de regulación que los incluidos, como modificaciones de las histonas y RNAs no codificantes largos, que podrían diluir la señal de CpGs, TFs y miRNAs, dificultando su detección. Contrastando las interacciones de las redes construidas con las relaciones regulatorias conocidas, se calcula un porcentaje de positivos verdaderos, que oscila entre 1.67 % y 11.47 % dependiendo de la red. El resto de las aristas puede representar regulaciones nuevas, asociaciones indirectas o incluso ruido. La posibilidad de encontrar relaciones regulatorias nuevas parece especialmente viable en el caso de los TFs, donde las interacciones validadas dependen de experimentos de ChIP-seq en tejidos que no necesariamente reflejan lo que pasa en los subtipos del cáncer de mama.

2.1. Los parámetros topológicos cambian entre niveles funcionales

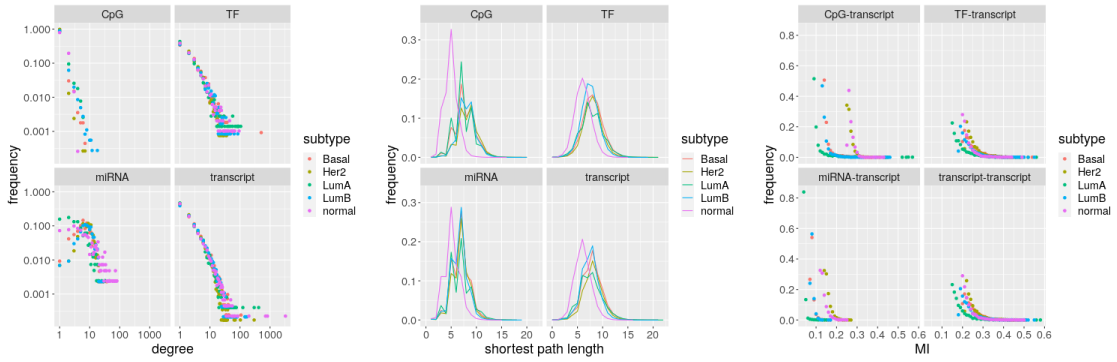
Una vez teniendo redes, el primer paso en el análisis de resultados es su caracterización topológica. Al tratarse de redes multipartitas, donde los nodos pueden ser CpGs, transcritos que codifican dianas regulatorias o TFs, y miRNAs, los parámetros explorados se agrupan de acuerdo al tipo de nodo, buscando características distintivas. Los descriptores más simples, del número de nodos, aristas y, sus umbrales de significancia pueden consultarse en la **tabla 7**.

Las redes están formadas por cientos de componentes aislados y el grado promedio es de tres en las distintas redes; pero las distribuciones varían entre ómicas. Incluso los transcritos y los transcritos que codifican para TFs tienen distribuciones de grado diferentes, a pesar de provenir

Tabla 7: Características de los redes de cada subtipo. La cantidad de interacciones validadas aparece entre paréntesis

	Basal	HER2E	LumA	LumB	normal
aristas					
CpG-transcrito	2456 (554)	3847 (88)	1932 (536)	4334 (708)	4732 (28)
TF-transcrito	2735 (5)	2498 (2)	1686 (5)	2746 (1)	2544 (14)
miRNA-transcrito	3483 (167)	3889 (226)	2065 (111)	4074 (201)	4953 (284)
transcrito-transcrito	4189	4523	2276	4709	5088
nodos					
sitios de CpG	2254	3769	1553	3638	3863
transcritos	4567	6356	2834	5235	4733
transcritos TFs	658	748	375	618	684
miRNAs	433	432	408	433	14
procesos biológicos	109	119	34	123	128
p-values					
CpG-mRNA	$\leq 10^{-6}$	$\leq 10^{-6}$	$\leq 10^{-6}$	$\leq 10^{-6}$	$\leq 10^{-6}$
mRNA-mRNA	$\leq 10^{-6}$	$\leq 10^{-6}$	$\leq 10^{-6}$	$\leq 10^{-6}$	$\leq 10^{-6}$
miRNA-mRNA	$\leq 10^{-3}$	$\leq 10^{-4}$	$\leq 10^{-4}$	$\leq 10^{-3}$	$\leq 10^{-4}$

de la misma plataforma de medición. La centralidad de grado de los sitios CpG y los transcritos siguen una distribución esperada en redes biológicas, donde la mayoría de los nodos tienen grado uno y conforme aumenta el grado, baja rápidamente la cantidad de nodos. Un promedio entre redes de 89.42 % de los CpGs se conectan sólamente con otro nodo, así que no participan en la comunicación de la red, más que como punto de partida o llegada. Por el contrario, el pico de la distribución de grado de los miRNAs, está cerca de 10, como se puede ver en la **figura 14a**.



(a) **(b)** **(c)**

Figura 14: Parametros de los distintos tios de nodos: (a) grado, (b) caminos más cortos, (c) información mutua

Esta manera en que se conectan los nodos repercute en la cantidad de transcritos accesibles

partiendo desde un miRNA o desde un CpG. El largo de los caminos más cortos, **figura 14b**, mide cuántos nodos hay que recorrer para llegar del nodo A al B y se calculó partiendo de los nodos de cada tipo, hasta los transcritos, que al traducirse en proteínas son lo que ejercen las funciones. Por un lado, el promedio de transcritos inaccesibles desde un CpG es de 32.23 %, mientras que el promedio para miRNAs es 19.71 %. Adicionalmente, las distribuciones son significativamente diferentes tanto al comparar entre ómicas, como al contrastar las redes de los subtipos con la del tejido normal. El cambio en la posición del pico de la distribución de los cuatro subtipos, respecto al tejido normal, sugiere una pérdida de la comunicación entre ómicas en el cáncer de mama, coherente con la perturbación de la metilación del DNA y la expresión [1, 2].

Finalmente se verificó la diferencia, previamente reportada, entre distribuciones de MI de los distintos tipos de aristas [84]; añadiendo una diferencia significativa entre los subtipos, **figura 14c**. De las tres medidas comparadas, este es el único caso en que los valores que acompañan a los transcritos codificantes de TFs siguen la misma distribución que el resto de los transcritos. Además es llamativo el rango tan bajo en que caen los valores de MI de las interacciones con miRNAs y la dispersión entre redes de las interacciones con CpGs.Aunque la información mutua entre transcritos y miRNAs es menor a lo observado por Drago-García et al., coincide con lo justificado por Setty et al. con la acción simultánea de otros reguladores y un efecto regulatorio modesto [122], que es superado por el efecto de los CpGs [126].

Considerando la diversidad de tamaños de muestra a partir de los cuales se infieren las redes, el cálculo de MI entre miRNAs y transcritos se repitió 100 veces con subconjuntos de las muestras. La **figura 15** muestra una tendencia de valores de sub-muestreo más ruidosos conforme la información mutua baja, que le da confiabilidad al top de mayor MI.

2.2. Los sitios de CpGs son exclusivos de los procesos, mientras que miRNAs y TFs se comparten

Con el objetivo de traducir las observaciones topológicas a una perspectiva más funcional, se calculó el índice de Jaccard entre cada par de procesos representados en una misma red, distinguiendo entre los distintos tipos de reguladores potenciales, para saber qué reguladores conectan funciones diferentes. A la par de la medida continua, cada regulador potencial fue clasificado como compartido o exclusivo, dependiendo de su conexión con uno o más procesos.

De manera consistente con lo observado mediante el grado y los caminos más cortos, los CpGs tienden a unirse exclusivamente con un proceso, mientras que los TFs y los miRNAs se comparten,

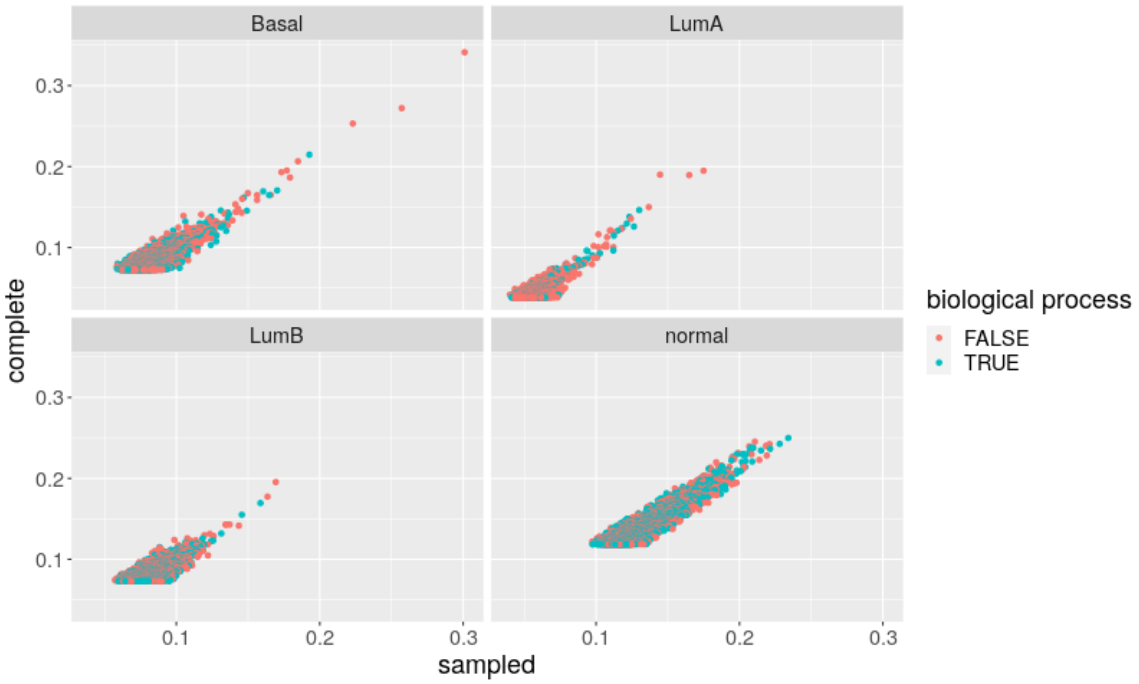


Figura 15: MI calculada con el total de muestras contra el promedio calculado con subconjuntos de 45 muestras tomadas al azar

figura 16a. Esto coincide con los mecanismos de regulación conocidos, pues se espera que los CpGs tengan un efecto local sobre los genes vecinos [35] y que tanto TFs, como miRNAs, actúen sobre distintas dianas dispersas por el genoma [92]. La exclusividad de los sitios CpG podría servir para el monitoreo temprano [196] de la perturbación de funciones específicas, como podría ser el control del daño a DNA en el subtipo HER2E, asociado a 19 sitios exclusivamente conectados con este proceso biológico, en este subtipo de prognosis intermedia. Sin embargo, para avanzar en ese sentido, habría que evaluar la predictibilidad de los sitios, lo que implica más y distintas pruebas; pero exhibe la utilidad de los análisis de integración multómica como generadores de hipótesis. Por su parte, la conexión de TFs y miRNAs con procesos diferentes, sugiere efectos pleiotrópicos como resultado de su alteración dirigida o esporádica.

Las distribuciones de índices de Jaccard son significativamente diferentes, **figura 16b**, y muestran que los CpGs se comparten menos que los otros reguladores potenciales. Contra lo sugerido por el grado y la clasificación en reguladores exclusivos o compartidos, el pico de las distribuciones de TFs supera el asociado a miRNAs, implicando más TFs compartidos más frecuentemente. Resaltan las distribuciones de CpGs del subtipo HER2E y de TFs del basal, por el rango que ocupan, respecto a las otras redes.

Los valores del índice de Jaccard ligados a los sitios CpG en el subtipo enriquecido de HER2,

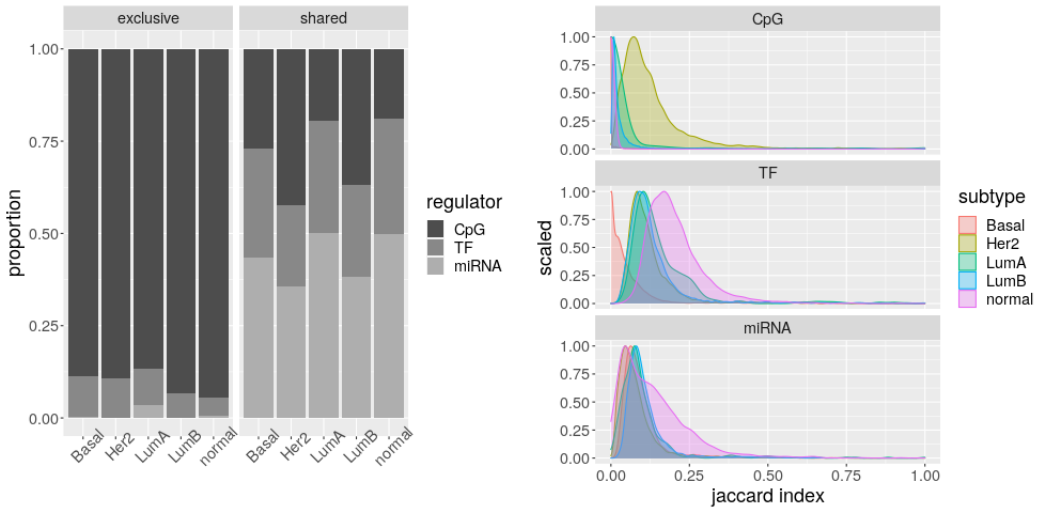


Figura 16: Conexión de los reguladores potenciales con los procesos biológicos: (a) Proporción de reguladores exclusivos y compartidos entre procesos. (b) Distribuciones del índice de Jaccard por tipo de nivel funcional.

indican más CpGs compartidos entre procesos biológicos, en el subtipo con la mayor variabilidad de patrones de metilación [1]. Revisando puntualmente esta red, hay 2112 sitios CpG compartidos, ligeramente concentrados en los cromosomas 1 y 17. Mientras se ha reportado que el cromosoma 1 está severamente afectado por la metilación diferencial [53], la amplificación característica del 17 no parece aportar al patrón. Del total de 1576 genes afectados por los sitios compartidos, sólo 76 llegan a amplificarse en conjunto con HER2 y sólo 361 tienen evidencia de ser regulados por AR, que interactúa con la amplificación de HER2 [19].

De manera opuesta, la distribución asociada a los TFs en la red del subtipo basal implica una proporción elevada de reguladores exclusivos respecto al resto de las redes. Esto no es consecuencia de una menor cantidad de transcritos que codifican para TFs, puesto que la representación de estos reguladores es similar, ni de los procesos enriquecidos, ya que todos aparecen en otras redes, salvo 6. En cambio, la explicación podría estar en la accesibilidad de los promotores, que divide al cáncer de mama en basal y no basal [3], con un firma pro-metástasica de cromatina abierta [197]. Integrando transcriptoma, proteoma y CNA, ya Koh et al. habían reportado que el subtipo basal es impulsado por TFs, aunque esto no aclara porque los procesos comparten transcritos que codifican TFs [149].

2.3. La proporción de reguladores potenciales en las redes cambia entre subtipos y respecto al tejido normal

Con la intención de explorar más a fondo las diferencias entre reguladores potenciales, se calculó su abundancia en cada proceso representado en las distintas redes. La proporción de reguladores asociados a los subtipos y el tejido normal puede apreciarse en la **figura 17**. Ahí es evidente el aumento de nodos CpG en las redes de cáncer de mama, respecto al tejido normal y una disminución de TFs y miRNAs. El gráfico del luminal A es menos ruidoso porque el subtipo tiene menos procesos representados en la red.

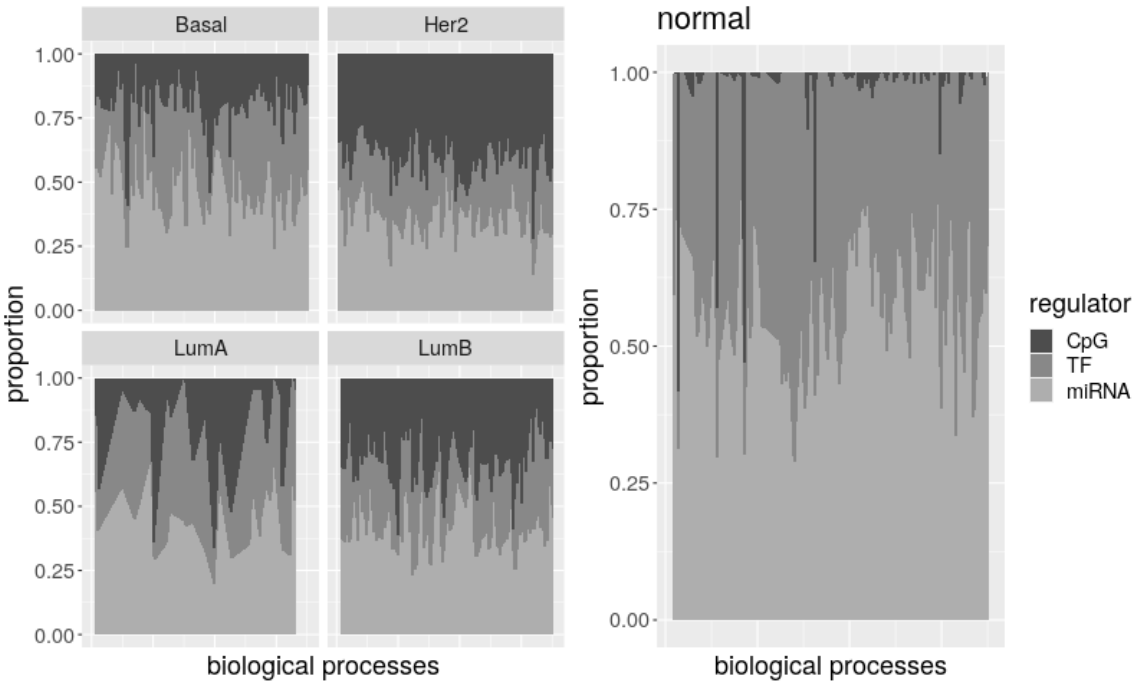


Figura 17: Abundancia de los reguladores potentiales por subtipo. Cada columna es un proceso diferente. En conjunto, todos los procesos representados en una red muestran que tan común es cada uno de los tres tipos reguladores potenciales

Considerando cada proceso, hay un enriquecimiento significativo de nodos CpG en casi todos los procesos del subtipos basal, HER2E y luminal B, y en menos de la mitad de los presentes en la red luminal A. Al mismo tiempo, la sub-representación de TFs y miRNAs es significativa en más de la mitad de los procesos en las redes HER2E y luminal B, y en 20 y 33 % de los procesos en basal y luminal A.

Si los reguladores potenciales encontrados efectivamente actúan sobre los transcritos funcionales, estos resultados implicarían que los mecanismos de regulación transcripcional y postrans-

cripcional son dominados en el cáncer de mama por la regulación epigenética.

2.4. Las interacciones potencialmente regulatorias del tejido normal no aparecen en las redes del cáncer de mama

Habiendo visto que la representación de las capas regulatorias completas cambia entre redes, se planteó el propósito de buscar interacciones específicas, para saber si los pares regulador-diana se conservan o son subtipo-específicos. Con esto en mente se calculó el índice de Jaccard de aristas comunes a cada par de redes. Del total de 176 procesos biológicos encontrados en alguna red, 86.36 % aparecen en dos o más redes y por lo tanto, pueden compartir aristas.

Como puede verse en la **figura 18**, las relaciones con miRNAs tienen los menores índices de Jaccard; mientras que los vínculos con TFs y CpGs ocupan rangos similares, pero siguen distribuciones diferentes. Globalmente, las interacciones con reguladores potenciales se repiten poco entre las redes. Más aún, la mayoría de los procesos encontrados en las redes de un subtipo y el tejido normal, tienen un índice de 0, indicando que las aristas de la red normal no están en las redes del cáncer de mama, con unas pocas excepciones que involucran transcritos codificantes de TFs. Considerando que la metilación preserva la identidad celular [43], el reemplazo de las interacciones normales con los sitios CpG por otras diferentes, apuntaría a una degeneración de los patrones definitorios de la glándula mamaria en procesos biológicos específicos, que pueden ser fácilmente señalados con las redes de MI.

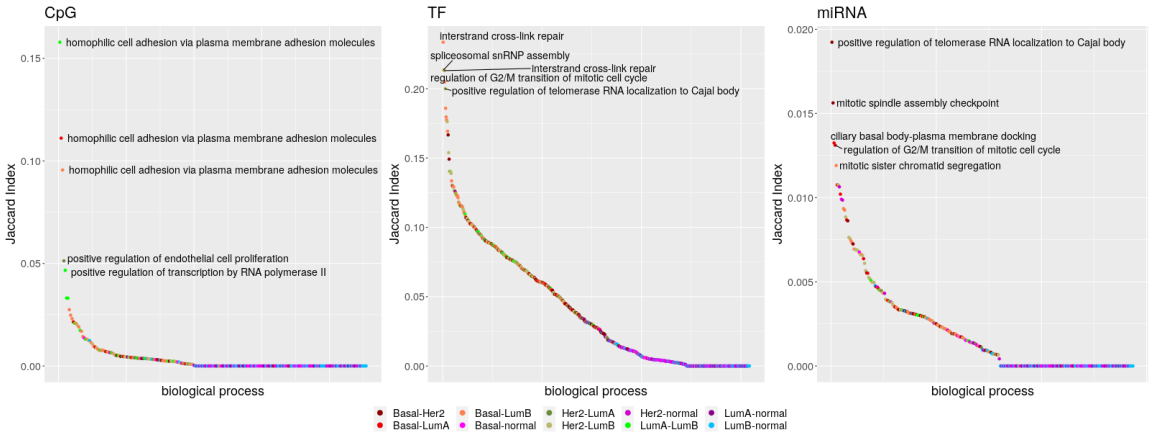


Figura 18: Aristas compartidas entre cada par de redes asociadas al mismo proceso biológico, en subtipos diferentes. Los puntos representan el valor del índice de Jaccard y se ordenan del mayor al menor. El color indica qué redes son contrastadas: los tonos de rojo involucran a la red basal, verde seco a HER2E, verde brillante a los luminales y, la comparación con las redes del tejido normal está en tonos de purpura y azul.

Los cinco procesos con los índices más altos por nivel funcional son señalados en la **figura 18**. Distintos puntos representan el mismo proceso, pero contrastando redes diferentes, como sucede en el gráfico correspondiente a los sitios CpG con la adhesión homofílica a través de moléculas de la membrana plasmática. La localización del RNA de la telomerasa (hTR) en los cuerpos de Cajal tiene el mayor índice de Jaccard, por las relaciones con miRNAs compartidas entre los redes basal y enriquecida de HER2 y, el quinto mayor índice respecto a los enlaces con TFs en HER2E y luminal B.

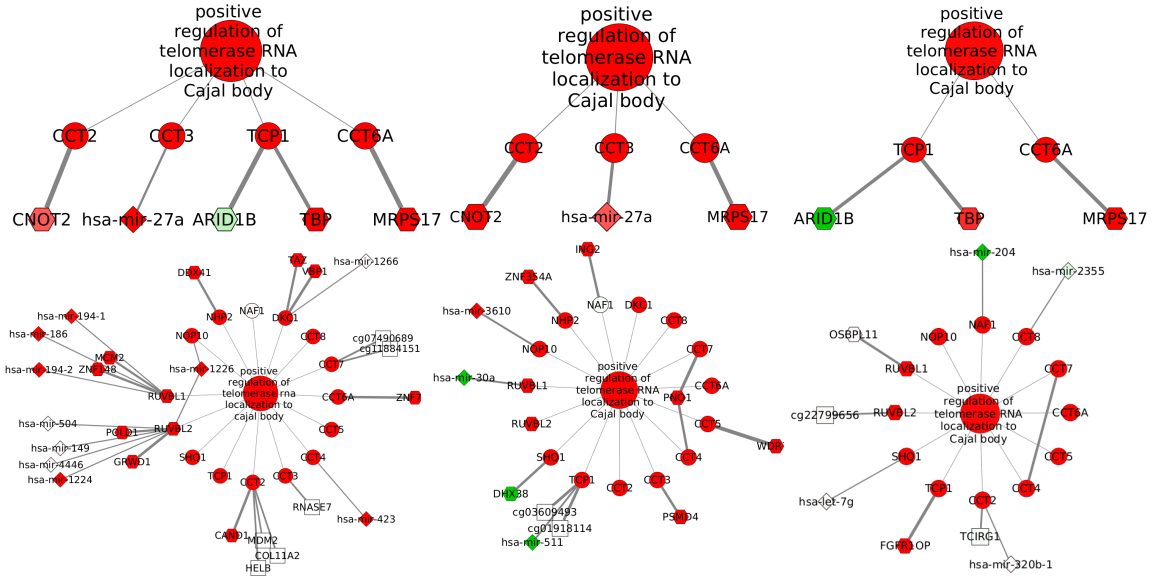


Figura 19: Redes asociadas con la localización de la telomerasa a los cuerpos de Cajal. De izquierda a derecha, subgrafos con aristas compartidas (arriba) y exclusivas (abajo) de los subtipos basal, enriquecido de HER2 y luminal B. Los nodos en rojo tienen valores de expresión diferencial o del puntaje normalizado de GSEA positivos, mientras que el verde representa valores negativos. La transparencia de los nodos muesta la significancia estadística. Los nodos del proceso biológico y de los transcritos son circulos, los transcrito codificantes de TFs son hexágonos, los miRNAs rombos y los sitios CpG cuadrados. Cuando se conoce el gen afectado por el CpG, su nombre está en la etiqueta del nodo, de otro modo aparece el identificador de la sonda CpG. El grosor de las aristas indica el valor de MI

Dado que los cuerpos de Cajal están implicados en la biogénesis de la telomerasa, la localización del hTR en los mismos está asociada a la división ilimitada del cáncer [198]. La **figura 19** muestra que el índice, en este caso particular, depende de una pocas interacciones compartidas entre redes pequeñas. La conexión entre TCP1 y MRPS17, común a los tres subtipos, podría ser un artilugio por la cercanía física de los dos genes. Como los nodos tienen valores similares de expresión diferencial y el proceso tiene puntajes GSEA parecidos en los distintos subtipos, pero las redes son diferentes, podría tratarse de una convergencia de esquemas regulatorios diferentes en un mismo desenlace. Sería importante encontrar este tipo de fenómenos, porque la manera en

que un tumor gana una firma de expresión crea susceptibilidades diferentes, como pasa con los tumores compatibles con la expresión HER2E que no tienen la amplificación del receptor [199].

2.5. Conclusión

A pesar de la disponibilidad relativamente amplia de herramientas de integración multi-ómica, las interacciones entre distintos niveles moleculares suelen mantenerse fuera de la discusión, que más bien se enfoca en nodos particulares o grupos de tumores. Las redes descritas en esta sección demuestran las posibilidades de este tipo de métodos de integración, que permiten hacer observaciones al nivel de capas regulatorias completas, como el grado de los CpGs, del contraste entre redes enteras, como el endemismo de las interacciones potencialmente regulatorias de la red del tejido normal, y puntuales, como los subgrafos de la localización del hTR en los cuerpos de Cajal.

La información mutua tiene un gran potencial en la integración multi-ómica, al carecer de restricciones en cuanto a los rangos dispares de las ómicas, porque depende de distribuciones de probabilidad, que entonces vuelven el tamaño de muestra la preocupación principal. Sin embargo, la repetición del cálculo de MI con subconjuntos de los datos, apoya el uso cuidadoso de datos con pocas muestras, ya que el orden de las interacciones se mantiene y sólo se vuelve ruidoso conforme baja el valor de MI. La gran desventaja de las redes presentadas es la imposibilidad de distinguir relaciones indirectas [155], ya que la diferencia entre distribuciones de MI de los distintos tipos de aristas, impide la aplicación del algoritmo estándar que desarma los triángulos en la red [142]. La alternativa de la representación de las redes como tensores [200], aún requiere una implementación accesible.

3. Funciones multi-ómicas de los subtipos del cáncer de mama

Los modelos de red elástica y las redes de información mutua, en conjunto, inciden sobre los cuatro objetivos específicos. Mientras las redes elásticas explican la expresión del PAM50 con la información multi-ómica, las redes de información mutua exponen las relaciones entre los distintos niveles funcionales y señalan funciones asociadas a esa interacción. En ambos casos, el contraste entre subtipos y con el tejido normal, toca el cuarto objetivo específico. Sin embargo, en ningún momento se satisface como tal la identificación de las funciones más afectadas por dichos niveles del tercer objetivo, tan vago como pueda sonar. Reconociendo desde

un principio la complementariedad entre la aproximación de redes y la multivariada escueta, el esquema de trabajo de la **figura 7** proyectaba una integración, ahora de los resultados de las dos aproximaciones, que se concreta en esta sección.

El flujo de trabajo que se construyó en torno a los resultados del SGCCA resuelve la disyuntiva dejada por las secciones 1 y 2. Mientras las redes elásticas implican un modelo por gen, demandando tiempo y capacidad de cómputo; las redes de MI conforman un sólo modelo por subtipo, pero incluyen aristas indirectas y sin dirección. Por su parte, el SGCCA sólo tiene que ajustarse una vez y, mediante un análisis de enriquecimiento funcional, permite reconocer las tan deseadas funciones más afectadas por la interacción multi-ómica. Esto es posible porque el SGCCA selecciona los predictores que mejor explican la varianza de las ómicas individuales y la covarianza con las otras ómicas. De modo que las funciones enriquecidas entre los predictores seleccionados serían las más afectadas por la relación entre ómicas y sus alteraciones. Aunque la diferencia con las funciones de la sección 2 es casi semántica, es relevante. No es lo mismo una función que podría ser regulada por, digamos CpGs, entre otros muchos mecanismos de regulación; que una función conectada con los CpGs más correlacionados con los niveles funcionales estudiados. Con el fin de lograr un discurso más sucinto, a partir de este momento, aparece en el texto el término "funciones multi-ómicas" para referirse a la funciones ligadas con los predictores seleccionados por el SGCCA, aunque la asociación aún tendría que comprobarse.

Una vez que se ajustó la penalización LASSO y se obtuvieron los predictores seleccionados y sus coeficientes para cada uno de los subtipos y el tejido normal, se descartaron los predictores con estabilidad por debajo del 70 %. Este paso es necesario por el uso del LASSO en lugar de la penalización de red elástica, pero es inherente a las implementaciones actuales del SGCCA. Aunque el filtro reduce severamente la cantidad de predictores seleccionados, **figura 20**, le da fiabilidad al enriquecimiento funcional, que entonces parte únicamente de los predictores verdaderamente correlacionados [120]. Tanto por el filtro por estabilidad, como por el problema del LASSO

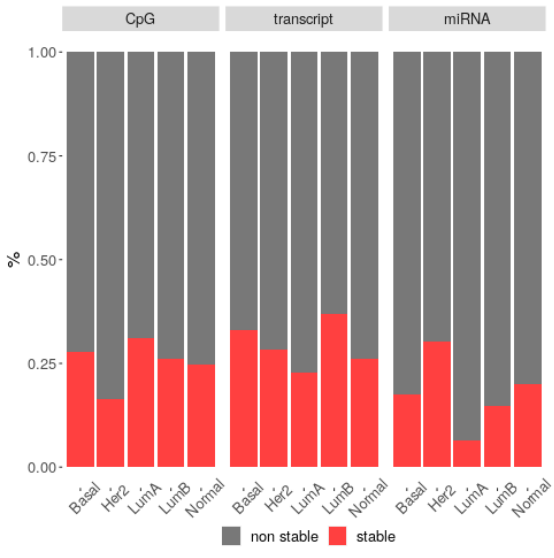


Figura 20: Proporción de predictores estables por ómica

con los grupos correlacionados dentro de cada
ómica, se puede aseverar que en realidad hay más CpGs, transcritos y miRNAs interdependientes de lo que se reporta y más funciones asociadas; pero el conjunto aquí descrito funciona como un punto de partida confiable.

Teniendo conjuntos de predictores interconectados confiables, que además explican distintas señales en la varianza de los datos, porque la selección es por componente latente, se procede al enriquecimiento funcional y a la inferencia de redes por función. Mientras el análisis de las funciones identificadas permite elucubrar sobre las características comunes y particulares de los subtipos; las redes constituyen modelos multi-ómicos de los mecanismos que la integración multi-ómica prometía descubrir [97]. Yendo por partes, a continuación se expone cada uno de estos resultados.

3.1. La funciones multi-ómicas son diferentes entre subtipos y con el tejido normal

La selección de predictores varía, como se había visto en las dos aproximaciones anteriores, con el tamaño de las ómicas, con más de 300 CpGs, 10 transcritos y menos de 5 miRNAs seleccionados por componente. Además hay una diferencia evidente en la selección de sitios CpG en el tejido normal, que tiene más de 400 CpGs por componente, lo que se opone a la sobre-representación de nodos CpG observada en las redes de la sección anterior.

El exceso de CpGs, respecto a los otros predictores, quizá explica que no haya transcritos ni miRNAs seleccionados en los cinco conjuntos de datos, pero sí seis sitios de CpG, que afectan la expresión de *MAPK8IP3*, *AFAP1*, *LFNG* y *VSTM2B*. Por otro lado, los transcritos que se seleccionan repetitivamente entre los componentes de un mismo subtipo están asociados o con sus características o con el cáncer de mama en general. Los tres transcritos más seleccionados para el subtipo basal son *MCL1*, *CTNNA1* y *NOTCH3*. *MCL1* es un miembro anti-apoptótico de la familia *BCL2* requerido para las células troncales mamarias [201] que se suele encontrar sobre-expresado en el subtipo [202]; mientras que la catenina alfa 1 se ha postulado como un supresor de tumores en los tumores basales negativos para E-caderina [203], y *NOTCH3* promueve la transición epitelial-mesenquimal [204].

El top de transcritos seleccionados en el subtipo HER2E, *CEACAM5*, *ACACA* y *PGK1*, también cumplen con esto. Los tumores del subtipo tienden a ser positivos a CEACAM5 [205], por lo que esta molécula de adhesión se ha propuesto como diana para terapias con anticuerpos bi-

específicos [206]. Por su parte, los inhibidores de la acetil-CoA carboxilasa, ACACA, interfieren con la biosíntesis de lípidos y el efecto Warburg de células MCF-7 que sobre-expresan HER2 [207]. Finalmente, la proteína PGK1 se sobre-expresa en el subtipo [208] y ha sido asociada con la infiltración de macrofagos y la estratificación de pacientes [209].

Es llamativo que la familia de miRNAs let-7 también esté dentro de los predictores más seleccionados entre los subtipos basal, HER2E y luminal B, además del tejido normal, ya que estos miRNAs regulan la vías de señalización de JAK-STAT3 y Myc, incidiendo sobre las células troncales y la metástasis [210]

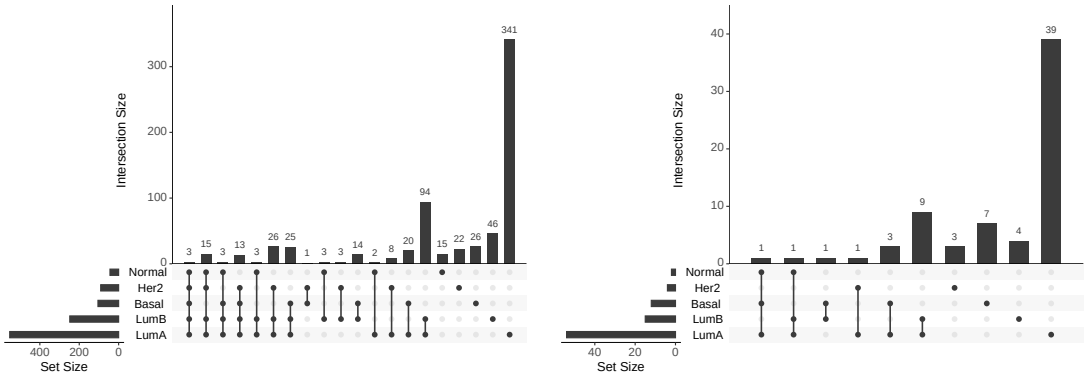


Figura 21: Intersección de (a) procesos biologicos y (b) vías KEGG enriquecidas en los subtipos y el tejido normal

Habiendo inspeccionado la salida de cada SGCCA, se realizaron los análisis de enriquecimiento funcional contra la ontología de procesos biológicos y las vías de KEGG, usando todos los predictores, no sólo los transcritos. En total se encontraron 683 procesos y 69 vías, aunque muy pocas funciones aparecen en los datos de los cuatro subtipos. Como muestra la **figura 21**, la mayoría de las funciones son exclusivas de un subtipo o comunes sólo a un par de ellos. En otras palabras, las funciones asociadas simultáneamente con la metilación del DNA y la expresión de transcritos y miRNA cambian entre subtipos.

Únicamente hay tres procesos biológicos significativamente sobre-representados en los cuatro subtipos y el tejido normal: desarrollo de la nefrona metanéfrica (GO:0072210), desarrollo del metanefros (GO:0001656) y especificación de patrones (GO:0007389). Dado que GO:0072210 es parte de GO:0001656, pueden tomarse como un sólo proceso. Aunque el enriquecimiento de procesos del desarrollo de los riñones pueda resultar sospechoso, ambos órganos requieren la morfogénesis ramificada de ductos [211]. Si bien los mecanismos precisos aún se desconocen,

se cree que los procesos responsables del desarrollo normal pueden ser cooptados durante la progresión del cáncer, lo que explicaría esta sobre-representación recurrente [212].

Los procesos comunes a los subtipos y el tejido normal además podrían compartir el circuito que conecta CpGs, transcritos y miRNAs, por lo que surge la pregunta, las funciones encontradas varias veces, ¿involucran a los mismos predictores e interacciones?

3.2. Los predictores responsables del enriquecimiento funcional cambian entre subtipos

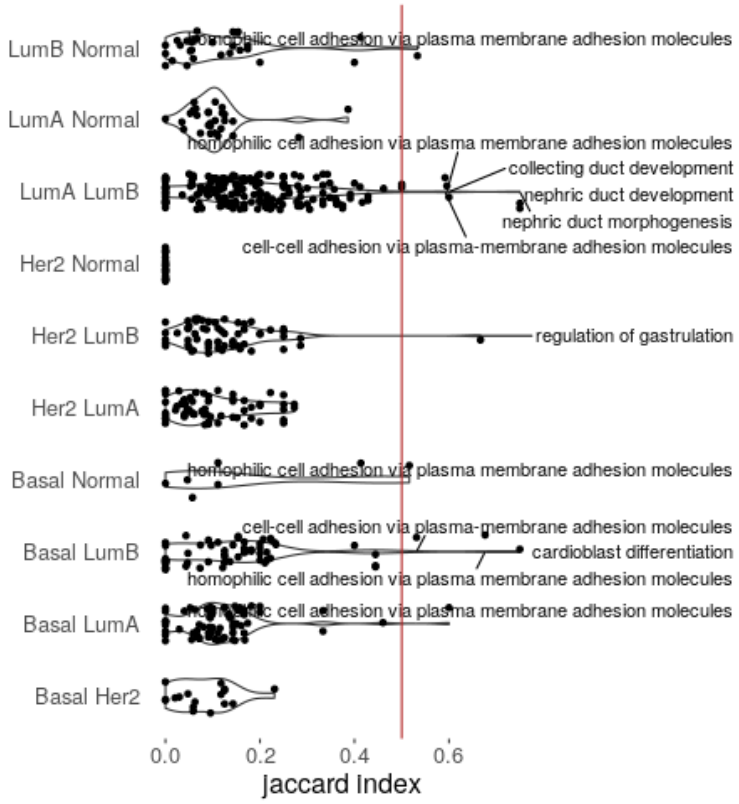


Figura 22: Predictores conectados con la misma función entre subtipos o en el tejido normal

El primer requisito para mantener el circuito de CpGs, transcritos y miRNAs, sería que los predictores responsables del enriquecimiento funcional sean los mismos o casi los mismos. Para verificar si este es el caso, se calculó el índice de Jaccard entre las distintas instancias en que una función aparece sobre-representada, obteniendo las distribuciones de la **figura 22**. Las distribuciones del índice de Jaccard bastan para asegurar que el circuito CpGs-transcritos-miRNAs es distinto en cada instancia de la función, ya que los predictores involucrados son diferentes. Solamente hay siete procesos biológicos enriquecidos en un par de subtipos, que comparten por

encima del 50 % de los predictores. Cinco de esos procesos se relacionan con el desarrollo y los otros dos estás vinculados con la adhesión celular. Estas son las funciones que podría compartir interacciones entre instancias.

Si este índice refleja la similitud entre subtipos respecto a la covariación de CpGs, transcritos y miRNAs, la distancia de HER2E es intrigante. Aunque podría ser ocasionada simplemente por el número de muestras, también podría reflejar la baja correlación que los patrones de metilación del DNA tienen con este subtipo [1]. Por el contrario, las funciones con más predictores en común están enriquecidas en los dos subtipos luminales.

3.3. Las funciones se conectan a través de los predictores dentro de cada subtipo

Al revisar los predictores responsables de la sobre-representación se descubrió que hay funciones diferentes asociadas a los mismos componentes latentes. Esto implica cierta comunicación entre las funciones, que se conectan a través de sus predictores asociados y que entonces, estarían representadas en el mismo componente de la red multi-ómica.

Analizando cada subtipo y al tejido normal por separado, las funciones fueron agrupadas de acuerdo a los componentes latentes en los que aparecen singificativamente sobre-representadas. Como ejemplo, el agrupamiento obtenido para el subtipo HER2E se muestra en la **figura 23**. Ahí se pueden ver once grupos y seis funciones que no pertenecen a ningún grupo porque involucran componentes latentes diferentes. Tomando las etiquetas más grandes como guía, los grupos púrpura, naranja y fucsia se relacionan con el desarrollo de las estructuras del riñón. Los grupos verde y azul en la parte inferior están conectados con el desarrollo del tejido conectivo. Los nodos en rosa se refieren a distintos procesos de la morfogénesis, mientras que los nodos en amarillo aluden al desarrollo de estructuras reproductivas. Los pequeños grupos en café y verde pálido se asocian con el músculo cardiaco y finalmente, los grupos en el centro en verde brillante y naranja pálido están ligados con el metabolismo y llenos de funciones exclusivas del subtipo. Las funciones con más predictores no forman parte de ningún grupo.

El agrupamiento exhibe un problema a considerar al analizar las funciones exclusivamente enriquecidas en un subtipo. Mientras las funciones exclusivas pueden revelar explicaciones mecanísticas de las alteraciones específicas de los subtipos, si las funciones están agrupadas con otras no-exclusivas y más representadas, puede tratarse de un artilugio. Así, los grupos pueden explicar algunos enriquecimientos inesperados, como el encontrado en el subtipo luminal A para

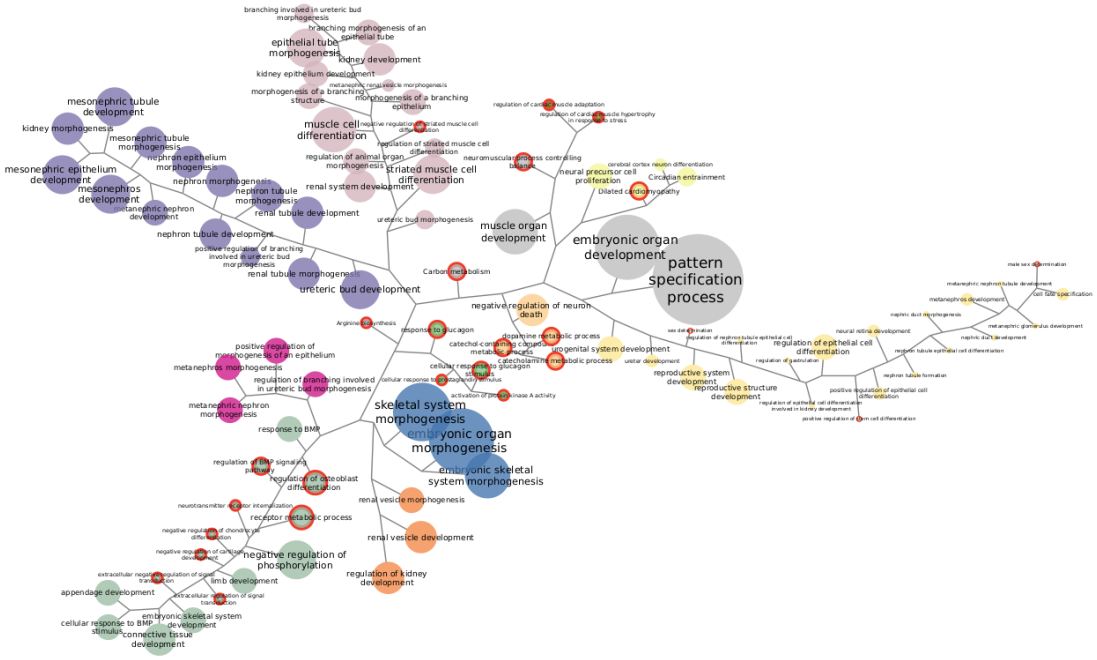


Figura 23: Funciones sobre-representadas en la salida de SGCCA del subtipo enriquecido de HER2. Tanto las vías de KEGG como los procesos biologicos están representados juntos. Los nodos del mismo color son funciones ligadas exactamente a los mismos componentes latentes. Los nodos en gris no pertenecen a ningún grupo. El tamaño de los nodos y sus etiquetas reflejan la cantidad de predictores sosteniendo el enriquecimiento. Las funciones exclusivamente encontradas en un subtipo se resaltan con un borde rojo.

la adicción a morfina. La adicción a la morfina ya se había observado sobre-representada entre genes regulados por metilación [56], pero depende de predictores que covarían con los predictores responsables de la interacción entre receptores y marcadores de la matriz extracelular, sugiriendo que la co-variación podría estar impulsando la sobre-representación de la adicción.

3.4. Las funciones exclusivas señalan un vínculo entre el subtipo basal y la invasión, y una perturbación de los procesos de modificación del DNA

Con el fin de elucidar si las funciones encontradas enriquecidas de manera exclusiva en un subtipo efectivamente explican las características de los subtipos, se analizó la sobre-representación de las categorías de proceso biológicos GOslim, que pueden apreciarse en la **figura 24**, y las clases de las vías KEGG.

No se encontró algún sesgo de las clases KEGG, pero si hay un enriquecimiento de las categorías: organización de los componentes celulares en los datos del subtipo basal, establecimiento de la localización en el luminal A y procesos metabólicos del DNA en el tejido normal. Hay siete

procesos biológicos sosteniendo la sobre-representación del subtipo basal, cinco están relacionados con la extensión del axón y se agrupan con el sexto, que es la regulación de la magnitud del crecimiento celular. El séptimo proceso es la organización de las fibras de colágeno, que sin estar ligada a los mismos componentes latentes, suma a vínculo del subtipo con la invasión tumoral.

