UNIVERSIDAD DE CHILE

FACULTAD DE CIENCIAS QUÍMICAS Y FARMACÉUTICAS

REDES DE INTERACCIÓN MOLECULAR COMO HERRAMIENTA PARA MEJORAR EL CÁLCULO DE PUNTAJES DE RIESGO POLIGÉNICO Y SU APLICACIÓN EN CÁNCER DE MAMA

Tesis presentada a la Universidad de Chile para optar al grado de Magister en Bioquímica área de Especialización en Clínica y Memoria para optar al Título de Bioquímico por:

ADOLFO ALEXIS ROJAS HIDALGO

Director de Tesis: Dr. Vinicius Maracaja-Coutinho

Codirector de Tesis: Dr. Michel Satya Naslavsky

Santiago, CHILE

Octubre 2022

UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS QUÍMICAS Y FARMACÉUTICAS

INFORME DE APROBACIÓN DE TESIS DE MAGÍSTER

Se informa a la Dirección de la Escuela de Postgrado de la Facultad de Ciencias Químicas y Farmacéuticas que la Tesis de Magíster y Memoria de Título presentada por el candidato

ADOLFO ALEXIS ROJAS HIDALGO

Ha sido aprobada por la Comisión de Evaluadora de Tesis como requisito para optar al grado de Magíster en Bioquímica, Área de Especialización Clínica y Título de Bioquímica, en el examen público rendido el día _____

Director de Tesis:
Dr. Vinicius Maracaja Coutinho _____

Co-director de Tesis:
Michel Satya Naslavsky _____

Comisión Evaluadora de Tesis:

Dra. Carmen Romero (Preside) _____

Dra. Katherine Marcelain _____

Dr. Alberto Martin _____

TABLA DE CONTENIDO

ABREVIATURAS

AUC : *Area under the ROC curve*

ceRNA : Ácido ribonucleico de competencia endógena

CGEMS : *Cancer Genetic Markers of Susceptibility*

CM, BC : Cáncer de mama, *Breast Cancer*

CPM : Cuentas por millón

dbGaP : *Database of Genotypes and Phenotypes*

DNA : Ácido desoxirribonucleico

e.g. : *Exempli gratia*, dado como ejemplo

FDR : *False discovery rate*

GTEx : *Genotype-Tissue Expression*

GSEA : *Gene set enrichment analysis*

GWAS : Estudios de asociación de genoma completo

Her2 : Receptor 2 del factor de crecimiento epidérmico humano

ER : Receptor de estrógeno

lncRNA : Ácido ribonucleico no codificante largo

Log2FC : Logaritmo base 2 de la veces de cambio (Log2 *Fold Change*)

MAF : *Minor allele frequency*

miRNA : Micro ácido ribonucleico

ncRNA : Ácido ribonucleico no codificante

NES : *Normalized enrichment score*

OR : *Odds Ratio*

PCA : *Principal component analysis*

PR : Receptor de progesterona

PRS : Puntuación de riesgo poligénico (polygenic risk score)

RNA : Ácido ribonucleico

ROC : *Receiver Operating Characteristic*

SNP : *Single nucleotide polymorphism*

TCGA : *The cancer genome atlas*

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

AGRADECIMIENTOS

Quiero agradecer a las fuentes de financiamiento de mis directores de Tesis. A FONDECYT Regular (1211731) y FONDAP (15130011) de la Agencia Nacional de Investigación y Desarrollo (ANID); Concurso Enlace Fondecyt 2020 (I0230/2020); y finalmente, al Concurso Apoyo a la Infraestructura para la Investigación 2019 (INFRA-021/01/2019) de la Vicerrectoría de Investigación y Desarrollo (VID) de la Universidad de Chile, otorgados al Dr. Vinicius Maracaja. También a los programas encargados de financiar al Dr. Michel Naslavsky coordinados por los profesores Mayana Zatz y Yeda Duarte: University of São Paulo (FAPESP CEPID 2013/08028-1, SABE 2014/50649-6, INCT 2014//50931-3, y Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq INCT 465355/2014-5).

A los integrantes del laboratorio de Bioinformática integrativa del Centro de Estudios Avanzados de Enfermedades Crónicas (ACCDiS), asi como a la Facultad de ciencias Químicas y Farmacéuticas de la Universidad de Chile, que proporcionaron recursos intelectuales e infraestructurales necesarios, para el desarrollo de esta tesis.

De manera personal también agradecer a mi familia y a cada persona que aporto de manera positiva, tanto en mi vida personal como profesional.

RESUMEN

El cáncer de mama (CM) es el cáncer más común en mujeres. A nivel nacional e internacional representa cerca del 15% de la mortalidad por cáncer en población femenina, resultando en un importante problema de salud pública. La forma óptima de abordarlo es el diagnóstico temprano, mediante métodos de tamizaje. La detección de factores de riesgo para el CM identificados (i.e, sexo, edad, historia familiar y factores hormonales) permiten un uso óptimo de los recursos de tamizaje, estos factores dependen directamente de la genética y los cambios en ella para cada individuo. A nivel genético, se identifican genes de susceptibilidad para el CM, que al presentar ciertas mutaciones generan alta predisposición. Sin embargo, estas mutaciones son de baja frecuencia, existiendo otras variantes genéticas de mayor frecuencia con menor efecto individual en el riesgo de CM, pero que permiten la acumulación de efectos patogénicos en la población. Estas variantes son detectables en estudios de asociación de genoma completo (GWAS), aunque no son capaces de detectar directamente todas las variantes causales, permiten la derivación de puntajes de riesgo poligénico (PRS) mediante el procesamiento de un efecto estimado de las variantes genéticas en GWAS. Los PRS captaron la atención del área biomédica y comercial, pero aún presentan margen de mejora. La teoría Omnigénica postula que la interconexión entre vías de regulación es suficiente para que todos los genes expresados afecten al desarrollo de la patología en tejidos relevantes para

esta, sugiriendo que integrar el efecto de variantes genéticas obtenidas de GWAS, junto a las interacciones moleculares de las regiones que las contienen a nivel de RNA y/o proteína, podrían mejorar el rendimiento de los PRS. En este trabajo, hipotetizamos que "*la integración de redes de interacción molecular, generadas en base a variantes genéticas asociadas a cáncer de mama, mejoran el rendimiento de puntajes de riesgo poligénico en la predicción de riesgo de la enfermedad"*. Para validar esta hipótesis, se trabajó con los siguientes objetivos: (i) selección y análisis de variantes genéticas conocidas y asociadas a CM, mediante su importancia a nivel estadístico poblacional; (ii) determinación de las redes de interacción molecular características de los genes o regiones en que se localizan las variantes genéticas seleccionadas; (iii) generación de un nuevo modelo de puntajes de riesgo poligénico para CM, a partir de una estrategia de priorización de genes vecinos dentro de las redes de interacción. En el desarrollo del trabajo, utilizamos herramientas bioinformáticas, bases de datos, además de información a nivel genómico y transcriptómico caso/control para CM, para la derivación de un puntaje poligénico integrativo. Se obtuvo un rendimiento predictivo con tendencia al aumento, levemente superior no significativo, al integrar la información de redes de interacción obtenida en los 2 modelos generados. Se concluye que los resultados mostraron cierta relación entre la interacción molecular de genes expresados en tejido mamario y un mejor PRS, pero son necesarios datos de mejor calidad, para un enfoque más específico y robusto.

ABSTRACT

Breast cancer (BC) is the most common cancer in women. Both in Chile and worldwide, it represents about 15% of cancer mortality in the female population, resulting in a major public health problem. The optimal way to address it is early diagnosis, by using screening methods. The detection of identified risk factors for BC (i.e., sex, age, family history and hormonal factors) allow optimal use of screening resources, these factors depend directly on genetics and changes in it for each individual. At the genetic level, susceptibility genes for BC were identified in families at risk, which when presenting certain mutations generate high predisposition (monogenic effects). However, these mutations are rare, and large combinations of variants of higher frequency and smaller individual variant effect sizes on the risk of BC, but that allow the distribution of pathogenic effects in the population. These variants and respective effects are detected in genome-wide association studies (GWAS), although they are not able to directly detect causal variants of the disease, they allow the derivation of polygenic risk scores (PRS) by processing an estimated effect of the genetic variants in GWAS. The PRS captured the attention of the biomedical and commercial area, but still have room for improvement. The Omnigenic theory postulates that the interconnection between regulatory pathways is sufficient for all the genes expressed to affect the development of pathology in tissues relevant to it, suggesting that integrating the effect of genetic variants obtained from GWAS, together with the molecular interactions of the regions that contain them at the

level of RNA and / or protein, could improve the performance of PRS. In this work, we hypothesized that "the integration of molecular interaction networks, generated based on genetic variants associated with breast cancer, improves the performance of polygenic risk scores in the prediction of disease risk." To validate this hypothesis, we worked with the following objectives: (i) selection and analysis of genetic variants known and associated with BC, through their importance at the population statistical level; (ii) determination of the molecular interaction networks characteristic of the genes or regions in which the selected genetic variants are located; (iii) generation of a new model of polygenic risk scores for BC, based on a strategy of prioritization of neighboring genes within interaction networks. In the development of the work, we used bioinformatics tools, databases, as well as information at the genomic and transcriptomic case/control level for BC, for the derivation of an integrative polygenic score. A predictive performance with a slightly higher non-significant tendency to increase was obtained when integrating the interaction network information obtained in the 2 generated models. It is concluded that the results showed some relationship between the molecular interaction of genes expressed in breast tissue and a better PRS, but better quality data are needed, for a more specific and robust approach.

1. INTRODUCCIÓN

1.1. Cáncer de mama

El cáncer de mama es el cáncer más común en mujeres en el mundo. Según valores estimados para el año 2020, representa cerca del 24% de los nuevos casos y 15% de la mortalidad por cáncer en mujeres. Es el cáncer de mayor diagnóstico en 157 de 185 países, y de mayor mortalidad en 110 países [1]. En Chile, para la población femenina las cifras rondan el 21% de nuevos diagnósticos y el 13% de mortalidad entre los casos de cáncer [2], convirtiéndose en un importante problema de salud pública tanto a nivel mundial como nacional. Para enfrentar esta problemática, el diagnóstico temprano es de vital importancia, ya que conduce a una mejor prognosis del desarrollo de la patología [3], sobre todo en etapas previas a la metástasis, proceso al cual se le atribuye cerca del 90% de la mortalidad asociada a cáncer [4], más aún si es diagnosticado como tumor primario removible a través de cirugía [3]. A causa de esto, algunos investigadores reportan que las tasas de sobrevida a 5 años en países desarrollados alcanzan cifras sobre el 80% (e.g. Estados Unidos 83,9%, Japón 81,6%), mientras que países de menor desarrollo rondan el 50% o inferior (e.g. India 52%, Argelia 38,8%). Estas cifras se dan principalmente debido al mejor acceso a recursos de diagnóstico y tratamiento, así como a mejores políticas públicas de control en los países desarrollados [5]. Entre las herramientas de tamizaje para el diagnóstico temprano del cáncer de mama se destaca la

mamografía, la cual ha sido asociada con importantes reducciones en las tasas de mortalidad (20 a 30% pertenecientes al 30% de descenso en la mortalidad desde 1990) [6]. Sin embargo, este recurso diagnóstico funciona una vez ya iniciado el potencial cáncer, aumentando el riesgo de detectar un cáncer en estadio avanzado por una consulta tardía, y no permite reconocer de manera previa la población que podría desarrollar la patología. Por lo tanto, es necesaria una clara identificación de los factores de riesgo asociados con esta enfermedad, para adelantarse al desarrollo de la patología e integrarlos a nuevos métodos diagnósticos, resultando en un uso óptimo de herramientas de tamizaje e imagen en poblaciones con alta predisposición para su desarrollo, permitiendo emplear los recursos de salud pública de manera eficaz.

1.2. Factores de riesgo.

Entre los factores de riesgo para el desarrollo de cáncer de mama se encuentra, en primer lugar el sexo, ya que se sabe que el cáncer de mama afecta principalmente a mujeres. En población estadounidense se reporta que el cáncer de mama en hombres comprende apenas el 1% de los casos totales [7].

La edad, ya que existe una alta relación entre la incidencia del cáncer mamario y el aumento en la edad de las mujeres [3]. En Chile, por ejemplo, la tasa de mortalidad por cáncer de mama comienza a ascender desde los 35 años, con una tasa de 31/100.000 (según quinquenio de edad), hasta una tasa 10 veces mayor en mujeres sobre los 80 años, según datos del periodo 2009-2013 [8], esta

relación es producida en parte por la acumulación de mutaciones somáticas en genes capaces de gatillar los procesos carcinogénicos [9].

El historial familiar de cáncer de mama, ya que se estima que existe un aumento cercano al doble del riesgo en mujeres con un familiar de primer grado que presente cáncer de mama, y aún mayor si este es diagnosticado antes de los 50 años de edad [10]. Este factor de riesgo está asociado con mutaciones germinales de largo efecto patogénico que se transmiten de modo Mendeliano a través de las generaciones. Por ejemplo, las alteraciones en genes BRCA son una de las fuentes principales, no solo en cáncer de mama, sino también en cáncer ovárico [11].

Factores hormonales, como se ha visto en estudios que relacionan procesos fisiológicos con alzas del riesgo de presentar cáncer de mama, como por ejemplo menarquia temprana y menopausia tardía, así como ausencia de embarazos, los cuales son procesos asociados a estados hormonales en la mujer [12]. Se identifican hormonas clave en el desarrollo tumoral en cáncer de mama, principalmente los estrógenos endógenos o exógenos y progesterona (ovariectomía reduce riesgo de cáncer de mama) [3], cuyos receptores funcionan como blanco farmacológico para terapia hormonal junto al receptor HER2, y la ausencia de estos define un cáncer de mama como triple negativo, reduciendo las opciones de tratamiento [13]. Además, se ha reportado que los patrones de expresión de estos receptores hormonales en conjunto con otros genes permiten clasificar el cáncer de mama en diferentes subtipos. Dentro de estas

clasificaciones, la más habitual es la conocida como PAM50, que utiliza 50 genes para la clasificación, y que comprende cinco subtipos: Luminal A, Luminal B, Basal, Her2 y Normal-like [14].

1.3. Predisposición genética a cáncer de mama.

En los factores de riesgo antes mencionados, la genética y los cambios en ella inciden de manera directa, y por esta razón el cáncer puede ser considerada una enfermedad genómica [15]. En la actualidad se identifican genes de susceptibilidad a CM como BRCA1, BRCA2 y TP53 (supresores de tumores), en los cuales las variantes genéticas relacionadas a la enfermedad son generalmente de alta penetrancia (cada variante patogénica presenta un gran tamaño de efecto). Lo que implica que la patología se presenta en un alto porcentaje de individuos que portan la variante genética. Sin embargo, dado que el efecto que conllevan es deletéreo, son de baja frecuencia en la población. Solo cerca de un 5% de los casos de cáncer de mama son producto de deficiencias en los genes de alta penetrancia [16], por otro lado, existen innumerables variantes genéticas de menor penetrancia (baja o intermedia), pero de una mayor frecuencia en la población, las cuales presentan una razón de probabilidades, del inglés *Odds Ratio* (OR) entre 1,2 y 1,5. Esta es una medida probabilística del riesgo de presentar una condición en una población, dada la presencia de un factor de riesgo, respecto de la ausencia de ese factor de riesgo (valores superiores a 1 indican mayor riesgo) [17]. Normalmente, estas variantes son identificadas mediante estudios de asociación de genoma completo (GWAS) [18].

1.4. Estudios de asociación de genoma completo.

Conocidos como GWAS, por su sigla en inglés (*Genome Wide Association Studies*), corresponden a estudios de gran escala, cuyo objetivo es genotipificar variantes en *loci* genómicos para detectar cuales están mayormente asociadas con rasgos o patologías [19]. GWAS detecta variantes genéticas comunes que podrían no ser las causales de la patología, pero sí determina una región génica en desequilibrio de ligamiento (fenómeno producido cuando hay alelos que debido a su cercanía física en un cromosoma se presentan juntos de una manera más frecuente de lo que se esperaría por azar [20]) con la variante genética causal [21]. Entre las limitaciones del uso de GWAS, los microarreglos para genotipado utilizados, usualmente identifican variantes que son comunes en una población en particular (Europea, principalmente). Como consecuencia, se destaca la sobrerrepresentación de población de ancestría europea en las bases de datos públicas utilizadas normalmente como referencia, limitando la precisión de diagnóstico genético en otras poblaciones, por diferencias en frecuencia alélica y estructuras de desequilibrio de ligamiento [22]. Entre los hallazgos que captaron la atención de los investigadores, se destaca que, a pesar de encontrar cientos de variantes genéticas asociadas a rasgos complejos, estos no conseguían explicar la totalidad de heredabilidad esperada, lo que es denominado heredabilidad perdida [23]. En cuanto al cáncer de mama, se han asociado sobre 1.200 variantes genéticas a la fecha en GWAS *Catalog* [24], una de las bases de datos que recopila estudios GWAS en un repositorio público de

alta calidad. A partir de esta modalidad de estudios, surgió la metodología de predicción de riesgo denominada puntajes de riesgo poligénico (PRS, del inglés *Polygenic Risk Scores*).

1.5. Puntajes de riesgo poligénico.

Esta metodología de predicción utiliza algoritmos que en general calculan la suma de las variantes genéticas de riesgo correspondientes a un fenotipo de interés en cada individuo, y luego la ponderada por la estimación del tamaño del efecto de GWAS para cada variante [25]. El tamaño de efecto específico para cada variante en el GWAS es en general expresado en OR para fenotipos binarios o en coeficiente beta de regresión para fenotipos continuos [26]. Existe esperanza en que los puntajes de riesgo poligénicos puedan mejorar las condiciones de salud al acelerar el diagnóstico y adaptar a los pacientes a tratamientos personalizados. Por esta razón, su uso se ha generalizado en las disciplinas biomédicas, con un creciente número de empresas empezando a comercializar esta tecnología [27]. Dado que son derivados de las estadísticas GWAS, estos mantienen el mismo problema de la sobrerrepresentación europea, generando rendimientos menos eficientes en poblaciones de otra ancestría [27]. La evaluación del rendimiento de los puntajes de riesgo poligénico como predictores tiene como procedimiento estándar construir curvas de característica operativa del receptor (ROC, del inglés *Receiver Operating Characteristic*), y evaluar el área bajo la curva [28]. Estudios recientes en cáncer de mama informan un rendimiento de 0,68 de área bajo la curva ROC [29]. Actualmente, estudios

continúan la búsqueda de métodos para mejorar su rendimiento. Algunos han logrado estratificar y otorgar correctamente valores de riesgo equiparables al de enfermedades monogénicas [29], otros también manejaron las diferencias entre poblaciones realizando correcciones por ancestría [30], mientras que otros han integrado el genoma personal y herramientas de *machine learning,* combinadas con el historial clínico del paciente, para predecir resultados clínicos relacionados a aneurisma aórtico abdominal [31].

