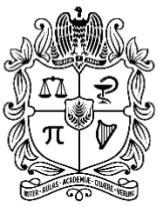


UNIVERSIDAD
NACIONAL
DE COLOMBIA

Modelo de aprendizaje automático de integración de datos genómicos, epigenómicos, transcriptómicos y clínicos provenientes de estudios de cáncer de endometrio y de cáncer de mama

Samuel Eyrolle-Cellier

Tesis presentada como requisito parcial para optar al título de:

Maestría en Bioinformática

Director:

Luis Fernando Niño Vásquez, PhD

Línea de Investigación:

Tecnologías Computacionales en Bioinformática

Grupo de Investigación:

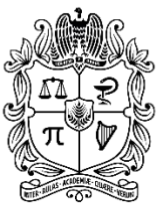
Laboratorio de Investigación en Sistemas Inteligentes

Universidad Nacional de Colombia

Facultad de Ingeniería

Bogotá, Colombia

2023



UNIVERSIDAD
NACIONAL
DE COLOMBIA

Machine learning model for integrating genomic, epigenomic, transcriptomic and clinical data from endometrial cancer and breast cancer studies

Samuel Eyrolle-Cellier

Declaración de obra original

Yo declaro lo siguiente:

He leído el Acuerdo 035 de 2003 del Consejo Académico de la Universidad Nacional. «Reglamento sobre propiedad intelectual» y la Normatividad Nacional relacionada al respeto de los derechos de autor. Esta disertación representa mi trabajo original, excepto donde he reconocido las ideas, las palabras, o materiales de otros autores.

Cuando se han presentado ideas o palabras de otros autores en esta disertación, he realizado su respectivo reconocimiento aplicando correctamente los esquemas de citas y referencias bibliográficas en el estilo requerido.

He obtenido el permiso del autor o editor para incluir cualquier material con derechos de autor (por ejemplo, tablas, figuras, instrumentos de encuesta o grandes porciones de texto).

Por último, he sometido esta disertación a la herramienta de integridad académica, definida por la universidad.

Bogotá, el 28 de julio de 2023.

Agradecimientos

En este momento de culminación de mi tesis, deseo expresar mi profundo agradecimiento a todas las personas que contribuyeron de manera significativa en la realización de este trabajo.

En primer lugar, quiero agradecer a mi supervisor, el doctor Luis Fernando Niño Vásquez, por su dedicación, paciencia y conocimientos compartidos a lo largo de todo el proceso de investigación. Su orientación experta y valiosos consejos han sido clave para el desarrollo de esta tesis. También quiero extender mi agradecimiento a dos estudiantes de pregrado en Ingeniería de Sistemas de la Universidad Nacional de Colombia y miembros del semillero de investigación del doctor Luis Fernando Niño Vásquez: Diego Felipe López Ávila por su ayuda en la gestión de los archivos de datos y Venus Estefanía Baquero Vargas por su apoyo en la implementación del modelo de aprendizaje multi-vista para la integración de los distintos tipos de datos. Cabe resaltar las recomendaciones y el apoyo brindado por los demás miembros del grupo de investigación LISI: Laboratorio de Investigación en Sistemas Inteligentes. En adición, quiero expresar mi gratitud a todas aquellas fuentes bibliográficas y trabajos de investigación que consulté durante este proyecto, ya que su invaluable conocimiento ha enriquecido significativamente mi trabajo.

Dedico esta tesis a mi madre, que recuerdo con mucho cariño y que supo transmitirme su interés por la investigación y el método científico. Agradezco a los demás miembros de mi familia, especialmente a mi padre, por su apoyo a lo largo de todo este camino académico. Quiero dar las gracias a mis amigos más cercanos, Sovio y Mar, por el apoyo en mi trayectoria estos dos últimos años, por las risas y todos los momentos compartidos. Finalmente, deseo expresar mi gratitud a todas las lindas personas encontradas en el camino: mis compañeras científicas Diana Vanegas y Diana Briñez a quien deseo muchos éxitos en la culminación de sus estudios de maestría, Laura Moreno, Karina, Lina, Jana, Marcela, Yesid, Germán, Laura Reyes, Daniela, Gabriela, Santiago, Mariana, Abby, Alejandra, Laura Mayorga, Sofía, Zohé e Ivon.

En síntesis, este logro no hubiera sido posible sin el apoyo y la colaboración de cada una de las personas mencionadas anteriormente. A todos ustedes, mi más sincero agradecimiento.

¡Gracias!

Resumen

Modelo de aprendizaje automático de integración de datos genómicos, epigenómicos, transcriptómicos y clínicos provenientes de estudios de cáncer de endometrio y de cáncer de mama

El cáncer de mama y el cáncer de endometrio son enfermedades complejas que presentan mucha heterogeneidad a nivel molecular e histológico. Ciertos pacientes de estos dos tipos de cáncer comparten tanto mecanismos moleculares y celulares, como factores causales, como lo es el hiperestrogenismo. Este proyecto de investigación buscó identificar biomarcadores tumorales compartidos entre ambas enfermedades.

565 pacientes con cáncer de mama y 348 con cáncer de endometrio de la plataforma *The Cancer Genome Atlas* fueron seleccionados según sus características histológicas, hormonales e inmunológicas. Sus datos ómicos fueron analizados de manera separada e integrada mediante el uso del algoritmo de aprendizaje multi-vista *Deep Generalized Canonical Correlation Analysis* y del método de reducción de dimensionalidad *Uniform Manifold Approximation and Projection*. Se extrajeron biomarcadores de cada grupo (*cluster*) a través del cálculo del puntaje de información mutua entre las variables iniciales y las variables sintéticas UMAP1 y UMAP2.

El análisis de los biomarcadores reveló que varios de estos genes tienen un rol en la proliferación celular, la apoptosis y la angiogénesis. Así mismo, el análisis reveló que la ausencia de metilación en las regiones promotoras de CLTC, importante en la organización del huso mitótico, y SON, involucrado en el empalme del ARN, es una característica compartida entre muchos pacientes de la cohorte. Por otro lado, FBXO11 y PTPN11 se caracterizan por niveles altos de expresión génica en ambos tipos de cáncer. FBXO11 codifica para una ubiquitina ligasa necesaria para la degradación proteica; mientras que PTPN11 codifica para una tirosina fosfatasa que actúa en la transducción de señales mediante una regulación positiva de la vía de señalización RAS/RAF/MAPK.

En conclusión, la estrategia de integración multi-ómica permitió descubrir biomarcadores que no aparecen en el análisis de datos ómicos de un solo tipo. Se inscribe como una prueba de concepto de integración de distintos tipos de datos provenientes de diferentes contextos patológicos en el campo de la oncología.

Palabras clave: Cáncer de mama, Cáncer de endometrio, Integración Multi-ómica, Aprendizaje multi-vista, *Clustering*, Biomarcadores.

Abstract

Machine learning model for integrating genomic, epigenomic, transcriptomic and clinical data from endometrial cancer and breast cancer studies

Breast cancer and endometrial cancer are complex diseases that show a high degree of molecular and histological heterogeneity. Certain patients with these two types of cancer share both molecular and cellular mechanisms, as well as causal factors such as hyperestrogenism. This research project aimed to identify shared tumor biomarkers between both diseases.

565 breast cancer patients and 348 endometrial cancer patients from *The Cancer Genome Atlas* platform were selected based on their histological, hormonal, and immunological characteristics. Their omics data was analyzed separately and integratively using the multi-view learning algorithm *Deep Generalized Canonical Correlation Analysis* and the dimensionality reduction method *Uniform Manifold Approximation and Projection*. Biomarkers were extracted from each cluster by calculating the mutual information score between the initial variables and the UMAP1 and UMAP2 synthetic variables.

The analysis of the biomarkers revealed that several of these genes play a role in cell proliferation, apoptosis, and angiogenesis. Additionally, the analysis showed that the absence of methylation in the promoter regions of CLTC, which is important in the organization of the mitotic spindle, and SON, involved in RNA splicing, is a shared characteristic among many patients in the cohort. On the other hand, FBXO11 and PTPN11 are characterized by high levels of gene expression in both types of cancer. FBXO11 encodes for a ubiquitin ligase necessary for protein degradation, while PTPN11 encodes for a tyrosine phosphatase that acts in signal transduction by positively regulating the RAS/RAF/MAPK signaling pathway.

In conclusion, the multi-omic integration strategy allowed the discovery of biomarkers that have not been identified in the omics data analysis of a single type. It serves as a proof of concept for integrating different types of data from different pathological contexts in the field of oncology.

Keywords: Breast Cancer, Endometrial Cancer, Multi-Omics Integration, Multi-view Learning, Clustering, Biomarkers.

Esta tesis de maestría se sustentó el 3 de noviembre de 2023 a las 11:00 a.m., y fue evaluada por los siguientes jurados:

Elizabeth León Guzmán, PhD
Universidad Nacional de Colombia, Facultad de Ingeniería

Andrés Mauricio Pinzón Velasco, PhD
Universidad Nacional de Colombia, Instituto de Genética

Contenido

Lista de Figuras

Anexos

Introducción

El cáncer de mama y el cáncer de endometrio son el primero y el cuarto cáncer más frecuentes en mujeres, respectivamente. Estos han sido ampliamente estudiados por médicos, biólogos y epidemiólogos, entre otros. Se han identificado varios subtipos histológicos y moleculares que difieren en cuanto a las células afectadas, factores causales, capacidad invasiva y biomarcadores. Por ejemplo, el adenocarcinoma endometrial y los cánceres de mama luminales A y B – positivos para los receptores de estrógenos y progesterona – afectan células epiteliales, están asociados con un desequilibrio del balance estrógenos/progestágenos, y presentan una baja capacidad invasiva, por tanto, un buen pronóstico (Harbeck et al., 2019; Makker et al., 2021).

La identificación de biomarcadores tumorales siempre ha sido un objetivo principal en investigación ya que permite mejorar la prognosis, la clasificación, la terapia y la predicción de la supervivencia o del riesgo en el ámbito de la oncología (Abeel et al., 2010). En efecto, la principal causa de la aparición de células malignas en el endometrio y en las mamas es la acumulación de mutaciones y de lesiones epigenéticas en el ADN de las células epiteliales sanas de estos tejidos. Estas mutaciones han sido reportadas en los siguientes genes: K-RAS, HER2, EGFR, PI3KCA, FGFR2, PTEN, y TP53, entre otros para el cáncer de endometrio (Banno et al., 2012); BRCA1, BRCA2, ESR1, PGR, HER2, y MKI67, entre otros, para el cáncer de mama (Harbeck et al., 2019). Estos biomarcadores han sido identificados a través del análisis de datos ómicos de un solo tipo o del uso de experimentos clásicos de biología molecular y celular.

En las dos últimas décadas, las ciencias ómicas han generado enormes cantidades de datos asociados con el perfil genómico, epigenómico y transcriptómico de los pacientes de cáncer. Estos datos están disponibles en línea y muchos grupos de investigación buscan desarrollar modelos de aprendizaje automático destinados a la integración de dichos datos (Reel et al., 2021). Uno de los objetivos de estas nuevas herramientas

bioinformáticas es identificar nuevos biomarcadores tumorales. Existe una gran variedad de modelos de aprendizaje automático de integración de datos multiómicos. Algunos modelos usan datos anotados para el entrenamiento (aprendizaje supervisado) mientras que otros modelos se caracterizan por datos de entrenamiento sin etiquetas, así que el algoritmo intenta interpretar la información suministrada por sí solo. Además, la estrategia de integración de datos varía considerablemente de un modelo a otro, ciertos modelos buscan concatenar los distintos tipos de datos y luego llevar a cabo el análisis (integración temprana) cuando otros modelos realizan primero el análisis de los distintos tipos de datos antes de integrar las representaciones obtenidas (integración tardía, métodos basados en similitud). La elección del modelo de aprendizaje automático depende, por un lado, de las características de los datos (tipo de datos, dimensionalidad, valores faltantes, distribución, etc.) y, por otro lado, del objetivo del proyecto (agrupación de pacientes, predicción de la supervivencia, elección del tratamiento, identificación de biomarcadores, etc.).

A continuación, se presenta el estado del arte de la investigación sobre cáncer de mama y cáncer de endometrio, así como los métodos de integración de datos. Luego se hace una presentación formal de la investigación con el planteamiento, la justificación del problema y los objetivos.

1. Estado del arte

El estado del arte desarrollado busca describir algunos factores causales asociados con el cáncer de endometrio y el cáncer de mama y, por otra parte, dar una idea clara sobre las dinámicas moleculares y celulares que rigen la iniciación y la progresión de ambos tipos de cáncer. De igual manera, esta investigación se enfoca en las estrategias bioinformáticas que sirven para integrar datos multiómicos (por ejemplo, datos genómicos, epigenómicos y transcriptómicos).

1.1. Presentación del cáncer de endometrio y del cáncer de mama

En esta sección, las principales características del cáncer de endometrio y del cáncer de mama desde datos epidemiológicos hasta factores de riesgo y mecanismos fisiopatológicos son presentadas con un enfoque en las hormonas sexuales femeninas (estrógenos y progestágenos).

1.1.1. Datos epidemiológicos

El cáncer de mama y el cáncer de endometrio son el primero y el cuarto cáncer más frecuentes en mujeres, respectivamente. En 2020, 417367 mujeres fueron diagnosticadas con cáncer de endometrio en el mundo (Makker et al., 2021). En 2018, 2.1 millones de mujeres fueron diagnosticadas con cáncer de mama. También, 626679 mujeres con cáncer de mama fallecieron en el mundo (Harbeck et al., 2019). La incidencia de estos dos tipos de cáncer ha estado subiendo desde los años 1980, con un incremento anual de 3.1% para el cáncer de mama (Bray et al., 2015). Las regiones más afectadas son América del Norte, Europa y Latinoamérica. Esto se explica por el nivel de acceso a atención médica de alta calidad, por la densidad de oncólogos y por el envejecimiento de la población en esas regiones (Chatterjee et al., 2016).

Estos dos tipos de cáncer presentan altas tasas de supervivencia a 5 años: 90.3% para el cáncer de mama (*Female Breast Cancer — Cancer Stat Facts*, n.d.) y 81.1% para el cáncer de endometrio (*Cancer of the Endometrium - Cancer Stat Facts*, n.d.). Sin embargo, la prognosis depende del grado de invasión, el cual difiere mucho según el subtipo molecular de cáncer. Tanto el cáncer de endometrio como el cáncer de mama son cánceres altamente heterogéneos molecularmente. En el caso del cáncer de endometrio, el cáncer de tipo II está asociado con una mayor tasa de mortalidad que el cáncer de tipo I (Makker et al., 2021). En el caso del cáncer de mama, el cáncer HER2-positivo está asociado con una mayor tasa de mortalidad, seguido por el cáncer triple negativo, el subtipo luminal A y luminal B (Harbeck et al., 2019).

1.1.2. Factores de riesgo

Un riesgo mayor de cáncer de endometrio está asociado con los siguientes factores: mayor edad, pertenencia a ciertas etnias (mujeres asiáticas), mayor índice de masa

corporal y obesidad, exposición a estrógenos endógenos o exógenos, menarquía precoz, menopausia tardía, menos paridad, síndrome metabólico, antecedentes familiares y predisposición genética (Colombo et al., 2016). De hecho, los desequilibrios hormonales aparecen como el principal impulsor de la carcinogénesis endometrial. En efecto, la terapia de reemplazo de estrógenos durante la menopausia, la anovulación crónica (causa de esterilidad femenina caracterizada por una ausencia de ovulación debido a que los ovarios no liberan óvulos en ningún momento del ciclo menstrual), y el tratamiento con tamoxifeno (medicamento que bloquea la actividad del estrógeno empleado como terapia complementaria para el cáncer de mama) son factores de riesgo muy conocidos. Generan un desequilibrio del balance estrógenos/progestágenos a favor de los estrógenos en el endometrio. Asimismo, alteraciones en los ovarios como la hipertecosis del estroma ovárico y el síndrome de ovario poliquístico pueden afectar el balance hormonal. De la misma manera, la obesidad promueve la carcinogénesis endometrial a través del aumento de la producción de estrógenos por conversión de andrógenos por los adipocitos (Fig. 1; Rodriguez et al., 2019).

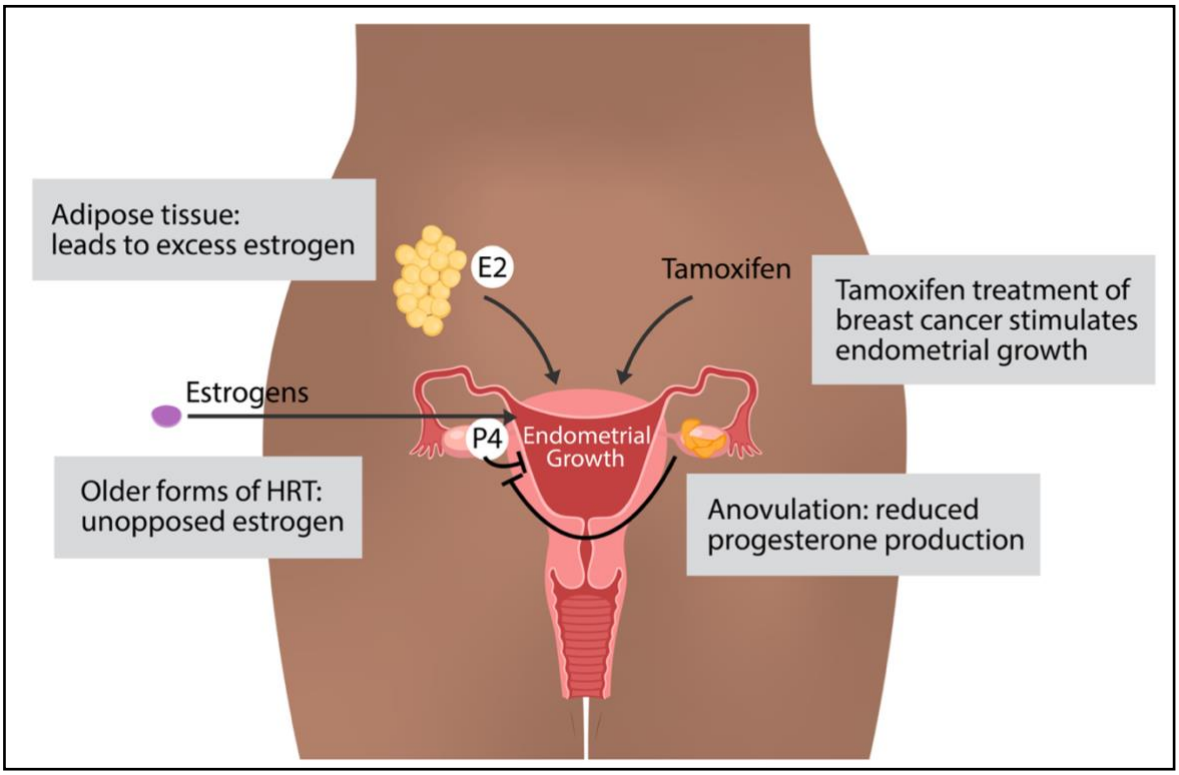


Figura 1: Cáncer de endometrio y desequilibrios hormonales (Rodriguez et al., 2019). Los estrógenos impulsan el crecimiento endometrial mediante la proliferación de las células epiteliales mientras que los progestágenos bloquean el crecimiento endometrial y promueven la diferenciación de las células epiteliales endometriales. Un exceso de estrógenos puede ser causado por la obesidad, la terapia de reemplazo de estrógenos y el tratamiento con tamoxifeno. Una pérdida de progestágenos (progesterona principalmente) puede ocurrir en el contexto de la anovulación.

Además, ciertas mutaciones de la línea germinal incrementan el riesgo de cáncer de endometrio. Dichas mutaciones están asociadas con el síndrome de Lynch y el síndrome de Cowden. El síndrome de Lynch es un trastorno genético que presenta una herencia autosómica dominante. Se caracteriza por mutaciones germinales en uno de los genes de reparación de emparejamiento: MLH1, MSH2, MSH6 o PMS2. Aproximadamente 3% de los pacientes con cáncer de endometrio padecen síndrome de Lynch (Ryan et al., 2019). Las mutaciones germinales de PTEN son raras y características del síndrome de Cowden (Ring et al., 2016). Finalmente, la asociación entre mutaciones germinales de BRCA1 y cáncer de endometrio sigue siendo tema de controversia (Makker et al., 2021).

Los factores de riesgo del cáncer de mama pueden ser clasificados de la siguiente manera:

- Factores de riesgo modificables: obesidad, inactividad física y consumo de alcohol. 20% de los casos están asociados a factores de riesgo modificables (Danaei et al., 2005).
- Factores de riesgo heredados: mutaciones en los dos genes supresores de tumores BRCA1 y BRCA2 involucrados en la reparación del ADN (Huen et al., 2010). Las mutaciones en BRCA1 y BRCA2 muestran un patrón de herencia autosómico dominante, están asociadas con un riesgo acumulado de desarrollo de cáncer de mama a la edad de 80 años de 72% y 69%, respectivamente (Brewer et al., 2017).
- Trastornos genéticos relacionados con mutaciones en la línea germinal: síndrome de Li-Fraumeni, PALB2, CHEK2, Ataxia telangiectosa, síndrome de Cowden (Harbeck et al., 2019).
- Afectaciones hormonales: exposición a estrógenos, menarquía precoz, menopausia tardía, ausencia de lactancia materna (Harbeck et al., 2019), y uso de algunos contraceptivos hormonales (Mørch et al., 2017). Los estrógenos son un promotor del cáncer de mama, a través de su unión a su receptor ERα ubicado en el núcleo celular. Así pueden modular la expresión génica. Los estrógenos estimulan el desarrollo de las mamas durante la pubertad, los ciclos menstruales y los embarazos. A lo largo de los ciclos menstruales, un desequilibrio del balance estrógenos/progestágenos aumenta la proliferación celular y puede provocar una acumulación de daños en el ADN. Con la repetición del proceso en cada ciclo, el proceso de reparación se puede ver afectado, lo que lleva a mutaciones en las

células premalignas y luego en las células malignas. Los estrógenos promueven el crecimiento de las células malignas y de algunas células del microentorno tumoral que apoyan el desarrollo del tumor (Harbeck et al., 2019).

1.1.3. Mecanismos fisiopatológicos

Muchas veces, el cáncer de endometrio es una enfermedad sensible a las hormonas que surge en un contexto de estimulación estrogénica excesiva de la capa endometrial del útero. Esta estimulación actúa como una señal mitogénica en las células epiteliales de las glándulas del endometrio y puede conducir a su transformación maligna. Así se establece el cáncer endometrioide (etapa I). El cáncer endometrioide evoluciona a través de la transformación maligna de las lesiones precursoras llamadas hiperplasias endometriales (García Ayala et al., 2010). Estas lesiones pueden exhibir mutaciones de PTEN. La adquisición de mutaciones en el gen ARID1A (Suryo Rahmanto et al., 2020) y la inactivación de TGFβ (Gao et al., 2017) promueven la progresión de las hiperplasias endometriales en carcinoma endometrial invasivo (etapa II). Luego, el carcinoma endometrial invasivo se puede propagar al aparato genital femenino completo (etapa III) y a la región abdominal y al organismo completo (etapa IV). Esta progresión se hace de manera progresiva por medio de una acumulación de mutaciones que promueve la transición epitelio-mesénquima de las células epiteliales del endometrio (Fig. 2; Makker et al., 2021). Cabe resaltar que existen otros subtipos histológicos de cáncer de endometrio que no se desarrollan a partir de células epiteliales como el adenocarcinoma que acaba de ser presentado. Por ejemplo, se puede mencionar el carcinoma seroso, el carcinoma de células claras y el carcinosarcoma. Estos subtipos no están asociados con hiperestrogenismo y pertenecen a los cánceres de endometrio de tipo II mientras que la mayoría de los adenocarcinomas de útero pertenecen al tipo I.

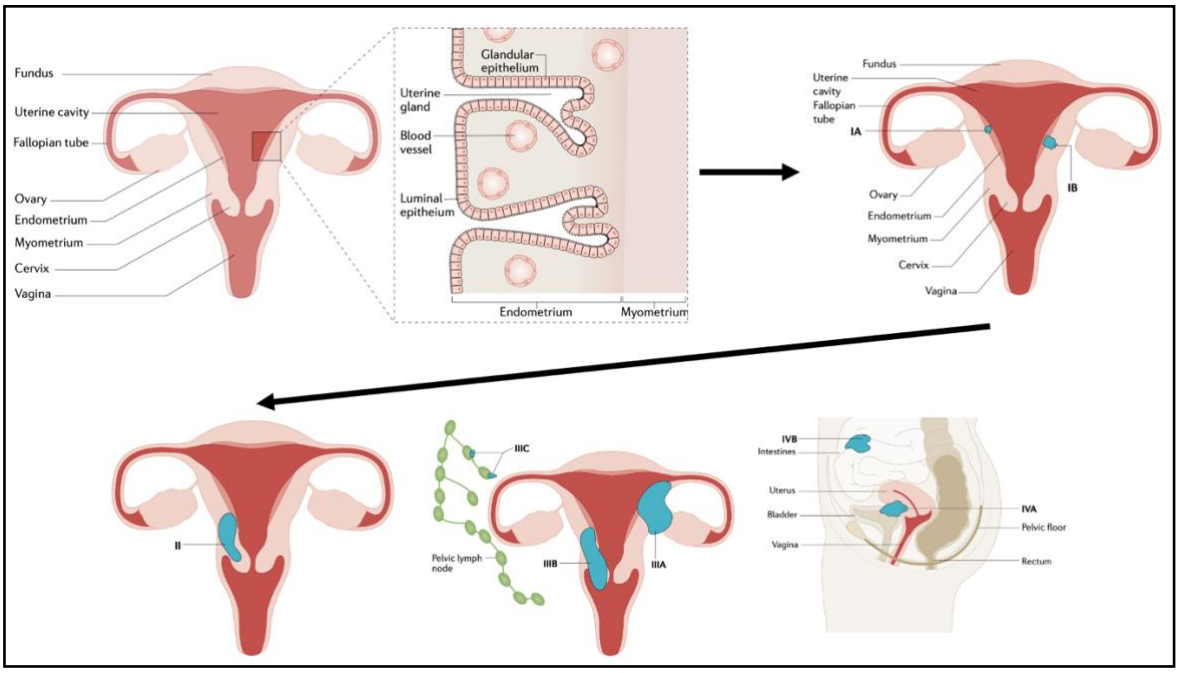


Figura 2: Etapas del adenocarcinoma de endometrio (Makker et al., 2021). La primera imagen corresponde a la organización celular de un útero sano mientras que las siguientes imágenes representan la progresión del cáncer de endometrio. IA: invasión inferior a la mitad del miometrio. IB: invasión superior o igual a la mitad del miometrio. II: invasión del estroma del cuello uterino sin extensión más allá del útero. IIIA: invasión de la parte superior de la serosa (capa de tejido que cubre el exterior del útero) adyacente a las trompas de Falopio y a los ovarios. IIIB: invasión de la parte inferior de la serosa y de la vagina. IIIC: formación de metástasis en los ganglios linfáticos pélvicos. IVA: metástasis en la vejiga y en la mucosa intestinal. IVB: metástasis distantes incluyendo metástasis intraabdominales y metástasis en los ganglios linfáticos inguinales.