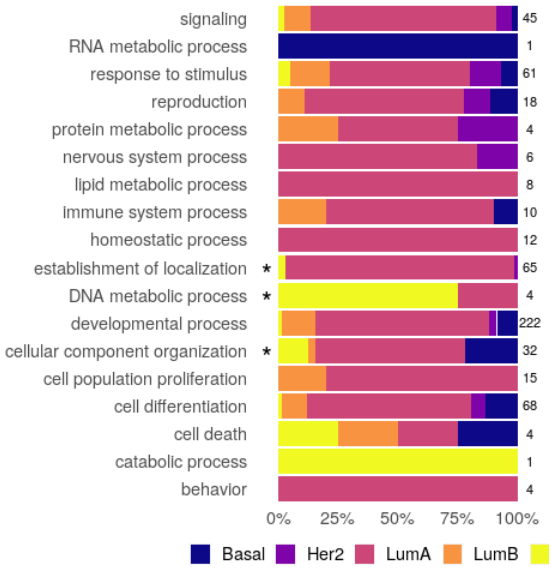


Figura 24: Sesgo de las funciones exclusivas. El asterisco marca las categorías con sobre-representación significativa (Prueba de Fisher con ajuste de Holm, p-value < 0.05)

En el caso del subtipo luminal A, hay 62 procesos detrás de la sobre-representación, ligados con transporte y secreción y estructurados en 32 grupos diferentes, de los que es difícil concluir algo más dada la abundancia de información. Por su parte, la sobre-representaciòn del tejido normal es interesante porque depende de tres procesos enriquecidos en los mismos componentes latentes: la alquilación del DNA, la metilación del DNA y la demetilación, quizás implicando una perturbación en los subtipos

3.5. Ejemplos de redes

Partiendo del agrupamiento se eligieron funciones de cada subtipo para explorar más fondo. Se le dio prioridad a las funciones enriquecidas exclusivamente en un subtipo y agrupadas con otras funciones exclusivas, pensando en que dos funciones asociadas al mismo conjunto de predictores covariantes no necesariamente involucran a los mismos actores. Para verificar si la comunicación entre funciones agrupadas es directa, se construyeron redes de información mutua con todos los conjunto de predictores co-seleccionados asociados al grupo. Las redes se pasaron por un filtro para mantener sólo las interacciones potencialmente regulatorias y se puso el foco sobre los componentes ligados a una función de interés

La intuición es que los predictores co-seleccionados con nodos funcionales pueden estar influyendo en la función y al tratarse de CpGs, transcritos que codifican para TFs y miRNAs, regularla. Se anticipan dos posibles escenarios: 1) componentes disconexos, cada uno con su pro-

pio conjunto de reguladores potenciales o 2) funciones que se conectan a través de predictores comunes, cuyo potencial como reguladores los vuelve de interés médico. Ambos escenarios están ejemplificados a continuación mediante redes de los cuatro subtipos y el tejido normal.

También se intentó la inferencia de redes para la adhesión celular, al tratarse de un proceso común a tres de los subtipos y al tejido normal, con un índice de Jaccard superior a 0.5, que indicaría una estructura medianamente compartida entre las distintas redes y que además siempre se encuentra sobre-representada en un conjunto único de componentes latentes -no forma parte de ningún grupo-. Sin embargo, el filtro de MI produce redes mínimas, cuya comparación se vuelve trivial.

3.5.1. Señalización por HIF-1 en el subtipo basal

La señalización del factor inducible de la hipoxia 1 (HIF-1) es una de las vías de KEGG enriquecidas exclusivamente en la salida de SGCCA basal. HIF-1 es un regulador maestro de la homeostasis del oxígeno, que induce la transcripción desde más de 100 elementos de respuesta a hipoxia [213]. La señalización HIF-1 se activa en los tumores por las condiciones de hipoxia, pero

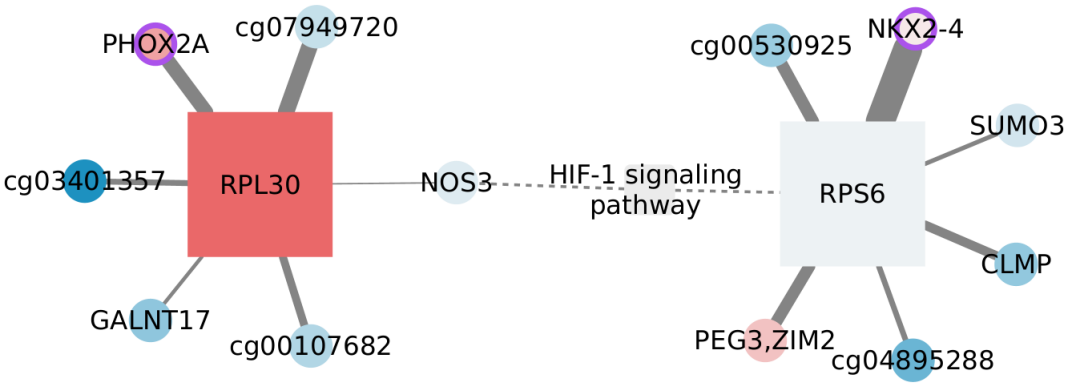


Figura 25: Predictores conectados con la señalización HIF-1 en el subtipo basal. Los circulos representan CpGs y los cuadrados transcritos. Cuando es posible los CpGs son identificados con el simbolo del gen afectado, de otro modo, aparece el ID de la sonda. Los tonos de rojo indican el nivel de sobre-expresión o sobre-metilación respecto al tejido normal; mientras que los tonos azules representan valores por debajo de lo esperado. El tamaño de los nodos refleja su grado. Los nodos cuya proteína es un factor transcripcional tienen un borde púrpura. El peso de las aristas representa la información mutua de los nodos. Las aristas discontinuas conectan las redes de MI con las funciones.

también por factores independientes al oxígeno, como las mutación de *TP53* y *BRCA1* [214], que son frecuentes en este subtipo [1]. La **figura 25** muestra la red correspondiente.

La señalización por AMPK no está en el mismo grupo que nuestra vía de interés, sino que se asocia con un subconjunto de los componentes latentes, lo que coincide con el intercambio que la vías tendrían durante la reprogramación del metabolismo del cáncer [215]. Sin embargo, el filtro de MI separa la vías en componentes disconexos.

Igualmente, sólo los predictores *NOS3* y *RPS6*, que participan en la vía, sobrepasan el filtro. Es importante recalcar que el enriquecimiento de la vía no depende solamente de dos predictores, sino estos son los únicos que forman relaciones lo suficientemente pesadas para permanecer después del filtro. Además de *NOS3* y *RPS6*, hay otros dos nodos ligados a la vía sin participar activamente. PEG3 se sobre-expresa en ratones una vez superada la hipoxia, y SUMO3 modifica la proteína HIF-1, afectando su estabilidad [216]. Mientras PEG3 y SUMO3 le dan credibilidad a la red, la naturaleza de las aristas es peculiar. La red está formada de conexiones entre sitios CpG y transcritos de proteínas ribosomales. Dado que los CpGs no forman parte del mismo cromosoma del que surge los transcritos, se descarta una relación directa. Para evaluar la posibilidad de que se trate de relaciones indirectas, impulsadas por transcritos vecinos a los CpGs, se estimó la información mutua correspondiente, aunque los transcritos no estén entre los predictores co-seleccionados. Los valores resultantes son menores de lo admitido por el filtro de MI y tampoco superan a los vínculos con sitios CpG. Entonces, los efectos indirectos no llegan a justificar completamente la relación de un transcrito con un CpG físicamente lejano.

Finalmente hay que mencionar que, en general, los nodos tienen patrones de expresión equivalentes a los observados en el tejido normal, de modo que el puntaje de GSEA no es significativo y se espera una vía de señalización normal. Por otro lado la vía es exclusiva del subtipo y su falta de representación entre los predictores del tejido normal, podría indicar un cambio en la relación entre ómicas.

3.5.2. Regulación positiva de la regulación de la diferenciación de las células troncales en el subtipo enriquecido de HER2

Las células troncales del cáncer son consideradas responsables de la recaída y la metástasis, y su mantenimiento se ha asociado con variantes de la proteína HER2, observadas en el subtipo [217], lo que podría justificar la asociación encontrada con la regulación de la diferenciación de células troncales. Aunque se trata de un proceso exclusivo, también hay enriquecimientos similares en

los otros tres subtipos. La **figura 26** muestra los primeros vecinos de los nodos funcionales y todos los procesos y vías conectados con los mismo predictores.

El CpG del factor transcripcional SOX9 es el único predictor ligado al proceso que pasa el filtro de MI y vincula en la red diferentes funciones relacionadas con la determinación celular y sexual. Siendo sus vecinos, los sitios CpG de los factores transcripcionales LHX1 y OSR1 los que conectan a la mayoría de las funciones. Ninguna de las interacciones del grafo ha sido reportada de manera directa en la literatura, pero por ejemplo CRISP3 sí se relaciona con la determinación sexual, específicamente con el tracto reproductivo masculino y los espermatozoides [218]; además de estar de-regulado en el cáncer prostático [219]. De manera similar, DNAH10 está ligado con la morfología del flagelo de los espermas [220], SHC3 participa en el desarrollo de la retina [221], CR1L en la activación de los linfocitos B [222] y quizá en las lesiones renales [223] y, la expresión diferencial de ITGB6 se ha reportado en el cáncer de prostata [224].

Además hay nodos cuya función está regulada por la metilación del DNA, dándole cierto

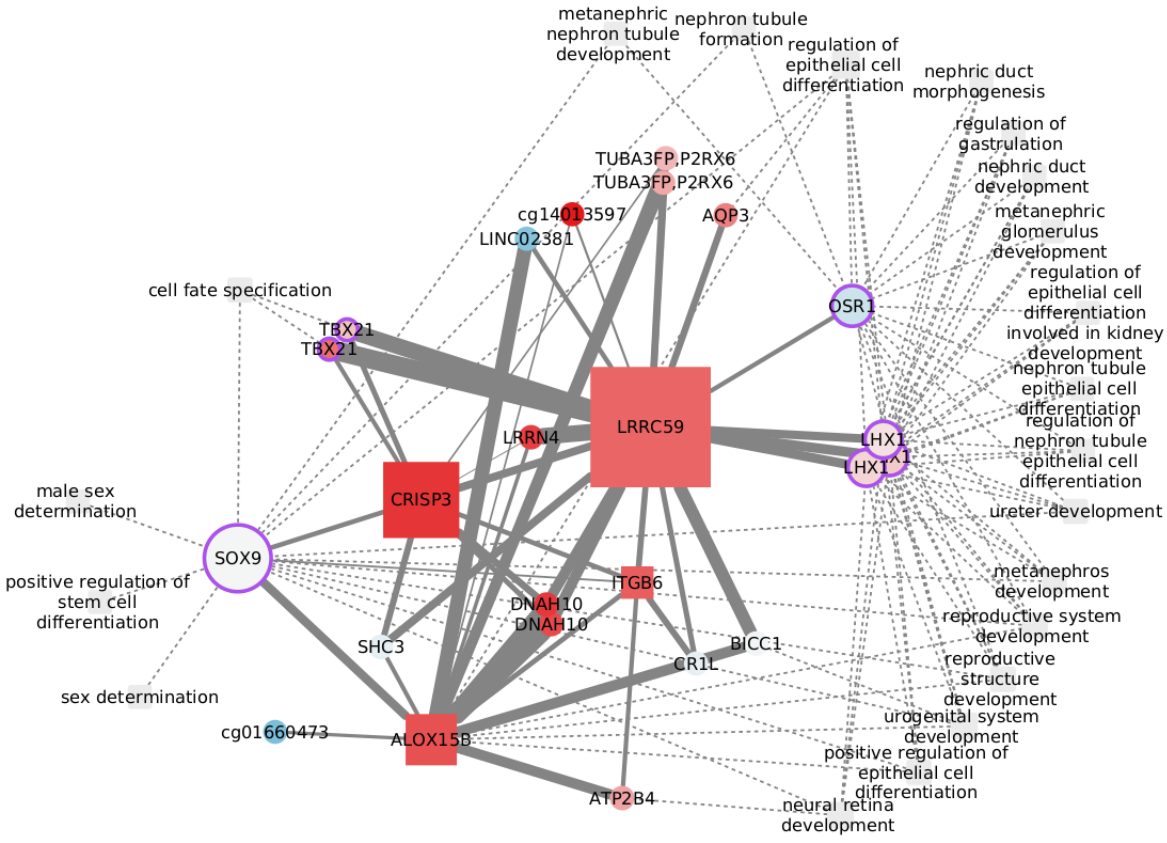


Figura 26: Predictores conectados con la regulación de la diferenciación de las células troncales en el subtipo HER2E. El tamaño de los nods refleja su *betweenness*

respaldo a su representación por nodos CpG. Tal es el caso de CR1L, cuya alteración epigenética está ligada con Alzheimer y demencia [225]; de DNAH10, conectada con un fenotipo metilador en carcinomas renales [226]; y del lncRNA LINC02381, que funciona como supresor de tumores en tumores colorectales, donde es silenciado por metilación [227].

Aunque la opción más evidente para investigar la comunicación entre procesos serían los factores transcripcionales mencionados, también puede ser importante analizar el efecto de nodos como *ALOX15B*, *CRISP3* y *LRRC59*, que tienen una centralidad *betweenness* elevada y entonces controlan el flujo de información en la red.

3.5.3. Señalización Ras en el subtipo luminal A

La señalización Ras es una de las muchas vías encontradas de manera exclusiva en los datos del subtipo luminal A. La vía afecta aspectos del cáncer como la proliferación celular, la supervivencia, la migración y la diferenciación. Aunque su activación no se observa de manera muy frecuente en este subtipo, se ha reportado como un indicador de mala prognosis en tumores luminales [228]. Revisando los puntajes GSEA, la señalización Ras estaría sub-activada respecto al tejido normal

El único nodo funcional que resiste el filtro de MI es la subunidad beta de la proteína G, GNB2, que enlaza la vía con distintos sitios CpG de la comunicación celular y la función del cerebro, a través del sensor de calcio SYT13.

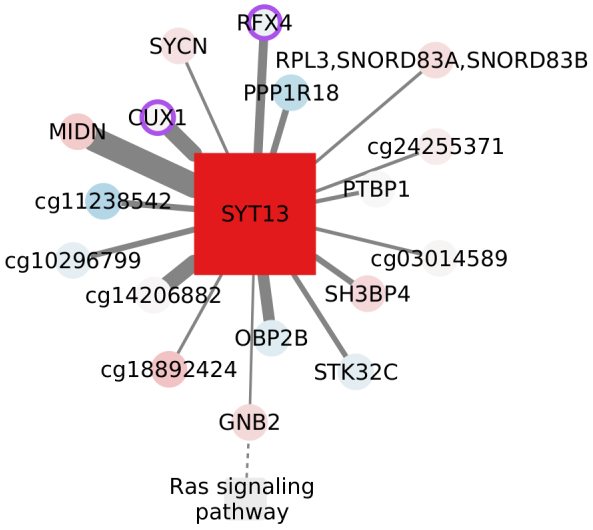


Figura 27: Predictores conectados con la señalización Ras en el luminal A. El tamaño del nodo representa su grado.

Los genes regulados por los CpGs incluyen la cinasa de expresión neuronal, *STK32C*; el factor transcripcional del desarrollo temprano del cerebro, *RFX4* y tres genes con funciones por demostrar: *MIDN*, que facilitaría la unión de las cinasas; *OBP2B*, que participaría en la unión de moléculas volátiles; y *SYCN*, con un rol predicho en la exocitosis. Respecto a las aristas, el vínculo con CUX1 concuerda con su cooperación con la mutante Kras-G12V en el cáncer de pulmón [229] y la sobre-expresión de *PTBP1* co-ocurre con las mutaciones de KRAS

en cáncer de colon [230]. Finalmente, la unión con la proteína de internalización del receptor de transferrina, SH3BP4, ha sido predicha previamente [231].

De nuevo, la red de la **figura 27** enlaza transcritos y sitios de CpG dispersos en el genoma. En este caso se ha propuesto que la señalización RAS controla la metilación aberrante [232], lo que podría causar esta abundancia de nodos CpGs, aunque parece improbable y en cambio falta una exploración puntal del modelo de red.

3.5.4. Regulación negativa de la vía de señalización Wnt en el subtipo luminal B

En las salidas del SGCCA, la regulación negativa de la vía Wnt sólo se encontró sobre-representada para el subtipo luminal B; pero hay otras funciones relacionadas con Wnt en el subtipo luminal A. Además de que esta función está bien establecida como importante para el cáncer de mama, al influir en la proliferación de tumores, la metástasis, la regulación del sistema inmune, la resistencia a tratamiento y el mantenimiento de las células troncales [233]; los genes de la vía están inesperadamente hipermetilados en una fracción de los tumores luminales B [1]. Las otras funciones en la **figura 28** no surgen de los mismos componentes latentes, sino de un subconjunto de ellos. El grafo sólo muestra los primeros vecinos de los nodos funcionales.

Como se podía esperar, la regulación negativa de la señalización Wnt y la regulación de Wnt comparten predictores. SOX9, de nuevo representado por su CpG, está en la intersección con distintos procesos del desarrollo, aunque además hay caminos indirectos conectándolos. Dado que COL4A2, subunidad del colágeno IV, y CEACAM6, molécula de adhesión celular, son dianas de SOX9 [234] y que SOX9 coopera con GLI3 [235], estas tres aristas son respaldadas por evidencia externa. Aunque no se puede hablar de mecanismos regulatorios del cáncer de mama, el enlace de COL4A2 con NFATC4 puede explicarse con la capacidad inhibitoria que COL4A2 ha mostrado sobre la translocación nuclear de NFATC4 en cardiomiocitos [236]. De manera similar, la conexión entre COL4A2 y IGF2 puede provenir de la desregulación coordinada de ambas proteínas extracelulares ante enfermedades que incluyen EMT [237]. Mientras que IGFBP4 estimula la sobre-expresión de SOX9 en células estromales de la médula ósea [238] e impide la diferenciación inducida por BMP2 [239].

En resumen, hay razones biológicas sólidas para esperar una dependencia estadìstica entre los nodos conectados. La cuestión es saber el efecto que tienen estas interacciones en la señalización Wnt, en la progresión del subtipo luminal B. Específicamente se resalta al nodo con mayor *betweenness*, el componente 2 del complejo de exocitosis, relacionado funcionalmente con Wnt

como efector de la señalización Hedgehog [240], y adicionalmente asociado con la metastasis y distintos tipos de cancer [241–243], pero no con el cáncer de mama.

3.5.5. Metilación del DNA en el tejido normal adyacente

La metilación está exclusivamente sobre-representada en el tejido normal, pero se discute por su relevancia para el cáncer [28]. Más aún, a diferencia de los otros ejemplos, esta red contiene microRNAs, incluyendo a let-7a-2.

Por consistencia se colorearon los nodos, aunque, al tomar al tejido normal como valor de referencia, se usaron los valores de cambio obtenidos del contraste con el subtipo basal. Los tumores basales tienen un puntaje GSEA significativo del proceso biológico y exhiben la mayor hipometilación del cáncer de mama [1]. Hay que enfatizar sin embargo, que la metilación del DNA no es una función enriquecida en los datos basales, de modo que el circuito de la **figura 29** no tiene porque corresponder con el subtipo.

Aunque ninguna de las aristas ha sido descrita, hay un par de nodos con relaciones indirectas con la maquinaria de metilación del DNA. AKAP8L interactúa con el complejo de metiltrans-

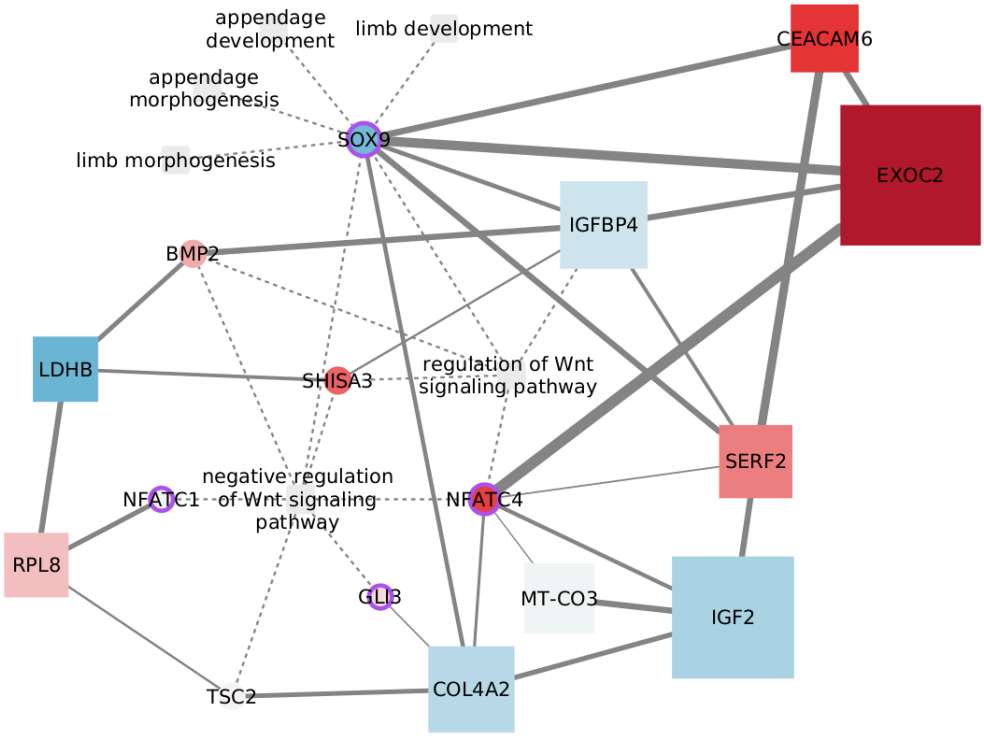


Figura 28: Predictores conectedos con la señalización Wnt en el luminal B. El tamaño de los nodos refleja la centralidad *betweenness*

ferasas H3K4 [244], que a su vez se relaciona con la modificación del DNA [245]. De manera similar, BCOR forma parte del complejo polycomb no canónico y está alterado en distintos tipos de cáncer [246]. Se ha observado que BANP puede abrir la cromatina en promotores CpG no metilados, activando genes esenciales en células pluripotentes y neuronas diferenciadas [247]. Finalmente, CUL1 interactúa con DNMT3b, de manera que efectivamente está conectado con la metilación aberrante [248].

En contraste, hay un conjunto de nodos que dependen del silenciamiento epigenético, como sucede con INPP5A en el adenocarcinoma de pulmón [249]. Además, junto con ATP11A y otros marcadores, la metilación de INPP5A tiene capacidad discriminatoria del cáncer colorrectal [250]. Del mismo modo, la metilación de ATP11A distingue varias enfermedades, incluyendo al cáncer de próstata metastásico [251]; la metilactión de GREB1L (*Growth Regulation By Estrogen In Breast Cancer 1 Like*) separa a los adenocarcinomas gástricos por sobrevivencia promedio y la de DBX2 marca al suero de los pacientes con cáncer hepatocelular [252]. Como su parálogo DNAH10, las aberraciones de DNAH2 son frecuentes en carcinomas renales con un fenotipo metilador [226]. El factor transcripcional del cerebro NPAS4, representado en la red por un sitio CpG, es regulado a través de la metilación [253] y se ha ligado con la prognósis del adenocarcinoma colón [254]. Por último, aunque la metilación de ITGB1 es constante en cáncer y tejido normal [255], la alteración de su expresión se ha reportado en tumores basales con mutaciones de BRCA, resaltando la relevancia de la migración y las propiedades mesenquimales para el subtipo [256].

Es interesante que los dos miRNAs de la red se asocien con la migración y la capacidad de invasión del tumor, pero de maneras opuestas. La familia let-7 funciona como supresores de tumores y es inhibida por la metilación del DNA [210]. Por el contrario, miR-103 actúa como un oncogen en tumores triple negativos, siendo su sobre-expresión la ligada con mala prognósis [257].

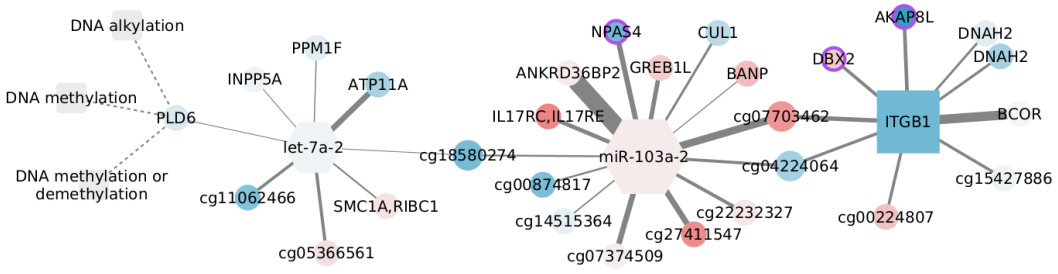


Figura 29: Predictores conectados con la metilación del DNA en el tejido normal. El color corresponde a la expresión diferencial del subtipo basal. Los hexagónos representan miRNAS

A pesar de que los cambios en la expresión son pequeños, son coherentes con lo que se esperaría en el subtipo.

3.6. Conclusión

El acoplamiento del SGCCA con ARACNE permite la inferencia de modelos de red de las funciones multi-ómicas previamente descritas. Para demostrar que los modelos tienen implicaciones regulatorias harían falta experimentos adicionales, sin embargo, la naturaleza de los nodos como sitios CpG, TFs y miRNAs debe ser considerada, tal y como se hizo en los ejemplos. Así, los modelos expuestos muestran el nivel de detalle posible en el camino hacia la validación experimental dirigida.

El SGCCA permite encontrar las funciones representadas por los predictores que mejor explican la covarianza entre la metilación, la expresión de transcritos y de miRNA. Esto no quiere decir que las funciones no puedan ser influenciadas por otros mecanismos regulatorios, sino que simplemente exhiben las funciones, como la señalización HIF en tumores basales, con mayor dependencia en la relación entre las tres ómicas. Por el contrario, la inestabilidad en la selección de predictores del LASSO, obliga a la incorporación de un filtro sobre los predictores y asegura que hay más predictores y funciones multi-ómicas de las encontradas. Aunque hay otras herramientas [112, 258] que permiten el mapeo multi-ómico de funciones sin el problema de la inestabilidad, al mismo tiempo restringen la posibilidad de reconocer nuevos roles dentro de las funciones.

En el siguiente paso, ARACNE con el filtro de MI, pone el foco sobre pares de predictores con asociaciones fuertes, potencialmente regulatorias. La desventaja aquí es el número de muestras inferior a lo recomendado del subtipo HER2E y del tejido normal, lo que potencialmende da lugar a ruido en la inferencia. Dado que la información mutua es insensible al rango, hay que advertir que, a pesar del filtro sobre los valores de MI, probablemente sólo algunas de las interacciones en la **figura 26** se mantengan relevantes si se consigue aumentar el tamaño de la muestra. Las herramientas basadas en PANDA [156, 157, 259] evitan el problema del tamaño de la muestra partiendo de redes de correlación de Pearson, pero en cambio requieren información a priori que no siempre está disponible. Durante la construcción del flujo de trabajo se intentó la construcción de redes PUMA, pero una vez que los predictores inestables se descartaron, no fue posible encontrar suficientes interacciones regulatorias validadas para correr la herramienta.

Antes de terminar hay que reconocer la escasez de interacciones con miRNAs y la abundancia de

proteínas ribosomales. Lo primero puede ser ocasionado por la información mutua de los enlaces con miRNAs, que es inferior a la reportada entre transcritos [84] o con CpGs y habla de una desventaja ante métodos dedicados [96]. Lo segundo hace sospechar una dependencia estadística trivial, incompatible con la baja centralidad mostrada en las redes, pero que igualmente amerita una mención.

Conclusiones generales

Justo como se planteó en el esquema de trabajo propuesto en un inicio, las dos aproximaciones a la integración multi-ómica ofrecen resultados complementarios, que en conjunto, cumplen con los cuatro objetivos específicos del proyecto y conforman los tres artículos que resumen los resultados de este proyecto doctoral. Mientras los modelos multivariados escuetos propician la identificación puntual de predictores de interés, las redes permiten evaluar niveles regulatorios completos, sin que sea difícil, posteriormente, encontrar nodos que por sus conectividad pueden ser relevantes. Al retomar ambas aproximaciones, este proyecto alcanza un rango muy amplio, desde observaciones globales en cuanto a la conectividad de una capa regulatoria como los CpGs, hasta la asociación consistente de los genes del PAM50 con miR-21. Incluye la foto panorámica del Amazonas y el acercamiento a la hormiga que lleva una hoja. Al acoplar las aproximaciones del SGCCA y la información mutua, se pierde perspectiva, pero en cambio se logra la extracción sistemática de modelos multi-ómicos centrados en funciones.

A pesar de la complejidad técnica y biológica, los distintos enfoques explorados coinciden en una perturbación de las capas regulatorias, especialmente de la metilación del DNA, en los subtipos del cáncer de mama respecto al tejido normal, que, sin embargo no es consistente entre subtipos. Si bien la proporcion de CpGs es distinta a la del tejido normal -en las redes de MI y en los componentes del SGCCA-, los circuítos difieren entre subtipos. Alimentando la pregunta teórica, subyacente, de la proporción de estados que convergen al atractor normal contra la proporción que cae en el atractor cáncer. Tristemente, esta conclusión general no avnza en el conocimiento, ya se sabía que los subtipos son diferentes. Hace falta entonces, volver a las conclusiones particulares de cada enfoque, que se resumen a continuación.

Para el primer artículo, descrito en la sección 1, se construyeron modelos de red elástica de los genes en el clasificador PAM50, asociando las patrones -alterados- de cada ómica en los cuatro subtipos del cáncer de mama y el tejido normal. Este análisis permite explorar el potencial regulatorio de los niveles funcionales completos. La transformación de los modelos de un

subtipo en una red direccional y pesada, aborda también la búsqueda de interacciones específicas entre ómicas y hace posible proponer mecanismos reguladores, como el interruptor formado por miR10a/b y miR-21. Aunque hay enriquecimiento funcional de los predictores seleccionados para cada gen en el PAM50, estas no son funciones multi-ómicas, ya que surgen solamente de la capa transcriptómica de predictores. Además de estar limitada al ámbito del PAM50, esta aproximación es lenta, porque requiere el ajuste de un modelo por gen, y carece de valores de significancia acompañando a los predictores.

El artículo descrito en la sección 2 discute redes multi-ómicas, de información mutua, centradas en transcritos functionales y estadísticamente ligados a los mecanismos regulatorios estudiados. El modelo de redes sólo necesita inferirse una vez por subtipo, lo que reduce el tiempo de cómputo, captura relaciones no lineales y no necesita de normalizaciones específicas. En este caso, también se extraen resultados sobre los niveles funcionales completos, como el grado de los CpGs, y se proponen interacciones puntuales potencialmente regulatorias; pero hay un sesgo claro hacía las observaciones sobre capas regulatorias enteras. Al mismo tiempo, el análisis de enriquecimiento funcional de las redes se acerca al objetivo que trata de funciones multi-ómicas, sin llegar a satisfacerlo del todo, por recuperar cualquier proceso biológico suficientemente representado entre los transcritos con cualquier tipo de asociación con las otras ómicas. En este sentido, las aristas de las redes no implican causalidad, no tienen dirección ni se pueden separar en directas e indirectas. En otras palabras, cada aproximación tiene ventajas y desventajas propias.

El acoplamiento de un modelo multivariado escueto con una red probabilística, que son el SGCCA y las redes de información mutua, se aborda en el último artículo, sección 3. Por fin, en este caso, se satisface el tercer objetivo y se identifican las funciones más afectadas por los niveles funcionales estudiados. Además se encuentran relaciones precisas que conforman los modelos multi-ómicos de la sección 3.5, un objetivo de la integración multi-ómica poco abordado. La elección del SGCCA, con penalización LASSO, sobre el ajuste de un modelo de red elástica por gen acelera el proceso, al requerir sólo un ajuste por subtipo y para el tejido normal, pero en cambio, obliga a la incorporación de un filtro de predictores de acuerdo a su estabilidad. La construcción de una red por función y el filtro de MI garantizan redes escuetas, más fáciles de interpretar que un grafo completo sin enriquecimiento funcional. Se conservan sin embargo los problemas en cuanto a la falta de un valor de significancia que acompañe a los predictores seleccionados y los propios de la información mutua: las aristas no tienen dirección y no es posible distinguir entre relaciones directas e indirectas, aunque el filtro de MI intenta paliar esto.

En todos los casos se optó consistentemente por la integración temprana, ya que se deseaba

capturar la interrelación de las ómicas. En consecuencia, los distintos datos reflejando CpGs, transcritos y miRNAs, se toman como uno solo, después de aplicar la normalización pertinente. Mientras la heterogeneidad entre plataformas no implica un problema para la información mutua, que parte de distribuciones de probabilidad, si podría serlo para los modelos multivariados escuetos. A diferencia de la implementación usada para los modelos de red elástica, el SGCCA tiene un parámetro de contracción por ómica, que protege los efectos de menor magnitud respecto a los otros niveles funcionales de ser descartados injustamente. Sin embargo, a pesar del parámetro de contracción menos estricto, el SGCCA recupera pocos miRNAs. Aunque esto podría obeder al efecto de menor magnitud de tales reguladores, también podría deberse al uso de precursores, en lugar de miRNAs maduros, que si bien permiten el mapeo preciso a coordenadas genómicas, no representan al efector final de la regulación.

Más allá de las diferencias técnicas, todas las aproximaciones reportan observaciones estadísticas sobre la biología de los subtipos del cáncer de mama, que son novedosas y potencialmente útiles en la búsqueda de alternativas clínicas. Probablemente el resultado más interesante, por conciso, es el intercambio de miR-10a/b por miR-21, en los cuatro subtipos. Aunque la baja predictibilidad de los miRNAs le quita fuerza al resultado, al sugerir una capacidad regulatoria despreciable, es interesante que el conjunto de genes en el PAM50, sin nada más en común que discernir correctamente los subtipos, se conecten a través de estos miRNAs. Considerando los coeficientes que acompañan las conexiones con miR-10a/b y miR-21, el efecto que tiene la alteración de su expresión y su selección en todos los modelos, parece sensato evaluar la posibilidad de que se trate de reguladores maestros de los genes del PAM50. Simultáneamente, hace falta aclarar qué efecto tiene este vínculo sobre la expresión del PAM50 y sobre el cáncer del mama.

Respecto a las redes de información mutua hay que resaltar la diferencia entre las capas regulatorias. Aunque no es sorprendente, es llamativo que los CpGs tengan un grado cercano a uno, en oposición a los miRNAs, coincidiendo con los mecanismos de acción de cada tipo de regulador. Si esta exclusividad de los CpGs se demostrará, podría ser útil para el control y monitoreo de las funciones cooptadas por el cáncer, al no reflejar más que a la función de interés, sin captar procesos alternativos. Por otro lado, si el grado y eventualmente los transcritos accesibles y los reguladores potenciales compartidos entre procesos, son consecuencia del mecanismo de regulación, ¿por qué habría enlaces con sitios CpG físicamente lejanos al transcrito? Al implicar una dependencia estadística sin filtro DPI, podría tratarse de relaciones indirectas, lo cual sólo desplaza la cuestión a buscar las razones de tal asociación indirecta. Respecto a la casi nula conservación de las aristas de la red normal y en general de las 5 redes, hay evidencia contradictoria,

con las interacciones entre transcritos y miRNAs siendo excluyentes entre tipos de cáncer [96]; pero con el 88.82 % de las interacciones con sitios CpG siendo compartidas en subtipos del cáncer de mama [127].

Cabe destacar que tanto los modelos de red elástica como las redes de información mutua, coinciden, a grandes rasgos, en una asociación superior entre sitios CpG y transcritos, que entre miRNAs y transcritos. Mientras las distribuciones de coeficientes de red elástica de los CpGs se concentran en valores absolutos más elevados; las distribuciones de MI tienen mayor rango. Analizando datos de cáncer de ovario, Sohn et al. ya habían observado que los mayores coeficientes en sus modelos LASSO correspondían a interacciones con sitios CpG [126]. Mientras que Setty et al. hablan de grandes coeficientes negativos para la metilación, usando sólo sitios CpG en el promotor del gen a predecir, y coeficientes variables para los miRNAs [122]. Aunque los valores no empatan con lo reportado por Drago-García et al., lo cual podría explicarse con la diferencia entre conjuntos de datos; comparativamente, las aristas con miRNAs sí tienen menor información mutua que, en ese caso, las aristas que sólo conectan transcritos. Eventualmente tanto los coeficientes como la información mutua reflejan la dependencia estadística entre las dos variables, la pregunta es si esta tendencia capta una diferencia biológica entre las dos capas regulatorias. Setty et al. explican los coeficientes que acompañan miRNAs con la acción simultánea de otros reguladores y un efecto modesto. Equivalentemente, las asociaciones fuertes con sitios CpGs, hablarían de un efecto mayor, con menos co-reguladores, que concuerda con el papel de la metilación como determinante del tipo celular [43].

Por su parte, la inferencia de redes a partir del SGCCA resalta un aspecto relevante para los tratamientos y es que, si dos funciones se conectan a través de predictores de la misma o distintas ómicas, es posible incidir sobre las dos simultáneamente. Si se trata de una intervención controlada podría, quizá, acentuarse el daño al tumor. Por el contrario, los caminos alternativos entre funciones a, por ejemplo, SOX9 en la red del subtipo luminal B, conllevan un nivel de robustez que dificulta la perturbación del todo y suponen una opción de resistencia a la intervención. Esto es similar a la transcripción desde los elementos de respuesta a estrógeno por la interacción del receptor con SP1 y no con la hormona, pero identificando posibles compañeros entre sitios CpG y miRNAs, además de entre TFs.

A lo largo de este trabajo se excluyeron las aristas ligando dos CpGs, a pesar de que esto añade un paso extra al cálculo de MI, que de otra manera pasaría por todos los pares posibles de predictores. Estrictamente este tipo de aristas no podían surgir de los modelos de red elástica, porque se trata de un análisis direccional para predecir la expresión genética, que entonces

siempre ocupa alguno de los extremos del enlace. Sin embargo, en los dos últimos artículos ignorar las interacciones entre sitios CpGs fue una decisión activa, pues no forman parte de los mecanismos regulatorios considerados, implican una interpretación más especulativa y ocupan una gran cantidad de espacio y tiempo, considerando alrededor de $1{,}5x10^{11}$ pares posibles. Descartando relaciones indirectas ocasionadas por la co-variación de los transcritos, la conexión entre CpGs podría exponer los programas epigenéticos. Estudios previos hablan de un nivel de co-metilación que decrece conforme aumenta la distancia genómica y similaridad funcional entre genes [42]; a la par de grupos de genes co-metilados que separan las muestras normales y con cáncer, independientemente del tejido [55]. Entonces, las aristas entre CpGs no son necesarias para el objetivo del proyecto, pero surgen durante los análisis desarrollados y son relevantes de explorar en otro contexto.

Por último, es necesario resaltar la necesidad de bases de datos públicas donde depositar los patrones capturados por las redes generadas, para pasar de la descripción estadística a la construcción de conocimiento. La producción sistemática de hipótesis demostrables aquí intentada, no tiene mayor impacto sin el intercambio entre laboratorios experimentales y computacionales. Sin embargo, la velocidad del trabajo computacional y el experimental es muy diferente, dados los requisitos tan dispares de cada uno. De modo que se hacen necesarios repositorios accesibles donde consultar las relaciones estadísticas ya encontradas, cuando así sea de interés. Lo que plantea a su vez la necesidad de criterios estandarizados para decidir qué vale la pena guardar y en qué formato. Al tratarse de análisis computacionales, podría argumentarse que basta con cuidar la reproducibilidad del código [98] y la preservación de los datos de partida; pero esto le suma una carga no necesariamente ligera a los posibles usuarios y restringe la viabilidad de estudios de meta-análisis. En este sentido, se han publicado sitios enfocados al trabajo de los grupos involucrados o a objetivos particulares [161, 162, 260], asentando posibles caminos.

frontiers in Oncology

ORIGINAL RESEARCH
published: 22 May 2020
doi: 10.3389/fonc.2020.00845

Multi-Omic Regulation of the PAM50 Gene Signature in Breast Cancer Molecular Subtypes

Soledad Ochoa[1,2], *Guillermo de Anda-Jáuregui*[1,3]* and *Enrique Hernández-Lemus*[1,4]*

[1] Computational Genomics Division, National Institute of Genomic Medicine, Mexico City, Mexico, [2] Graduate Program in Biomedical Sciences, Universidad Nacional Autónoma de México, Mexico City, Mexico, [3] Cátedras Conacyt para Jóvenes Investigadores', National Council on Science and Technology, Mexico City, Mexico, [4] Center for Complexity Sciences, Universidad Nacional Autónoma de México, Mexico City, Mexico

Breast cancer is a disease that exhibits heterogeneity that goes from the genomic to the clinical levels. This heterogeneity is thought to be captured (at least partially) by the so-called breast cancer molecular subtypes. These molecular subtypes were initially defined based on the unsupervised clustering of gene expression and its correlate with histological, morphological, phenotypic and clinical features already known. Later, a 50-gene signature, PAM50, was defined in order to identify the biological subtype of a given sample within the clinical setting. The PAM50 signature was obtained by the use of unsupervised statistical methods, and therefore no limitation was set on the biological relevance (or lack of) of the selected genes beyond its predictive capacity. An open question that remains is what are the regulatory elements that drive the various expression behaviors of this set of genes in the different molecular subtypes. This question becomes more relevant as the measurement of more biological layers of regulation becomes accessible. In this work, we analyzed the gene expression regulation of the 50 genes in the PAM50 signature, in terms of (a) gene co-expression, (b) transcription factors, (c) micro-RNAs, and (d) methylation. Using data from the Cancer Genome Atlas (TCGA) for the Luminal A and B, Basal, and HER2-enriched molecular subtypes as well as normal tumor adjacent tissue, we identified predictors for gene expression through the use of an elastic net model. We compare and contrast the sets of identified regulators for the gene signature in each molecular subtype, and systematically compare them to current literature. We also identified a unique set of predictors for the expression of genes in the PAM50 signature associated with each of the molecular subtypes. Most selected predictors are exclusive for a PAM50 gene and predictors are not shared across subtypes. There are only 13 coding transcripts and 2 miRNAs selected for the four subtypes. *MiR-21* and *miR-10b* connect almost all the PAM50 genes in all the subtypes and normal tissue, but do it in an exclusive manner, suggesting a cancer switch from *miR-10b* coordination in normal tissue to *miR-21*. The PAM50 gene sets of selected predictors that enrich for a function across subtypes, support that different regulatory molecular mechanisms are taking place. With this study we aim to a wider understanding of the regulatory mechanisms that differentiate the expression of the PAM50 signature, which in turn could perhaps help understand the molecular basis of the differences between the molecular subtypes.

Keywords: multi-omic approaches, breast cancer subtypes, PAM50, elastic net, data integration

OPEN ACCESS

Edited by:
Chiara Romualdi,
University of Padova, Italy

Reviewed by:
Tanja Kunej,
University of Ljubljana, Slovenia
Valentina Silvestri,
Sapienza University of Rome, Italy

***Correspondence:**
Guillermo de Anda-Jáuregui
gdeanda@inmegen.edu.mx
Enrique Hernández-Lemus
ehernandez@inmegen.gob.mx

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: *01 December 2019*
Accepted: *29 April 2020*
Published: *22 May 2020*

Citation:
Ochoa S, de Anda-Jáuregui G and Hernández-Lemus E (2020) Multi-Omic Regulation of the PAM50 Gene Signature in Breast Cancer Molecular Subtypes. Front. Oncol. 10:845. doi: 10.3389/fonc.2020.00845

1. INTRODUCTION

Breast cancer is the most common cause of cancer death among females (1). Breast tumors have been classified in molecular subtypes with distinctive clinical characteristics and a recognizable gene expression signature (2). Such signature has been reduced to 50 genes that achieve the best separation of subtypes, attaining the PAM50 classifier (3). However, the physiological implications of the difference in gene expression, if any, are not well-understood.

Given that gene expression is regulated by several interconnected mechanisms (4–7), differences across subtypes are expected for these mechanisms. Evidence of this was found in the form of distinguishable patterns of DNA methylation, mutation and miRNA expression that shape groups partially equivalent to the molecular subtypes (8). These patterns imply a link between the different omics and PAM50 gene expression, but do not clarify which genomic, epigenetic or post transcriptional changes drive the expression signature of such molecular subtypes. To advance in the identification of such drivers of molecular subtypes expression, we propose the use of a sparse model of PAM50 gene expression.

Sparse models achieve the selection of the best predictors of an independent variable by fitting penalized linear models. The penalization of the regression coefficients aim is to shrink them toward zero in such a way that predictors contributing lowly to prediction i.e., poorly associated with the independent variable, end up with null coefficient values and get filtered out of the model (9). Ridge Regression, Least Absolute Shrinkage and Selection Operator, and Elastic Network methods apply different penalizations. The elastic network approach selects groups of pairwise correlated variables instead of choosing a single predictor from the group (10, 11), augmenting the space of predictors of interest but also incrementing false positive rates (12).

Sparse models have been proposed for multi-omic sample classification (13, 14) and biomarker identification (15–17); but their capacity to simplify multi-omics co-interpretation has only been tested in the evaluation of the extent of different omics effects over a phenotype (18, 19). Here, the predictor selection capability of the elastic network approach is exploited to identify the CpGs, coding transcripts, and miRNAs most associated with the expression of the PAM50 genes in order to outline molecular differences behind the gene expression patterns characterizing breast cancer subtypes within a true multi-omic framework. The hypothesis is that PAM50 gene expression patterns are accompanied by distinctive regulatory elements, reflecting the way gene expression is controlled in the different breast cancer subtypes.

2. METHODS

2.1. Data Acquisition

Concurrent experimental samples of DNA methylation, transcript and miRNA expression were downloaded from the GDC (https://portal.gdc.cancer.gov/repository) at May 2019. Only samples with Illumina Human Methylation 450, RNA-seq and miRNA-seq measures were kept; filtering out samples quantified with the Illumina Human Methylation 27 BeadChip, which covers a smaller portion of the genome than the one we wanted to target. Subtype classification was also downloaded from the GDC trough TCGABiolinks R package (20).

After preprocessing them according to Aryee er al. (21), Tarazona et al. (22), and Tam et al. (23), and biomaRt v95, values of methylation for 384,575 probes and expression for 16,475 coding transcripts and 433 miRNA precursors were obtained for 45 unique samples of Her2, 395 LumA, 128 LumB, and 125 Basal subtypes, plus 75 samples of non-tumor (normal adjacent) tissue.

2.2. Elastic Network Implementation

The three different data types were concatenated and normalized to have mean $= 0$ and standard deviation $= 1$. Eighty percent of the samples for each subtype were used for training, leaving the rest for testing as in Liu et al. (13). Using the R package glmnet (24), elastic network models were fitted per subtype for each gene in the PAM50 classifier with the linked script https://github.com/CSB-IG/PAM50multiomics/blob/master/enetGLMNET.R. The mixing parameter was held fixed at 0.5 because such value has shown a good performance (10), but shrinkage parameter (λ) was optimized between values from 0.001 and 1,000 through repeated cross-validation.

Cross-validation was repeated 100 times with $k = 3$-folds for the subtypes with <100 training samples (Her2+ subtype and normal tissue) and $k = 5$ for the more represented subtypes (Luminal A, Luminal B, and Basal). Chosen λ parameters were used to predict testing data and root mean squared error (RMSE) was calculated per model. Fitting was repeated with the same specifications, for only 40 samples per subtype to verify the effect of data set size.

2.3. Omics Comparison

For each PAM50 gene model, RMSE was calculated for the testing data either with (1) the complete set of selected predictors, (2) only with selected CpGs, (3) just with selected coding transcripts, or (4) solely with selected miRNAs. Omic's specific RMSE were evaluated by zeroing all coefficients not associated to the omic of interest in the already fitted models with the linked script https://github.com/CSB-IG/PAM50multiomics/blob/master/RMSEperOmics.R, in an approach similar to the one used by Setty et al. (25) to search for key regulators. Obtained values shape RMSE distributions per omic which were compared via Kolmorogov–Smirnov test. This was done both per subtype per omic and mixing all the subtypes in a distribution per omic. *P*-values obtained were corrected for multiple testing with the FDR method.