1.6. Teoría Omnigénica y redes de interacción molecular.

Con los esfuerzos por entender la incógnita de la heredabilidad perdida en GWAS, en adición a las observaciones de varios grupos, que indicaban que los SNPs asociados a una patología están enriquecidos en la cromatina activa y en particular en tipos celulares relevantes para la patología, surge la teoría omnigénica, que postula que las vías de regulación están lo suficientemente interconectadas de manera que todos los genes expresados en células relevantes para el desarrollo de una patología son responsables de perturbar funciones en los genes principales de la patología. Además, afirma que la mayor parte de la heredabilidad puede ser explicada por genes fuera de las principales vías de la patología [32]. Estas perturbaciones funcionales, dependiendo del tipo de gen afectado por la variante genética, pueden desarrollarse a distintos niveles de interacción. En primer lugar, interacciones del tipo DNA-proteína, mermando acción de proteínas de unión a DNA; RNA-RNA, para genes no codificantes, que incluyen entre otros los lncRNAs y miRNAs que actúan en la regulación de RNAs

mensajeros; RNA-proteína, como ejemplo mRNAs que son sometidos a splicing y presentan regiones de unión a elementos reguladores de splicing; y proteína-proteína, ya sea interacción física o enzimática, donde el rango de efecto para el gen varía entre pérdida parcial y total de función.

Basado en lo enunciado por la teoría omnigénica y la diversidad de los efectos en la funcionalidad de las variantes genéticas, se propone analizar y caracterizar las variantes genéticas asociadas a cáncer de mama, en conjunto a las interacciones conocidas a nivel molecular que presenta el gen o la región genómica donde se localizan estas variantes genéticas y sus productos (RNA y/o proteína), considerando la generación de redes de co-expresión para obtener el contexto en el que se expresan los genes afectados, ya que a menudo, genes que participan en las mismas vías o generan complejos proteicos, son co-regulados [33], realizando un catastro de las potenciales perturbaciones que cada variante realiza a las vías relacionadas con el cáncer de mama, y de esta manera otorgar robustez al efecto de la variante e integrar abordajes de redes moleculares a los puntajes de riesgo poligénico.

2. HIPÓTESIS Y OBJETIVOS

Hipótesis

"La integración de redes de interacción molecular, generadas en base a variantes genéticas asociadas a cáncer de mama, mejoran el rendimiento de puntajes de riesgo poligénico en la predicción de riesgo de la enfermedad".

Objetivo General

Evaluar si la integración de redes de interacción molecular mejora el rendimiento en el cálculo de puntajes de riesgo poligénico, en la predicción del riesgo de cáncer de mama

Objetivos específicos:

1. Seleccionar y analizar variantes genéticas conocidas, asociadas a cáncer de mama, mediante su importancia a nivel estadístico poblacional.

2. Determinar las redes de interacción molecular características de los genes o regiones en que se localizan las variantes genéticas seleccionadas.

3. Generar un nuevo modelo de puntajes de riesgo poligénico para cáncer de mama, a partir de una estrategia de priorización de genes vecinos dentro de las redes de interacción.

3. METODOLOGÍA

3.1. Objetivo 1: Seleccionar y analizar variantes genéticas conocidas, asociadas a cáncer de mama, mediante su importancia a nivel estadístico poblacional.

3.1.1. Selección y preprocesamiento de datos genéticos

Para la generación de un puntaje de riesgo inicial, se utilizaron las métricas estadísticas de asociación resumidas de GWAS para cáncer de mama definidas a partir del estudio "Pan-UK Biobank" [34], que entregan las estadísticas de asociación de variantes genéticas con cáncer de mama de población europea, y además contiene información sobre poblaciones no-europeas, y otros rasgos o enfermedades. Se utilizó la información genotípica y fenotípica de poblaciones caso/control para cáncer de mama disponible en el estudio accedido a través de un estudio disponible en el repositorio dbGaP [35], llamado "*Cancer Genetic Markers of Susceptibility*" (CGEMS), el cual comprendió en sus primeras etapas el genotipado de 528.173 SNPs en población europea con el chip HumanHap 550 de Illumina [36], con lo que posteriormente, se realizó la imputación de 31.326.389 variantes genéticas utilizando MaCH [37][38].

Se realizó la conversión de las variantes genéticas imputadas desde los formatos de salida (extensiones: ".dose" e ".info") de Minimac2 [39], productos de la imputación empleada en el estudio, obtenidos desde dbGaP, a uno de los

formatos ampliamente usados de la suite de herramientas genéticas PLINK (extensiones: “.bed”, “.bim” y “.fam”), con el programa “GCTA” [40]. Además, se realizó un filtro preliminar según la calidad de imputación informada de RSQ > 0,3, que corresponde a la correlación al cuadrado entre los genotipos imputados y los genotipos verdaderos no observados [41]. Posteriormente, se realizó un control de calidad a las variantes genéticas presentes con el programa PLINK versión 1.9, que incluyen: remoción de variantes genéticas con MAF (*Minor Allele Frequency*) menor a 0,01; filtrado de muestras según heterocigosidad y según nivel de parentesco entre muestras [26].

3.1.2. Generación de PRS base

Se utilizó como herramienta de generación de PRS a LDpred2, disponible en el paquete de R “bigsnpr” [42], a partir de las estadísticas de asociación de GWAS mencionadas anteriormente, utilizando 1.500 individuos, correspondientes a un 66,43% del total de la población caso/control que superó los controles de calidad con PLINK, los que fueron seleccionados aleatoriamente para ser usados en una validación cruzada de 10 iteraciones, como conjunto de entrenamiento/testeo, utilizando en cada iteración un 10% de población (150 individuos) para el testeo, con el objetivo de seleccionar uno de los modelos de generación de PRS incluidos en LDPred2. Los 3 modelos de ejecución principales de LDPred2 son: (i) el modelo infinitesimal [43], que asume que todas las variantes genéticas tienen un efecto y son causales [44]; (ii) el modelo “Grid”, el cual requiere de un conjunto de datos de validación para ajustar hiper-

parámetros (*p*, la proporción de variantes causales; y h^2, que corresponde a la heredabilidad de SNP*)*, este modelo se subdivide en 2 (GRID-SP y GRID-NOSP), considerando un tercer hiper-parámetro llamado ”*Sparsity*”, el cual estima la proporción de variantes genéticas con efecto cero; y finalmente (iii) el modelo automático, que realiza el ajuste de hiper-parámetros sin necesidad de un conjunto de datos de validación [42]. LDPred2 también es capaz de realizar el control de calidad de SNPs duplicados, ambigüedad de hebras y filtrar variantes genéticas comunes para ambos conjuntos de datos (GWAS y CGEMS). Una vez obtenidos los nuevos efectos beta, se generaron los puntajes de riesgos individuales con la función “*big_prodVec()*”, y finalmente “*AUCBoot()*”, para la generación de la curva ROC y la obtención del rendimiento asociado a la AUC. La comparación entre rendimientos de modelos de LDPred2, para elegir el modelo a emplear para generar el PRS base, se realizó con el paquete de R “pROC” [45], que emplea por defecto el método “delong” descrito en DeLong et al. (1988) [46], para realizar la comparación de 2 curvas ROC, esto utilizando la población total de entrenamiento/testeo (1500 individuos).

3.1.3. Anotación genómica de las variantes seleccionadas

Se realizó el análisis de las variantes genéticas seleccionadas con LDPred2 utilizando la herramienta *Variant effect Predictor* de Ensembl [47], en su versión web. La búsqueda de elementos genómicos, se designó en una región de 5.000 pares de bases río arriba y río abajo de cada variante genética.

3.2. Objetivo 2: Determinar las redes de interacción características de los genes o regiones en que se localizan las variantes genéticas seleccionadas

3.2.1. Expresión de genes en tejido mamario

En primer lugar, se utilizó los datos generados por la iniciativa "*The Genotype-Tissue Expression*" (GTEx), que tiene como objetivo analizar la expresión en múltiples tejidos [48], para obtener la información referente a la expresión normal en tejido mamario. Los datos de expresión provenientes de GTEx fueron filtrados según sexo, manteniendo los datos de muestras femeninas, quedando así 168 perfiles de expresión, con mujeres que van de los 20 a los 79 años. A su vez, de la matriz fueron removidos los genes *PAR Y*, que dan cuenta de la expresión pseudoautosómica del cromosoma Y, dejando un total de 56.156 genes, esta cantidad de genes se definió por la cantidad en la matriz de ensembl geneIDs, notación empleada por el proyecto GENCODE [49]. La matriz de datos construida a partir de estos perfiles fue filtrada, manteniendo los genes por abundancia de RNAs, según el umbral CPM > 0,035, dado que, para este valor de cuentas por millón, se alcanza el valor de 2.2 cuentas en el tamaño promedio de las librerías de muestras (aproximadamente 63 millones de cuentas), estableciendo que este valor de CPM se alcance en al menos el 10% de las muestras. De esta manera, se estableció una lista de genes expresados normalmente en el tejido mamario.

Por otro lado, para analizar los genes expresados en el cáncer de mama, se utilizaron muestras de tejido mamario normal adyacente al tumor y tumor primario, disponibles en la iniciativa "*The Cancer Genome Atlas"* (TCGA), desde el proyecto de estudio TCGA-BRCA [50]. La obtención de los datos se realizó a través del paquete de R "TCGAbiolinks", junto a metadatos con información clínica de los individuos vinculados a las muestras, para ambos RNA-seq y microRNA-seq [51]. En adición, también se utilizó la cuantificación de isoformas de microRNA-seq para TCGA-BRCA, el cual informa la cuantificación respecto a las diferencias en la posición genómica, y permite realizar el conteo de expresión según hebra. La generación de la matriz se realizó con un script *in house* de Python, que toma la información de isoformas que integran cada muestra (**Ilustración 1**), realiza la suma de las cuentas de las isoformas que corresponden a un mismo identificador del tipo MIMAT00000XX, y la guarda en la matriz, identificando a las muestras de origen en cada columna. La recuperación de la hebra 5p o 3p se realizó con un script en R que convierte los IDs del tipo MIMAT00000XX a hsa-let-XX-5p o hsa-let-XX-3p según corresponda, empleando el paquete "miRBaseConverter" [52].

miRNA_ID	isoform_coords	read_count	reads_per_million_miRNA_mapped	cross-mapped	miRNA_region
hsa-let-7a-1	hg38:chr9:94175942-94175962:+	1	0.075462	N	precursor
hsa-let-7a-1	hg38:chr9:94175961-94175982:+	15	1.131923	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175961-94175983:+	23	1.735616	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175961-94175984:+	62	4.678616	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175962-94175981:+	298	22.487540	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175962-94175982:+	13869	1046.576148	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175962-94175983:+	38195	2882.253656	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175962-94175984:+	96603	7289.811492	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175962-94175985:+	2467	186.163628	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175962-94175986:+	42	3.169385	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175962-94175992:+	1	0.075462	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175963-94175983:+	2	0.150923	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175963-94175984:+	16	1.207385	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175963-94175985:+	1	0.075462	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175965-94175983:+	5	0.377308	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175965-94175984:+	7	0.528231	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175966-94175983:+	1	0.075462	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175966-94175984:+	1	0.075462	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175967-94175991:+	1	0.075462	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175984-94176005:+	1	0.075462	N	stemloop
hsa-let-7a-1	hg38:chr9:94175984-94176006:+	2	0.150923	N	stemloop

Ilustración 1. Formato cuantificación de isoformas. Formato de archivo que informa la cuantificación de isoformas para microRNA-seq de cada muestra en el proyecto TCGA-BRCA.

Se realizó un filtro de los individuos en los conjuntos de datos obtenidos desde TCGA, manteniendo sólo los individuos femeninos que presenten a la vez datos de RNA-seq y microRNA-seq, con muestras pareadas (tumor primario y tejido sólido normal) en ambos casos, generando un subconjunto de 198 muestras, con la información de 99 individuos.

3.2.2. Análisis de expresión diferencial en subtipos de cáncer de mama

Los genes diferencialmente expresados se obtuvieron con el paquete de R DESeq2 [53], estableciendo como punto de corte de expresión diferencial un Log2FC de ±2 en general para RNA-seq, con excepción de los lncRNAs, en los cuales se estableció el mismo punto de corte utilizado para microRNA-seq de Log2FC ±1. Para ambos casos, se usó un *p-value* ajustado (también llamado FDR, por "*False Discovery Rate*") <= 0,05. Esta métrica es utilizada para reducir el número de falsos positivos en los análisis que generan gran cantidad de datos. Derivado de este análisis, se extrajo la matriz de cuentas normalizada por

DESeq2, con la que se realizó un análisis de componentes principales para reducir la dimensionalidad de los datos. Para esta, se utilizaron los 500 genes (valor por defecto) que mostraron mayor varianza en su expresión en la totalidad de las muestras, tanto para RNA-seq, como para miRNA-seq.

3.2.3. RNAs de competencia endógena (ceRNAs)

En colaboración con el Dr. Ignacio Wichmann (Pontificia Universidad Católica, Chile), quien trabajó con generación de redes de competencia endógena en su tesis doctoral titulada "*Identification of long noncoding RNAs in competing endogenous RNA network throughout the gastric precancerous cascade*" [54], se desarrolló un script de R *in house* para descubrir los lncRNAs participantes en la regulación de la expresión proteica, por medio de mecanismos de competencia endógena. Para tanto, se analizaron en conjunto la expresión de RNA total y de isoformas de miRNAs desde TCGA utilizando el paquete de R "SPONGE", el cual permite identificar rápidamente redes de interacción de ceRNAs, y a su vez, cuantificar la participación de múltiples miRNAs a cada interacción ceRNA [55]. Para esto, se utilizó además una matriz de interacciones del tipo miRNA-mRNAs y miRNA-lncRNAs, previamente generada en nuestro laboratorio en el trabajo de tesis magistral de Allan Peñaloza [56], la cual emplea información de interacción de 4 bases de datos de interacciones miRNAs-mRNAs blanco, de las cuales 3 son de interacciones por predicción (TargetMiner [57], miRDB [58] y TargetScanHuman [59]), en donde se mantuvieron aquellas interacciones presentes simultáneamente en las 3 bases de datos, y una base

de datos con interacciones validadas(TarBase [60]); finalmente, otras 2 bases de datos de interacciones miRNA-lncRNA, lncBase [61] y starBase [62]. Previo al análisis con SPONGE, las matrices de expresión fueron filtradas por abundancia, manteniendo las muestras con CPM (*counts per million*) superiores a 0,035 para RNA y de 1 para miRNA (valores dependen del tamaño promedio de las librerías), en al menos el 5% de las muestras, y finalmente normalizadas a TMM con transformación logarítmica, aplicando el paquete edgeR [63]. La generación de redes ceRNA con SPONGE en general consta de 4 etapas: (i) Se identifican aquellos miRNAs que probablemente tienen un efecto regulador a través de la búsqueda de correlaciones negativas con genes marcados como objetivos en la matriz de interacción; (ii) luego se filtra los pares de genes que compartan regulación por uno o más miRNAs, generando el coeficiente "*mscor*" (del inglés, *multiple miRNAs sensitivity correlation*), implementado por los autores de SPONGE, que da cuenta del efecto de múltiples miRNAs en la correlación entre 2 genes evaluados; (iii) en seguida se establece las significancias de las interacciones encontradas por medio de un muestreo de la distribución nula de mscor; (iv) y finalmente, se filtran las interacciones por grado de significancia, donde se estableció el umbral de *p-value* ajustado < 0,05 [55]. La visualización de las redes generadas se realizó en Cytoscape [64]. Para tener una idea del rol que estarían cumpliendo estos lncRNAs por medio del mecanismo de competencia endógena, se realizó un análisis de enriquecimiento funcional utilizando las extensiones de Cytoscape: BINGO para la búsqueda de términos

GO asociados a las proteinas [65], EnrichmentMap para visualización de vías enriquecidas en los términos (umbral establecido de $5*10^{-7}$ > FDR) [66] y Autoannotate para agrupar los términos más frecuentes entre las vías más enriquecidas [67].

3.2.4. Redes de interacción proteína-proteína

Se realizó la búsqueda en la base de datos STRING (versión 11), la cual presenta datos de interacción entre proteínas, ya sean físicas o funcionales, las que se filtraron por puntuación combinada (valor obtenido al evaluar los orígenes y análisis que respaldan a cada interacción) ≥ 0,4, para obtener interacciones de confiabilidad media a superior [68], un filtro adicional se realizó a las proteínas interactuantes obtenidas, usando la lista de genes generada a partir de los datos de expresión de GTEx. Se reemplazó la anotación original de Ensembl ID, utilizada en la base de datos de STRING, a "*Gene* Symbol", con el paquete de R "EnsDb.Hsapiens.v86" [69].

3.2.5. Redes de regulación de factores de transcripción

Se utilizó el paquete de R "DoRoThea", el cual es una colección de regulones (grupo de genes regulados por un factor de transcripción), que almacena la información referente a genes blanco y sus factores de transcripción a partir de 4 fuentes de evidencia: (i) levantamiento bases de datos o literatura con una curación de datos manual por un experto humano; (ii) experimentos de ChIP-seq; (iii) inferencia en base a datos de GTEx; y, finalmente, (iv) predicción computacional en base a motivos de unión en promotores, pertenecientes a

humano y ratón. Con eso, DoRoThea define 5 niveles de confianza, designados de la “A” hasta la “E”, según el respaldo de la información en las 4 fuentes de evidencia, donde “A” constituye información curada y “E” solo contempla predicciones [70]. Se realizó una selección de aquellas interacciones pertenecientes a humano, descartando el nivel “E” de confianza, y se generó el filtrado en base a la abundancia de expresión en tejido mamario para los genes blanco y factores de transcripción, según los datos de GTEx, de la misma manera en que se realizó para las interacciones proteína-proteína.

3.2.6. Redes de interacción lncRNA-DNA y lncRNA-proteína

Para enriquecer la cantidad de interacciones de ncRNAs, se hizo descarga de los datos disponibles en la base de datos lncRNAfunc [71], un repositorio dedicado a compilar información sobre la función de los lncRNAs en cáncer humano, integrando, por ejemplo, interacciones lncRNA-DNA y lncRNA-proteína, entre otros, haciendo uso de los datos depositados en TCGA. Se filtraron aquellas interacciones que no aparezcan respaldadas por experimentos, y que cuyos interactores no estén en la lista de expresión generada a partir de los datos de GTEx, exceptuando miRNAs, no incluidos en dicha lista de expresión.