El cáncer de endometrio admite cuatro subtipos moleculares definidos por la carga mutacional del tumor (número total de mutaciones que se encuentran en el ADN de las células cancerosas) y por las alteraciones del número de copias de genes (Fig. 3; The Cancer Genome Atlas Research Network & Levine, 2013): *POLE (ultramutated)*, *MSI (hypermutated)*, *CN low (endometrioid)* y *CN high (serous-like)*.

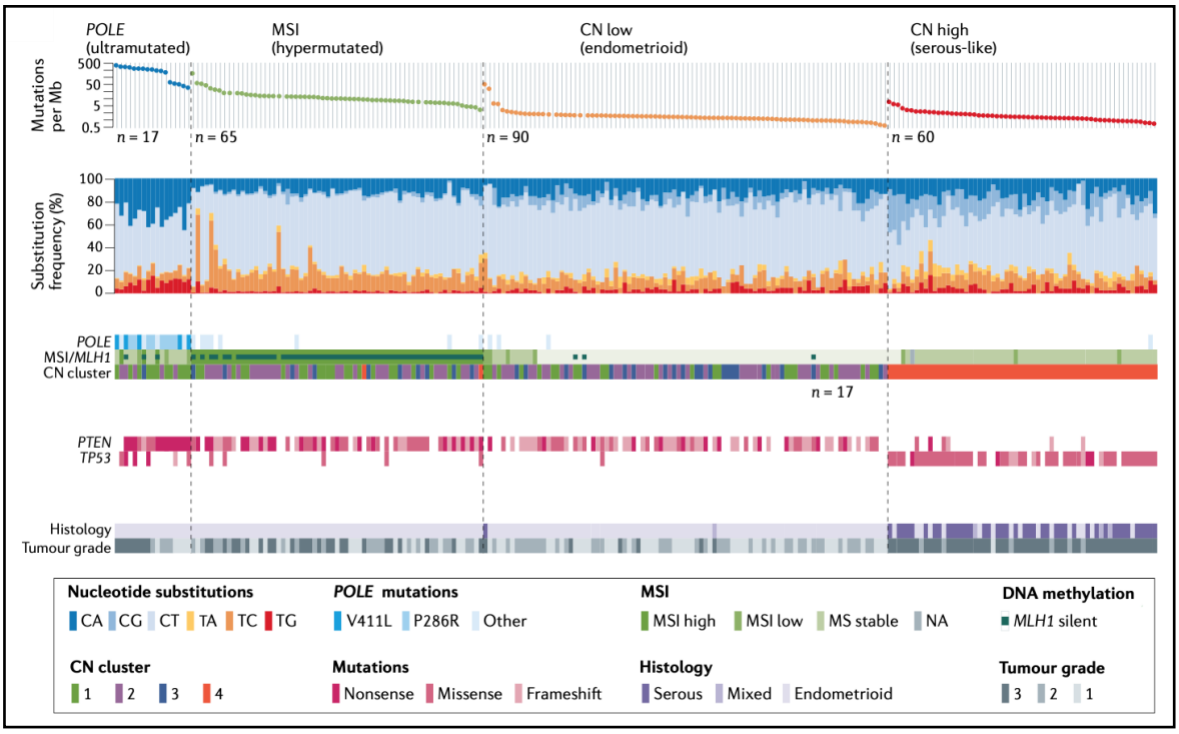


Figura 3: Identificación de cuatro subtipos moleculares de cáncer de endometrio (The Cancer Genome Atlas Research Network & Levine, 2013).

Todos los cánceres de mama surgen en las unidades lobulillares de los conductos de leche (donde se produce la leche materna) o en los conductos de leche (tubos delgados que permiten el transporte de la leche de las unidades lobulillares a los pezones). Existen varios subtipos histológicos de cáncer de mama según su ubicación y sus capacidades invasivas. Primero, el carcinoma ductal *in situ* que inicia en los conductos de leche, se propaga a través de los conductos, distorsiona la arquitectura ductal y puede progresar a cáncer invasivo. También existe el carcinoma lobulillar *in situ* que inicia en las unidades lobulillares de los conductos de leche y no distorsiona la arquitectura ductal. Puede ser clasificado de factor de riesgo en lugar de precursor de cáncer invasivo. Importantemente, el carcinoma ductal *in situ* puede volverse invasivo y producir metástasis en el sistema linfático y en la sangre. Paralelamente, el carcinoma lobulillar invasivo produce metástasis preferiblemente en las vísceras (Fig. 4; Harbeck et al., 2019).

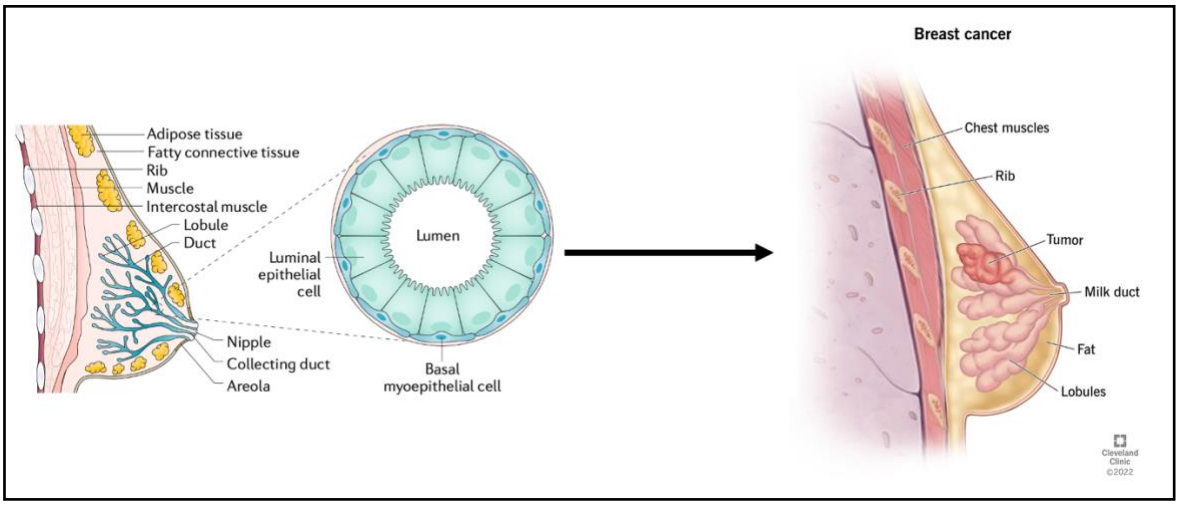


Figura 4: Anatomía del seno y del carcinoma de mama (*Breast Cancer Overview*, n.d.; Harbeck et al., 2019).

En cuanto a las alteraciones moleculares características del cáncer de mama, los genes más frecuentemente mutados o amplificados en las células tumorales son TP53, PIK3CA, MYC, PTEN, CCND1, ERBB2, FGFR1 y GATA3 (Nik-Zainal et al., 2016). El cáncer de mama admite cinco subtipos definidos por características histológicas y moleculares tales como la expresión del receptor de estrógenos, del receptor de progesterona, de la proteína HER2 y del marcador de proliferación Ki67. Los cánceres luminales A y B son positivos para la expresión de los receptores de hormonas femeninas, presentan bajas tasas de proliferación y están asociados a buenos pronósticos (Fig. 5; Harbeck et al., 2019).

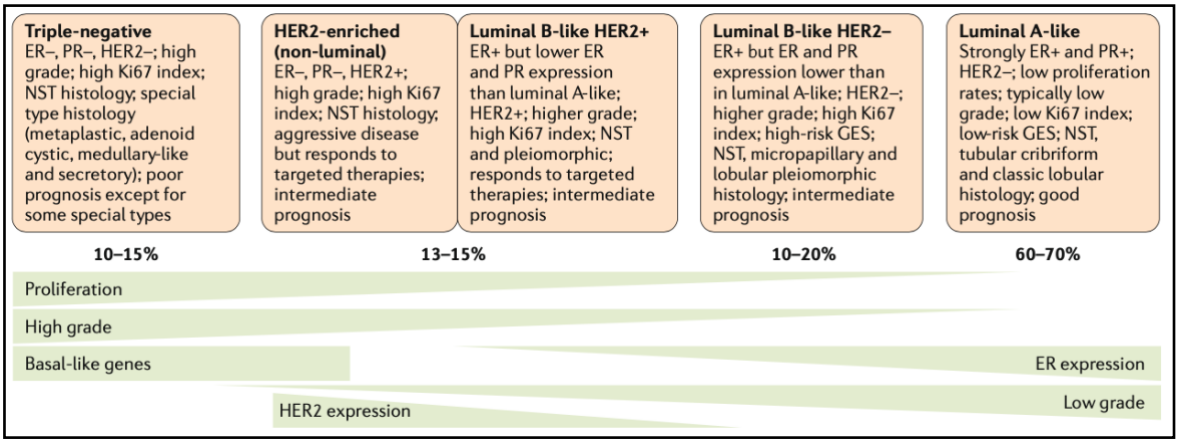


Figura 5: Identificación de cinco subtipos moleculares de cáncer de mama (Harbeck et al., 2019).

Para concluir, a pesar de que haya mucha heterogeneidad histológica y molecular en los cánceres de endometrio y de mama, se pueden identificar subtipos que comparten

tanto mecanismos moleculares y celulares como factores de riesgo. En efecto, los adenocarcinomas endometriales y los cánceres de mama luminales A y B están asociados con desregulaciones hormonales (balance estrógenos/progestágenos), proliferación descontrolada de células epiteliales y pronóstico favorable. Los estudios enfocados en cánceres ginecológicos y cáncer de mama suelen incluir todos los subtipos de tumores (Berger et al., 2018). Una mejor delimitación de los subtipos tumorales incluidos podría permitir mejorar la identificación de biomarcadores compartidos entre cáncer de endometrio y cáncer de mama.

1.2. Métodos de integración de datos ómicos y no ómicos

En esta sección, tras una presentación del dogma central de la biología molecular y de las ciencias ómicas, se van a discutir las diferentes estrategias de integración de datos multiómicos.

1.2.1. Presentación del dogma central de la biología molecular y de las ciencias ómicas

El dogma central de la biología molecular fue propuesto por Francis Crick en el año 1970. Afirma que la expresión de la información contenida en los genes de una célula se hace de forma unidireccional, es decir, que el ADN se usa como molde para la generación del ARN mediante el proceso de transcripción y que el ARN se transforma en proteínas a través del proceso de traducción (Crick, 1970). Sin embargo, este dogma se vio debilitado, primero, en el mismo año, cuando se descubrió el proceso de transcripción inversa el cual permite la generación de ADN complementario a partir de ARN (Fig. 6; Baltimore, 1970; Temin & Mizutani, 1970).

Figura 6: Dogma central de la biología molecular.

Más adelante, los avances tecnológicos en biología clásica así como el desarrollo de las ciencias ómicas permitieron revelar el retrocontrol ejercido por las proteínas y por ciertas moléculas de ARN sobre el genoma. En las dos últimas décadas, se han

desarrollado nuevas técnicas que permiten estudiar el conjunto de ciertas moléculas en un tejido o en una célula en un instante dado. Este campo de investigación se llama ciencias ómicas. Dependiendo de las moléculas estudiadas, se habla de genómica (ADN), epigenómica (metilación del ADN y modificación de las histonas), transcriptómica (ARN), proteómica (proteínas), metabolómica (metabolitos) o microbiómica (microbiota) (Fig. 7; Hasin et al., 2017).

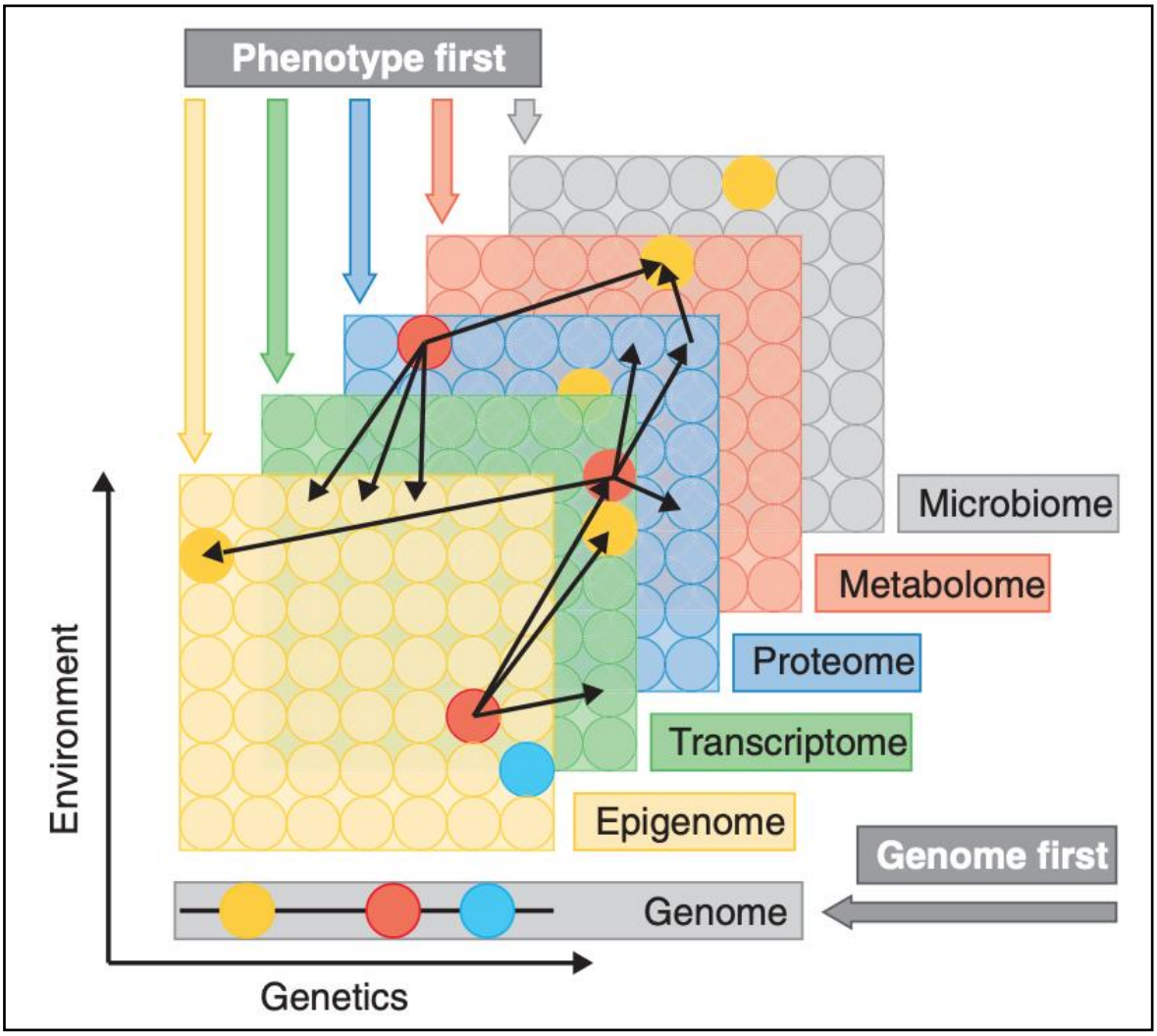


Figura 7: Presentación de los distintos tipos de datos ómicos (Hasin et al., 2017). Aparte del genoma, las demás capas reflejan tanto la regulación intrínseca (epigenoma, transcriptoma, proteoma y metaboloma) como la regulación ejercida por el entorno (microbiota). Las flechas negras representan posibles interacciones o correlaciones detectadas entre moléculas en diferentes capas.

Se volvió común el uso de estas técnicas para mejorar el entendimiento de los mecanismos moleculares implicados en diferentes etapas (diagnosis, respuesta al tratamiento, recaída, metástasis) del cáncer (Reel et al., 2021). Estas técnicas producen

tremendas cantidades de datos ya que no se enfocan en una molécula, sino en el conjunto de moléculas presentes en las condiciones de interés. Es importante resaltar que muchos datos ómicos o multiómicos de estudios de cáncer están disponibles en línea en las siguientes plataformas: *The Cancer Genome Atlas* (TCGA) que tiene más de 2.5 petabytes de datos, *International Cancer Genome Consortium* (ICGC), *Catalogue of Somatic Mutations In Cancer* (COSMIC) o *The Pathology Atlas* (TPA) (Das et al., 2020).

1.2.2. Integración de datos multiómicos

Últimamente se han desarrollado muchas herramientas bioinformáticas para la integración de datos multiómicos, es decir datos ómicos de distintas naturalezas; por ejemplo, datos de genómica y datos de transcriptómica obtenidos a partir de las muestras de los mismos pacientes. Estas herramientas son llamadas algoritmos de aprendizaje multi-vista. Estos esfuerzos han permitido el descubrimiento de nuevos biomarcadores que no aparecían en el análisis de datos ómicos de un solo tipo. Dichas herramientas bioinformáticas usan distintos algoritmos de aprendizaje profundo (Fig. 8; Reel et al., 2021). El aprendizaje profundo es un campo del aprendizaje de máquinas en el cual los algoritmos se basan en redes neuronales artificiales que tratan de imitar el cerebro humano con una cascada de capas. Estas capas ordenan y filtran los conjuntos de datos utilizados para “entrenar” el modelo de aprendizaje profundo, también comunican entre ellas, lo que permite a cada capa refinar la salida. De manera general, existen dos tipos de aprendizaje: supervisado y no supervisado. Por un lado, el aprendizaje supervisado es aquel en el cual los datos para el entrenamiento están anotados, es decir, que incluyen la solución deseada. Por otro lado, el aprendizaje no supervisado es el que no comprende etiquetas en los datos de entrenamiento, así que el algoritmo intentará interpretar la información suministrada por sí solo.

Family	Models	Comparative Accuracy	Overfitting Risk	Samples needed	Explainability	Hyper-parameter Tuning	Complexity	Implementation Time	Computation Cost
Probability-based (Bayesian)	Bayesian Network	2	2	2	2	3	3	2	3
	Naive Bayes	2	2	2	2	2	3	2	3
Information based (Tree)	Decision Tree	2	3	2	3	2	2	1	2
	Random Forest	3	2	1	3	3	2	1	2
	Gradient Boosting	3	3	2	1	4	4	2	3
Error based (Linear)	Linear Regression	1	3	2	3	1	2	1	2
	Logistic Regression	1	3	2	3	1	2	1	2
	Partial Linear Regression	2	1	3	3	2	2	1	2
Similarity-based (Instance)	K nearest neighbour	2	3	2	2	2	3	1	1
	Self-Organising Maps	2	3	2	2	3	3	1	1
Support Vectors	Linear SVM	3	3	3	1	3	2	2	2
	Non-linear (Kernel) SVM	3	3	3	1	3	3	3	3
Neural Network-based	Artificial Neural Network	3	3	2	1	3	3	3	3
	Deep Learning (Neural Network)	4	1	4	1	4	4	4	4

Figura 8: Algoritmos de aprendizaje de máquinas usados para la integración de datos multiómicos y sus características de desempeño (Reel et al., 2021). 1: bajo, 2: medio, 3: alto y 4: muy alto.

Existen cinco estrategias de integración de datos multiómicos. Tradicionalmente, han sido empleadas para agrupar muestras o pacientes según sus perfiles ómicos, es decir, encontrar grupos coherentes de muestras en los datos para luego poder descubrir módulos de genes co-expresados en ciertos grupos o identificar subtipos de cáncer (Fig. 9; Rappoport & Shamir, 2018). Cada estrategia va a ser explicada a continuación:

- Integración temprana: Se establece una matriz concatenada que incluye los datos de todas las ómicas para cada muestra y se aplican los mismos algoritmos de análisis de datos ómicos de un solo tipo a los datos multiómicos de la matriz concatenada. Por lo tanto, esta estrategia permite el uso de algoritmos existentes. Sin embargo, admite varias desventajas. Primero, la construcción de la matriz concatenada se traduce en un aumento significativo de la dimensionalidad así que correr los algoritmos de análisis se vuelve una tarea computacionalmente exigente. Además, no se consideran las diferencias de distribución de los datos de las diferentes ómicas ya que se aplican los mismos algoritmos a todos los datos. Finalmente, se puede dar más peso a la ómica con más características. Por ejemplo, en el caso de que haya datos de número de copias para 20.000 genes, de metilación del ADN para 8.000 regiones promotoras y de conteo de 25.000 transcritos, es probable que los datos de transcriptómica (25.000 transcritos) tengan más peso en el análisis.

- Integración tardía: Se aplican algoritmos adaptados a cada matriz que contienen datos ómicos de un solo tipo para obtener agrupaciones de las muestras en cada ómica. Luego, las representaciones obtenidas son integradas. Generalmente, el algoritmo de integración requiere los datos originales y no solo las representaciones obtenidas para cada ómica. Es el caso de PINS (*perturbation clustering for data integration and disease subtyping*) que realiza perturbaciones en los datos originales (Nguyen et al., 2017). Otro algoritmo de integración es CoCA (*cluster-of-clusters analysis*), ha sido utilizado para integrar representaciones obtenidas a partir de datos genómicos (mutaciones y número de copias de genes), epigenómicos (metilación del ADN), transcriptómicos (ARNm y miARN), y proteómicos de pacientes con cáncer de mama y cánceres ginecológicos (endometrio, cuello uterino y ovarios). La agrupación de pacientes resultante estaba fuertemente dominada por el tipo de tumor (Berger et al., 2018). La mayor desventaja de esta estrategia de integración es que se puede perder señales débiles en cada ómica durante la fase de integración.

- Métodos basados en similitud: Esta estrategia se parece a los métodos de integración tardía excepto que se calculan las similitudes entre muestras para cada ómica en lugar de llevar a cabo una agrupación. Las representaciones obtenidas antes de la integración exhiben nodos que corresponden a las muestras. El grosor de los enlaces que unen los nodos depende de la similitud entre las muestras para una ómica específica (Fig. 9). Este tipo de algoritmos puede admitir fácilmente datos categóricos y ordinales, por ejemplo, datos clínicos.

- Métodos basados en reducción de la dimensionalidad: En esta estrategia se busca reducir el número de variables aleatoria tratadas para cada ómica antes de llevar a cabo la integración. El método de aprendizaje de varios núcleos en entornos no supervisados rMKL-LPP (*regularized multiple kernel learning with locality preserving projections*) lleva a cabo una reducción de la dimensionalidad en los datos ómicos de entrada de modo que las similitudes entre cada muestra y sus vecinos cercanos se mantengan en una baja dimensionalidad (Speicher & Pfeifer, 2015). La aplicación de estos métodos a datos binarios discretos o a datos de conteo es técnicamente posible, pero a menudo inapropiada. Importantemente, el

uso de esta estrategia proporciona alguna interpretación para las características moleculares dominantes en cada grupo cuando los algoritmos basados en la similitud ignoran las características originales una vez que se calcularon las similitudes entre las muestras. Por lo tanto, los métodos de reducción de la dimensionalidad pueden ser útiles cuando se necesita una asociación entre grupos y características moleculares.

- Métodos basados en modelación probabilística: Estos métodos modelan la distribución probabilística de los datos. Algunos de estos métodos consideran que las muestras se originan en diferentes grupos y que cada grupo define la distribución de los datos, mientras que otros métodos no utilizan explícitamente la estructura de los grupos en el modelo. Una ventaja de esta estrategia es que permite incluir conocimiento biológico previo como parte del modelo al determinar las funciones de distribución de los datos.

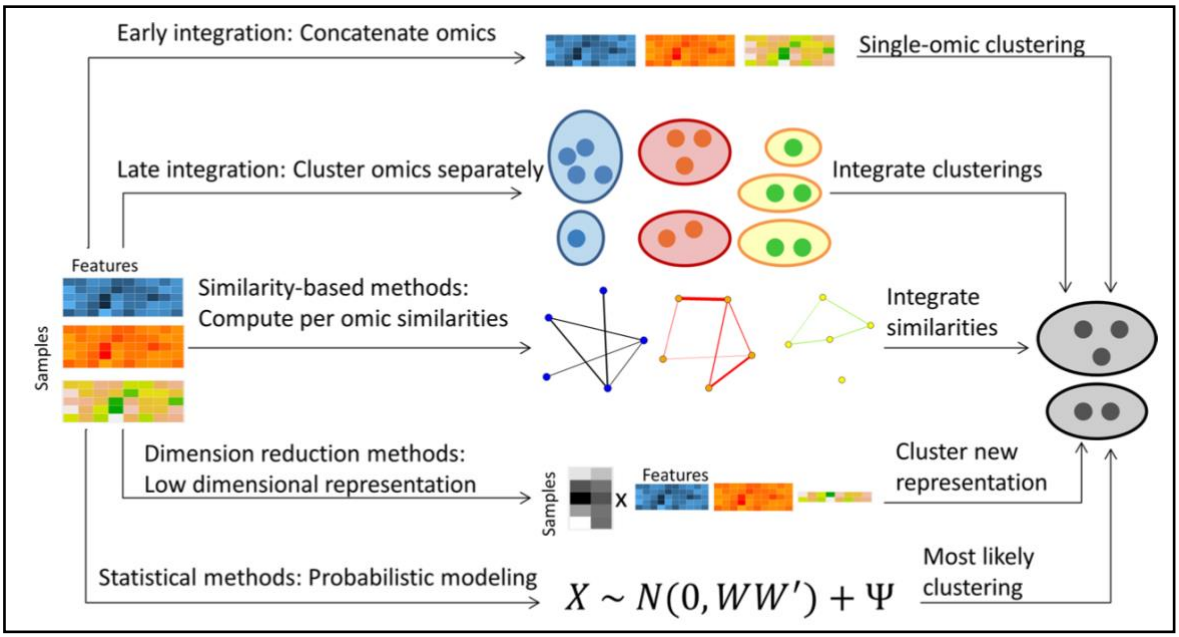


Figura 9: Descripción general de las estrategias de integración de datos multiómicos (Rappoport & Shamir, 2018).

Cabe resaltar que la integración de datos multiómicos presenta mayores desafíos debido a varios factores. La complejidad de los algoritmos de aprendizaje de máquina usados hace que los análisis se demoren mucho y que las necesidades en recursos computacionales sean muy exigentes. Además, la alta dimensionalidad, así como las

diferencias en la naturaleza (diferencias en la escala, por ejemplo) entre los conjuntos de datos ómicos constituyen otro obstáculo a la integración de datos ómicos.

2. Presentación del proyecto de investigación

2.1. Planteamiento del problema

En el estado actual de la investigación en ciencias biomédicas, pocos proyectos buscan desarrollar herramientas bioinformáticas de integración de datos multiómicos provenientes de estudios de distintos tipos de cáncer.

Lo anterior se debe a que los distintos tipos de datos ómicos presentan muchas diferencias respecto a su dimensionalidad y naturaleza. Por otra parte, las diferencias de mecanismos moleculares y de biomarcadores entre los distintos tipos de cáncer componen un obstáculo adicional (Martinez-Ledesma et al., 2015). Por lo tanto, es importante hacer un esfuerzo en la selección de los pacientes en el caso de la búsqueda de biomarcadores compartidos entre dos contextos patológicos mediante el aprovechamiento del conocimiento disponible sobre los factores causales y los mecanismos moleculares. Finalmente, la exigencia computacional de los modelos de integración de datos constituye otro desafío. Sin embargo, se puede lidiar con este inconveniente mediante una elección apropiada de la estrategia de integración y el uso de técnicas como la reducción de la dimensionalidad de los datos y la selección de variables antes de la fase de integración.