2.4. Test vs. Reported Links Between Predictors and PAM50 Genes

Enrichment for previously reported regulatory links between PAM50 genes and CpGs, TFs, and miRNAs were tested by simple Fisher's Exact Test. Tests repeated by subtypes had *p*-values adjusted by FDR. Regulatory targets were taken from Illumina's annotation in the case of CpGs and from databases accessible through R packages in the case of TFs

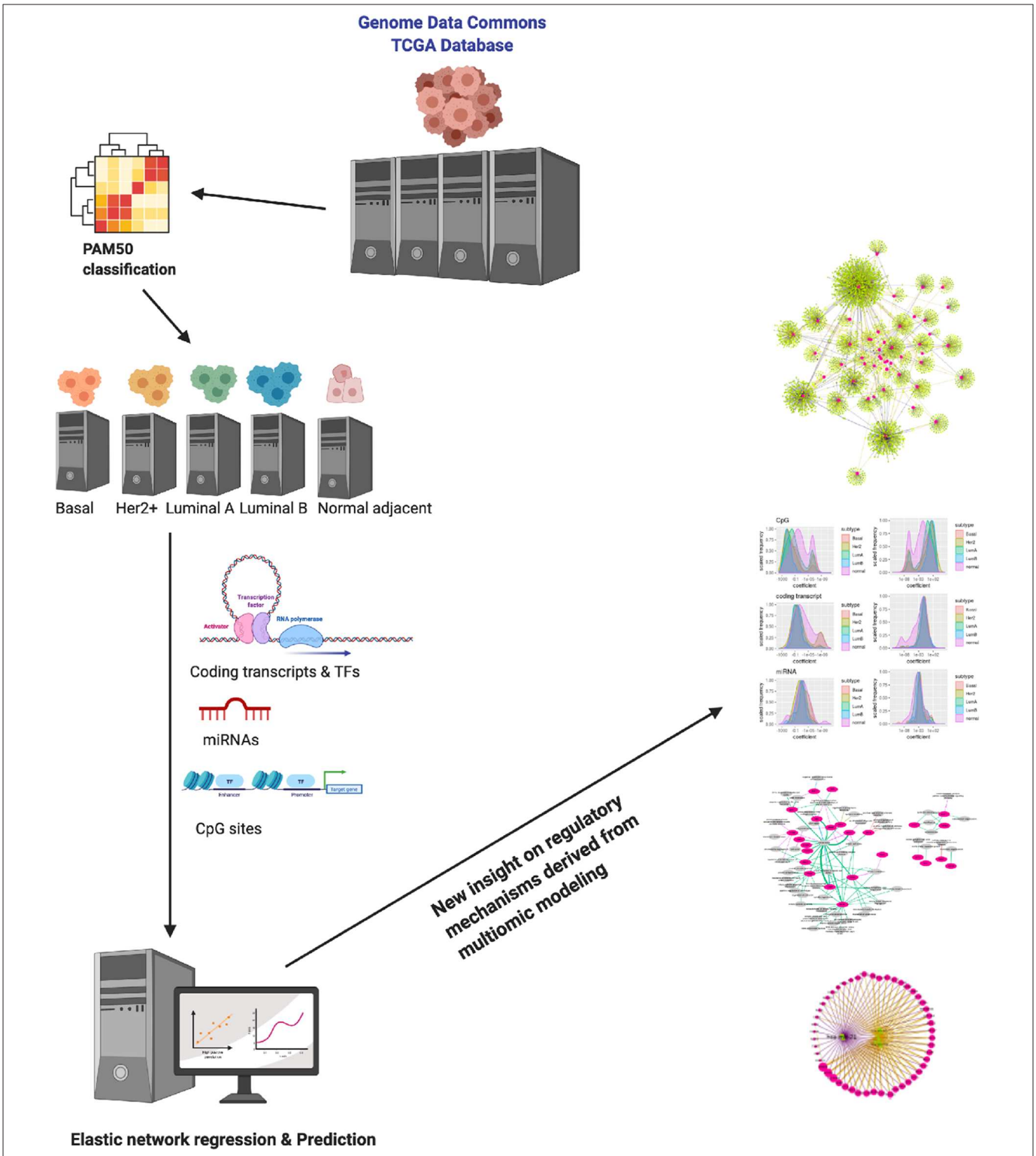


FIGURE 1 | Schematic depiction of this work. By analyzing multiomic data from the TCGA/Genome Data Commons collaboration for the different breast cancer molecular subtypes and healthy adjacent breast tissue via a generalized elastic network modeling, we have been able to derive some insight on the way the PAM50 genes are regulated (as predicted by the model). Results may shine some new light on the way PAM50 genes are able to capture intrinsic features of these phenotypes.

and miRNAs, with the linked script https://github.com/CSB-IG/PAM50multiomics/blob/master/validateInteractions.R. tftargets https://github.com/slowkow/tftargets is the package used to retrieve TF targets. It queries both predicted and validated data from TRED(2007), ITFP(2008), ENCODE(2012), and TRRUST(2015) databases at the date specified in parentheses

next to each resource, plus the lists curated by Neph et al. (26) and Marbach et al. (27).

The package used to retrieve miRNA targets is multiMiR v2.2 (28), it queries DIANA-microT-CDS, ElMMo, MicroCosm, miRanda, miRDB, PicTar, PITA, TargetScan, miRecords, miRTarBase, and TarBase, also reporting both experimentally validated and predicted results. Universe size for enrichment tests were taken from these databases, constrained to regulators measured in the input datasets. The hypothesis is that models selected reported associations between a PAM50 gene and a regulator measured in the input dataset more than expected.

2.5. Analysis of the Selected Predictors

Selected predictors and associated coefficient values were loaded to Cytoscape to construct a network of PAM50 gene predictors per subtype. PAM50 genes are taken as targets while predictors are sources, this makes a directed network were out and indegree are estimated. Predictors with the largest outdegree were submitted to an analysis of differential expression and their coefficient value distributions were compared to the global miRNA distribution via Kolmorogov–Smirnov tests. The differential analysis of miRNA expression was done per subtype by limma's package *treat* function in order to control for both fold change and significance (29). A minimum fold change of 1.1 was used.

2.6. Gene Enrichment Analysis

Every set of predictors selected for a PAM50 gene was submitted to functional enrichment analysis with the R package *HTSanalyzeR* v2.13.1 (30) versus the GO-BP with the linked script https://github.com/CSB-IG/PAM50multiomics/blob/master/enrichment.R. Sets enriched across subtypes were further tested via Fisher's Exact Test with the alternative hypothesis that selection in one subtype is exclusive with regards to selection another subtype.

The code to perform all previous analyses (see **Figure 1**) can be found at the following GitHub repository: https://github.com/CSB-IG/PAM50multiomics

3. RESULTS

Elastic network models were fitted per gene, regressing PAM50 gene expression to DNA methylation, miRNA and coding transcript expression. Elastic networks model shrink the regression coefficients toward 0, filtering predictors by its strength of association with the variable of interest. This ability for feature selection was exploited versus unfiltered omic data to identify the CpGs, coding transcripts and miRNAs most related to the PAM50 genes in cancer subtypes and normal tissue.

We fitted five models for each PAM50 gene, one per subtype and one for the normal tissue, since differences are expected for each of the 5 phenotypes. Descriptors of models per subtype and omic are reported in **Table 1**.

The output of the model are lists of associations between PAM50 genes and the selected predictors. Each selected predictor has a coefficient of regression whose value reflects the extent of association with the PAM50 gene. Coefficients are never zero, since this value means predictors can be filtered out of the prediction; but can be both negative and positive indicating an opposite effect over the predicted value. Lists of associations shape networks like the one represented in **Figure 2**. Networks for the other subtypes and the normal tissue can be found at **Figures S1–S4**.

TABLE 1 | Size of input and output of the models per subtype: Basal, Her2+, Luminal A, Luminal B as well as normal (i.e. tumor-adjacent healthy tissue).

	Basal	Her2+	LumA	LumB	Normal
Samples	125	45	395	128	75
Selected CpGs	3,090	2,514	7,173	1,485	5,373
Known CpGs selected	9	0	21	12	0
Selected coding transcripts	1,525	591	3,115	888	2,340
Selected TFs	207	91	465	133	327
Selected TFs predicted by another software	15	2	49	11	19
Selected TFs experimentally observed	4	3	25	7	9
miRNAs	101	85	174	116	123
Selected miRNAs predicted by another software	7	3	7	2	4
Selected miRNAs experimentally observed	8	5	8	12	5

From observation of networks of selected predictors to PAM50 genes, it is evident that CpGs are the most selected predictors, followed by transcripts and with only a few miRNAs selected. It can also be seen that most predictors are exclusive of a PAM50 gene but all the PAM50 genes share predictors whose pattern of expression or methylation links one gene to another. This suggests the complete set of PAM50 expression is coordinated, independently of the gene being of luminal expression, basal, or any other signature.

3.1. Omics Contribute Differently to PAM50 Gene Expression Prediction in Normal Tissue and Cancer

In order to test the reliability of the fitted models, we checked the prediction error and the selection of previously reported associations. Regulation through DNA methylation, miRNA, or TF targeting is hence regarded as true positive and compared to model's results.

The proportion of selected predictors can not be explained solely by the size of the omics taken as input (χ^2, p-value $<$ 2.2e-16, **Figure 3**), specifically, coding transcripts and miRNAs are overrepresented in the models (Fisher's Exact Test, p-value $<$ 2.2e-16). Concordantly, there are more true TF (Fisher's Exact Test, p-value $\leq$ 1.942846e-05) and miRNA (Fisher's Exact Test, p-value $\leq$ 7.573200e-11) relations than expected but less CpGs (Fisher's Exact Test, p-value $\leq$ 4.311267e-03). The exception is LumB subtype which has as many true positive CpGs as expected.

Given the difference between input and selected proportion of omics, we hypothesized a discrepant prediction power of

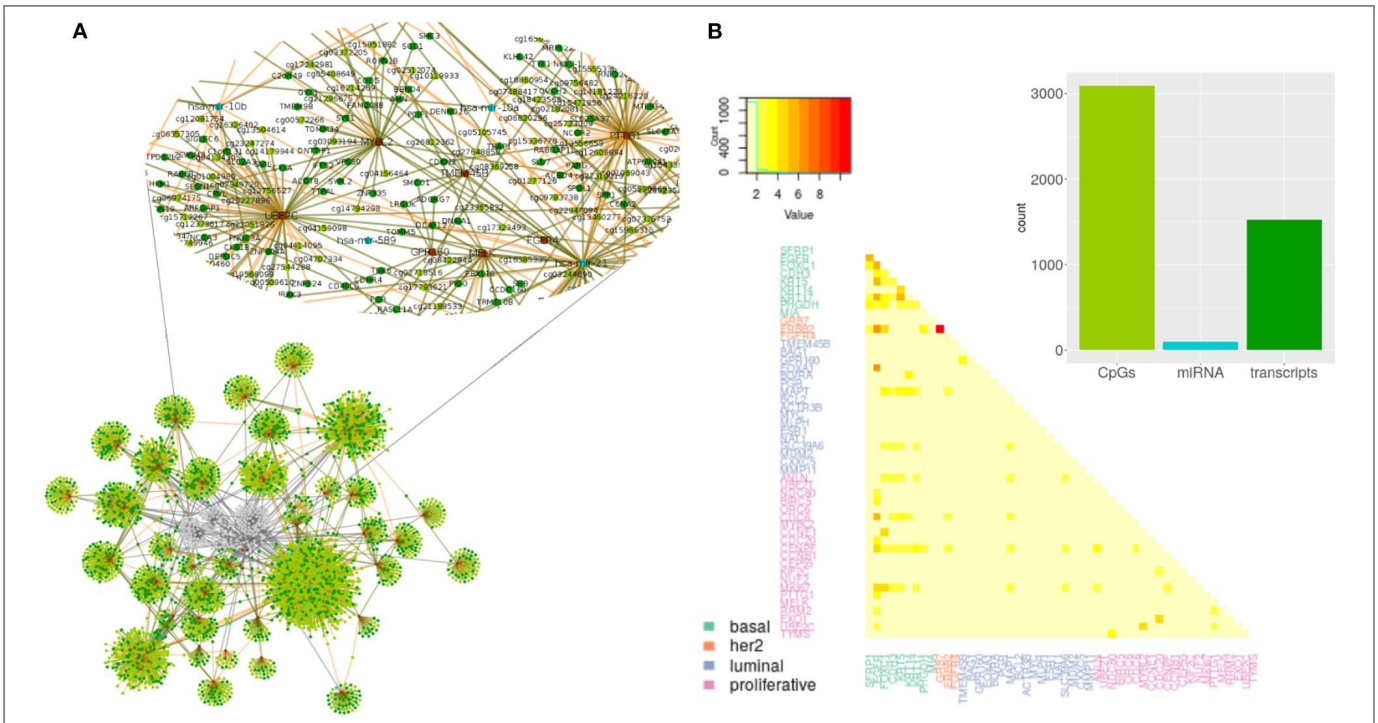


FIGURE 2 | Predictors selected per PAM50 gene for Basal subtype. **(A)** Topology of selected predictors and associated PAM50 genes. Brown circles are PAM50 genes. Transcripts are colored in dark green, miRNAs in blue, and CpGs in light green. Edges link each PAM50 gene with its selected predictors. The color of the line indicates the sign of the coefficient value associated with the predictor; negative values are in brown and positive ones in green. Zoom of the gray area shows the predictors selected for *MYBL2*. **(B)** Summary of the network. Barplot with the total representation of each omic plus heatmap of the count of predictors shared by PAM50 genes.

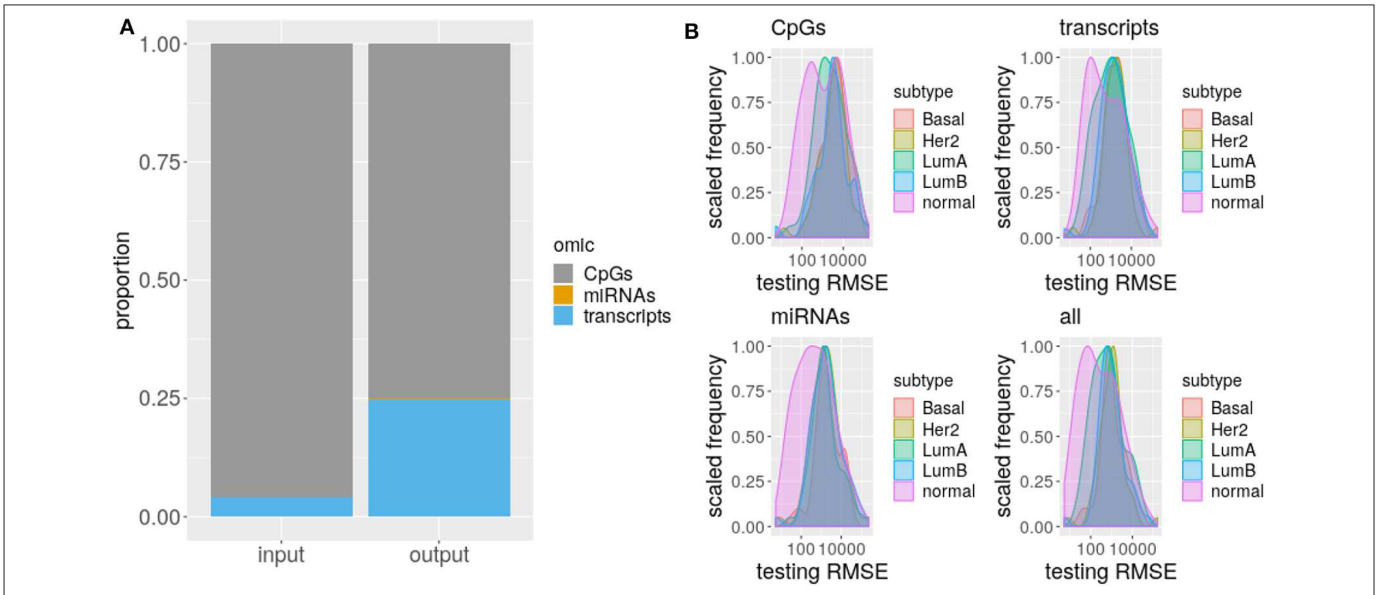


FIGURE 3 | Omics differ on selection and RMSE. **(A)** The proportion of each omic is shown for input and selected sets. The different omics are represented by a different color. **(B)** Distribution of testing RMSE per subtypes are displayed for single and multi-omics.

CpGs, coding transcripts, and miRNAs. To test this, we evaluated models carrying the complete set of selected predictors or just the predictors from each omic.

As RMSE is a standard measure to compare regression models that measures how far is the model prediction from the observed data in response variable units, then, the lower its value the better.

Normally, the error decreases the more independent predictors are included in the model, so we choose not to fit again with the selected predictor per omic, but to test the exact same model with the jointly fitted coefficient values, just zeroing predictor's coefficients from other than the omic of interest. This way, the RMSE distribution of a model containing only predictors of a given omic, represents how much of the total prediction is contributed by the predictors from that omic.

As suggested by the difference with the input proportions, DNA methylation is the less predictive omic for all the subtypes, thought this difference is not always significant (CpGs vs. coding transcripts ks. test p-value $\leq$ 0.03192 for LumB, Her2+, and Basal and CpGs vs. miRNAs ks. test p-value $\leq$ 0.02222 for Her2+ and Basal). This disagrees with the great prediction improvement reported by Huang et al. (16) for methylation data, a fact that could be driven by the much larger and heterogenous input data used here, that we believe captures better the heterogeneity of breast cancer subtypes. Meanwhile, coding transcript and miRNAs contribute the same, with no significant difference between their distributions for all the subtypes.

Remarkably, the error distribution obtained with the complete set of predictors significantly outperforms CpGs and some subtype miRNAs (ks.test p-value $\leq$ 0.02222 for LumA and Basal) but never outweighs coding transcripts. Single omics can not beat multi-omics error due to the design of the test, thus the outperforming of CpGs and miRNAs is unsurprising, what is startling is the complete statistical agreement between multi-omics prediction power and coding transcripts prediction power, which supports gene expression as the current best biomarker of molecular subtypes. We must note however that this may be related to (1) more info on RNA and (2) PAM50 was derived from expression signatures.

Finally, there is no significant difference across subtypes RMSE distributions for both single-omics and multi-omics, but CpGs (ks.test p-value $\leq$ 0.01601952), miRNAs (ks.test p-value $\leq$ 0.002834981), and multi-omics (ks.test p-value $\leq$ 0.03919459) distributions of normal tissue differ from the distribution of each subtype, suggesting these omics represent a distinct amount of PAM50 gene expression in normal tissue than in cancer, that is, the association of DNA methylation and miRNA expression with PAM50 gene expression is altered in cancer.

3.2. The Association Strength Distributions of Predictors Are Different for Each Subtype

The difference between omics extends to coefficient values, shown in **Figure 4**. Since coefficients represent the strength of association between predictors and PAM50 expression (16), coefficient values suggest that each omic has a specific association with PAM50 gene expression. Coefficient value distributions are significantly different between subtypes (ks.test p-value $\leq$ 2.82E-02) and omics (ks.test p-value $\leq$ 0.01535) with few exceptions for coding transcripts and miRNAs. Basal, Her2+, and LumB coding transcripts coefficients are not significantly different. Neither are miRNA coefficients of pairs LumA and normal tissue, LumB and Basal subtype, and Basal and Her2.

According to these distributions, DNA methylation has a strong but noisy association with PAM50 gene expression while miRNA (Fisher test p-values $\leq$ 0.001403597) and coding transcript (Fisher test p-values $\leq$ 1.086031e-29) association tends to be positive (**Figure S3**) and more stable. The elevated association between DNA methylation and PAM50 genes expression explains why so many CpGs get selected in spite of its low prediction power. A stronger association between DNA methylation and gene expression than between gene and miRNA expression had previously been found for ovarian cancer by Sohn et al. (18) using a different penalization modeling.

3.3. *miR-21* and *miR-10b* Are the Only Relevant Predictors Selected Across Subtypes

Next, we wanted to see how variable is actually the association between one predictor and the predicted PAM50 gene, that is, the specific coefficient values, not their distributions. For this, we wanted to focus on the predictors selected for a PAM50 gene across subtypes, shown in **Figure 5**. However, as noted before, models selected a great quantity of predictors exclusive for each gene, 93.45% of the selected CpGs, 74.24% of the coding transcript, and 81.37% the miRNAs are not shared between any two genes. In consequence, there are no CpGs associated with any gene for all the subtypes but there are 14 relations with coding transcripts and 51 with miRNAs satisfying this.

The 13 coding transcripts selected across subtypes as predictors of a specific PAM50 gene are trivial, since they just portray physical linkage. *ELP2* and *SLC39A6* are coded in opposite strands of the same locus while the rest of pairs are contiguous. Most of the associations, 84.77%, connect a PAM50 gene with a coding transcript in another chromosome, but these are not repeatedly selected across subtypes. It is worth mentioning that although all coefficients values are positive, even close predictors, like *YEATS4* and *SLC35E3* carry distinct coefficients.

Regarding miRNAs, there are only two miRNAs repeatedly selected among subtypes, *miR-10b* and *miR-21*. These are known breast cancer markers targeting some PAM50 genes (31). *Mir-21* has been experimentally linked with *BCL2*, *MYC*, *EGFR*, and *ERBB2* expression (32–35) and predicted to target *ESR1* and *FOXA1* (36, 37). On the other hand, *miR-10b* has been linked to *CDC6*, *EGFR*, and *SFRP1* (38, 39). There is no particular pattern among validated associations or coefficients, other than *miR-21* carrying mostly positive coefficient values and *miR-10b* selection extending up to normal tissue (for the full set of validated interactions please see **Supplementary Table S1**).

3.4. Micro-RNA *miR-21* and *miR-10b* Are Universal PAM50 Predictors in Cancer and Health

Next we wanted to check the role of *miR-21* and *miR-10b* per subtype. With this in mind, we revisited the models derived networks, that link PAM50 genes and predictors per subtype.

The networks show that genes overexpressed in each subtype get larger models. About 30% of the luminal genes have models

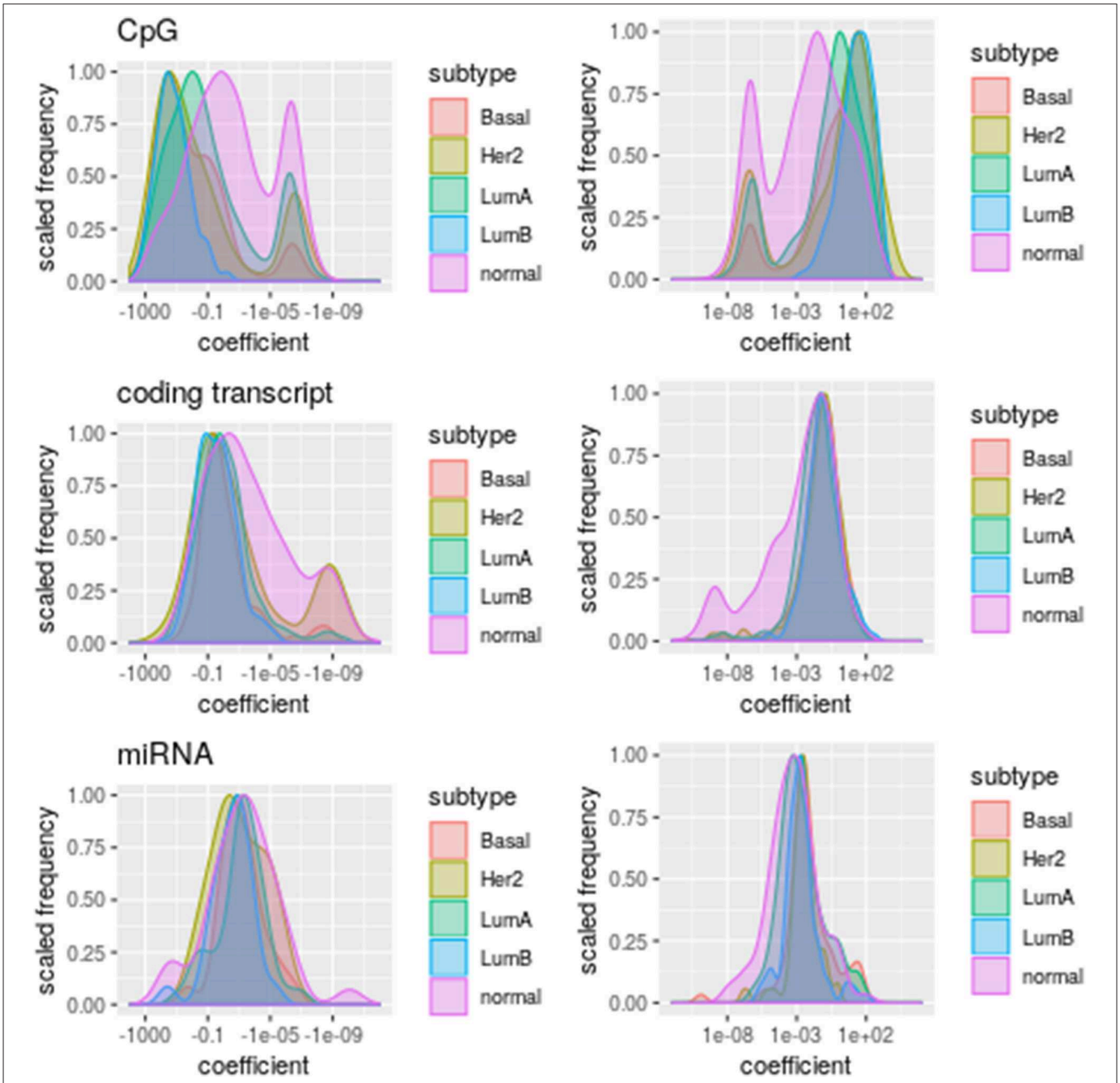


FIGURE 4 | Distribution of coefficient values per omic. The subtypes are represented by different colors. Plots at the **(Left)** column represent negative coefficients whereas those at the **(Right)** column stand for positive coefficient values.

larger than average for LumA subtype, while almost 90% of basal genes have the equivalent for Basal subtype. Her2+ subtype and normal tissue have no clear pattern, but for LumB subtype, half the luminal genes and 28% of the proliferative ones have increased size models.

Predictors that bridge between PAM50 genes can proceed from any omic, but CpGs are significantly underrepresented (Fisher test p-values $\leq$ 1.81E-88). CpGs are at most, selected for two subtypes as predictors of a specific PAM50 gene. There are just 24 CpGs in this situation, of which 15 are shared between Her2+ and another subtype or the normal tissue, including nine CpGs associated with *ERBB2* but placed in other loci than chromosome 17.

Meanwhile, coding transcripts and miRNAs fulfill this role more often (Fisher test p-values $\leq$ 5.84E-03) than solely input proportions would explain. This is no surprise since both pertain to the same level of molecular features, that of transcripts, as the PAM50 gene expression signature; as such, coding transcript and miRNA may be subject to the same biomolecular pressures. The stunning observation is that one miRNA can link almost all of

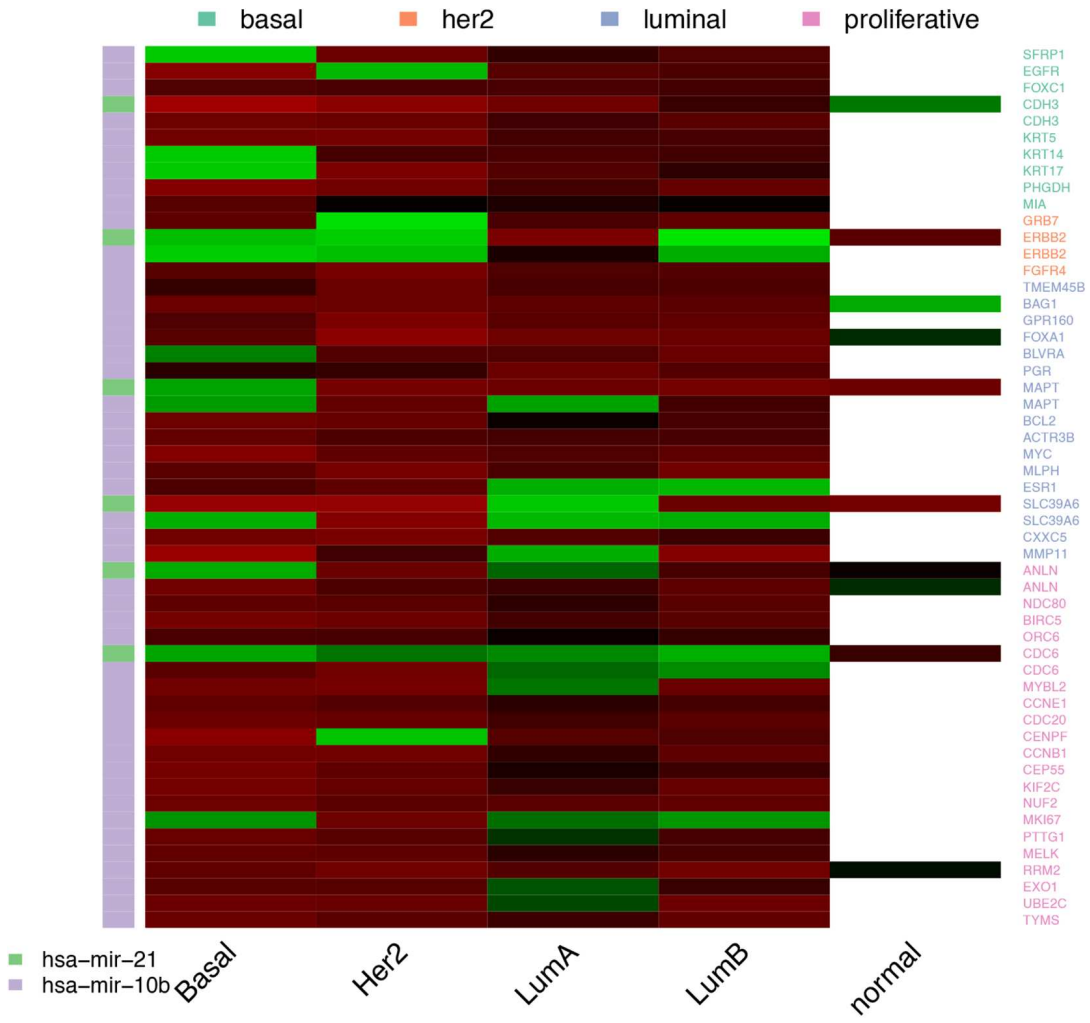


FIGURE 5 | miRNAs selected across the four cancer subtypes. miRNAs are represented by the color bar at the left. The highest positive coefficient value appears in bright red, the lowest negative coefficient is in bright green. Near zero coefficients look black. The white cells in the column of normal tissue means the predictors was not selected.

the PAM50 genes for all the cases (**Figure 6**). The outstanding miRNAs are again *miR-21* and *miR-10b*.

For normal tissue *miR-10b* was selected as predictor of all PAM50 genes while *miR-21* is linked to only four genes. On the contrary, *miR-21* is connected to most genes in the all the breast cancer subtypes, while *miR-10b* is poorly linked. For LumA subtype, shown in **Figure 6B**, both *miR-10b* and *miR-10a* are highly connected, but still can not reach genes like *FOXC1*, which is connected instead with *miR-21*.

Both *miR-10a* and *miR-10b* are members of the miR-10 family encoded within the Hox genes genomic clusters; *miR-10a* resides upstream from *HOXB4* and *miR-10b* upstream from *HOXD4* (40). Due to their relatedness they will be referred as *miR-10a/b*.

The hub-like behavior of these miRNAs agrees with previous observations of our group of highly connected miRNAs per subtype (41), which are important for network cohesion (42). Although the coefficients networks maintain a large connected component when removing *miR-10a/b* and *miR-21*, tens to hundreds of predictors are needed to link all the PAM50 genes; when only one of these miRNAs is required to achieve the same.

Given that each miRNA has the potential to target hundreds of genes (43), *miR-10a/b* and *miR-21* are not so exceptional in this regard. However, as explained earlier, only a fraction of PAM50 genes have a regulatory relation with these miRNAs, suggesting most of the detected associations are indirect. Indirectness is consistent with the low values of the coefficients, which range from −0.2938690 to 0.4333184, when miRNAs coefficient values range within two orders of magnitude higher. Coefficient value distributions of *miR-10a/b* and *miR-21* are also significantly different than the rest of miRNA coefficients (ks.test p-value $\leq$ 9.068e-05).

3.5. PAM50 Genes Enrich for Different Functions per Subtype

The selection of predictors we have presented is based on a statistical association with the pattern of expression of a

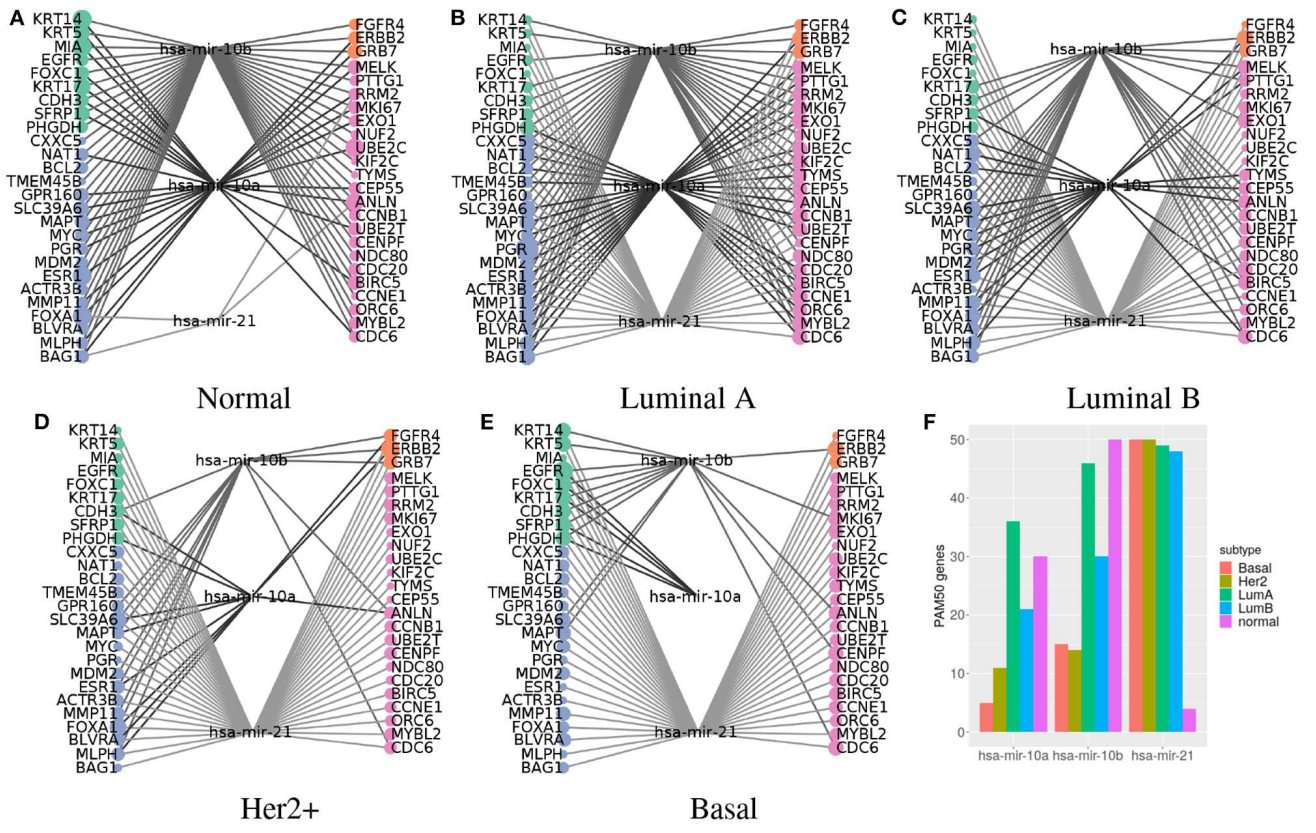


FIGURE 6 | Predictors connecting most PAM50 genes transition from *miR-10a/b* in normal tissue to *miR-21* in cancer subtypes. **(A–E)** PAM50 genes are at the sides, colored according to their pattern of expression: green for basal expression, orange for Her2 enriched, blue for luminal expression, and pink for proliferation favorable. Node size represents the number of predictors selected. Predictors are in the middle. Lines connect PAM50 genes to the associated predictors. Lines are in a gray gradient to distinguish different predictors. **(F)** The bar chart represents the number of PAM50 genes connected to the miRNAs per subtype.

PAM50 gene. The covariation sustaining such an association may respond to how a specific group of predictors is able to attain some biological function. To test this, functional enrichment was done with the set of selected predictors per gene per subtype, versus Gene Ontology Biological Processes categories (GO-BP) (**Figure 7**).

Only two PAM50 genes are enriched for some process in the Basal subtype, *FOXC1* (basal cluster) and *ANLN*(proliferative cluster). Neither the *ANLN* enrichment for telomere protection nor the *FOXC1* linkage to transforming growth factor response are within these genes immediate annotated processes. Though *FOXC1* is actually related with *TGFβ* since both are able to regulate EMT (44).

In the case of Her2+, just *ORC6* (proliferative cluster) is enriched for the totally unexpected process of synapse assembly, but, despite the significant *p*-value, we must notice that this is based on only two genes.

LumA is the most enriched subtype. This is not surprising since it has the largest number of selected coding transcripts, which is the starting material for enrichment. The 20 enriched genes are mostly linked to distinct cellular division aspects. The exception are the three keratins, genes with basal expression, which are connected through their normal processes, suggesting selected predictors respond to the normal gene's function. *MYC* and *UBE2T* are linked to rather wide categories (45) while *MLPH* associates with other than its normal processes. The remaining 14 genes are connected through categories consistent with their proliferative expression, which again alludes to a selection that followed the normal function of the genes. This is again consistent with the available evidence.

For LumB subtype, *MELK* and *CCNB1* enrich for cell division as would be normally expected; while *MYBL2* is unintuitively linked to negative regulation of epithelial cell proliferation, which however, has been reported (46). Finally, the normal tissue shows different cell division aspects coherent with the proliferative expression of its enriched genes.

Altogether, few genes have predictors with significant enrichment extended across subtypes. Eight genes enriched in two subtypes, including *CCNB1*, *MKI67*, and *UBE2C*, that connect with the same processes, the expected ones, for the two subtypes. *MELK* also connects with its normal process for two subtypes but in LumA and LumB subtypes plus normal tissue. *ANLN*, *CEP55*, *KRT17*, *MYBL2*, and *ORC6*, enrich for different processes across subtypes, that is, a fifth of the genes with any kind of enrichment, but five of the nine genes enriched for more than one subtype.

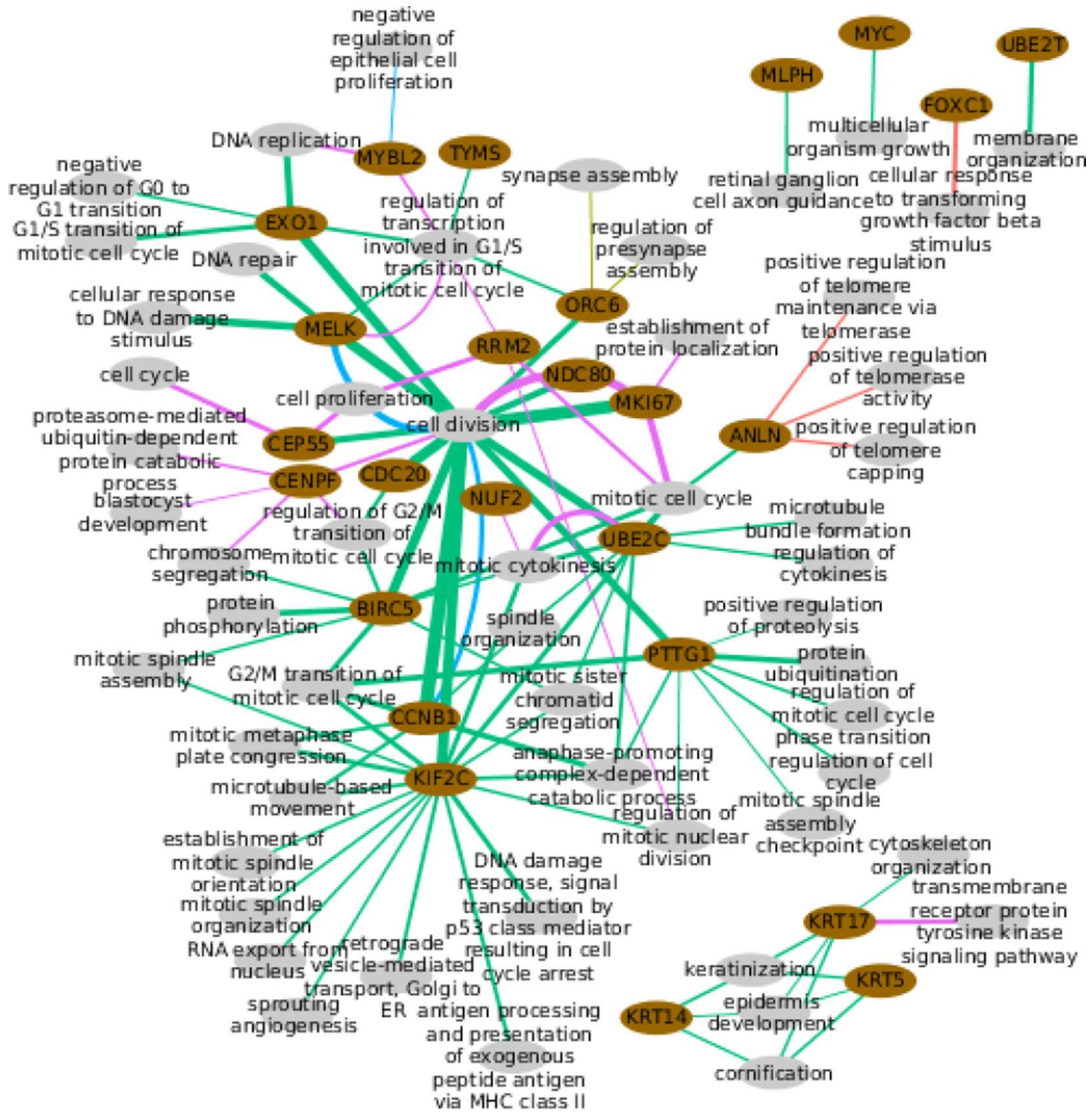


FIGURE 7 | Functional enrichment of the predictors selected per PAM50 gene. Pink ellipses represent PAM50 genes while the gray ellipses represent biological processes. Colored lines link genes with the processes they are significantly enriched to the corresponding subtype. Wider lines indicate a higher number of PAM50 gene predictors in the process.

To further test the functional enrichment per subtype, we compared the sets of predictors selected per subtype for each one of the 9 genes that enrich for several subtypes. Genes enriched for cell division across subtypes, *CCNB1*, *MKI67*, and *MELK* connect to the process via distinct sets of selected predictors. From the beginning, these genes bear different predictors (Fisher's Exact Test H1: less, p-value $\leq$ 1.281e-09), with a small intersection whose removal does not change the significant enrichment for cell division. This reflects the robustness of the process, which is so important that distinct subsets of the 603 genes annotated in the category are enough to call it.

The other two genes enriched for the same process across subtypes, *UBE2C* for mitotic cytokinesis and, *MELK* for regulation of transcription involved in G1/S transition of mitotic cell cycle, lost the functional enrichment when the predictors selected in both LumA and normal tissue (the intersection) were removed. This implies LumA mitotic cytokinesis and regulation of transcription may be involved in G1/S transition of mitotic cell cycle relying on the normal tissue mechanism.

The quantity of shared predictors between the sets selected for *CEP55*, indicates that predictor selection in the LumA subtype is exclusive for normal tissue selection (Fisher's Exact Test H1: less, p-value = 1.141e-10). This means that the differential enrichment between LumA and normal tissue is sustained by different predictors, suggesting *CEP55* fulfills divergent roles in these phenotypes. This matches differences observed between cancer and normal tissue (47) but, to our knowledge, not reported for LumA subtype.

The same reasoning supports *KRT17* and *ORC6* divergent roles across subtypes. It is odd that *KRT17* is linked to kinase signaling for normal tissue and not for a breast cancer subtype, when this has been described for another cancer (48) but this may be associated to tumor incidence over adjacent tissue (49). For *ANLN* and *MYBL2*, selection exclusion between subtypes is not significant, meaning that differential enrichment of these genes could settle on the same predictors, suggesting functional diversity.

4. DISCUSSION

Sparse penalized models have already proven useful to discover molecular mechanisms, cluster samples, and predict outcomes such as survival (50). Penalization permits the fitting of models otherwise unattainable given the relatively small sample sizes and huge number of variables measured by the omics. Here, the elastic network approach was used for integrated interpretation of different omics measuring DNA methylation and expression of both coding transcripts and miRNAs.

However, a large training set is always preferable, and not all breast cancer subtypes have been extensively sampled, which is reflected in the models. For Luminal A, the most frequent and sampled subtype, the highest number of predictors were selected by the models; while Her2+, with only 45 samples, got the lowest number of selected predictors. To assure comparability across subtypes we trained the models again, but now using the same number of samples, 40 samples, for all the subtypes. Patterns found with this subset persist in the analysis of the whole set of data, supporting comparability (**Figures S5–S8**). Nevertheless, the absence of predictors found for LumA in the smaller subtype's models due to a lack of representation can not be ruled out. This could specifically affect the functional enrichment of PAM50 neighborhoods of predictors and so, the functional divergence between subtypes is not definitive but should be experimentally tested.

Multi-omic modeling of PAM50 gene expression is no better than the sole use of coding transcripts, supporting gene expression as the best biomarker of molecular subtypes. However, our point in using the sparse model was not to predict PAM50 but to identify the molecular differences associated with PAM50 signatures that may lead to functional differences.

At the global level, a reduced prediction power of DNA methylation and miRNAs containing models was observed for all subtypes vs. normal tissue, indicating that the influence of this omics on PAM50 gene expression is reduced for cancer. Although this may be born out of incomplete knowledge or incipient technology, an alteration of these omics has been effectively reported; specifically, a generalized hypomethylation has been observed for breast and other cancers (51).

Different predictors were expected per cancer subtype, but the exclusivity of predictors from all the omics was surprisingly high. Only 13 coding transcripts and 2 miRNAs were selected for the four subtypes. The lack of CpGs selected across subtypes is consistent with the high strength of association it has with PAM50 gene expression. If the pattern of expression is different between subtypes, the highly associated CpGs should be different.

The ubiquitous selection of *miR-10b* and *miR-21* across subtypes suggests a central role for these miRNAs in breast cancer, which is actually supported by the literature. Proliferation, cell migration, and *in vivo* tumor growth of MCF7 and MDA-MB-231 cell lines implanted in nude mice is inhibited through antagomiR-21 (52) demonstrating the relevance of this miRNA, at least for luminal A and triple negative subtypes. In turn, both sub and overexpression of *miR-10* are oncogenic. *MiR-10b* overexpression enhances cell migration and invasion by targeting *HOXD10*; while subexpression of *miR-10b-3p*, coded in the same *miR-10b* locus, participates in breast cancer onset by upregulating the cell cycle regulators *BUB1*, *PLK1*, and *CCNA2* (53).

Coherent with the ubiquitous selection of *miR-21* breast cancer subtypes and its replacement by *miR-10a/b* in normal tissue. *MiR-21* is significantly overexpressed for all cancer subtypes while *miR-10b* is underexpressed, as previous reports say (31). *Mir-10a* is significantly underexpressed in Basal and Her2+ subtypes and slightly overexpressed in luminal subtypes, but this is not significant in LumB case. The proposal is that when *miR-10b* coordinates PAM50 genes, normal tissue expression is predicted; when *miR-10b* is sub expressed and *miR-21* is overexpressed, this second miRNA gains *miR-10b* place, coordinating cancer expression of the PAM50 genes. Since *miR-10b* has a known role in metastasis (31), it would be interesting to observe the dynamics of the networks throughout the evolution of the disease.