3.2.7. Módulos de co-expresión

Para generar módulos de co-expresión y evaluar vías involucradas con potencial interacción con los genes afectados por cada variante genética, se utilizó la herramienta CEMiTool [72], usando como entrada la matriz de expresión de RNA-seq normalizada en TMM, aplicando también la red de interacción de

proteínas, obtenida a partir de STRING [68] como red de referencia. La herramienta CEMiTool permite realizar una reducción de elementos según su abundancia en la matriz de expresión de manera automática (25% de menor expresión promedio), requerida para hacer uso de la función VST (*Variance Stabilizing Transformation*), recomendada en la publicación original de la herramienta, para datos de RNA-Seq. CEMiTool realiza una selección de genes basados en p-values obtenidos a través del modelado de la varianza como una distribución gamma inversa, determina un criterio de similitud entre pares de genes y finalmente estos son separados en módulos utilizando de manera interna el paquete de R "Dynamic Tree Cut package" [72]. El análisis con CEMiTool también genera un análisis de enriquecimiento de conjuntos de genes (GSEA, *Gene Set Enrichment Analysis*), el cual da cuenta de la actividad transcripcional de los módulos detectados, informando una puntuación normalizada de enriquecimiento (NES), basada en cuan representados están los integrantes de estos módulos en los extremos de una lista ordenada de genes, según cambios de expresión en las condiciones presentadas, para lograr esto, CEMiTool emplea otro paquete de R llamado fgsea [73]. El enriquecimiento funcional de los integrantes de cada módulo se llevó a cabo con el paquete de R "clusterProfiler" [74], utilizando el conjunto de datos de MSigDB hallmarks 2020 [75], accedidos desde la web de enrichR [76], el enriquecimiento de términos, se realizó considerando un p-value ajustado menor a 0,05.

3.3. Objetivo 3: Generar un nuevo modelo de puntajes de riesgo poligénico para cáncer de mama, a partir de una estrategia de priorización de genes vecinos dentro de las redes de interacción

3.3.1. Creación del modelo y estrategias de priorización de genes

La matriz de interacción se generó con un script *in house* escrito en python3, el cual transforma la lista de interacciones obtenida de los análisis anteriores o de bases de datos, haciendo uso de la función *crosstab* del módulo de Python "pandas" [77], en una matriz que suma las interacciones encontradas entre genes. Cabe destacar que los diferentes tipos de interacción encontrados (proteína-proteina, ceRNA, miRNA-mRNA, etc.), se ponderaran con el mismo valor (1), ya que de otra manera la complejidad de los modelos y la información requerida para generarlos, aumentan significativamente. En adición, se generó una tabla de priorización, la que contiene los genes que se expresan diferencialmente en alguno de los subtipos PAM50 y/o que forman parte de alguno de los módulos de co-expresión detectados por CEMiTool.

Con eso, se generaron 2 modelos, el primero basado netamente en la cantidad de interacciones a 2 niveles desde los genes afectados, y sin priorización de genes:

$$IS_G = INT_G + \sum_{i=1}^{INT_G} \frac{INT_{gi}}{2}$$

Donde IS_g representa el puntaje de interacción de cada gen potencialmente afectado, INT_g las interacciones de primer grado del gen, INT_{gi} el número de interacciones de los interactores en primer grado del gen.

El segundo modelo generado corresponde a una extensión del primer modelo, aplicando una priorización de genes, en la cual se integró la información obtenida sobre la expresión diferencial a lo largo de los subtipos PAM50 de cáncer mamario, y pertenencia a alguno de los módulos de co-expresión relacionados a procesos tumorales. Para esto se derivó un puntaje para cada gen con base en su grado de expresión diferencial general (valor absoluto de Log2FC en la comparación de todas las muestras tumorales contra los controles), amplificando según su estado de expresión diferencial en los subtipos PAM50 de cáncer de mama, esta amplificación parte del valor base 1, hasta un valor máximo de 6, que considera 4 subtipos (+1 por cada subtipo), y la expresión diferencial en la comparación general (+1). La bonificación por pertenencia a un módulo de co-expresión relacionado a cáncer se estableció en 10%, la relación a cáncer de los módulos se estableció según el enriquecimiento funcional generado a través de CEMiTool [72] y MSigDB hallmarks 2020 [75]. Las interacciones próximas aportan al puntaje de cada gen, usando el mismo criterio que para el primer modelo, reduciendo el valor del puntaje de cada interactor del gen evaluado a la mitad. Este planteamiento se resume en la siguiente ecuación:

$$IS_G = INT_G * DES_G * M_G + \sum_{I=1}^{INT_G} \frac{INT_{Gi} * DES_{Gi} * M_{Gi}}{2}$$

Donde se añaden los términos $M_G|M_{Gi}$, que representan la bonificación otorgada, según la inclusión del gen en un módulo de co-expresión relacionado a cáncer, y finalmente $DES_G|DES_{Gi}$ que representan factores generados con la información de expresión diferencial.

Una vez obtenidos los puntajes de interacción de cada gen (IS_G), se procedió a realizar la normalización del rango de estos puntajes en un rango de valores de 1 a 6, en donde el gen con mayor IS_G tomó el valor de 6 y aquellos sin información de interacción disponible, tomaron el valor de 1 (neutro multiplicativo), los valores normalizados se representan con el factor Ω.

Finalmente, se utilizó una versión ampliada de la formula básica de generación de PRS individual para cada modelo, incluyendo el valor normalizado (Ωj) máximo de entre los genes cercanos de cada variante, determinados en el análisis previamente realizado con VEP de Ensembl [47]. La ecuación que resume lo expuesto se presenta a continuación.

$$\hat{S} = \sum_{j=1}^{m} X_j LDPred2\beta_j * \Omega_{maxj}$$

Donde $\hat{S}$ representa el puntaje de riesgo individual, m el número total de variantes de riesgo que presenta un individuo, X_j la variante de riesgo evaluada,

LDPred2β_j el valor del efecto de la variante calculado por LDPred2, y Ω_{maxj} el valor normalizado máximo entre los genes cercanos a la variante X_j.

3.3.2. Evaluación del modelo generado

Los modelos generados se compararon en rendimiento frente a los resultados generados por LDPred2, haciendo uso de la población de comparación previamente definida de 758 individuos, la cual se subdividió en tres grupos, dos grupos de 250 individuos y un grupo de 258, por selección aleatoria, para comprobar el desempeño en diferentes configuraciones poblacionales de una misma ancestría.

4. RESULTADOS

4.1. Objetivo 1: Seleccionar y analizar variantes genéticas conocidas, asociadas a cáncer de mama, mediante su importancia a nivel estadístico poblacional.

4.1.1. Generación de un conjunto de variantes genéticas de referencia y su nivel de efecto base en cáncer de mama

El trabajo se inició estableciendo el modelo base, utilizando el conjunto de herramientas para la generación de puntajes de riesgo poligénico LDPred2 [42], la cual integra cuatro abordajes posibles de ejecución en la generación de PRS: (i) Infinitesimal, (ii) Grid-NOSP, (iii) Grid-SP y (iv) Automático. La herramienta fue aplicada en los datos del estudio CGEMS [36], haciendo una división aleatoria de los individuos del estudio en dos grupos para determinar los mejores rendimientos. El primer grupo estaba compuesto por 1.500 individuos, con el cual se realizó la validación cruzada de 10 iteraciones (**Tabla 1**); mientras que el segundo grupo contemplaba 758 individuos, los cuales serán utilizados en la comparación con los modelos generados más adelante en esta tesis, construidos a partir de las redes de interacción analizadas.

El universo de variantes genéticas inicialmente de 31.326.389, posterior al control de calidad realizado con la herramienta PLINK, se redujo a 6.303.051, las cuales fueron utilizadas como entrada para los análisis comparativos utilizando los cuatro abordajes disponibles en LDPred2. Este tipo de análisis depende de la

intersección de las variantes en la cohorte caso/control (CGEMS) y las que componen los conjuntos de datos del resumen estadístico del GWAS de asociación a cáncer de mama (Pan-UK Biobank). De manera adicional, por recomendación de los autores de LDPred2, se consideraron solo variantes incluidas en HapMap3 [78]. Con eso, el universo final de variantes genéticas que entraron al análisis de efecto en cáncer de mama luego de ser filtradas para mantener solamente aquellas presentes en HapMap3 fue de 766.518.

Validación Cruzada 10-fold				
K	Infinitesimal	Grid-NOSP	Grid-SP	Automático
1	0,589523	0,601834	0,601254	0,584032
2	0,539019	0,62455	0,643568	0,650704
3	0,544759	0,6015	0,600406	0,58271
4	0,619832	0,668406	0,661355	0,632194
5	0,574441	0,593873	0,597906	0,610211
6	0,609489	0,681744	0,687562	0,669993
7	0,606196	0,608209	0,60184	0,584449
8	0,503522	0,569172	0,569349	0,578427
9	0,569325	0,576254	0,560265	0,555608
10	0,588375	0,544391	0,526343	0,504714
Promedio	0,5744481	**0,6069933**	**0,6049848**	0,5953042

Tabla 1. **Validación cruzada k=10**. Valores de AUC en cada iteración para los 4 modelos de generación de PRS de LDPred2, y promedio final del rendimiento.

En este análisis, dado que los valores del efecto beta de cada variante cambia entre cada iteración, se utilizó como valor final de efecto, el promedio de las 10 iteraciones para cada variante genética. Como resultado, se obtuvo que para este conjunto de datos el rendimiento promedio de área bajo la curva ROC (AUC) más alto fue alcanzado por el modelo Grid-NOSP con 0,606, seguido del modelo Grid-SP (0,604), el cual, dado que es capaz de reducir el efecto de ciertas variantes genéticas a 0, reduce el universo de variantes genéticas en el PRS a 500.689. Los otros dos modelos presentaron un rendimiento promedio inferior a 0,6.

Para la generación del PRS Base, comparativamente se reutilizó el conjunto de 1.500 individuos del entrenamiento/testeo, y de esta manera decidir entro los modelos GRID-SP y GRID-NOSP, dado que su diferencia de

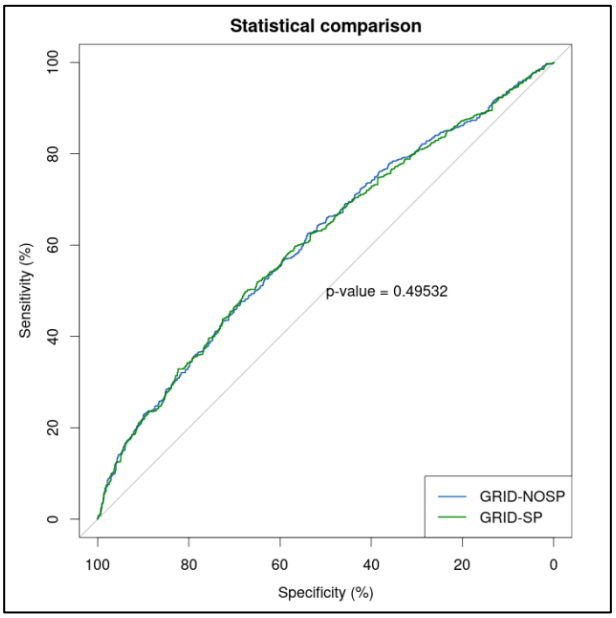

Figura 1. Comparación modelos GRID-SP y GRID-NOSP. Gráfica de curvas ROC, según su puntaje de riesgo poligénico individual. Método “delong”, paquete de R pROC. N=1500.

rendimiento promedio en la validación cruzada es baja (-0,002). Se determinó que la diferencia entre las curvas ROC de ambos modelos, no es significativa (**Figura 1**).

Con estos datos, se seleccionó el segundo modelo de mayor rendimiento (Grid-SP), como modelo de generación del PRS Base. Esta selección va en la dirección de estrategias en el área de los PRS como *fine-mapping*, que buscan

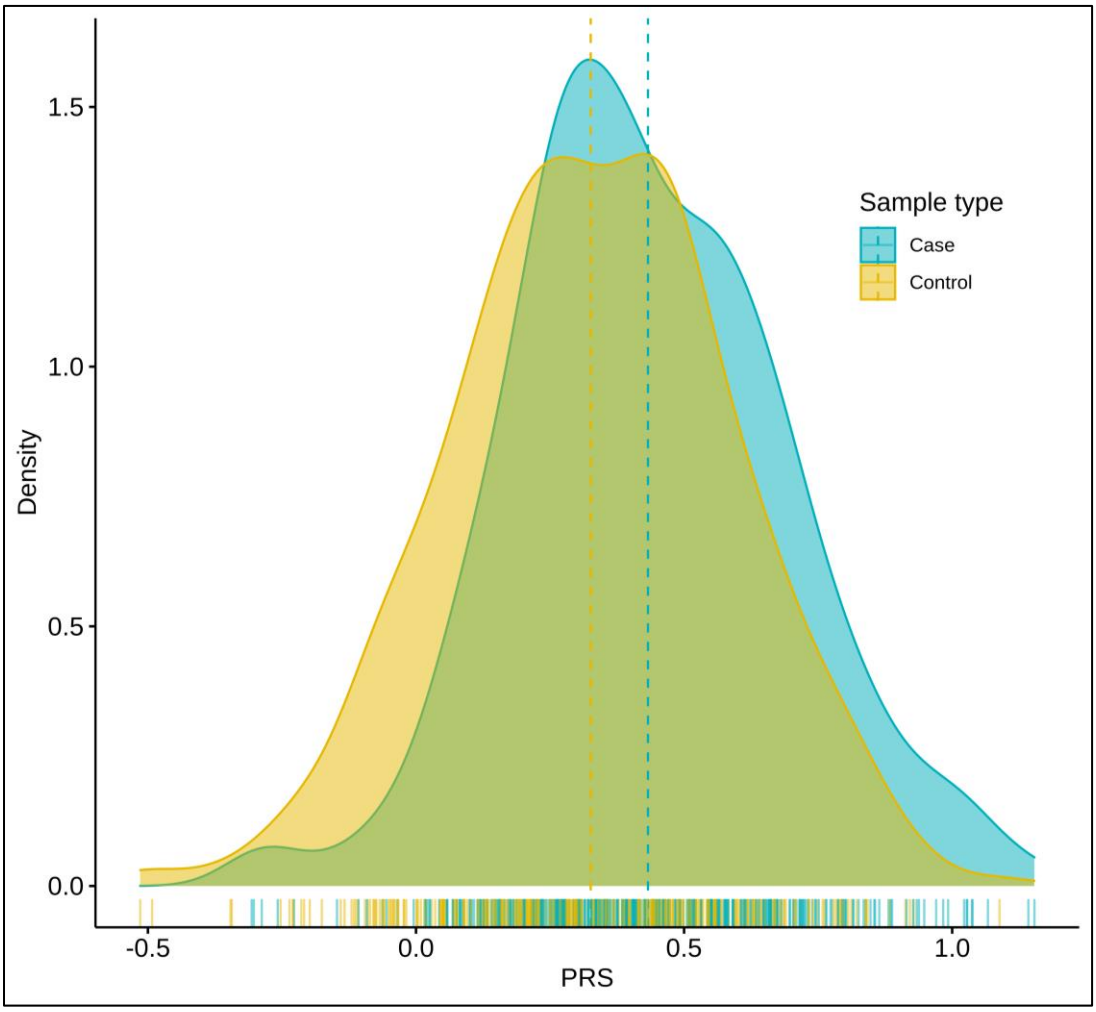

Figura 2. Distribución en PRS base. Gráfico de densidad mostrando la distribución de los individuos en la población de comparación, según su puntaje de riesgo poligénico individual, en amarillo se representan los controles, en azul se representan los casos. N=758.

reducir el número de variantes genéticas y determinar la variante causal más probable en una región genómica [79]. En seguida, se procedió a calcular las puntuaciones individuales de riesgo poligénico en la población de comparación, obteniendo una distribución mostrada en la **Figura 2**, con lo que además se calculó el riesgo de presentar el cáncer de mama en los percentiles de mayor puntaje. Por ejemplo, para el percentil 5% superior se alcanzó un *odds ratio* aproximado de 2,944.

4.1.2. Anotación de las variantes encontradas

Una vez establecidas las 500.689 variantes genéticas de trabajo, utilizando sus rsID, se realizó la anotación de las variantes usando la herramienta de Ensembl VEP (*Variant Effect Predictor*) [47], la cual entrega información sobre los genes cercanos a las variantes genéticas de estudio y sus respectivas anotaciones. VEP logró encontrar 500.656 variantes genéticas del total, de las cuales 138.381 fueron categorizadas como intergénicas dentro del rango establecido (±5kb, ver metodología). En cuanto a genes potencialmente afectados, estos alcanzan la suma de 44.933, dentro de los cuales 16.638 se consideran genes codificantes de proteínas, 14.050 lncRNAs, 1.198 miRNAs, el resto se divide en pseudogenes, otros RNAs pequeños (como miscRNA, snoRNA, snRNA y scaRNA), y genes codificantes relacionados a las cadenas constante y variable de los receptores de células T o inmunoglobulinas.

4.2. Resultados del Objetivo N°2

“Determinar las redes de interacción molecular características de los genes o regiones en que se localizan las variantes genéticas seleccionadas”.

4.2.1. Evaluación de los datos obtenidos desde el repositorio TCGA

El universo total de muestras disponible en el repositorio de datos de TCGA para cáncer de mama correspondientes al proyecto TCGA-BRCA, incluye a 1.097 individuos, de los cuales 1.085 son mujeres. Dado que los análisis realizados con estos datos tienen un fin comparativo, el universo de individuos se redujo en función de la disponibilidad de los análisis de RNA-seq (RNA total) y microRNA-seq (RNAs pequeños, con la respectiva muestra control en cada caso, permaneciendo muestras referentes a 99 mujeres en el análisis. En adición, se realizaron análisis de reducción de dimensionalidad según la expresión, mediante componentes principales (PCA), para corroborar que las muestras control presentaran patrones de expresión coherentes con su estado de muestras control (**Figura 3**). Con eso, se descartaron 2 muestras, en base al perfil de expresión de miRNAs: TCGA-BH-A18U-11, una muestra control; y TCGA-BH-A0BZ-01, esta última corresponde a una muestra tumoral, que agrupó con las muestras normales (Fig. Anexo 1). Cabe resaltar que, en ambos PCA generados, se observa una amplia dispersión de las muestras normales a lo largo del eje de la componente número 2, mayor a la mostrada por las muestras tumorales.