En consecuencia, la mayoría de los modelos actuales son elaborados a partir de conjuntos de datos ómicos de un solo tipo (datos genómicos comúnmente) y no se aprovechan los conjuntos de datos multiómicos disponibles en línea (Das et al., 2020). El estudio de datos genómicos (polimorfismos y variación del número de copias de genes) ha brindado mucho conocimiento sobre los genes involucrados en la carcinogénesis, la invasión local y la colonización metastásica. Sin embargo, estos estudios no toman en cuenta otros conjuntos de moléculas como el ARN o las proteínas que interactúan con el

genoma y modifican sus características funcionales. Asimismo, no toman en cuenta modificaciones epigenéticas como la metilación del ADN y modificaciones en las histonas involucradas en la expresión génica. Por lo tanto, los modelos basados en una sola ómica no logran desentrañar completamente la complejidad de una enfermedad como el cáncer. No contar con herramientas bioinformáticas de integración de datos multiómicos frena el avance de la investigación en el ámbito de la oncología.

De este problema surgió la siguiente preguntas de investigación:

- ¿Cómo integrar datos genómicos, epigenómicos y transcriptómicos provenientes de estudios de cáncer de endometrio y de cáncer de mama mediante un modelo de aprendizaje automático para identificar biomarcadores compartidos?

2.2. Justificación del problema

La búsqueda de biomarcadores tumorales se ha hecho principalmente a través del análisis de datos ómicos de un solo tipo (Rappoport & Shamir, 2018). Sin embargo, con el desarrollo de las ciencias ómicas, se planteó la hipótesis de que las señales relevantes para la detección de un cáncer o para el pronóstico pueden provenir de múltiples vías e involucrar una gran cantidad de biomarcadores ómicos, cuyo efecto puede ser visible solo cuando se agregan las ómicas (Reel et al., 2021). De ahí nació la necesidad de integrar datos multiómicos. Descubrir nuevos biomarcadores permite mejorar la prognosis, la clasificación, la terapia o la predicción de la supervivencia o del riesgo en el ámbito de la oncología (Abeel et al., 2010).

En el caso del cáncer de endometrio y del cáncer de mama, a pesar de que la tasa de supervivencia a cinco años sea alta (81.1% y 90.3%, respectivamente) y que existan tanto tratamientos (quimioterapia, radioterapia e inmunoterapia) como procedimientos quirúrgicos (histerectomía o mastectomía, respectivamente) eficientes, el entendimiento de las dinámicas moleculares asociadas a desequilibrios hormonales y compartidas entre ambas enfermedades podría permitir una mejora en la prevención y la detección de dichos cánceres. Esta tesis de maestría, aunque enfocada en los cánceres de endometrio y de mama, podría inscribirse como una prueba de concepto de integración de distintos tipos de datos provenientes de diferentes contextos patológicos en el campo de la oncología.

2.3. Objetivo general

El objetivo general de este proyecto de investigación es diseñar un modelo de aprendizaje automático de integración de datos genómicos, epigenómicos y transcriptómicos de pacientes con cáncer de endometrio y con cáncer de mama con el fin de identificar biomarcadores compartidos.

2.4. Objetivos específicos

Cuatro objetivos específicos fueron identificados:

- Seleccionar los conjuntos de datos genómicos, epigenómicos, transcriptómicos y clínicos de pacientes con cáncer de endometrio y con cáncer de mama susceptibles de compartir tanto mecanismos moleculares, celulares e inmunológicos como factores causales en la plataforma *The Cancer Genome Atlas*.
- Determinar el tipo de modelo de aprendizaje automático más adecuado recurriendo a una exploración de los datos seleccionados.
- Desarrollar un modelo de aprendizaje automático de integración de los datos multiómicos seleccionados.
- Caracterizar los biomarcadores identificados mediante el uso del método *Gene Ontology*.

Metodología

1. Metodología general

La primera fase del proyecto corresponde a la selección de los datos y está asociada al primer objetivo específico de selección de los conjuntos de datos genómicos, epigenómicos, transcriptómicos y clínicos de pacientes con cáncer de endometrio y con cáncer de mama susceptibles de compartir tanto mecanismos moleculares, celulares e inmunológicos como factores causales en la plataforma TCGA. Esta fase tiene tres actividades:

- Selección preliminar de los pacientes de interés mediante la explotación de los datos clínicos: se busca seleccionar los pacientes mujeres con cáncer ductal o lobulillar infiltrante positivos para los receptores de estrógenos y progesterona (cáncer de mama) y los pacientes con adenocarcinoma endometrioide (cáncer de endometrio).
- Selección de los pacientes con un perfil hormonal de interés mediante la explotación de los datos transcriptómicos: primero, se determina la distribución del nivel de expresión de los receptores de estrógenos y progesterona en los pacientes; luego, se seleccionan los pacientes cuyo tumor tiene una expresión de ambos receptores de hormonas superior al umbral de positividad establecido.
- Selección de los pacientes con una composición tumoral similar a nivel inmunológico mediante el uso de la herramienta *Cibersort*: se caracteriza el microentorno tumoral y el fenómeno de infiltración inmune en los pacientes.

El entregable de esta fase es un listado de los pacientes seleccionados.

La segunda fase del proyecto es el análisis exploratorio de los datos y busca cumplir con el segundo objetivo específico: determinar el tipo de modelo de aprendizaje automático

más adecuado recurriendo a una exploración de los datos seleccionados. Esta fase también está divida en tres actividades:

- Análisis de datos ómicos de un solo tipo mediante el uso de algoritmos existentes: se busca agrupar los pacientes seleccionados según su perfil genómico, epigenómico, o transcriptómico e identificar los patrones moleculares que rigen las agrupaciones obtenidas.
- Análisis correlacional entre los distintos tipos de datos ómicos: se busca establecer una relación matemática entre el número de copias de los genes o la metilación del ADN y la expresión génica.
- Análisis exploratorio de los datos clínicos: se identifican las variables clínicas completas para el conjunto de pacientes y se busca caracterizar los grupos de pacientes obtenidos durante el análisis de datos ómicos de un solo tipo por sus características clínicas.

Los entregables de esta fase son: un reporte del análisis exploratorio para cada tipo de datos y tres listados de biomarcadores resultantes de los análisis de datos ómicos de un solo tipo.

La tercera fase del proyecto es la integración de los datos. Es la fase crítica del proyecto y está asociada al tercer objetivo específico: desarrollar un modelo de aprendizaje automático de integración de los datos ómicos seleccionados. Esta fase contiene tres actividades:

- Recolección y preparación de los datos: primero, se establece un consenso sobre el nombre de las características (genes, regiones metiladas y transcritos) para poder relacionarlas posteriormente; luego, se da el formato adecuado a los datos para la implementación del modelo de aprendizaje automático.
- Desarrollo del modelo de aprendizaje automático: es un proceso cíclico, compuesto por tres etapas recurrentes, destinado a optimizar el modelo. La primera etapa es el entrenamiento del modelo. La segunda etapa corresponde a la prueba del modelo de aprendizaje automático, se evalúa el desempeño del modelo a través del cálculo de las métricas adecuadas. En la tercera etapa se lleva a cabo el ajuste del modelo, es decir, la optimización de los hiperparámetros del modelo (configuraciones utilizadas durante la etapa de entrenamiento).

- Identificación de los biomarcadores compartidos mediante el uso del modelo optimizado y análisis clínico de los grupos de pacientes obtenidos en el proceso.

Los entregables de la tercera fase son: un reporte de evaluación del modelo desarrollado y un listado de biomarcadores resultantes del uso del modelo.

La cuarta fase de este estudio es la caracterización de los resultados obtenidos. Busca cumplir con el cuarto objetivo específico: caracterizar los biomarcadores compartidos identificados mediante el uso del método *Gene Ontology*. Esta fase admite dos actividades:

- Caracterización biológica de los biomarcadores resultantes de los análisis de datos ómicos de un solo tipo y de los biomarcadores resultantes del uso del modelo de aprendizaje automático de integración de datos: se buscan los términos ontológicos (de la categoría proceso biológico) asociados con los biomarcadores identificados mediante el uso del método *Gene Ontology*.
- Comparación de los biomarcadores identificados con las dos metodologías (análisis de datos ómicos de un solo tipo y uso del modelo): se profundiza el análisis de los biomarcadores encontrados tanto en el análisis de datos ómicos de un solo tipo como con el uso del modelo.

El entregable de esta última fase es un reporte de la caracterización biológica de los biomarcadores identificados previamente (en la segunda y la tercera fase).

La metodología general de esta tesis fue representada de forma visual (Fig. 10). En las siguientes secciones se describe de manera detalla la aplicación de la metodología presentada.

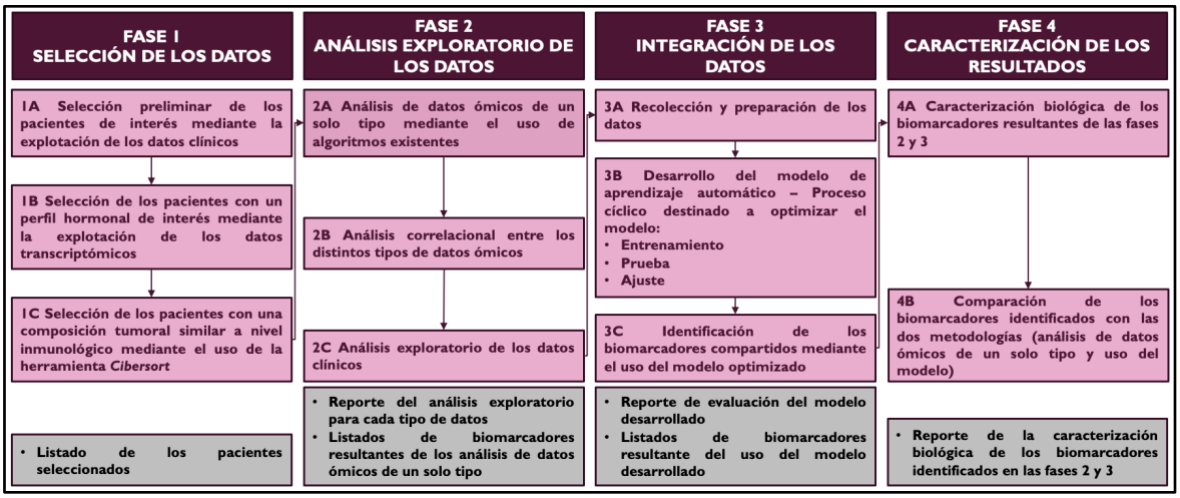


Figura 10: Metodología general del proyecto de investigación.

2. Métodos empleados

2.1. Preparación de los datos

Los datos usados para este proyecto provienen de la plataforma *The Cancer Genome Atlas* (TCGA). Dos conjuntos de datos fueron usados: el conjunto TCGA-BRCA, constituido por 1097 pacientes con cáncer de mama, y el conjunto TCGA-UCEC, constituido por 509 pacientes con cáncer de endometrio. Cuatro tipos de datos fueron recolectados directamente en la plataforma de TCGA:

- Datos clínicos: conjunto de variables asociadas a datos básicos y a características clínicas;
- Datos genómicos: variación del número de copias de los genes;
- Datos epigenómicos: metilación del ADN;
- Datos transcriptómicos: conteo de los transcritos.

Por un lado, los archivos de datos ómicos correspondientes a los análisis realizados en tumores sólidos primarios fueron conservados mientras que los archivos relativos a

tumores sólidos recurrentes, metástasis y tejido normal fueron eliminados. De la misma manera, en el caso de que hubiera varios archivos de análisis ómico del mismo tipo para un mismo paciente, todos los archivos fueron borrados y así el paciente eliminado. Este proceso dejó 1059 archivos genómicos, 1089 archivos epigenómicos y 1090 archivos transcriptómicos para el conjunto TCGA-BRCA; 532 archivos genómicos, 543 archivos epigenómicos y 541 archivos transcriptómicos para el conjunto TCGA-UCEC. Por otro lado, solo el archivo de datos clínicos más completo fue conservado. Se unieron las matrices de datos de los dos conjuntos de datos TCGA-BRCA y TCGA-UCEC para cada tipo de datos. Distintos procesos de preparación de los datos fueron aplicados a cada tipo de datos ómicos.

Datos clínicos

Las siguientes variables consideradas como teniendo un interés en cuanto a diagnóstico, pronóstico, clasificación o predicción de supervivencia fueron conservadas: *data_id*, *gender*, *menopause_status*, *race*, *ethnicity*, *history_other_malignancy*, *history_neoadjuvant_treatment*, *tumor_status*, *vital_status*, *last_contact_days_to*, *death_days_to*, *age_at_diagnosis*, *lymph_nodes_axillary_he_count*, *ajcc_nodes_pathologic_pn*, *ajcc_nodes_pathologic_pm*, *clinical_stage*, *er_status_by_ihc*, *pr_status_by_ihc*, *histologic_diagnosis*, *tumor_invasion_percent*, *lymph_nodes_pelvic_he_count*, *bmi_kg_m2_at_diagnosis* y *clinical_stage_agreg*. La variable *bmi_kg_m2_at_diagnosis* fue creada usando la fórmula:

$$\text{weight\_kg\_at\_diagnosis}/(\text{height\_cm\_at\_diagnosis}/100)^2 \qquad \textit{(Fórmula 1)}$$

La variable *clinical_stage_agreg* corresponde a una modificación de la variable *clinical_stage*: agrupación de los pacientes en Fase IA, IB y IC en Fase I; Fase IIA y IIB en Fase II; Fase IIIA, IIIB, IIIC, IIIC1 y IIIC2 en Fase III; y Fase IVA y IVB en Fase IV.

Datos genómicos

La matriz de datos genómicos antes del procesamiento tenía un 0.61% de valores ausentes NAs y 60623 variables. 3179 variables tenían por lo menos un valor ausente. Esto corresponde a un 5.2% de las variables. Dado que la matriz de datos genómicos tiene una dimensionalidad muy alta, estas variables fueron eliminadas. La matriz de datos genómicos final está compuesta por 57444 variables (sin valores ausentes) que son genes identificados por el ID de la base de datos Ensembl.

Datos epigenómicos

La matriz de datos epigenómicos inicial está compuesta por 25978 variables que son dinucleótidos CG ubicados en todo el genoma. Los valores corresponden al nivel de metilación normalizado de los dinucleótidos CG. Los dinucleótidos CG fueron asociados a los genes cercanos bajo la hipótesis que su nivel de metilación puede afectar el reclutamiento de la maquinaría de transcripción y así afectar la expresión génica de los genes del entorno. Así fue creada una segunda matriz de datos epigenómicos compuesta por 24796 variables que son parejas dinucleótido CG/gen. Es importante notar que algunos genes están asociados a varios dinucleótidos CG. Finalmente, la matriz de datos epigenómicos fue modificada para asignar un nivel de metilación promedio a los genes y así eliminar los dinucleótidos CG de la matriz (Fig. 11).

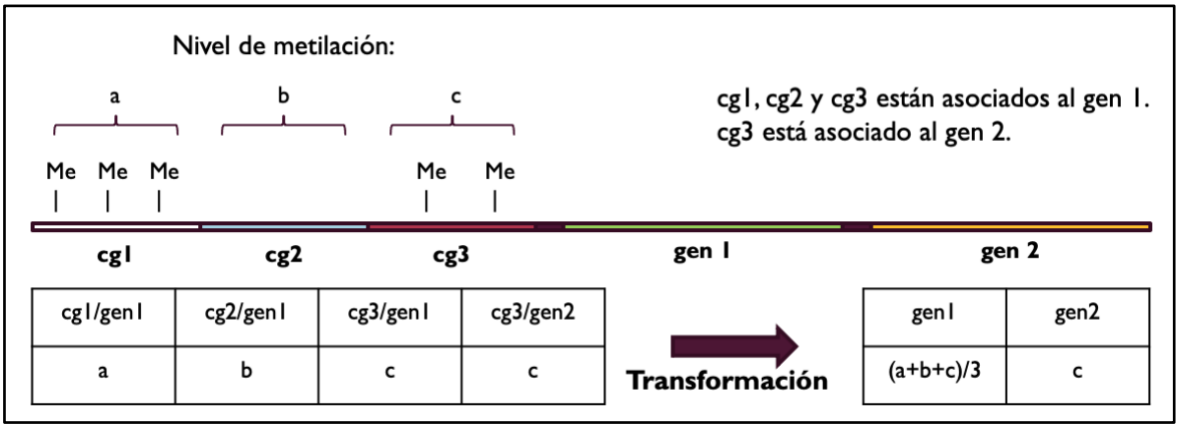


Figura 11: Transformación de la matriz de datos epigenómicos.

Esta matriz de datos epigenómicos antes del procesamiento tenía un 0.62% de valores ausentes NAs y 12746 variables (genes). 1500 variables tenían por lo menos un valor ausente. Esto corresponde a un 11.8% de las variables. Dado que la matriz de datos epigenómicos no tiene una dimensionalidad tan alta como las matrices de datos genómicos y transcriptómicos, estas variables no fueron eliminadas para tratar de conservar el número máximo de variables compartidas entre los distintos tipos de datos ómicos. Los valores ausentes fueron imputados mediante una estrategia basada en el modelo de PCA (*Principal Component Analysis*) llamada *imputePCA* en R. Esta estrategia busca imputar datos minimizando la modificación de los componentes principales computados a partir del conjunto de datos. Primero, se aplicó el algoritmo PCA a la matriz de datos epigenómicos sin las variables que presentan valores ausentes. PCA es una técnica estadística de reducción de la dimensionalidad la cual transforma el conjunto de datos inicial en otro conjunto de nuevas variables llamadas componentes principales. Los componentes

principales son combinaciones no lineales de las variables del conjunto de datos inicial, que no son correlacionados entre sí. Se encontró que 406 componentes principales son necesarios para representar el 90% de la varianza de los datos. Luego, se aplicó la estrategia *imputePCA* a la matriz de datos epigenómicos con valores ausentes. Es una estrategia iterativa la cual se divide en 3 etapas:

- Inicialización: Los valores ausentes son reemplazados por la media de la variable considerada. A partir de ahí los valores ausentes pueden ser calificados de valores imputados.
- PCA: Se corre el algoritmo de PCA sobre el conjunto de datos obtenido.
- Actualización: Se proyectan los valores imputados sobre los componentes principales para actualizarlos. En este caso, los valores imputados fueron proyectados sobre 406 componentes principales para su actualización.

Las etapas de PCA y Actualización se repiten hasta llegar a la convergencia de los valores imputados. La matriz de datos epigenómicos final está compuesta por 12746 variables (sin valores ausentes) que son genes identificados por el ID de la base de datos Ensembl.

Datos transcriptómicos

Se aplicaron procesos de normalización y suavización al conteo de transcritos (*unstranded*) de la matriz de datos transcriptómicos la cual no presenta valores ausentes. Primero, se usó el método de normalización TMM (*Trimmed mean of M values*) el cual se basa en el supuesto que “la mayoría de los genes no se encuentran diferencialmente expresados” y calcula los factores de normalización para llevar a una misma escala los valores de expresión promedio entre las muestras. Las unidades de CPM (*counts per million*) fueron usadas. Este proceso permite una comparación de los valores de expresión de los diferentes genes tanto dentro de un mismo paciente como entre diferentes pacientes (P. Li et al., 2015). Luego, los datos normalizados fueron suavizados usando la función logarítmica base 2. Dado que no existe el logaritmo de 0, se sumó 1 a los valores normalizados antes de aplicar la función logarítmica. La matriz de datos transcriptómicos final está compuesta por 60660 variables (sin valores ausentes) que son genes identificados por el ID de la base de datos Ensembl.

2.2. Selección de los pacientes con una expresión de ESR1 y PGR intermedia o alta

Los pacientes fueron visualizados en un espacio de dos dimensiones según su expresión de los genes ESR1 y PGR. Se llevó a cabo una agrupación usando el algoritmo *K-Medoids* con un número de grupos (*clusters*) igual a 3. Este número de *clusters* fue determinado calculando el coeficiente de Silueta. El coeficiente de Silueta es una métrica utilizada para calcular el número óptimo de *clusters* y para evaluar la calidad del agrupamiento. Para cada objeto x, se calculan los valores $a(x)$ y $b(x)$. $a(x)$ es la distancia promedio del objeto x a todos los demás objetos de este *cluster* mientras que $b(x)$ es la distancia promedio del objeto x a todos los demás objetos del *cluster* más cercano. El coeficiente de Silueta para el objeto x, $s(x)$, es dado por:

$$s(x) = (b(x) - a(x)) / \max[a(x) - b(x)] \quad \textit{(Fórmula 2)}$$

El coeficiente de Silueta para todo el agrupamiento corresponde al promedio de $s(x)$ para todos los objetos del conjunto de datos. Se busca maximizar el coeficiente de Silueta promedio.

El algoritmo *K-Medoids* sigue un proceso iterativo dividido en cuatro fases:

1. k centroides iniciales son generados dentro del conjunto de datos.
2. Paso de asignación: Los objetos son asignados al centroide más cercano. La métrica de distancia usada es la distancia euclidiana. En consecuencia, se generan k grupos.
3. Paso de actualización: El centroide de cada grupo es actualizado tomando como nuevo centroide el punto cuyas coordenadas corresponden a la mediana de las coordenadas de los objetos que pertenecen al grupo.
4. Se repiten los pasos de asignación y de actualización hasta llegar a la convergencia, es decir hasta que la composición de los grupos sea estable.

Los dos *clusters* correspondientes a una expresión intermedia o alta de ESR1 y PGR fueron extraídos. Finalmente, los pacientes con una expresión de ESR1 o PGR atípica baja en los dos *clusters* seleccionados fueron eliminados. Se usó la siguiente fórmula para la definición de los valores atípicos bajos:

$$q < Q_1 - 1.5 * \text{Rango Intercuartil} \quad \textit{(Fórmula 3)}$$

2.3. Análisis de la infiltración inmunológica

Para el análisis de la infiltración inmunológica, el modelo de aprendizaje informático *Cibersort* fue implementado en modo absoluto con 100 permutaciones. La matriz LM22 fue usada para el entrenamiento del modelo (Newman et al., 2015). *Cibersort* determina la abundancia absoluta de 22 tipos de células del sistema inmune: linfocitos T, linfocitos B, células NK, macrófagos, células dendríticas, eosinófilos, neutrófilos, monocitos, mastocitos y células plasmáticas en los pacientes a partir de los datos transcriptómicos (Chen et al., 2018). *Cibersort* dio una matriz con una estimación de la abundancia absoluta de los 22 tipos de células para cada paciente. Se llevó a cabo una reducción de la dimensionalidad mediante PCA para visualizar los datos. Finalmente, los pacientes con una infiltración inmunológica total (suma de la abundancia absoluta de cada tipo de células del sistema inmune) atípica extrema fueron eliminados. Se usó la siguiente fórmula para la definición de los valores atípicos extremos:

$$q < Q_1 - 3 * \text{Rango Intercuartil o } q > Q_3 + 3 * \text{Rango Intercuartil} \qquad \textit{(Fórmula 4)}$$

2.4. Análisis de datos ómicos de un solo tipo

El mismo protocolo fue usado para el análisis de los tres tipos de datos ómicos: reducción de la dimensionalidad usando UMAP (*Uniform Manifold Approximation and Proyection*), agrupación de los pacientes usando *K-Means*, implementación del algoritmo *Mutual Information* para determinar la dependencia entre cada variable inicial y las variables sintéticas UMAP1 y UMAP2 y extracción de 100 biomarcadores por *cluster*.

UMAP es un algoritmo de reducción de la dimensionalidad cuyo objetivo es crear una representación de baja dimensionalidad de los datos (dos variables sintéticas: UMAP1 y UMAP2) que preserva los *clusters* de alta dimensionalidad y su relación entre sí y que también permita una visualización de los *clusters*. La idea principal del algoritmo UMAP es construir una representación gráfica de alta dimensionalidad de los datos y luego optimizar la representación gráfica de baja dimensionalidad maximizando la similitud estructural entre las dos representaciones. La construcción de la representación gráfica de alta dimensionalidad de los datos se hace a partir de un simplex: gráfico ponderado donde los pesos representan la probabilidad de que dos objetos estén conectados. La conectividad entre los objetos es determinada extendiendo un radio hacia fuera desde cada objeto, conectando los objetos cuando esos radios se superponen. La elección del radio se hace

localmente basado en la distancia al enésimo vecino más cercano de cada objeto. El tamaño del vecindario local (número de objetos vecinos) es clave dado que, entre más alto, se conservan las estructuras globales de los datos, y, entre más bajo, se da prioridad a las estructuras locales en la proyección final. Otro parámetro clave de UMAP es la distancia mínima entre los objetos en el espacio de baja dimensionalidad. Este parámetro controla la fuerza con la que UMAP agrupa los objetos. Valores bajos conducen a representaciones más compactas mientras que valores altos dan agrupaciones más flexibles enfocándose en la preservación de la estructura topológica global. En este caso, se usaron valores intermedios para el número de objetos vecinos: 15, y para la distancia mínima entre los objetos en el espacio de baja dimensionalidad: 0.1, para obtener un balance satisfactorio entre conservación de las estructuras globales y locales.

La agrupación de los pacientes se realizó en los conjuntos de datos de dos dimensiones (UMAP1 y UMAP2) obtenidos por UMAP mediante *KMeans*. *KMeans* es un algoritmo semejante a *KMedoids*: la diferencia es que el paso de actualización de los centroides se hace tomando como nuevo centroide la posición promedio de los objetos perteneciendo al grupo considerado, es decir que se usa la media y no la mediana para actualizar el centroide. El número óptimo de *clusters* fue determinado calculando el coeficiente de Silueta. El número óptimo de *clusters* encontrado fue de 2, 3 y 6 para los datos transcriptómicos, epigenómicos y genómicos, respectivamente.

La dependencia entre las variables sintéticas UMAP1 y UMAP2 y las variables iniciales que corresponden a los genes fue evaluada con el algoritmo *Mutual Information*. La información mutua entre dos variables X e Y llamada $I(X; Y)$ corresponde a la cantidad compartida de información entre X e Y. $I(X; Y)$ se calcula sumando la entropía de X y la entropía de Y, y restándole la entropía conjunta, como se expresa en la siguiente fórmula:

$$I(X; Y) = H(X) + H(Y) - H(X, Y) \quad \textit{(Fórmula 5)}$$

El cálculo de las entropías se hace mediante una estrategia de *binning* o agrupamiento en porciones de los datos ordenados. Se hace un conteo de los objetos de X o Y en cada *bin*. La entropía y la entropía conjunta se calculan usando las siguientes fórmulas:

$$H = -\sum_{i=1}^{n} p_i \log(p_i) \quad \textit{(Fórmula 6)}$$

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} P(x, y) \log_2[P(x, y)] \quad \textit{(Fórmula 7)}$$

La información mutua entre dos variables toma valores entre 0 y 1. 0 corresponde a una independencia total, mientras que 1 corresponde a una dependencia total entre las dos

variables. Este análisis fue llevado a cabo para todos los *clusters*. Las variables iniciales fueron ordenadas por su puntaje de información mutua máximo con las variables sintéticas UMAP1 o UMAP2. Las 100 variables iniciales con mayor puntaje fueron extraídas. Estos genes son considerados como biomarcadores.