Additionally, the small coefficients associated with these miRNAs are consistent with indirect associations. Considering all these pieces, the transition from hub *miR-10a/b* in normal tissue to *miR-21* in breast cancer through the luminal subtypes, evokes a switch between two master regulators. Master regulators are genes needed for the specification of a lineage by its capacity to regulate downstream genes either directly or not, whose missexpression can re-specify the fate of cells (54).

Nonetheless, sparse models can not select regulators naively, they need to feed on known regulators (16, 25, 55). Then, the regulatory capacity of selected predictor can not be stated, leaving *miR-10a/b* and *miR-21* just as universal predictors of PAM50 genes.

Another limitation of the study is the absence of an estimator of significance or accuracy intrinsic to the methodology (56). Regression models quality is described in terms of RMSE, without an indication of how well the selected predictors describe PAM50 expression. A ROC curve is not feasible, since models would have to be turned into the classification setting, and even this is unreachable, because true negative regulators can not be ascertained, as non regulators could simply be regulators yet to discover.

Finally, it is important to mention that applying the same shrinkage to inherently different molecular levels, like CpG methylation and transcript expression, could shrink to zero all the coefficients of subtler effect predictors (13). Thus, the next implementation of sparse multiomic models on PAM50 expression should adopt multiple penalizations, which could

even ameliorate the bias on subtype representation (57). Distinct values for the mixing parameter should also be probed, as well as data decomposition into latent variables (58).

Future Directions

Apart from exploration of alternative frameworks, the immediate follow up should be the experimental assessment of the observations described here. Specifically, silencing and expression of *miR-10a/b* and *miR-21* need to be tested for each breast cancer subtype. Disection of interaction between the miRNAs and the PAM50 genes is required too.

Then, more omics could be included in the models. Copy number variation is the first candidate to be incorporated since it is already available in the databases and has a proven effect on Her2+ subtype, in particular regarding the effect of the *Her2* amplicon since it has been associated to regulation of growth and survival processes. But single nucleotide variation and chromatin accessibility are also available for some samples.

Other phenotypes with discriminant patterns of expression could benefit from sparse modeling. There could be significant predictors linked to the glioblastoma subtypes as was observed for breast cancer. Predictors represent potential regulators of the mechanisms behind subtype heterogeneity and, as such, are interesting markers of cancer. In this sense, predictor selection across stages, not subtypes, could illuminate the driving forces behind disease development. Alternative methods like A–JIVE (59) and sPLS (60) would have also exciting outcomes in this settings.

A relevant mid to long term future direction will be the implementation of experimental assays to test for multi-omic synergistic or cooperative phenomena, aiming at providing some mechanistic clues of the biological functions behind. There is however a strong challenge on this given the combinatorial mixture of effects that may be complex to disentangle. Some promissory (yet preliminary) advances are starting to arise.

5. CONCLUSION

Holistic studies of cancer are needed to dissect its complexity. Initiatives like The Cancer Genome Atlas have delivered the distinct molecular perspectives that need to be interpreted as a whole. The elastic net models subject of this work, approach such an integration in a rather simplistic linear form. Yet, the methodology is powerful enough to prove the intuition that PAM50 gene expression patterns are accompanied by distinctive potentially regulatory elements. Predictors are selected in an almost exclusive manner, heavily dictated by the omic of origin, with CpGs strongly associated to PAM50 expression not selected across subtypes. The way *miR-10a/b* and *miR-21*, the only relevant predictors selected for all subtypes, are connected and differentially expressed, suggest an specific regulatory difference between breast cancer and normal tissue that merits further research.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in the Genome Data Commons site https://bit.ly/2Itoi2e. The code to perform all previous analyses can be found at the following GitHub repository: https://github.com/CSB-IG/PAM50multiomics.

AUTHOR CONTRIBUTIONS

SO organized the database, performed the statistical analysis, and wrote the first draft of the manuscript. GA-J contributed to design of the study, generated programming code, and contributed to the writing of the manuscript. EH-L conceived the study, contributed to design of the study, provided funding, discussed findings, and reviewed the writing of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported by the Consejo Nacional de Ciencia y Tecnología [SEP-CONACYT-2016-285544 and FRONTERAS-2017-2115], and the National Institute of Genomic Medicine, México. Additional support has been granted by the Laboratorio Nacional de Ciencias de la Complejidad, from the Universidad Nacional Autónoma de México. EH-L is recipient of the 2016 Marcos Moshinsky Fellowship in the Physical Sciences.

ACKNOWLEDGMENTS

This paper constitutes a partial fulfilment of the Graduate Program in Biomedical Sciences of the National Autonomous University of México (UNAM) requirements of SO (María de la Soledad Ochoa-Méndez). She acknowledges the scholarship and support provided by the National Council of Science and Technology (CONACyT) and UNAM. **Figure 1** was generated using Biorender (https://biorender.com/).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2020.00845/full#supplementary-material

Figures S1–S4 depict the topology of the networks for the non-basal subtypes that were not shown. **Table S1** contains a list of all validated interactions.

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Cancer J Clin.* (2018) 68:394–424. doi: 10.3322/caac.21492
2. Prat A, Pineda E, Adamo B, Galván P, Fernández A, Gaba L, et al. Clinical implications of the intrinsic molecular subtypes of

breast cancer. *Breast*. (2015) 24:S26–35. doi: 10.1016/j.breast.2015.07.008
3. Perou CM, Søørlie T, Eisen MB, Van De Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. *Nature*. (2000) 406:747. doi: 10.1038/35021093
4. Cheng C, Alexander R, Min R, Leng J, Yip KY, Rozowsky J, et al. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res*. (2012) 22:1658–67. doi: 10.1101/gr.136838.111
5. Vimalraj S, Miranda P, Ramyakrishna B, Selvamurugan N. Regulation of breast cancer and bone metastasis by microRNAs. *Dis Mark*. (2013) 35:369–87. doi: 10.1155/2013/451248
6. Cao J, Luo Z, Cheng Q, Xu Q, Zhang Y, Wang F, et al. Three-dimensional regulation of transcription. *Protein Cell*. (2015) 6:241–53. doi: 10.1007/s13238-015-0135-7
7. Liu X, Chen X, Yu X, Tao Y, Bode AM, Dong Z, et al. Regulation of microRNAs by epigenetics and their interplay involved in cancer. *J Exp Clin Cancer Res*. (2013) 32:96. doi: 10.1186/1756-9966-32-96
8. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*. (2012) 490:61–70. doi: 10.1038/nature11412
9. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. Vol. 112. New York, NY: Springer (2013). doi: 10.1007/978-1-4614-7138-7
10. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B*. (2005) 67:301–20. doi: 10.1111/j.1467-9868.2005.00503.x
11. Neto EC, Bare JC, Margolin AA. Simulation studies as designed experiments: the comparison of penalized regression models in the "large p, small n" setting. *PLoS ONE*. (2014) 9:e107957. doi: 10.1371/journal.pone.0107957
12. Kirpich A, Ainsworth EA, Wedow JM, Newman JR, Michailidis G, McIntyre LM. Variable selection in omics data: a practical evaluation of small sample sizes. *PLoS ONE*. (2018) 13:e0197910. doi: 10.1371/journal.pone.0197910
13. Liu J, Liang G, Siegmund KD, Lewinger JP. Data integration by multi-tuning parameter elastic net regression. *BMC Bioinformatics*. (2018) 19:369. doi: 10.1186/s12859-018-2401-1
14. Tini G, Marchetti L, Priami C, Scott-Boyer MP. Multi-omics integration-a comparison of unsupervised clustering methodologies. *Brief Bioinformatics*. (2019) 20:1269–79. doi: 10.1093/bib/bbx167
15. Bravo-Merodio L, Williams JA, Gkoutos GV, Acharjee A. -Omics biomarker identification pipeline for translational medicine. *J Transl Med*. (2019) 17:155. doi: 10.1186/s12967-019-1912-5
16. Huang S, Xu W, Hu P, Lakowski TM. Integrative analysis reveals subtype-specific regulatory determinants in triple negative breast cancer. *Cancers*. (2019) 11:507. doi: 10.3390/cancers11040507
17. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omic assays. *Bioinformatics*. (2019) 35:3055–62. doi: 10.1093/bioinformatics/bty1054
18. Sohn KA, Kim D, Lim J, Kim JH. Relative impact of multi-layered genomic data on gene expression phenotypes in serous ovarian tumors. *BMC Syst Biol*. (2013) 7:S9. doi: 10.1186/1752-0509-7-S6-S9
19. Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann Appl Stat*. (2013) 7:523. doi: 10.1214/12-AOAS597
20. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*. (2016) 44:e71. doi: 10.1093/nar/gkv1507
21. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. (2014) 30:1363–9. doi: 10.1093/bioinformatics/btu049
22. Tarazona S, Furió-Tarí- P, Turrà D, Pietro AD, Nueda MJ, Ferrer A, et al. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res*. (2015) 43:e140. doi: 10.1093/nar/gkv711
23. Tam S, Tsao MS, McPherson JD. Optimization of miRNA-seq data preprocessing. *Brief Bioinformatics*. (2015) 16:950–63. doi: 10.1093/bib/bbv019
24. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. (2010) 33:1–22. doi: 10.18637/jss.v033.i01
25. Setty M, Helmy K, Khan AA, Silber J, Arvey A, Neezen F, et al. Inferring transcriptional and microRNA-mediated regulatory programs in glioblastoma. *Mol Syst Biol*. (2012) 8:605. doi: 10.1038/msb.2012.37
26. Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoyannopoulos JA. Circuitry and dynamics of human transcription factor regulatory networks. *Cell*. (2012) 150:1274–86. doi: 10.1016/j.cell.2012.04.040
27. Marbach D, Lamparter D, Quon G, Kellis M, Kutalik Z, Bergmann S. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat Methods*. (2016) 13:366–70. doi: 10.1038/nmeth.3799
28. Ru Y, Kechris KJ, Tabakoff B, Hoffman P, Radcliffe RA, Bowler R, et al. The multiMiR R package and database: integration of microRNA-target interactions along with their disease and drug associations. *Nucleic Acids Res*. (2014) 42:e133. doi: 10.1093/nar/gku631
29. McCarthy DJ, Smyth GK. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics*. (2009) 25:765–71. doi: 10.1093/bioinformatics/btp053
30. Wang X, Terfve C, Rose JC, Markowetz F. HTSanalyzeR: an R/Bioconductor package for integrated network analysis of high-throughput screens. *Bioinformatics*. (2011) 27:879–80. doi: 10.1093/bioinformatics/btr028
31. O'Day E, Lal A. MicroRNAs and their target gene networks in breast cancer. *Breast Cancer Res*. (2010) 12:201. doi: 10.1186/bcr2484
32. Si ML, Zhu S, Wu H, Lu Z, Wu F, Mo YY. miR-21-mediated tumor growth. *Oncogene*. (2007) 26:2799–803. doi: 10.1038/sj.onc.1210083
33. Bhat-Nakshatri P, Wang G, Collins NR, Thomson MJ, Geistlinger TR, Carroll JS, et al. Estradiol-regulated microRNAs control estradiol response in breast cancer cells. *Nucleic Acids Res*. (2009) 37:4850–61. doi: 10.1093/nar/gkp500
34. Barker A, Giles KM, Epis MR, Zhang PM, Kalinowski F, Leedman PJ. Regulation of ErbB receptor signalling in cancer cells by microRNA. *Curr Opin Pharmacol*. (2010) 10:655–61. doi: 10.1016/j.coph.2010.08.011
35. Huang TH, Wu F, Loeb GB, Hsu R, Heidersbach A, Brincat A, et al. Up-regulation of miR-21 by HER2/neu signaling promotes cell invasion. *J Biol Chem*. (2009) 284:18515–24. doi: 10.1074/jbc.M109.006676
36. Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M. Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*. (2007) 8:69. doi: 10.1186/1471-2105-8-69
37. Maragkakis M, Reczko M, Simossis VA, Alexiou P, Papadopoulos GL, Dalamagas T, et al. DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Res*. (2009) 37:W273–6. doi: 10.1093/nar/gkp292
38. Kishore S, Jaskiewicz L, Burger L, Hausser J, Khorshid M, Zavolan M. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat Methods*. (2011) 8:559–64. doi: 10.1038/nmeth.1608
39. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. MicroRNA targets in Drosophila. *Genome Biol*. (2003) 5:R1. doi: 10.1186/gb-2003-5-1-r1
40. Lund AH. miR-10 in development and cancer. *Cell Death Differ*. (2010) 17:209–14. doi: 10.1038/cdd.2009.58
41. de Anda-Jáuregui G, Espinal-Enríquez J, Drago-García D, Hernández-Lemus E. Nonredundant, highly connected microRNAs control functionality in breast cancer networks. *Int J Genomics*. (2018) 2018:9585383. doi: 10.1155/2018/9585383
42. Drago-García D, Espinal-Enríquez J, Hernández-Lemus E. Network analysis of EMT and MET micro-RNA regulation in breast cancer. *Sci Rep*. (2017) 7:13534. doi: 10.1038/s41598-017-13903-1
43. Helwak A, Kudla G, Dudnakova T, Tollervey D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*. (2013) 153:654–65. doi: 10.1016/j.cell.2013.03.043
44. Yu M, Bardia A, Wittner BS, Stott SL, Smas ME, Ting DT, et al. Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. *Science*. (2013) 339:580–4. doi: 10.1126/science.1228522
45. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res*. (2019) 47:D419–26. doi: 10.1093/nar/gky1038

46. Martin FT, Dwyer RM, Kelly J, Khan S, Murphy JM, Curran C, et al. Potential role of mesenchymal stem cells (MSCs) in the breast tumour microenvironment: stimulation of epithelial to mesenchymal transition (EMT). *Breast Cancer Res Treat*. (2010) 124:317–26. doi: 10.1007/s10549-010-0734-1
47. Jeffery J, Sinha D, Srihari S, Kalimutho M, Khanna KK. Beyond cytokinesis: the emerging roles of CEP55 in tumorigenesis. *Oncogene*. (2016) 35:683–90. doi: 10.1038/onc.2015.128
48. Sankar S, Tanner JM, Bell R, Chaturvedi A, Randall RL, Beckerle MC, et al. A novel role for keratin 17 in coordinating oncogenic transformation and cellular adhesion in Ewing sarcoma. *Mol Cell Biol*. (2013) 33:4448–60. doi: 10.1128/MCB.00241-13
49. Aran D, Camarda R, Odegaard J, Paik H, Oskotsky B, Krings G, et al. Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nat Commun*. (2017) 8:1077. doi: 10.1038/s41467-017-01027-z
50. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics*. (2016) 17(Suppl. 2):15. doi: 10.1186/s12859-015-0857-9
51. Vidal Ocabo E, Sayols S, Moran S, Guillaumet-Adkins A, Schroeder MP, Royo R, et al. A DNA methylation map of human cancer at single base-pair resolution. *Oncogene*. (2017) 36:5648–57. (2017). doi: 10.1038/onc.2017.176
52. Wang SE, Lin RJ. MicroRNA and HER2-overexpressing cancer. *MicroRNA*. (2013) 2:137–47. doi: 10.2174/22115366113029990011
53. Biagioni F, Bossel Ben-Moshe N, Fontemaggi G, Yarden Y, Domany E, Blandino G. The locus of microRNA-10b: a critical target for breast cancer insurgence and dissemination. *Cell Cycle*. (2013) 12:2371–5. doi: 10.4161/cc.25380
54. Chan SSK, Kyba M. What is a master regulator? *J Stem Cell Res Ther*. (2013) 3:114. doi: 10.4172/2157-7633.1000e114
55. Li W, Zhang S, Liu CC, Zhou XJ. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*. (2012) 28:2458–66. doi: 10.1093/bioinformatics/bts476
56. Pineda S, Real FX, Kogevinas M, Carrato A, Chanock SJ, Malats N, et al. Integration analysis of three omics data using penalized regression methods: an application to bladder cancer. *PLoS Genet*. (2015) 11:e1005689. doi: 10.1371/journal.pgen.1005689
57. Lee G, Bang L, Kim SY, Kim D, Sohn KA. Identifying subtype-specific associations between gene expression and DNA methylation profiles in breast cancer. *BMC Med Genomics*. (2017) 10:28. doi: 10.1186/s12920-017-0268-z
58. Lê Cao KA, Martin PGP, Robert-Granié C, Besse P. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*. (2009) 10:34. doi: 10.1186/1471-2105-10-34
59. Feng Q, Jiang M, Hannig J, Marron J. Angle-based joint and individual variation explained. *J Multivar Anal*. (2018) 166:241–65. doi: 10.1016/j.jmva.2018.03.008
60. Rohart F, Gautier B, Singh A, Le Cao KA. mixOmics: An R package for-omics feature selection and multiple data integration. *PLoS Comput Biol*. (2017) 13:e1005752. doi: 10.1371/journal.pcbi.1005752

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 18 March 2021
doi: 10.3389/fgene.2021.617512

An Information Theoretical Multilayer Network Approach to Breast Cancer Transcriptional Regulation

Soledad Ochoa[1], *Guillermo de Anda-Jáuregui*[1,2,3]* *and Enrique Hernández-Lemus*[1,2]*

[1] Computational Genomics Division, National Institute of Genomic Medicine, Mexico City, Mexico, [2] Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Mexico City, Mexico, [3] Conacyt Research Chairs, National Council on Science and Technology, Mexico City, Mexico

Breast cancer is a complex, highly heterogeneous disease at multiple levels ranging from its genetic origins and molecular processes to clinical manifestations. This heterogeneity has given rise to the so-called intrinsic or molecular breast cancer subtypes. Aside from classification, these subtypes have set a basis for differential prognosis and treatment. Multiple regulatory mechanisms—involving a variety of biomolecular entities—suffer from alterations leading to the diseased phenotypes. Information theoretical approaches have been found to be useful in the description of these complex regulatory programs. In this work, we identified the interactions occurring between three main mechanisms of regulation of the gene expression program: transcription factor regulation, regulation via noncoding RNA, and epigenetic regulation through DNA methylation. Using data from The Cancer Genome Atlas, we inferred probabilistic multilayer networks, identifying key regulatory circuits able to (partially) explain the alterations that lead from a healthy phenotype to different manifestations of breast cancer, as captured by its molecular subtype classification. We also found some general trends in the topology of the multi-omic regulatory networks: Tumor subtype networks present longer shortest paths than their normal tissue counterpart; epigenomic regulation has frequently focused on genes enriched for certain biological processes; CpG methylation and miRNA interactions are often part of a regulatory core of conserved interactions. The use of probabilistic measures to infer information regarding theoretical-derived multilayer networks based on multi-omic high-throughput data is hence presented as a useful methodological approach to capture some of the molecular heterogeneity behind regulatory phenomena in breast cancer, and potentially other diseases.

Keywords: breast cancer, probabilistic multilayer networks, information theory, co-expression networks, multiomics analysis

OPEN ACCESS

Edited by:
Marieke Lydia Kuijjer, University of Oslo, Norway

Reviewed by:
Giuseppe Jurman, Bruno Kessler Foundation, Italy
Tatiana Belova, University of Oslo, Norway

***Correspondence:**
Enrique Hernández-Lemus
ehernandez@inmegen.gob.mx
Guillermo de Anda-Jáuregui
gdeanda@inmegen.edu.mx

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Genetics

Received: *14 October 2020*
Accepted: *05 February 2021*
Published: *18 March 2021*

Citation:
Ochoa S, de Anda-Jáuregui G and Hernández-Lemus E (2021) An Information Theoretical Multilayer Network Approach to Breast Cancer Transcriptional Regulation. Front. Genet. 12:617512. doi: 10.3389/fgene.2021.617512

1. INTRODUCTION

Cancer is a collection of complex diseases characterized by uncontrolled proliferation (GM., 2000). The complexity of cancer comes, among other sources, from the interaction of different molecular layers and the environment and results in both intra- and inter-tumor heterogeneity (Tian et al., 2011; Burrell et al., 2013; Turashvili and Brogi, 2017). In the case of breast cancer, this heterogeneity has been intended to be captured by tumor sub-classification. Breast cancer has been

thus classified into subtypes with specific molecular signatures and treatment options (Prat et al., 2015), though each altered molecular layer groups differently (Cancer Genome Atlas Network, 2012). Some of these layers, such as gene expression and DNA methylation, have been intensively studied, while others like chromatin accessibility are still gaining attention (Liu, 2020). However, all these layers are interrelated (Wang et al., 2014) and the study of their collective effect calls for multi-omic approaches.

Multi-omic approaches have become possible only recently due to their more stringent methodological requirements. A (relatively large) minimal number of samples are required to find significant patterns, and the needed sample size increases with the noise added per each additional omic. Measurements must refer to the same set of samples, with sustained quality, no matter the differences in data type and range (Kristensen et al., 2014; Bersanelli et al., 2016; Tarazona et al., 2020).

The ability to model heterogeneous and high-dimensional data has made networks a promising tool for multi-omics integration (Vaske et al., 2010; Kim et al., 2012; Wang et al., 2014). For instance, mutual information (MI) networks combining miRNA and gene expressions have been built to gain insight on the regulatory mechanisms behind breast cancer (Drago-García et al., 2017). Such networks pinpointed miR-200 and miR-199 as regulators of the acquisition of epithelial and mesenchymal traits. Another example is the coupling of promoter methylation, transcription factors (TFs), and gene expression in several cancers proposed by Liu et al. Based on those networks, they fitted per target regression models that suggest key cancer processes are jointly regulated by TFs and CpG sites, not by either one alone. Those processes turned out to be different than the processes dominated by copy number variants (Liu et al., 2019).

Gene co-expression networks have been extensively studied in the context of breast cancer subtypes, both from our group (de Anda-Jáuregui et al., 2016; de Anda-Jáuregui et al., 2019; Espinal-Enriquez et al., 2017; Dorantes-Gilardi et al., 2020; García-Cortés et al., 2020; Ochoa et al., 2020) and others (Tang et al., 2018; Bhuva et al., 2019). Here, we are presenting the results on the incorporation of CpG methylation in addition to the study of coding transcripts (for both TFs and other genes) and miRNA expression analyzed in each breast cancer subtype. The goal is to identify CpG sites, TF transcripts (referred to as TF-genes from here on) and miRNAs associated with the biological processes differentially activated in breast cancer, since these may perform potential roles as regulators of the phenotype. Integrated analyses may thus provide us with additional hints toward the possible discovery of synergistic or cooperative effects of these different regulators.

2. MATERIALS AND METHODS

2.1. Data Acquisition

Concurrent-sample measurements of DNA methylation, transcript abundance, and miRNA expression were downloaded from the GDC (https://portal.gdc.cancer.gov/repository) in May 2019. Samples quantified with the Illumina Human Methylation 27 BeadChip, which covers a smaller portion of the genome, were discarded. Instead, we used data obtained with the Infinium HumanMethylation450 BeadChip, which covers 99% of RefSeq genes, at both transcription repressive sites around promoters and transcription favorable sites on the body of genes (Dedeurwaerder et al., 2011). Since these measurements pertain to three distinct techniques: methylation beadchip, RNAseq, and miRNAseq; we treat them as separate omics, here on identified as CpG sites, transcripts, and miRNAs. By including the whole set of features, we wanted to recover the highest possible number of interactions. Subtype classification was also downloaded from the GDC metadata using the `TCGABiolinks` R package (Colaprico et al., 2016).

Each omic was pre-processed independently according to Aryee et al. (2014), Tarazona et al. (2015), and Tam et al. (2015) by using `biomaRt` v95. Preprocessing included filtering of transcripts and miRNAs with low counts, TMM normalization and batch effect correction with ARSyN. Low count thresholds are less than 10 counts per million for transcripts and, less than 5 counts for 25% or more of the samples for every subtype, in the case of miRNAs. Transcripts were also normalized for length and GC content via full method. Annotation was downloaded to tag transcripts coding for TFs (TF-genes).

For methylation data, we discarded sites with over 75% missing values, nonmapped or located within sexual chromosomes or SNPs. Remaining missing values were imputed via nearest neighbors. Resulting beta value matrices were transformed into *M*-value matrices. This way, values of 384,575 methylation probes, 16,475 coding transcripts, and 433 miRNA precursors were obtained for 45 unique samples belonging to the Her2+ subtype, 395 of LumA, 128 of LumB, and 125 of Basal subtypes, plus 75 samples of nontumor (normal adjacent) tissue. All samples correspond to women, ranging in age at diagnosis between 26 and 91 years, and further details can be found in the **Supplementary File 1**.

2.2. Inference of MI Networks

Normalized data matrices for methylation data, coding transcripts, and miRNA expression were merged by sample and used as input to the MI-based ARACNE network deconvolution algorithm (Margolin et al., 2006).

ARACNE calculates mutual information between every pair of features and returns values above a threshold, set either as an MI value or as a permutation *p*-value. There is no restriction on the features that get paired by MI calculation, and it was not required for CpG sites to be on the same chromosome than targets, nor that target promoters carry some TF motif. The only restriction made was for CpG-CpG interactions, which were not calculated due to the space needed to save all possible combinatoria. In a nutshell, pairwise mutual information calculations were performed for the expression patterns for all genes and miRNAs, as well as the beta values for genomewide CpG methylation. Co-expression networks on the different layers were built from the most significant interactions as follows:

Since MI distribution has been shown to change depending on the type of molecules (Drago-García et al., 2017), a unique threshold cannot be set. A unique MI threshold has the risk of discarding significant interactions between molecules whose values simply fall in a lower range or accepting nonsignificant interaction between molecules exhibiting values on a higher than the threshold range. A threshold based on *p*-values induces a

similar problem because MI and *p*-values are roughly inversely proportional. For example, it is possible to see that setting the threshold value to 0.1 in **Figure 2C** would discard most miRNA to transcript interactions while retaining all the interactions among transcripts, and that such pruning of edges would affect differently the distinct networks, producing disparate results due to methodology. Mutual information distributions and their respective threshold values have a direct impact on the topology of the underlying networks and in particular in the degree distributions. So, by choosing MI cutoffs one is indeed imposing an associated network topology.

To overcome this issue, top 10,000 interaction for each type of molecules paired were selected, that is, the 10,000 interactions with the highest MI values linking CpG sites and transcripts (both genes and TF-genes), CpG sites and miRNAs, transcripts (both genes and TF-genes) and miRNAs, and interactions within these two last groups. This way, the topology resulting from such a set of interactions is comparable among cancer subtypes and normal tissue. Thus, we take the focus from the varying MI distributions to a defined topology size. This strategy has been previously validated and used by our group for the reconstruction of biologically relevant networks from high-throughput data (de Anda-Jáuregui et al., 2016).

Fixed bandwidth ARACNE calculations ran with kernel width parameter (h) of 0.165024 for Basal data, 0.211612 for Her2+, 0.12527 for LumA, 0.16567 for LumB, and 0.18679 for normal tissue. To check the significance of the interactions in these networks, maximal MI for each pair of molecules was registered for different *p*-value thresholds. Thresholds with MI values larger than those observed in a network contain the network's interactions. The *p*-value upper limits for the final networks are reported in **Supplementary Table 1**. Finally, MI distributions were compared via Kolmogorov-Smirnov tests with False Discovery Rate (FDR) correction.

Kernel width variation between subtypes can be attributed to the size of the datasets. We estimated z-scores with subsets of the data to evaluate how size differences are affecting the networks. To this end, 100 subsamples of size 45 were taken from luminal and Basal subtypes, and from the normal tissue data. The subsample size was set to 45 for direct comparison with Her2-associated networks. MI was calculated using these subsets and resulting distributions served for z-score calculation. Results can be observed in **Supplementary Table 2**.

By keeping the same number of links in each layer, we are able to directly compare network parameters between layers. However, it should be noted that since the number of possible links increases (quadratically) with the number of nodes, there may be differences in the statistical significance. However, all our networks have an equivalent *p*-value of less than 1E-6 (corresponding to the CpG layer in Her2+ samples, i.e., the layer with more features analyzed for the subtype with the lowest number of samples).

2.3. Functional Enrichment

Independently of network construction, differential expression vs. normal tissue was calculated for every subtype using `limma`'s treat (McCarthy and Smyth, 2009) function with null fold change equal to 1.5. Afterwards, the complete rank of differential expression *t*-values was used as input for a GSEA on each subtype, as implemented in the R package `fgsea` (Sergushichev, 2016), vs. the biological process gene ontology.

Processes with Benjamini and Hochberg adjusted *p*-value lesser than 0.01 were subject to over-representation analysis on the corresponding subtype network. Processes with Benjamini and Hochberg adjusted *p*-value over 0.05 were regarded as nonrepresented in the network. The rest was examined for CpG sites, miRNAs, and TF-genes associated via their MI value with the functionally annotated transcripts, since these serve as potential regulators of the function. For the normal tissue, all the processes significant for a subtype were submitted to the over-representation analysis. There are processes present in a subtype network, but absent from the normal tissue network. This results in a total of 176 processes over-represented in at least one subtype network, from which only 128 have a match in the normal tissue network. In this step, a mean of 59.05% nodes was removed from the MI networks, a breakdown of which can be found in **Supplementary Table 3**.

Resulting networks were visualized using `Cytoscape` (Shannon et al., 2003) with a prefuse force directed layout. Nodes were added to account for the enriched functions in order to find out which biological processes were potentially regulated. Hereafter, these networks are denominated as *final networks* or *functionally enriched networks* to distinguish them from the purely probabilistically inferred networks. These focus on the processes whose expression is the most associated with the subtype, and that rely on interactions with the highest MI; these functions are potentially relevant for the subtypes and so it may be useful to elucidate the associated regulatory patterns.

2.4. Validation of MI Interactions

To check for additional support for the interactions in the final networks, regulator-target databases were reviewed per omic. CpG annotation was taken from Illumina's manifest file, and the genes affected by each site are considered as *validated*. CpG sites on the same chromosome than the target gene are considered as plausible regulators and regarded when adding predictions. These are distinguished from one another as *mapped* and *same chromosome* sites in **Supplementary Table 1**.

Transcription factor targets were downloaded via `tftargets` https://github.com/slowkow/tftargets, a package that queries TRED, ITFP, ENCODE, and TRRUST databases, and the lists compiled by (Neph et al., 2012; Marbach et al., 2016). Only TRRUST TF-targets are considered as validated, since those were manually curated from PubMed articles. The associations between transcripts and miRNAs were sought on DIANA-microT-CDS, ElMMo, MicroCosm, miRanda, miRDB, PicTar, PITA, TargetScan, miRecords, miRTarBase, and TarBase via multiMiR (Ru et al., 2014).

Targets for both TF and miRNA were searched in the tables obtained from each package. The only tuning needed for TF's search was to track ENTREZ gene IDs, HGNC symbols, and Ensembl IDs; this was done according to biomaRt data. Since GDC measurements are identified by precursor miRNA IDs,

TABLE 1 | Networks description.

Edges	Basal	Her2+	LumA	LumB	Normal
CpG–mRNA	2,456	3,847	1,932	4,334	4,732
	(554)	(88)	(536)	(708)	(28)
TF-genes–mRNA	2,735	2,498	1,686	2,746	2,544
	(5)	(2)	(5)	(1)	(14)
miRNA–mRNA	3,483	3,889	2,065	4,074	4,953
	(167)	(226)	(111)	(201)	(284)
mRNA–mRNA	4,189	4,523	2,276	4,709	5,088
Nodes					
Biological processes	109	119	34	123	128
CpG sites	2,254	3,769	1,553	3,638	3,863
Transcripts	4,567	6,356	2,834	5,235	4,733
TF-genes	658	748	375	618	684
miRNAs	433	432	408	433	14

Validated interactions appear between parentheses. Edges correspond to significant statistical dependencies inferred via MI calculations.

while databases use mature miRNA tags, this search requires translation from one to the other using mirBase records.

2.5. Characterization of the Potential Regulators

Looking for differences between subtypes, total regulators of each type were added for every process. Retrieved counts were compared between each subtype and the normal tissue via Fisher tests with FDR correction. Enrichment is only considered if the process has associated regulators of any type, in both the normal tissue and the subtype under evaluation. Statistical tests were one-tailed. Null hypothesis is set to be opposite to expected trends, that is, "greater" for the CpG nodes and "less" for both TF-genes and miRNAs.

To weight the abundance of each regulatory layer, counts per regulator type were divided by the total number of regulators associated with the process, obtaining the percentages displayed in **Supplementary File 2**.

Node topological parameters were calculated over the MI networks, that is, ignoring the *biological processes* nodes, which have to be excluded given the different nature of their associated edges: probabilistically inferred or database curated. Distributions were compared via Wilcoxon rank sum test with continuity correction and p-values were FDR corrected.

2.6. Potential Regulators Comparison

Both intra and inter-subtype comparisons were made. To this end, Jaccard index was calculated for each pair of processes from the same subtype for the intra-subtype comparison and for the same process in different subtypes for the inter-subtype comparison. Inter-subtype contrasts count edges instead of nodes, because in this case, the interest is on conserved regulatory interactions. Obtained distributions were evaluated via Kolmogorov–Smirnov tests with FDR correction.

The number of potential regulators either shared or exclusive between processes of the same subtype was evaluated via Fisher tests with the corresponding alternative hypothesis set "greater" for the CpG sites, and "less" for TF-genes and miRNAs, as previously stated.

All the code used for the described analysis is available at https://github.com/CSB-IG/MI-MultiOmics.git.

3. RESULTS

MI networks were constructed for each breast cancer subtype and for normal tissue combining three different omics: CpG methylation, transcript gene expression, and miRNA expression. The second omic includes two layers of information, regulated genes, and TF-genes. No restriction was made on the features that can get paired by MI calculation, CpG sites can get linked to targets on a different chromosome, and TFs may associate with targets without the akin binding motifs. Let us recall that mutual information does not assume any a priori mechanism and relies instead on statistical dependencies. **Table 1** presents all the different networks of MI-inferred potential gene regulators (CpG-mRNA, TF-gene-mRNA, miRNA-mRNA, mRNA-mRNA) plus the biological processes associated with them.

MI networks went through two pruning steps, first by edge significance (see section 2.2) and then by functional annotation of the nodes (see section 2.3). The first one retains only the most significant interactions, i.e., those with the largest MI. For the second pruning, biological processes with significant GSEA enrichment scores were mapped to the networks, keeping only the nodes involved in an enriched process and their first neighbors. For the normal tissue, all the processes significant for a subtype were subjected to over-representation analysis. This way, only nodes linked to transcripts involved in a process altered in the subtype are kept. Then, final networks carry only CpG–transcript, miRNA–transcript, and transcript–transcript interactions with the highest MI. The hypothesis is that nodes with gene expression regulatory roles may regulate the associated biological process. This would be partially explained, if regulators co-vary (even in a nonlinear fashion) with their targets, thus becoming detectable as MI statistical dependencies. It is relevant

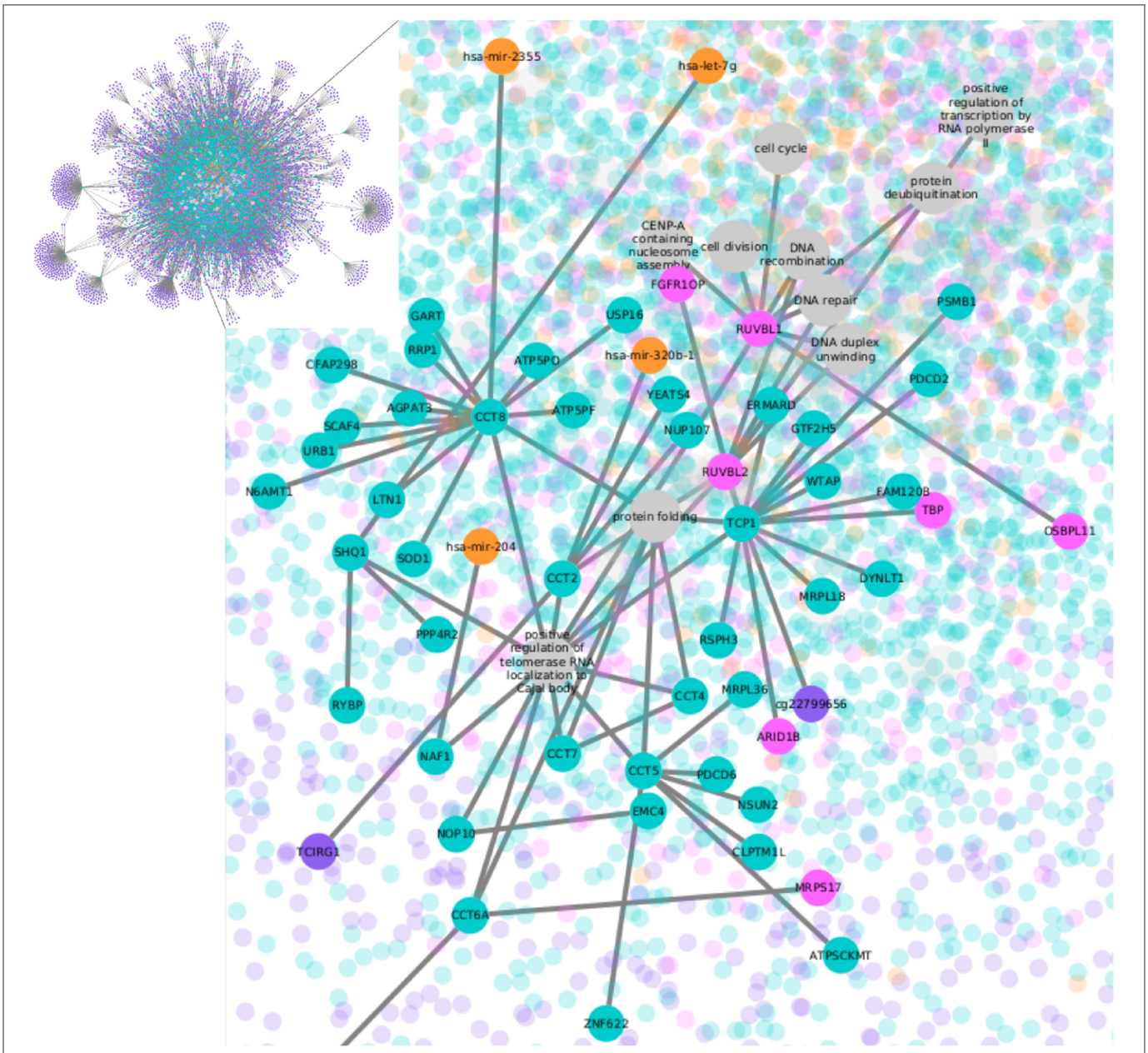


FIGURE 1 | LumB subtype network. Nodes represent CpG sites in purple, transcripts in green, TF-genes in pink, miRNAs in orange, and biological processes in gray. The whole network is shown in the upper left box; the rest of the figure contains a zoom-in.

to recall, however, that regulatory mechanisms are proxied here by the information given by the omics under study. Other regulatory mechanisms—including those of (explicit) chromatin remodeling, as well as post-transcriptional and post-translational modifications among others—may not be fully accounted by the statistical dependencies structures just outlined.

To assess the contribution of linear correlation measures, we are including further calculations in **Supplementary Figures 1–5** to show how many of the MI edges would be lost if the criterion was instead an FDR-corrected Pearson correlation with an associated $p < 0.05$.

To identify unequivocally the functions linked to each transcript, nodes representing the biological processes were added, resulting in multipartite graphs as the one shown in **Figure 1**. The multipartite nature of the network comes from the three different molecules (CpG sites, transcripts, and miRNAs) associated with the biological process nodes. There are also two kinds of edges: (1) MI edges, which indicate molecule covariation, and (2) functional annotation edges, which make explicit the link of a transcript and a process. All the five networks, four for the breast cancer subtypes and one for the normal tissue, consist of one giant single connected component. As expected,

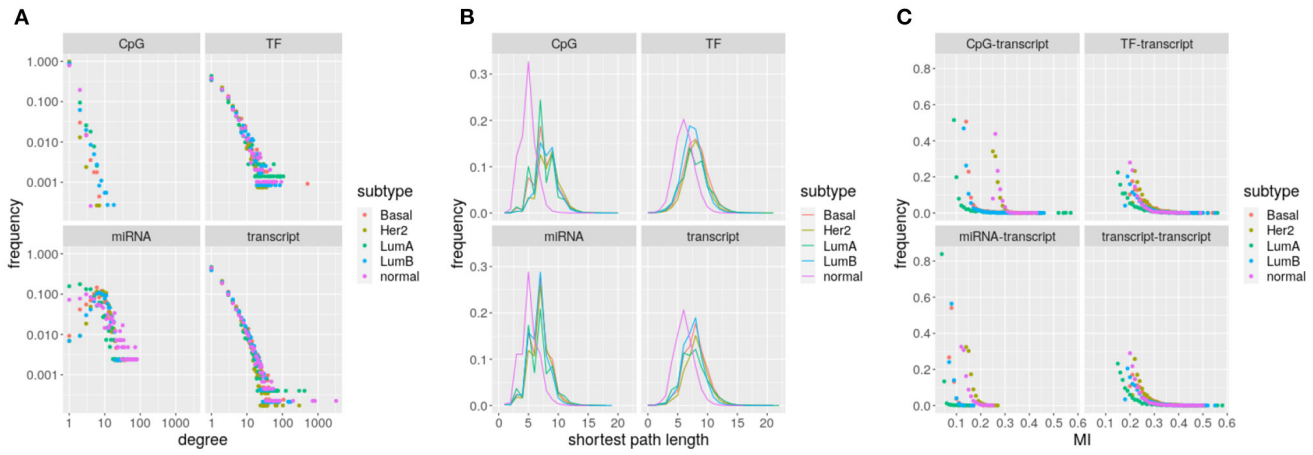


FIGURE 2 | Networks parameters: **(A)** degree, **(B)** shortest path, and **(C)** edges with mutual information. The subtype depicted by each network is indicated by the color code.

CpG methylation, which has the largest number of features, is the most represented omic in the networks.

By contrasting the molecules paired with databases on regulator-target, we can see how many of the found interactions were already known. Interactions absent from the databases can be new, previously unknown relationships, or simply indirect associations caused by the statistical co-variation of the molecules. Between 1.67 and 11.47% of the interactions linking a transcript with a potential regulator, that is a CpG, a TF-gene, or a miRNA, have been validated. The number of validated edges per subtype is shown in **Table 1**. If predictions are included (see section 2.4), 8.26–23.52% of the interactions have additional support. The effect on the networks of considering only some of the potential regulatory CpGs can be seen in **Supplementary Figure 6**. A large number of TF target predictions are based on ChIP-seq experiments, not necessarily performed on breast tissue, which may lower such matches.

Having described the general features of the five networks (one for each tumor subtype plus the one for normal tissue), we proceeded to search for differences between the behavior of the different omics among subtypes. Focus was made on differences on the potential regulators, since this could translate to regulatory features behind the subtypes.

3.1. Network Parameters Vary Between Omics

As stated earlier, there are two types of edges in the networks, edges that account for co-expression (i.e., significant statistical dependency) with a given value of MI, and edges that record functional annotation as presented in curated databases. Given the difference of meaning, interactions need to be analyzed separately.

Focusing only on MI edges, the number of components grows from 1 to hundreds. Average degree is around 3 for all the networks, but distributions vary between omics (Wilcoxon rank sum test q-value $\leq$ 1.666712e-22, **Figure 2A**). Though TF-genes and gene transcripts are measured by the same omic, distributions are significantly different (Wilcoxon rank sum test q-value $\leq$ 0.0237) for the five networks. The case of miRNAs stands out because distributions are not scale-free like. CpG sites show the lowest degrees, with an average of 89.42% nodes connected only with another node. Thus, most CpG sites do not contribute to network communication as they do not interlink paths.

The constrained (bounded) degree distribution of CpGs translates into a large portion of unreachable target nodes, an average of 32.23% of targets cannot be reached from some CpGs. Consistently, miRNAs have an average of 19.71% of unreachable targets, which is the lowest frequency. Despite range similarity, distributions change significantly across omics and between tumor subtypes and normal tissue (Wilcoxon rank sum test q-value $\simeq 0$). Again, distributions for TF-genes and gene transcripts are significantly different (see **Figure 2B**, Wilcoxon rank sum test q-value $\leq$ 0.0002). The shift in the position of the peak in breast cancer subtypes relative to normal tissue suggests a loss of communication.

Edges also differ depending on the omics involved. Differences on mutual information distributions between omics and subtypes are significant (Kolmogorov–Smirnov q-value $\leq$ 5.53264e-06). TF-genes and gene transcripts follow the same distribution on each network. It is noticeable how small is the range of miRNA interactions and how CpG distributions segregate.

In **Table 1** and **Figure 2**, we have characterized the interactions occurring within and between different omics in each molecular subtype of breast cancer. We may appreciate that both intra-layer and inter-layer interaction sets are specific to each biological condition. In what follows, we will now leverage both the monolayer and multilayer interactions to further elucidate biological functions associated with each molecular subtype.

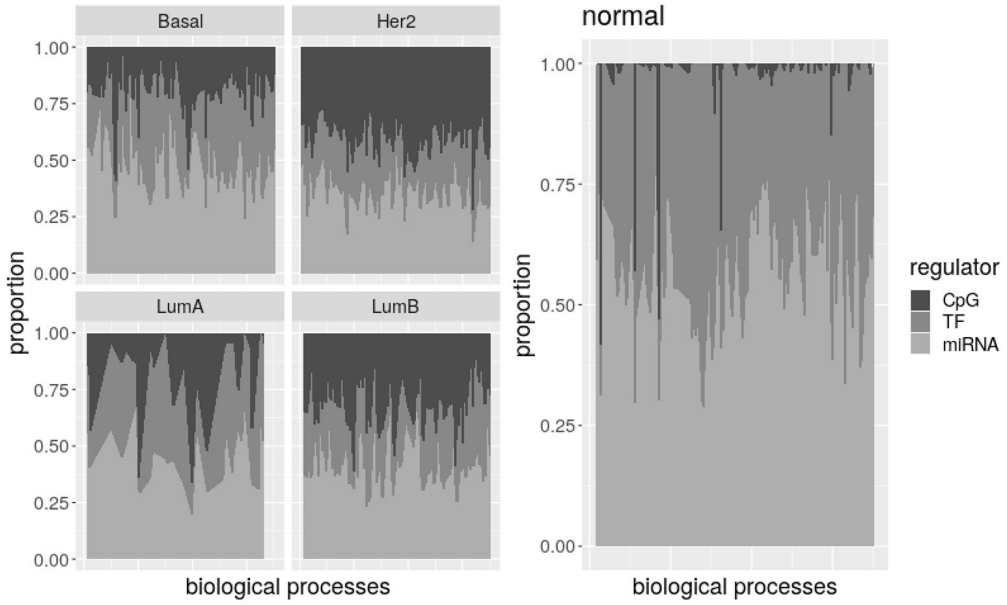


FIGURE 3 | Potential regulators per biological process. There is an area plot per network. Each column is a process. The proportion of regulators of each type associated with the process is denoted by the grayscale. All the processes together show how common are the three potential regulators in the subtype.

3.2. Representation of Potential Regulators Changes With the Subtype

To further explore the differences among potential regulators, its abundance per biological processes was calculated. To this end, total number of CpG, TF-genes, and miRNA nodes were obtained for each biological process. The proportion of regulators of each type is shown in **Figure 3** as a simple measure of the impact a regulatory layer has in a given subtype. A version of this figure with labels for biological processes and the corresponding table are available as **Supplementary Material**.