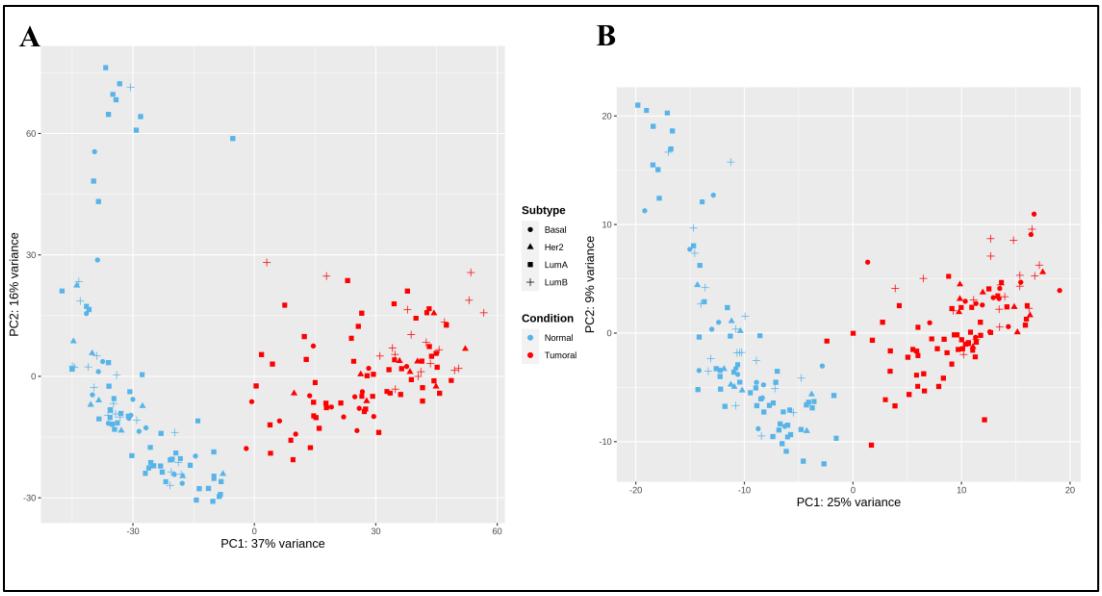

Figura 3. Análisis de componentes principales. Graficas con las 2 primeras componentes principales. **A:** Datos de RNA-seq, componentes resumen el 53% de la varianza. **B:** Datos de miRNA-seq, componentes resumen el 34% de la varianza. N=194 muestras.

La información clínica de las 194 muestras (miRNA-seq y RNA-seq) provenientes de 97 individuos, obtenidas desde el proyecto TCGA-BRCA, y utilizadas en este proyecto de investigación, permitió utilizar la clasificación de las muestras según el test genético PAM50, el cual evalúa la actividad de 50 genes para clasificar el tipo de cáncer de mama. Esta clasificación en general considera 5 subtipos: Luminal A, Luminal B, Basal, Her2 y Normal-like (no incluido en la investigación debido a que solo presentaba 2 muestras control, y su cada vez menor consideración en la literatura [80]). Además, se recuperó el estadio de cada tumor al momento de tomarse la muestra, observándose que la mayoría de

las muestras provienen de tumores en etapa II (57 de 97). La distribución de ambas informaciones puede observarse en la **Figura 4**.

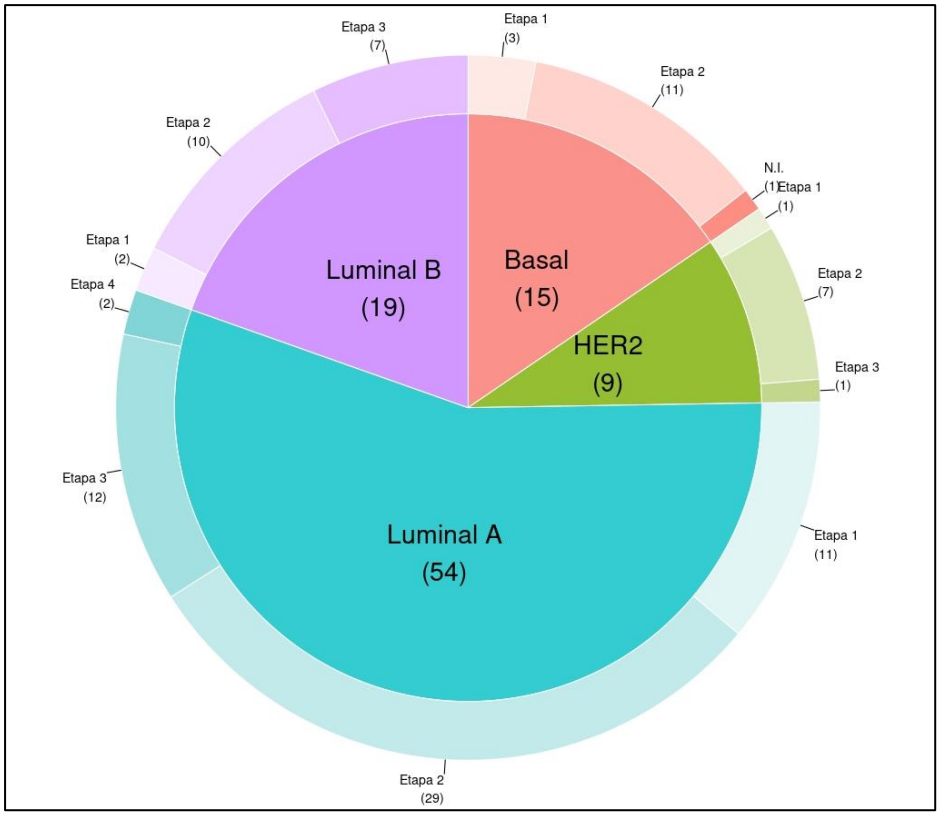

Figura 4. Distribución muestral subtipos PAM50. Composición de universo de muestras de cáncer mamario utilizadas, según clasificación de prueba genética PAM50 del paciente (Luminal A, Luminal B, Basal y Her2) y estadio tumoral (I a IV) para cada una de ellas. N=97

4.2.2. Análisis de expresión diferencial en subtipos de cáncer de mama

Una vez generada la clasificación de las muestras, se procedió a realizar diversos análisis de expresión diferencial utilizando la herramienta DESeq2 [53]. Entre las comparaciones, se evaluó el total de las muestras tumorales *versus* los controles normales, y luego se subdividieron según el subtipo PAM50 *versus* la

totalidad de los controles normales (dado el bajo número de controles en algunos subtipos). La comparación general entre muestras normales y tumorales (**Figura 5**), arrojó un total de 5.662 elementos expresados diferencialmente (3.443 sobreexpresados y 2.219 subexpresados) para los datos de RNA-seq (RNA total), mientras que para datos de microRNAs, se presentaron 306 expresados diferencialmente (203 sobreexpresados y 103 subexpresados).

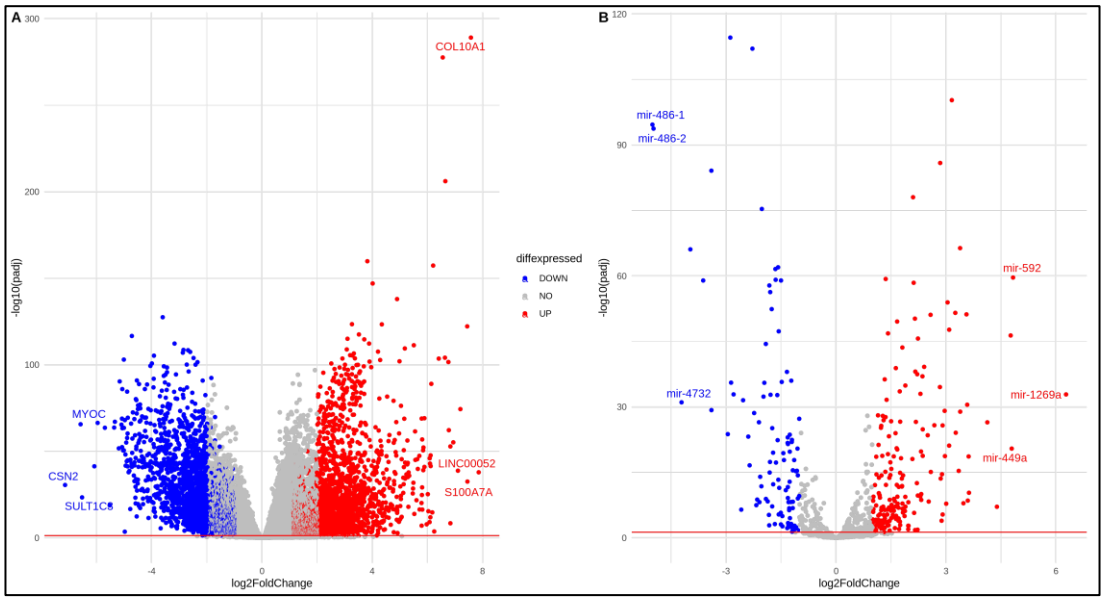

Figura 5. Expresión diferencial en cáncer de mama. Análisis de expresión diferencial en cáncer de mama, mediante DESeq2. **A:** Expresión diferencial en datos de RNA-seq. **B:** Expresión diferencial en datos de miRNA-seq.

En cuanto a la cantidad total de genes diferencialmente expresados por subtipo (**Figura 6**), se desglosó el análisis de RNA-seq en los biotipos de anotación "*Protein coding*", "lncRNA" y "*Others*", según la clasificación de GENCODE v38 [49]. Este último incluye otros biotipos de RNAs, tales como scaRNA, snoRNA, snRNA y relacionados a pseudogenes). Además, también tenemos en nuestros datos el biotipo miRNAs, aportado por el análisis de miRNA-

seq. En estos análisis, se observa que el subtipo Basal es el que tiene una mayor cantidad de genes sobreexpresados respecto a la condición normal (5.270); y que el subtipo Luminal B es el que posee una mayor cantidad de genes subexpresados (5.359). Para ambos casos, el mayor número de genes expresados diferencialmente corresponden a RNAs no codificantes largos (2.035 y 2.357, respectivamente). El biotipo con menor cantidad de elementos expresados diferencialmente corresponde a los miRNAs, y es donde el subtipo

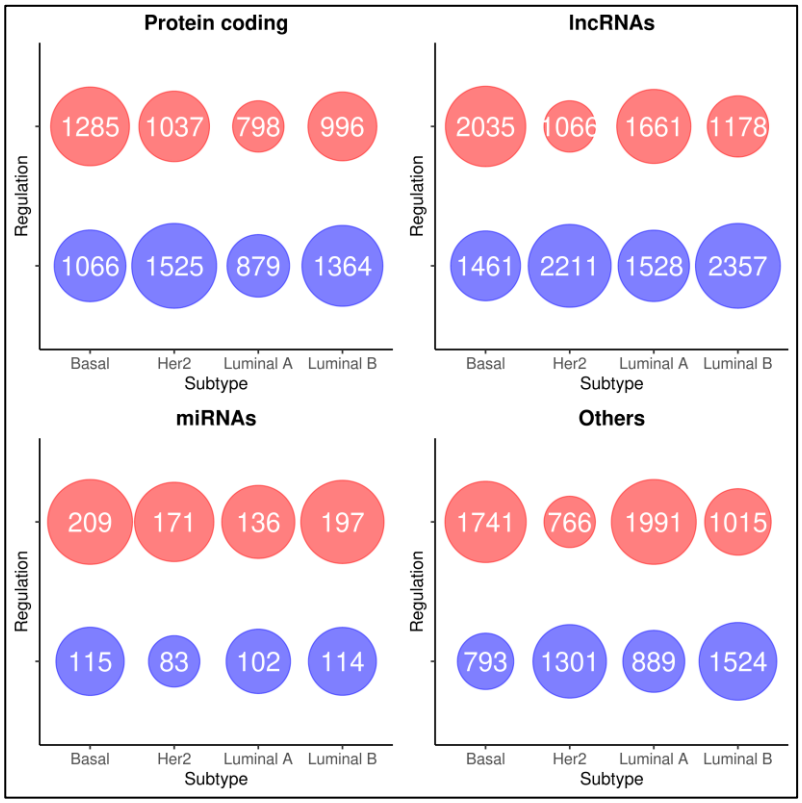

Figura 6. Genes diferencialmente expresados por subtipo y biotipo. Se muestra la cantidad total de genes expresados diferencialmente separados en paneles según biotipo de RNA, incluyendo la expresión aumentada (rojo) y expresión reducida (azul), por subtipo PAM50 de cáncer de mama. El panel "*Others*" incluye las cantidades de genes para biotipos de RNAs menos abundantes.

Basal presenta la menor diferencia en términos numéricos, respecto a los demás subtipos PAM50.

Se analizó también, cuan compartidos están los genes diferencialmente expresados entre los distintos subtipos de cáncer (**Figura 7**), para determinar genes que podrían actuar como núcleo del cáncer de mama, así como genes únicos de cada subtipo, con la finalidad de encontrar elementos a priorizar en las redes de interacción. Este análisis, se realizó para los 3 principales biotipos de RNA, partiendo por los RNAs codificantes de proteínas (**Figura 7A**), en donde se obtuvo que hay 704 elementos compartidos y el subtipo PAM50 Basal presentó una notable mayor cantidad de elementos sobreexpresados únicos (602). En el caso de los lncRNAs (**Figura 7B**), se observa que hay 871 de estos, compartidos por todos los subtipos, y nuevamente el subtipo PAM50 Basal presentó una muy mayor cantidad de elementos sobreexpresados únicos respecto al resto (1123). Cabe destacar que tanto para lncRNAs, como para codificantes de proteínas, el subtipo PAM50 Her2 fue el de mayor cantidad de elementos únicos subexpresados (494 y 329, respectivamente). Finalmente, para los miRNAs (**Figura 7C**), el subtipo Basal presenta la mayor cantidad de miRNAs sobreexpresados y subexpresados únicos (60 y 23, respectivamente), en cuanto a los compartidos por todos los subtipos, estos alcanzan valores de 46 para los subexpresados, y 75 para sobreexpresados. Estos resultados, muestran que el subtipo Basal es el que más difiere del resto, lo que es coherente con lo descrito en literatura, dado que a este se le relaciona con los tumores triple negativo,

mientras que los otros 3 subtipos restantes tienen carácter hormonal, combinando de distintas maneras la presencia de los receptores Her2, ER y PR.

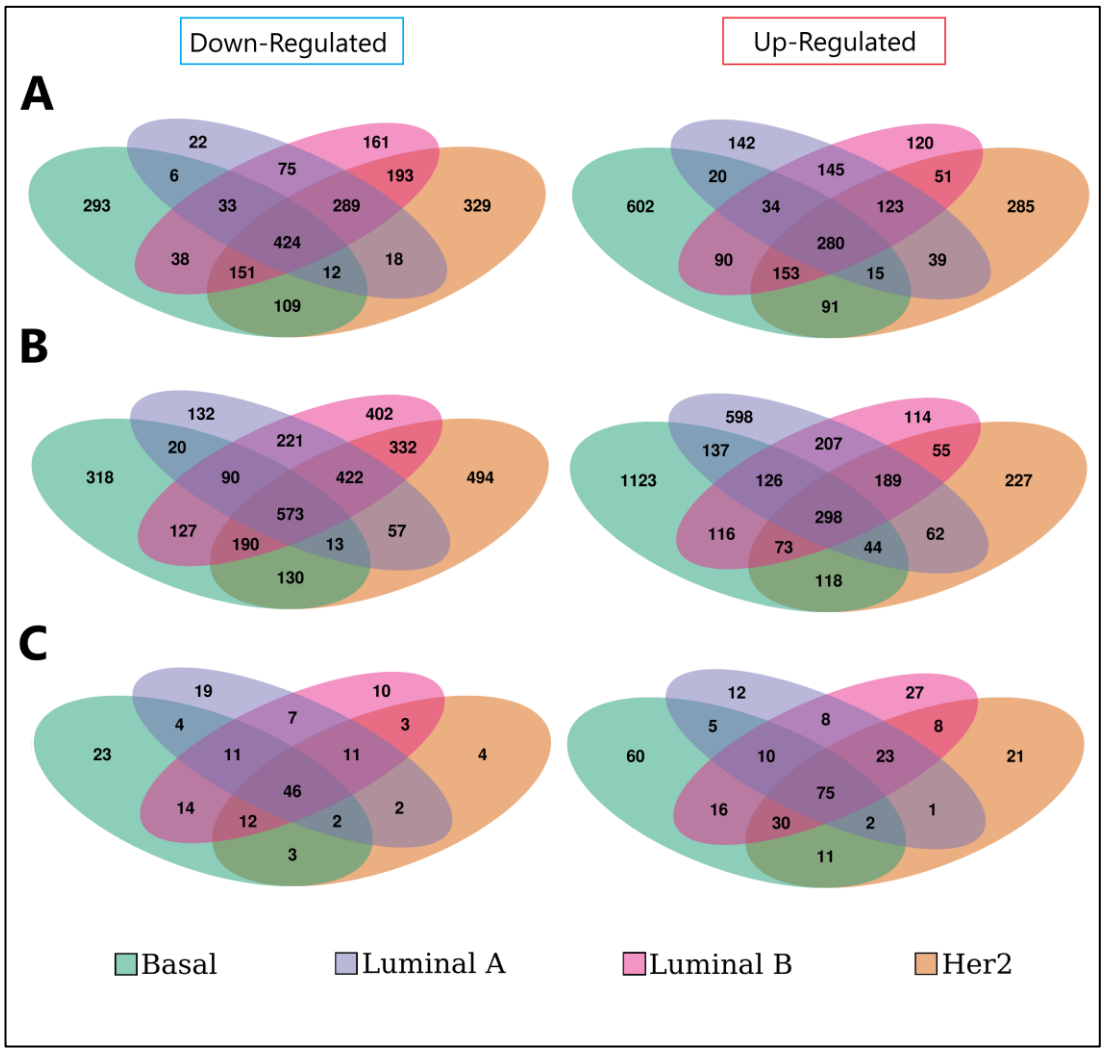

Figura 7. RNAs expresados diferencialmente compartidos entre subtipos. Diagramas de Venn indicando la cantidad de miRNAs expresados diferencialmente y su nivel de repartición entre los subtipos tumorales. Izquierda, cantidades de miRNAs de expresión reducida; derecha, las cantidades asociadas a miRNAs de expresión aumentada en cáncer de mama. **A**: RNAs codificantes de proteínas. **B**: lncRNAs. **C:** miRNAs.

4.2.3. Redes de competencia endógena de RNAs en cáncer de mama

La generación de redes de ceRNAs, las cuales constan de 2 o más RNAs que compiten en la unión de miRNAs, y que afectan la regulación de los RNAs mensajeros, generan un acercamiento al rol que pueden estar cumpliendo algunos lncRNAs en el desarrollo o combate del cáncer de mama. Los resultados de este análisis realizado con la herramienta SPONGE arrojaron como resultados 5.141.658 de interacciones, de las cuales 687.132 son consideradas estadísticamente significativas (FDR < 0,05).