Además, se calculó para cada variable genómica, epigenómica o transcriptómica el promedio de número de copias, metilación del ADN o expresión génica, respectivamente, en cada *cluster*. Estos datos fueron visualizados en un diagrama de cajas. Los promedios fueron analizados estadísticamente. El análisis está compuesto por dos etapas: la comprobación de supuestos y la prueba estadística en sí. Los supuestos de normalidad y de homogeneidad de las varianzas fueron comprobados en cada grupo (*cluster*) mediante el uso de las pruebas de Anderson-Darling para la normalidad y el *F-test* (2 *clusters*) o la prueba de Levene (más de 2 *clusters*) para la homogeneidad de las varianzas. Dado que las hipótesis de normalidad y homogeneidad de las varianzas fueron descartadas, la prueba no paramétrica de Mann-Whitney-Wilcoxon para datos pareados fue usada entre los distintos *clusters*.

2.5. Análisis correlacional entre los distintos tipos de datos ómicos

Se realizó un análisis correlacional, por un lado, entre los datos genómicos y transcriptómicos, y, por otro lado, entre los datos epigenómicos y transcriptómicos. El análisis correlacional se hizo entre los valores tomados por un gen para el número de copias o la metilación del ADN y los valores tomados por este mismo gen para la expresión génica. Por lo tanto, este análisis fue realizado para las 12594 variables compartidas entre los distintos tipos de datos ómicos.

Primero, se modificaron los conjuntos de datos genómicos y transcriptómicos para mejorar el análisis correlacional entre estos dos tipos de datos. Los datos genómicos originales – que comportan 913 registros (pacientes) – fueron reducidos a 81 valores entre 0 y 80 para todas las variables dado que en los datos originales el menor y el mayor número de copias para un gen eran de 0 y 80, respectivamente. En cuanto a los datos transcriptómicos, para cada variable, se calculó la mediana de los valores transcriptómicos originales para cada número de copias. De la misma manera, los datos transcriptómicos

modificados presentan 81 registros, cada registro corresponde a un número de copias entre 0 y 80 y el valor a la expresión mediana del gen para este número de copias. Importantemente, el análisis correlacional entre los datos epigenómicos y transcriptómicos se hizo a partir de los 913 registros (pacientes) de los datos originales.

Se llevó a cabo una prueba de Shapiro-Wilk para averiguar la normalidad de las variables. Dado que ninguna pareja (genómica/transcriptómica o epigenómica/transcriptómica) de genes mostró una distribución normal, el coeficiente de correlación de Spearman (medida no paramétrica) fue usado para el análisis correlacional. Los coeficientes de Spearman fueron calculados para cada gen y luego visualizados en diagramas de caja.

2.6. Análisis exploratorio de los datos clínicos

Las variables numéricas y categóricas fueron separadas para el análisis exploratorio de los datos clínicos. El análisis exploratorio de los datos clínicos fue realizado a nivel general y a nivel de *cluster*. A nivel general, los datos clínicos fueron simplemente visualizados en diagramas de caja para las variables numéricas y diagramas de barra para las variables categóricas y ninguna prueba estadística fue llevada a cabo. A nivel de *cluster*, los datos clínicos fueron visualizados de la misma manera, pero separando los pacientes por *cluster*. Se usó el análisis de chi-cuadrado para las variables categóricas.

El análisis estadístico de las variables numéricas está compuesto por dos etapas: la comprobación de supuestos y la prueba estadística en sí. Los supuestos de normalidad y de homogeneidad de las varianzas fueron comprobados en cada grupo (*cluster*) mediante el uso de las pruebas de Shapiro-Wilk para la normalidad y el *F-test* (2 *clusters*) o la prueba de Levene (más de 2 *clusters*) para la homogeneidad de las varianzas. Estos resultados condicionan la elección de la prueba estadística para comparar los *clusters*. Las pruebas paramétricas t de Student (dos *clusters*) o ANOVA (*Analysis of Variance*; más de dos *clusters*) fueron usadas cuando la normalidad (en todos los *clusters*) y la homogeneidad de varianzas fueron comprobados. En el caso contrario, se usaron las pruebas no paramétricas de Mann-Whitney-Wilcoxon (2 *clusters*) o de Kruskal-Wallis (más de 2 *clusters*). Cuando diferencias significativas fueron encontrados por las pruebas ANOVA o de Kruskal-Wallis, las pruebas *post hoc* de Tukey y de Mann-Whitney-Wilcoxon,

respectivamente, fueron empleadas para determinar entre cuales *clusters* se encuentran las diferencias significativas.

Es importante notar que en el caso de los *clusters* mixtos, es decir compuestos tanto por pacientes con cáncer de mama como por pacientes con cáncer de endometrio, los mismos análisis fueron realizados, pero separando los pacientes según su tipo de cáncer.

2.7. Implementación del modelo de aprendizaje automático DGCCA

La integración de los datos ómicos se hizo mediante el uso del algoritmo de aprendizaje automático DGCCA (*Deep Generalized Canonical Correlation Analysis*) (Fig. 12; Benton et al., 2019).

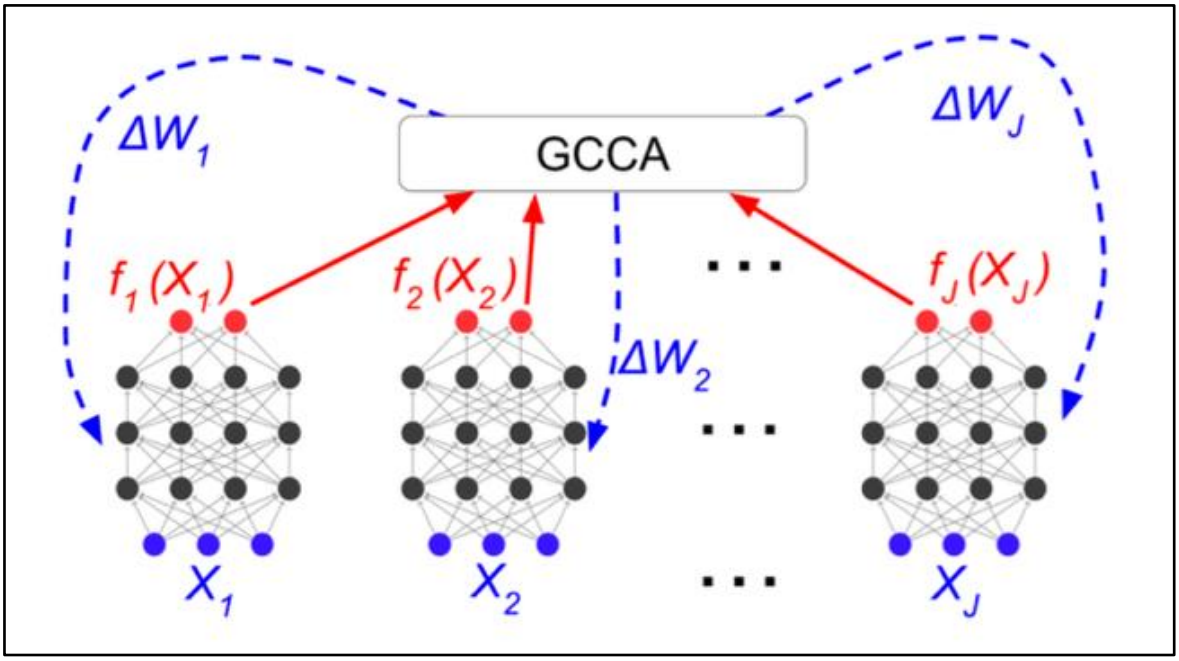


Figura 12: Implementación del algoritmo DGCCA para J vistas (Benton et al., 2019).

El algoritmo de aprendizaje multi-vista no supervisado DGCCA lleva a cabo el aprendizaje de un mapeo no lineal para cada vista que maximiza la correlación entre las vistas. DGCCA pasa los vectores de cada vista a través de múltiples capas ocultas de transformaciones no lineales y propaga hacia atrás el gradiente de la función objetivo para ajustar la red de cada vista. La función objetivo de DGCCA es la siguiente:

$$\text{minimize} \sum_{i=1}^{m} ||G - U_i^T f_i(X_i)||_F^2 \text{ con } GG^T = I \qquad \textit{(Fórmula 8)}$$

- X_J corresponde a la vista de entrada J.
- $f_J(X_J)$ corresponde a la salida de la última capa de la red neuronal profunda para la vista J.
- U_J es la transformación lineal de $f_J(X_J)$.
- G es la representación integrada de las vistas de entrada.
- W_J corresponde a la matriz de pesos para cada capa k de la red neuronal profunda para la vista J. W_J también es denominado θ_J o parámetros de f_J.

La salida de cada capa k de la red neuronal para la vista J es llamada h^J_k y su fórmula es la siguiente:

$$h^J_k = s(w^J_k\, h^J_{k-1}) \qquad \textit{(Fórmula 9)}$$

s corresponde a la función de activación del nodo. Puede ser la función ReLu o *Sigmoid*.

El ajuste del modelo se hace entrenando U_J y G para reducir el error de reconstrucción de GCCA (*Generalized Canonical Correlation Analysis*), y, para actualizar θ_J, los gradientes son propagados hacia atrás a través de la red neuronal. El gradiente que se propaga a $f_J(X_J)$ es definido de la siguiente manera:

$$2U_J G - 2U_J U_J^T f_J(X_J) \qquad \textit{(Formula 10)}$$

θ_J es actualizado por retro propagación para minimizar la función objetivo de DGCCA. A medida que θ_J se actualiza, el valor de $f_J(X_J)$ cambia. Por lo tanto, para resolver la función objetivo de DGCCA, se realizan alternativamente la actualización de U_J y G, y la actualización de θ_J.

Los hiperparámetros del algoritmo DGCCA fueron ajustados empíricamente. A continuación, tenemos el detalle para la versión de DGCCA con el error de reconstrucción más bajo al finalizar.

- Tasa de aprendizaje: 0.01
- Tamaño del *batch*: 1000
- rcov (pequeña cantidad de regularización aplicada a la matriz de covarianza de cada vista proyectada): 0.000001
- Algoritmo de optimización: Adam
- Regularización L1: 0.0001
- Regularización L2: 0.01
- Número de épocas: 200

- Número de vistas: 3
- Vistas: Datos genómicos, Datos epigenómicos y Datos transcriptómicos. Solo las variables compartidas entre las tres vistas fueron conservadas. Cada matriz de datos está compuesta por 913 registros y 12594 variables. Se realizó un escalamiento lineal de las tres vistas entre 0 y 1 usando el método MinMax:

  $X_{escala} = (X - X_{min})/(X_{max} - X_{min})$ (Fórmula 11)

  El escalamiento de los datos permite reducir el sesgo de la diferencia de naturaleza entre los tres tipos de datos ómicos.
- Arquitectura de las redes neuronales: Tres capas (capa de entrada, una capa oculta, capa de salida) con la siguiente arquitectura: [12594, 250, 500]
- k (dimensionalidad de la representación integrada G): 500
- Función de activación: ReLu
- Pesos atribuidos a cada vista: [1 1 1]

La representación integrada G fue analizada usando los mismos pasos que para las matrices de datos ómicos: reducción de la dimensionalidad usando UMAP, agrupación de los pacientes por *KMeans* (3 *clusters*), computación del puntaje de información mutua entre las variables iniciales de los conjuntos de datos genómicos, epigenómicos y transcriptómicos después del escalamiento y las variables sintéticas UMAP1 y UMAP2, identificación de los biomarcadores y análisis comparativo de las variables clínicas entre *clusters*. La identificación de los biomarcadores se hizo calculando el puntaje de información mutua entre las variables sintéticas UMAP1 y UMAP2 y las 12594 variables de cada vista (genómica, epigenómica y transcriptómica) después del escalamiento lineal. Para cada uno de los 12594 genes considerados fue calculada la media del puntaje de información mutua obtenida para cada vista tanto para UMAP1 como para UMAP2. Como se describió previamente, se calculó para cada variable genómica, epigenómica o transcriptómica el promedio de número de copias, metilación del ADN o expresión génica, respectivamente, en cada *cluster* de la representación integrada G. Estos datos fueron visualizados en diagramas de caja. Se hicieron las respectivas pruebas estadísticas para comparar el promedio de número de copias, metilación del ADN o expresión génica, para cada variable, entre *clusters*. Por otro lado, la evolución del error de entrenamiento (error de reconstrucción) a través de las 200 épocas fue visualizado en un diagrama.

2.8. Caracterización biológica de los biomarcadores

El método de ontología de genes fue usado para caracterizar las 14 listas de 100 biomarcadores (6 listas para los datos genómicos, 3 listas para los datos epigenómicos, 2 listas para los datos transcriptómicos y 3 listas para los datos transcriptómicos). Un enriquecimiento de los términos GO (*Gene Ontology*) relativos a procesos biológicos a partir de los biomarcadores fue llevado a cabo en la base de datos DAVID (*Database for Annotation, Visualization and Integrated Discovery*). Los 10 términos GO enriquecidos con menor *p-value* fueron representados junto a los biomarcadores asociados en diagramas de red. Los términos GO más interesantes – es decir, relacionados con procesos biológicos desregulados en el contexto del cáncer – así como los biomarcadores asociados fueron seleccionados manualmente.

Se realizó un análisis comparativo entre *clusters* de los biomarcadores genómicos, epigenómicos y transcriptómicos seleccionados previamente. Los resultados fueron visualizados en diagramas de caja. El análisis estadístico fue realizado como descrito anteriormente en el análisis de las variables numéricas de los datos clínicos. Finalmente, los biomarcadores comunes a varios tipos de datos fueron representados en diagramas de Venn.

Resultados

A continuación, se presentan los resultados de cada fase y sus respectivas actividades como establecido en la metodología general. Para mayor legibilidad, se usan figuras para la visualización de los resultados. Las diferencias significativas son cuantificadas por su *p-value*: en algunos casos, las pruebas estadísticas no pudieron arrojar un *p-value* exacto y solo expusieron un *p-value* menor a 2.2e-16, el cual indica una diferencia significativa.

1. Selección de los datos

Tres filtros fueron aplicados a los conjuntos de datos iniciales compuestos por 1097 pacientes de cáncer de mama y 509 pacientes de cáncer de endometrio. Se llevó a cabo una selección preliminar de los pacientes de interés mediante la explotación de los datos clínicos, una selección de los pacientes con un perfil hormonal de interés (expresión intermedia o alta de los receptores a los estrógenos y a la progesterona) mediante la explotación de los datos transcriptómicos y una selección de los pacientes con una composición tumoral semejante. A continuación, se describe cada filtro y el número de pacientes eliminados en cada uno. El objetivo de estos filtros era seleccionar los pacientes con cáncer ductal o lobulillar infiltrante (cáncer de mama) y los pacientes con adenocarcinoma endometrioide (cáncer de endometrio) positivos para los receptores de estrógenos y progesterona y con una composición tumoral homogénea a nivel inmunológico.

1.1. Selección preliminar de los pacientes de interés mediante la explotación de los datos clínicos

Con el objetivo de disminuir la heterogeneidad clínica y así eliminar algunos sesgos, los pacientes fueron seleccionados sucesivamente según su sexo (Fig. 13), la histología de su tumor (Fig. 14) y la expresión de los receptores de estrógenos y progesterona medida por inmunohistoquímica (Fig. 15-16). Solo los pacientes mujeres fueron conservados en el filtro del sexo. En esta etapa, 12 pacientes hombres con cáncer de mama fueron eliminados del conjunto de datos (Fig. 13).

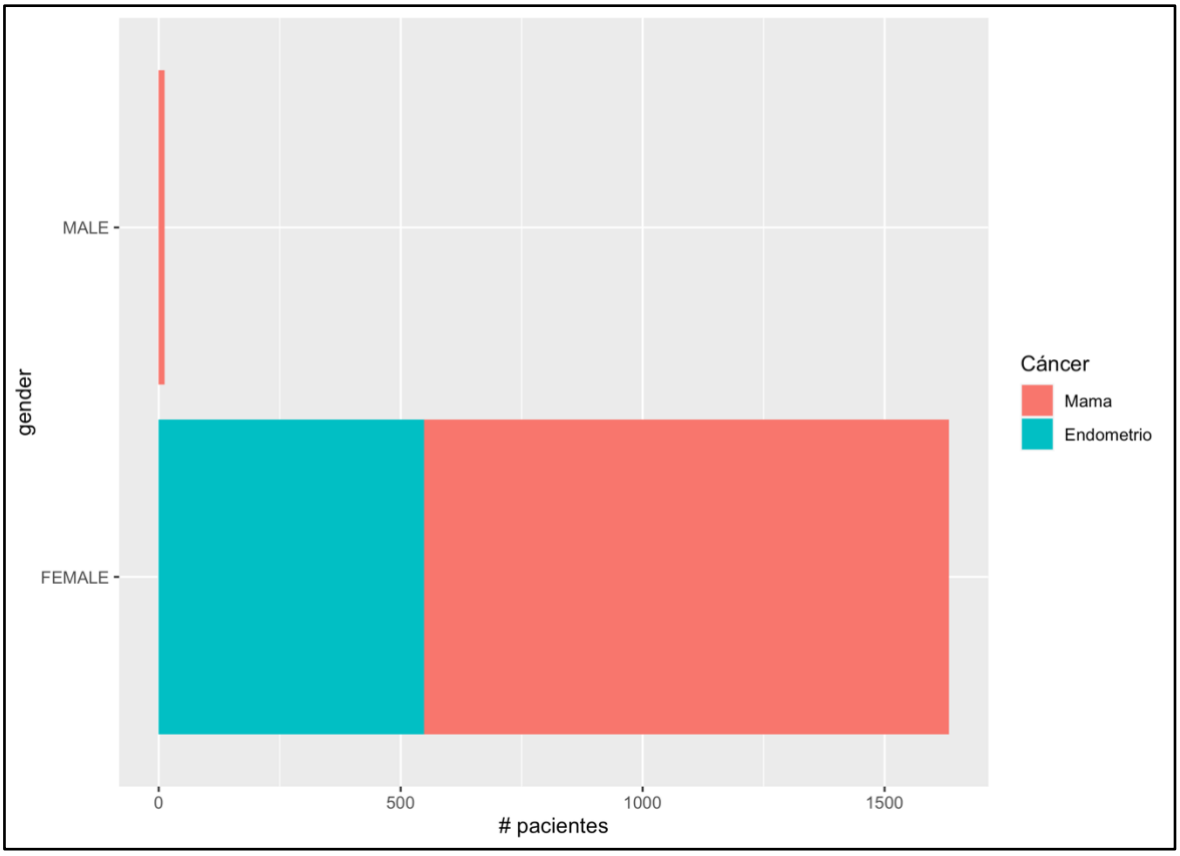


Figura 13: Sexo de los pacientes de cáncer de mama y de cáncer de endometrio.

Para conseguir más homogeneidad en las características clínicas de los pacientes y eliminar los pacientes con carcinoma seroso que corresponde al cáncer de endometrio tipo II, el cual no está asociado con hiperestrogenismo, los pacientes con las siguientes histologías tumorales fueron conservados:

- Carcinoma lobulillar infiltrante (cáncer de mama);
- Carcinoma ductal infiltrante (cáncer de mama);
- Adenocarcinoma endometrioide (cáncer de endometrio).

Este filtro eliminó 98 pacientes con cáncer de mama y 137 pacientes con cáncer de endometrio (Fig. 14).

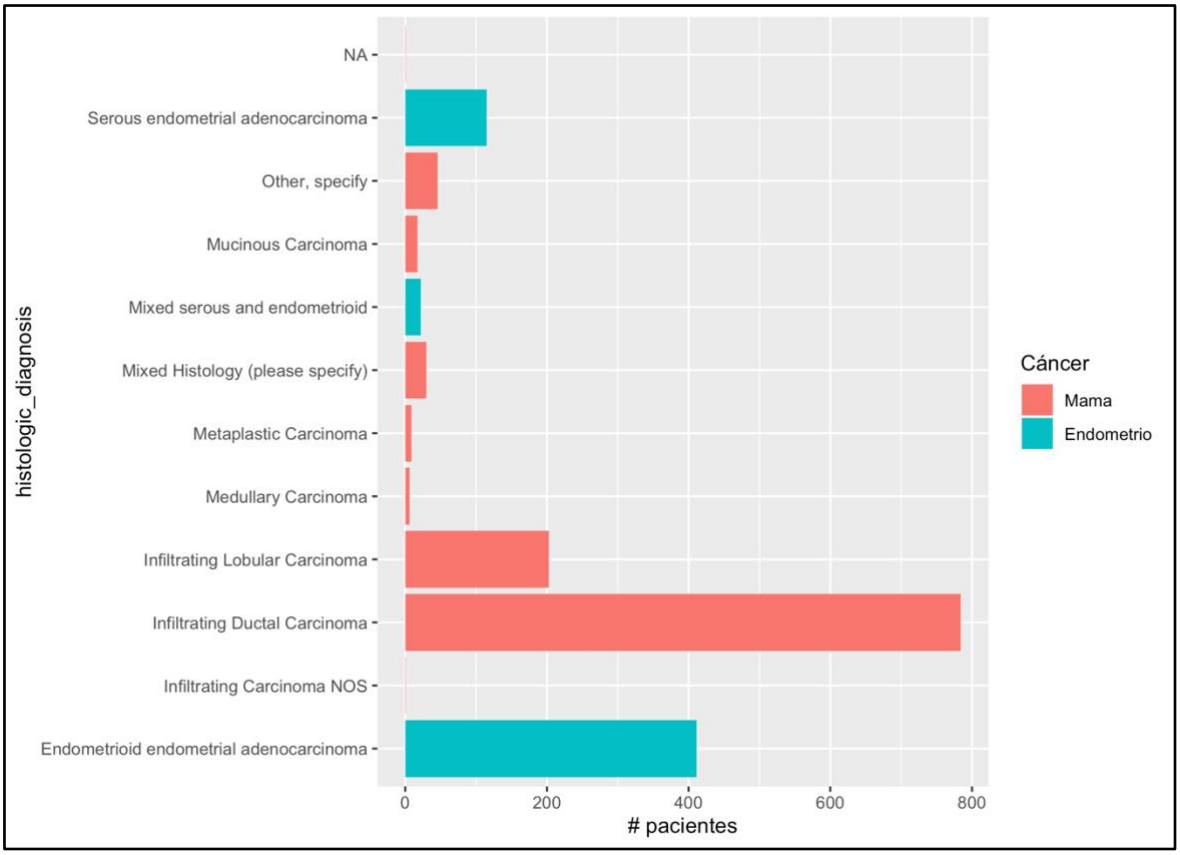


Figura 14: Histología del tumor de los pacientes de cáncer de mama y de cáncer de endometrio.

Finalmente, dado que este proyecto está enfocado en los pacientes con cáncer de mama y cáncer de endometrio presentando hiperestrogenismo, los pacientes cuyo tumor es negativo para la expresión de los receptores de estrógenos o progesterona medida por inmunohistoquímica fueron eliminados. Esta información solo estaba presente para los pacientes con cáncer de mama, 305 de ellos fueron eliminados (Fig. 15-16).

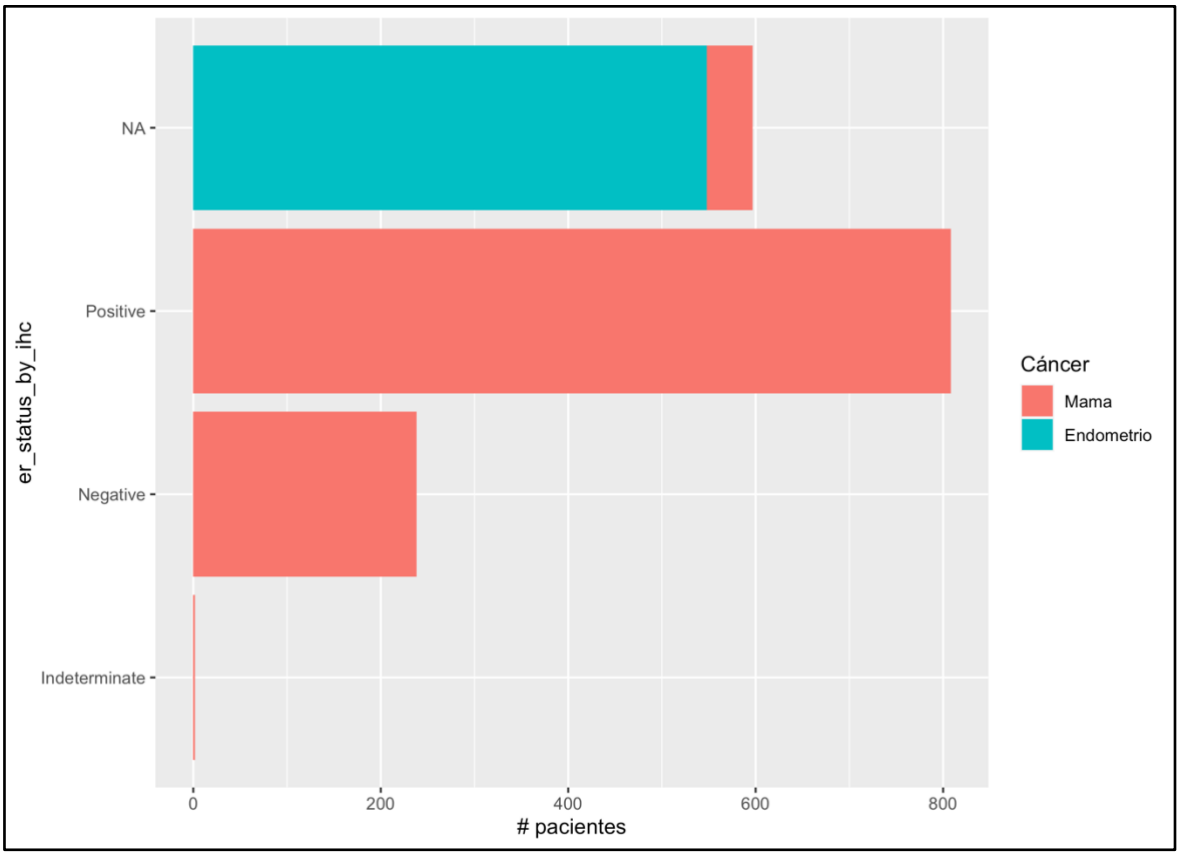


Figura 15: Nivel de expresión del receptor de estrógenos en el tumor de los pacientes de cáncer de mama y de cáncer de endometrio medido por inmunohistoquímica. La expresión del receptor de estrógenos por el tumor de los pacientes con cáncer de endometrio y por el tumor de algunos pacientes con cáncer de mama no fue medida por inmunohistoquímica.