Despite variability, it is evident that the number of CpG nodes increases on breast cancer subtypes relative to normal tissue, while TF-genes and miRNA numbers of nodes are lower. The plot for Luminal A subtype is less noisy because this subtype has less processes on its network. Nevertheless, by comparing processes represented in each subtype and normal tissue, we found most processes are significantly enriched of CpG nodes in the Basal, Her2+, and LumB subtypes. Simultaneously, TF-genes and miRNAs are significantly under-represented on more than half of the processes in the Her2+ and LumB networks. Additionally, between 20 and 33% of the Basal- and LumA-associated processes show under-representation of TF-genes and miRNAs, and almost half of LumA processes are enriched of CpG nodes.

If potential regulators are actually regulating their associated processes, this may indicate transcriptional and post-transcriptional regulations are subdue in breast cancer subtypes while epigenetic regulation gains strength. By considering the combined effect across layers (inter-layer regulation) as well as the effects on a single type of molecular interaction, as given by each omic dataset (intra-layer regulation), it is possible to develop a deeper understanding of cross-regulatory effects. This will be considered in the next subsections in the context of the different tumor subtypes.

3.3. Normal Interactions With Potential Regulators Are Almost Absent in Breast Cancer Networks

Having seen that the abundance of complete regulatory layers is not maintained across subtypes, we wondered what happens to specific regulatory interactions. With this in mind, we calculated the extent to which interactions with potential regulators are shared among networks by calculating their associated Jaccard indices. The Jaccard index weights the size of the intersection between two sets with the size of their union. In other words, it counts what fraction of the elements is shared from the total. This way, sets of different extensions are assigned values between 0 and 1, and can be objectively compared.

From the total of 176 biological processes enriched in any network, 86.36% appear in at least two subtypes and also are able to share edges. Interactions with miRNAs are poorly shared, while TF-genes and CpG-edges reach a similar maximum but following different distributions (Kolmogorov–Smirnov test q-value $\leq$ 2.498002e-16). Links with any regulator are almost not shared between the breast cancer subtypes and the normal tissue (Kolmogorov–Smirnov test q-value $\leq$ 1.541449e-06), but TF-genes are visibly more shared. The five biological processes with the highest Jaccard index are shown in **Figure 4**.

Localization of telomerase RNA (hTR) to the Cajal body has the highest index for miRNAs for the sharing among Basal and Her2+ networks. This process is also the fifth for TF-genes, but pairing Her2+ and LumB. **Figure 5** shows that the elevated Jaccard indices are driven by only few shared interactions among sets of small size. Although potential regulation changes, the

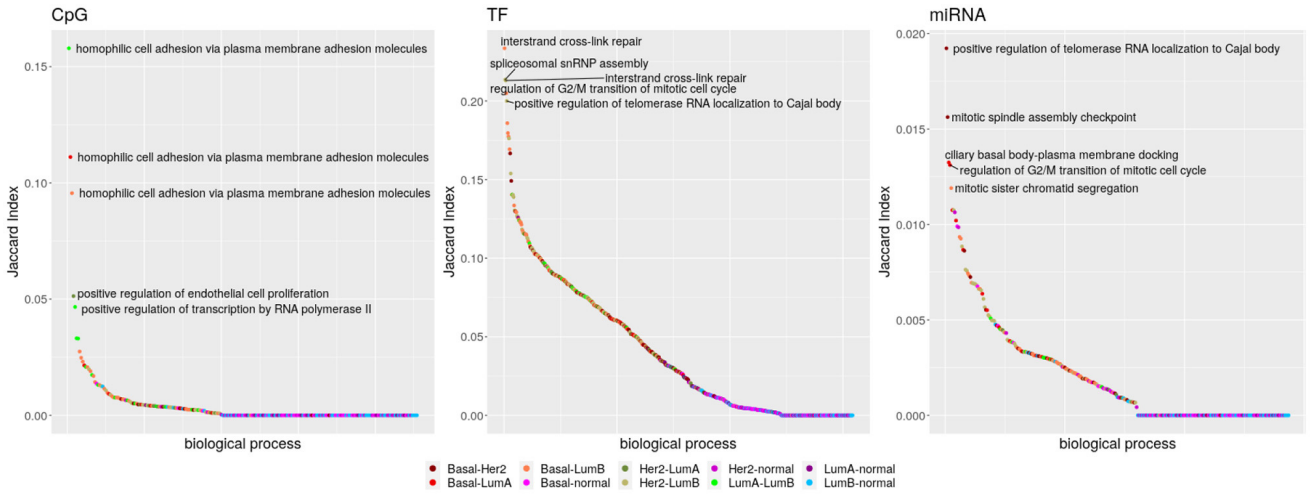


FIGURE 4 | Inter-subtype sharing. Biological processes are symbolized by dots, ordered from those sharing more potentially regulatory interactions to those sharing less. The color code indicates which networks are contrasted. Comparisons with the Basal network are in reddish colors, and those with Her2+ are in dry greens; indexes involving the normal tissue go from pink to blue and those between the luminal subtypes are in bright green.

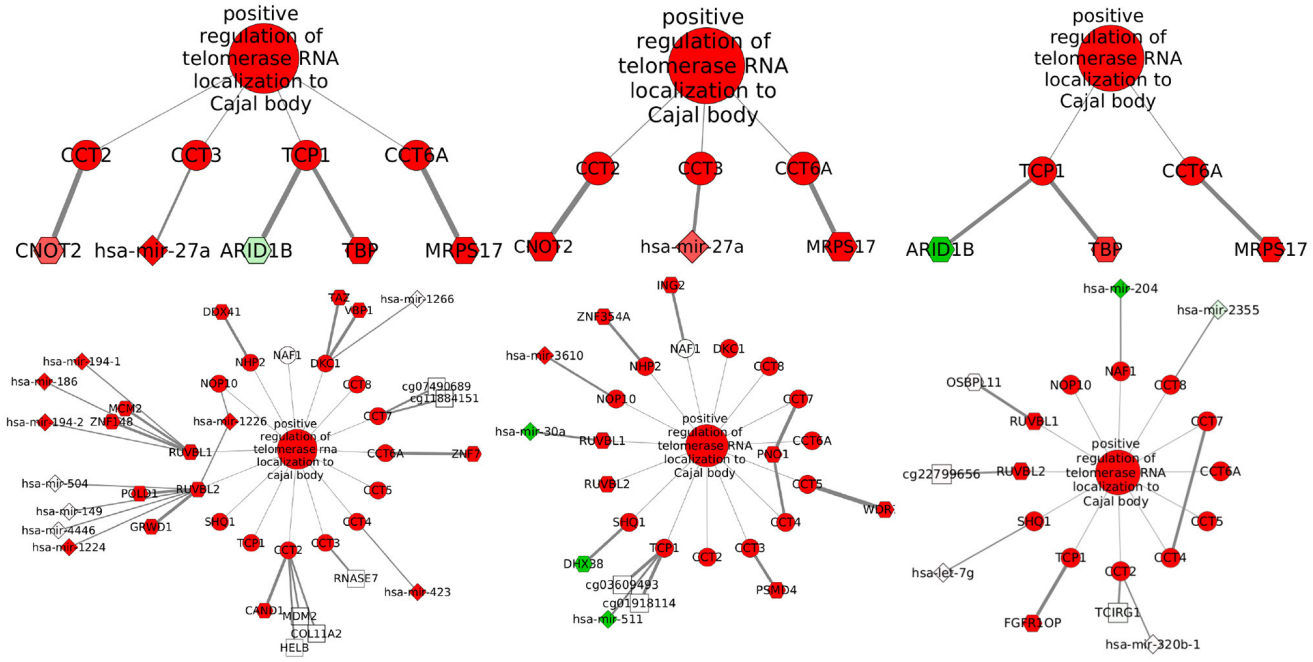


FIGURE 5 | Potential regulators associated with telomerase localization to Cajal body. From left to right, subgraphs with shared **(top)** and exclusive **(bottom)** interactions are shown for Basal, Her2+, and LumB subtypes. Nodes are colored in red if differential expression values or GSEA normalized enrichment scores are positive or in green if values are negative. Node transparency represents statistical significance. Nodes for biological processes and transcripts are circles, TF-genes are hexagons, miRNAs are diamonds, and CpG sites are squares. When the altered gene is known, its name is on the CpG label. Edge thickness represents MI values.

process is equivalently activated in these three subtypes. The interaction linking Chaperonin Containing TCP1 Subunit 6A (CCT6) with Mitochondrial Ribosomal Protein S17 (MRPS17) is shared across these three subtypes, but may be an artifact of the physical proximity of the genes.

3.4. Within Subtypes, CpG Nodes Are Exclusive of Processes, but miRNAs Do Not

For a complementary perspective, we checked if regulators are shared between the distinct biological processes enriched in a

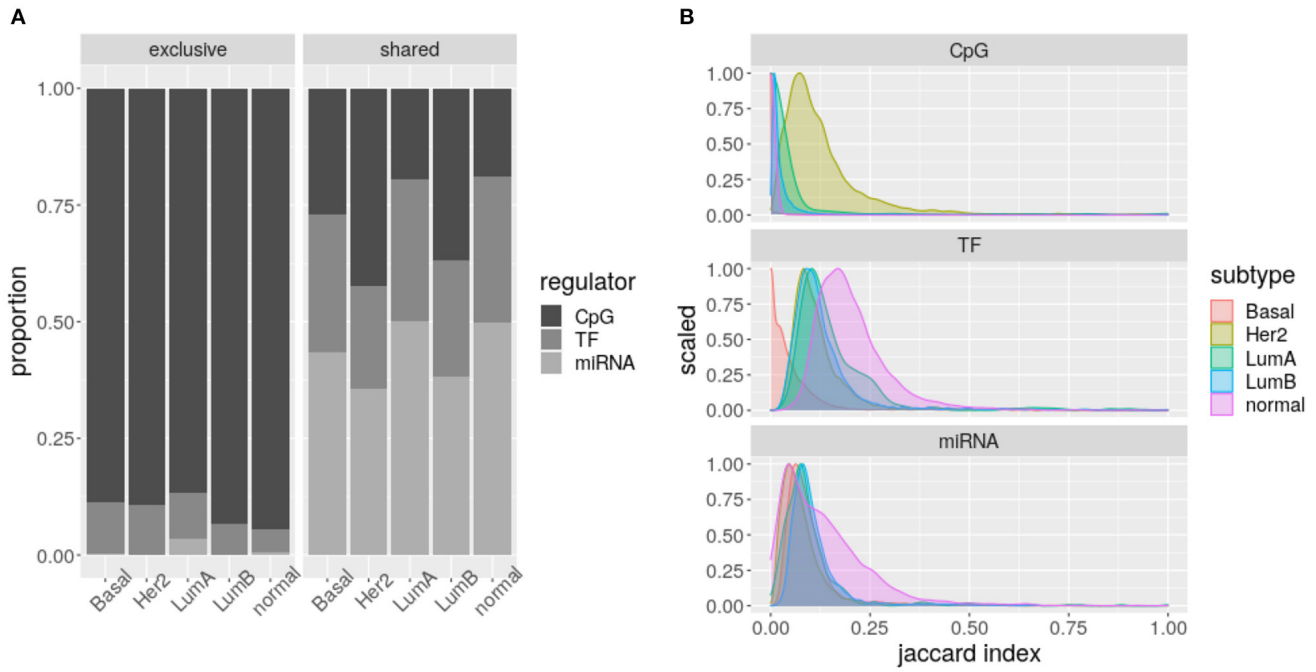


FIGURE 6 | Intra-subtype sharing: **(A)** Proportion of regulators shared or exclusive, across biological processes per subtype. Each column represents a different network. The grayscale indicates the potential regulator. **(B)** Jaccard Index distributions per regulator. The color code indicates the network represented.

single subtype. Degree distributions suggest that CpG sites are exclusive, while miRNAs and TF-genes are shared.

Figure 6A shows how CpG sites are mostly exclusive of one biological process (Fisher test q-value $\leq$ 1.949349e-67), while TF-genes and specially miRNAs are shared between various processes (Fisher test q-value $\leq$ 1.411310e-11). That is, miRNA expression seems to connect different biological processes, while for CpG methylation this effect is much lower.

When calculating the Jaccard index of the biological processes enriched for each subtype and regulator, significantly different distributions are obtained (Kolmogorov–Smirnov test q-value $\leq$ 0.0221, **Figure 6B**). Consistently, as presented in **Figure 6A**, these distributions show CpG sites are less shared, but TF-genes seem to be more shared than miRNAs. The CpG sites of Her2+ and the TF-genes of Basal subtypes call for attention.

4. DISCUSSION

With the aim of exploring potential regulatory patterns of breast cancer subtype expression, we reconstructed via mutual information, multi-omics networks, functionally enriched in GO biological processes. The hypothesis is that there may be a transitive property between the regulators of a transcript and the function associated with the transcript.

This way, potential regulators emerging from the networks are associated with the biological processes significantly enriched. Potential regulators separate domains topologically from non-regulatory transcripts and from each other. Degree distributions are coherent with the pattern of exclusivity and sharing across processes, observed later for CpG sites and TF-genes–miRNAs, respectively. Both results coincide with what is known for the molecule types. Namely, CpG sites have a rather local effect (Li and Zhang, 2014), while TF-genes and miRNAs are *promiscuous*, spanning through a much wider chromosome range (Cho, 2007).

Given the pattern of sharing/exclusivity across processes, one could expect that targeting DNA methylation may drive focused changes, while miRNAs and TF-genes targeting may show pleiotropy. However, current modulators of DNA methylation act over the whole genome, making impossible to change sites related to specific processes. On the contrary, CpG sites linked to specific processes may have potential as predictors of process alteration. Such potential is promising given the early timing of methylation alterations in other cancer types (Vrba and Futscher, 2019). For example, there are 19 CpG sites associated with DNA damage checkpoint in Her2+ subtype, suggesting a possible monitoring mechanism. Nevertheless, it would be necessary to have a whole new project to test the predictability of such sites. The value of the multilayer networks presented here is to propose this kind of hypothesis among all possible combinations, though they need further testing.

To verify that CpG exclusivity per process is not induced by the omission of CpG–miRNA and miRNA–miRNA interactions, non-functionally enriched networks were revisited (**Supplementary Figure 7**). Distributions still change per omic (Wilcoxon rank sum test q-value $\leq$ 4.657478e-16), while the percentage of CpG nodes with degree equal to one is maintained above 90%, indicating that observations made for the first neighbors are relevant when considering farther neighbors. By considering the top 10,000 MI interactions per paired molecules,

we observed that CpG sites do not significantly participate in the regulatory circuitry flow but are often endpoints.

Shortest-paths distributions point out to a decrease in communication independently of the omic observed. This is in line with the under-representation of TF-genes and miRNAs detected specially in Her2+ and Luminal B associated processes. To reconcile communication reduction with over-representation of CpG sites on the subtypes, it is necessary to remember that most CpG nodes do not participate in network connection. These layer level patterns consistently match literature reports on alteration of CpG methylation (Cancer Genome Atlas Network, 2012; Berger et al., 2018), and miRNA expression in breast cancer (O'Day and Lal, 2010; Bertoli et al., 2015; Klinge, 2018).

Two subtype-specific patterns attracted our attention, elevated sharing of CpG nodes between the processes enriched for the subtype Her2+, and decreased sharing of Basal TF-genes. The 2,112 CpG sites shared by Her2+ processes are all over the genome, with a slight increase in chromosomes 1 and 17. While chromosome 1 has been reported as severely affected by differential methylation (Lindqvist et al., 2014), the characteristic amplification of chromosome 17 cannot be fully accounted for the excess sharing. Only 76 from the 1576 genes affected by shared CpG sites co-amplify with the *Her2* gene. Similarly, only 22.91% of affected genes have evidence of AR regulation, a TF postulated to crosstalk with Her2 amplification (Daemen and Manning, 2018).

The other pattern that caught our attention is the decrease in TF-genes linking any two processes in the network for the Basal subtype. This is not caused by a decrease in TF-genes, since the quantity of TF-gene nodes associated with the processes is equivalent for all the networks. Uniqueness of biological processes in the Basal network are neither responsible, seeing that only 6 processes are exclusive for this subtype. Instead, we speculate the pattern is related to promoter accessibility because of ATAC-seq data groups tumors in Basal and non-basal networks (Corces et al., 2018). Further characterization finds a pro-metastasis open-chromatin signature elevated in the Basal subtype (Cai et al., 2020). By its side, protein level measures integrated with copy number normalized gene expression suggest TF-genes as relevant drivers of this subtype (Koh et al., 2019).

Only one edge level pattern was found, but it is a remarkable one. Interactions with regulatory potential are poorly shared among all networks, but the edges of the normal tissue network are almost endemic, especially in the case of CpG sites and miRNAs. If we conform to the idea that DNA methylation preserves cell type identity (Szyf, 2012), our results advert mammary gland defining methylation has been lost in processes like T-cell receptor signaling pathway and inflammatory response.

Localization of hTR to the Cajal body is a biological process linked with cancer cell's unlimited division, given that these organelles have been implicated in the biogenesis of telomerase (Tomlinson et al., 2008). Associated subgraphs exhibit how few edges are shared across subtypes and suggest a convergence of different regulatory schemes to a single outcome. The relative uniformity of enrichment scores across subtypes (**Supplementary Figure 8**) indicates this could be common. Such pattern is important because the way a tumor gains an expression signature might create different vulnerabilities. An example is given by tumors compatible with Her2-enriched expression, but lacking the mutation that makes tumors sensitive to targeted treatment (Godoy-Ortiz et al., 2019).

We must, however, stress that one limitation of the current approach resides on the relatively small sample size. This is a constraint due to lack of availability of a larger dataset comprising the same types of multi-omic data. Limited availability of additional independent datasets also precluded us to validate our findings on an independent cohort. To partially alleviate this, we have resorted to subsampling procedures and null models. The effect of data size differences can be seen in **Supplementary Figure 9** and **Supplementary Table 2**. **Supplementary Table 2** and **Supplementary Figure 9** show the dispersion between MI values estimated with the whole set of samples as well as values obtained through subsampling, for the interactions with the lowest, most varying significance, those between miRNAs and transcripts. Though subsampling repetition is low (100), it catches a tendency toward small z-scores and noisier low subsampled MI values. This means higher z-scores are not necessarily bad, since the large difference between complete and subsampled values maintains points at the top of the range. Altogether, subsampling suggests adding samples would reach higher MI values, but would not alter the ranking dramatically, which supports the (cautious) usage of datasets such as the one used for Her2+. Nevertheless, our analysis could only take advantage of an increase of the number available samples.

As with other areas of molecular biology, one driving force behind the development of multi-omics is the expectation that the results from these technologies may lead to novel pharmacological interventions (de Anda-Jáuregui and Hernández-Lemus, 2020). Nevertheless, the translation from the identification of a perturbation to clinical implementation is not straightforward (Silverman et al., 2020). In this regard, pharmaceutical interventions in each of the analyzed layers are unevenly distributed: drugs that have effects on epigenetic modifications such as methylation have not attained the efficacy that was expected (Buocikova et al., 2020), although they remain an important research area. Meanwhile, gene expression has been able to identify biomarkers as well as drug repurposing opportunities (Mejía-Pedroza et al., 2018; Koudijs et al., 2019). In this context, the type of analyses that we present here provides the opportunity to identify not only the deregulation features in each regulatory layer but also the way it connects to other molecular elements. As such, the opportunity to modulate virtually undruggable targets through the control of its neighbors may help unblock therapeutic opportunities. However, as we mentioned previously, the path from these initial data analyses toward a translational and eventually a clinical setting is long and not necessarily direct.

4.1. Summary of Findings

In brief, the main findings that have been derived from our analysis may be summarized as follows:

- For networks associated with tumor subtypes:
  - Shortest paths are longer for the four subtypes than for the normal tissue.
  - Most biological processes (over 85%) are enriched for CpG nodes in Basal, Her2+, and LumB. Only 41.38% of the processes in LumA are enriched for CpG nodes.
  - Most biological processes (over 50%) are under-represented of TF-gene and miRNA nodes in Her2+ and LumB.
  - Interactions with CpGs and miRNAs found in normal tissue network are near endemic.
  - Her2+ CpG nodes are more shared between processes than expected.
  - Basal TF-gene nodes are less shared between processes than expected.
- For differences in the representation of different omics:
  - CpG nodes tend to show degree = 1, which translates into exclusivity for each process.
  - TF-genes have fewer nodes with degree = 1, and miRNAs have even less. Consistently, these nodes are more shared between processes thus participating in concerted network communication.
  - miRNAs degree distribution shape is remarkably different.
- For shared interactions:
  - Those with CpGs and miRNAs are less maintained than those with TF-genes.

5. CONCLUSIONS

Together, the observations made from multi-omic mutual information networks for the different breast cancer subtypes build a landscape of the differential influence the distinct regulatory layers may exert over the phenotypes. This expands our understanding of breast cancer associated regulatory phenomena and poses possible treatment alternatives to be further explored. For example, now that there is evidence that CpG methylation coordinates with the expression of Her2-associated genes involved in most biological processes more than in any other subtype, experiments with de-methylation agents on this specific subtype seem relevant to analyze.

So far, the interaction between regulatory layers has been overlooked due to the paucity of data and inadequacy of methods. Yet, mutual information calculations and the available algorithms just presented have no formal restriction to handle different omics, unlike other correlation measures MI allows to handle variables with disparate dynamic ranges as it relies in the probability distributions, and has proven capable to retrieve single omics regulatory interactions. Results obtained with the multi-omic setting are encouraging, though refinement of post-MI analysis is needed and is indeed a further avenue of research within our group.

In order to capture CpG methylation and miRNAs linked to biological processes via the interaction with one another, a more sophisticated method would be needed. For example, a computationally expensive recovery of all the paths between transcripts associated with functions. Another possible improvement would be the implementation of a multi-omics data processing inequality (DPI). DPI states that the edge with the smaller MI in a triangle can be filtered out as indirect. However, MI distribution changes for every type of omics paired complicating MI comparisons. Perhaps a better alternative will be to resort to tensor representations of probabilistic multilayer networks (Hernández-Lemus, 2020).

It is also pertinent to recall that higher mutual information does not translate into causal interactions. The so-called *potential regulators* may simply co-vary with transcript expression, or causality may be dependent on an intermediate node. Even if linked CpGs sites regulated gene expression, omics that are not included like copy number variation may also play relevant roles. To identify the potential regulators whose patterns are most related to transcripts expression, there are other strategies available (Lê Cao et al., 2009), which may benefit from MI interaction scores (Koh et al., 2019). There are however more insights to be extracted from the multi-omics networks yet.

With the set of potential regulators associated with a biological process, we aspire to multi-layer regulatory models that include examples like the one described for miRNA processing enzymes Drosha and Dicer (Rupaimoole et al., 2014). Here, we present general results, but particular cases can be further examined within this general approach. When the focus is on particular models, the distinct regulators connected to single gene allow the proposal of hypothesis about synergy and antagonism among regulation layers. Nevertheless, this approach calls for a much more detailed scrutiny.

All in all, due to the relative simplicity and generalizability of the approach, the use of combined probabilistic modeling and knowledge discovery in databases presented here allows for the inference of regulatory models that may be refined by resorting to more specialized techniques, both experimental and computational.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

SO organized data, developed code, performed calculations, analyzed data, and drafted the manuscript. GA-J contributed to the methodological approach, analyzed data, discussed results, and co-supervised the project. EH-L envisioned the project, devised the methodological strategy, designed the study, contributed to the methodological approach, analyzed data, discussed results, reviewed the manuscript, and supervised the project. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Consejo Nacional de Ciencia y Tecnología [SEP-CONACYT-2016-285544 and FRONTERAS-2017-2115], and the National Institute of Genomic Medicine, México. Additional support has been granted by the Laboratorio Nacional de Ciencias de la Complejidad, from the Universidad Nacional Autónoma de México. EH-L is recipient of the 2016 Marcos Moshinsky Fellowship in the Physical Sciences.

ACKNOWLEDGMENTS

The authors want to thank Gabriela Graham for her support with language editing and proofreading of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.617512/full#supplementary-material

REFERENCES

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., et al. (2014). Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium DNA methylation microarrays. *Bioinformatics* 30, 1363–1369. doi: 10.1093/bioinformatics/btu049

Berger, A. C., Korkut, A., Kanchi, R. S., Hegde, A. M., Lenoir, W., Liu, W., et al. (2018). A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell* 33, 690–705.e9.

Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., et al. (2016). Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics* 17(Suppl. 2):15. doi: 10.1186/s12859-015-0857-9

Bertoli, G., Cava, C., and Castiglioni, I. (2015). MicroRNAs: New biomarkers for diagnosis, prognosis, therapy prediction and therapeutic tools for breast cancer. *Theranostics* 5, 1122–1143. doi: 10.7150/thno.11543

Bhuva, D. D., Cursons, J., Smyth, G. K., and Davis, M. J. (2019). Differential co-expression-based detection of conditional relationships in transcriptional data: comparative analysis and application to breast cancer. *Genome Biol.* 20, 1–21. doi: 10.1186/s13059-019-1851-8

Buocikova, V., Rios-Mondragon, I., Pilalis, E., Chatziioannou, A., Miklikova, S., Mego, M., et al. (2020). Epigenetics in breast cancer therapy-new strategies and future nanomedicine perspectives. *Cancers* 12:3622. doi: 10.3390/cancers12123622

Burrell, R. A., McGranahan, N., Bartek, J., and Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* 501, 338–345. doi: 10.1038/nature12625

Cai, W. L., Greer, C. B., Chen, J. F., Arnal-Estapé, A., Cao, J., Yan, Q., et al. (2020). Specific chromatin landscapes and transcription factors couple breast cancer subtype with metastatic relapse to lung or brain. *BMC Med. Genomics* 13:33. doi: 10.1186/s12920-020-0695-0

Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70. doi: 10.1038/nature11412

Cho, W. C. S. (2007). Oncomirs: the discovery and progress of microRNAs in cancers. *Mol. Cancer* 6:60. doi: 10.1186/1476-4598-6-60

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., et al. (2016). TCGAbiolinks: an R/bioconductor package for integrative analysis of TCGA data. *Nucl. Acids Res.* 44:e71. doi: 10.1093/nar/gkv1507

Corces, M. R., Granja, J. M., Shams, S., Louie, B. H., Seoane, J. A., Zhou, W., et al. (2018). The chromatin accessibility landscape of primary human cancers. *Science* 362:eaav1898. doi: 10.1126/science.aav1898

Daemen, A., and Manning, G. (2018). Her2 is not a cancer subtype but rather a pan-cancer event and is highly enriched in AR-driven breast tumors. *Breast Cancer Res.* 20:8. doi: 10.1186/s13058-018-0933-y

de Anda-Jáuregui, G., Alcalá-Corona, S. A., Espinal-Enríquez, J., and Hernández-Lemus, E. (2019). Functional and transcriptional connectivity of communities in breast cancer co-expression networks. *Appl. Netw. Sci.* 4, 1–13. doi: 10.1007/s41109-019-0129-0

de Anda-Jáuregui, G., and Hernández-Lemus, E. (2020). Computational oncology in the multi-omics era: state of the art. *Front. Oncol.* 10:423. doi: 10.3389/fonc.2020.00423

de Anda-Jáuregui, G., Velázquez-Caldelas, T. E., Espinal-Enríquez, J., and Hernández-Lemus, E. (2016). Transcriptional network architecture of breast cancer molecular subtypes. *Front. Physiol.* 7:568. doi: 10.3389/fphys.2016.00568

Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C., and Fuks, F. (2011). Evaluation of the infinium methylation 450k technology. *Epigenomics* 3, 771–784. doi: 10.2217/epi.11.105

Dorantes-Gilardi, R., García-Cortés, D., Hernández-Lemus, E., and Espinal-Enríquez, J. (2020). Multilayer approach reveals organizational principles disrupted in breast cancer co-expression networks. *Appl. Netw. Sci.* 5, 1–23. doi: 10.1007/s41109-020-00291-1

Drago-García, D., Espinal-Enríquez, J., and Hernández-Lemus, E. (2017). Network analysis of EMT and MET micro-RNA regulation in breast cancer. *Sci. Rep.* 7:13534. doi: 10.1038/s41598-017-13903-1

Espinal-Enriquez, J., Fresno, C., Anda-Jáuregui, G., and Hernández-Lemus, E. (2017). RNA-seq based genome-wide analysis reveals loss of inter-chromosomal regulation in breast cancer. *Sci. Rep.* 7, 1–19. doi: 10.1038/s41598-017-01314-1

García-Cortés, D., de Anda-Jáuregui, G., Fresno, C., Hernandez-Lemus, E., and Espinal-Enríquez, J. (2020). Gene co-expression is distance-dependent in breast cancer. *Front. Oncol.* 10:1232. doi: 10.3389/fonc.2020.01232

GM., C. (2000). *The Development and Causes of Cancer. The Cell: A Molecular Approach, 2nd Edn*. Sunderland: Sinauer Associates.

Godoy-Ortiz, A., Sanchez-Muñoz, A., Chica Parrado, M. R., Álvarez, M., Ribelles, N., Rueda Dominguez, A., et al. (2019). Deciphering her2 breast cancer disease: biological and clinical implications. *Front. Oncol.* 9:1124. doi: 10.3389/fonc.2019.01124

Hernández-Lemus, E. (2020). On a class of tensor Markov fields. *Entropy* 22:451. doi: 10.3390/e22040451

Kim, D., Shin, H., Song, Y. S., and Kim, J. H. (2012). Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *J. Biomed. Inform.* 45, 1191–1198. doi: 10.1016/j.jbi.2012.07.008

Klinge, C. M. (2018). Non-coding RNAs: long non-coding RNAs and microRNAs in endocrine-related cancers. *Endocr. Relat. Cancer* 25, R259–R282. doi: 10.1530/ERC-17-0548

Koh, H. W. L., Fermin, D., Vogel, C., Choi, K. P., Ewing, R. M., and Choi, H. (2019). iomicspass: network-based integration of multiomics data for predictive subnetwork discovery. *NPJ Syst. Biol. Appl.* 5:22. doi: 10.1038/s41540-019-0099-y

Koudijs, K. K. M., Terwisscha van Scheltinga, A. G. T., Böhringer, S., Schimmel, K. J. M., and Guchelaar, H.-J. (2019). Transcriptome signature reversion as a method to reposition drugs against cancer for precision oncology. *Cancer J.* 25, 116–120. doi: 10.1097/PPO.0000000000000370

Kristensen, V. N., Lingjærde, O. C., Russnes, H. G., Vollan, H. K. M., Frigessi, A., and Børresen-Dale, A.-L. (2014). Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer* 14, 299–313. doi: 10.1038/nrc3721

Lê Cao, K.-A., Martin, P. G. P., Robert-Granié, C., and Besse, P. (2009). Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics* 10:34. doi: 10.1186/1471-2105-10-34

Li, E., and Zhang, Y. (2014). DNA methylation in mammals. *Cold Spring Harb. Perspect. Biol.* 6:a019133. doi: 10.1101/cshperspect.a019133

Lindqvist, B. M., Wingren, S., Motlagh, P. B., and Nilsson, T. K. (2014). Whole genome dna methylation signature of Her2-positive breast cancer. *Epigenetics* 9, 1149–1162. doi: 10.4161/epi.29632

Liu, Y. (2020). Clinical implications of chromatin accessibility in human cancers. *Oncotarget* 11, 1666–1678. doi: 10.18632/oncotarget.27584

Liu, Y., Liu, Y., Huang, R., Song, W., Wang, J., Xiao, Z., et al. (2019). Dependency of the cancer-specific transcriptional regulation circuitry on the promoter DNA methylome. *Cell Rep*. 26, 3461–3474.e5. doi: 10.1016/j.celrep.2019.02.084

Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., and Bergmann, S. (2016). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat. Methods* 13, 366–370. doi: 10.1038/nmeth.3799

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., et al. (2006). Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7(Suppl. 1):S7. doi: 10.1186/1471-2105-7-S1-S7

McCarthy, D. J., and Smyth, G. K. (2009). Testing significance relative to a fold-change threshold is a treat. *Bioinformatics* 25, 765–771. doi: 10.1093/bioinformatics/btp053

Mejía-Pedroza, R. A., Espinal-Enríquez, J., and Hernández-Lemus, E. (2018). Pathway-based drug repositioning for breast cancer molecular subtypes. *Front. Pharmacol*. 9:905. doi: 10.3389/fphar.2018.00905

Neph, S., Stergachis, A. B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J. A. (2012). Circuitry and dynamics of human transcription factor regulatory networks. *Cell* 150, 1274–1286. doi: 10.1016/j.cell.2012.04.040

Ochoa, S., de Anda-Jáuregui, G., and Hernández-Lemus, E. (2020). Multi-omic regulation of the pam50 gene signature in breast cancer molecular subtypes. *Front. Oncol*. 10:845. doi: 10.3389/fonc.2020.00845

O'Day, E., and Lal, A. (2010). Micrornas and their target gene networks in breast cancer. *Breast Cancer Res*. 12:201. doi: 10.1186/bcr2484

Prat, A., Pineda, E., Adamo, B., Galván, P., Fernández, A., Gaba, L., et al. (2015). Clinical implications of the intrinsic molecular subtypes of breast cancer. *Breast* 24, S26–S35. doi: 10.1016/j.breast.2015.07.008

Ru, Y., Kechris, K. J., Tabakoff, B., Hoffman, P., Radcliffe, R. A., Bowler, R., et al. (2014). The multimir R package and database: integration of microRNA-target interactions along with their disease and drug associations. *Nucl. Acids Res*. 42:e133. doi: 10.1093/nar/gku631

Rupaimoole, R., Wu, S. Y., Pradeep, S., Ivan, C., Pecot, C. V., Gharpure, K. M., et al. (2014). Hypoxia-mediated downregulation of miRNA biogenesis promotes tumour progression. *Nat. Commun*. 5:5202. doi: 10.1038/ncomms6202

Sergushichev, A. A. (2016). An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv*. 1–40. doi: 10.1101/060012

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 13, 2498–2504. doi: 10.1101/gr.1239303

Silverman, E. K., Schmidt, H. H. H. W., Anastasiadou, E., Altucci, L., Angelini, M., Badimon, L., et al. (2020). Molecular networks in network medicine: development and applications. *Wiley Interdisc. Rev*. 12:e1489. doi: 10.1002/wsbm.1489

Szyf, M. (2012). Dna methylation signatures for breast cancer classification and prognosis. *Genome Med*. 4:26. doi: 10.1186/gm325

Tam, S., Tsao, M.-S., and McPherson, J. D. (2015). Optimization of miRNA-seq data preprocessing. *Brief. Bioinformatics* 16, 950–963. doi: 10.1093/bib/bbv019

Tang, J., Kong, D., Cui, Q., Wang, K., Zhang, D., Gong, Y., et al. (2018). Prognostic genes of breast cancer identified by gene co-expression network analysis. *Front. Oncol*. 8:374. doi: 10.3389/fonc.2018.00374

Tarazona, S., Balzano-Nogueira, L., Gómez-Cabrero, D., Schmidt, A., Imhof, A., Hankemeier, T., et al. (2020). Harmonization of quality metrics and power calculation in multi-omic studies. *Nat. Commun*. 11:3092. doi: 10.1038/s41467-020-16937-8

Tarazona, S., Furió-Tarí, P., Turrà, D., Pietro, A. D., Nueda, M. J., Ferrer, A., et al. (2015). Data quality aware analysis of differential expression in RNA-seq with noiseq R/bioc package. *Nucl. Acids Res*. 43:e140. doi: 10.1093/nar/gkv711

Tian, T., Olson, S., Whitacre, J. M., and Harding, A. (2011). The origins of cancer robustness and evolvability. *Integr. Biol*. 3, 17–30. doi: 10.1039/C0IB00046A

Tomlinson, R. L., Abreu, E. B., Ziegler, T., Ly, H., Counter, C. M., Terns, R. M., et al. (2008). Telomerase reverse transcriptase is required for the localization of telomerase RNA to cajal bodies and telomeres in human cancer cells. *Mol. Biol. Cell* 19, 3793–3800. doi: 10.1091/mbc.e08-02-0184

Turashvili, G., and Brogi, E. (2017). Tumor heterogeneity in breast cancer. *Front. Med*. 4:227. doi: 10.3389/fmed.2017.00227

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., et al. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics* 26, i237–i245. doi: 10.1093/bioinformatics/btq182

Vrba, L., and Futscher, B. W. (2019). Dna methylation changes in biomarker loci occur early in cancer progression. *F1000Research* 8:2106. doi: 10.12688/f1000research.21584.1

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11, 333–337. doi: 10.1038/nmeth.2810

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 05 January 2023
DOI 10.3389/fgene.2022.1078609

OPEN ACCESS

EDITED BY
Dominik Heider,
University of Marburg, Germany

REVIEWED BY
Paolo Martini,
University of Brescia, Italy
Markus List,
Technical University of Munich, Germany

*CORRESPONDENCE
Enrique Hernández-Lemus,
ehernandez@inmegen.gob.mx

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 24 October 2022
ACCEPTED 15 December 2022
PUBLISHED 05 January 2023

CITATION
Ochoa S and Hernández-Lemus E (2023), Functional impact of multi-omic interactions in breast cancer subtypes. *Front. Genet.* 13:1078609. doi: 10.3389/fgene.2022.1078609

Functional impact of multi-omic interactions in breast cancer subtypes

Soledad Ochoa[1,2] and Enrique Hernández-Lemus[1,3]*

[1]Computational Genomics Division, National Institute of Genomic Medicine, Mexico City, Mexico, [2]Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México, Mexico City, Mexico, [3]Center for Complexity Sciences, Universidad Nacional Autónoma de México, Mexico City, Mexico

Multi-omic approaches are expected to deliver a broader molecular view of cancer. However, the promised mechanistic explanations have not quite settled yet. Here, we propose a theoretical and computational analysis framework to semi-automatically produce network models of the regulatory constraints influencing a biological function. This way, we identified functions significantly enriched on the analyzed omics and described associated features, for each of the four breast cancer molecular subtypes. For instance, we identified functions sustaining over-representation of invasion-related processes in the basal subtype and DNA modification processes in the normal tissue. We found limited overlap on the omics-associated functions between subtypes; however, a startling feature intersection within subtype functions also emerged. The examples presented highlight new, potentially regulatory features, with sound biological reasons to expect a connection with the functions. Multi-omic regulatory networks thus constitute reliable models of the way omics are connected, demonstrating a capability for systematic generation of mechanistic hypothesis.

KEYWORDS

multi-omics, breast cancer, network biology, HIF, RAS, WNT, SOX9, DNA methylation

1 Introduction

The establishment of high-throughput technologies has made possible a systems biology approach to cancer through multi-omics integration (Kristensen et al., 2014). The multi-omics perspective takes advantage of the complementarity between different molecular levels of description. However, the promise of attaining mechanistic explanations (Bersanelli et al., 2016) has not settled yet.

Although there is a plethora of statistical approximations (Huang et al., 2017), sparse multivariate methods are arguably nearer to the mechanistic explanation goal, given their capacity to pinpoint potential regulators (Li et al., 2012; Sohn et al., 2013; Bose et al., 2022). These approaches have even identified potential key regulators for each breast cancer subtype (Huang et al., 2019), and for the subgroups of the triple-negative breast cancer subtype (Chappell et al., 2021). The networks shown in some of these works (Li et al., 2012; Sohn et al., 2013; Huang et al., 2019) constitute hypothesized models of the way regulators are connected, demonstrating a capability for systematic production of testable regulatory mechanisms.

Here, we applied the sparse generalized canonical correlation analysis (SGCCA) to data on DNA methylation and gene and miRNA expression from TCGA. The SGCCA is a statistical method that outputs correlated features among a large collection by the use of LASSO penalization (Tenenhaus et al., 2014). The SGCCA has been successfully used for biomarker discovery from cancer (Fan et al., 2020) and non-cancer contexts (Garali et al., 2018). In order to find not just the features but the connections between them, SGCCA was

coupled with ARACNE (Margolin et al., 2006), a method for inference of transcriptional networks, that has allowed our group to find transcriptional master regulators (Tapia-Carrillo et al., 2019), to document a loss of long-distance co-expression (García-Cortés et al., 2020; Dorantes-Gilardi et al., 2021; García-Cortés et al., 2021), and to evaluate the role that relevant miRNAs play in some oncogenic processes (Drago-García et al., 2017; Zamora-Fuentes et al., 2022), among other applications in the large-scale molecular study of cancer. As an outcome, we describe some of the reconstructed networks and their implications, highlighting their relevance to understand cancer biology and potentially impact treatment. The general pipeline is described in Figure 1.

2 Methods

All the analyses described hereafter were performed with R programming language version 4.1.1 (R Core Team, 2021) and can be found at http://csbig.inmegen.gob.mx/SGCCA/. Release 105 of `biomaRt` was used all along and plots were produced with `ggplot2` (Wickham, 2016).

2.1 Data acquisition

TCGA data were obtained through the `TCGAbiolinks` R package. We only used samples with Illumina Human Methylation 450, RNA-seq, and miRNA-seq data from unique patients. This constraints the number of samples to 128 from the basal subtype, 46 from Her2-enriched, 416 from luminal A, 140 from luminal B, and 75 samples from normal adjacent tissue.

Pre-processing has been described before (Ochoa et al., 2021) and follows published guidelines (Aryee et al., 2014; Tam et al., 2015; Tarazona et al., 2015). As a first step, only protein-coding transcripts were kept since for our purposes, these were considered the main functional effectors. This restriction toward the study of non-coding features was chosen in order to focus on the expression regulatory layers of DNA methylation, miRNA expression and, hidden among the transcripts, the layer of transcription factors. Length and GC content biases were checked with the `NOISeq` package (Tarazona et al., 2015) and alleviated using `EDASeq` (Risso et al., 2011) full normalization. Genes with zero counts were (the only ones) discarded at the low count filter, TMM normalization was applied between samples, and the batch effect was corrected. Since batch effects can be induced by *a priori*-unknown factors, `ARSyNseq` was used to remove all systematic noise not associated with the subtypes (Nueda et al., 2012). Preprocessing of microRNAs is the same, except there is no length or GC bias and the normalization used between samples is the median method.

Finally, CpG probes with over 25% missing values and non-mapped or overlapping SNPs were discarded. The remaining missing values were imputed *via* nearest neighbors and transformed into M-value matrices. This way, datasets account for 393,132 methylation probes, 17,077 coding transcripts, and 604 miRNA precursors.

2.2 Sparse generalized canonical correlation analysis

Once pre-processing was performed, we normalized each omic by the square root of the first eigenvalue and concatenated them patient-wise, obtaining one matrix per breast cancer subtype and one for the normal tissue. Using this normalization ensures the influence of each omic over upcoming analysis depends on its variance (De Tayrac et al., 2009).

Afterward, we approached the SGCCA as implemented in the `mixOmics` package (Rohart et al., 2017) and largely followed the Garali et al. guidance (Garali et al., 2018). The analysis takes as input the different blocks of data and a sparsity parameter per block, the number of components to recover (`ncomp`), a design matrix, and a function to maximize the covariance. Sparsity parameters were chosen for each omic from the sequence [0.01, 0.02, ..., 0.09, 0.1, ..., 0.9], by cross-validation. With this purpose, a balanced dataset, composed of 10 samples per tumor subtype and 10 samples from normal tissue, was randomly taken from the original data, 10 times per each sparsity parameter value. Each time, a simple SGCCA was run, recovering only one component and taking note of the selected number of features and the average variance explained (AVE). Summing the different combinations, in total, every value was tested 11,340 times per omic. Sparsity parameters were chosen in order to obtain the largest AVE with the lowest number of features (Supplementary Figure S1), namely, 0.02 for CpG sites and transcripts and 0.05 for microRNAs.

Data analytics included several stages: independent pre-processing to deal with factors specific to the platforms, while normalization and penalization concern appropriate data integration. Eigenvalue normalization was further performed to equilibrate the still disparate rank of the different values. Separate penalization considers the different signal sizes the distinct omics may have. Shrinking the same CpG coefficients and miRNA coefficients may over-penalize relevant associations yet with effects smaller than those coming from other omics Liu et al. (2018). After the fitting process, we noticed that miRNAs are slightly less penalized than the other omics.

The definite SGCCA for each subtype and the normal tissue was run using the fitted values. The smaller the sparsity value, the fewer features get selected. For each subtype, we used the number of samples minus 1 as ncomp, the default design matrix, and the `centroid` function, which enables negative correlation.

Feature selection attained by SGCCA is expected to be a bit unstable due to the LASSO penalization. Mimicking the filter used in `miRDriver` (Bose and Bozdag, 2019), we re-run SGCCA 100 times per subtype, or the normal tissue, using a random subset of half the samples each time. We only kept those features selected at least 70% of the time.

2.3 Functional enrichment analysis

SGCCA results include a matrix of the loadings a feature has in each component. The said matrix is quite sparse, except for the features summarizing the relevant information between and within omics. These non-zero loadings indicate co-selected features that can be tested for functional enrichment.

With the idea of exploiting the full set of co-selected features, and not just the transcripts, all the features, being CpG probes, miRNA precursors, or transcripts, were mapped to Entrez gene IDs. Both transcripts and miRNAs have a direct annotation at Entrez, (e.g., hsa-mir-34b becomes MIR34B). To translate CpG probes to Entrez IDs, we recovered the genes affected by each probe from the microarray annotation file. This results in an amplification of CpG representation

since one site can be associated with a whole cluster of genes and assumes a methylation effect on overlapping genes, which is not necessarily true. Both are cons of this mapping that need to be considered.

Then, the group of features with non-zero loading in every SGCCA component was submitted to a separate over-representation analysis, taking Entrez IDs as input. Enrichment was run using the `clusterProfiler` package (Wu et al., 2021) against the pathways from the KEGG database (Kanehisa and Goto, 2000) and against the biological process gene ontology (Consortium, 2021). A significance threshold of FDR-corrected p-values < 0.01 was set. The intersection between sets of enriched functions was plotted with the `UpSetR` package (Gehlenborg, 2019). Functions exclusively enriched in one dataset were tested for over-representation. With this purpose, exclusively enriched functions were grouped according to GOslim and KEGG classes. Dependence between grouped categories and the subtypes was assessed with Fisher's test, and p-values were adjusted for multiple testing using the Bonferroni method.

In an independent manner, we ran a gene set enrichment analysis (GSEA), only with transcript data, to check for functions affected by differential expression. GSEA was also performed with the `clusterProfiler` package, in this case, without a p-value cutoff. The idea is to recover a GSEA enrichment score for every one of the functions over-represented in the SGCCA results. Such scores would answer if functions over-represented among the features related through different omics are also enriched among genes with altered expression. We must stress, however, that all discussed functions are significantly over-represented (p-value < 0.01), but only the specified ones also have a significant GSEA score.

2.4 Network reconstruction

Chosen functions were represented as networks to draw potentially regulatory models. To achieve this, we estimated mutual information (MI) between every pair of nodes using ARACNE software (Margolin et al., 2006) and then filtered out all the pairs with lower MI than the median value observed for known regulatory interactions. Thus, for each chosen function, we recovered all the features co-selected (co-varying) with the features responsible for the functional enrichment and focused on this set.