Dado que el enfoque de estos análisis es el rol de los lncRNAs, con parte de estas interacciones se representan visualmente las redes de interacción de ceRNAs en el software de visualización de redes Cytoscape [64], mostrando aquellas interacciones donde participan los lncRNAs determinados anteriormente como expresados diferencialmente en la comparación general de muestras de cáncer de mama (**Figura 8A**). Cabe señalar que los miRNAs involucrados han sido omitidos para mejorar la visualización. Se observa que la mayoría de las interacciones encontradas, son moduladas por lncRNAs que están subexpresados, esto se puede constatar, por la mayor cantidad de elementos que presentan al menos 200 interacciones dentro de la red ceRNA (etiquetados con nombre en **Figura 8A**). Los lncRNA, que cumplen esta condición son: *HOXA-AS3*, *CAHM*, *EMX2OS*, *CARMN*, *ADAMTS9-AS2*, *LINC02607*, *HOTAIRM1*, *MAGI2-AS3*, *MEG3*, *MIR99AHG*, *PCAT19*, *RP11-679B19.1*, *MIR100HG*, *SLC26A4-AS1*, *LINC02693*, *LINC00667* y *HCG11*. En concordancia, la mayoría

de las proteínas que interactúan con estos, están subexpresadas o no presentan una expresión diferencial significativa. Por otro lado, para los lncRNAs sobreexpresados, solo hay 2 de ellos que presentan una conectividad de al menos 200 interacciones, los cuales son *LINC00511* y *LCT-AS1*. Al igual que con los lncRNAs subexpresados, hay en mayoría concordancia entre proteínas sobreexpresadas que interactúan con lncRNAs sobreexpresados.

Para tener una visión indirecta de la funcionalidad y procesos biológicos en que podrían estar participando estos ceRNAs, se analizaron las proteínas involucradas en las redes de ceRNAs obtenidas de manera comparativa según la expresión diferencial. Las proteínas subexpresadas (**Figura 8B**), se relacionaron con regulación de procesos, entre ellos la regulación negativa de la proliferación celular, adhesión celular, formación de vasos sanguíneos, respuesta hormonal. Por otro lado, las proteínas sobreexpresadas (**Figura 8C**) mostraron relación con procesos de proliferación celular, organización del huso mitótico y replicación del DNA.

Otros importantes participantes de estas redes de interacción son los miRNAs, por quienes se produce esta competencia de unión, en este ámbito se detectaron 546 miRNAs participantes, de los cuales, los que participan en una mayor cantidad de interacciones son: hsa-miR-27a-3p (1.819), hsa-miR-16-5p (1.588) y hsa-miR-1-3p (1.419).

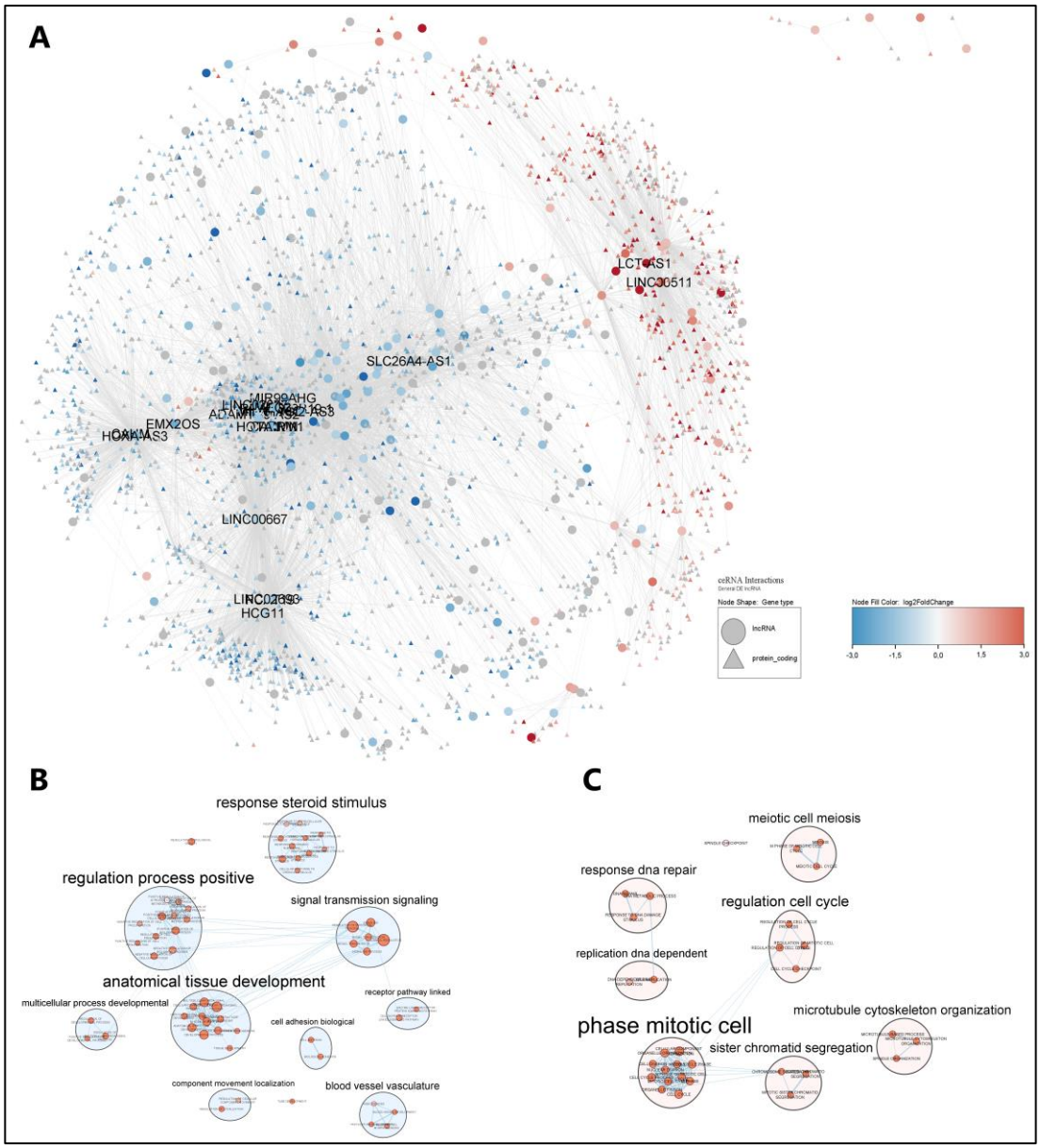

Figura 8. Red de Interacciones ceRNAs general. Red de interacción con lncRNAs expresados diferencialmente en análisis general, y enriquecimiento a nivel de vías biológicas de proteínas involucradas. **A:** Visualización de red en Cytoscape, se muestran proteínas y lncRNAs, colores indican expresión diferencial, subexpresión (azul), sobreexpresión (rojo) y no significativa (gris), se etiquetan lncRNAs con más de 200 interacciones. **B:** Enriquecimiento de vías biológicas en proteínas subexpresadas. **C:** Enriquecimiento de vías biológicas en proteínas sobreexpresadas. FDR < $5*10^{-7}$ (B, C)

4.2.4. Redes de regulación e interacción molecular en el tejido mamario

Se trabajó con la base de datos GTEx para obtener perfiles de expresión normales en tejido mamario. Se estableció un umbral de expresión de 0,035 CPM con un mínimo de presencia en las muestras del 10% (**Figura 9**), con el cual, de los 56.156 genes que contemplaban las librerías de secuenciación de GTEx, se consideraron como expresados de manera relevante 33.871 genes.

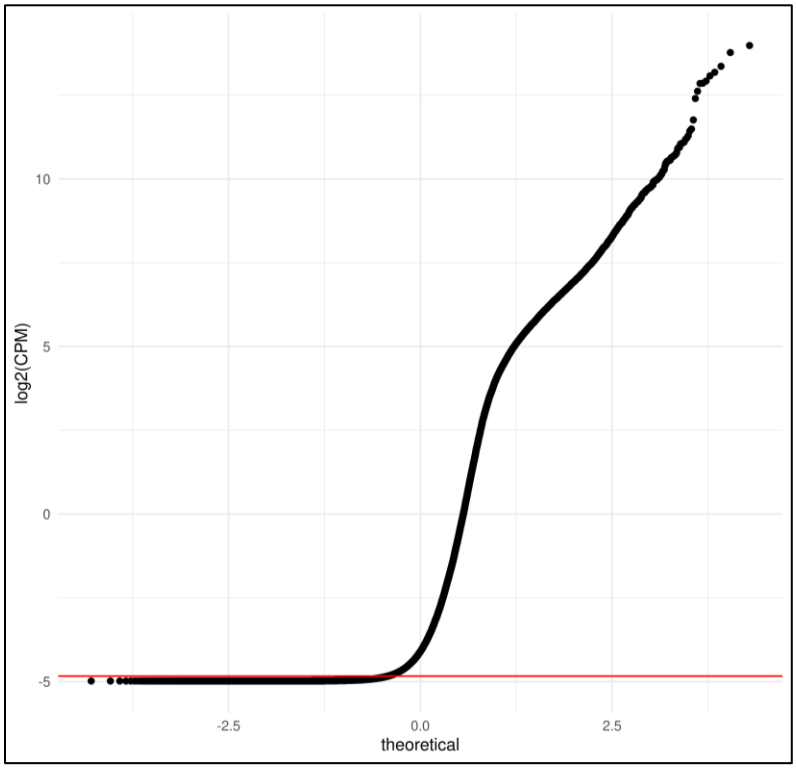

Figura 9. Perfil de expresión promedio genes GTEx. QQ-plot indicando niveles de expresión promedio en log(CPM) de los genes en el conjunto de datos de tejido mamario femenino. Línea roja indica el umbral de CPM aplicado.

La versión 11 de la base de datos STRING para interacción de proteínas en humanos contiene 11.759.455 interacciones, de las cuales se realizó una búsqueda de aquellas que presentaran una puntuación combinada superior a 0,4,

dando cuenta de una confiabilidad intermedia, cumpliendo 1.985.196 interacciones este requisito. Se filtraron, además, según la abundancia en la expresión de cada interactor que componen las interacciones proteína-proteína en los datos provenientes desde GTEx para tejido mamario femenino. Con este filtro, la cantidad de interacciones proteína-proteína que estarían ocurriendo de manera normal alcanza el valor de 1.832.315 interacciones.

La red regulatoria a nivel transcripcional por parte de factores de transcripción en humanos se obtuvo desde la base de datos DoRoThea [70], y arrojó un total de 486.751 interacciones, donde se incluyen regulaciones tanto positivas como inhibitorias. Luego de realizar el filtro de expresión con GTEx, en los factores de transcripción y en los blancos regulatorios asociados a estos, el número de interacciones bajó a 463.170, divididas según grados de confiabilidad en 5 categorías, desde la A hasta la E, considerando E como el grado de confiabilidad más bajo, el que fue descartado por presentar solo regulaciones basadas en inferencia desde datos de expresión utilizando GTEx o por motivos

Grado de confiabilidad	**Cantidad**
Grado A	5.250
Grado B	1.017
Grado C	7.178
Grado D	8.975
TOTAL	22.420

Tabla 2. **Cantidad de regulaciones TF-blanco por grado de confiabilidad**.

de unión en promotores, por lo cual 22.420 regulaciones que pasan los filtros de expresión se consideraron (**Tabla 2**)**.**

Finalmente, la última base de datos pública referente a interacciones moleculares a utilizar fue lncRNAfunc, en donde se pueden recuperar interacciones entre lncRNAs y otras moléculas validadas experimentalmente [71]. Desde este repositorio fuimos capaces de recopilar un número inicial de interacciones de 2.801, que posteriormente al filtro de expresión con GTEx fueron reducidas a 2.758 interacciones.

4.2.5. Módulos de co-expresión de genes en el cáncer de mama

El análisis de co-expresión fue utilizado para encontrar módulos de genes que comparten patrones de expresión similares. El paquete CEMiTool [72] detectó 13 módulos en los datos de expresión de las muestras provenientes de TCGA, los cuales están conformados por 5.273 genes, habiendo 248 genes que no clusterizan, los que son incluidos en el módulo *Not.Correlated.* El detalle de cantidad de genes asociado a cada módulo puede ser observado en la **Tabla 3**. Con los módulos detectados, se generó un análisis de enriquecimiento de conjuntos de genes (GSEA, *Gene Set Enrichment Analysis*) (**Figura 10**), en el cual se muestran los 13 módulos. Este análisis indica los cambios en la actividad transcripcional de estos a través de las muestras, basado en el puntaje de enriquecimiento normalizado (NES).

Modulo	Cantidad de genes
M1	2.053
M2	1.206
M3	696
M4	423
M5	229
M6	148
M7	139
M8	94
M9	81
M10	65
M11	64
M12	40
M13	35
Not.correlated	248

Tabla 3. **Cantidad de genes por módulo de co-expresión detectado**.

En primer lugar, se muestra la comparación de la actividad transcripcional de las muestras normales *versus* el conjunto de las muestras tumorales (**Figura 10A**), apreciándose cambios más notorios, que consisten en un alza en la actividad transcripcional respecto a las muestras control, en los módulos M4 y M6, y algo más leve en los módulos M5, M9 y M10. Se determino por medio de un enriquecimiento funcional de términos GO (**Figura 11**), que los primeros módulos están integrados por genes relacionados con progresión del ciclo celular

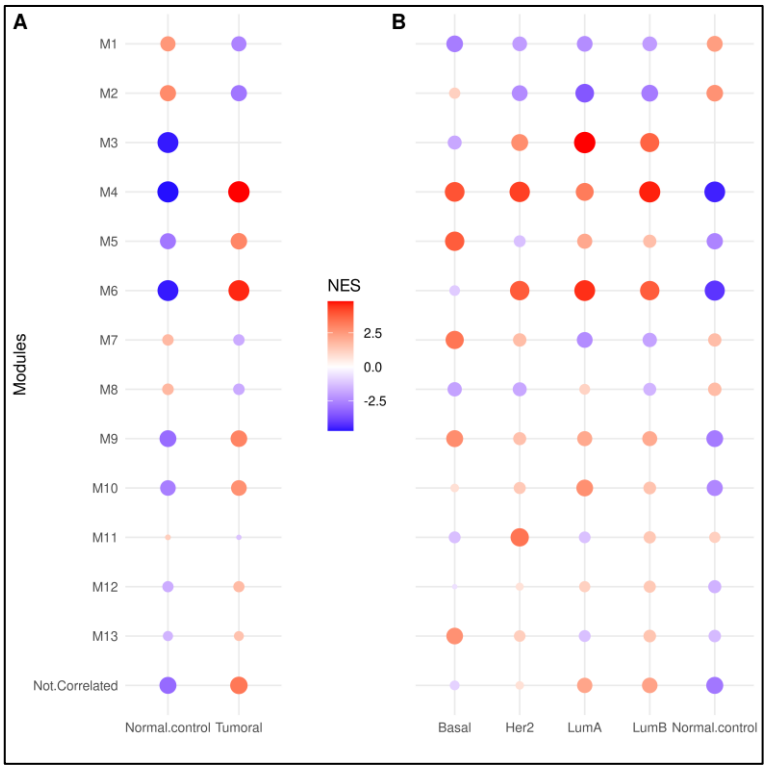

Figura 10. Análisis de enriquecimiento de conjuntos de genes. Mapas de calor en función del puntaje de enriquecimiento normalizado (NES), mostrando actividad de módulos. **A:** mapa de calor perteneciente al análisis de muestras control vs el total de muestras tumorales. **B:** mapa de calor perteneciente al análisis de muestras control vs muestras tumorales agrupadas según subtipo PAM50.

(M4), elementos relacionados con transición epitelio-mesénquima y angiogénesis (M6); mientras que los 3 últimos con vinculación a respuesta inmune. Por otro lado, el análisis de co-expresión se realizó también generando el agrupamiento de las muestras tumorales según la clasificación PAM50 (**Figura 10B**), en donde se aprecian las diferencias en la actividad transcripcional entre los diferentes subtipos. De los módulos de co-expresión antes mencionados, el que presenta la mayor diferencia en cuanto a la actividad entre subtipos tumorales PAM50, es

el módulo M3, relacionado a las respuestas a estrógenos temprana y tardía en el enriquecimiento funcional (**Figura 11**), el cual tiene una alta actividad en los subtipos de carácter hormonal (Luminal A, Luminal B y Her2) y baja actividad en el subtipo Basal. Los módulos 1 y 2 presentan mayor cantidad de elementos, a pesar de que su NES presenta un cambio entre condiciones menor al mostrado por los módulos M4 y M6, el análisis de enriquecimiento funcional los vincula a

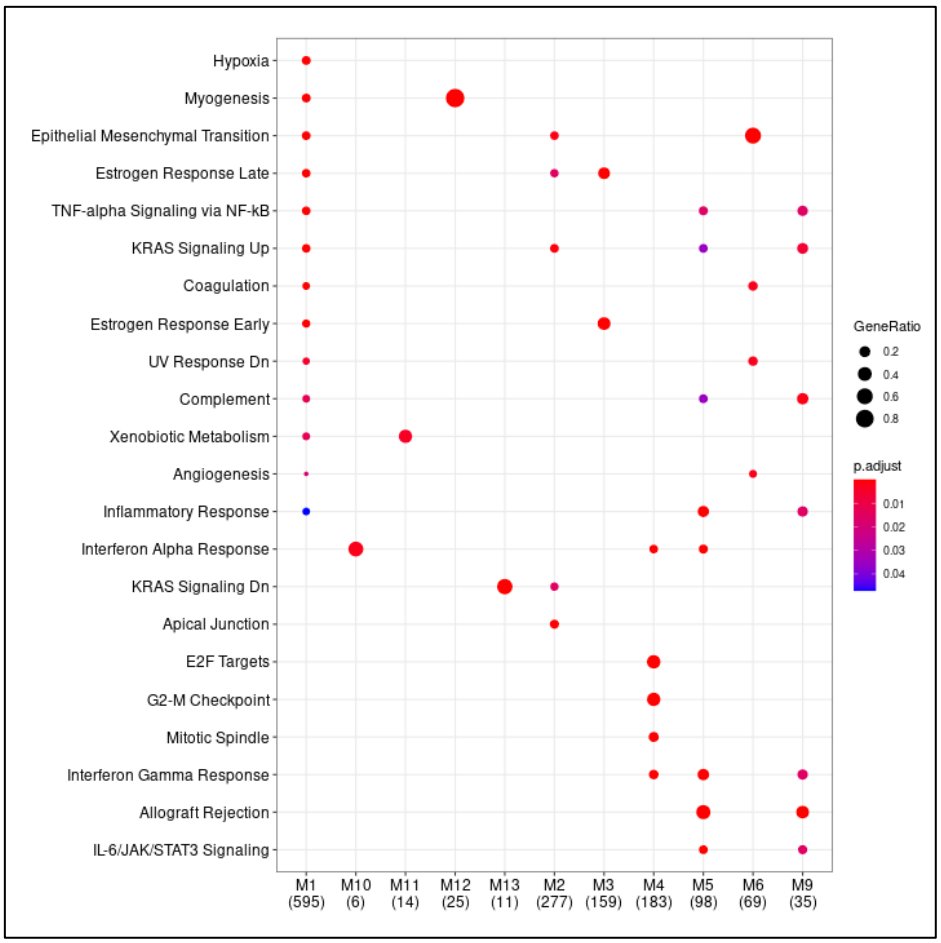

Figura 11. Enriquecimiento términos GO módulos de co-expresión. Significancia de asociación de términos GO a módulos de co-expresión seleccionados y la proporción de cobertura del término GO por los integrantes del módulo. p-value ajustado ≤ 0,05.

procesos como la transición epitelio mesénquima, angiogénesis, hipoxia y genes activados por la sobreexpresión de KRAS (**Figura 11**), un importante regulador y conocido por facilitar la progresión de diversos tipos de cáncer [81].