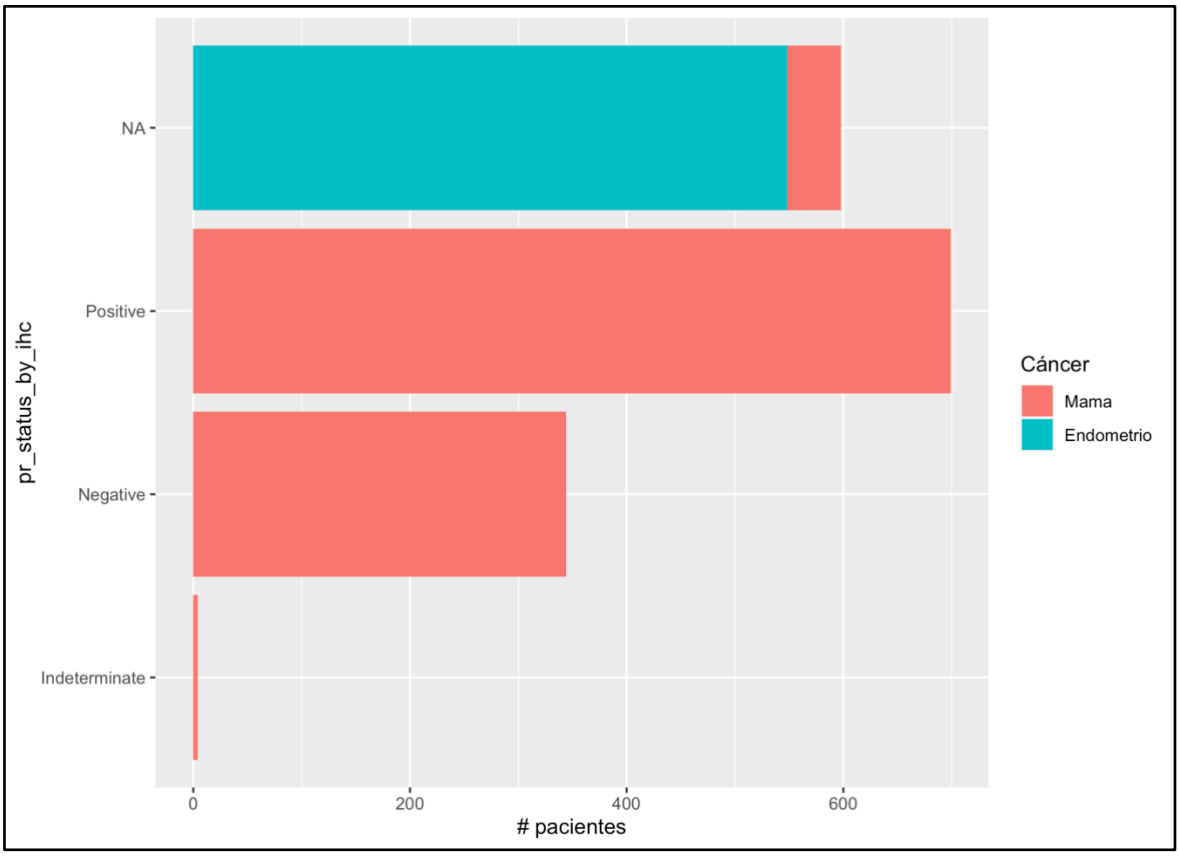


Figura 16: Nivel de expresión del receptor de progesterona en el tumor de los pacientes de cáncer de mama y de cáncer de endometrio medido por inmunohistoquímica. La expresión del receptor de progesterona por el tumor de los pacientes con cáncer de endometrio y por el tumor de algunos pacientes con cáncer de mama no fue medida por inmunohistoquímica.

Los diferentes filtros clínicos aplicados a los pacientes iniciales resultaron en la selección de 1093 pacientes: 682 con cáncer de mama y 411 con cáncer de endometrio. Debido a la ausencia de medida de la expresión de los receptores de estrógenos y progesterona por inmunohistoquímica para el tumor de los pacientes con cáncer de endometrio, fue necesario implementar otro método para poder evaluar la expresión de los receptores a las hormonas femeninas.

1.2. Selección de los pacientes con un perfil hormonal de interés mediante la explotación de los datos transcriptómicos

ESR1 y PGR son los genes que codifican para los receptores de los estrógenos (ERα) y de la progesterona (PR), respectivamente. Estos receptores pertenecen a la familia de los receptos nucleares, actúan como factores de transcripción cuando son

estimulados por su respectivo ligando. Es decir que modulan la expresión de otros genes. Por un lado, la expresión del receptor ERα ha sido reportada en el endometrio, en las células cancerígenas mamarias, en el estroma ovárico, en el hipotálamo y en el epitelio de los conductos eferentes (Hess, 2003; Yaghmaie et al., 2005). La estimulación del receptor ERα por los estrógenos endógenos o exógenos promueve la proliferación celular. Por otro lado, el receptor PR también está expresado en el endometrio y las mamas. También es considerado como un marcador tanto de los cánceres de mama luminales A y B como del cáncer de endometrio tipo I: 92% de los cánceres de endometrio tipo I son positivos para PR (Harbeck et al., 2019; Shen et al., 2017).

Los datos de expresión génica para ESR1 y PGR fueron comparados con la expresión de los receptores ERα y PR medida por inmunohistoquímica (Fig. 17). Se observó una expresión génica mediana de ESR1 tres veces más alta para los pacientes cuyo tumor es positivo para la expresión de ERα a comparación de los pacientes cuyo tumor es negativo (9 contra 2.8). De la misma manera, la expresión génica mediana de PGR para los pacientes cuyo tumor es positivo para la expresión de PR supera más de cuatro veces la expresión génica mediana de PGR de los pacientes cuyo tumor es negativo (6.8 contra 1.5). Por lo tanto, se puede concluir una correspondencia satisfactoria entre los datos transcriptómicos y los datos clínicos. En cuanto a los pacientes cuyo tumor no fue analizado por inmunohistoquímica para la expresión de ERα o PGR, o sea todos los pacientes con cáncer de endometrio para ambos receptores, 51 pacientes con cáncer de mama para ERα y 54 pacientes de cáncer de mama para PGR, se observó una expresión génica de ESR1 o PGR parecida a la del grupo positivo para ERα o PGR, respectivamente. Sin embargo, a pesar de que la expresión mediana sea semejante entre los dos grupos (8.2 contra 9 para ESR1 y 6.5 contra 6.8 para PGR), la distribución de la expresión génica de los pacientes cuyo tumor no fue analizado por inmunohistoquímica para la expresión de ERα o PGR presenta más valores atípicos bajos para la expresión génica de ESR1 y un primer cuartil inferior sobre todo para la expresión génica de PGR (3.5 contra 4.9).

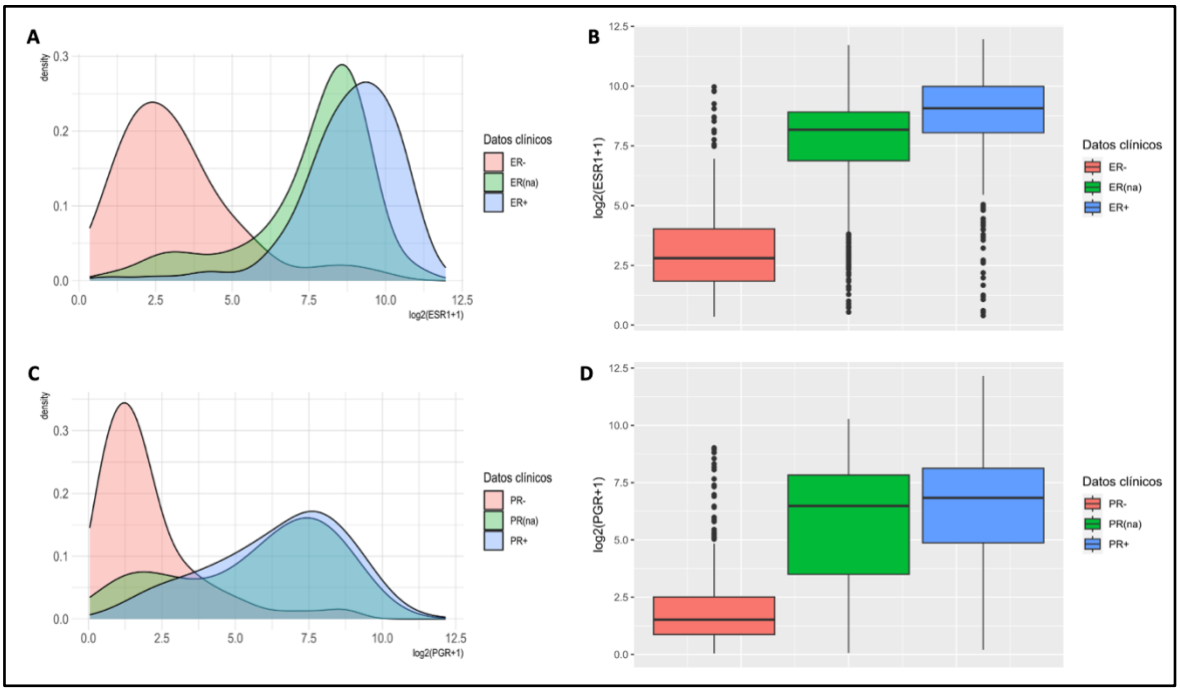


Figura 17: Nivel de expresión de los transcritos codificantes para los receptores de estrógenos (ESR1) y progesterona (PGR) en el tumor de los pacientes de cáncer de mama y de cáncer de endometrio comparado con la expresión medida por inmunohistoquímica. A-B: Nivel de expresión del transcrito ESR1. C-D: Nivel de expresión del transcrito PGR. Los tres grupos corresponden a la expresión de ERα y PR medida por inmunohistoquímica. La expresión de los receptores de estrógenos y progesterona por el tumor de los pacientes con cáncer de endometrio y por el tumor de algunos pacientes con cáncer de mama no fue medida por inmunohistoquímica. Por lo tanto, aparece un tercer grupo de pacientes – ER(na) y PR(na) – cuyo tumor no fue analizado por inmunohistoquímica.

Los datos de expresión génica para ESR1 y PGR fueron visualizados en un espacio de dos dimensiones (Fig. 18). Los pacientes fueron agrupados según su nivel de expresión de los genes ESR1 y PGR. Los 350 pacientes del grupo correspondiente a la expresión baja de ambos genes fueron eliminados. Finamente, 7 pacientes del grupo de expresión intermedia fueron eliminados por tener un valor atípico bajo para la expresión del gen ESR1.

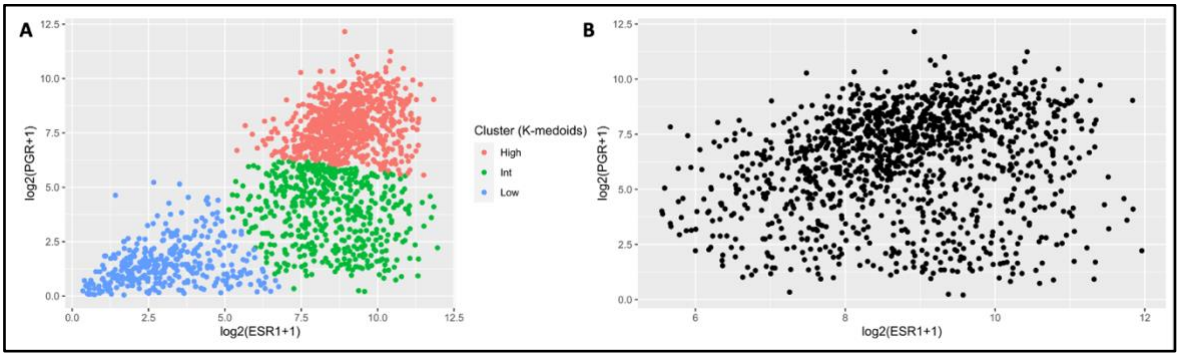


Figura 18: Agrupación *KMedoids* de los pacientes de cáncer de mama y de cáncer de endometrio según el nivel de expresión de los transcritos codificantes para los receptores de estrógenos (ESR1) y progesterona (PGR) en el tumor. A: Agrupación (*clustering*) de los pacientes (k = 3). B: Pacientes seleccionados mediante la extracción de los clusters de expresión intermedia y alta de ESR1 y PGR seguida por la eliminación de los pacientes con un valor atípico bajo para la expresión del gen ESR1.

El filtro transcriptómico aplicado a los pacientes iniciales resultó en la selección de 1273 pacientes: 826 con cáncer de mama y 447 con cáncer de endometrio.

1.3. Selección de los pacientes con una composición tumoral similar a nivel inmunológico mediante el uso de la herramienta *Cibersort*

La infiltración inmunológica en el tumor de los pacientes con cáncer de mama o con cáncer de endometrio fue estimada mediante el uso de la herramienta *Cibersort* con el objetivo de identificar y eliminar los pacientes cuya composición tumoral a nivel inmunológico difiere excesivamente de la composición tumoral de los demás pacientes. Es crucial establecer este filtro antes de analizar los datos ómicos de manera separada (sección 2 de los resultados) o integrada (sección 3 de los resultados) ya que una heterogeneidad excesiva a nivel de la composición tumoral de los pacientes constituiría un sesgo. En efecto, los perfiles de expresión génica y de metilación del ADN varían grandemente entre distintos tipos celulares como ha sido demostrado por experimentos de secuenciación de células individuales (*single-cell RNA-sequencing*) y de caracterización del metiloma del ADN de células individuales (*single-cell DNA methylation profiling*) (Heo et al., 2022; Kotliar et al., 2019).

La estimación de los tipos celulares inmunológicos reveló la presencia abundante de macrófagos sobre todo M0, pero también M2 y M1, y de linfocitos T citotóxicos (CD8) y colaboradores (CD4) en el tumor de los pacientes con cáncer de mama o con cáncer de endometrio (Fig. 19). Para cada tipo celular inmunológico estimado, se puede identificar pacientes con una abundancia atípica alta. Esta etapa busca identificar estos pacientes y eliminarlos.

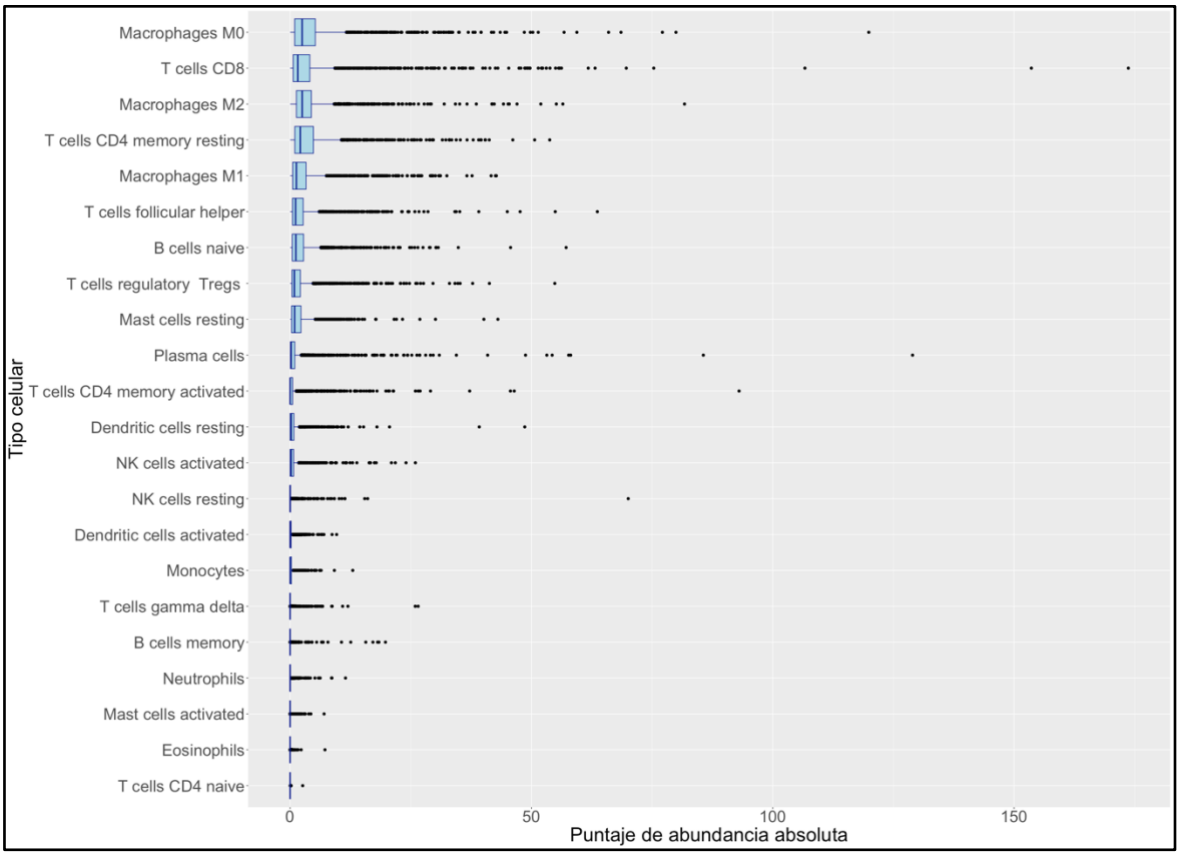


Figura 19: Abundancia absoluta de 22 tipos celulares asociados con la inmunidad en el tumor de los pacientes de cáncer de mama y de cáncer de endometrio estimada por *CIBERSORT*.

El análisis de componentes principales realizado en los datos de abundancia absoluta de los 22 tipos celulares inmunológicos resultó en dos componentes principales: PC1 y PC2, que representan el 74% de la variabilidad de los datos de abundancia absoluta. Se puede observar que la mayoría de los pacientes están agrupados en el espacio bidimensional PC1/PC2. Los demás pacientes – alejados del *cluster* – son los que presentan una composición tumoral significativamente diferente a nivel inmunológico de los pacientes agrupados. Notablemente, las variables iniciales que más contribuyen a los componentes principales PC1 y PC2 son la abundancia absoluta estimada de macrófagos y linfocitos (Fig. 20).

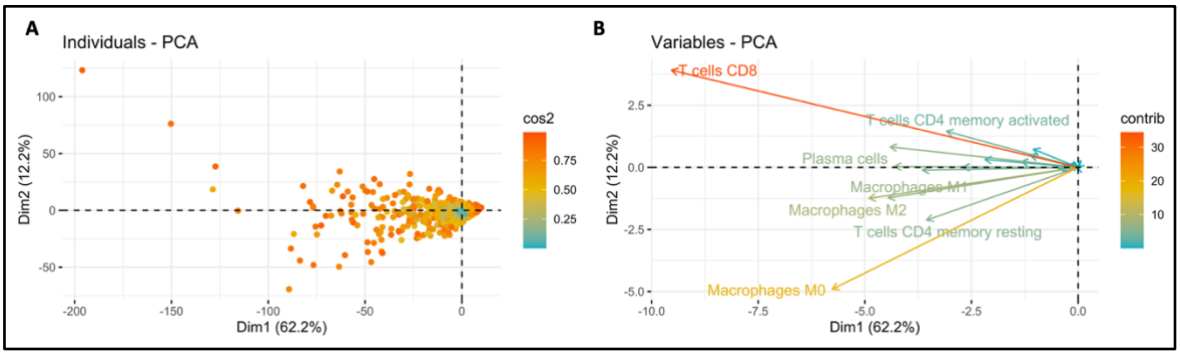


Figura 20: Análisis de componentes principales realizado en los datos de abundancia de células inmunitarias en el tumor de los pacientes de cáncer de mama y de cáncer de endometrio. A: Visualización de los pacientes en un espacio bidimensional PC1 vs. PC2. B: Contribución de las variables iniciales a PC1 y PC2.

Los pacientes fueron visualizados según su puntaje de abundancia de leucocitos infiltrados totales (Fig. 21). El filtro usado para homogeneizar los pacientes en cuanto a su composición tumoral a nivel inmunológico fue la eliminación de los pacientes con valores atípicos extremos para la abundancia de leucocitos infiltrados totales. Este filtro eliminó 101 pacientes. Los pacientes agrupados cuyo tumor es homogéneo a nivel inmunológico fueron conservados (Fig. 22). El filtro inmunológico aplicado a los pacientes iniciales resultó en la selección de 1529 pacientes: 1014 con cáncer de mama y 515 con cáncer de endometrio.

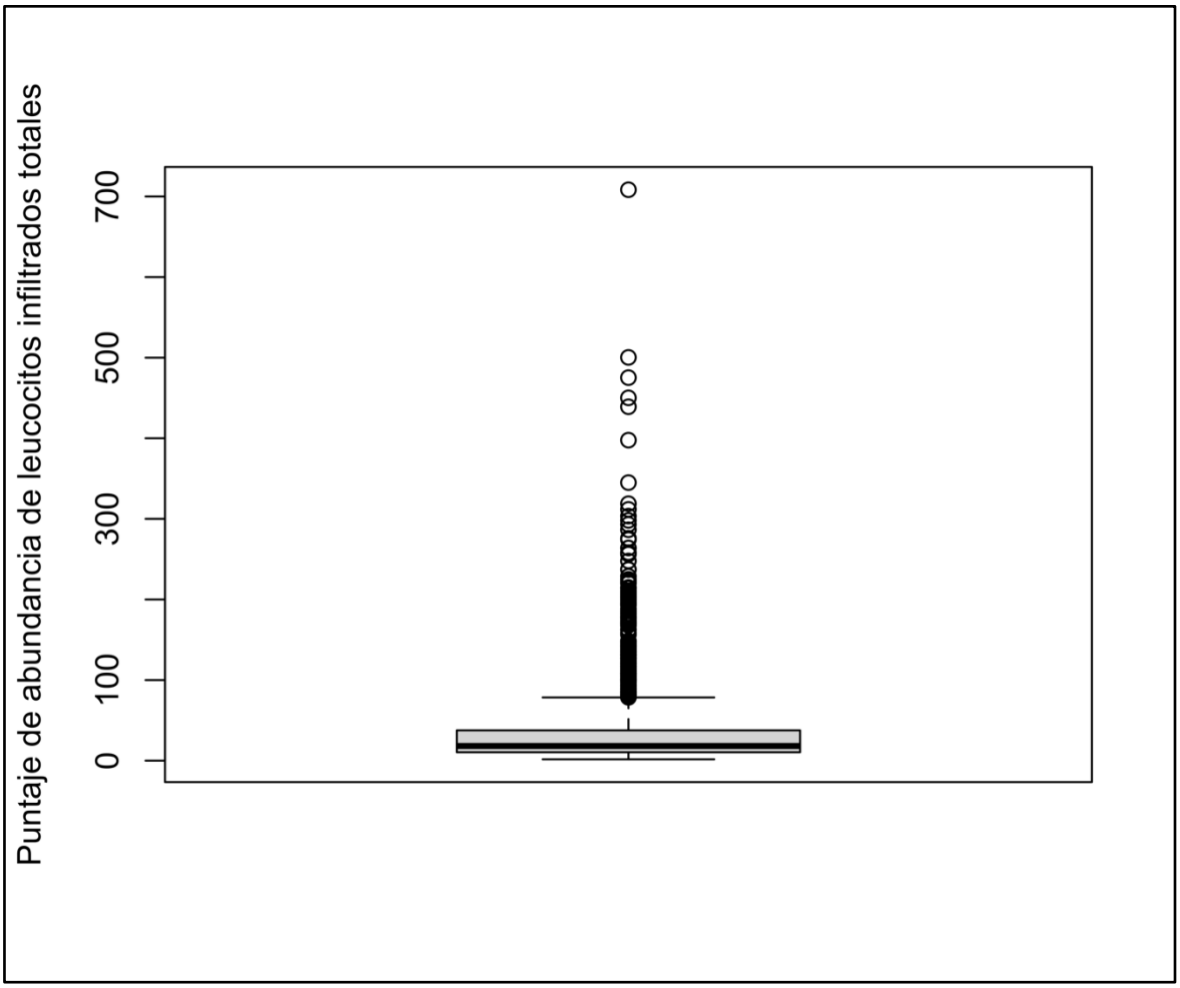


Figura 21: Abundancia absoluta de leucocitos infiltrados totales en el tumor de los pacientes de cáncer de mama y de cáncer de endometrio.

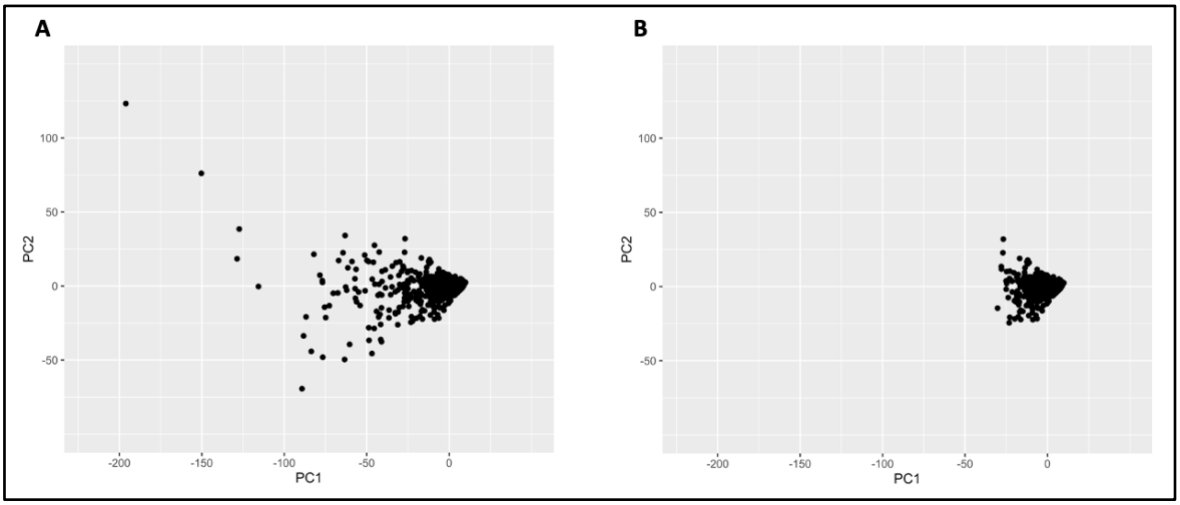


Figura 22: Eliminación de los pacientes de cáncer de mama y de cáncer de endometrio con una infiltración inmunológica total calificada de atípica extrema. A-B: Las dos imágenes corresponden a la visualización de los pacientes en el espacio bidimensional PC1 vs. PC2 antes de la selección (A) y después de la selección de los pacientes (B).

La intersección de los 1093 pacientes seleccionados por el filtro clínico, de los 1273 pacientes seleccionados por el filtro transcriptómico y de los 1529 pacientes seleccionados por el filtro inmunológico resultó en la selección de 913 pacientes: 565 pacientes con cáncer de mama y 348 con cáncer de endometrio. El proceso de aplicación de filtros a los pacientes iniciales permitió seleccionar un conjunto de pacientes relativamente homogéneo: mujeres padeciendo de cáncer ductal o lobulillar infiltrante o de adenocarcinoma endometrioide con un tumor positivo para la expresión de los receptores de estrógenos y progesterona y con una infiltración inmunológica baja. El listado de pacientes seleccionados se encuentra en el Anexo 1.

2. Análisis exploratorio de los datos

El análisis exploratorio de los datos de los 913 pacientes seleccionados corresponde a la caracterización de los pacientes según sus distintos perfiles ómicos: perfil transcriptómico, epigenómico y genómico y según su perfil clínico. La presente sección está dividida en tres partes: el análisis de los tres perfiles ómicos concluido por la extracción de biomarcadores, la búsqueda de correlación entre datos ómicos y el estudio de las variables clínicas comparando los pacientes entre los *clusters* obtenidos previamente.