1. We extracted a sub-matrix from the original dataset and run ARACNE.
2. We retrieved regulatory interactions involving the selected features. Again, this was performed with the microarray annotation file for the CpGs, assuming position overlap is enough to affect gene expression. The `multiMiR` package (Ru et al., 2014) was used in the case of miRNAs and `TFtargets` (github.com/slowkow) for the transcript coding for transcription factors. This latter package queries several resources, namely, TRED, ITFP, ENCODE, TRRUST, and the databases from Neph et al., 2012; Marbach et al., 2016 (Jiang et al., 2007; Zheng et al., 2008; Consortium et al., 2012; Neph et al., 2012; Han et al., 2015; Marbach et al., 2016), which include validated and predicted interactions. We considered those hits coming from ChIP-seq, DNaseI footprinting, and small-scale experiments as validated.
3. We obtained MI values for such regulatory interactions, using the `infotheo` package (Meyer, 2014) (the use of this specific tool obeys the need to focus on a reduced set of given pairs, instead of estimating all the pairs with a feature of interest in the adjacency matrix, as ARACNE would perform).
4. We took the median MI value for the regulatory interactions as the threshold. Since MI is expected to differ between the distinct kinds of pairs, different thresholds were obtained for the different types of edges: CpG–transcript, CpG–miRNAs, TF transcript–transcript, and miRNA–transcript. The median was preferred over the mean to avoid outliers dominating the threshold.
5. The MI value distribution obtained with ARACNE was contrasted between types of edges, *via* Kolmorogov–Smirnov tests. If distributions were not significantly different, the lowest median MI from regulatory interactions—obtained with infotheo—was chosen as the unique threshold to pass, no matter the edge-type, relaxing the threshold and increasing the MI interactions accepted in the final network.

The output of these items is a table with predicted interactions and weights that illustrate the largest statistical dependencies between the features selected by the SGCCA.

2.5 Network analysis

Mutual information networks were analyzed with the `igraph` package (Csardi and Nepusz, 2006) and represented with `Cytoscape` (Shannon et al., 2003), making use of the RCy3 package (Gustavsen et al., 2019).

Node colors represent logFC values between every subtype and the normal tissue. MiRNA differential expression went through *voom* normalization and *eBayes* `limma` function. Since the batch effect was not corrected in methylation data, we used the `missMethyl` package for the differential analysis. This tool removes systematic errors of unknown origin, bypassing the lack of batch-effect correction (Maksimovic et al., 2015).

The node degree was calculated for the whole network; however, only those network components with features annotated as players of a function are shown in the corresponding figures. Since Her2+ and luminal B subtypes produce large networks, we further zoomed in the graph by selecting only the first neighbors of functional features. Such subnetworks may serve as a model of the regulatory pressures influencing the function.

Every neighbor of a functional node was searched in PubMed, together with the associated functions, to find out if some biological role has already been suggested. PubMed was also queried with every pair of interacting nodes, as well as the databases for predicted regulatory links accessible through `multiMir`. Transcription factor-related features are identified according to the list from humantfs (Lambert et al., 2018). This achieves a fairly automated way to build a regulatory model for the functions enriched in the SGCCA.

3 Results and discussion

By applying SGCCA, we have identified, for each one of the breast cancer subtypes, transcripts whose expression patterns better reflect the variance in its own block, while also co-varying with the other

blocks of data. The pattern of selected features by omics and subtype is provided in Supplementary Figure S2.

SGCCA uses a LASSO penalization, which may select inconsistent sets of features. Since this could affect the reliability of functional enrichment, identifying functions dependent on unstable features, we just proceeded with the features most consistently selected, whose proportion is shown in Supplementary Figure S3. There are no individual transcripts or miRNAs selected simultaneously across all five datasets, but there are six CpG sites in this situation which potentially affect MAPK8IP3, AFAP1, LFNG, and VSTM2B.

The transcripts repeatedly selected in the same subtype have known associations with breast cancer. The top three transcripts selected more often for the basal subtype are MCL1, CTNNA1, and NOTCH3. MCL1 is an anti-apoptotic member of the BCL2 family that is required for mammary stem cell function (Fu et al., 2015), and it is expected to be overexpressed in tumors of this subtype (Farrugia et al., 2015). Meanwhile, catenin alpha 1 is postulated to act as a tumor suppressor in E-cadherin-negative basal-like breast cancer cells (Piao et al., 2014), and NOTCH3 seems to function as a promoter of the epithelial–mesenchymal transition (Liang et al., 2018).

Her2 enriched has also been clearly associated with its most selected transcripts: CEACAM5, ACACA, and PGK1. Though heterogeneously expressed, Her2-enriched tumors tend to be positive for CEACAM5 (Bechmann et al., 2020) and so this adhesion molecule has been suggested as a target for T-cell bi-specific antibodies (Messaoudene et al., 2019). Inhibitors of acetyl-CoA carboxylase, ACACA, work over MCF-7 cells overexpressing Her2 by interfering with cancer stem cell lipid biosynthesis and the Warburg effect (Corominas-Faja et al., 2014). At last, PGK1 protein has been found overexpressed in these tumors (Schulz et al., 2009), while being linked to macrophages and stratifying patients at higher risk (Li et al., 2021).

Interestingly, microRNAs from the let-7 family were among the top selected for basal, Her2+, and luminal B subtypes, as well as for normal breast tissue. These miRNAs regulate JAK-STAT3 and Myc signaling pathways, thus affecting stemness and metastasis (Thammaiah and Jayaram, 2016).

3.1 Functions enriched on SGCCA output differ between datasets

After inspecting the overall output of SGCCA, we wanted to know if there are functions involving the co-varying features. Enrichment against GO biological processes and KEGG pathways allows us to identify functions affected by the specific regulatory mechanisms identified.

A total of 683 GO biological processes and 69 KEGG pathways were found significantly over-represented (FDR adjusted p. value < 0.01) among the SGCCA co-selected features. Figure 2 shows the intersections between subtypes. Few functions were found enriched across all subtypes, and most of them are either exclusive or shared only by a pair of subtypes. That is, functions associated with DNA methylation and miRNA expression are not the same for all subtypes.

There are three biological processes significantly enriched (FDR-corrected p-value ≤.0099, for the specific values, see Supplementary Table S1) in the four subtypes and the normal tissue. These are the developmental processes: metanephric nephron development (GO: 0072210), metanephros development (GO:0001656), and pattern specification process (GO:0007389). Since GO:0072210 is a part of GO:0001656, they may be considered the same.

Then, we wondered if functions linked with DNA methylation and miRNA expression in cancer and normal tissue maintain an intact circuitry connecting CpGs, transcripts, and miRNAs. In more general terms, does a function enriched twice involve identical features and interactions?

3.2 Features responsible for the same functional enrichment differ across subtypes

The first step toward a shared circuitry connecting CpGs, transcripts, and miRNAs in different phenotypes would be to have the same (or similar) features behind the functional enrichment. To verify if this happens, we calculated the Jaccard index for every pair of functions enriched more than once. The Jaccard index divides the size of intersection between two sets by their union, measuring similarity with a normalized value between 0 (fully disjoint sets) and 1 (the same set). Distributions for the Jaccard index are shown in Figure 3A.

The obtained distributions are enough to state that, for most functions, the CpG–transcript–miRNA circuitry is not the same across datasets since the features involved are not the same. Only seven biological processes enriched in a given pair of SGCCA results share more than 50% of the involved features. Five of them are related to development, while the other two are related to cell adhesion. These are the functions that may share the interactions between CpG sites, transcripts, and miRNAs.

If this index hints at the similarity between subtypes pertaining to CpG–transcript–miRNA co-variation, the distance with Her2-enriched subtype results are intriguing. This may be caused by a bias induced by the low number of samples. Or perhaps this is associated with the lower correlation with DNA methylation patterns (Network et al., 2012). Not surprisingly, the pair with the most similarly enriched functions corresponds to the two luminal subtypes.

3.3 Exclusive category over-representation

To answer if functions exclusively found in one dataset bring to light subtype-specific properties, we analyzed over-representation of GOslim categories and KEGG classes. The proportion of biological processes found for each dataset in every one of the categories is given in Figure 3B, while the equivalent plot for KEGG pathways is found in Supplementary Figure S4.

None of the KEGG classes is biased toward a given subtype, but there is an enrichment for the categories: *cellular component organization* in the basal SGCCA components, *establishment of localization* in luminal A, and *DNA metabolic process* in the normal tissue. There are seven biological processes behind the *cellular component organization* over-representation, comprising five processes related to axon extension, which are clustered with regulation of the extent of cell growth. Collagen fibril organization is not in the cluster and is the seventh process, suggesting a potential bond between the basal subtype and invasiveness.

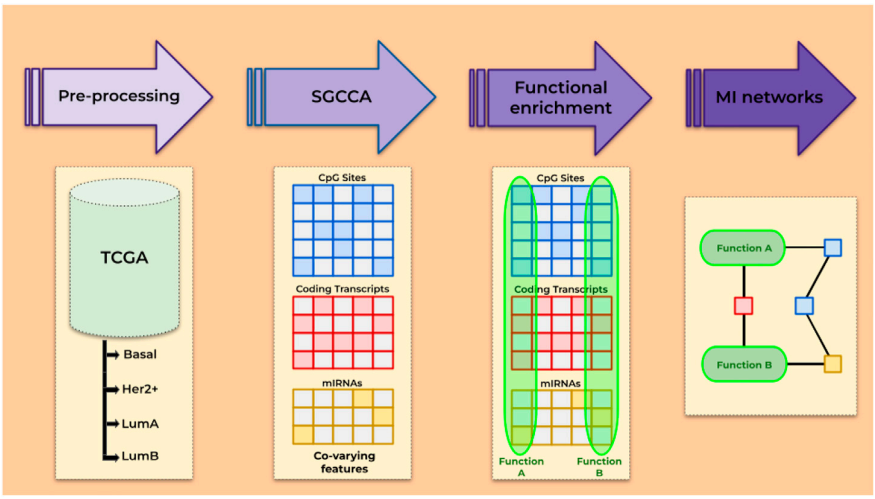


FIGURE 1
Overview of the steps followed.

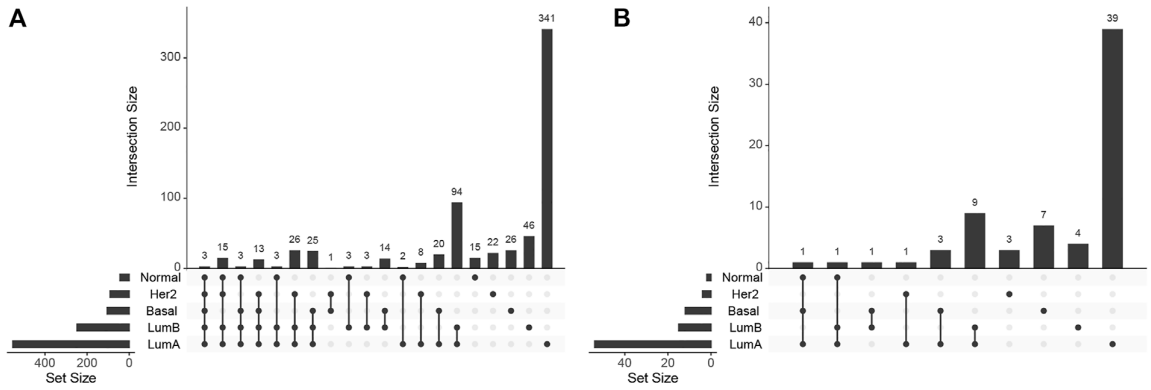


FIGURE 2
UpSet plot for **(A)** biological processes and **(B)** KEGG pathways enrichment.

In the case of luminal A, there are 62 biological processes behind the over-representation of *establishment of localization*. These processes affect transport and secretion and conform to 32 different clusters. Regarding over-representation in the normal tissue, it is interesting that it is related to DNA alkylation and methylation processes, perhaps implying that these processes are somehow disarranged on the tumor subtypes.

3.4 Within subtypes, different functions can be connected through correlated features

When checking the features responsible for the enrichment of a given function, we discovered that several functions are enriched in the exact same set of co-varying features, that is, the same set of SGCCA components. This suggests some level of crosstalk between functions that can be connected through correlated features. This observation has been made subtype-wise and implies that a single network of correlated features may actually span several functions.

Going through each subtype separately, we clustered functions by the proportion of SGCCA components shared. Figure 4 shows Her2 clusters. There are 11 clusters and six functions that cannot be grouped since they involve features that are not related with the clusters. Taking the bigger labels as a guide, purple, orange, and fuchsia clusters are related with development of kidney structures. Green and blue clusters at the bottom are linked with connective tissue development. Pale pink nodes refer to distinct processes of morphogenesis, while the nodes in yellow allude development of reproductive structures. The small brown and pale green clusters are related to cardiac muscle and neural cells, respectively. Finally, the small clusters in the center, in bright green and pale orange, are linked with metabolism and loaded with functions exclusively found in this subtype, a fact that may be

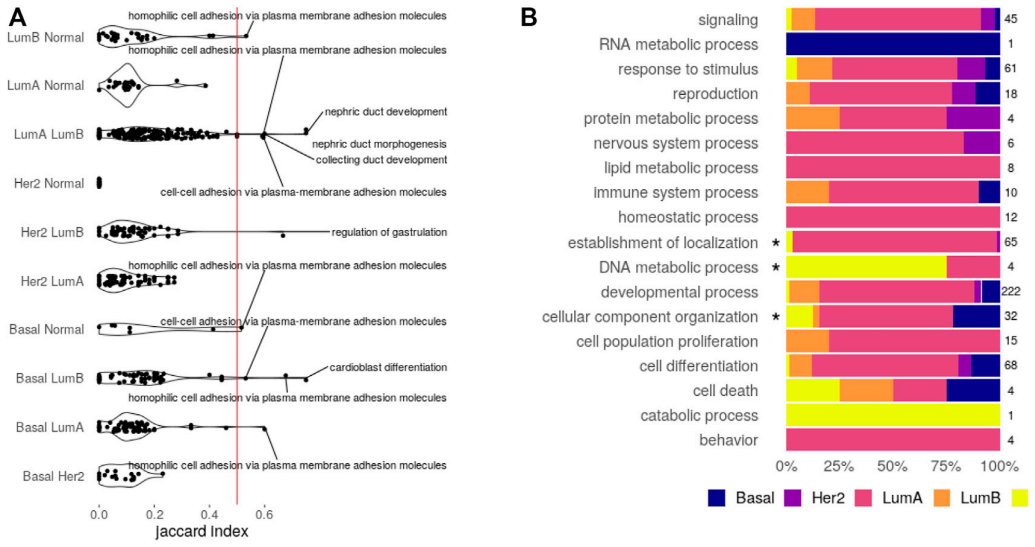


FIGURE 3
Enriched functions. **(A)** Feature similarity between functions shared by the pair of datasets indicated. Functions with similarity over 0.5 are displayed. **(B)** Bias of exclusive functions. An asterisk marks categories with significant over-representation (Fisher's test, Bonferroni adjusted p-value < 0.05).

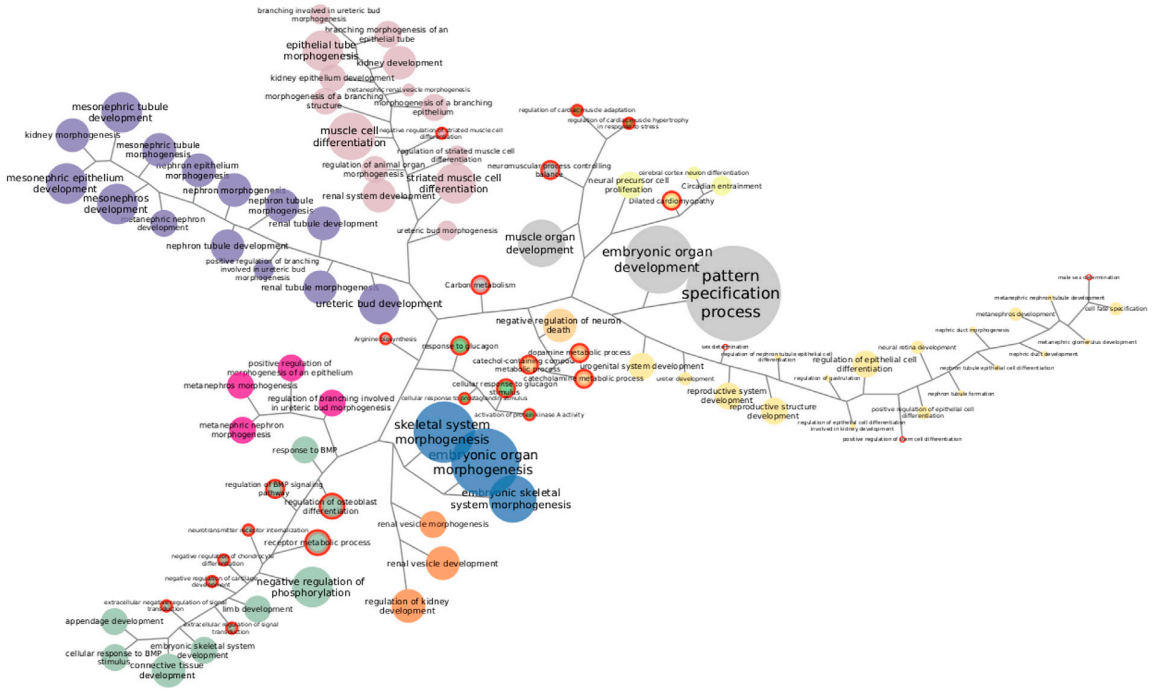


FIGURE 4
Functions enriched in Her2 SGCCA components. Both KEGG pathways and GO biological processes are represented together. Same-color clustered nodes are enriched in the same components. Nodes in gray do not belong to any cluster. The size of nodes and labels reflect the number of features behind the enrichment. Functions exclusively found in this subtype are highlighted with a red border.

interesting to explore further. The functions enriched with the most genes do not form a part of any cluster.

Clustering exposes information that needs to be accounted when discussing one particular enrichment. Functions exclusively found in one subtype may reveal mechanistic explanations of subtype-specific alterations, but, if exclusive functions are clustered with others that are non-exclusive and better represented, relevance may be debatable. Similarly, clusters may help explain some odd enrichments, like the one found in the luminal A dataset for morphine addiction. Morphine addiction has been found enriched on methylation-driven genes (Xu et al., 2019) but depends on features correlated with those responsible for ECM–receptor interaction, suggesting co-variation may be pulling up the enrichment for this addiction. Even after considering clusters, there are enrichments hard to figure out fully; however, some specific features can be actually tracked (Supplementary Tables S1, S2).

In order to select functions to explore further, we repeated the analysis described with Her2+ for each SGCCA result. While not all clustering are displayed here, full groups and enrichment results are supplied as Supplementary Files. A filtering step was necessary because, even with the clustering, there are almost 500 sets of functionally related features. It is interesting that the two cell adhesion processes with the Jaccard index over 0.5 appear consistently out of any cluster in the subtypes with such enrichment.

3.5 Network examples

In our path to answer if a function enriched twice involves identical features and interactions, we found that a given function is commonly enriched through distinct sets of features in two different datasets. At the same time, we observed several functions over-represented among the same sets of co-selected features and wondered how functions were connected. Functions involving the same features are already identifiable in the annotation databases, but by means of this multi-omic integration strategy, we have been able to find cross-linking paths across single layers and maybe even connect seemingly independent functions through multi-omics pattern co-variation. To check how this appears, we built mutual information (MI) networks. The networks went through a stringent threshold to keep just the interactions that are most likely regulatory. To this end, we obtained the MI values accompanying true regulatory interactions and took the median value as the minimum MI required to consider an edge as possibly regulatory. Within these reduced sets of interactions, the following figures show the network components that contain those features annotated as participating in the functions, though some of the obtained networks extend further.

The intuition is that co-selected features, whose patterns are correlated with those of functional features, may also be participating in a given function. Beyond that, nodes for miRNAs, CpGs, and transcripts that ultimately code for transcription factors may be playing regulatory roles. The stringent threshold attempts to filter out the interactions owed to simple co-variation. Two broad possible scenarios are expected, 1) disconnected components per function, each with its own potential regulators, or 2) functions that crosstalk through common features, whose potential regulators could be of medical interest. The different scenarios are exemplified through the four subtypes and the normal tissue in the coming sections.

3.5.1 HIF-1 signaling in the basal subtype

Hypoxia-inducible factor 1 (HIF-1) signaling is one of the KEGG pathways enriched exclusively in the basal SGCCA results. HIF-1 is the master regulator of oxygen homeostasis since it induces transcription from at least 100 hypoxia-responsive elements (Corrado and Fontana, 2020). HIF-1 signaling is activated in tumors not only under hypoxic conditions but also by oxygen-independent factors, like TP53 and BRCA mutations (de Heer et al., 2020), which have been associated with the basal subtype (Network et al., 2012).

The network we identified for this function is given in Figure 5. AMPK signaling is enriched in a subset of the same SGCCA components such as HIF-1 signaling, which is consistent with the idea that these two pathways interplay in cancer metabolism re-programming (Moldogazieva et al., 2020). However, after applying the MI threshold, each pathway occupies disconnected components.

Only two functional features, that is, annotated as participants of the function, pass the MI threshold, NOS3 and RPS6. It is important to clarify that the enrichment does not rest only on these two features, but we only find interactions over the threshold for them. There are also two nodes that have been linked with the signaling pathway without being participants as such. PEG3 gets upregulated after hypoxia in mouse lungs (Wollen et al., 2013), while SUMO3 would be one of the modifiers affecting HIF-1 stability (Matic et al., 2008). Thus, nodes seem to be associated with the function.

On the other hand, the complete network is formed by CpG–transcript interactions, more specifically, by edges linking a CpG with a transcript coding for a ribosomal protein. Since CpG sites are not in the same chromosome as the transcript, a direct regulatory influence can be discarded. To account for indirect relations, we estimated the mutual information between the corresponding transcripts, even when these were not originally in the SGCCA set of co-selected features. Obtained MI values are smaller than the global threshold and smaller than the edges between CpGs and ribosomal protein-coding transcripts. Hence, indirect effects going through the transcript linked with the CpG do not seem to fully explain the phenomenon.

Most nodes are not significantly different from the normal tissue, either regarding expression or methylation values. This is consistent with the lack of significance of the pathway GSEA score (NES = 0.9252, adjusted p-value: 0.7937). HIF-1 signaling in the basal subtype is transcriptionally comparable with that of the normal tissue. Nevertheless, the pathway is not found enriched in the normal tissue SGCCA output, suggesting a change in the correlation between omics.

3.5.2 Positive regulation of stem cell differentiation in the Her2-enriched subtype

Cancer stem cells are largely responsible for relapse and metastasis. Her2 variants, observed in Her2+ patients with poor clinical outcomes, have been reported to drive maintenance and enrichment of breast cancer stem cells (Pupa et al., 2021). *Positive regulation of stem cell differentiation* was found enriched exclusively in Her2+ data, but related processes also appear in the other three subtypes. The process is clustered with several other functions, as shown in Figure 6, where we have focused on the first neighbors of the functional features.The transcription factor SOX9 is the only feature

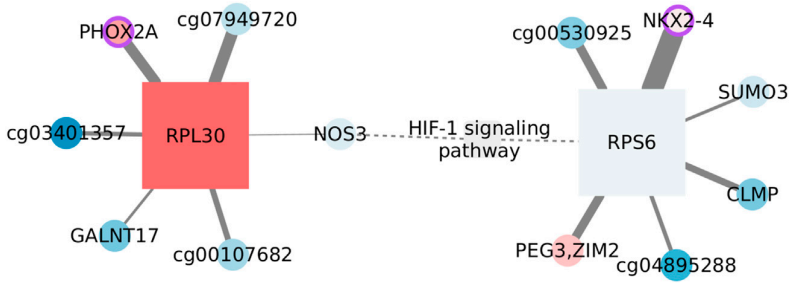


FIGURE 5
Features connected with HIF-1 signaling in the basal subtype. Circles represent CpGs, and squares are transcripts. When possible, CpGs are identified with the symbol of the gene they affect; otherwise, the ID of the probe is used. The shades of red indicate the level of overexpression/methylation against the normal tissue, while blue tones represent values under what is expected. The node size reflects its degree. A purple border identifies nodes whose protein plays a transcription factor role. The weight of the link is the extent of mutual information between connected nodes. Dashed edges link MI components with prior information.

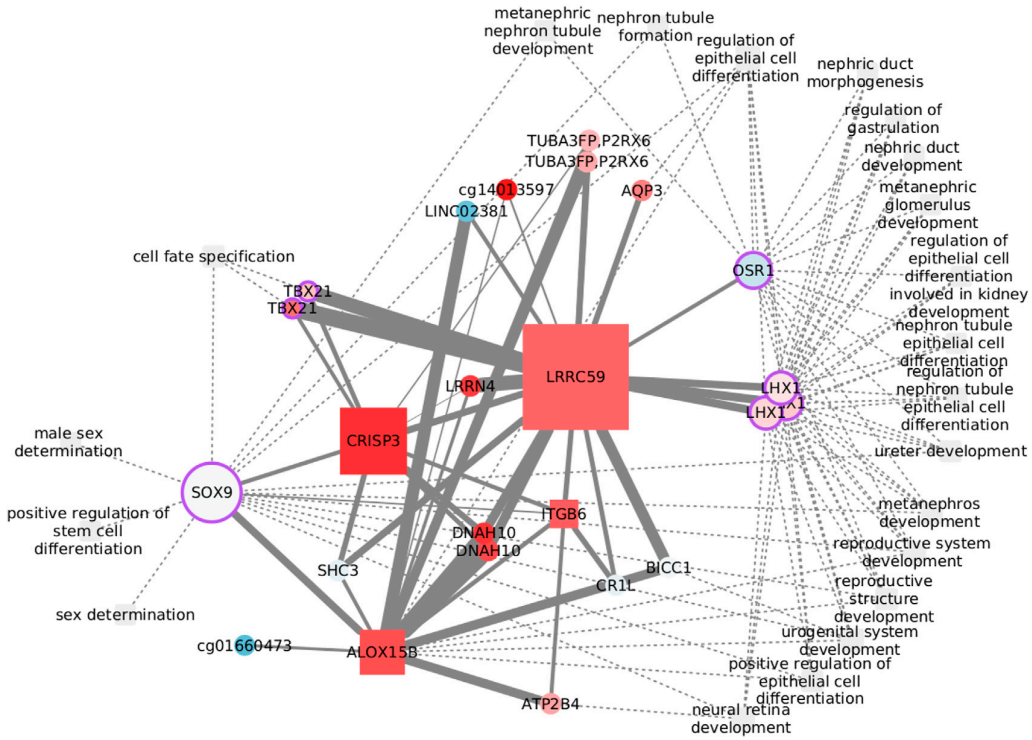


FIGURE 6
Features connected with the regulation of stem cell differentiation in the Her2-enriched subtype. Node size reflects the betweenness centrality.

from *positive regulation of stem cell differentiation* with edges passing the MI threshold. SOX9 binds functions related with cell fate and sex determination, while LHX1 and OSR1 are at the crossroads of most functions. None of the edges has been previously reported, but several nodes have known links with these functions. The relation between CRISP3 and sex determination, for instance, may be explained by the role of the protein in sperm function (Weigel Muñoz et al., 2019) and its up regulation in prostate cancer (Pathak et al., 2016). DNAH10 is another feature with a known bond with sex determination, specifically with sperm flagella morphological abnormalities (Li et al., 2022). The connection with ITGB6 is perhaps weaker since it rests only on differential expression analysis of prostate cancer (Li et al., 2013). CR1L is involved in B lymphocyte activation (Fernández-Centeno et al., 2000) and may have a role in renal injury (He et al., 2005). Finally, the somehow unexpected *neural retina development* is related with the function of SHC3 (Nakazawa et al., 2002).

The functional implications of some of these nodes are specifically dependent on DNA methylation. Although epigenetically altered

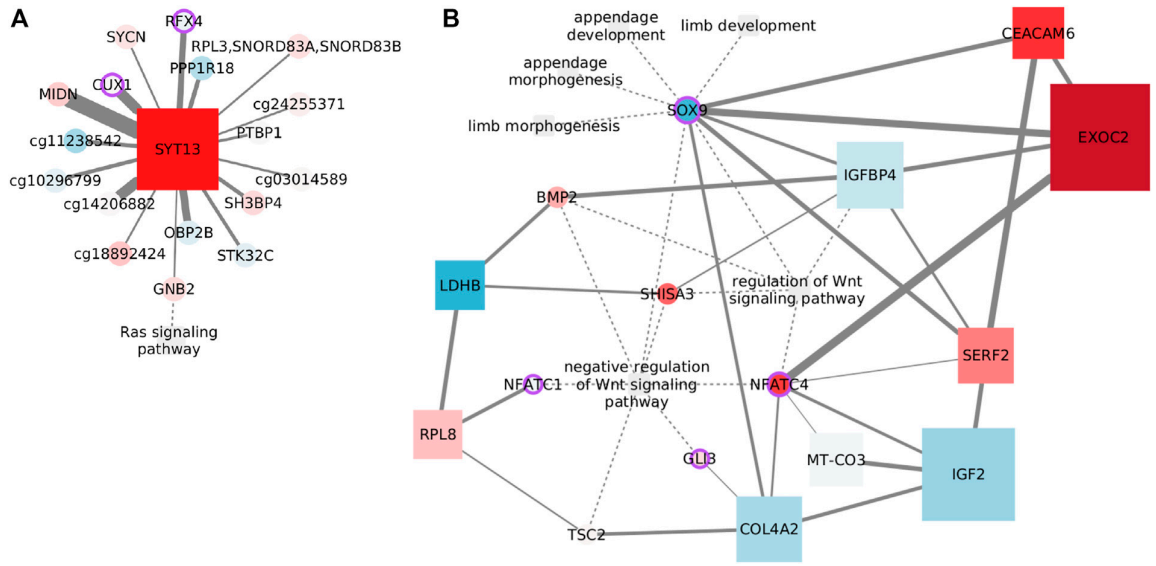


FIGURE 7
Example networks for the luminal subtypes. **(A)** Features connected with the Ras signaling pathway in luminal A data. Node size indicates degree. **(B)** Features connected with Wnt signaling in the luminal B subtype. Node size reflects betweenness centrality.

CR1L is linked with Alzheimer's and dementia (Bahado-Singh et al., 2021), DNAH10 has emerged when studying renal carcinomas with a CpG-island methylator phenotype (Arai et al., 2015). Finally, CpG methylation of the lncRNA LINC02381 functions as a tumor suppressor in colorectal cancer (Jafarzadeh et al., 2020). While all of these features are represented by CpG sites in the network, LINC02381 appearance highlights the complexity of transcription regulation and the need to widen multi-omic analysis to include more data layers.

Despite that transcription factors may be the obvious option to explore the crosstalk between biological processes, less explored options, like ALOX15B, CRISP3, and LRRC59, with elevated graph betweenness, may result of interest.

3.5.3 Ras signaling pathway in the luminal A subtype

Ras signaling is one of the many pathways exclusively found enriched in the luminal A subtype. It is a well-documented pathway influencing cancer aspects like cell proliferation, survival, migration, and differentiation. Although the pathway is more frequently activated in the other subtypes, it has been reported as an indicator of poor prognosis in luminal tumors (Wright et al., 2015). Not surprisingly, Ras signaling components are under-expressed relative to the normal tissue (NES: 1.5796, adjusted *p*-value: 0.0084) in this analysis.

Only one functional feature endures the MI threshold, GNB2. The subunit beta 2 of G protein links the signaling pathway with a set of CpGs associated with cell communication and brain function, through the calcium sensor SYT13. Genes affected by the CpGs include the brain active kinase, STK32C; MIDN, that is predicted to enable kinase binding; OBP2B, which is supposed to enable binding of small volatile molecules; a TF from early brain development, RFX4; and SYCN, which is predicted to be active in exocytosis.

Among the remaining nodes, the connection with the CUX1 CpG site agrees with the cooperation observed between this transcription factor and Kras-G12V mutant in lung cancer (Ramdzan et al., 2014). In a similar way, PTBP1 overexpression is known to co-occur with oncogenic KRAS mutations in colon cancer (Hollander et al., 2016). Finally, a connection with the transferrin receptor internalization protein, SH3BP4, has been predicted before by a random forest classifier (Xin et al., 2021).

Again, the network shown in Figure 7A links a transcript with CpG sites all over the genome. Although it has been proposed that Ras signaling controls aberrant DNA methylation (Patra, 2008), the specific influence nodes may have over the signaling pathway remains unclear.

3.5.4 Negative regulation of the Wnt signaling pathway in the luminal B subtype

Wnt signaling normally controls organ development. In breast cancer, Wnt signaling is involved in tumor proliferation and metastasis, immune microenvironment regulation, stemness maintenance, and therapeutic resistance (Xu et al., 2020). The relevance of this function does not end here, but it has also been associated explicitly with the luminal B subtype. Though generalized DNA hypomethylation is common in cancer (Vidal Ocabo et al., 2017), a fraction of luminal B tumors exhibit hypermethylation, specifically affecting Wnt signaling (Network et al., 2012).

In our results, *negative regulation of* the *Wnt signaling pathway* is exclusively found enriched in this subtype, but related Wnt pathways were also found for luminal A. The cross-talking functions shown in Figure 7B are not in the same cluster but are found in a subset of the SGCCA components, where negative regulation of Wnt appears. Since these related functions makeup the largest network—after the threshold—we have, and this network consists of a large single

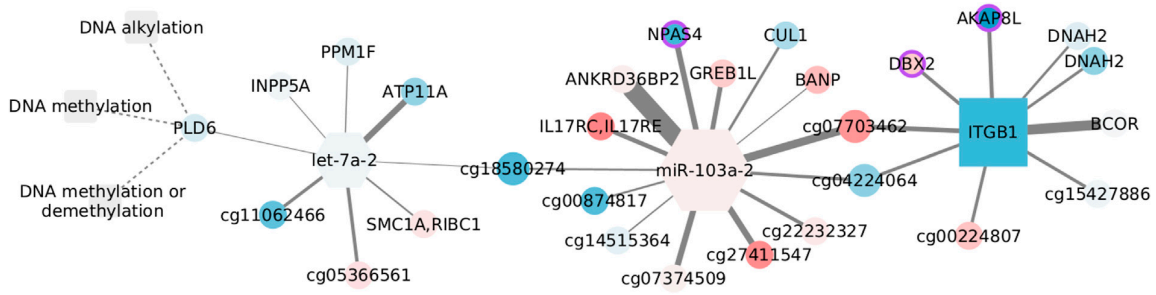


FIGURE 8
Features connected with DNA methylation in the normal adjacent tissue. Color corresponds to log fold changes in the basal subtype.

component, we decided to focus on the first neighbors of the functional features.

As expected, *negative regulation of Wnt signaling* and *regulation of Wnt* share functional features. The transcription factor for skeletal development, Sox9, is represented by its CpG at the crossroad between Wnt signaling with the developmental processes, but there are also multiple indirect paths. Since the genes coding collagen subunit Col4a2 and cell adhesion molecule Ceacam6 are targeted by Sox9 (Sumi et al., 2007), and Sox9 acts in cooperation with Gli3 (Tan et al., 2018), that pair of edges are easy to justify. Similarly, the link between COL4A2 and NFATC4 could be explained by the inhibition of the nuclear translocation of NFATc4 by Col4a2 in cardiomyocytes (Sugiyama et al., 2020), while both COL4A2 and IGF2 code for extracellular proteins deregulated under diseases with EMT (Bueno et al., 2011). Additionally, bone marrow stromal cells induced with IGFBP4, among other factors, overexpress SOX9 (Liu et al., 2012). Insulin-like growth factor-binding protein 4 is also connected with BMP2, as IGFBP4 overexpression impairs BMP2-induced osteogenic differentiation (Wu et al., 2017).

In summary, there are sound biological reasons to expect co-variation of the connected features. The question to solve is how such connections affect Wnt signaling and luminal B cancer progression, specifically what is the role of the node with the highest betweenness. Exocyst complex component 2 is related with the Wnt pathway as an effector of Hedgehog signaling (Arraf et al., 2020) and has been associated with metastasis and different cancer types (Cerhan et al., 2014; Hazelett and Yeaman, 2012; D'Aloia et al., 2018), but not with breast cancer.

3.5.5 DNA methylation in the normal adjacent tissue

DNA methylation is exclusively enriched in the normal tissue, but we choose to discuss it because of its relevance for cancer (Baylin and Jones, 2016). In addition, unlike the other examples, this network does contain microRNAs, including the top selected let-7a-2.

For consistency, we colored the nodes in Figure 8. However, since the normal tissue is our reference value, we used the log fold changes obtained by contrasting basal and normal tissue expression. This subtype has significant overexpression of related genes (NES = 1.9251, adjusted p-values = 0.0031) and has been linked with hypomethylation (Network et al., 2012). Yet, we have to emphasize that DNA methylation is not enriched in the basal data, and so, the relation between CpGs, miRNAs, and transcripts may not follow what is suggested in this figure.

Despite none of the interactions has been reported, a couple of nodes are somehow connected with the DNA methylation machinery. AKAP8L interacts with core subunits of the H3K4 histone methyltransferase complexes (Bieluszewska et al., 2018), whose action is interrelated with DNA modification (Rose and Klose, 2014). BCOR is part of the non-canonical polycomb repressive complex 1 and is altered in distinct cancer types (Astolfi et al., 2019). It has been observed that BANP can open the chromatin at unmethylated CpG-island promoters, thus activating essential genes in pluripotent stem and differentiated neuronal cells (Grand et al., 2021). Finally, *de novo* DNA methyltransferase, DNMT3b, can interact with CUL1, involving this node in aberrant methylation (Shamay et al., 2010).

In contrast, another set of nodes hinges on epigenetic silencing, as is the case of INPP5A in lung adenocarcinoma (Ke et al., 2020). Together with ATP11A, INPP5A CpG methylation has shown discriminatory capacity for colorectal cancer (Izquierdo et al., 2021). In the same manner, ATP11A methylation distinguishes several diseases including metastatic-lethal prostate cancer (Zhao et al., 2017), while a methylation signature including the growth regulation by estrogen in breast cancer 1 like GREB1L separates gastric adenocarcinoma cases by overall survival, and DBX2 methylation marks the serum from hepatocellular cancer patients (Zhang et al., 2013). Similar to its paralog DNAH10, DNAH2 aberrations are frequent in renal carcinomas with a CpG-island methylator phenotype (Arai et al., 2015). Although unexpected, the brain-specific transcription factor NPAS4, present in the form of a CpG site, is known to be regulated by DNA methylation (Furukawa-Hibi et al., 2015) and has been linked with colon adenocarcinoma survival (Luo et al., 2021). Last, though ITGB1 methylation is expected to be constant both in cancer and normal tissue (Strelnikov et al., 2021), alteration of the gene expression has been observed in basal-like tumors and cells with BRCA mutation, highlighting the relevance of migration and mesenchymal properties for this subtype (Privat et al., 2018).

Interestingly, the two miRNAs in the network are associated with migration and invasion, although in opposite ways. The let-7 family works as a tumor suppressor and is inhibited by DNA methylation and several regulators (Thammaiah and Jayaram, 2016). Contrastingly, miR-103 acts as an oncogene in triple-negative tumors, and its over-expression is linked with poor prognosis (Xiong et al., 2017). In spite of the low fold changes, the expression of both miRNAs is coherent with what would be expected in the basal subtype.

4 Conclusion

Here, we have described the kind of multi-omic network models that can be obtained through the sequential application of SGCCA and ARACNE. The collection of interactions shown in any of these networks suggests a multi-omic model that may or may not have regulatory implications. To asseverate regulation, wet laboratory testing would be needed. However, the nature of nodes as CpG sites, microRNAs, or transcript coding for functional proteins must be considered, as shown in the examples. Although further testing is required, the examples embody the level of details we can get in the way toward targeted experimental validation of multi-omic regulatory phenomena.

Though the interactions encountered seem to be subtype-specific, given the low values of the Jaccard index, there is no restriction to believe these same associations could not be repeated in other contexts, with somehow equivalent patterns of methylation and expression. Instead, an interesting question arises about the traceability of tissue and disease signals. A fair attempt to carry out would be to compare cancer and tissue networks with the same nodes, even if the edge weights are disparate, which were not produced here. Also, it has to be noticed that the normal adjacent tissue may not be the best control since it carries detected alterations across tissues (Aran et al., 2017).

The use of SGCCA allowed us to identify the functions enriched in features co-varying across DNA methylation, transcript, and miRNA expression. This does not mean such functions may not be influenced by other regulatory mechanisms: this simply indicates the functions, like HIF signaling in the basal subtype, depending the furthest on features whose methylation and expression co-vary. The con of the method is the instability of the LASSO, which forced us to keep just the features identified in over 70% of subsamples. Even when other tools (Hernández-de Diego et al., 2018; Meng et al., 2019) could achieve the multi-omic functional enrichment without the instability issue, we prefer the sparse method exactly because of the stable portion of the feature set. Then, possible improvements include the elastic network penalization, which overcomes the stability problem.

mixOmics output for the SGCCA includes a complete graph connecting all the features selected in a component. However, having found the same functions over-represented in different components, we wanted to further explore the relations among all the features co-varying with those associated with a given function. The mutual information statistical dependency measure has desirable properties for multi-omic integration, such as being able to capture non-linear relations and being a parameterization invariant. Moreover, we wanted to discern likely regulatory interactions, a task that has been successfully achieved with ARACNE for transcriptomics. With edges linking different types of nodes, such discerning becomes harder because ARACNE's data processing inequality (DPI) cannot be used in a straightforward manner. Thus, the setting of varying thresholds based on regulatory interactions is established. In this case, MI ability to recover non-linear relations may not be fully profited, being posterior to the lineal filter of SGCCA. MI is, however, used as a way to bring together all the results concerning a function and highlight some potentially interesting pairs of nodes.

The DPI posed with ARACNE discards the lowest weighted edge from a triad, as a likely indirect interaction driven by the other pair of nodes. The difficulty of using it comes from the observation that mutual information distributions change with the different omics (Drago-García et al., 2017). While maintaining the treatment of lower weighted interactions as indirect, the threshold we applied accounts for the difference between omics by estimating MI values from known regulatory interactions.

It is worth considering that MI has a dependency on the number of observations, which varies between subtypes and the normal tissue. Her2 enriched has a smaller number of samples than recommended, and so special care must be taken with it. Given that MI is rank-invariant, it is expected that, even with the stringent threshold, only a subset of the interactions in Figure 6 keep relevance when increasing dataset size. By progressing from a set where every feature is correlated with one another to highly significant interactions (Pethel and Hahs, 2014; Mukherjee et al., 2020), we pursue an automatic assembly of regulatory models. Tools better suited to find regulatory interactions (Kuijjer et al., 2020; Sonawane et al., 2021) require prior information not always available or heavier calculations (Weighill et al., 2021), making the approach described here an accessible solution.

To end with the pros and cons' discussion, here, we have overlooked interactions between CpG sites because those are beyond described regulatory mechanisms. Nevertheless, links between CpG sites are accompanied by large MI values that would surpass our threshold and may become of relevance in the cancer context (Akulenko and Helms, 2013; Zhang and Huang, 2017). On the other hand, links with miRNAs were expected but only appeared in the normal tissue example. Drago-García et al. had already reported lower MI values for these types of links (Drago-García et al., 2017). Despite the threshold attempted to incorporate this difference on the MI, our multi-omic pipeline does not recover miRNA interactions as well as other dedicated methods (Bose et al., 2022).

The networks produced in this way capture statistical dependencies that may guide further work. However, such a hypothetical future work depends on a user being able to find these kinds of networks and research the reasons behind a statistical dependency. Article databases can serve this purpose, as we have done here, but may become unspecific. Instead, network databases (Arif et al., 2021; Ben Guebila et al., 2022) may offer a smoother connection between wet and dry laboratories, in order to transcend statistical description toward actual knowledge acquisition.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

SO organized data, performed calculations, analyzed data, discussed results, and drafted the manuscript; EH-L designed the study, contributed to the methodological approach, discussed results, reviewed the manuscript, and supervised the project. Both authors read and approved the final manuscript.

Funding

SO is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and received fellowship 615847 from CONACYT. This work was partially performed at cluster INMEGEN and received technical support from Israel Aguilar-Ordoñez. The results published here are based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1078609/full#supplementary-material

References

Akulenko, R., and Helms, V. (2013). Dna co-methylation analysis suggests novel functional associations between gene pairs in breast cancer samples. *Hum. Mol. Genet.* 22, 3016. doi:10.1093/hmg/ddt158

Arai, E., Gotoh, M., Tian, Y., Sakamoto, H., Ono, M., Matsuda, A., et al. (2015). Alterations of the spindle checkpoint pathway in clinicopathologically aggressive c p g island methylator phenotype clear cell renal cell carcinomas. *Int. J. cancer* 137, 2589–2606. doi:10.1002/ijc.29630

Aran, D., Camarda, R., Odegaard, J., Paik, H., Oskotsky, B., Krings, G., et al. (2017). Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nat. Commun.* 8, 1077–1114. doi:10.1038/s41467-017-01027-z

Arif, M., Zhang, C., Li, X., Güngör, C., Çakmak, B., Arslantürk, M., et al. (2021). Inetmodels 2.0: an interactive visualization and database of multi-omics data. *Nucleic acids Res.* 49, W271–W276. doi:10.1093/nar/gkab254

Arraf, A. A., Yelin, R., Reshef, I., Jadon, J., Abboud, M., Zaher, M., et al. (2020). Hedgehog signaling regulates epithelial morphogenesis to position the ventral embryonic midline. *Dev. Cell.* 53, 589–602. doi:10.1016/j.devcel.2020.04.016

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., et al. (2014). Minfi: A flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays. *Bioinforma. Oxf. Engl.* 30, 1363–1369. doi:10.1093/bioinformatics/btu049

Astolfi, A., Fiore, M., Melchionda, F., Indio, V., Bertuccio, S. N., and Pession, A. (2019). Bcor involvement in cancer. *Epigenomics* 11, 835–855. doi:10.2217/epi-2018-0195

Bahado-Singh, R. O., Vishweswaraiah, S., Aydas, B., Yilmaz, A., Metpally, R. P., Carey, D. J., et al. (2021). Artificial intelligence and leukocyte epigenomics: Evaluation and prediction of late-onset alzheimer's disease. *PloS one* 16, e0248375. doi:10.1371/journal.pone.0248375

Baylin, S. B., and Jones, P. A. (2016). Epigenetic determinants of cancer. *Cold Spring Harb. Perspect. Biol.* 8, a019505. doi:10.1101/cshperspect.a019505

Bechmann, M. B., Brydholm, A. V., Codony, V. L., Kim, J., and Villadsen, R. (2020). Heterogeneity of ceacam5 in breast cancer. *Oncotarget* 11, 3886–3899. doi:10.18632/oncotarget.27778

Ben Guebila, M., Lopes-Ramos, C. M., Weighill, D., Sonawane, A. R., Burkholz, R., Shamsaei, B., et al. (2022). Grand: A database of gene regulatory network models across human conditions. *Nucleic acids Res.* 50, D610–D621. doi:10.1093/nar/gkab778

Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., et al. (2016). Methods for the integration of multi-omics data: Mathematical aspects. *BMC Bioinforma.* 17, S15. doi:10.1186/s12859-015-0857-9

Bieluszewska, A., Weglewska, M., Bieluszewski, T., Lesniewicz, K., and Poreba, E. (2018). Pka-binding domain of akap 8 is essential for direct interaction with dpy 30 protein. *FEBS J.* 285, 947–964. doi:10.1111/febs.14378

Bose, B., and Bozdag, S. (2019). "mirdriver: A tool to infer copy number derived mirna-gene networks in cancer," in *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, 366.