4.3. Resultados del Objetivo N°3

“Generar un nuevo modelo de puntajes de riesgo poligénico para cáncer de mama, a partir de una estrategia de priorización de genes vecinos dentro de las redes de interacción”.

4.3.1. Generación y prueba de un modelo PRS basado en la teoría omnigénica

El total de interacciones obtenidas destinadas a nutrir los modelos es de 2.592.742 (**Tabla 4**), lo que involucra la participación de 17.835 genes codificantes de proteínas, 1.763 lncRNAs y 562 miRNAs únicos. Estas abarcan 17.715 genes cercanos a loci relacionados a cáncer de mama en GWAS, mientras que, otros 26.552 genes cercanos a estos loci, no tienen representación en las interacciones encontradas, de los cuales 9.252, pertenecen al grupo de genes con expresión relevante, según el umbral establecido para los datos de GTEx, y otros 947 corresponden a miRNAs. El resumen de las interacciones generadas y recopiladas se muestran en la **Tabla 4**.

Tipo de interacción	Fuente	Cantidad	Codificante de proteínas	lncRNAs	miRNAs
Proteína-Proteína	STRING db [93]	1.832.315	17.557	39	-
miRNA-mRNA	Tesis A. Peñaloza [76] – filtrado correlaciones SPONGE	45.369	12.946	1.498	546
ceRNA	Este trabajo - SPONGE	687.132	6.298	365	-
TF-Target	DoRoThea [24]	22.420	7.391	222	2
lncRNA – mRNA/proteínas/DNA	lncRNAfunc [110]	5.506	531	287	19
TOTAL		2.592.742	17.835	1.763	562

Tabla 4. **Resumen interacciones utilizadas en la generación de modelos**.

Se generaron 2 modelos de re-ponderación de tamaño de efecto para las variantes genéticas previamente seleccionadas. El primero, llamado “INT Model”, comprende sólo información de interacciones, mientras que el segundo, “DE Model”, integra priorización por expresión diferencial y los siguientes módulos de co-expresión relevantes: M1, M2, M4 y M6. Se seleccionaron en función de su grado de relación con cáncer en el enriquecimiento funcional realizado y su NES. Para los modelos generados, se evaluó el rendimiento en la población de comparación compuesta de 758 individuos, la cual fue dividida aleatoriamente en 3 subpoblaciones. Los rendimientos obtenidos para los modelos generados se

muestran en la **Tabla 5**, en donde se obtuvo un rendimiento promedio de 0,614 en el primer modelo, que considera sólo las interacciones. Este es levemente superior al registrado por el PRS Base (+0,0058). Por otro lado, el segundo modelo basado en interacción con priorización según los valores de expresión diferencial y módulos importantes de co-expresión, tuvo un rendimiento promedio también levemente superior al registrado por el PRS Base (+0,0056). Ambos rendimientos no son significativamente diferentes del rendimiento base, como se puede apreciar en la **Figura 12**, utilizando los valores de AUC de la totalidad de la población de comparación. Los *odd ratios* del percentil 5% superior de PRS en la población de comparación subieron de 2,943 a 3,406 en el primer modelo y

Modelo	**Población**	**AUC**
PRS Base	CGEMS – Subpob 1	0,6152640
	CGEMS – Subpob 2	0,5816825
	CGEMS – Subpob 3	0,6278964
	Promedio	**0,608281**
INT Model	CGEMS – Subpob 1	0,6104354
	CGEMS – Subpob 2	0,6015180
	CGEMS – Subpob 3	0,6303866
	Promedio	**0,6141133**
DE Model	CGEMS – Subpob 1	0,6149633
	CGEMS – Subpob 2	0,5937446
	CGEMS – Subpob 3	0,6330009
	Promedio	**0,6139029**

Tabla 5. **Rendimientos de los modelos en las 3 subpoblaciones de comparación**.

hasta 3,986 en el segundo, esto dado que en este 5% superior el número de casos subió a 29 y 30 respectivamente, desde 28 en el PRS Base.

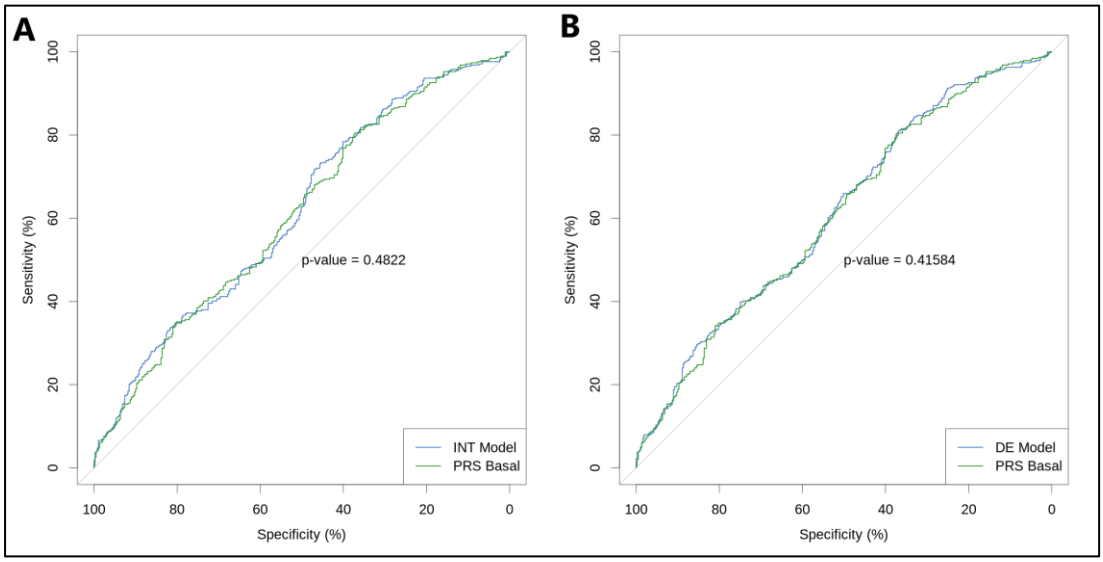

Figura 12. Comparación modelos INT Model y DE Model. Gráfica de curvas ROC, según su puntaje de riesgo poligénico individual. Método “delong”, paquete de R pROC. **A:** INT Model, rendimiento AUC = 0,61396. **B:** DE Model, rendimiento AUC = 0,6144. N=758.

5. DISCUSION

5.1. Rendimientos PRS

El presente trabajo de investigación busca una aproximación a la importancia que pueden tener las redes de interacción que presentan los elementos próximos a los *loci* de las variantes genéticas asociadas al riesgo de padecer con una enfermedad, en este caso, cáncer de mama. Revisando los resultados presentados en el objetivo 1 de esta tesis, el rendimiento obtenido por los mejores modelos de LDPred2 sobrepasa por poco los 0,6 de área bajo la curva ROC, rendimientos más bajos que lo previamente informado por Khera et al, en 2018 para cáncer de mama (AUC = 0,68) [29]. Esto se puede explicar en parte por la diferencia en calidad de los conjuntos de datos utilizados, ya que para la publicación antes mencionada se utilizaron datos desde UK Biobank [82], los cuales presentan genotipos de sobre 400.000 individuos, favoreciendo poblaciones de validación y entrenamiento de mayor tamaño, lo que produce resultados más confiables. En adición a la diferencia en el tamaño muestral, los individuos en UK Biobank fueron genotipados con principalmente 2 microarreglos de sobre 807.000 marcadores genéticos (UK BiLEVE Axiom Array y UK Biobank Axiom Array), lo que también permite un mejor rendimiento en la posterior imputación de variantes genéticas, ya que esto aumenta la densidad de los SNPs que cubren cada cromosoma [83], frente a los 550.000 marcadores genéticos que detectaban los microarreglos en CGEMS (HumanHap550 de Illumina), el

estudio empleado para generar el PRS en este trabajo investigativo. Implementamos este conjunto de datos como alternativa a UK Biobank, debido a los costes y tiempo necesario para su adjudicación, pero sería ideal refinar los modelos propuestos en conjuntos de datos con una población de mayor tamaño, como un paso a futuro.

5.2. Redes de Interacción y priorización de genes

En el área de priorización de genes, uno de los primeros resultados son los perfiles de expresión, en el análisis de reducción de dimensionalidad con componentes principales, destaca la dispersión de las muestras normales en la componente 2, esto entre otras cosas podría deberse a la manipulación de las muestras en la fase preanalítica de la secuenciación. Cabe mencionar que el proyecto del cual proceden dichas muestras es de carácter internacional, por lo cual hay diferentes centros de obtención de muestra, así como centros de secuenciación que colaboraron para generar el proyecto TCGA-BRCA, de modo que la manipulación humana y el estado de los equipos utilizados pueden tener cierto grado de impacto, desafortunadamente. Además, a diferencia de otros proyectos en TCGA, como de cáncer gástrico, en que se indicaba la ciudad o país de adquisición [54], para las muestras de cáncer de mama estos datos no están informados, lo que nos impide generar una apropiada corrección del efecto *batch*, el cual a menudo no puede ser corregido en su totalidad por los métodos de normalización [84]. Otro factor para considerar en este ámbito es que las muestras normales consideradas en TCGA son adyacentes al tumor, extraídas a

más de 2 cm y/o no debe contener tumor según una revisión histopatológica. No obstante, esto no descarta que haya perturbaciones por los microambientes que generan los tumores a nivel molecular [85].

Las diferencias en cuanto a la expresión diferencial entre subtipos obtenidas concuerdan con lo esperado, dado que en general el subtipo PAM50 Basal fue el que presentó mayor cantidad de elementos sobreexpresados únicos. Este subtipo tiene relación con los tumores triple negativo [86], que no son susceptibles a terapia hormonal, lo que los hace clínicamente de peor prognosis, mientras que para los otros 3 subtipos PAM50, presentan susceptibilidad a terapia hormonal, por utilizar para el desarrollo tumoral uno o más de los receptores ER, PR o Her2, esperándose que presentaran más elementos en común.

Dado el alto número de genes involucrados en los resultados de expresión diferencial, co-expresión y redes de interacción, la discusión de estos estará enfocada en aquellos elementos que se destacan en sus respectivos análisis. Entre los genes con mayor nivel de expresión diferencial, como se aprecia en la **Figura 5**, se detectó entre los miRNAs de mayor sobreexpresión a mir-592, mir-1269a y mir-449a. Este último incluso presentó 67 interacciones verificadas por correlación negativa en SPONGE, sugiriendo una participación importante posiblemente regulada por RNAs de competencia endógena. A pesar de que este miRNA está asociado a cáncer de mama en literatura, se detectaron puntos de vista contrarios en lo que se respecta a su función en este cáncer. Por un lado,

se menciona que funcionaria como un supresor de la migración y la invasión, a través de PLAGL2 (no encontrada SPONGE) [87]; mientras que por otro, se describe que promueve la progresión del cáncer a través de CRIP2 (también no encontrada SPONGE) [88], a su vez la visión de expresión diferencial de este miRNA, también es contraria en esas publicaciones, existiendo discrepancia con los datos generados en esta tesis respecto a la primera publicación [87], la que indica subexpresión de mir-449a en cáncer. Esto podría deberse al subtipo de cáncer mamario usado en aquella publicación, el cual no es especificado, mientras que, en los datos generados en este trabajo a nivel de subtipo, solo se encontró una reducción relacionada al subtipo PAM50 Basal (Log2FC = -1,298), la cual no fue significativa (p-value ajustado = 0,286) para las 15 muestras de este subtipo. Por otro lado, los subtipos de carácter hormonal presentaron una sobreexpresión significativa (p-value ajustado < 4,12e-5) sobre 4,3 Log2FC. Entre los miRNAs subexpresados, podemos destacar a mir-4732 y mir-486. El primero no presenta interacciones verificadas por correlación negativa en SPONGE, pero ha sido estudiado previamente en cáncer de mama, donde en general está subexpresado, pero han detectado una expresión mayor en tejidos positivos para metástasis de nodo linfático, sugiriendo, que cambia de un rol supresor en etapas tempranas a un rol promotor en etapas tardías de cáncer de mama [89]. El segundo presentó 52 interacciones respaldadas por SPONGE, 20 de estas por su hebra 5p y 30 por la hebra 3p; ambos han sido ampliamente estudiados en varios cánceres, en específico para cáncer de mama se ha visto

que es un importante supresor y que además mejora el reconocimiento por el sistema inmune de células cancerígenas [90].

Entre los datos de redes de competencia endógena se resaltaron 19 lncRNAs que tenían una alta conectividad en la red general (**Figura 8**, **Tabla 6**), entre ellos solo dos se encuentran sobreexpresados, y corresponden a *LCT-AS1* y *LINC00511.* Estos interaccionan como ceRNA con una alta cantidad de genes codificantes de proteínas que también se encuentran sobreexpresados y, a su vez, presentan este mismo tipo de interacción entre sí, compartiendo principalmente dos miRNAs: hsa-miR-497-5p y hsa-miR-195-5p, los que comprenden la totalidad de los miRNAs respaldados por SPONGE para *LCT-AS1* y dos de los seis miRNAs que interactúan con *LINC00511.* Además, estos lncRNAs están presentes en cientos de interacciones miRNA-target respaldadas por SPONGE, 176 para hsa-miR-497-5p y 365 para hsa-miR-195-5p, estando ambos ya relacionados a cáncer en literatura. A diferencia de *LINC00511*, el cual aparece en diversas publicaciones asociadas a cáncer [91][92][93]. Además, a través de una búsqueda en PubMed y Google Scholar por *LCT-AS1* no se encontraron publicaciones que relacionen a este lncRNA con cáncer, existiendo una publicación en que este es mencionado como resultado de análisis en el área de la inmunología [94]. Por otro lado, este lncRNA solo presenta una asociación en la base de datos de GWAS Catalog, que no está directamente relacionada a cáncer, sino que al conteo de células dendríticas plasmacitoides. Otros lncRNAs de alta conectividad que están subexpresados y de los que además presentan

escasas publicaciones que los involucran son RP11-679B19.1, *LINC02693* y

Gen	Cáncer de mama (PMID)	Otros tipos de cáncer (PMID)	Rasgos asociados a cáncer GWAS	Otros rasgos asociados en GWAS
SLC26A4-AS1	34880202	32939012	0	2
RP11-679B19.1	---	---	0	0
MAGI2-AS3	32730644	34026068	0	3
LINC02693	---	---	0	1
MIR99AHG	---	32874129	2	19
HOTAIRM1	32284737	33650656; 31853186	0	4
LINC02607	---	---	0	26
CARMN	34162418	32305636	0	5
EMX2OS	34703931	32273754	1 (CM)	23
HCG11	33739352	33215418	0	4
LCT-AS1	---	---	0	1
LINC00511	30482236	32042282	5 (CM)	20
HOXA-AS3	---	34659534	0	10
PCAT19	---	31819778	1	2
LINC00667	31897133	34313922	0	1
MIR100HG	33088216	30886062	0	32
MEG3	33845141	28975980	1 (CM)	10
CAHM	32190687	24799664	0	14
ADAMTS9-AS2	34178640	32516127	1 (CM)	50

Tabla 6. **lncRNAs con alta conectividad en red ceRNA general, ejemplos de caracterización en cáncer disponibles en literatura, y asociaciones en GWAS.**

LINC02607. El primero no presenta asociación en GWAS catalog, mientras que el segundo presenta una asociación relacionada al conteo de proteínas en hígado; y el último con varias asociaciones (26), las cuales abarcan entre otros al índice de masa corporal, conteo de neutrófilos, desarrollo cognitivo.

En el análisis de co-expresión realizado se eligieron 4 módulos para bonificar en el segundo modelo (DE Model): M1, M2, M4 y M6, dado que guardan mayor relación al proceso cancerígeno. Otros módulos que también estaban enriquecidos en procesos biológicos relacionados a cáncer, como M5 y M9, guardaban relación con procesos propios de células inmunitarias, por lo que podría tener procedencia desde células inmunes presentes en los tejidos tumorales extraídos por parte del proyecto TCGA-BRCA, evidencia de esto es recopilada desde los genes *Hubs* de co-expresión mostrados en la **Fig. Anexo 2**, en donde para el módulo M9 se aprecian genes codificantes de proteínas, que codifican para el receptor LILRB4, el cual interactúa con el MHC1 [95]; CD53, proteína involucrada en transducción de señales en linfocitos T y células NK [96]; y finalmente NCKAP1L, que tiene una expresión específica en células hematopoyéticas [97]. Para el módulo M5 se observan genes codificantes para CD2, el cual es un marcador de linfocitos T de sangre periférica [98]; SLAMF6, expresada en células NK y linfocitos B y T [99]; IL2RG, componente importante de los receptores de interleucinas [100]; CD3D, involucrado en el desarrollo de células T [101]; y SLA2, rol en la regulación de respuesta mediada por linfocitos B y T [102].

5.3. Fortalezas, limitaciones y futuras perspectivas del estudio

Sobre este abordaje de mejora de PRS, la idea de usar elementos de transcriptómica en el ajuste de PRS fue ejecutado en la tesis “Refining Polygenic Risk Score Models Through Fine Mapping and Functional Gene Modules” [103], en la cual se utilizó *fine mapping* y la generación de módulos de expresión, pero cuyo enfoque fue para reducir el número de variantes genéticas y usar p-values de asociación re-calculados en base al fine-mapping, manteniendo el efecto de las variantes genéticas calculado en GWAS, obteniendo así, rendimientos AUC menores (AUC = 0,6176) al PRS base del GWAS original (AUC = 0,6656) en enfermedad inflamatoria intestinal. A diferencia de lo antes mencionado, en este trabajo de investigación, se obtuvieron rendimientos levemente superiores al PRS base, lo que sugiere que la integración de interacciones presenta un efecto positivo en la reponderación de los tamaños de efecto de las variantes genéticas asociadas a cáncer mamario, sin embargo, dado que estos aumentos no son significativos, aún queda trabajo que hacer en cuanto a la manera en que estos son integrados y formas en las cuales refinar estos puntajes. En este ámbito, hay publicaciones que hacen uso de datos de transcriptómica para refinar los PRS [104], aplicación de multi-omicas en análisis de GWAS [105], así como integración de vías biológicas por enriquecimiento en PRS [106], por lo que la premisa de esta investigación y sus métodos son de interés en el área de los puntajes de riesgo poligénico y su futura aplicación en el área clínica, una vez estos alcancen un alto grado de confiabilidad. Por otro lado, trabajar con

interacciones y la generación de estas posibilita encontrar blancos novedosos, como los antes mencionados lncRNAs, que no se han descrito para cáncer o cáncer de mama, lo que forma parte a su vez de los aspectos a mejorar de este trabajo de investigación, como la aplicación de validaciones experimentales, sobre todo en esta área más novedosa de las interacciones generadas, las redes de competencia endógena, área en que se pueden encontrar genes con potencial farmacológico o como marcadores, para apoyar diagnósticos o hacer seguimientos.