2.1. Análisis de datos ómicos de un solo tipo mediante el uso de algoritmos existentes

Datos transcriptómicos

El análisis de los datos transcriptómicos de los 913 pacientes resultó en dos *clusters*: el *cluster* transcriptómico 1 está compuesto exclusivamente por pacientes con cáncer de mama mientras que el *cluster* transcriptómico 2 está compuesto exclusivamente por pacientes con cáncer de endometrio (Fig. 23). La composición de los *clusters* transcriptómicos fue reportada en el Anexo 2. Este resultado se explica por las diferencias de expresión génica entre distintos tejidos. Por un lado, existen genes calificados de constitutivos o *housekeeping* que tienen niveles de expresión constante en todas las células, tejidos y condiciones (Eisenberg & Levanon, 2013). Por otro lado, algunos genes son expresados específicamente en un tejido o presentan una expresión diferencial entre distintos tejidos (GTEx Consortium, 2017). Se habla de firma génica cuando un gen se encuentra sobre expresado específicamente en algún tipo de cáncer. Los genes compartidos entre varias firmas génicas de cáncer de mama son: MAD2L1, CCNB2, CCNA2, DTL, NUSAP1, MLF1IP, CMC2, CX3CR1, CEP55, UBE25, GTSE1, CCNE2, BIRC5, RACGAP1, BUB1B, PRC1, MELK, KPNA2, CDC20, CDC2, RRM2, CCNB1, CENPA, MYBL2 y MKI67 (Huang et al., 2018). Existen diferentes firmas génicas de cáncer

de endometrio; algunos de los genes reportados en estas firmas son BUB1B, NDC80, TPX, TTK (Huang et al., 2021), NBAT1, GFRA4, PTPRT, DLX4, RANBP3L, UNQ6494, KLRB1, PRAC1 (Gu et al., 2023), CTSW, PCSK4, LRRC8D, TNFRSF18, IHH, CDKN2A (Wang et al., 2018), CCNB1, CDC20 y NCAPG (Bian et al., 2020). Es importante resaltar que algunos genes son compartidos entre firmas génicas de cáncer de mama y de cáncer de endometrio; es el caso de BUB1B, CDC20 y CCNB1.

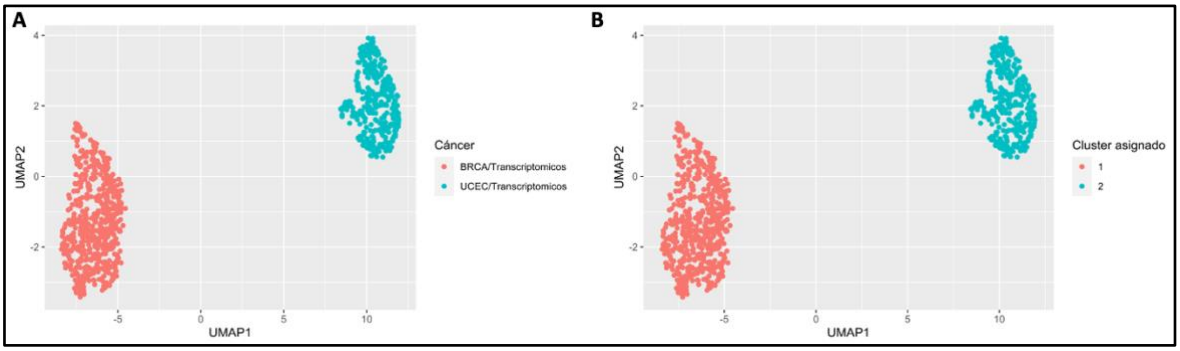


Figura 23: Visualización UMAP de los pacientes de cáncer de mama y cáncer de endometrio obtenida a partir de su perfil transcriptómico y agrupación *KMeans*. A: Visualización de los pacientes en el espacio bidimensional UMAP1 vs. UMAP2. B: *Clustering KMeans* de los pacientes (k = 2).

La comparación de la expresión génica promedio de los 60660 genes de la matriz de datos transcriptómicos entre los dos *clusters* transcriptómicos usando la prueba estadística no paramétrica de Mann-Whitney-Wilcoxon para datos pareados resultó en una diferencia significativa entre el *cluster* 1 y el *cluster* 2 (*p-value* < 2.2e-16) con una expresión génica superior en el *cluster* 1. Dado que la prueba estadística fue usada para comparar un número muy alto de variables (60660 genes) entre los *clusters* transcriptómicos, es crucial llevar a cabo un análisis visual de los resultados. A pesar de la diferencia significativa a nivel estadístico, este resultado no puede ser transpuesto a nivel biológico ya que la mayoría de los genes solo presentan una expresión residual. En efecto, la mediana de la expresión génica promedio de los 60660 genes de la matriz de datos transcriptómicos es de 0.040 en el *cluster* 1 contra 0.053 en el *cluster* 2: corresponden a niveles de expresión génica muy bajos (Fig. 24).

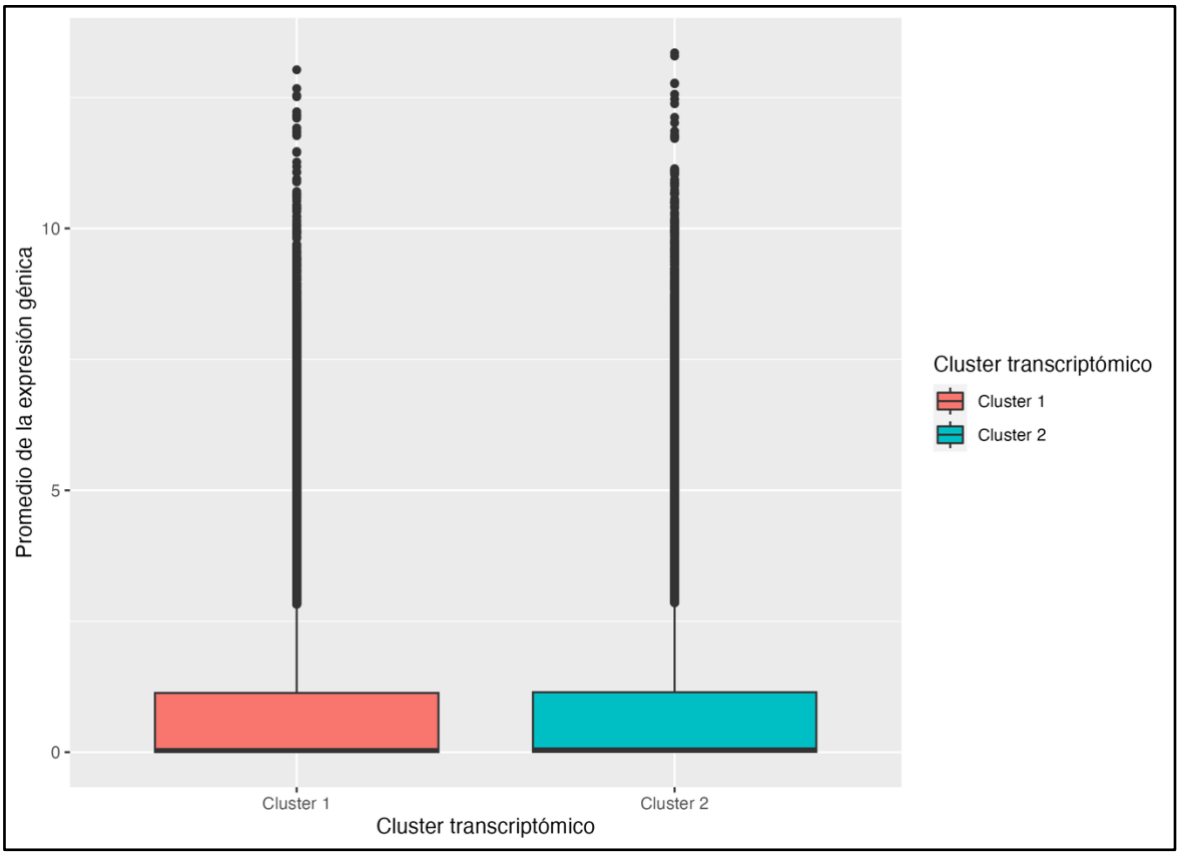


Figura 24: Comparación de la expresión génica promedio para cada gen por *cluster* transcriptómico.

100 biomarcadores fueron identificados para cada *cluster* transcriptómico por tener un puntaje de información mutua alto contra las variables sintéticas UMAP1 y UMAP2 (Fig. 25). Estos fueron reportados en el Anexo 3. Estos biomarcadores no pertenecen a las firmas génicas de cáncer de mama o de endometrio reportadas en la literatura científica. Este resultado es lógico dado que las firmas génicas mencionadas fueron establecidas comparando datos transcriptómicos entre tumores y tejidos sanos, a diferencia de este trabajo que no considera datos de pacientes sanos.

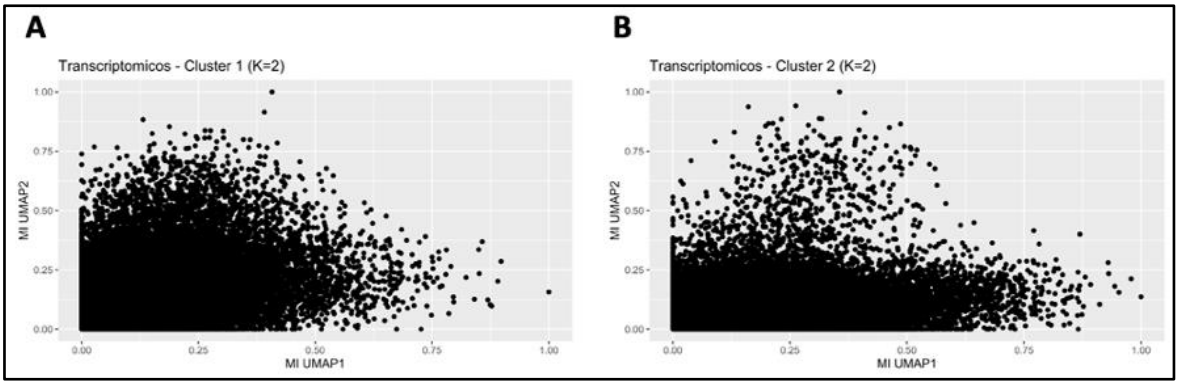


Figura 25: Visualización de las variables transcriptómicas según su puntaje de información mutua contra las variables sintéticas UMAP1 y UMAP2 en cada cluster transcriptómico. Puntaje de información mutua contra UMAP1 y UMAP2 para cada variable transcriptómica en el *cluster* transcriptómico 1 (A) y 2 (B).

Datos epigenómicos

El estudio de los datos epigenómicos de los 913 pacientes resultó en tres *clusters*: los *clusters* epigenómicos 1 y 2 están compuestos por 169 y 396 pacientes con cáncer de mama, respectivamente, mientras que el *cluster* transcriptómico 3 está compuesto exclusivamente por pacientes con cáncer de endometrio (Fig. 26). La composición de los *clusters* epigenómicos fue reportada en el Anexo 4. La lejanía entre el *cluster* epigenómico 3 y los *clusters* epigenómicos 1 y 2 se puede explicar por la existencia de patrones de metilación del ADN específicos para un tejido. Por ejemplo, en la literatura científica fue reportada una hipermetilación de las regiones promotoras de varios genes como BRCA1, p16, GSTP1, RASSF1A, Cyclin D2 en los pacientes con cáncer de mama (Phuong et al., 2015) y de las regiones promotoras de APC, CDH1, ESR1, MGMT, hMLH1, p16, PR-B, PTEN, RASSF1A y RARβ2 en los pacientes con cáncer de endometrio (Tao & Freudenheim, 2010). La hipermetilación de las regiones promotoras de un gen puede conducir a su silenciamiento. En el ámbito del cáncer, la hipermetilación de las regiones promotoras de genes supresores de tumor es un evento que favorece la carcinogénesis. La separación de los pacientes con cáncer de mama en dos *clusters* epigenómicos requiere profundización.

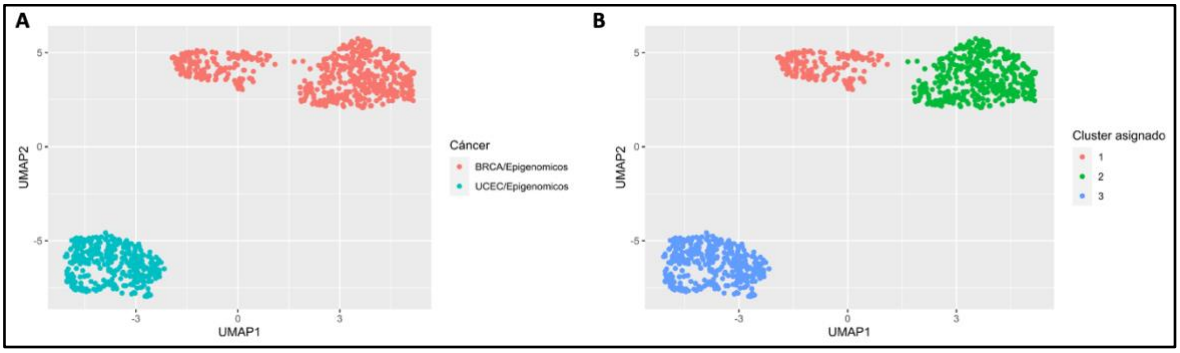


Figura 26: Visualización UMAP de los pacientes de cáncer de mama y cáncer de endometrio obtenida a partir de su perfil epigenómico y agrupación *KMeans*. A: Visualización de los pacientes en el espacio bidimensional UMAP1 vs. UMAP2. B: *Clustering* por *KMeans* de los pacientes (k = 3).

La comparación de la metilación promedio de las regiones promotoras de los 12746 genes de la matriz de datos epigenómicos entre los tres *clusters* epigenómicos resultó en diferencias significativas entre cada *cluster* (*p-value* < 2.2e-16 para todas las comparaciones) con una metilación promedio mayor en el *cluster* 2, intermedia en el *cluster*

3 y menor en el *cluster* 1. De la misma manera que para los datos transcriptómicos, ya que la prueba estadística fue usada para comparar un número alto de variables (12746 genes) entre los *clusters* epigenómicos, es crítico realizar un análisis visual de los resultados. La mediana de la metilación del ADN promedio de los 12746 genes de la matriz de datos epigenómicos es de 0.173 en el *cluster* 2 contra 0.129 en el *cluster* 1 y 0.134 en el *cluster* 3; la diferencia de metilación promedio del ADN entre los *clusters* epigenómicos 1 y 3 no está clara (Fig. 27). Se puede concluir que la metilación promedio de las regiones promotoras es mayor en el *cluster* 2 en comparación con los demás *clusters* epigenómicos.

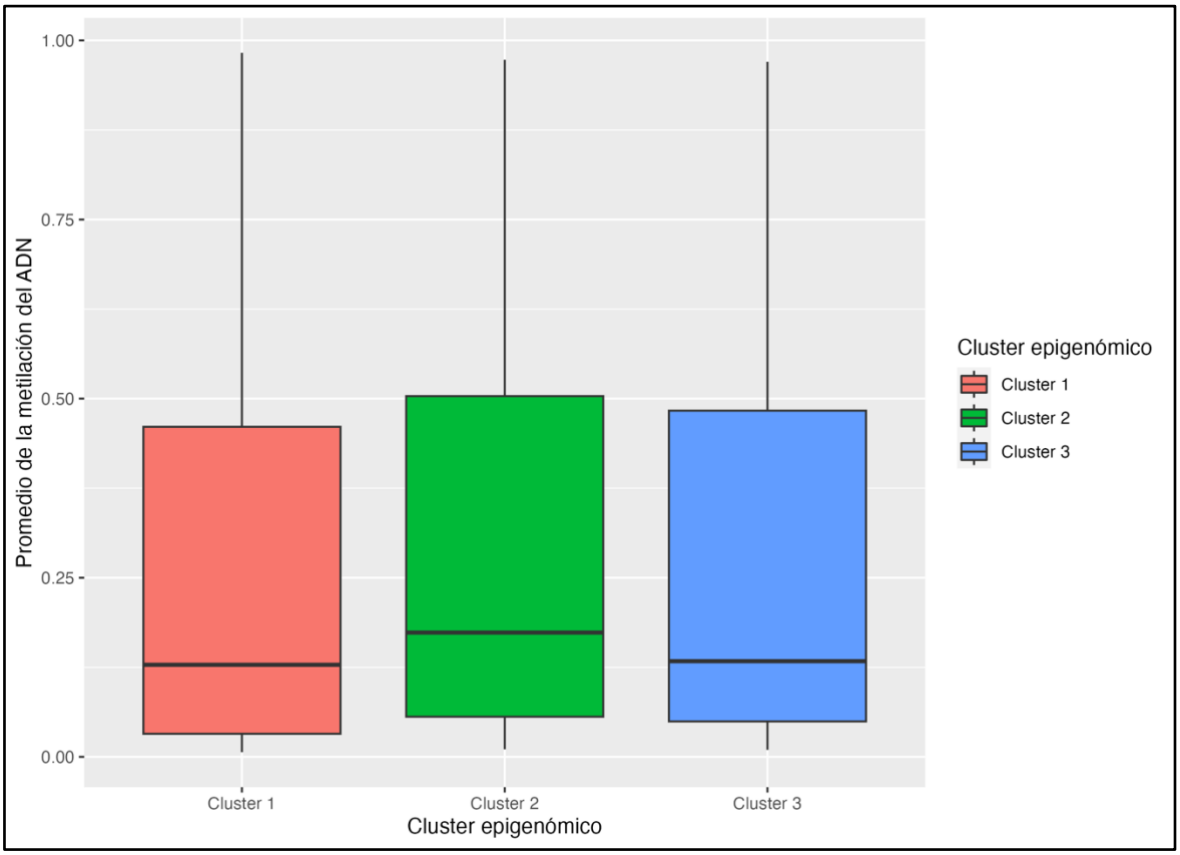


Figura 27: Comparación de la metilación promedio del ADN para cada gen por *cluster* epigenómico.

100 biomarcadores fueron identificados para cada *cluster* epigenómico por tener un puntaje de información mutua alto contra las variables sintéticas UMAP1 y UMAP2 (Fig. 28). Estos biomarcadores fueron reportados en el Anexo 5. Dichos biomarcadores no pertenecen a las firmas epigenéticas de cáncer de mama o de endometrio reportadas en la literatura científica. Así como para los datos transcriptómicos, este resultado se explica por el método usado para establecer firmas epigenéticas de ambos tipos de cáncer, o sea

la comparación de tejidos sanos con tejidos patológicos mientras que este proyecto solo contempla datos de tejidos patológicos.

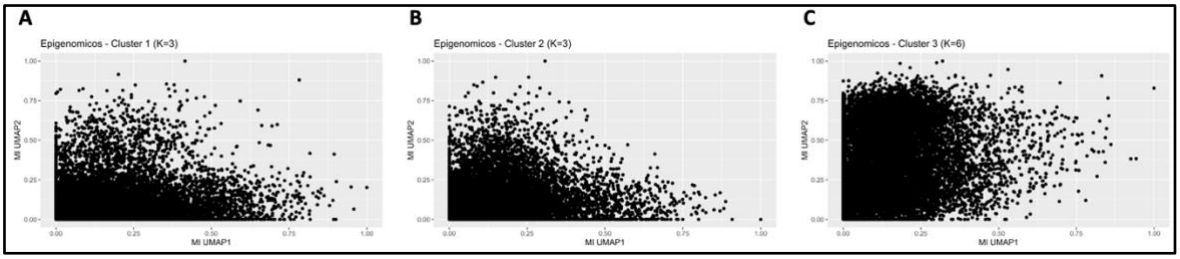


Figura 28: Visualización de las variables epigenómicas según su puntaje de información mutua contra las variables sintéticas UMAP1 y UMAP2 en cada *cluster* epigenómico. Puntaje de información mutua contra UMAP1 y UMAP2 para cada variable epigenómica en el *cluster* epigenómico 1 (A), 2 (B) y 3 (C).

Datos genómicos

El análisis de los datos genómicos de los 913 pacientes resultó en seis *clusters* mixtos. El *cluster* genómico 1 está compuesto por 83 pacientes con cáncer de mama y 13 pacientes con cáncer de endometrio; el *cluster* genómico 2 por 19 pacientes con cáncer de mama y 80 con cáncer de endometrio; el *cluster* genómico 3 por 29 pacientes con cáncer de mama y 96 con cáncer de endometrio; e *cluster* genómico 4 por 85 pacientes con cáncer de mama y 66 con cáncer de endometrio; el *cluster* genómico 5 por 104 pacientes con cáncer de mama y 53 con cáncer de endometrio; y, el *cluster* genómico 6 por 245 pacientes con cáncer de mama y 40 con cáncer de endometrio (Fig. 29). La composición de los *clusters* genómicos fue reportada en el Anexo 6. En la literatura han sido reportadas varias firmas de número de copias compartidas entre cáncer de mama y cáncer tales como CN1, CN2, CN6, CN8, CN9 y CN17 (Steele et al., 2022). Esto podría explicar la obtención de *clusters* mixtos para los datos genómicos.

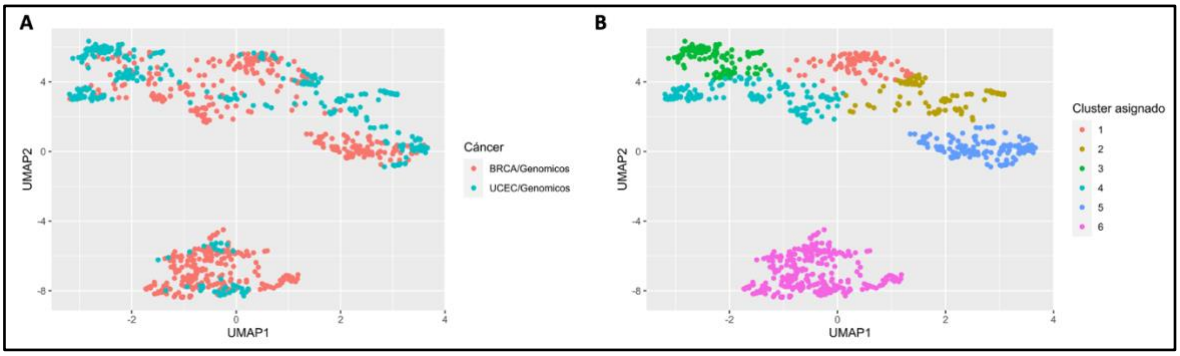


Figura 29: Visualización UMAP de los pacientes de cáncer de mama y cáncer de endometrio obtenida a partir de su perfil genómicos y agrupación por *KMeans*. A: Visualización de los pacientes en el espacio bidimensional UMAP1 vs. UMAP2. B: *Clustering por KMeans* de los pacientes (k = 6).

La comparación del número de copias promedio de los 57444 genes de la matriz de datos genómicos entre los seis *clusters* genómicos resultó en diferencias significativas entre cada *cluster* (*p-value* < 2.2e-16 para todas las comparaciones) con un número de copias promedio mayor en el *cluster* 6, seguido por el *cluster* 2, el *cluster* 5, el *cluster* 4, el *cluster* 3 y, finalmente, menor en el *cluster* 1. La prueba estadística fue usada para comparar un número alto de variables (57444 genes) entre los *clusters* genómicos, así que es clave efectuar un análisis visual de los resultados. La mediana del número de copias promedio de los 57444 genes de la matriz de datos genómicos es de 3.582 en el *cluster* 6, 2.040 en el *cluster* 2, 2.038 en el *cluster 5*, 1.987 en el *cluster* 4, 1.976 en el *cluster* 3 y de 1.927 en el *cluster* 1. La diferencia en el número de copias entre los *clusters* genómicos 1, 2, 3, 4 y 5 no está clara (Fig. 30). La conclusión es que el número de copias promedio de los genes es mayor en el *cluster* 6 en comparación con los demás *clusters* genómicos.

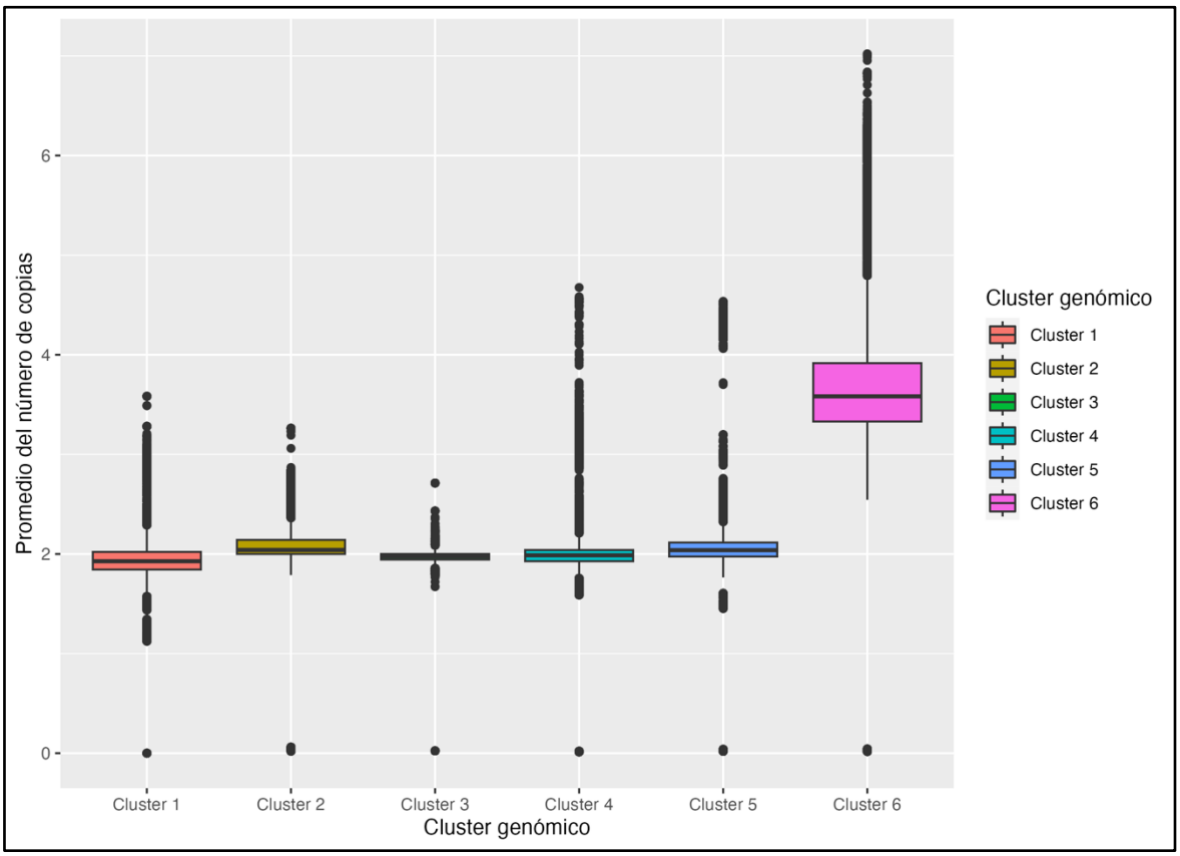


Figura 30: Comparación del número de copias promedio para cada gen por *cluster* genómico.

100 biomarcadores fueron identificados para cada *cluster* genómico por tener un puntaje de información mutua alto contra las variables sintéticas UMAP1 y UMAP2 (Fig. 31). Estos fueron reportados en el Anexo 7.