Bose, B., Moravec, M., and Bozdag, S. (2022). Computing microrna-gene interaction networks in pan-cancer using mirdriver. *Sci. Rep.* 12, 3717–17. doi:10.1038/s41598-022-07628-z

Bueno, D. F., Sunaga, D. Y., Kobayashi, G. S., Aguena, M., Raposo-Amaral, C. E., Masotti, C., et al. (2011). Human stem cell cultures from cleft lip/palate patients show enrichment of transcripts involved in extracellular matrix modeling by comparison to controls. *Stem Cell. Rev. Rep.* 7, 446–457. doi:10.1007/s12015-010-9197-3

Cerhan, J. R., Berndt, S. I., Vijai, J., Ghesquières, H., McKay, J., Wang, S. S., et al. (2014). Genome-wide association study identifies multiple susceptibility loci for diffuse large b cell lymphoma. *Nat. Genet.* 46, 1233–1238. doi:10.1038/ng.3105

Chappell, K., Manna, K., Washam, C. L., Graw, S., Alkam, D., Thompson, M. D., et al. (2021). Multi-omics data integration reveals correlated regulatory features of triple negative breast cancer. *Mol. Omics* 17, 677–691. doi:10.1039/d1mo00117e

Consortium, E. P., et al. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature* 489, 57–74. doi:10.1038/nature11247

Consortium, G. O. (2021). The gene ontology resource: Enriching a gold mine. *Nucleic Acids Res.* 49, D325–D334. doi:10.1093/nar/gkaa1113

Corominas-Faja, B., Cuyàs, E., Gumuzio, J., Bosch-Barrera, J., Leis, O., Martin, Á. G., et al. (2014). Chemical inhibition of acetyl-coa carboxylase suppresses self-renewal growth of cancer stem cells. *Oncotarget* 5, 8306–8316. doi:10.18632/oncotarget.2059

Corrado, C., and Fontana, S. (2020). Hypoxia and hif signaling: One axis with divergent effects. *Int. J. Mol. Sci.* 21, 5611. doi:10.3390/ijms21165611

Csardi, G., and Nepusz, T. (2006). *The igraph software package for complex network research*. Cambridge, MA: NECSI, 1695.

D'Aloia, A., Berruti, G., Costa, B., Schiller, C., Ambrosini, R., Pastori, V., et al. (2018). Ralgps2 is involved in tunneling nanotubes formation in 5637 bladder cancer cells. *Exp. Cell. Res.* 362, 349–361. doi:10.1016/j.yexcr.2017.11.036

de Heer, E. C., Jalving, M., Harris, A. L., et al. (2020). Hifs, angiogenesis, and metabolism: Elusive enemies in breast cancer. *J. Clin. investigation* 130, 5074–5087. doi:10.1172/JCI137552

De Tayrac, M., Lê, S., Aubry, M., Mosser, J., and Husson, F. (2009). Simultaneous analysis of distinct omics data sets with integration of biological knowledge: Multiple factor analysis approach. *BMC genomics* 10, 32. doi:10.1186/1471-2164-10-32

Dorantes-Gilardi, R., García-Cortés, D., Hernández-Lemus, E., and Espinal-Enríquez, J. (2021). k-core genes underpin structural features of breast cancer. *Sci. Rep.* 11, 16284–16317. doi:10.1038/s41598-021-95313-y

Drago-García, D., Espinal-Enríquez, J., and Hernández-Lemus, E. (2017). Network analysis of emt and met micro-rna regulation in breast cancer. *Sci. Rep.* 7, 13534. doi:10.1038/s41598-017-13903-1

Fan, Z., Zhou, Y., and Ressom, H. W. (2020). Mota: Network-based multi-omic data integration for biomarker discovery. *Metabolites* 10, 144. doi:10.3390/metabo10040144

Farrugia, M., Sharma, S., Lin, C., McLaughlin, S., Vanderbilt, D., Ammer, A., et al. (2015). Regulation of anti-apoptotic signaling by kruppel-like factors 4 and 5 mediates lapatinib resistance in breast cancer. *Cell. death Dis.* 6, e1699. doi:10.1038/cddis.2015.65

Fernández-Centeno, E., de Ojeda, G., Rojo, J. M., and Portolés, P. (2000). Crry/p65, a membrane complement regulatory protein, has costimulatory properties on mouse t cells. *J. Immunol.* 164, 4533–4542. doi:10.4049/jimmunol.164.9.4533

Fu, N. Y., Rios, A. C., Pal, B., Soetanto, R., Lun, A. T., Liu, K., et al. (2015). Egf-mediated induction of mcl-1 at the switch to lactation is essential for alveolar cell survival. *Nat. Cell. Biol.* 17, 365–375. doi:10.1038/ncb3117

Furukawa-Hibi, Y., Nagai, T., Yun, J., and Yamada, K. (2015). Stress increases dna methylation of the neuronal pas domain 4 (npas4) gene. *Neuroreport* 26, 827–832. doi:10.1097/WNR.0000000000000430

Garali, I., Adanyeguh, I. M., Ichou, F., Perlbarg, V., Seyer, A., Colsch, B., et al. (2018). A strategy for multimodal data integration: Application to biomarkers identification in spinocerebellar ataxia. *Briefings Bioinforma.* 19, 1356–1369. doi:10.1093/bib/bbx060

García-Cortés, D., de Anda-Jáuregui, G., Fresno, C., Hernández-Lemus, E., and Espinal-Enríquez, J. (2020). Gene co-expression is distance-dependent in breast cancer. *Front. Oncol.* 10, 1232. doi:10.3389/fonc.2020.01232

García-Cortés, D., Hernández-Lemus, E., and Espinal-Enríquez, J. (2021). Luminal a breast cancer co-expression network: Structural and functional alterations. *Front. Genet.* 12, 629475. doi:10.3389/fgene.2021.629475

Gehlenborg, N. (2019). *UpSetR: A more scalable alternative to venn and euler diagrams for visualizing intersecting sets*. Oxford, England: Oxford Academic.

Grand, R. S., Burger, L., Gräwe, C., Michael, A. K., Isbel, L., Hess, D., et al. (2021). Banp opens chromatin and activates cpg-island-regulated genes. *Nature* 596, 133–137. doi:10.1038/s41586-021-03689-8

Gustavsen, A., J., Pai, S., Isserlin, R., et al. (2019). Rcy3: Network biology using cytoscape from within r. *F1000Research* doi:10.12688/f1000research.20887.3

Han, H., Shim, H., Shin, D., Shim, J. E., Ko, Y., Shin, J., et al. (2015). Trrust: A reference database of human transcriptional regulatory interactions. *Sci. Rep.* 5, 11432–11511. doi:10.1038/srep11432

Hazelett, C. C., and Yeaman, C. (2012). Sec5 and exo84 mediate distinct aspects of rala-dependent cell polarization. *PLoS One* 7, e39602. doi:10.1371/journal.pone.0039602

He, C., Imai, M., Song, H., Quigg, R. J., and Tomlinson, S. (2005). Complement inhibitors targeted to the proximal tubule prevent injury in experimental nephrotic syndrome and demonstrate a key role for c5b-9. *J. Immunol.* 174, 5750–5757. doi:10.4049/jimmunol.174.9.5750

Hernández-de Diego, R., Tarazona, S., Martínez-Mira, C., Balzano-Nogueira, L., Furió-Tarí, P., Pappas, G. J., et al. (2018). Paintomics 3: A web resource for the pathway analysis and visualization of multi-omics data. *Nucleic acids Res.* 46, W503-W509. doi:10.1093/nar/gky466

Hollander, D., Donyo, M., Atias, N., Mekahel, K., Melamed, Z., Yannai, S., et al. (2016). A network-based analysis of colon cancer splicing changes reveals a tumorigenesis-favoring regulatory pathway emanating from elk1. *Genome Res.* 26, 541–553. doi:10.1101/gr.193169.115

Huang, S., Chaudhary, K., and Garmire, L. X. (2017). More is better: Recent progress in multi-omics data integration methods. *Front. Genet.* 8, 84. doi:10.3389/fgene.2017.00084

Huang, S., Xu, W., Hu, P., and Lakowski, T. M. (2019). Integrative analysis reveals subtype-specific regulatory determinants in triple negative breast cancer. *Cancers* 11, 507. doi:10.3390/cancers11040507

Izquierdo, A. G., Boughanem, H., Diaz-Lagares, A., Arranz-Salas, I., Esteller, M., Tinahones, F. J., et al. (2021). Dna methylome in visceral adipose tissue can discriminate patients with and without colorectal cancer. *Epigenetics* 1–12, 665–676. doi:10.1080/15592294.2021.1950991

Jafarzadeh, M., Soltani, B. M., Soleimani, M., and Hosseinkhani, S. (2020). Epigenetically silenced linc02381 functions as a tumor suppressor by regulating pi3k-akt signaling pathway. *Biochimie* 171, 63–71. doi:10.1016/j.biochi.2020.02.009

Jiang, C., Xuan, Z., Zhao, F., and Zhang, M. Q. (2007). Tred: A transcriptional regulatory element database, new entries and other development. *Nucleic acids Res.* 35, D137–D140. doi:10.1093/nar/gkl1041

Kanehisa, M., and Goto, S. (2000). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic acids Res.* 28, 27–30. doi:10.1093/nar/28.1.27

Ke, H., Wu, Y., Wang, R., and Wu, X. (2020). Creation of a prognostic risk prediction model for lung adenocarcinoma based on gene expression, methylation, and clinical characteristics. *Med. Sci. Monit. Int. Med. J. Exp. Clin. Res.* 26, 9258333–e925841. doi:10.12659/MSM.925833

Kristensen, V. N., Lingjærde, O. C., Russnes, H. G., Vollan, H. K. M., Frigessi, A., and Børresen-Dale, A.-L. (2014). Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer* 14, 299–313. doi:10.1038/nrc3721

Kuijjer, M. L., Fagny, M., Marin, A., Quackenbush, J., and Glass, K. (2020). Puma: Panda using microrna associations. *Bioinformatics* 36, 4765–4773. doi:10.1093/bioinformatics/btaa571

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., et al. (2018). The human transcription factors. *Cell.* 172, 650–665. doi:10.1016/j.cell.2018.01.029

Li, J., Xu, Y.-H., Lu, Y., Ma, X.-P., Chen, P., Luo, S.-W., et al. (2013). Identifying differentially expressed genes and small molecule drugs for prostate cancer by a bioinformatics strategy. *Asian Pac. J. cancer Prev.* 14, 5281–5286. doi:10.7314/apjcp.2013.14.9.5281

Li, W., Zhang, S., Liu, C.-C., and Zhou, X. J. (2012). Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics* 28, 2458–2466. doi:10.1093/bioinformatics/bts476

Li, Y., Wang, Y., Wen, Y., Zhang, T., Wang, X., Jiang, C., et al. (2022). Whole-exome sequencing of a cohort of infertile men reveals novel causative genes in teratozoospermia that are chiefly related to sperm head defects. *Hum. Reprod.* 37, 152–177. doi:10.1093/humrep/deab229

Li, Y., Zhao, X., Liu, Q., and Liu, Y. (2021). Bioinformatics reveal macrophages marker genes signature in breast cancer to predict prognosis. *Ann. Med.* 53, 1019–1031. doi:10.1080/07853890.2021.1914343

Liang, Y.-K., Lin, H.-Y., Dou, X.-W., Chen, M., Wei, X.-L., Zhang, Y.-Q., et al. (2018). Mir-221/222 promote epithelial-mesenchymal transition by targeting notch3 in breast cancer cell lines. *NPJ breast cancer* 4, 20–29. doi:10.1038/s41523-018-0073-7

Liu, J., Liang, G., Siegmund, K. D., and Lewinger, J. P. (2018). Data integration by multi-tuning parameter elastic net regression. *BMC Bioinforma.* 19, 369. doi:10.1186/s12859-018-2401-1

Liu, J., Liu, X., Zhou, G., Xiao, R., and Cao, Y. (2012). Conditioned medium from chondrocyte/scaffold constructs induced chondrogenic differentiation of bone marrow stromal cells. *Anatomical Rec. Adv. Integr. Anat. Evol. Biol.* 295, 1109–1116. doi:10.1002/ar.22500

Luo, Y., Sun, F., Peng, X., Dong, D., Ou, W., Xie, Y., et al. (2021). Integrated bioinformatics analysis to identify abnormal methylated differentially expressed genes for predicting prognosis of human colon cancer. *Int. J. General Med.* 14, 4745–4756. doi:10.2147/IJGM.S324483

Maksimovic, J., Gagnon-Bartsch, J. A., Speed, T. P., and Oshlack, A. (2015). Removing unwanted variation in a differential methylation analysis of illumina humanmethylation450 array data. *Nucleic acids Res.* 43, e106. doi:10.1093/nar/gkv526

Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., and Bergmann, S. (2016). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat. methods* 13, 366–370. doi:10.1038/nmeth.3799

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., et al. (2006). Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinforma. Biomed. Cent.* 7, S7. doi:10.1186/1471-2105-7-S1-S7

Matic, I., van Hagen, M., Schimmel, J., Macek, B., Ogg, S. C., Tatham, M. H., et al. (2008). *In vivo* identification of human small ubiquitin-like modifier polymerization sites by high accuracy mass spectrometry and an *in vitro* to *in vivo* strategy. *Mol. Cell. proteomics* 7, 132–144. doi:10.1074/mcp.M700173-MCP200

Meng, C., Basunia, A., Peters, B., Gholami, A. M., Kuster, B., and Culhane, A. C. (2019). Mogsa: Integrative single sample gene-set analysis of multiple omics data. *Mol. Cell. Proteomics* 18, S153-S168–S168. doi:10.1074/mcp.TIR118.001251

Messaoudene, M., Mourikis, T., Michels, J., Fu, Y., Bonvalet, M., Lacroix-Trikki, M., et al. (2019). T-Cell bispecific antibodies in node-positive breast cancer: Novel therapeutic avenue for mhc class i loss variants. *Ann. Oncol.* 30, 934–944. doi:10.1093/annonc/mdz112

Meyer, P. E. (2014). *Infotheo: Information-Theoretic measures*. Princeton, NJ: R. package.

Moldogazieva, N. T., Mokhosoev, I. M., and Terentiev, A. A. (2020). Metabolic heterogeneity of cancer cells: An interplay between hif-1, gluts, and ampk. *Cancers* 12, 862. doi:10.3390/cancers12040862

Mukherjee, S., Asnani, H., and Kannan, S. (2020). "Ccmi: Classifier based conditional mutual information estimation," in Proceedings of Machine Learning Research.

Nakazawa, T., Nakano, I., Sato, M., Nakamura, T., Tamai, M., and Mori, N. (2002). Comparative expression profiles of trk receptors and shc-related phosphotyrosine adapters during retinal development: Potential roles of n-shc/shcc in brain-derived neurotrophic factor signal transduction and modulation. *J. Neurosci. Res.* 68, 668–680. doi:10.1002/jnr.10259

Neph, S., Stergachis, A. B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J. A. (2012). Circuitry and dynamics of human transcription factor regulatory networks. *Cell.* 150, 1274–1286. doi:10.1016/j.cell.2012.04.040

Network, C. G. A., et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70. doi:10.1038/nature11412

Nueda, M. J., Ferrer, A., and Conesa, A. (2012). Arsyn: A method for the identification and removal of systematic noise in multifactorial time course microarray experiments. *Biostatistics* 13, 553–566. doi:10.1093/biostatistics/kxr042

Ochoa, S., de Anda-Jáuregui, G., and Hernández-Lemus, E. (2021). An information theoretical multilayer network approach to breast cancer transcriptional regulation. *Front. Genet.* 12, 617512. doi:10.3389/fgene.2021.617512

Pathak, B. R., Breed, A. A., Apte, S., Acharya, K., and Mahale, S. D. (2016). Cysteine-rich secretory protein 3 plays a role in prostate cancer cell invasion and affects expression of psa and anxa1. *Mol. Cell. Biochem.* 411, 11–21. doi:10.1007/s11010-015-2564-2

Patra, S. K. (2008). Ras regulation of dna-methylation and cancer. *Exp. Cell. Res.* 314, 1193–1201. doi:10.1016/j.yexcr.2008.01.012

Pethel, S. D., and Hahs, D. W. (2014). Exact test of independence using mutual information. *Entropy* 16, 2839–2849. doi:10.3390/e16052839

Piao, H.-l., Yuan, Y., Wang, M., Sun, Y., Liang, H., and Ma, L. (2014). α-catenin acts as a tumour suppressor in e-cadherin-negative basal-like breast cancer by inhibiting nf-κb signalling. *Nat. Cell. Biol.* 16, 245–254. doi:10.1038/ncb2909

Privat, M., Rudewicz, J., Sonnier, N., Tamisier, C., Ponelle-Chachuat, F., and Bignon, Y.-J. (2018). Antioxydation and cell migration genes are identified as potential therapeutic targets in basal-like and brca1 mutated breast cancer cell lines. *Int. J. Med. Sci.* 15, 46–58. doi:10.7150/ijms.20508

Pupa, S. M., Ligorio, F., Cancila, V., Franceschini, A., Tripodo, C., Vernieri, C., et al. (2021). Her2 signaling and breast cancer stem cells: The bridge behind her2-positive breast cancer aggressiveness and therapy refractoriness. *Cancers* 13, 4778. doi:10.3390/cancers13194778

R Core Team (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramdzan, Z. M., Vadnais, C., Pal, R., Vandal, G., Cadieux, C., Leduy, L., et al. (2014). Ras transformation requires cux1-dependent repair of oxidative dna damage. *PLoS Biol.* 12, e1001807. doi:10.1371/journal.pbio.1001807

Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). GC-content normalization for RNA-seq data. *BMC Bioinforma.* 12, 480. doi:10.1186/1471-2105-12-480

Rohart, F., Gautier, B., Singh, A., and Le Cao, K.-A. (2017). mixomics: An r package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* 13, e1005752. doi:10.1371/journal.pcbi.1005752

Rose, N. R., and Klose, R. J. (2014). Understanding the relationship between dna methylation and histone lysine methylation. *Biochimica Biophysica Acta (BBA)-Gene Regul. Mech.* 1839, 1362–1372. doi:10.1016/j.bbagrm.2014.02.007

Ru, Y., Kechris, K. J., Tabakoff, B., Hoffman, P., Radcliffe, R. A., Bowler, R., et al. (2014). The multimir r package and database: Integration of microrna–target interactions along with their disease and drug associations. *Nucleic acids Res.* 42, e133. doi:10.1093/nar/gku631

Schulz, D. M., Bollner, C., Thomas, G., Atkinson, M., Esposito, I., Hofler, H., et al. (2009). Identification of differentially expressed proteins in triple-negative breast carcinomas using dige and mass spectrometry. *J. proteome Res.* 8, 3430–3438. doi:10.1021/pr900071h

Shamay, M., Greenway, M., Liao, G., Ambinder, R. F., and Hayward, S. D. (2010). De novo dna methyltransferase dnmt3b interacts with nedd8-modified proteins. *J. Biol. Chem.* 285, 36377–36386. doi:10.1074/jbc.M110.155721

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi:10.1101/gr.1239303

Sohn, K.-A., Kim, D., Lim, J., and Kim, J. H. (2013). Relative impact of multi-layered genomic data on gene expression phenotypes in serous ovarian tumors. *BMC Syst. Biol.* 7, S9. doi:10.1186/1752-0509-7-S6-S9

Sonawane, A. R., DeMeo, D. L., Quackenbush, J., and Glass, K. (2021). Constructing gene regulatory networks using epigenetic data. *npj Syst. Biol. Appl.* 7, 45–13. doi:10.1038/s41540-021-00208-3

Strelnikov, V. V., Kuznetsova, E. B., Tanas, A. S., Rudenko, V. V., Kalinkin, A. I., Poddubskaya, E. V., et al. (2021). Abnormal promoter dna hypermethylation of the integrin, nidogen, and dystroglycan genes in breast cancer. *Sci. Rep.* 11, 2264–2314. doi:10.1038/s41598-021-81851-y

Sugiyama, A., Okada, M., and Yamawaki, H. (2020). Canstatin suppresses isoproterenol-induced cardiac hypertrophy through inhibition of calcineurin/nuclear factor of activated t-cells pathway in rats. *Eur. J. Pharmacol.* 871, 172849. doi:10.1016/j.ejphar.2019.172849

Sumi, E., Iehara, N., Akiyama, H., Matsubara, T., Mima, A., Kanamori, H., et al. (2007). Sry-related hmg box 9 regulates the expression of col4a2 through transactivating its enhancer element in mesangial cells. *Am. J. pathology* 170, 1854–1864. doi:10.2353/ajpath.2007.060899

Tam, S., Tsao, M.-S., and McPherson, J. D. (2015). Optimization of mirna-seq data preprocessing. *Briefings Bioinforma.* 16, 950–963. doi:10.1093/bib/bbv019

Tan, Z., Niu, B., Tsang, K. Y., Melhado, I. G., Ohba, S., He, X., et al. (2018). Synergistic co-regulation and competition by a sox9-gli-foxa phasic transcriptional network coordinate chondrocyte differentiation transitions. *PLoS Genet.* 14, e1007346. doi:10.1371/journal.pgen.1007346

Tapia-Carrillo, D., Tovar, H., Velazquez-Caldelas, T. E., and Hernandez-Lemus, E. (2019). Master regulators of signaling pathways: An application to the analysis of gene regulation in breast cancer. *Front. Genet.* 10, 1180. doi:10.3389/fgene.2019.01180

Tarazona, S., Furió-Tarí, P., Turrà, D., Pietro, A. D., Nueda, M. J., Ferrer, A., et al. (2015). Data quality aware analysis of differential expression in rna-seq with noiseq r/bioc package. *Nucleic acids Res.* 43, e140. doi:10.1093/nar/gkv711

Tenenhaus, A., Philippe, C., Guillemot, V., Le Cao, K.-A., Grill, J., and Frouin, V. (2014). Variable selection for generalized canonical correlation analysis. *Biostatistics* 15, 569–583. doi:10.1093/biostatistics/kxu001

Thammaiah, C. K., and Jayaram, S. (2016). Role of let-7 family microrna in breast cancer. *Non-coding RNA Res.* 1, 77–82. doi:10.1016/j.ncrna.2016.10.003

Vidal Ocabo, E., Sayols, S., Moran, S., Guillaumet-Adkins, A., Schroeder, M. P., Royo, R., et al. (2017). A dna methylation map of human cancer at single base-pair resolution. *Oncogene* 36 (40), 5648–5657. doi:10.1038/onc.2017.176

Weigel Muñoz, M., Carvajal, G., Curci, L., Gonzalez, S. N., and Cuasnicu, P. S. (2019). Relevance of crisp proteins for epididymal physiology, fertilization, and fertility. *Andrology* 7, 610–617. doi:10.1111/andr.12638

Weighill, D., Burkholz, R., Guebila, M. B., Zacharias, H. U., Quackenbush, J., and Altenbuchinger, M. (2021). DRAGON: Determining regulatory associations using graphical models on multi-omic networks. *Oxford, England: Nucleic Acids Res.* [Epub ahead of print]. doi:10.1093/nar/gkac1157

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.

Wollen, E. J., Sejersted, Y., Wright, M. S., Bik-Multanowski, M., Madetko-Talowska, A., Günther, C.-C., et al. (2013). Transcriptome profiling of the newborn mouse lung after hypoxia and reoxygenation: Hyperoxic reoxygenation affects mtor signaling pathway, dna repair, and jnk-pathway regulation. *Pediatr. Res.* 74, 536–544. doi:10.1038/pr.2013.140

Wright, K. L., Adams, J. R., Liu, J. C., Loch, A. J., Wong, R. G., Jo, C. E., et al. (2015). Ras signaling is a key determinant for metastatic dissemination and poor survival of luminal breast cancer patients. *Cancer Res.* 75, 4960–4972. doi:10.1158/0008-5472.CAN-14-2992

Wu, J., Wang, C., Miao, X., Wu, Y., Yuan, J., Ding, M., et al. (2017). Age-related insulin-like growth factor binding protein-4 overexpression inhibits osteogenic differentiation of rat mesenchymal stem cells. *Cell. Physiology Biochem.* 42, 640–650. doi:10.1159/000477873

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., et al. (2021). Clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* 2, 100141. doi:10.1016/j.xinn.2021.100141

Xin, S., Fang, W., Li, J., Li, D., Wang, C., Huang, Q., et al. (2021). Impact of stat1 polymorphisms on crizotinib-induced hepatotoxicity in alk-positive non-small cell lung cancer patients. *J. Cancer Res. Clin. Oncol.* 147, 725–737. doi:10.1007/s00432-020-03476-4

Xiong, B., Lei, X., Zhang, L., and Fu, J. (2017). mir-103 regulates triple negative breast cancer cells migration and invasion through targeting olfactomedin 4. *Biomed. Pharmacother.* 89, 1401–1408. doi:10.1016/j.biopha.2017.02.028

Xu, N., Wu, Y.-P., Ke, Z.-B., Liang, Y.-C., Cai, H., Su, W.-T., et al. (2019). Identification of key dna methylation-driven genes in prostate adenocarcinoma: An integrative analysis of tcga methylation data. *J. Transl. Med.* 17, 311–315. doi:10.1186/s12967-019-2065-2

Xu, X., Zhang, M., Xu, F., and Jiang, S. (2020). Wnt signaling in breast cancer: Biological mechanisms, challenges and opportunities. *Mol. cancer* 19, 165–235. doi:10.1186/s12943-020-01276-5

Zamora-Fuentes, J. M., Hernández-Lemus, E., and Espinal-Enríquez, J. (2022). Oncogenic role of mir-217 during clear cell renal carcinoma progression. *Front. Oncol.* 12, 934711. doi:10.3389/fonc.2022.934711

Zhang, J., and Huang, K. (2017). Pan-cancer analysis of frequent dna co-methylation patterns reveals consistent epigenetic landscape changes in multiple cancers. *Bmc Genomics* 18, 1045–1114. doi:10.1186/s12864-016-3259-0

Zhang, P., Wen, X., Gu, F., Deng, X., Li, J., Dong, J., et al. (2013). Methylation profiling of serum dna from hepatocellular carcinoma patients using an infinium human methylation 450 beadchip. *Hepatol. Int.* 7, 893–900. doi:10.1007/s12072-013-9437-0

Zhao, S., Geybels, M. S., Leonardson, A., Rubicz, R., Kolb, S., Yan, Q., et al. (2017). Epigenome-Wide tumor DNA methylation profiling identifies novel prognostic biomarkers of metastatic-lethal progression in men diagnosed with clinically localized prostate cancer. *Clin. Cancer Res.* 23, 311–319. doi:10.1158/1078-0432.CCR-16-0549

Zheng, G., Tu, K., Yang, Q., Xiong, Y., Wei, C., Xie, L., et al. (2008). Itfp: An integrated platform of mammalian transcription factors. *Bioinformatics* 24, 2416–2417. doi:10.1093/bioinformatics/btn439

Bibliografía

[1] C. G. A. Network *et al.*, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, 2012.

[2] A. C. Berger, A. Korkut, R. S. Kanchi, A. M. Hegde, W. Lenoir, W. Liu, Y. Liu, H. Fan, H. Shen, V. Ravikumar, *et al.*, "A comprehensive pan-cancer molecular study of gynecologic and breast cancers," *Cancer Cell*, 2018.

[3] M. R. Corces, J. M. Granja, S. Shams, B. H. Louie, J. A. Seoane, W. Zhou, T. C. Silva, C. Groeneveld, C. K. Wong, S. W. Cho, A. T. Satpathy, M. R. Mumbach, K. A. Hoadley, A. G. Robertson, N. C. Sheffield, I. Felau, M. A. A. Castro, B. P. Berman, L. M. Staudt, J. C. Zenklusen, P. W. Laird, C. Curtis, C. G. A. A. Network, W. J. Greenleaf, and H. Y. Chang, "The chromatin accessibility landscape of primary human cancers.," *Science (New York, N.Y.)*, vol. 362, Oct. 2018.

[4] C. Buccitelli and M. Selbach, "mrnas, proteins and the emerging principles of gene expression control," *Nature Reviews Genetics*, vol. 21, no. 10, pp. 630–644, 2020.

[5] C. M. Perou, T. Sørlie, M. B. Eisen, M. Van De Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, *et al.*, "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, p. 747, 2000.

[6] V. N. Kristensen, O. C. Lingjærde, H. G. Russnes, H. K. M. Vollan, A. Frigessi, and A.-L. Børresen-Dale, "Principles and methods of integrative genomic analyses in cancer," *Nature Reviews Cancer*, vol. 14, no. 5, pp. 299–313, 2014.

[7] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries.," *CA: a cancer journal for clinicians*, Feb. 2021.

[8] INEGI, "Estadísticas a propósito del día mundial de la lucha contra el cáncer de mama (19 de octubre)."

[9] Consenso cancer mamario, *Consenso Mexicano sobre diagnóstico y tratamiento del cáncer mamario*, Oct. 2019.

[10] Q. Feng, M. Jiang, J. Hannig, and J. Marron, "Angle-based joint and individual variation explained," *Journal of multivariate analysis*, vol. 166, pp. 241–265, 2018.

[11] F. F. Parl, *The etiology of breast cancer: Endogenous and exogenous causes*. F.F. Parl, 2014.

[12] C. Wu, E. W. Demerath, J. S. Pankow, J. Bressler, M. Fornage, M. L. Grove, W. Chen, and W. Guan, "Imputation of missing covariate values in epigenome-wide analysis of dna methylation data.," *Epigenetics*, vol. 11, pp. 132–139, 2016.

[13] D. Pellacani, S. Tan, S. Lefort, and C. J. Eaves, "Transcriptional regulation of normal human mammary cell heterogeneity and its perturbation in breast cancer," *The EMBO journal*, vol. 38, no. 14, p. e100330, 2019.

[14] A. H. Sims, A. Howell, S. J. Howell, and R. B. Clarke, "Origins of breast cancer subtypes and therapeutic implications.," *Nature clinical practice. Oncology*, vol. 4, pp. 516–525, Sept. 2007.

[15] G. Turashvili and E. Brogi, "Tumor heterogeneity in breast cancer," *Frontiers in medicine*, vol. 4, p. 227, 2017.

[16] X. Dai, L. Xiang, T. Li, and Z. Bai, "Cancer hallmarks, biomarkers and breast cancer molecular subtypes," *Journal of Cancer*, vol. 7, no. 10, p. 1281, 2016.

[17] J. S. Parker, M. Mullins, M. C. U. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, J. F. Quackenbush, I. J. Stijleman, J. Palazzo, J. S. Marron, A. B. Nobel, E. Mardis, T. O. Nielsen, M. J. Ellis, C. M. Perou, and P. S. Bernard, "Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes," *Journal of Clinical Oncology*, vol. 27, pp. 1160–1167, Mar. 2009.

[18] A. Prat, E. Pineda, B. Adamo, P. Galván, A. Fernández, L. Gaba, M. Díez, M. Viladot, A. Arance, and M. Muñoz, "Clinical implications of the intrinsic molecular subtypes of breast cancer," *The Breast*, vol. 24, pp. S26–S35, 2015.

[19] A. Daemen and G. Manning, "Her2 is not a cancer subtype but rather a pan-cancer event and is highly enriched in ar-driven breast tumors," *Breast Cancer Research*, vol. 20, no. 1, pp. 1–16, 2018.

[20] B. Lehmann, J. Bauer, X. Chen, and M. Sanders, "Chakravart hy ab, shyr y, pietenpol ja. identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies," *J clin Invest*, vol. 121, pp. 2750–67, 2011.

[21] B. D. Lehmann, A. Colaprico, T. C. Silva, J. Chen, H. An, Y. Ban, H. Huang, L. Wang, J. L. James, J. M. Balko, *et al.*, "Multi-omics analysis identifies therapeutic vulnerabilities in triple-negative breast cancer subtypes," *Nature communications*, vol. 12, no. 1, pp. 1–18, 2021.

[22] X. Li, J. Zhou, M. Xiao, L. Zhao, Y. Zhao, S. Wang, S. Gao, Y. Zhuang, Y. Niu, S. Li, *et al.*, "Uncovering the subtype-specific molecular characteristics of breast cancer by multiomics analysis of prognosis-associated genes, driver genes, signaling pathways, and immune activity," *Frontiers in Cell and Developmental Biology*, vol. 9, 2021.

[23] Y. Xiao, D. Ma, S. Zhao, C. Suo, J. Shi, M.-Z. Xue, M. Ruan, H. Wang, J. Zhao, Q. Li, *et al.*, "Multi-omics profiling reveals distinct microenvironment characterization and suggests immune escape mechanisms of triple-negative breast cancermicroenvironment landscape of triple-negative breast cancer," *Clinical cancer research*, vol. 25, no. 16, pp. 5002–5014, 2019.

[24] J. N. Rich, "Cancer stem cells: understanding tumor hierarchy and heterogeneity," *Medicine*, vol. 95, no. Suppl 1, 2016.

[25] D. García-Cortés, E. Hernández-Lemus, and J. Espinal-Enríquez, "Luminal a breast cancer co-expression network: Structural and functional alterations," *Frontiers in genetics*, vol. 12, p. 629475, 2021.

[26] E. A. Serrano-Carbajal, J. Espinal-Enríquez, and E. Hernández-Lemus, "Targeting metabolic deregulation landscapes in breast cancer subtypes," *Frontiers in Oncology*, vol. 10, p. 97, 2020.

[27] B. Vogelstein and K. W. Kinzler, "Cancer genes and the pathways they control," *Nat Med*, vol. 10, p. 789–799, Aug 2004.

[28] S. B. Baylin and P. A. Jones, "Epigenetic determinants of cancer," *Cold Spring Harbor perspectives in biology*, vol. 8, no. 9, p. a019505, 2016.

[29] B. Bose and S. Bozdag, "mirdriver: A tool to infer copy number derived mirna-gene networks in cancer," in *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 366–375, 2019.

[30] D. Hanahan, "Hallmarks of cancer: new dimensions," *Cancer discovery*, vol. 12, no. 1, pp. 31–46, 2022.

[31] D. W. Bell, "Our changing view of the genomic landscape of cancer," *The Journal of Pathology*, p. n/a–n/a, 2009.

[32] A. R. Anderson and P. K. Maini, "Mathematical oncology," *Bulletin of mathematical biology*, vol. 80, no. 5, pp. 945–953, 2018.

[33] J. Cao, Z. Luo, Q. Cheng, Q. Xu, Y. Zhang, F. Wang, Y. Wu, and X. Song, "Three-dimensional regulation of transcription," *Protein & cell*, vol. 6, no. 4, pp. 241–253, 2015.

[34] G. Kustatscher, P. Grabowski, and J. Rappsilber, "Pervasive coexpression of spatially proximal genes is buffered at the protein level," *Molecular systems biology*, vol. 13, no. 8, p. 937, 2017.

[35] E. Li and Y. Zhang, "Dna methylation in mammals," *Cold Spring Harbor perspectives in biology*, vol. 6, no. 5, p. a019133, 2014.

[36] C.-L. Xiao, S. Zhu, M. He, D. Chen, Q. Zhang, Y. Chen, G. Yu, J. Liu, S.-Q. Xie, F. Luo, *et al.*, "N6-methyladenine dna modification in the human genome," *Molecular cell*, vol. 71, no. 2, pp. 306–318, 2018.

[37] P. A. Jones and S. B. Baylin, "The fundamental role of epigenetic events in cancer," *Nature reviews genetics*, vol. 3, no. 6, pp. 415–428, 2002.

[38] C. Cava, G. Bertoli, and I. Castiglioni, "Integrating genetics and epigenetics in breast cancer: biological insights, experimental, computational methods and therapeutic potential," *BMC systems biology*, vol. 9, no. 1, p. 62, 2015.

[39] C. Leygo, M. Williams, H. C. Jin, M. W. Chan, W. K. Chu, M. Grusch, and Y. Y. Cheng, "Dna methylation as a noninvasive epigenetic biomarker for the detection of cancer," *Disease markers*, vol. 2017, 2017.

[40] R. Pidsley, E. Zotenko, T. J. Peters, M. G. Lawrence, G. P. Risbridger, P. Molloy, S. Van Djik, B. Muhlhausler, C. Stirzaker, and S. J. Clark, "Critical evaluation of the illumina methylationepic beadchip microarray for whole-genome dna methylation profiling," *Genome biology*, vol. 17, no. 1, pp. 1–17, 2016.

[41] T. Vavouri and B. Lehner, "Human genes with cpg island promoters have a distinct transcription-associated chromatin organization," *Genome biology*, vol. 13, no. 11, p. R110, 2012.

[42] R. Akulenko and V. Helms, "Dna co-methylation analysis suggests novel functional associations between gene pairs in breast cancer samples," *Human molecular genetics*, vol. 22, no. 15, pp. 3016–3022, 2013.

[43] M. Szyf, "Dna methylation signatures for breast cancer classification and prognosis," *Genome medicine*, vol. 4, no. 3, p. 26, 2012.

[44] N. Loyfer, J. Magenheim, A. Peretz, G. Cann, J. Bredno, A. Klochendler, I. Fox-Fisher, S. Shabi-Porat, M. Hecht, T. Pelet, *et al.*, "A dna methylation atlas of normal human cell types," *Nature*, pp. 1–10, 2023.

[45] J. Li, C. Sun, W. Cai, J. Li, B. P. Rosen, and J. Chen, "Insights into s-adenosyl-l-methionine (sam)-dependent methyltransferase related diseases and genetic polymorphisms," *Mutation Research/Reviews in Mutation Research*, vol. 788, p. 108396, 2021.

[46] H. Y. Zoghbi and A. L. Beaudet, "Epigenetics and human disease," *Cold Spring Harbor perspectives in biology*, vol. 8, no. 2, p. a019497, 2016.

[47] S. Kriaucionis and M. Tahiliani, "Expanding the epigenetic landscape: novel modifications of cytosine in genomic dna," *Cold Spring Harbor perspectives in biology*, vol. 6, no. 10, p. a018630, 2014.

[48] G. Wang, X. Luo, J. Wang, J. Wan, S. Xia, H. Zhu, J. Qian, and Y. Wang, "Medreaders: a database for transcription factors that bind to methylated dna," *Nucleic acids research*, vol. 46, no. D1, pp. D146–D151, 2018.

[49] A. Blattler and P. J. Farnham, "Cross-talk between site-specific transcription factors and dna methylation states," *Journal of Biological Chemistry*, vol. 288, no. 48, pp. 34287–34294, 2013.

[50] D. Huilgol, P. Venkataramani, S. Nandi, and S. Bhattacharjee, "Transcription factors that govern development and disease: An achilles heel in cancer," *Genes*, vol. 10, no. 10, p. 794, 2019.

[51] L. J. Brown, J. Achinger-Kawecka, N. Portman, S. Clark, C. Stirzaker, and E. Lim, "Epigenetic therapies and biomarkers in breast cancer," *Cancers*, vol. 14, no. 3, p. 474, 2022.

[52] E. Vidal Ocabo, S. Sayols, S. Moran, A. Guillaumet-Adkins, M. P. Schroeder, R. Royo, M. Orozco, M. Gut, I. G. Gut, N. López Bigas, *et al.*, "A dna methylation map of human cancer at single base-pair resolution," *Oncogene. 2017 Oct 5; 36 (40): 5648-5657*, 2017.

[53] B. M. Lindqvist, S. Wingren, P. B. Motlagh, and T. K. Nilsson, "Whole genome dna methylation signature of her2-positive breast cancer," *Epigenetics*, vol. 9, no. 8, pp. 1149–1162, 2014.

[54] I. Mallona, S. Aussó, A. Díez-Villanueva, V. Moreno, and M. A. Peinado, "Dna co-methylation networks outline the structure and remodeling dynamics of colorectal cancer epigenome," *BioRxiv*, p. 428730, 2020.

[55] J. Zhang and K. Huang, "Pan-cancer analysis of frequent dna co-methylation patterns reveals consistent epigenetic landscape changes in multiple cancers," *Bmc Genomics*, vol. 18, no. 1, pp. 1–14, 2017.

[56] N. Xu, Y.-P. Wu, Z.-B. Ke, Y.-C. Liang, H. Cai, W.-T. Su, X. Tao, S.-H. Chen, Q.-S. Zheng, Y. Wei, *et al.*, "Identification of key dna methylation-driven genes in prostate adenocarcinoma: an integrative analysis of tcga methylation data," *Journal of translational medicine*, vol. 17, no. 1, pp. 1–15, 2019.

[57] C. Cheng, R. Alexander, R. Min, J. Leng, K. Y. Yip, J. Rozowsky, K. Yan, X. Dong, S. Djebali, Y. Ruan, *et al.*, "Understanding transcriptional regulation by integrative analysis of transcription factor binding data," *Genome research*, vol. 22, no. 9, pp. 1658–1667, 2012.

[58] S. A. Lambert, A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T. R. Hughes, and M. T. Weirauch, "The human transcription factors," *Cell*, vol. 172, no. 4, pp. 650–665, 2018.

[59] H. E. Pratt, G. R. Andrews, N. Phalke, J. D. Huey, M. J. Purcaro, A. van der Velde, J. E. Moore, and Z. Weng, "Factorbook: an updated catalog of transcription factor motifs and candidate regulatory motif sites," *Nucleic acids research*, vol. 50, no. D1, pp. D141–D149, 2022.

[60] S. Frietze and P. J. Farnham, "Transcription factor effector domains," in *A handbook of transcription factors*, pp. 261–277, Springer, 2011.

[61] D. L. Fulton, S. Sundararajan, G. Badis, T. R. Hughes, W. W. Wasserman, J. C. Roach, and R. Sladek, "Tfcat: the curated catalog of mouse and human transcription factors," *Genome biology*, vol. 10, no. 3, pp. 1–14, 2009.

[62] J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe, "A census of human transcription factors: function, expression and evolution," *Nature Reviews Genetics*, vol. 10, no. 4, pp. 252–263, 2009.

[63] P. F. Johnson and S. L. McKnight, "Eukaryotic transcriptional regulatory proteins," *Annual review of biochemistry*, vol. 58, no. 1, pp. 799–839, 1989.

[64] K. R. Nitta, A. Jolma, Y. Yin, E. Morgunova, T. Kivioja, J. Akhtar, K. Hens, J. Toivonen, B. Deplancke, E. E. Furlong, *et al.*, "Conservation of transcription factor binding specificities across 600 million years of bilateria evolution," *elife*, vol. 4, p. e04837, 2015.

[65] O. Fornes, J. A. Castro-Mondragon, A. Khan, R. Van der Lee, X. Zhang, P. A. Richmond, B. P. Modi, S. Correard, M. Gheorghe, D. Baranašić, *et al.*, "Jaspar 2020: update of the open-access database of transcription factor binding profiles," *Nucleic acids research*, vol. 48, no. D1, pp. D87–D92, 2020.

[66] T. H. Kim, Z. K. Abdullaev, A. D. Smith, K. A. Ching, D. I. Loukinov, R. D. Green, M. Q. Zhang, V. V. Lobanenkov, and B. Ren, "Analysis of the vertebrate insulator protein ctcf-binding sites in the human genome," *Cell*, vol. 128, no. 6, pp. 1231–1245, 2007.

[67] W. J. de Jonge, H. P. Patel, J. V. Meeussen, and T. L. Lenstra, "Following the tracks: how transcription factor binding dynamics control transcription," *Biophysical Journal*, 2022.

[68] H. Chen, S. Chung, and S. Sukumar, "Hoxa5-induced apoptosis in breast cancer cells is mediated by caspases 2 and 8," *Molecular and cellular biology*, vol. 24, no. 2, pp. 924–935, 2004.

[69] D. Tapia-Carrillo, H. Tovar, T. E. Velazquez-Caldelas, and E. Hernandez-Lemus, "Master regulators of signaling pathways: an application to the analysis of gene regulation in breast cancer," *Frontiers in genetics*, p. 1180, 2019.

[70] M. Inoue and K. Horimoto, "Relationship between regulatory pattern of gene expression level and gene function," *PloS one*, vol. 12, no. 5, p. e0177430, 2017.

[71] S. S.-K. Chan and M. Kyba, "What is a master regulator?," *Journal of stem cell research & therapy*, vol. 3, May 2013.

[72] H. Tovar, R. García-Herrera, J. Espinal-Enríquez, and E. Hernández-Lemus, "Transcriptional master regulator analysis in breast cancer genetic networks," *Computational biology and chemistry*, vol. 59, pp. 67–77, 2015.

[73] V. Vinnitsky, "Oncogerminative hypothesis of tumor formation," *Medical hypotheses*, vol. 40, no. 1, pp. 19–27, 1993.

[74] M. Lambert, S. Jambon, S. Depauw, and M.-H. David-Cordonnier, "Targeting transcription factors for cancer treatment," *Molecules*, vol. 23, no. 6, p. 1479, 2018.

[75] M. Chen, Y. Yang, K. Xu, L. Li, J. Huang, and F. Qiu, "Androgen receptor in breast cancer: from bench to bedside," *Frontiers in Endocrinology*, p. 573, 2020.

[76] N. Lara-Castillo, "Estrogen signaling in bone," *Applied Sciences*, vol. 11, no. 10, p. 4439, 2021.

[77] I. Paterni, C. Granchi, J. A. Katzenellenbogen, and F. Minutolo, "Estrogen receptors alpha (erα) and beta (erβ): subtype-selective ligands and clinical potential," *Steroids*, vol. 90, pp. 13–29, 2014.

[78] R. Jia, P. Chai, H. Zhang, and X. Fan, "Novel insights into chromosomal conformations in cancer," *Molecular cancer*, vol. 16, no. 1, p. 173, 2017.

[79] P. K. Singh and M. J. Campbell, "The interactions of microrna and epigenetic modifications in prostate cancer," *Cancers*, vol. 5, no. 3, pp. 998–1019, 2013.

[80] G. Bertoli, C. Cava, and I. Castiglioni, "Micrornas: new biomarkers for diagnosis, prognosis, therapy prediction and therapeutic tools for breast cancer," *Theranostics*, vol. 5, no. 10, p. 1122, 2015.

[81] C. M. Klinge, "Non-coding rnas: long non-coding rnas and micrornas in endocrine-related cancers," *Endocrine-related cancer*, vol. 25, no. 4, pp. R259–R282, 2018.

[82] X. Liu, X. Chen, X. Yu, Y. Tao, A. M. Bode, Z. Dong, and Y. Cao, "Regulation of micrornas by epigenetics and their interplay involved in cancer," *Journal of Experimental & Clinical Cancer Research*, vol. 32, no. 1, p. 96, 2013.

[83] E. O'Day and A. Lal, "Micrornas and their target gene networks in breast cancer," *Breast cancer research*, vol. 12, no. 2, p. 201, 2010.

[84] D. Drago-García, J. Espinal-Enríquez, and E. Hernández-Lemus, "Network analysis of emt and met micro-rna regulation in breast cancer," *Scientific reports*, vol. 7, no. 1, p. 13534, 2017.

[85] C. Baer, R. Claus, L. P. Frenzel, M. Zucknick, Y. J. Park, L. Gu, D. Weichenhan, M. Fischer, C. P. Pallasch, E. Herpel, *et al.*, "Extensive promoter dna hypermethylation and hypomethylation is associated with aberrant microrna expression in chronic lymphocytic leukemia," *Cancer research*, vol. 72, no. 15, pp. 3775–3785, 2012.

[86] R. Rupaimoole, G. A. Calin, G. Lopez-Berestein, and A. K. Sood, "mirna deregulation in cancer cells and the tumor microenvironment," *Cancer discovery*, vol. 6, no. 3, pp. 235–246, 2016.

[87] R. Martienssen and D. Moazed, "Rnai and heterochromatin assembly," *Cold Spring Harbor perspectives in biology*, vol. 7, no. 8, p. a019323, 2015.

[88] T. Hulf, T. Sibbritt, E. D. Wiklund, S. Bert, D. Strbenac, A. L. Statham, M. D. Robinson, and S. J. Clark, "Discovery pipeline for epigenetically deregulated mirnas in cancer: integration of primary mirna transcription," *BMC genomics*, vol. 12, no. 1, p. 54, 2011.