Uno de los elementos que dificulta la integración ideal de las redes de interacción, es el sesgo natural que ocurre con el descubrimiento de estas, generalmente los genes de mayor importancia para ciertas patologías o procesos fisiológicos están mejor descritos y presentan un número mayor de interacciones conocidas, como es la diferencia en cuanto a conocimiento entre proteínas y lncRNAs. Estos últimos y otros ncRNAs han atraído más atención en el último tiempo, por lo que actualmente las proteínas tienen una base de información más extensa, como ejemplo el lncRNA *LINC00052*, a pesar de ser uno de los top 3 en cuanto a sobreexpresión (**Figura 5**), no presenta información de interacciones validadas en bases de datos, como por ejemplo en la recientemente publicada base de datos de lncRNAfunc, utilizada en este trabajo, y que además tiene foco en la información sobre cáncer aportada por TCGA [71].

La falta de acceso a datos de transcriptómica a nivel de secuencias es otro de los factores limitantes, una vez que estos permitirían verificar la presencia de

las variantes genéticas seleccionadas en el PRS en los transcritos, para cuantificar los cambios que pueden producir en la expresión de manera específica, en vez de solo hacer un trabajo asociativo con los cambios en la expresión genética en cáncer de mama de manera general. A su vez, esto también implica que los datos dependen del conjunto de genes que se haya empleado por los investigadores, ya sea del proyecto TCGA o GTEx, al momento de realizar el conteo de lecturas posterior al mapeo, generando pérdida de información.

Por otro lado, las nuevas tecnologías que se han expandido en los últimos años aumentan las opciones de mejorar los modelos haciéndolos más precisos. Como ejemplo de esto tenemos las tecnologías de *single cell sequencing*, o secuenciación de célula individual, la cual podría entregarnos perfiles de expresión de tejido ya sea normal o tumoral, más precisos, al ser capaces de diferenciar con mayor resolución la fuente de los RNAs, pudiendo clasificar según tipo celular y, de esta forma, filtrar células que no sean de interés, removiendo variabilidad introducida en la fase preanalítica dada por las diferentes poblaciones celulares que pueden ser capturadas desde un tejido como muestra. Esta heterogeneidad en las poblaciones celulares puede diluir señales relacionadas a tumorigenesis u otros procesos importantes en la investigación [107]. El proyecto GTEx se actualizó a mediados de 2021 a la versión 9, que contiene datos de *single cell* para múltiples tejidos, entre ellos tejido mamario [108]. También ha habido avances en nuevas anotaciones de regiones con

importancia en regulación, que tienen efecto en el desarrollo de enfermedades o rasgos complejos. Ejemplo de esto es una reciente publicación, en que desarrollan un mapa de restricción mutacional que abarca regiones no codificantes, basado en datos de 76.156 genomas humanos [109]. Previamente, este mismo proyecto (denominado gnomAD) realizó esta tarea con regiones codificantes utilizando datos de exomas y genomas, encontrando que las regiones codificantes se encuentran más restringidas que las no codificantes [110], y a su vez aquellas regiones no codificantes, que se encuentran más restringidas, se enriquecen con elementos reguladores y variantes genéticas implicadas en enfermedades y rasgos complejos [109]. Parte de esto ha sido posible por avances en el proyecto ENCODE, el cual ha añadido nuevas características en base a la accesibilidad del DNA y las modificaciones a la cromatina, generando un registro de elementos cis-reguladores candidatos (cCREs, *candidate cis-regulatory elements*) [111]. Finalmente, todos estos avances tienen potencial de mejorar el rendimiento y generar un mejor sustento a la hipótesis planteada, pero a su vez, complejizan la generación de los modelos al aumentar las variables. De esta manera, lo más óptimo probablemente sea aplicar un enfoque de *machine learning*, con la respectiva “caja negra” en el modelado que implican [112], si se tiene la cantidad suficiente de individuos para entrenar y validar los modelos generados.

6. CONCLUSIONES

- Se confirmó que hay un sesgo de conocimiento en cuanto a interacciones a favor de los genes codificantes en desmedro de los no codificantes.
- Se detectaron lcnRNAs de alta conectividad en redes de ceRNAs (por ejemplo: *LCT-AS1*, *LINC02607* y *LINC02693*), que pueden ser importantes para validar y caracterizar en cáncer de mama.
- Ambos modelos generados presentaron un leve aumento de rendimiento respecto al PRS base, no significativos, pero marcan una tendencia al aumento en la población de comparación total.
- La tendencia al aumento sugiere que existe margen de mejora en la reponderación de los tamaños de efecto entre genes con evidencia de participación en el PRS.
- Se detectaron 13 módulos de co-expresión, de los cuales 4 presentan la mayor relación a cáncer a nivel de tejido mamario.
- Es necesario emplear conjuntos de datos de mayor tamaño, para evaluar con mejor resolución los rendimientos de los modelos generados.

Finalmente se concluye que a pesar de presentar una tendencia de mejora en el rendimiento del PRS, los modelos generados no logran sustentar la hipótesis presentada: “La integración de redes de interacción molecular, generadas en base a variantes genéticas asociadas a cáncer de mama, mejoran el rendimiento de puntajes de riesgo poligénico en la predicción de riesgo de cáncer de mama”.

7. ANEXOS

ANEXO 1: Complementario preparación de datos TCGA

Muestras removidas provenientes de TCGA, por presentar patrones de expresión demasiado alejados de su condición (normal o tumoral) en el análisis de miRNA-seq normalizado con DESeq2, son mostradas en la **Fig. anexo 1**.

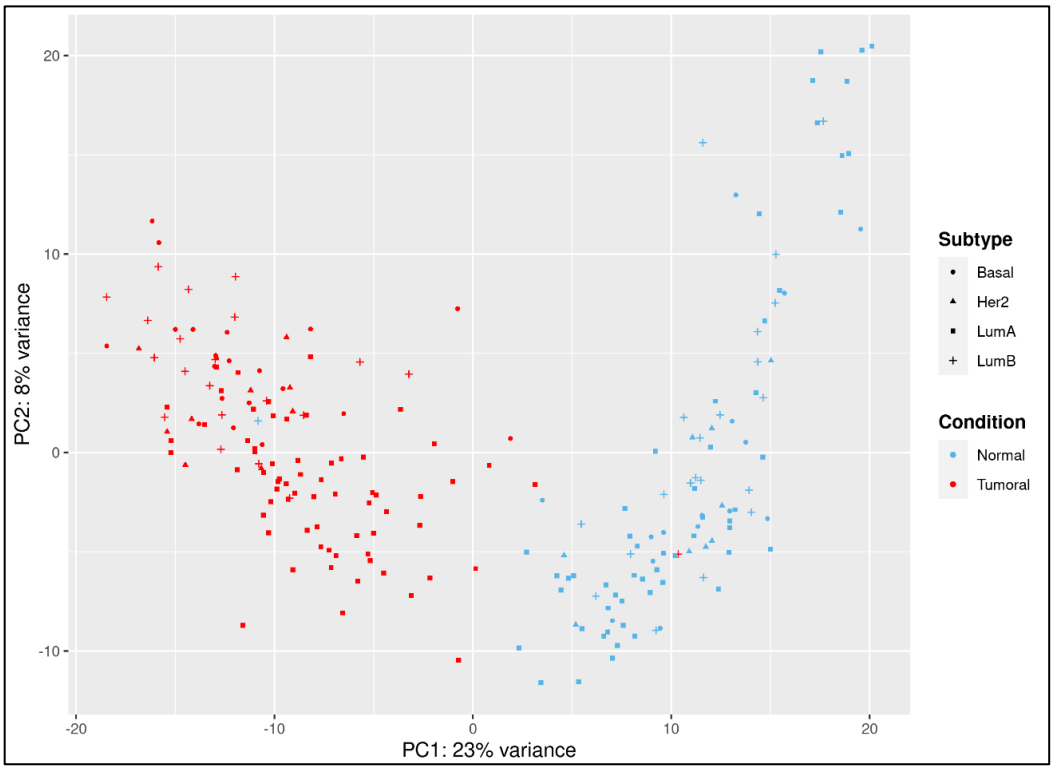

Fig. Anexo 1. Análisis de componentes principales miRNA-seq inicial. Graficas con las 2 primeras componentes principales previa a la remoción de 2 individuos. Datos de miRNA-seq, componentes resumen el 31% de la varianza. N=198 muestras.

ANEXO 2: Complementario resultados redes de co-expresión

Genes con alta conectividad en los principales módulos de co-expresión detectados, en relación con el proceso tumoral, **Fig. Anexo 2**.

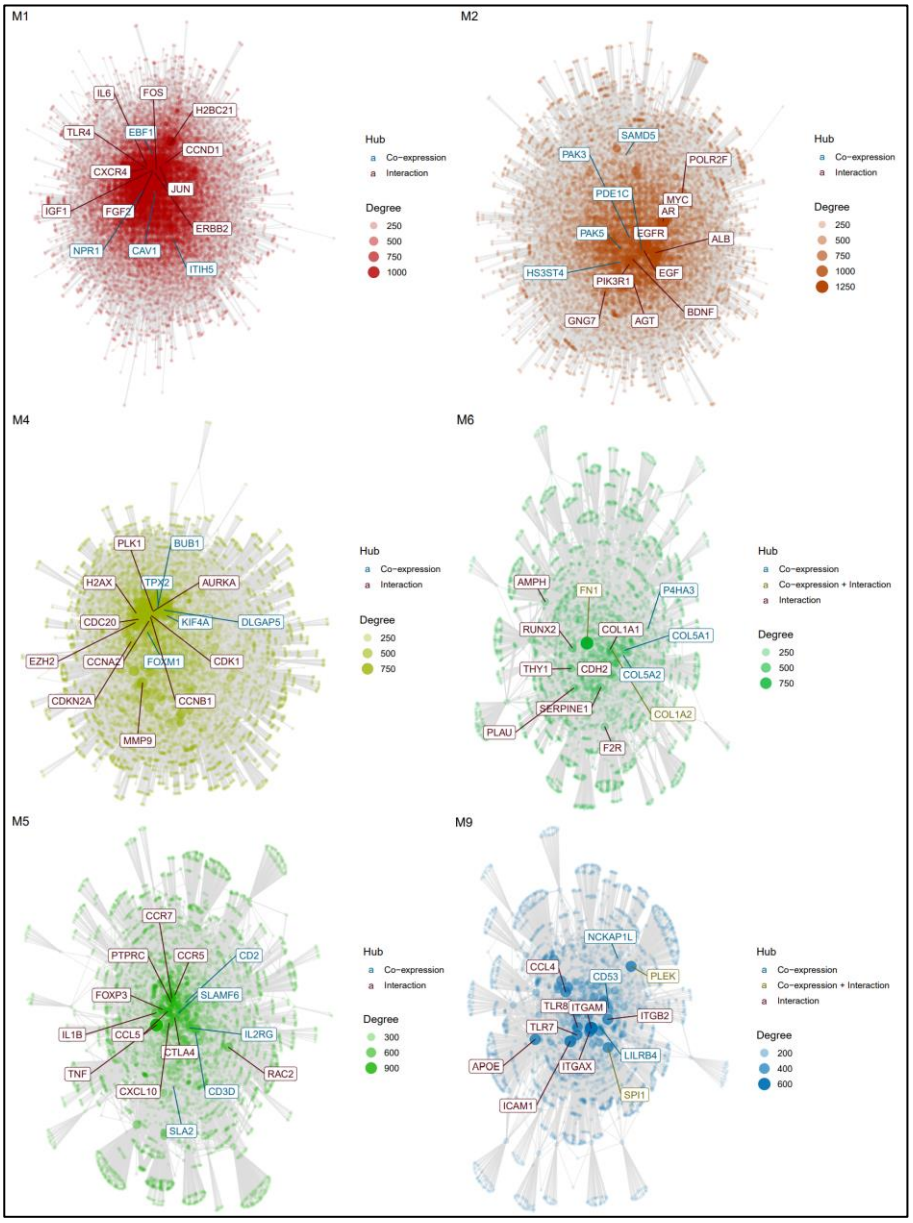

Fig. Anexo 2. Módulos de co-expresión enriquecidos con interacciones proteína-proteína. Se muestran los módulos con relación al proceso carcinogénico identificados en el enriquecimiento de proceso biológicos (M1, M2, M4, M5, M6, M9).

8. BIBLIOGRAFÍA

1. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: a cancer journal for clinicians, 71(3), 209–249. doi:10.3322/caac.21660
2. Ferlay, J., Ervik, M., Lam, F., Colombet, M., Mery, L., Piñeros, M., . . . Bray, F. (2020). International Agency for Research on Cancer. Retrieved January 20, 2022, from Global Cancer Observatory: Cancer Today: https://gco.iarc.fr/today
3. Sun, Y. S., Zhao, Z., Yang, Z. N., Xu, F., Lu, H. J., Zhu, Z. Y., . . . Zhu., H. P. (2017). Risk Factors and Preventions of Breast Cancer. International Journal of Biological Sciences, 13(11), 1387–1397. doi:10.7150/ijbs.21635
4. Valastyan, S., & Weinberg, R. A. (2011). Tumor metastasis: molecular insights and evolving paradigms. Cell, 147(2), 275-292. doi:10.1016/j.cell.2011.09.024
5. Rivera-Franco, M. M., & Leon-Rodriguez, E. (2018). Delays in Breast Cancer Detection and Treatment in Developing Countries. Breast cancer: basic and clinical research, 12, 1178223417752677. doi:10.1177/1178223417752677
6. Lee, T. C., Reyna, C., Shaughnessy, E., & Lewis, J. D. (2019). Screening of populations at high risk for breast cancer. Journal of Surgical Oncology, 120 (5), 820-830. doi:10.1002/jso.25611
7. Leon-Ferre, R. A., Giridhar, K. V., Hieken, T. J., Mutter, R. W., Couch, F. J., Jimenez, R. E., . . . Ruddy, K. J. (2018). A contemporary review of male breast cancer: current evidence and unanswered questions. Cancer and Metastasis Reviews, 37(4), 599-614. doi:10.1007/s10555-018-9761-x
8. Icaza, G., Núñez, L., & Bugueño, H. (2017). Epidemiological analysis of breast cancer mortality in women in Chile. Revista médica de Chile, 145(1), 106-114. doi:10.4067/S0034-98872017000100014
9. Takeshima, H., & Ushijima, T. (2019). Accumulation of genetic and epigenetic alterations in normal cells and cancer risk. NPJ precision oncology, 3, 7. doi:10.1038/s41698-019-0079-0
10. Brewer, H. R., Jones, M. E., Schoemaker, M. J., Ashworth, A., & Swerdlow, A. J. (2017). Family history and risk of breast cancer: an analysis accounting for family structure. Breast cancer research and treatment, 165(1), 193-200. doi:10.1007/s10549-017-4325-2
11. Rashid, M. U., Muhammad, N., Naeemi, H., Khan, F. A., Hassan, M., Faisal, S., . . . Hamann, U. (2019). Spectrum and prevalence of BRCA1/2 germline mutations in Pakistani breast cancer patients: results from a large comprehensive study. Hereditary cancer in clinical practice, 17, 27. doi:10.1186/s13053-019-0125-5
12. Parsa, P., & Parsa, B. (2009). Effects of Reproductive Factors on Risk of Breast Cancer: A Literature Review. Asian Pacific Journal of Cancer Prevention, 10(4), 545-550.
13. Cleator, S., Heller, W., & Coombes, R. C. (2007). Triple-negative breast cancer: therapeutic options. The Lancet Oncology, 8(3), 235–244. doi:10.1016/s1470-2045(07)70074-8
14. Pu, M., Messer, K., Davies, S. R., Vickery, T. L., Pittman, E., Parker, B. A., . . . Natarajan, L. (2020). Research-based PAM50 signature and long-term breast cancer survival. Breast cancer research and treatment, 179(1), 197–206. doi:10.1007/s10549-019-05446-y
15. Cline, M. S., Craft, B., Swatloski, T., Goldman, M., Ma, S., Haussler, D., & Zhu, J. (2013). Exploring TCGA Pan-Cancer data at the UCSC Cancer Genomics Browser. Scientific reports, 3, 2652. doi:10.1038/srep02652
16. Mahdavi, M., Nassiri, M., Kooshyar, M. M., Vakili-Azghandi, M., Avan, A., Sandry, R., . . . Gopalan, V. (2018). Hereditary breast cancer; Genetic penetrance and current status with BRCA. Journal of cellular physiology, 234(5), 5741-5750. doi:10.1002/jcp.27464

17. Szumilas, M. (2010). Explaining odds ratios. Journal of the Canadian Academy of Child and Adolescent Psychiatry, 19(3), 227-229.

18. Ledwoń, J. K., Hennig, E. E., Maryan, N., Goryca, K., Nowakowska, D., Niwińska, A., & Ostrowski, J. (2013). Common low-penetrance risk variants associated with breast cancer in Polish women. BMC cancer, 13, 510. doi:10.1186/1471-2407-13-510

19. MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., . . . Parkinson, H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog. Nucleic acids research, 45(Database issue), D896-D901.