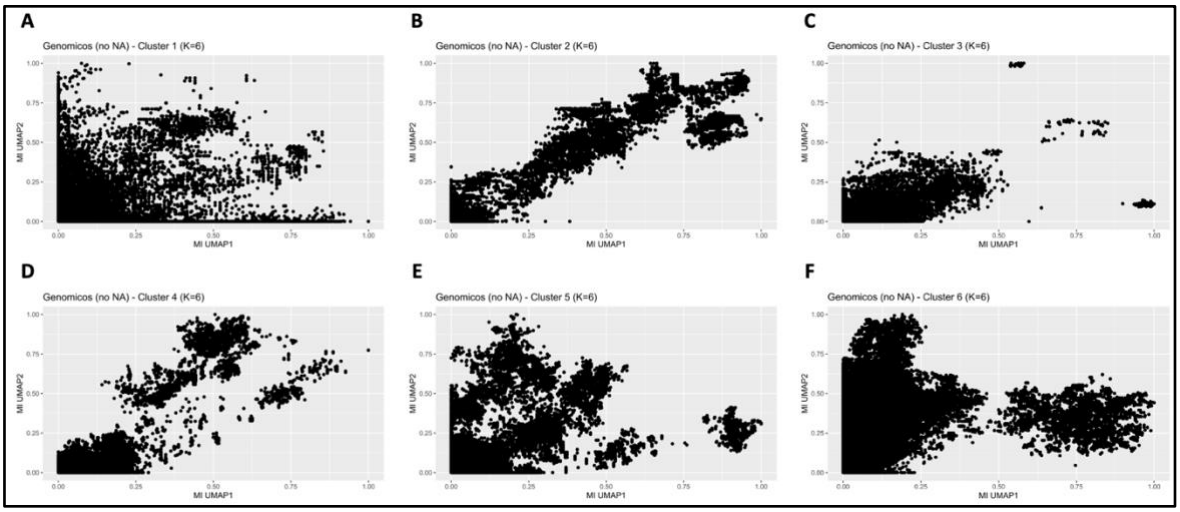


Figura 31: Visualización de las variables genómicas según su puntaje de información mutua contra las variables sintéticas UMAP1 y UMAP2 en cada *cluster* genómico. Puntaje de información mutua contra UMAP1 y UMAP2 para cada variable genómica en el *cluster* genómico 1 (A), 2 (B), 3 (C), 4 (D), 5 (E) y 6 (F).

2.2. Análisis correlacional entre los distintos tipos de datos ómicos

Las potenciales relaciones entre distintos tipos de datos ómicos fueron estudiadas para mejorar el conocimiento de los datos ómicos y, más específicamente, definir si el número de copias, por un lado, y si la metilación de las regiones promotoras, por otro lado, condicionan la expresión génica. La correlación de Spearman fue medida para cada gen entre el número de copias y la expresión génica mediana para cada número de copias. El coeficiente de correlación de Spearman entre datos genómicos y transcriptómicos difiere mucho según el gen considerado con un mínimo de -0.96 para el gen TNNI2 y un máximo de 1 para 578 genes. Sin embargo, la mayoría de los 12594 genes estudiados presentan una correlación positiva entre su número de copias y su expresión génica. En efecto, el 50% de los genes estudiados (4198 genes) muestra un coeficiente de correlación superior o igual a 0.71 y el 25% de los genes estudiados (3149 genes) muestra un coeficiente de correlación superior o igual a 0.90. Un comportamiento similar fue observado enfocándose en los biomarcadores genómicos y transcriptómicos (ver Fig. 32). Para la mayoría de los genes, entre más copias de un gen, mayor expresión génica. Semejante observación ya había sido reportada en la literatura en un estudio integrando todos los tipos de cáncer disponibles en la plataforma TCGA (Shao et al., 2019).

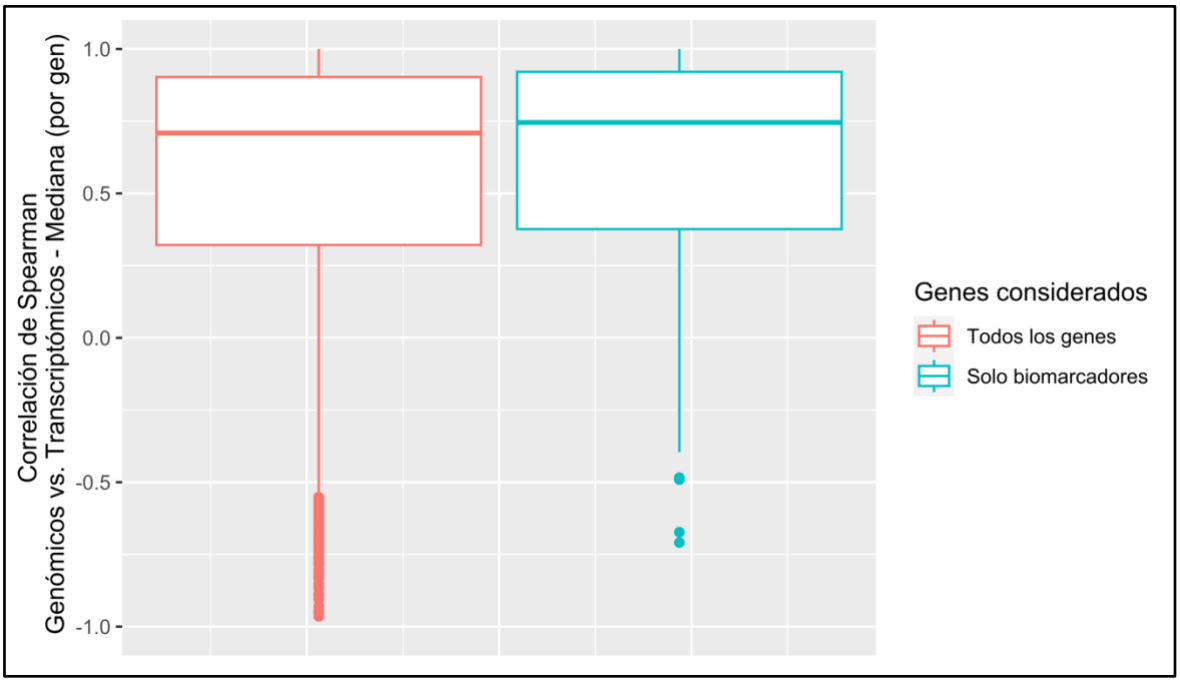


Figura 32: Visualización de los coeficientes de correlación de Spearman entre el número de copias y la expresión génica mediana para cada número de copias por gen.

Además, la correlación de Spearman fue medida para cada gen entre la metilación de las regiones promotoras del gen y su expresión. El coeficiente de correlación de Spearman entre datos epigenómicos y transcriptómicos no varía tanto como el coeficiente de Spearman entre datos genómicos y transcriptómicos. La mayoría de los 12594 genes estudiados no presentan ninguna correlación entre la metilación de sus regiones promotoras y su expresión génica. En efecto, la mediana del coeficiente de correlación de Spearman de los 12594 genes estudiados es igual a -0.05. Un comportamiento similar fue observado seleccionando únicamente los biomarcadores genómicos y transcriptómicos (ver Fig. 33). Se puede concluir que, para los genes estudiados y en el ámbito de los cánceres de mama y endometrio, la metilación del ADN no regula tanto negativamente como positivamente la expresión génica. Anteriormente, era comúnmente aceptado que la metilación del ADN regula negativamente la expresión génica al afectar, durante el proceso de transcripción, las interacciones entre el ADN y las proteínas de la cromatina y factores de transcripción (Razin & Cedar, 1991). Sin embargo, una posible regulación positiva entre la metilación del ADN y la expresión génica fue avanzada últimamente en un trabajo sobre los datos de pacientes con cáncer provenientes de TCGA. Los autores observaron que dentro de las regiones promotoras había una cantidad sustancial de correlación positiva entre la metilación y la expresión génica. Es importante mencionar que esta correlación

positiva no fue generalizada al genoma entero (Spainhour et al., 2019). Nuevos análisis son necesarios para mejor el entendimiento del impacto de la hipometilación o de la hipermetilación del ADN en la expresión génica en el contexto del cáncer.

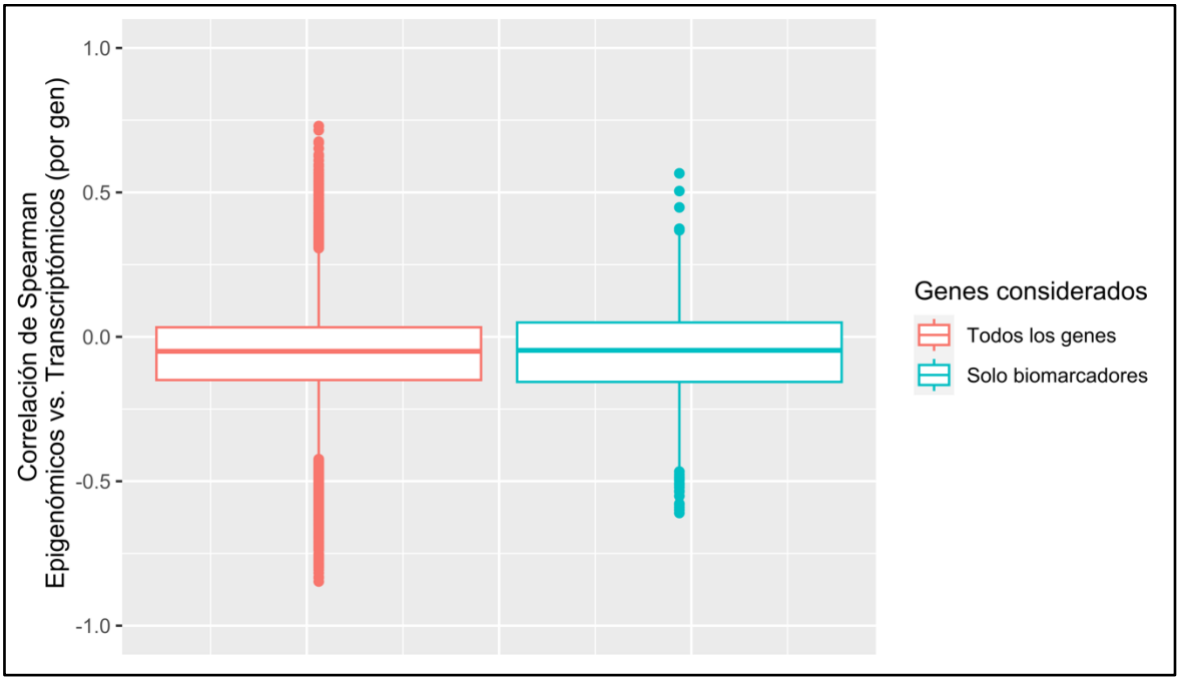


Figura 33: Visualización de los coeficientes de correlación de Spearman entre la metilación del ADN y la expresión génica por gen.

2.3. Análisis exploratorio de los datos clínicos

Los *clusters* transcriptómicos, epigenómicos y genómicos obtenidos previamente fueron caracterizados a nivel clínico. A continuación, son presentadas las variables que mostraron diferencias significativas entre *clusters*. El objetivo de esta etapa es entender el perfil clínico de los pacientes según su *cluster* ómico asignado.

Datos transcriptómicos

La comparación de las variables clínicas de los pacientes según su *cluster* transcriptómico asignado equivale a comparar los pacientes con cáncer de mama (*cluster* transcriptómico 1) y los pacientes con cáncer de endometrio (*cluster* transcriptómico 2). El análisis de la estadificación del cáncer entre los *clusters* transcriptómicos reveló una diferencia significativa (*p-value* = 7.8e-25) entre pacientes con cáncer de mama y pacientes con cáncer de endometrio. En efecto, la mayoría de los pacientes con cáncer de endometrio están clasificados en etapa 1 (71.6%) mientras que la mayoría de los pacientes

con cáncer de mama están clasificados en etapa II (52.6 %) o III (23.7%). Hay que resaltar que la estadificación del cáncer no fue medida para algunos pacientes (Fig. 34). Se puede concluir de esta figura que los pacientes con cáncer de mama tienen tumores más avanzados que los pacientes con cáncer de endometrio.

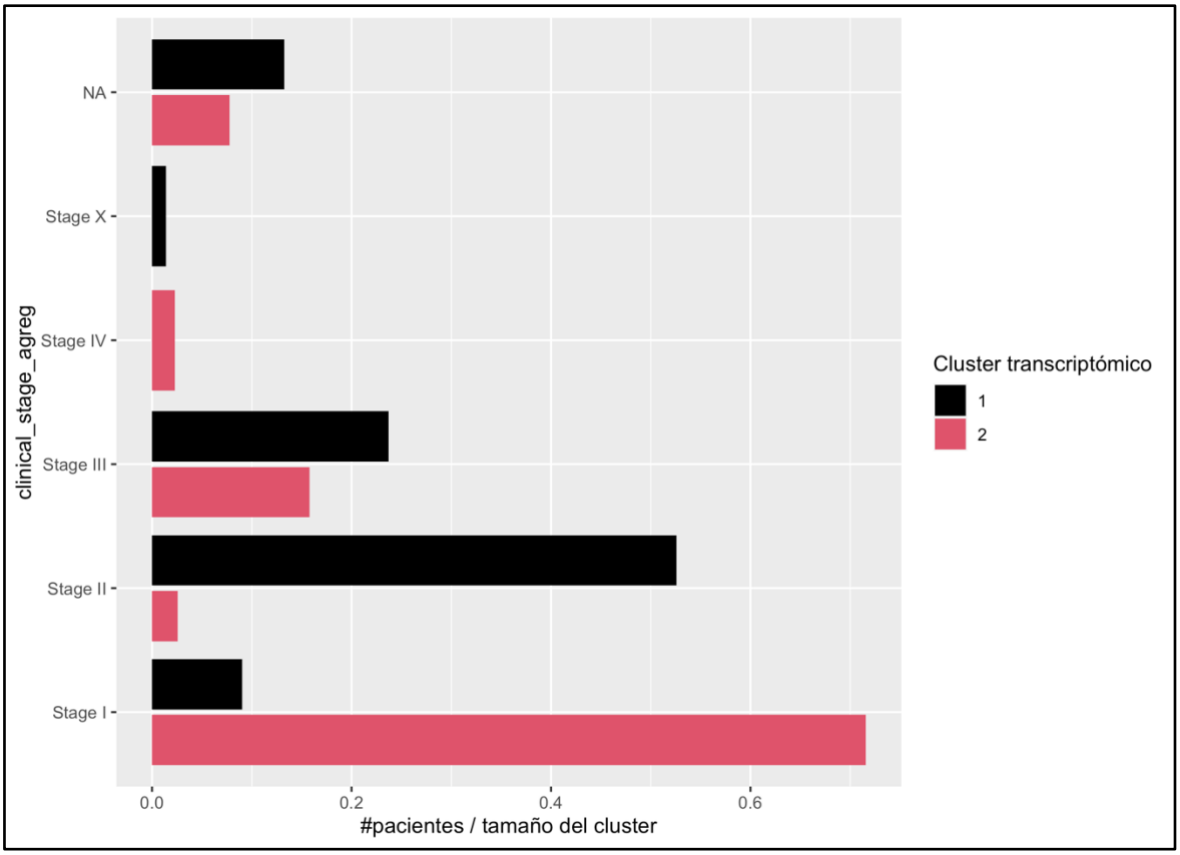


Figura 34: Comparación de la estadificación del cáncer entre los *clusters* transcriptómicos.

La comparación del estado vital de los pacientes entre los *clusters* transcriptómicos mostró una diferencia significativa (*p-value* = 0.004) entre pacientes con cáncer de mama y pacientes con cáncer de endometrio. Los pacientes con cáncer de mama presentan una tasa de mortalidad mayor: 7.6% de los pacientes fallecieron contra 4.9% para el cáncer de endometrio (Fig. 35). Este resultado está en adecuación con lo observado para la estadificación del cáncer: en esta cohorte de pacientes, los que tienen cáncer de mama presentan tumores más avanzados y, por consiguiente, una mayor tasa de mortalidad.

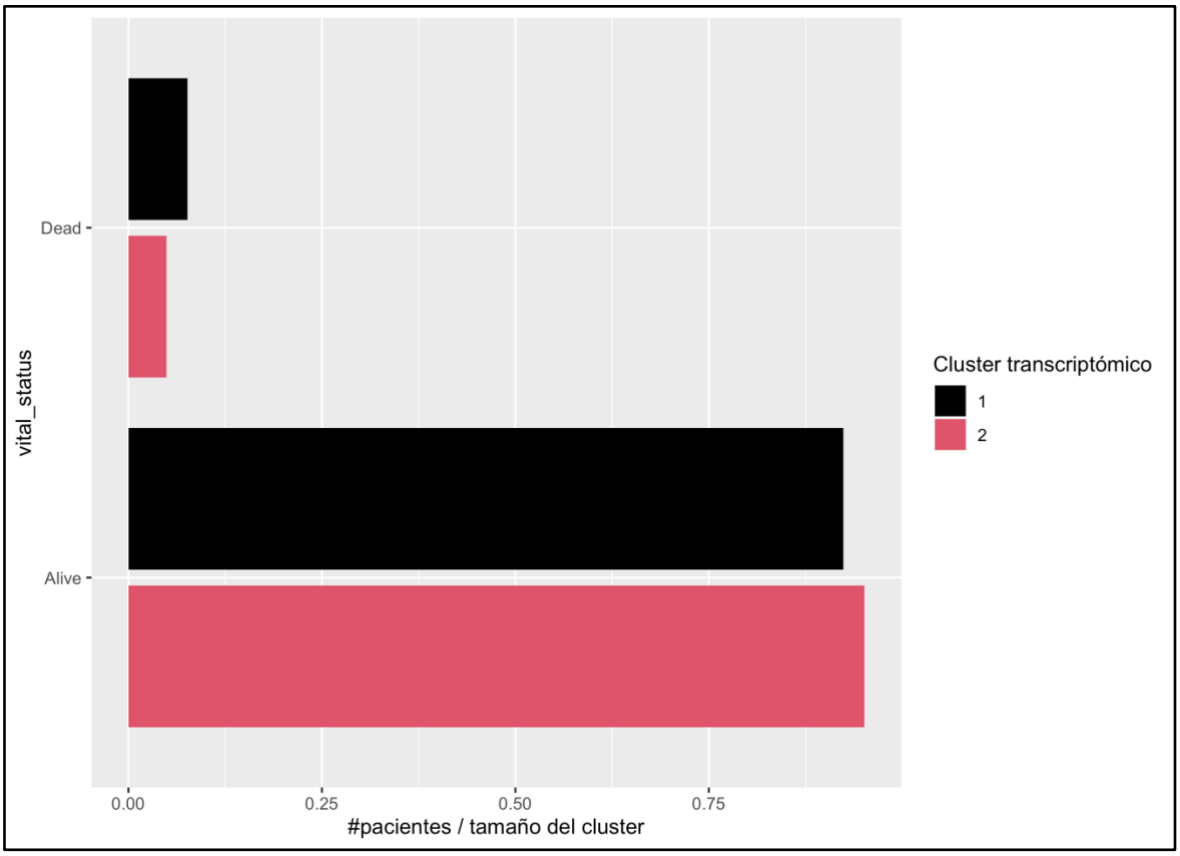


Figura 35: Comparación del estado vital de los pacientes según su *cluster* transcriptómico asignado.

El estudio comparativo de la edad en el diagnóstico de los pacientes entre los *clusters* transcriptómicos evidenció una diferencia significativa (*p-value* = 3.8e-5) entre pacientes con cáncer de mama y pacientes con cáncer de endometrio. Los pacientes con cáncer de endometrio presentan una edad mediana de 62 años, mayor que la edad mediana de los pacientes con cáncer de mama: 59 años (Fig. 36).

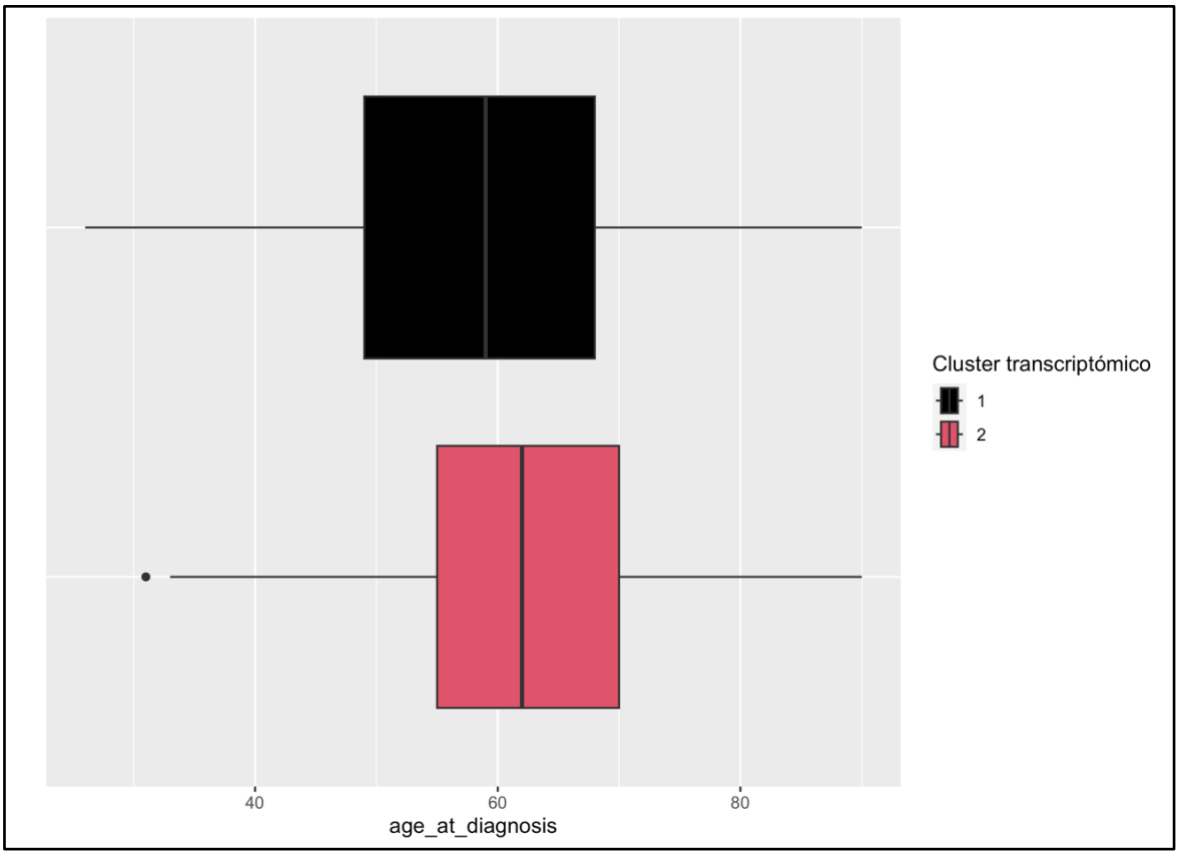


Figura 36: Comparación de la edad en el diagnóstico de los pacientes según su *cluster* transcriptómico asignado.

Datos epigenómicos

La comparación de las variables clínicas de los pacientes según su *cluster* epigenómico asignado se enfocó en los *clusters* 1 y 2 compuestos por 169 y 396 pacientes con cáncer de mama, respectivamente. El *cluster* 3 está compuesto exclusivamente por pacientes con cáncer de endometrio (Fig. 37). Cabe recordar que el *cluster* 2 se caracteriza por una metilación promedio de las regiones promotoras mayor.

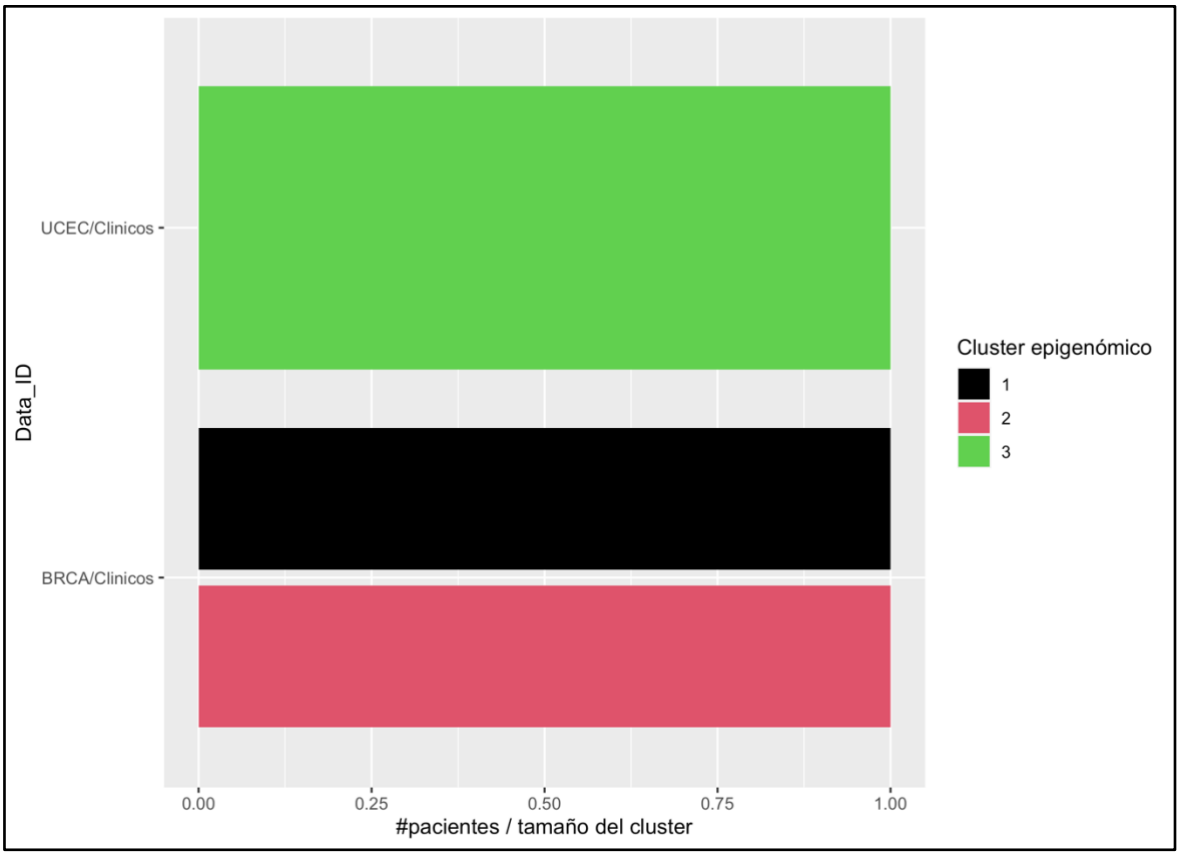


Figura 37: Comparación del tipo de cáncer de los pacientes según su *cluster* epigenómico asignado.

El análisis de la histología del cáncer de los pacientes entre los *clusters* epigenómicos expuso una diferencia significativa (*p-value* = 7.7e-208) entre *clusters*. El subtipo histológico de carcinoma ductal infiltrante predomina en el *cluster* epigenómico 1 (88.8%). En paralelo, el *cluster* epigenómico 2 presenta un balance más equilibrado entre el subtipo histológico de carcinoma ductal infiltrante: 65.2% de los pacientes, y el subtipo histológico de carcinoma lobulillar infiltrante: 34.8% de los pacientes (Fig. 38).