[89] B. Humphries and C. Yang, "The microrna-200 family: small molecules with novel roles in cancer development, progression and therapy," *Oncotarget*, vol. 6, no. 9, p. 6472, 2015.

[90] V. V. Pham, J. Zhang, L. Liu, B. Truong, T. Xu, T. T. Nguyen, J. Li, and T. D. Le, "Identifying mirna-mrna regulatory relationships in breast cancer with invariant causal prediction," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–12, 2019.

[91] Y. Ru, K. J. Kechris, B. Tabakoff, P. Hoffman, R. A. Radcliffe, R. Bowler, S. Mahaffey, S. Rossi, G. A. Calin, L. Bemis, and D. Theodorescu, "The multimir r package and database: integration of microrna-target interactions along with their disease and drug associations.," *Nucleic acids research*, vol. 42, p. e133, 2014.

[92] W. C. Cho, "Oncomirs: the discovery and progress of micrornas in cancers," *Molecular cancer*, vol. 6, no. 1, p. 60, 2007.

[93] Z. Ali Syeda, S. S. S. Langden, C. Munkhzul, M. Lee, and S. J. Song, "Regulatory mechanism of microrna expression in cancer," *International journal of molecular sciences*, vol. 21, no. 5, p. 1723, 2020.

[94] H. Suzuki, S. Takatsuka, H. Akashi, E. Yamamoto, M. Nojima, R. Maruyama, M. Kai, H.-o. Yamano, Y. Sasaki, T. Tokino, *et al.*, "Genome-wide profiling of chromatin signatures reveals epigenetic regulation of microrna genes in colorectal cancer," *Cancer research*, 2011.

[95] B. Pal, Y. Chen, A. Bert, Y. Hu, J. M. Sheridan, T. Beck, W. Shi, K. Satterley, P. Jamieson, G. J. Goodall, *et al.*, "Integration of microrna signatures of distinct mammary epithelial cell types with their gene expression and epigenetic portraits," *Breast Cancer Research*, vol. 17, no. 1, p. 85, 2015.

[96] B. Bose, M. Moravec, and S. Bozdag, "Computing microrna-gene interaction networks in pan-cancer using mirdriver," *Scientific reports*, vol. 12, no. 1, pp. 1–17, 2022.

[97] M. Bersanelli, E. Mosca, D. Remondini, E. Giampieri, C. Sala, G. Castellani, and L. Milanesi, "Methods for the integration of multi-omics data: mathematical aspects," *BMC bioinformatics*, vol. 17, no. 2, p. S15, 2016.

[98] B. Ulfenborg, "Vertical and horizontal integration of multi-omics data with miodin," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–10, 2019.

[99] F. Rohart, B. Gautier, A. Singh, and K.-A. Le Cao, "mixomics: An r package for 'omics feature selection and multiple data integration," *PLoS computational biology*, vol. 13, no. 11, p. e1005752, 2017.

[100] D. Kim, H. Shin, K.-A. Sohn, A. Verma, M. D. Ritchie, and J. H. Kim, "Incorporating interrelationships between different levels of genomic data into cancer clinical outcome prediction," *Methods*, vol. 67, no. 3, pp. 344–353, 2014.

[101] G. Tini, *The influence of the inclusion of biological knowledge in statistical methods to integrate multi-omics data*. PhD thesis, UNIVERSITÀ DEGLI STUDI DI TRENTO Department of Mathematics, 2017.

[102] I. Garali, I. M. Adanyeguh, F. Ichou, V. Perlbarg, A. Seyer, B. Colsch, I. Moszer, V. Guillemot, A. Durr, F. Mochel, *et al.*, "A strategy for multimodal data integration: application to biomarkers identification in spinocerebellar ataxia," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1356–1369, 2018.

[103] E. F. Lock, K. A. Hoadley, J. S. Marron, and A. B. Nobel, "Joint and individual variation explained (jive) for integrated analysis of multiple data types," *The annals of applied statistics*, vol. 7, no. 1, p. 523, 2013.

[104] S. Ciucci, Y. Ge, C. Durán, A. Palladini, V. Jiménez-Jiménez, L. M. Martínez-Sánchez, Y. Wang, S. Sales, A. Shevchenko, S. W. Poser, *et al.*, "Enlightening discriminative network functional modules behind principal component analysis separation in differential-omic science studies," *Scientific reports*, vol. 7, no. 1, pp. 1–24, 2017.

[105] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112. Springer, 2013.

[106] S. Tarazona, L. Balzano-Nogueira, D. Gómez-Cabrero, A. Schmidt, A. Imhof, T. Hankemeier, J. Tegnér, J. A. Westerhuis, and A. Conesa, "Harmonization of quality metrics and power calculation in multi-omic studies," *Nature communications*, vol. 11, no. 1, pp. 1–13, 2020.

[107] E. F. Lock and D. B. Dunson, "Bayesian consensus clustering," *Bioinformatics*, vol. 29, no. 20, pp. 2610–2616, 2013.

[108] D. M. Swanson, T. Lien, H. Bergholtz, T. Sørlie, and A. Frigessi, "A bayesian two-way latent structure model for genomic data integration reveals few pan-genomic cluster subtypes in a breast cancer cohort," *Bioinformatics*, vol. 35, no. 23, pp. 4886–4897, 2019.

[109] M. Paczkowska, J. Barenboim, N. Sintupisut, N. S. Fox, H. Zhu, D. Abd-Rabbo, M. W. Mee, P. C. Boutros, and J. Reimand, "Integrative pathway enrichment analysis of multivariate omics data," *Nature communications*, vol. 11, no. 1, pp. 1–16, 2020.

[110] C. Meng, O. A. Zeleznik, G. G. Thallinger, B. Kuster, A. M. Gholami, and A. C. Culhane, "Dimension reduction techniques for the integrative analysis of multi-omics data," *Briefings in bioinformatics*, vol. 17, no. 4, pp. 628–641, 2016.

[111] M. De Tayrac, S. Lê, M. Aubry, J. Mosser, and F. Husson, "Simultaneous analysis of distinct omics data sets with integration of biological knowledge: Multiple factor analysis approach," *BMC genomics*, vol. 10, no. 1, p. 32, 2009.

[112] C. Meng, A. Basunia, B. Peters, A. M. Gholami, B. Kuster, and A. C. Culhane, "Mogsa: integrative single sample gene-set analysis of multiple omics data," *Molecular & Cellular Proteomics*, vol. 18, no. 8, pp. S153–S168, 2019.

[113] M. J. O'Connell and E. F. Lock, "R. jive for exploration of multi-source molecular data," *Bioinformatics*, vol. 32, no. 18, pp. 2877–2879, 2016.

[114] L. Cantini, P. Zakeri, C. Hernandez, A. Naldi, D. Thieffry, E. Remy, and A. Baudot, "Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer," *Nature communications*, vol. 12, no. 1, pp. 1–12, 2021.

[115] Z. Fan, Y. Zhou, and H. W. Ressom, "Mota: Network-based multi-omic data integration for biomarker discovery," *Metabolites*, vol. 10, no. 4, p. 144, 2020.

[116] A. Tenenhaus, C. Philippe, V. Guillemot, K.-A. Le Cao, J. Grill, and V. Frouin, "Variable selection for generalized canonical correlation analysis," *Biostatistics*, vol. 15, no. 3, pp. 569–583, 2014.

[117] I. González, K.-A. Lê Cao, M. J. Davis, and S. Déjean, "Visualising associations between paired 'omics' data sets," *BioData mining*, vol. 5, no. 1, p. 19, 2012.

[118] A. Singh, C. P. Shannon, B. Gautier, F. Rohart, M. Vacher, S. J. Tebbutt, and K.-A. Lê Cao, "Diablo: an integrative approach for identifying key molecular drivers from multi-omic assays," *Bioinformatics*, 2019.

[119] K.-A. Lê Cao, D. Rossouw, C. Robert-Granié, and P. Besse, "A sparse pls for variable selection when integrating omics data.," *Statistical applications in genetics and molecular biology*, vol. 7, p. Article 35, 2008.

[120] K.-A. Lê Cao, S. Boitard, and P. Besse, "Sparse pls discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems," *BMC bioinformatics*, vol. 12, no. 1, p. 253, 2011.

[121] S. Huang, W. Xu, P. Hu, and T. M. Lakowski, "Integrative analysis reveals subtype-specific regulatory determinants in triple negative breast cancer," *Cancers*, vol. 11, no. 4, p. 507, 2019.

[122] M. Setty, K. Helmy, A. A. Khan, J. Silber, A. Arvey, F. Neezen, P. Agius, J. T. Huse, E. C. Holland, and C. S. Leslie, "Inferring transcriptional and microrna-mediated regulatory programs in glioblastoma," *Molecular systems biology*, vol. 8, no. 1, p. 605, 2012.

[123] W. Li, S. Zhang, C.-C. Liu, and X. J. Zhou, "Identifying multi-layer gene regulatory modules from multi-dimensional genomic data," *Bioinformatics*, vol. 28, no. 19, pp. 2458–2466, 2012.

[124] K. Chappell, K. Manna, C. L. Washam, S. Graw, D. Alkam, M. D. Thompson, M. K. Zafar, L. Hazeslip, C. Randolph, A. Gies, *et al.*, "Multi-omics data integration reveals correlated regulatory features of triple negative breast cancer," *Molecular Omics*, 2021.

[125] A. Rossnerova, A. Izzotti, A. Pulliero, A. Bast, S. Rattan, and P. Rossner, "The molecular mechanisms of adaptive response related to environmental stress," *International Journal of Molecular Sciences*, vol. 21, no. 19, p. 7053, 2020.

[126] K.-A. Sohn, D. Kim, J. Lim, and J. H. Kim, "Relative impact of multi-layered genomic data on gene expression phenotypes in serous ovarian tumors," *BMC systems biology*, vol. 7, no. 6, p. S9, 2013.

[127] G. Lee, L. Bang, S. Y. Kim, D. Kim, and K.-A. Sohn, "Identifying subtype-specific associations between gene expression and dna methylation profiles in breast cancer," *BMC medical genomics*, vol. 10, no. 1, p. 28, 2017.

[128] P. K. Mankoo, R. Shen, N. Schultz, D. A. Levine, and C. Sander, "Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles," *PloS one*, vol. 6, no. 11, p. e24709, 2011.

[129] L. Bravo-Merodio, J. A. Williams, G. V. Gkoutos, and A. Acharjee, "-omics biomarker identification pipeline for translational medicine," *Journal of translational medicine*, vol. 17, no. 1, p. 155, 2019.

[130] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society, Series B*, vol. 67, pp. 301–320, 2005.

[131] A. Kirpich, E. A. Ainsworth, J. M. Wedow, J. R. Newman, G. Michailidis, and L. M. McIntyre, "Variable selection in omics data: A practical evaluation of small sample sizes," *PloS one*, vol. 13, no. 6, p. e0197910, 2018.

[132] S. Pineda, F. X. Real, M. Kogevinas, A. Carrato, S. J. Chanock, N. Malats, and K. Van Steen, "Integration analysis of three omics data using penalized regression methods: an application to bladder cancer," *PLoS genetics*, vol. 11, no. 12, p. e1005689, 2015.

[133] E. C. Neto, J. C. Bare, and A. A. Margolin, "Simulation studies as designed experiments: the comparison of penalized regression models in the "large p, small n" setting," *PloS one*, vol. 9, no. 10, p. e107957, 2014.

[134] J. Liu, G. Liang, K. D. Siegmund, and J. P. Lewinger, "Data integration by multi-tuning parameter elastic net regression," *BMC bioinformatics*, vol. 19, no. 1, p. 369, 2018.

[135] E. I. Vlachavas, J. Bohn, F. Ückert, and S. Nürnberg, "A detailed catalogue of multi-omics methodologies for identification of putative biomarkers and causal molecular networks in translational cancer research," *International journal of molecular sciences*, vol. 22, no. 6, p. 2822, 2021.

[136] J. S. Hawe, F. J. Theis, and M. Heinig, "Inferring interaction networks from multi-omics data," *Frontiers in genetics*, vol. 10, p. 535, 2019.

[137] D. Weighill, R. Burkholz, M. B. Guebila, H. U. Zacharias, J. Quackenbush, and M. Altenbuchinger, "Dragon: determining regulatory associations using graphical models on multi-omic networks," *arXiv preprint arXiv:2104.01690*, 2021.

[138] G. Ciriello, M. L. Miller, B. A. Aksoy, Y. Senbabaoglu, N. Schultz, and C. Sander, "Emerging landscape of oncogenic signatures across human cancers," *Nature genetics*, vol. 45, no. 10, pp. 1127–1133, 2013.

[139] Z. Wang, J. Yin, W. Zhou, J. Bai, Y. Xie, K. Xu, X. Zheng, J. Xiao, L. Zhou, X. Qi, *et al.*, "Complex impact of dna methylation on transcriptional dysregulation across 22 human cancer types," *Nucleic Acids Research*, vol. 48, no. 5, pp. 2287–2302, 2020.

[140] E. Hernández-Lemus, "Systems biology and integrative omics in breast cancer," in *Omics Approaches in Breast Cancer*, pp. 333–352, Springer, 2014.

[141] I. Lee, U. M. Blom, P. I. Wang, J. E. Shim, and E. M. Marcotte, "Prioritizing candidate disease genes by network-based boosting of genome-wide association data," *Genome research*, vol. 21, no. 7, pp. 1109–1121, 2011.

[142] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano, "Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," in *BMC bioinformatics*, vol. 7, p. S7, BioMed Central, 2006.

[143] Z. Mousavian, K. Kavousi, and A. Masoudi-Nejad, "Information theory in systems biology. part i: Gene regulatory and metabolic networks," in *Seminars in cell & developmental biology*, vol. 51, pp. 3–13, Elsevier, 2016.

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

[145] Y. Liu, Y. Liu, R. Huang, W. Song, J. Wang, Z. Xiao, S. Dong, Y. Yang, and X. Yang, "Dependency of the cancer-specific transcriptional regulation circuitry on the promoter dna methylome.," *Cell reports*, vol. 26, pp. 3461–3474.e5, Mar. 2019.

[146] A. Joshi, Y. Van de Peer, and T. Michoel, "Analysis of a gibbs sampler method for model-based clustering of gene expression data," *Bioinformatics*, vol. 24, no. 2, pp. 176–183, 2007.

[147] E. Bonnet, T. Michoel, and Y. Van de Peer, "Prediction of a gene regulatory network linked to prostate cancer from gene expression, microrna and clinical data," *Bioinformatics*, vol. 26, no. 18, pp. i638–i644, 2010.

[148] E. Bonnet, L. Calzone, and T. Michoel, "Integrative multi-omics module network inference with lemon-tree," *PLoS computational biology*, vol. 11, no. 2, p. e1003983, 2015.

[149] H. W. Koh, D. Fermin, C. Vogel, K. P. Choi, R. M. Ewing, and H. Choi, "iomicspass: network-based integration of multiomics data for predictive subnetwork discovery," *NPJ systems biology and applications*, vol. 5, no. 1, pp. 1–10, 2019.

[150] N. Biswas, K. Kumar, S. Bose, R. Bera, and S. Chakrabarti, "Analysis of pan-omics data in human interactome network (apodhin).," *Frontiers in genetics*, vol. 11, p. 589231, 2020.

[151] C. Cava, S. Sabetian, and I. Castiglioni, "Patient-specific network for personalized breast cancer therapy with multi-omics data," *Entropy*, vol. 23, no. 2, p. 225, 2021.

[152] C. Dimitrakopoulos, S. K. Hindupur, L. Häfliger, J. Behr, H. Montazeri, M. N. Hall, and N. Beerenwinkel, "Network-based integration of multi-omics data for prioritizing cancer genes," *Bioinformatics*, vol. 34, no. 14, pp. 2441–2448, 2018.

[153] Y.-X. Chen, H. Chen, Y. Rong, F. Jiang, J.-B. Chen, Y.-Y. Duan, D.-L. Zhu, T.-L. Yang, Z. Dai, S.-S. Dong, *et al.*, "An integrative multi-omics network-based approach identifies key regulators for breast cancer," *Computational and structural biotechnology journal*, vol. 18, pp. 2826–2835, 2020.

[154] L. Shu, Y. Zhao, Z. Kurt, S. G. Byars, T. Tukiainen, J. Kettunen, L. D. Orozco, M. Pellegrini, A. J. Lusis, S. Ripatti, *et al.*, "Mergeomics: multidimensional data integration to identify pathogenic perturbations to biological systems," *BMC genomics*, vol. 17, no. 1, p. 874, 2016.

[155] K. Glass, C. Huttenhower, J. Quackenbush, and G.-C. Yuan, "Passing messages between biological networks to refine predicted interactions," *PloS one*, vol. 8, no. 5, p. e64832, 2013.

[156] M. L. Kuijjer, M. Fagny, A. Marin, J. Quackenbush, and K. Glass, "Puma: Panda using microrna associations," *Bioinformatics*, vol. 36, no. 18, pp. 4765–4773, 2020.

[157] A. R. Sonawane, D. L. DeMeo, J. Quackenbush, and K. Glass, "Constructing gene regulatory networks using epigenetic data," *npj Systems Biology and Applications*, vol. 7, no. 1, pp. 1–13, 2021.

[158] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg, "Similarity network fusion for aggregating data types on a genomic scale," *Nature methods*, vol. 11, no. 3, p. 333, 2014.

[159] N. Rappoport, R. Safra, and R. Shamir, "Monet: multi-omic module discovery by omic selection," *PLoS computational biology*, vol. 16, no. 9, p. e1008182, 2020.

[160] D. L. Gibbs, L. Gralinski, R. S. Baric, and S. K. McWeeney, "Multi-omic network signatures of disease," *Frontiers in genetics*, vol. 4, p. 309, 2014.

[161] M. Arif, C. Zhang, X. Li, C. Güngör, B. Çakmak, M. Arslantürk, A. Tebani, B. Özcan, O. Subaş, W. Zhou, *et al.*, "inetmodels 2.0: an interactive visualization and database of multi-omics data," *Nucleic acids research*, vol. 49, no. W1, pp. W271–W276, 2021.

[162] M. Ben Guebila, C. M. Lopes-Ramos, D. Weighill, A. R. Sonawane, R. Burkholz, B. Shamsaei, J. Platig, K. Glass, M. L. Kuijjer, and J. Quackenbush, "Grand: a database of gene regulatory network models across human conditions," *Nucleic acids research*, vol. 50, no. D1, pp. D610–D621, 2022.

[163] M. J. Aryee, A. E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A. P. Feinberg, K. D. Hansen, and R. A. Irizarry, "Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays.," *Bioinformatics (Oxford, England)*, vol. 30, pp. 1363–1369, May 2014.

[164] S. Tarazona, P. Furió-Tarí, D. Turrà, A. D. Pietro, M. J. Nueda, A. Ferrer, and A. Conesa, "Data quality aware analysis of differential expression in rna-seq with noiseq r/bioc package.," *Nucleic acids research*, vol. 43, p. e140, Dec. 2015.

[165] S. Tam, M.-S. Tsao, and J. D. McPherson, "Optimization of mirna-seq data preprocessing," *Briefings in bioinformatics*, vol. 16, no. 6, pp. 950–963, 2015.

[166] C. Soneson and M. Delorenzi, "A comparison of methods for differential expression analysis of rna-seq data," *BMC bioinformatics*, vol. 14, no. 1, p. 91, 2013.

[167] A. Chu, G. Robertson, D. Brooks, A. J. Mungall, I. Birol, R. Coope, Y. Ma, S. Jones, and M. A. Marra, "Large-scale profiling of microrna for the cancer genome atlas," *Nucleic acids research*, vol. 44, no. 1, pp. e3–e3, 2016.

[168] S. Dedeurwaerder, M. Defrance, E. Calonne, H. Denis, C. Sotiriou, and F. Fuks, "Evaluation of the infinium methylation 450k technology.," *Epigenomics*, vol. 3, pp. 771–784, Dec. 2011.

[169] P. Du, X. Zhang, C.-C. Huang, N. Jafari, W. A. Kibbe, L. Hou, and S. M. Lin, "Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis.," *BMC bioinformatics*, vol. 11, p. 587, Nov. 2010.

[170] C. W. Law, M. Alhamdoosh, S. Su, X. Dong, L. Tian, G. K. Smyth, and M. E. Ritchie, "Rna-seq analysis is easy as 1-2-3 with limma, glimma and edger," *F1000Research*, vol. 5, 2016.

[171] D. J. McCarthy and G. K. Smyth, "Testing significance relative to a fold-change threshold is a treat," *Bioinformatics*, vol. 25, no. 6, pp. 765–771, 2009.

[172] U. D. Bioinformatics, "2018 june rna seq workshop." `https://ucdavis-bioinformatics-training.github.io/2018-June-RNA-Seq-Workshop/thursday/DE.html`, 2018.

[173] J. Maksimovic, J. A. Gagnon-Bartsch, T. P. Speed, and A. Oshlack, "Removing unwanted variation in a differential methylation analysis of illumina humanmethylation450 array data," *Nucleic acids research*, vol. 43, no. 16, pp. e106–e106, 2015.

[174] B. Phipson, J. Maksimovic, and A. Oshlack, "missmethyl: an r package for analyzing data from illumina's humanmethylation450 platform," *Bioinformatics*, vol. 32, no. 2, pp. 286–288, 2016.

[175] S. Neph, A. B. Stergachis, A. Reynolds, R. Sandstrom, E. Borenstein, and J. A. Stamatoyannopoulos, "Circuitry and dynamics of human transcription factor regulatory networks," *Cell*, vol. 150, no. 6, pp. 1274–1286, 2012.

[176] D. Marbach, D. Lamparter, G. Quon, M. Kellis, Z. Kutalik, and S. Bergmann, "Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases.," *Nature methods*, vol. 13, pp. 366–370, Apr. 2016.

[177] G. de Anda-Jáuregui, T. E. Velázquez-Caldelas, J. Espinal-Enríquez, and E. Hernández-Lemus, "Transcriptional network architecture of breast cancer molecular subtypes," *Frontiers in physiology*, vol. 7, p. 568, 2016.

[178] S. Ochoa, G. de Anda-Jáuregui, and E. Hernández-Lemus, "Multi-omic regulation of the pam50 gene signature in breast cancer molecular subtypes," *Frontiers in oncology*, vol. 10, p. 845, 2020.

[179] S. Ochoa, G. de Anda-Jáuregui, and E. Hernández-Lemus, "An information theoretical multilayer network approach to breast cancer transcriptional regulation," *Frontiers in genetics*, vol. 12, p. 617512, 2021.

[180] M.-L. Si, S. Zhu, H. Wu, Z. Lu, F. Wu, and Y.-Y. Mo, "mir-21-mediated tumor growth.," *Oncogene*, vol. 26, pp. 2799–2803, Apr. 2007.

[181] P. Bhat-Nakshatri, G. Wang, N. R. Collins, M. J. Thomson, T. R. Geistlinger, J. S. Carroll, M. Brown, S. Hammond, E. F. Srour, Y. Liu, and H. Nakshatri, "Estradiol-regulated micrornas control estradiol response in breast cancer cells.," *Nucleic acids research*, vol. 37, pp. 4850–4861, Aug. 2009.

[182] A. Barker, K. M. Giles, M. R. Epis, P. M. Zhang, F. Kalinowski, and P. J. Leedman, "Regulation of erbb receptor signalling in cancer cells by microrna.," *Current opinion in pharmacology*, vol. 10, pp. 655–661, Dec. 2010.

[183] T.-H. Huang, F. Wu, G. B. Loeb, R. Hsu, A. Heidersbach, A. Brincat, D. Horiuchi, R. J. Lebbink, Y.-Y. Mo, A. Goga, and M. T. McManus, "Up-regulation of mir-21 by her2/neu signaling promotes cell invasion.," *The Journal of biological chemistry*, vol. 284, pp. 18515–18524, July 2009.

[184] D. Gaidatzis, E. van Nimwegen, J. Hausser, and M. Zavolan, "Inference of mirna targets using evolutionary conservation and pathway analysis.," *BMC bioinformatics*, vol. 8, p. 69, Mar. 2007.

[185] M. Maragkakis, M. Reczko, V. A. Simossis, P. Alexiou, G. L. Papadopoulos, T. Dalamagas, G. Giannopoulos, G. Goumas, E. Koukis, K. Kourtis, T. Vergoulis, N. Koziris, T. Sellis, P. Tsanakas, and A. G. Hatzigeorgiou, "Diana-microt web server: elucidating microrna functions through target prediction.," *Nucleic acids research*, vol. 37, pp. W273–W276, July 2009.

[186] S. Kishore, L. Jaskiewicz, L. Burger, J. Hausser, M. Khorshid, and M. Zavolan, "A quantitative analysis of clip methods for identifying binding sites of rna-binding proteins.," *Nature methods*, vol. 8, pp. 559–564, May 2011.

[187] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks, "Microrna targets in drosophila.," *Genome biology*, vol. 5, p. R1, 2003.

[188] A. H. Lund, "mir-10 in development and cancer.," *Cell death and differentiation*, vol. 17, pp. 209–214, Feb. 2010.

[189] S. E. Wang and R.-J. Lin, "Microrna and her2-overexpressing cancer.," *MicroRNA (Shariqah, United Arab Emirates)*, vol. 2, pp. 137–147, 2013.

[190] F. Biagioni, N. Bossel Ben-Moshe, G. Fontemaggi, Y. Yarden, E. Domany, and G. Blandino, "The locus of microrna-10b: a critical target for breast cancer insurgence and dissemination.," *Cell cycle (Georgetown, Tex.)*, vol. 12, pp. 2371–2375, Aug. 2013.

[191] G. de Anda-Jáuregui, J. Espinal-Enríquez, D. Drago-García, and E. Hernández-Lemus, "Nonredundant, highly connected micrornas control functionality in breast cancer networks.," *International journal of genomics*, vol. 2018, p. 9585383, 2018.

[192] A. Helwak, G. Kudla, T. Dudnakova, and D. Tollervey, "Mapping the human mirna interactome by clash reveals frequent noncanonical binding.," *Cell*, vol. 153, pp. 654–665, Apr. 2013.

[193] M. Yu, A. Bardia, B. S. Wittner, S. L. Stott, M. E. Smas, D. T. Ting, S. J. Isakoff, J. C. Ciciliano, M. N. Wells, A. M. Shah, K. F. Concannon, M. C. Donaldson, L. V. Sequist, E. Brachtel, D. Sgroi, J. Baselga, S. Ramaswamy, M. Toner, D. A. Haber, and S. Maheswaran, "Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition.," *Science (New York, N.Y.)*, vol. 339, pp. 580–584, Feb. 2013.

[194] J. Jeffery, D. Sinha, S. Srihari, M. Kalimutho, and K. K. Khanna, "Beyond cytokinesis: the emerging roles of cep55 in tumorigenesis.," *Oncogene*, vol. 35, pp. 683–690, Feb. 2016.

[195] S. Sankar, J. M. Tanner, R. Bell, A. Chaturvedi, R. L. Randall, M. C. Beckerle, and S. L. Lessnick, "A novel role for keratin 17 in coordinating oncogenic transformation and cellular adhesion in ewing sarcoma.," *Molecular and cellular biology*, vol. 33, pp. 4448–4460, Nov. 2013.

[196] L. Vrba and B. W. Futscher, "Dna methylation changes in biomarker loci occur early in cancer progression.," *F1000Research*, vol. 8, p. 2106, 2019.

[197] W. L. Cai, C. B. Greer, J. F. Chen, A. Arnal-Estapé, J. Cao, Q. Yan, and D. X. Nguyen, "Specific chromatin landscapes and transcription factors couple breast cancer subtype with metastatic relapse to lung or brain.," *BMC medical genomics*, vol. 13, p. 33, Mar. 2020.

[198] R. L. Tomlinson, E. B. Abreu, T. Ziegler, H. Ly, C. M. Counter, R. M. Terns, and M. P. Terns, "Telomerase reverse transcriptase is required for the localization of telomerase rna to cajal bodies and telomeres in human cancer cells.," *Molecular biology of the cell*, vol. 19, pp. 3793–3800, Sept. 2008.

[199] A. Godoy-Ortiz, A. Sanchez-Muñoz, M. R. Chica Parrado, M. Álvarez, N. Ribelles, A. Rueda Dominguez, and E. Alba, "Deciphering her2 breast cancer disease: Biological and clinical implications.," *Frontiers in oncology*, vol. 9, p. 1124, 2019.

[200] E. Hernández-Lemus, "On a Class of Tensor Markov Fields," *Entropy*, vol. 22, no. 4, p. 451, 2020.

[201] N. Y. Fu, A. C. Rios, B. Pal, R. Soetanto, A. T. Lun, K. Liu, T. Beck, S. A. Best, F. Vaillant, P. Bouillet, *et al.*, "Egf-mediated induction of mcl-1 at the switch to lactation is essential for alveolar cell survival," *Nature Cell Biology*, vol. 17, no. 4, pp. 365–375, 2015.

[202] M. Farrugia, S. Sharma, C. Lin, S. McLaughlin, D. Vanderbilt, A. Ammer, M. Salkeni, P. Stoilov, Y. Agazie, C. Creighton, *et al.*, "Regulation of anti-apoptotic signaling by kruppel-like factors 4 and 5 mediates lapatinib resistance in breast cancer," *Cell death & disease*, vol. 6, no. 3, pp. e1699–e1699, 2015.

[203] H.-l. Piao, Y. Yuan, M. Wang, Y. Sun, H. Liang, and L. Ma, "α-catenin acts as a tumour suppressor in e-cadherin-negative basal-like breast cancer by inhibiting nf-κb signalling," *Nature cell biology*, vol. 16, no. 3, pp. 245–254, 2014.

[204] Y.-K. Liang, H.-Y. Lin, X.-W. Dou, M. Chen, X.-L. Wei, Y.-Q. Zhang, Y. Wu, C.-F. Chen, J.-W. Bai, Y.-S. Xiao, *et al.*, "Mir-221/222 promote epithelial-mesenchymal transition by targeting notch3 in breast cancer cell lines," *NPJ breast cancer*, vol. 4, no. 1, pp. 1–9, 2018.

[205] M. B. Bechmann, A. V. Brydholm, V. L. Codony, J. Kim, and R. Villadsen, "Heterogeneity of ceacam5 in breast cancer," *Oncotarget*, vol. 11, no. 43, p. 3886, 2020.

[206] M. Messaoudene, T. Mourikis, J. Michels, Y. Fu, M. Bonvalet, M. Lacroix-Trikki, B. Routy, A. Fluckiger, S. Rusakiewicz, M. Roberti, *et al.*, "T-cell bispecific antibodies in node-positive breast cancer: novel therapeutic avenue for mhc class i loss variants," *Annals of Oncology*, vol. 30, no. 6, pp. 934–944, 2019.

[207] B. Corominas-Faja, E. Cuyàs, J. Gumuzio, J. Bosch-Barrera, O. Leis, Á. G. Martin, and J. A. Menendez, "Chemical inhibition of acetyl-coa carboxylase suppresses self-renewal growth of cancer stem cells," *Oncotarget*, vol. 5, no. 18, p. 8306, 2014.

[208] D. M. Schulz, C. Bollner, G. Thomas, M. Atkinson, I. Esposito, H. Hofler, and M. Aubele, "Identification of differentially expressed proteins in triple-negative breast carcinomas using dige and mass spectrometry," *Journal of proteome research*, vol. 8, no. 7, pp. 3430–3438, 2009.

[209] Y. Li, X. Zhao, Q. Liu, and Y. Liu, "Bioinformatics reveal macrophages marker genes signature in breast cancer to predict prognosis," *Annals of medicine*, vol. 53, no. 1, pp. 1020–1032, 2021.

[210] C. K. Thammaiah and S. Jayaram, "Role of let-7 family microrna in breast cancer," *Non-coding RNA research*, vol. 1, no. 1, pp. 77–82, 2016.

[211] E. Hannezo, C. L. Scheele, M. Moad, N. Drogo, R. Heer, R. V. Sampogna, J. Van Rheenen, and B. D. Simons, "A unifying theory of branching morphogenesis," *Cell*, vol. 171, no. 1, pp. 242–255, 2017.

[212] V. Vinnitsky, "Oncogerminative hypothesis of tumor formation," *Medical hypotheses*, vol. 40, no. 1, pp. 19–27, 1993.

[213] C. Corrado and S. Fontana, "Hypoxia and hif signaling: One axis with divergent effects," *International Journal of Molecular Sciences*, vol. 21, no. 16, p. 5611, 2020.

[214] E. C. de Heer, M. Jalving, A. L. Harris, *et al.*, "Hifs, angiogenesis, and metabolism: elusive enemies in breast cancer," *The Journal of clinical investigation*, vol. 130, no. 10, pp. 5074–5087, 2020.

[215] N. T. Moldogazieva, I. M. Mokhosoev, and A. A. Terentiev, "Metabolic heterogeneity of cancer cells: an interplay between hif-1, gluts, and ampk," *Cancers*, vol. 12, no. 4, p. 862, 2020.

[216] I. Matic, M. van Hagen, J. Schimmel, B. Macek, S. C. Ogg, M. H. Tatham, R. T. Hay, A. I. Lamond, M. Mann, and A. C. Vertegaal, "In vivo identification of human small ubiquitin-like modifier polymerization sites by high accuracy mass spectrometry and an in vitro to in vivo strategy," *Molecular & cellular proteomics*, vol. 7, no. 1, pp. 132–144, 2008.

[217] S. M. Pupa, F. Ligorio, V. Cancila, A. Franceschini, C. Tripodo, C. Vernieri, and L. Castagnoli, "Her2 signaling and breast cancer stem cells: The bridge behind her2-positive breast cancer aggressiveness and therapy refractoriness," *Cancers*, vol. 13, no. 19, p. 4778, 2021.

[218] M. Weigel Muñoz, G. Carvajal, L. Curci, S. N. Gonzalez, and P. S. Cuasnicu, "Relevance of crisp proteins for epididymal physiology, fertilization, and fertility," *Andrology*, vol. 7, no. 5, pp. 610–617, 2019.

[219] B. R. Pathak, A. A. Breed, S. Apte, K. Acharya, and S. D. Mahale, "Cysteine-rich secretory protein 3 plays a role in prostate cancer cell invasion and affects expression of psa and anxa1," *Molecular and cellular biochemistry*, vol. 411, no. 1, pp. 11–21, 2016.

[220] Y. Li, Y. Wang, Y. Wen, T. Zhang, X. Wang, C. Jiang, R. Zheng, F. Zhou, D. Chen, Y. Yang, *et al.*, "Whole-exome sequencing of a cohort of infertile men reveals novel causative genes in teratozoospermia that are chiefly related to sperm head defects," *Human Reproduction*, vol. 37, no. 1, pp. 152–177, 2022.

[221] T. Nakazawa, I. Nakano, M. Sato, T. Nakamura, M. Tamai, and N. Mori, "Comparative expression profiles of trk receptors and shc-related phosphotyrosine adapters during retinal development: potential roles of n-shc/shcc in brain-derived neurotrophic factor signal transduction and modulation," *Journal of neuroscience research*, vol. 68, no. 6, pp. 668–680, 2002.

[222] E. Fernández-Centeno, G. de Ojeda, J. M. Rojo, and P. Portolés, "Crry/p65, a membrane complement regulatory protein, has costimulatory properties on mouse t cells," *The Journal of Immunology*, vol. 164, no. 9, pp. 4533–4542, 2000.

[223] C. He, M. Imai, H. Song, R. J. Quigg, and S. Tomlinson, "Complement inhibitors targeted to the proximal tubule prevent injury in experimental nephrotic syndrome and demonstrate a key role for c5b-9," *The Journal of Immunology*, vol. 174, no. 9, pp. 5750–5757, 2005.

[224] J. Li, Y.-H. Xu, Y. Lu, X.-P. Ma, P. Chen, S.-W. Luo, Z.-G. Jia, Y. Liu, and Y. Guo, "Identifying differentially expressed genes and small molecule drugs for prostate cancer by a bioinformatics strategy," *Asian Pacific journal of cancer prevention*, vol. 14, no. 9, pp. 5281–5286, 2013.

[225] R. O. Bahado-Singh, S. Vishweswaraiah, B. Aydas, A. Yilmaz, R. P. Metpally, D. J. Carey, R. C. Crist, W. H. Berrettini, G. D. Wilson, K. Imam, *et al.*, "Artificial intelligence and leukocyte epigenomics: Evaluation and prediction of late-onset alzheimer's disease," *PloS one*, vol. 16, no. 3, p. e0248375, 2021.

[226] E. Arai, M. Gotoh, Y. Tian, H. Sakamoto, M. Ono, A. Matsuda, Y. Takahashi, S. Miyata, H. Totsuka, S. Chiku, *et al.*, "Alterations of the spindle checkpoint pathway in clinicopathologically aggressive c p g island methylator phenotype clear cell renal cell carcinomas," *International journal of cancer*, vol. 137, no. 11, pp. 2589–2606, 2015.

[227] M. Jafarzadeh, B. M. Soltani, M. Soleimani, and S. Hosseinkhani, "Epigenetically silenced linc02381 functions as a tumor suppressor by regulating pi3k-akt signaling pathway," *Biochimie*, vol. 171, pp. 63–71, 2020.

[228] K. L. Wright, J. R. Adams, J. C. Liu, A. J. Loch, R. G. Wong, C. E. Jo, L. A. Beck, D. R. Santhanam, L. Weiss, X. Mei, *et al.*, "Ras signaling is a key determinant for metastatic dissemination and poor survival of luminal breast cancer patients," *Cancer research*, vol. 75, no. 22, pp. 4960–4972, 2015.

[229] Z. M. Ramdzan, C. Vadnais, R. Pal, G. Vandal, C. Cadieux, L. Leduy, S. Davoudi, L. Hulea, L. Yao, A. N. Karnezis, *et al.*, "Ras transformation requires cux1-dependent repair of oxidative dna damage," *PLoS biology*, vol. 12, no. 3, p. e1001807, 2014.

[230] D. Hollander, M. Donyo, N. Atias, K. Mekahel, Z. Melamed, S. Yannai, G. Lev-Maor, A. Shilo, S. Schwartz, I. Barshack, *et al.*, "A network-based analysis of colon cancer splicing changes reveals a tumorigenesis-favoring regulatory pathway emanating from elk1," *Genome research*, vol. 26, no. 4, pp. 541–553, 2016.

[231] S. Xin, W. Fang, J. Li, D. Li, C. Wang, Q. Huang, M. Huang, W. Zhuang, X. Wang, and L. Chen, "Impact of stat1 polymorphisms on crizotinib-induced hepatotoxicity in alk-positive non-small cell lung cancer patients," *Journal of Cancer Research and Clinical Oncology*, vol. 147, no. 3, pp. 725–737, 2021.

[232] S. K. Patra, "Ras regulation of dna-methylation and cancer," *Experimental cell research*, vol. 314, no. 6, pp. 1193–1201, 2008.

[233] X. Xu, M. Zhang, F. Xu, and S. Jiang, "Wnt signaling in breast cancer: biological mechanisms, challenges and opportunities," *Molecular cancer*, vol. 19, no. 1, pp. 1–35, 2020.

[234] E. Sumi, N. Iehara, H. Akiyama, T. Matsubara, A. Mima, H. Kanamori, A. Fukatsu, D. J. Salant, T. Kita, H. Arai, *et al.*, "Sry-related hmg box 9 regulates the expression of col4a2 through transactivating its enhancer element in mesangial cells," *The American journal of pathology*, vol. 170, no. 6, pp. 1854–1864, 2007.

[235] Z. Tan, B. Niu, K. Y. Tsang, I. G. Melhado, S. Ohba, X. He, Y. Huang, C. Wang, A. P. McMahon, R. Jauch, *et al.*, "Synergistic co-regulation and competition by a sox9-gli-foxa phasic transcriptional network coordinate chondrocyte differentiation transitions," *PLoS genetics*, vol. 14, no. 4, p. e1007346, 2018.

[236] A. Sugiyama, M. Okada, and H. Yamawaki, "Canstatin suppresses isoproterenol-induced cardiac hypertrophy through inhibition of calcineurin/nuclear factor of activated t-cells pathway in rats," *European Journal of Pharmacology*, vol. 871, p. 172849, 2020.

[237] D. F. Bueno, D. Y. Sunaga, G. S. Kobayashi, M. Aguena, C. E. Raposo-Amaral, C. Masotti, L. A. Cruz, P. L. Pearson, and M. R. Passos-Bueno, "Human stem cell cultures from cleft lip/palate patients show enrichment of transcripts involved in extracellular matrix modeling by comparison to controls," *Stem cell reviews and reports*, vol. 7, no. 2, pp. 446–457, 2011.

[238] J. Liu, X. Liu, G. Zhou, R. Xiao, and Y. Cao, "Conditioned medium from chondrocyte/scaffold constructs induced chondrogenic differentiation of bone marrow stromal cells," *The Anatomical Record: Advances in Integrative Anatomy and Evolutionary Biology*, vol. 295, no. 7, pp. 1109–1116, 2012.

[239] J. Wu, C. Wang, X. Miao, Y. Wu, J. Yuan, M. Ding, J. Li, and Z. Shi, "Age-related insulin-like growth factor binding protein-4 overexpression inhibits osteogenic differentiation of rat mesenchymal stem cells," *Cellular Physiology and Biochemistry*, vol. 42, no. 2, pp. 640–650, 2017.

[240] A. A. Arraf, R. Yelin, I. Reshef, J. Jadon, M. Abboud, M. Zaher, J. Schneider, F. K. Vladimirov, and T. M. Schultheiss, "Hedgehog signaling regulates epithelial morphogenesis to position the ventral embryonic midline," *Developmental Cell*, vol. 53, no. 5, pp. 589–602, 2020.

[241] J. R. Cerhan, S. I. Berndt, J. Vijai, H. Ghesquières, J. McKay, S. S. Wang, Z. Wang, M. Yeager, L. Conde, P. I. De Bakker, *et al.*, "Genome-wide association study identifies multiple susceptibility loci for diffuse large b cell lymphoma," *Nature genetics*, vol. 46, no. 11, pp. 1233–1238, 2014.

[242] C. C. Hazelett and C. Yeaman, "Sec5 and exo84 mediate distinct aspects of rala-dependent cell polarization," *PLoS One*, vol. 7, no. 6, p. e39602, 2012.

[243] A. D'Aloia, G. Berruti, B. Costa, C. Schiller, R. Ambrosini, V. Pastori, E. Martegani, and M. Ceriani, "Ralgps2 is involved in tunneling nanotubes formation in 5637 bladder cancer cells," *Experimental Cell Research*, vol. 362, no. 2, pp. 349–361, 2018.

[244] A. Bieluszewska, M. Weglewska, T. Bieluszewski, K. Lesniewicz, and E. Poreba, "Pka-binding domain of akap 8 is essential for direct interaction with dpy 30 protein," *The FEBS journal*, vol. 285, no. 5, pp. 947–964, 2018.

[245] N. R. Rose and R. J. Klose, "Understanding the relationship between dna methylation and histone lysine methylation," *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, vol. 1839, no. 12, pp. 1362–1372, 2014.

[246] A. Astolfi, M. Fiore, F. Melchionda, V. Indio, S. N. Bertuccio, and A. Pession, "Bcor involvement in cancer," *Epigenomics*, vol. 11, no. 7, pp. 835–855, 2019.

[247] R. S. Grand, L. Burger, C. Gräwe, A. K. Michael, L. Isbel, D. Hess, L. Hoerner, V. Iesmantavicius, S. Durdu, M. Pregnolato, *et al.*, "Banp opens chromatin and activates cpg-island-regulated genes," *Nature*, vol. 596, no. 7870, pp. 133–137, 2021.

[248] M. Shamay, M. Greenway, G. Liao, R. F. Ambinder, and S. D. Hayward, "De novo dna methyltransferase dnmt3b interacts with nedd8-modified proteins," *Journal of Biological Chemistry*, vol. 285, no. 47, pp. 36377–36386, 2010.

[249] H. Ke, Y. Wu, R. Wang, and X. Wu, "Creation of a prognostic risk prediction model for lung adenocarcinoma based on gene expression, methylation, and clinical characteristics," *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, vol. 26, pp. e925833–1, 2020.

[250] A. G. Izquierdo, H. Boughanem, A. Diaz-Lagares, I. Arranz-Salas, M. Esteller, F. J. Tinahones, F. F. Casanueva, M. Macias-Gonzalez, and A. B. Crujeiras, "Dna methylome in visceral adipose tissue can discriminate patients with and without colorectal cancer," *Epigenetics*, pp. 1–12, 2021.

[251] S. Zhao, M. S. Geybels, A. Leonardson, R. Rubicz, S. Kolb, Q. Yan, B. Klotzle, M. Bibikova, A. Hurtado-Coll, D. Troyer, *et al.*, "Epigenome-wide tumor dna methylation profiling identifies novel prognostic biomarkers of metastatic-lethal progression in men diagnosed with clinically localized prostate cancerdna methylation biomarkers and prostate cancer prognosis," *Clinical Cancer Research*, vol. 23, no. 1, pp. 311–319, 2017.

[252] P. Zhang, X. Wen, F. Gu, X. Deng, J. Li, J. Dong, J. Jiao, and Y. Tian, "Methylation profiling of serum dna from hepatocellular carcinoma patients using an infinium human methylation 450 beadchip," *Hepatology international*, vol. 7, no. 3, pp. 893–900, 2013.

[253] Y. Furukawa-Hibi, T. Nagai, J. Yun, and K. Yamada, "Stress increases dna methylation of the neuronal pas domain 4 (npas4) gene," *Neuroreport*, vol. 26, no. 14, pp. 827–832, 2015.

[254] Y. Luo, F. Sun, X. Peng, D. Dong, W. Ou, Y. Xie, and Y. Luo, "Integrated bioinformatics analysis to identify abnormal methylated differentially expressed genes for predicting prognosis of human colon cancer," *International Journal of General Medicine*, vol. 14, p. 4745, 2021.

[255] V. V. Strelnikov, E. B. Kuznetsova, A. S. Tanas, V. V. Rudenko, A. I. Kalinkin, E. V. Poddubskaya, T. V. Kekeeva, G. G. Chesnokova, I. D. Trotsenko, S. S. Larin, *et al.*, "Abnormal promoter dna hypermethylation of the integrin, nidogen, and dystroglycan genes in breast cancer," *Scientific reports*, vol. 11, no. 1, pp. 1–14, 2021.

[256] M. Privat, J. Rudewicz, N. Sonnier, C. Tamisier, F. Ponelle-Chachuat, and Y.-J. Bignon, "Anti-oxydation and cell migration genes are identified as potential therapeutic targets in basal-like and brca1 mutated breast cancer cell lines," *International journal of medical sciences*, vol. 15, no. 1, p. 46, 2018.

[257] B. Xiong, X. Lei, L. Zhang, and J. Fu, "mir-103 regulates triple negative breast cancer cells migration and invasion through targeting olfactomedin 4," *Biomedicine & Pharmacotherapy*, vol. 89, pp. 1401–1408, 2017.

[258] R. Hernández-de Diego, S. Tarazona, C. Martínez-Mira, L. Balzano-Nogueira, P. Furió-Tarí, G. J. Pappas, and A. Conesa, "Paintomics 3: a web resource for the pathway analysis and visualization of multi-omics data.," *Nucleic acids research*, vol. 46, pp. W503–W509, July 2018.

[259] M. L. Kuijjer, M. G. Tung, G. Yuan, J. Quackenbush, and K. Glass, "Estimating sample-specific regulatory networks," *Iscience*, vol. 14, pp. 226–240, 2019.

[260] S. V. Vasaikar, P. Straub, J. Wang, and B. Zhang, "Linkedomics: analyzing multi-omics data within and across 32 cancer types," *Nucleic acids research*, vol. 46, no. D1, pp. D956–D963, 2018.