20. National Cancer Institute. (n.d.). Definition of LD. Retrieved January 10, 2022, from NCI Dictionary of Genetics Terms : https://www.cancer.gov/publications/dictionaries/genetics-dictionary/def/ld

21. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., . . . Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic acids research, 42(Database issue), D1001-D1006. doi:10.1093/nar/gkt1229

22. Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., . . . Kenny, E. E. (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. American journal of human genetics, 100(4), 635-649. doi:10.1016/j.ajhg.2017.03.004

23. Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., . . . Visscher, P. M. (2009). Finding the missing heritability of complex diseases. Nature, 461(7265), 747-753. doi:10.1038/nature08494

24. Buniello, A., MacArthur, J., Cerezo, M., Harris, L., Hayhurst, J., Malangone, C., . . . Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Research, 47(Database issue), D1005-D1012. doi:10.1093/nar/gky1120

25. Sugrue, L. P., & Desikan, R. S. (2019). What Are Polygenic Scores and Why Are They Important? JAMA, 321(18), 1820-1821. doi:10.1001/jama.2019.3893

26. Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., & Derks, E. M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. International journal of methods in psychiatric research, 27(2), e1608. doi:10.1002/mpr.1608

27. Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., . . . Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. Nature communications, 10(1), 3328. doi:10.1038/s41467-019-11112-0

28. Lello, L., Raben, T. G., Yong, S. Y., Tellier, L. C., & Hsu, S. D. (2019). Genomic Prediction of 16 Complex Disease Risks Including Heart Attack, Diabetes, Breast and Prostate Cancer. Scientific reports, 9(1), 15286. doi:10.1038/s41598-019-51258-x

29. Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., . . . Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nature genetics, 50 (9), 1219-1224. doi:10.1038/s41588-018-0183-z

30. Khera, A. V., Chaffin, M., Zekavat, S. M., Collins, R. L., Roselli, C., Natarajan, P., . . . Kathiresan, S. (2019). Whole-Genome Sequencing to Characterize Monogenic and Polygenic Contributions in Patients Hospitalized With Early-Onset Myocardial Infarction. Circulation, 139(13), 1593-1602. doi:10.1161/CIRCULATIONAHA.118.035658

31. Li, J., Pan, C., Zhang, S., Spin, J. M., Deng, A., Leung, L. L., . . . Snyder, M. (2018). Decoding the Genomics of Abdominal Aortic Aneurysm. Cell, 174(6), 1361-1372. doi:10.1016/j.cell.2018.07.021

32. Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell, 169(7), 1177-1186. doi:10.1016/j.cell.2017.05.038

33. Stuart, J. M., Segal, E., Koller, D., & Kim, S. K. (2003). A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science, 302(5643), 249-255. doi:10.1126/science.1087447

34. Pan-UKB team. (2020). Retrieved February 23, 2021, from Pan-UK Biobank Website: https://pan.ukbb.broadinstitute.org

35. Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., . . . Sherry, S. T. (2007). The NCBI dbGaP database of genotypes and phenotypes. Nature genetics, 39(10), 1181-1186. doi:10.1038/ng1007-1181

36. Hunter, D. J., Kraft, P., Jacobs, K. B., Cox, D. G., Yeager, M., Hankinson, S. E., . . . Chanock, S. J. (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of. Nature genetics, 39(7), 870-874. doi:10.1038/ng2075

37. Guo, Q., Schmidt, M. K., Kraft, P., Canisius, S., Chen, C., Khan, S., . . . Pharoah, P. D. (2015). Identification of novel genetic markers of breast cancer survival. Journal of the National Cancer Institute, 107(5), djv081. doi:10.1093/jnci/djv081

38. Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes. Genetic epidemiology, 34(8), 816–834. doi:10.1002/gepi.20533

39. Fuchsberger, C., Abecasis, G. R., & Hinds, D. A. (2015). minimac2: faster genotype imputation. Bioinformatics, 31(5), 782–784. doi:10.1093/bioinformatics/btu704

40. Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. American journal of human genetics, 88(1), 76-82. doi:10.1016/j.ajhg.2010.11.011

41. Lindström, S., Loomis, S., Turman, C., Huang, H., Huang, J., Aschard, H., . . . Kraft, P. (2017). A comprehensive survey of genetic variation in 20,691 subjects from four large cohorts. PLoS One, 12(3), e0173997. doi:10.1371/journal.pone.0173997

42. Privé, F., Arbel, J., & Vilhjálmsson, B. J. (2020). LDpred2: better, faster, stronger. Bioinformatics. doi:10.1093/bioinformatics/btaa1029

43. Fisher, R. (1918). The Correlation between Relatives on the Supposition of Mendelian Inheritance. Transactions of the Royal Society of Edinburgh, 52, 399-433.

44. Vilhja´lmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindstro¨m, S., Ripke, S., . . . Price, A. L. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. American journal of human genetics, 97(4), 576-592. doi:10.1016/j.ajhg.2015.09.001

45. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics, 12, 77. doi:10.1186/1471-2105-12-77

46. DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, 44(3), 837–845.

47. McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome biology, 17 (1), 122. doi:10.1186/s13059-016-0974-4

48. The GTEx Consortium. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science, 369(6509), 1318-1330. doi:10.1126/science.aaz1776

49. Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., García Girón, C., … Flicek, P. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic acids research, 47 (Database issue), D766–D773. doi:10.1093/nar/gky955

50. The Cancer Genome Atlas Research Network. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature, 455(7216), 1061-1068. doi:10.1038/nature07385
51. Colaprico, A., Silva, T., Olsen, C., Garofano, L., Cava, C., Garolini, D., . . . Noushmehr, H. (2015). TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Research., 44(8), e71. doi:10.1093/nar/gkv1507
52. Xu, T., Su, N., Liu, L., Zhang, J., Wang, H., Zhang, W., . . . Le, T. D. (2018). miRBaseConverter: an R/Bioconductor package for converting and retrieving miRNA name, accession, sequence and family information in different versions of miRBase. BMC bioinformatics,, 19, 514. doi:10.1186/s12859-018-2531-5
53. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology, 15 (12), 550. doi:10.1186/s13059-014-0550-8
54. Wichmann, I. (2020). Identification of long noncoding RNAs in competing endogenous RNA networks through out the gastric precancerous cascade. Tesis Doctor en Ciencias Médicas. Pontificia universidad Catolica de Chile, Escuela de Medicina, 133p. Retrieved from https://repositorio.uc.cl/xmlui/handle/11534/48244
55. List, M., Dehghani Amirabad, A., Kostka, D., & Schulz, M. H. (2019). Large-scale inference of competing endogenous RNA networks with sparse partial correlation. Bioinformatics, 35(14), i596-i604. doi:10.1093/bioinformatics/btz314
56. Peñaloza, A. (2020). Caracterización de los Competing Endogenous RNAs en la hipertensión arterial primaria. Tesis Magíster en Genética. Universidad de Chile, Facultad de Medicina, 95p. Retrieved from https://repositorio.uchile.cl/handle/2250/184582
57. Bandyopadhyay, S., & Mitra, R. (2009). TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. Bioinformatics (Oxford, England), 25(20), 2625–2631. doi:10.1093/bioinformatics/btp503
58. Wong, N., & Wang, X. (2015). miRDB: an online resource for microRNA target prediction and functional annotations. Nucleic acids research, 43(Database issue), D146–D152. doi:10.1093/nar/gku1104
59. Agarwal, V., Bell, G. W., Nam, J. W., & Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. eLife, 4, e05005. doi:10.7554/eLife.05005
60. Karagkouni, D., Paraskevopoulou, M. D., Chatzopoulos, S., Vlachos, I. S., Tastsoglou, S., Kanellos, I., Papadimitriou, D., Kavakiotis, I., Maniou, S., Skoufos, G., Vergoulis, T., Dalamagas, T., & Hatzigeorgiou, A. G. (2018). DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions. Nucleic acids research, 46(D1), D239–D245. doi:10.1093/nar/gkx1141
61. Paraskevopoulou, M. D., Vlachos, I. S., Karagkouni, D., Georgakilas, G., Kanellos, I., Vergoulis, T., Zagganas, K., Tsanakas, P., Floros, E., Dalamagas, T., & Hatzigeorgiou, A. G. (2016). DIANA-LncBase v2: indexing microRNA targets on non-coding transcripts. Nucleic acids research, 44(D1), D231–D238. doi:10.1093/nar/gkv1270
62. Li, J. H., Liu, S., Zhou, H., Qu, L. H., & Yang, J. H. (2014). starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic acids research, 42(Database issue), D92–D97. doi:10.1093/nar/gkt1248
63. Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1), 139-140. doi:10.1093/bioinformatics/btp616
64. Shannon, P., Markiel, A., Ozier, O., Baliga, N., Wang, J., Ramage, D., . . . Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research, 13(11), 2498-2504. doi:10.1101/gr.1239303
65. Maere, S., Heymans, K., & Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks. Bioinformatics, 21(16), 3448-3449. doi:10.1093/bioinformatics/bti551

66. Merico, D., Isserlin, R., Stueker, O., Emili, A., & Bader, G. D. (2010). Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PLoS One, 5(11), e13984. doi:10.1371/journal.pone.0013984

67. Kucera, M., Isserlin, R., Arkhangorodsky, A., & Bader, G. (2016). AutoAnnotate: A Cytoscape app for summarizing networks with semantic annotations. F1000Research, 5, 1717. doi:10.12688/f1000research.9090.1

68. Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., . . . Mering, C. V. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research, 47(Database issue), D607-D613. doi:10.1093/nar/gky1131

69. Rainer, J. (2017). EnsDb.Hsapiens.v. Retrieved February 25, 2021, from Bioconductor: https://bioconductor.org/packages/release/data/annotation/html/EnsDb.Hsapiens.v86.html

70. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D., & Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. Genome research, 29(8), 1363-1375. doi:10.1101/gr.240663.118

71. Yang, M., Lu, H., Liu, J., Wu, S., Kim, P., & Zhou, X. (2022). lncRNAfunc: a knowledgebase of lncRNA function in human cancer. Nucleic acids research, 50(D1), D1295–D1306. doi:10.1093/nar/gkab1035

72. Russo, P. S., Ferreira, G. R., Cardozo, L. E., Bürger, M. C., Arias-Carrasco, R., Maruyama, S. R., . . . Nakaya, H. I. (2018). CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses. BMC bioinformatics, 19(1), 56. doi:10.1186/s12859-018-2053-1

73. Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M., & Sergushichev, A. (n.d.). Fast gene set enrichment analysis. bioRxiv. doi:10.1101/060012

74. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., & Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Cambridge (Mass.)), 2(3), 100141. doi:10.1016/j.xinn.2021.100141

75. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics (Oxford, England), 27(12), 1739–1740. doi:10.1093/bioinformatics/btr260

76. Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W., & Ma'ayan, A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic acids research, 44(W1), W90–W97. doi:10.1093/nar/gkw377

77. The pandas development team. (2020). pandas-dev/pandas: Pandas. doi:10.5281/zenodo.3509134

78. The International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. Nature, 467(7311), 52–58. doi:10.1038/nature09298

79. Hormozdiari, F., Eskin, E., Price, A. L., Kraft, P., & Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS genetics, 10(10), e1004722. doi:10.1371/journal.pgen.1004722

80. Raj-Kumar, P.-K., Liu, J., Hooke, J. A., Kovatich, A. J., Kvecher, L., Shriver, C. D., & Hu, H. (2019). PCAPAM50 improves consistency between breast cancer intrinsic and clinical subtyping reclassifying a subset of luminal A tumors as luminal B. Scientific Reports, 9(1), 7956. doi:10.1038/s41598-019- 44339-4

81. Herdeis, L., Gerlach, D., McConnell, D. B., & Kessler, D. (2021). Stopping the beating heart of cancer: KRAS reviewed. Current opinion in structural biology, 71, 136–147. doi:10.1016/j.sbi.2021.06.013

82. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., . . . Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Medicine, 12(3), e1001779. doi:10.1371/journal.pmed.1001779

83. Zheng, H. F., Rong, J. J., Liu, M., Han, F., Zhang, X. W., Richards, J. B., & Wang, L. (2015). Performance of genotype imputation for low frequency and rare variants from the 1000 genomes. PloS one, 10(1), e0116487. doi:10.1371/journal.pone.0116487

84. Zhang, Y., Parmigiani, G., & Johnson, W. E. (2020). ComBat-seq: batch effect adjustment for RNA-seq count data. NAR genomics and bioinformatics, 2(3), lqaa078. doi:10.1093/nargab/lqaa078

85. Aran, D., Camarda, R., Odegaard, J., Paik, H., Oskotsky, B., Krings, G., . . . Butte, A. J. (2017). Comprehensive analysis of normal adjacent to tumor transcriptomes. Nature communications, 8(1), 1077. doi:10.1038/s41467-017-01027-z

86. Alluri, P., & Newman, L. A. (2014). Basal-like and triple-negative breast cancers: searching for positives among many negatives. Surgical oncology clinics of North America, 23(3), 567–577. doi:10.1016/j.soc.2014.03.003

87. Xu, B., Zhang, X., Wang, S., & Shi, B. (2018). MiR-449a suppresses cell migration and invasion by targeting PLAGL2 in breast cancer. Pathology, research and practice, 214(5), 790–795. doi:10.1016/j.prp.2017.12.012

88. Shi, W., Bruce, J., Lee, M., Yue, S., Rowe, M., Pintilie, M., . . . Liu, F. F. (2016). MiR-449a promotes breast cancer progression by targeting CRIP2. Oncotarget, 7(14), 18906–18918. doi:10.18632/oncotarget.7753

89. Wang, Y. W., Zhao, S., Yuan, X. Y., Liu, Y., Zhang, K., Wang, J., . . . Ma, R. (2019). miR-4732-5p promotes breast cancer progression by targeting TSPAN13. Journal of cellular and molecular medicine, 23(4), 2549–2557. doi:10.1111/jcmm.14145

90. ElKhouly, A. M., Youness, R. A., & Gad, M. Z. (2020). MicroRNA-486-5p and microRNA-486-3p: Multifaceted pleiotropic mediators in oncological and non-oncological conditions. Non-coding RNA research, 5(1), 11–21. doi:10.1016/j.ncrna.2020.01.001

91. Liu, L., Zhu, Y., Liu, A. M., Feng, Y., & Chen, Y. (2019). Long noncoding RNA LINC00511 involves in breast cancer recurrence and radioresistance by regulating STXBP4 expression via miR-185. European review for medical and pharmacological sciences, 23(17), 7457–7468. doi:10.26355/eurrev_201909_18855

92. Mahmoud, M. M., Sanad, E. F., Elshimy, R., & Hamdy, N. M. (2021). Competitive Endogenous Role of the LINC00511/miR-185-3p Axis and miR-301a-3p From Liquid Biopsy as Molecular Markers for Breast Cancer Diagnosis. Frontiers in oncology, 11, 749753. doi:10.3389/fonc.2021.749753

93. Agbana, Y. L., Abi, M. E., Ni, Y., Xiong, G., Chen, J., Yun, F., . . . Zhu, Y. (2020). LINC00511 as a prognostic biomarker for human cancers: a systematic review and meta-analysis. BMC cancer, 20(1), 682. doi:10.1186/s12885-020-07188-3

94. Orrù, V., Steri, M., Sidore, C., Marongiu, M., Serra, V., Olla, S., . . . Cucca, F. (2020). Complex genetic signatures in immune cells underlie autoimmunity and inform therapy. Nature genetics, 52(10), 1036–1045. doi:10.1038/s41588-020-0684-4

95. Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] –. Accession No. 11006, LILRB4 leukocyte immunoglobulin like receptor B4 [Homo sapiens (human)]; [Retrieved March 8, 2022]. Available from: https://www.ncbi.nlm.nih.gov/gene/11006

96. Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] –. Accession No. 963, CD53 CD53 molecule [Homo sapiens (human)]; [Retrieved March 8, 2022]. Available from: https://www.ncbi.nlm.nih.gov/gene/963

97. Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] –. Accession No. 3071, NCKAP1L NCK associated protein 1 like [Homo sapiens (human)]; [Retrieved March 8, 2022]. Available from: https://www.ncbi.nlm.nih.gov/gene/3071

98. Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – . Accession No. 914, CD2 CD2 molecule [Homo sapiens (human)]; [Retrieved March 8, 2022]. Available from: https://www.ncbi.nlm.nih.gov/gene/914

99. Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] –. Accession No. 114836, SLAMF6 SLAM family member 6 [Homo sapiens (human)]; [Retrieved March 8, 2022]. Available from: https://www.ncbi.nlm.nih.gov/gene/114836

100. Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – . Accession No. 3561, IL2RG interleukin 2 receptor subunit gamma [Homo sapiens (human)]; [Retrieved March 8, 2022]. Available from: https://www.ncbi.nlm.nih.gov/gene/3561

101. Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – . Accession No. 915, CD3D CD3 delta subunit of T-cell receptor complex [Homo sapiens (human)]; [Retrieved March 8, 2022]. Available from: https://www.ncbi.nlm.nih.gov/gene/915

102. Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] –. Accession No. 84174, SLA2 Src like adaptor 2 [Homo sapiens (human)]; [Retrieved March 8, 2022]. Available from: https://www.ncbi.nlm.nih.gov/gene/84174

103. Hu, E. (2020). Refining Polygenic Risk Score Models Through Fine Mapping and Functional Gene Modules. Thesis Master of Engineering in Computer Science and Molecular Biology. Massachusetts Institute of Technology, 30p. Retrieved from https://dspace.mit.edu/handle/1721.1/130692

104. Liang, Y., Pividori, M., Manichaikul, A., Palmer, A. A., Cox, N. J., Wheeler, H. E., & Im, H. K. (2022). Polygenic transcriptome risk scores (PTRS) can improve portability of polygenic risk scores across ancestries. Genome biology, 23(1), 23. doi:10.1186/s13059-021-02591-w

105. Assum, I., Krause, J., Scheinhardt, M. O., Müller, C., Hammer, E., Börschel, C. S., Völker, U., Conradi, L., Geelhoed, B., Zeller, T., Schnabel, R. B., & Heinig, M. (2022). Tissue-specific multi-omics analysis of atrial fibrillation. Nature communications, 13(1), 441. doi:10.1038/s41467-022-27953-1

106. Maj, C., Salvi, E., Citterio, L., Borisov, O., Simonini, M., Glorioso, V., Barlassina, C., Glorioso, N., Thijs, L., Kuznetsova, T., Cappuccio, F. P., Zhang, Z. Y., Staessen, J. A., Cusi, D., Lanzani, C., & Manunta, P. (2022). Dissecting the Polygenic Basis of Primary Hypertension: Identification of Key Pathway-Specific Components. Frontiers in cardiovascular medicine, 9, 814502. doi:1010.3389/fcvm.2022.814502

107. Li, X., & Wang, C. Y. (2021). From bulk, single-cell to spatial RNA sequencing. International journal of oral science, 13(1), 36. doi:10.1038/s41368-021-00146-0

108. Eraslan, G., Drokhlyansky, E., Anand, S., Subramanian, A., Fiskin, E., Slyper, M., . . . Regev, A. (2021). Single-nucleus cross-tissue molecular reference maps to decipher disease gene function. bioRxiv, 2021.07.19.452954. doi:10.1101/2021.07.19.452954

109. Chen, S., Francioli, L. C., Goodrich, J., Collins, R., Wang, Q., Alföldi, J., . . . Karczewski, K. (2022). A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. bioRxiv, 2022.03.20.485034. doi:10.1101/2022.03.20.485034

110. Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., . . . MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature, 581(7809), 434–443. doi:10.1038/s41586-020-2308-7

111. ENCODE Project Consortium, Moore, J. E., Purcaro, M. J., Pratt, H. E., Epstein, C. B., Shoresh, N., . . . Weng, Z. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature, 583(7818), 699–710. doi:10.1038/s41586-020-2493-4

112. Azodi, C. B., Tang, J., & Shiu, S. H. (2020). Opening the Black Box: Interpretable Machine Learning for Geneticists. Trends in genetics, 36(6), 442–455. doi:10.1016/j.tig.2020.03.005