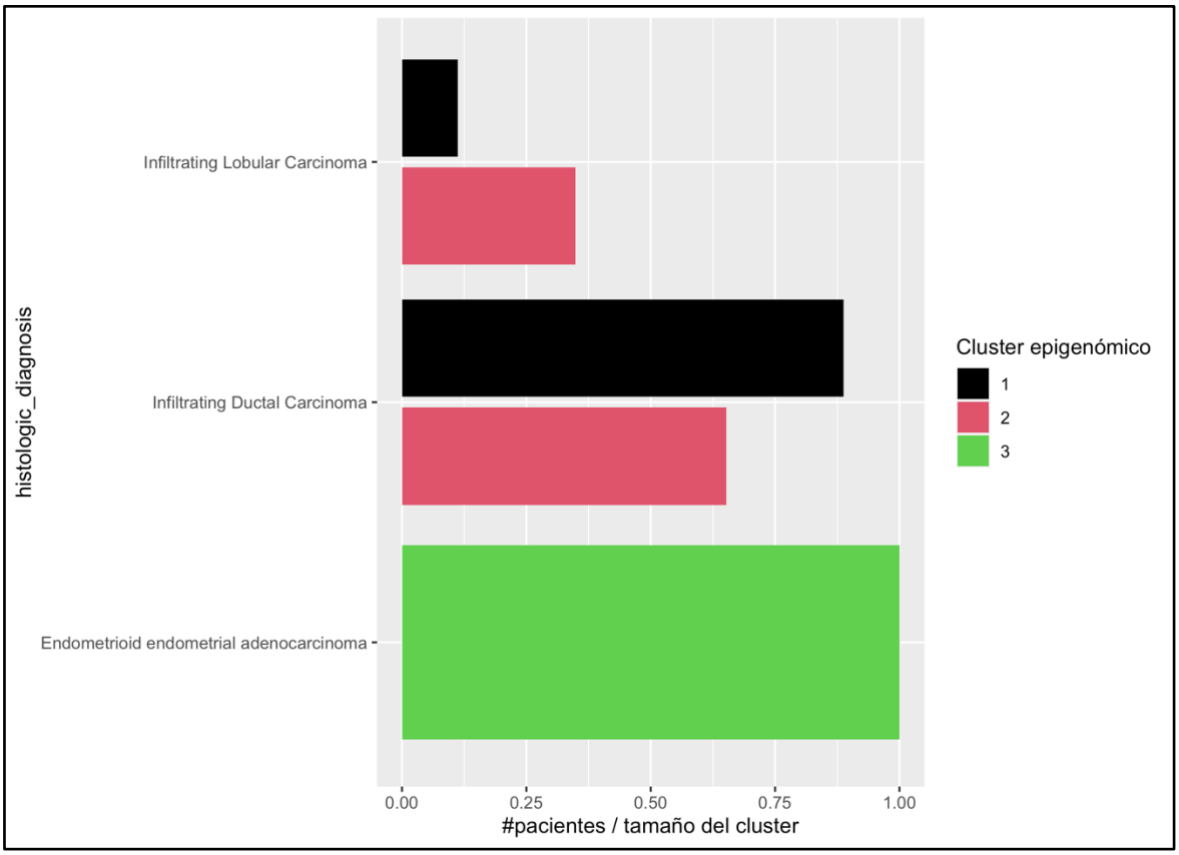


Figura 38: Comparación de la histología del tumor de los pacientes según su *cluster* epigenómico asignado.

La comparación del estado menopaúsico de los pacientes según su *cluster* epigenómico asignado resultó en una diferencia significativa (*p-value* = 7.7e-208). Una leve diferencia fue encontrada entre los *clusters* epigenómicos 1 y 2: el *cluster* epigenómico 1 presenta un menor porcentaje de pacientes en estado premenopáusico que el *cluster* epigenómico 2: 16.6% contra 23.5%, y un mayor porcentaje de pacientes en estado posmenopáusico que el *cluster* epigenómico 2: 69.2% contra 62.6% (Fig. 39).

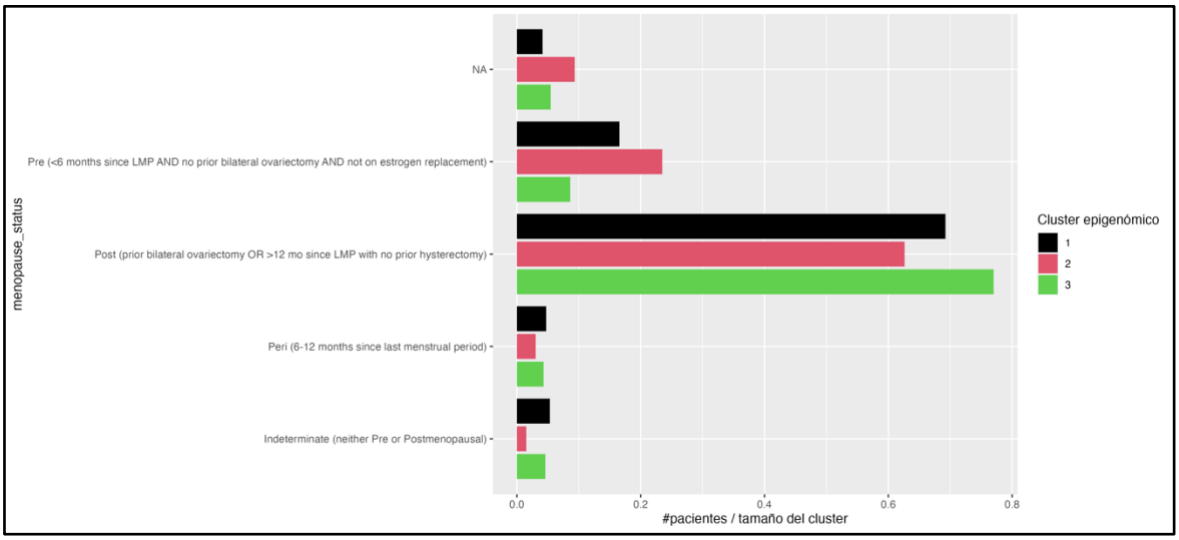


Figura 39: Comparación del estado menopaúsico de los pacientes según su *cluster* epigenómico asignado.

El análisis comparativo de la edad en el diagnóstico de los pacientes entre los *clusters* epigenómicos dio lugar a una diferencia significativa (*p-value* = 1.4e-5) entre *clusters*. Los pacientes del *cluster* 1 presentan una edad mediana de 61 años, significativamente mayor que la edad mediana de los pacientes del *cluster* 2: 59 años (*p-value post hoc* = 0.045; Fig. 40). Esta observación está en adecuación con los resultados del análisis del estado menopaúsico de los pacientes de los *clusters* epigenómicos 1 y 2.

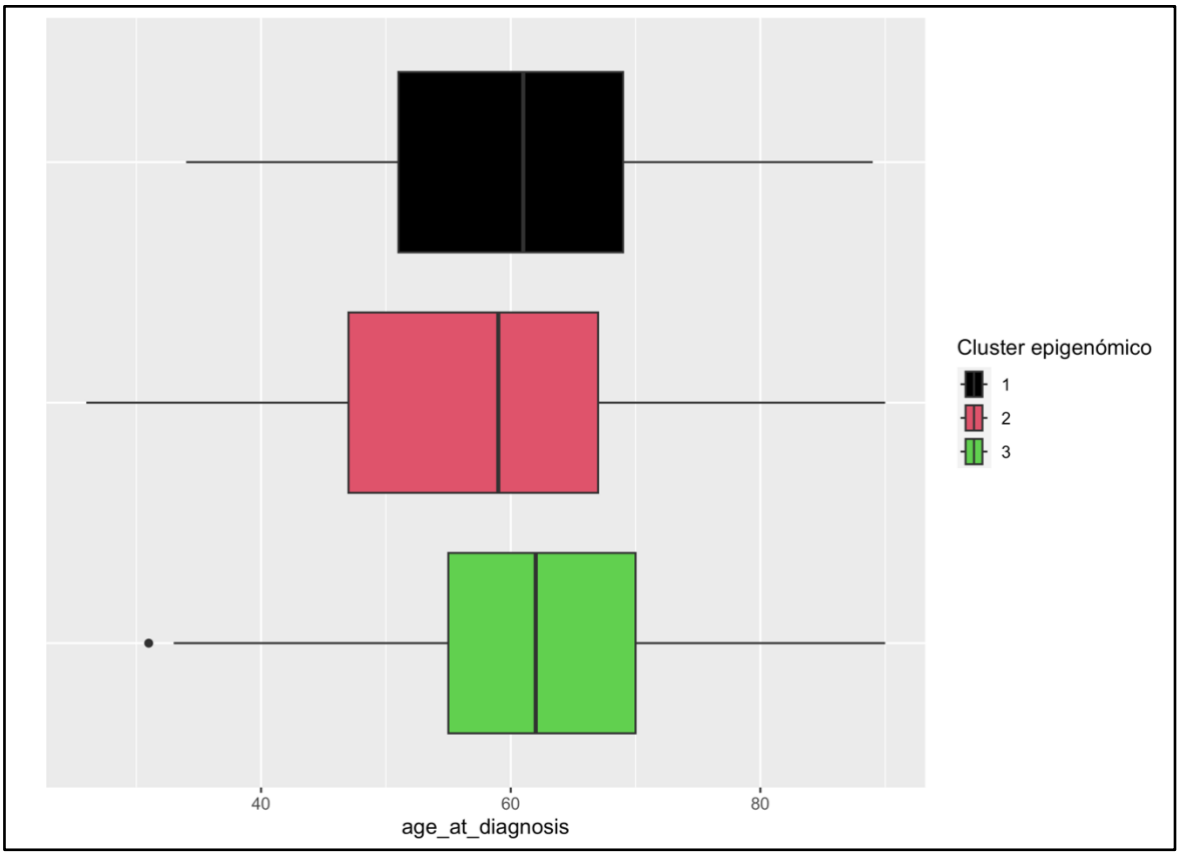


Figura 40: Comparación de la edad en el diagnóstico de los pacientes según su *cluster* epigenómico asignado.

El análisis de la estadificación del cáncer entre los *clusters* epigenómicos reveló una diferencia significativa (*p-value* = 5.8e-92) entre *clusters*. Se puede observar un porcentaje mayor de pacientes clasificados en etapa III en el *cluster* 2 (27.0%) que en el *cluster* 1 (16.0%). Es necesario subrayar que la estadificación del cáncer no fue medida para algunos pacientes de cada *cluster* (ver Fig. 41). Se puede concluir de esta figura que el *cluster* epigenómico 2 tiene más pacientes con un tumor muy avanzado que el *cluster* 1.

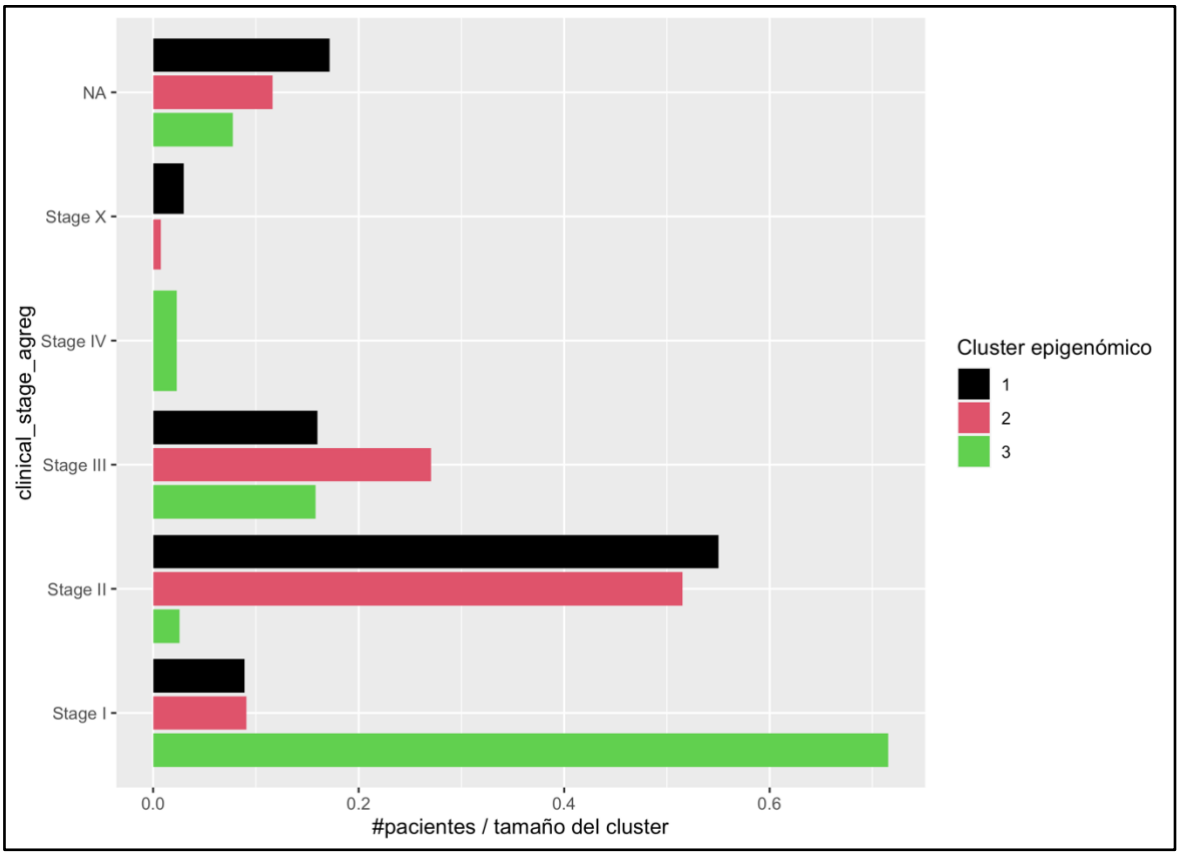


Figura 41: Comparación de la estadificación del cáncer de los pacientes según su *cluster* epigenómico asignado.

La comparación del conteo de ganglios linfáticos positivos para la presencia de micrometástasis o residuos tumorales – distinguidos por tinción con hematoxilina eosina – entre los *clusters* epigenómicos reveló una diferencia significativa (*p-value* = 1.7e-4) entre *clusters*. Se puede observar un número mediano de ganglios linfáticos positivos para la presencia de células cancerígenas en etapa III mayor en el *cluster* 2: 1, que en el *cluster* 1: 0. Los ganglios linfáticos axilares de los pacientes con cáncer de endometrio no fueron estudiados ya que la distancia entre el aparato genital y las axilas es muy grande y que los ganglios linfáticos que suelen ser positivos para la presencia de residuos tumorales en este contexto son los ganglios pélvicos (Fig. 42). Los pacientes con cáncer de mama del *cluster* epigenómico 2 que presentan un ganglio linfático con células cancerígenas o más podrían ser los pacientes clasificados en etapa III, más numerosos que en el *cluster* epigenómico 1.

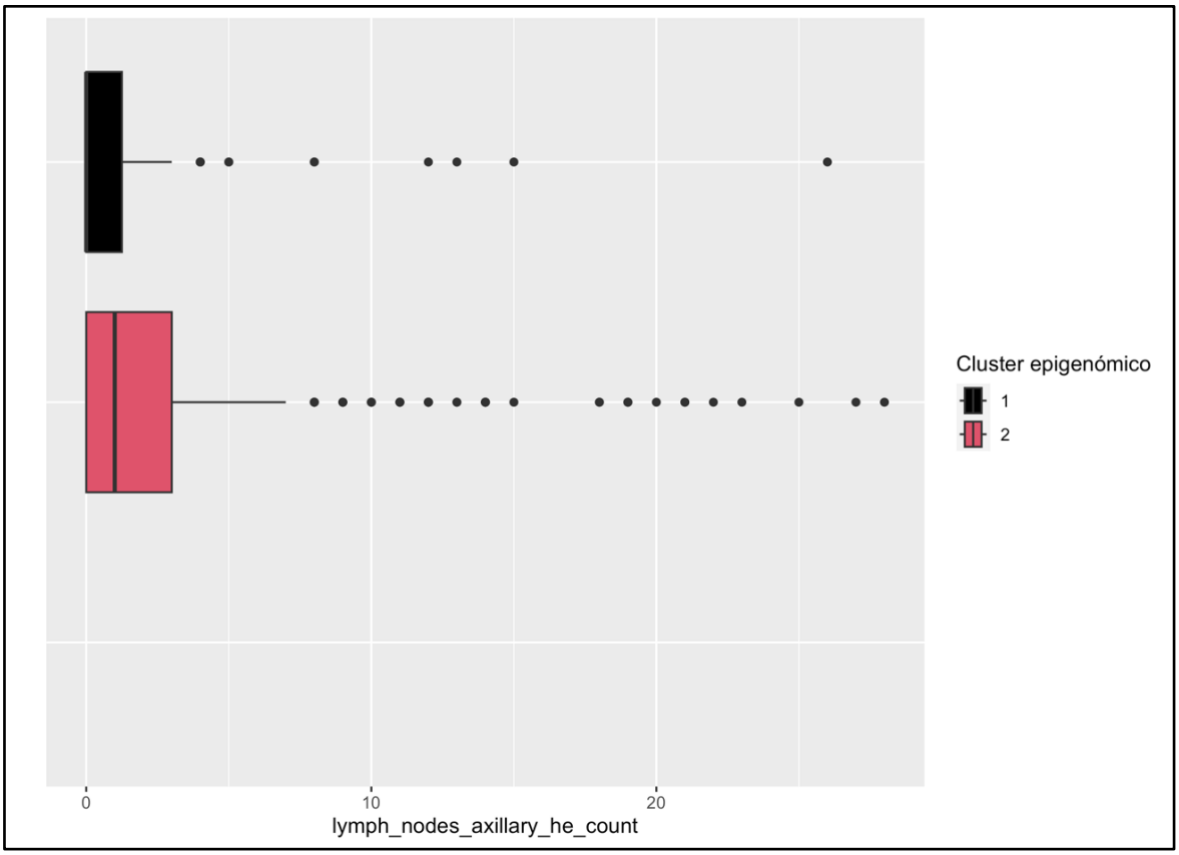


Figura 42: Comparación del conteo de ganglios linfáticos axilares positivos para la presencia de células cancerígenas de los pacientes según su *cluster* epigenómico asignado.

El análisis comparativo de los pacientes con cáncer de mama que pertenecen a los *clusters* epigenómicos 1 y 2 permitió emitir las siguientes conclusiones:

- El *cluster* epigenómico 1 se caracteriza, primero, por un fuerte desequilibrio en la proporción entre pacientes con carcinoma ductal y lobulillar infiltrante hacia el carcinoma ductal infiltrante, luego, por una edad mediana mayor y, por lo tanto, más pacientes en estado posmenopáusico; finalmente, por una menor cantidad de pacientes en etapas cancerígenas avanzadas y, en consecuencia, menos pacientes cuyos ganglios linfáticos son positivos para la presencia de células cancerígenas.
- Los pacientes del *cluster* epigenómico 2 presentan, una mayor metilación promedio de las regiones promotoras de los genes, una edad mediana menor, asociada con más pacientes en estado premenopáusico, y, más ganglios linfáticos positivos para micrometástasis o residuos tumorales, lo que podría explicar la cantidad mayor de pacientes clasificados en etapa III, una etapa tardía del cáncer.

Datos genómicos

La comparación de las variables clínicas de los pacientes según su *cluster* genómico asignado se hizo separando los pacientes con cáncer de mama de los pacientes con cáncer de endometrio en cada *cluster* para eliminar los sesgos relativos a la estadificación, al estado vital y a la edad en el diagnóstico que fueron descritos en el análisis comparativo de los dos *clusters* transcriptómicos o sea de los pacientes con cáncer de mama y de los pacientes con cáncer de endometrio. A pesar de que los *clusters* genómicos sean mixtos, la proporción entre pacientes con cáncer de mama y pacientes con cáncer de endometrio difiere considerablemente según el *cluster* considerado (Fig. 43). Ninguna diferencia significativa fue encontrada comparando los pacientes con cáncer de endometrio entre los seis *clusters* genómicos. Para los pacientes con cáncer de mama, diferencias significativas fueron exhibidas en el estado vital y la histología tumoral.

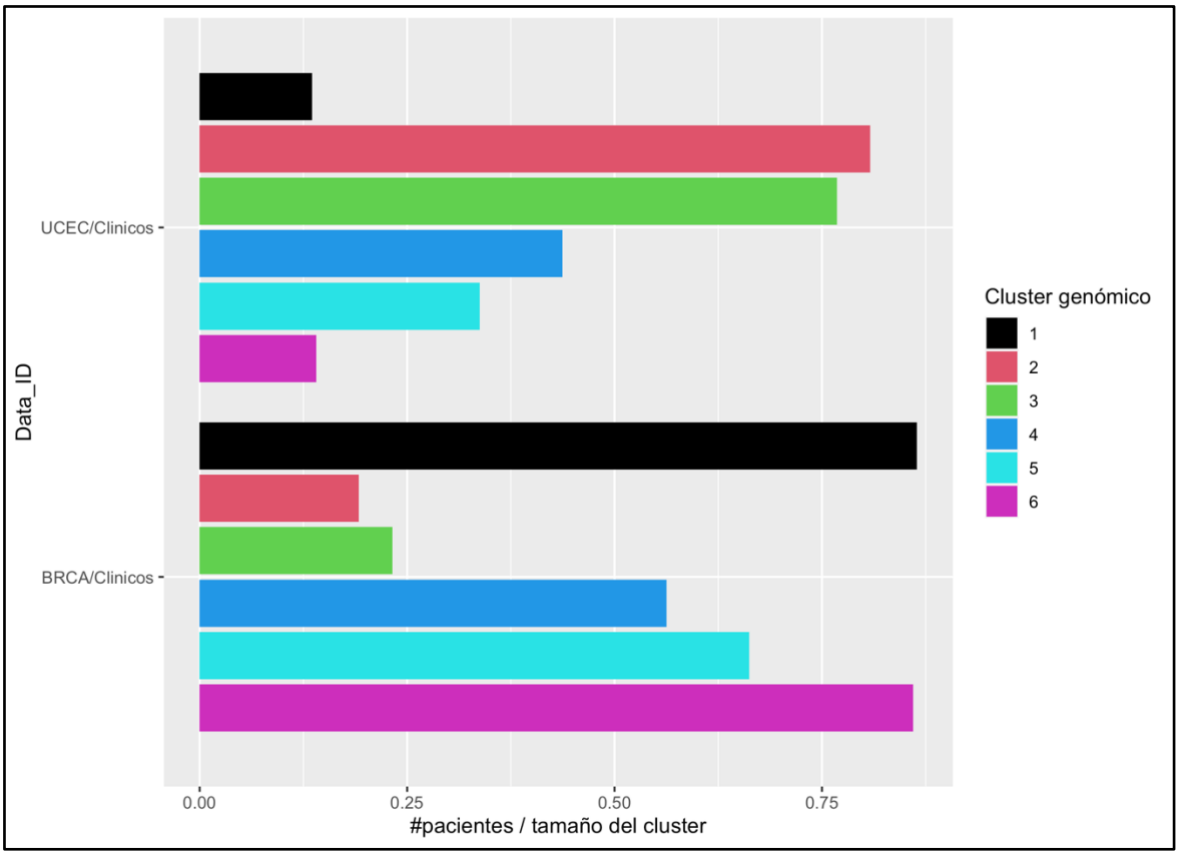


Figura 43: Comparación del tipo de cáncer de los pacientes según su *cluster* genómico asignado.

Datos genómicos – Pacientes con cáncer de mama

El análisis del estado vital de los pacientes entre los *clusters* genómicos reveló una diferencia significativa (*p-value* = 0.020) entre *clusters*. En efecto, los pacientes de cáncer

de mama de los *clusters* genómicos 2 y 4 presentan una tasa de mortalidad muy alta en comparación con los demás *clusters*: 15.8% para el *cluster* 2 y 15.3% para el *cluster* 4, contra 0% para el *cluster* 3, 4.8% para los *cluster*s 1 y 5 y 7.3% para el *cluster* 6 (ver Fig. 44). Vale la pena mencionar que un mayor número de copias de genes no parece estar relacionado con una mayor tasa de mortalidad observando el *cluster* 6.

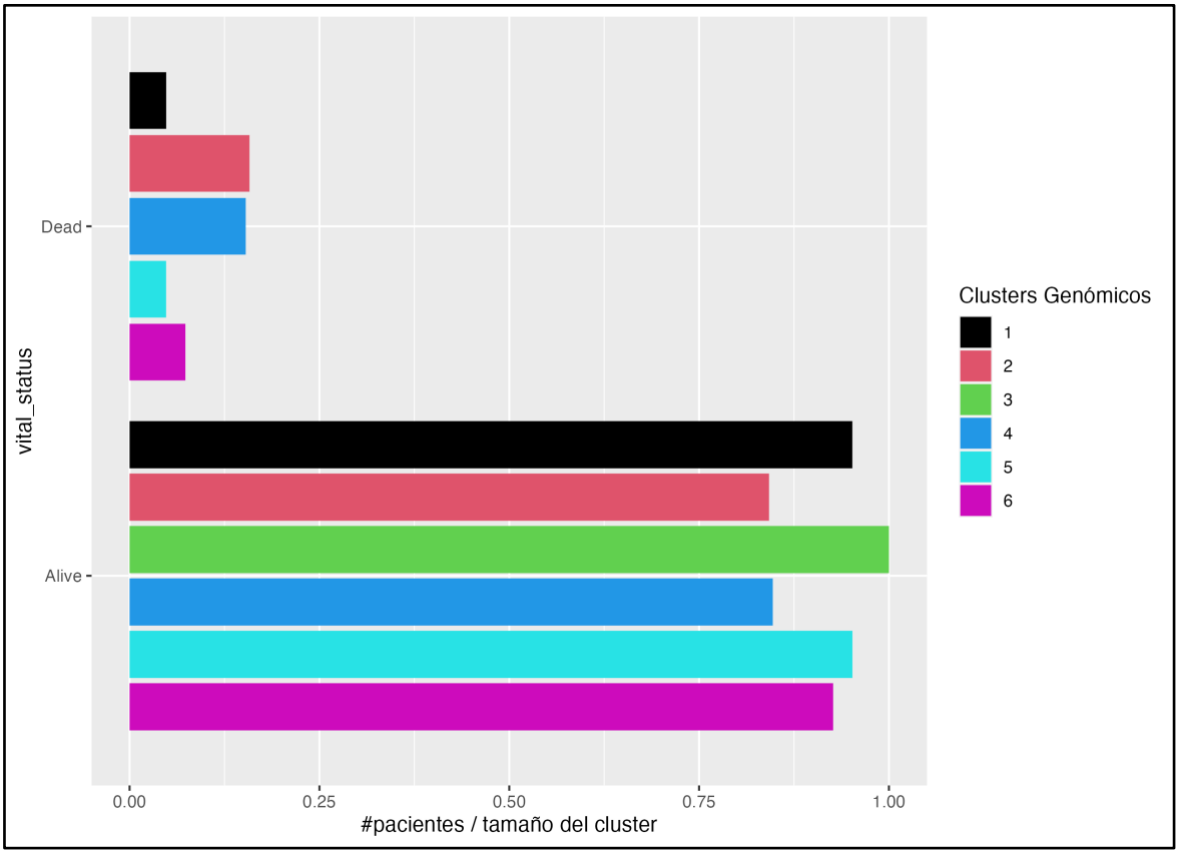


Figura 44: Comparación del estado vital de los pacientes con cáncer de mama según su *cluster* genómico asignado.

Una diferencia significativa (*p-value* = 2.7e-4) entre *clusters* surgió de la comparación de la histología del cáncer de los pacientes con cáncer de mama entre los *clusters* genómicos. Los *clusters* genómicos 4 y 6 presentan un desequilibrio en el balance histológico: carcinoma ductal infiltrante/carcinoma lobulillar infiltrante con un porcentaje mayor de pacientes con un carcinoma ductal infiltrante (82.4% y 78.8%, respectivamente) en comparación con los demás *clusters* (ver Fig. 45).

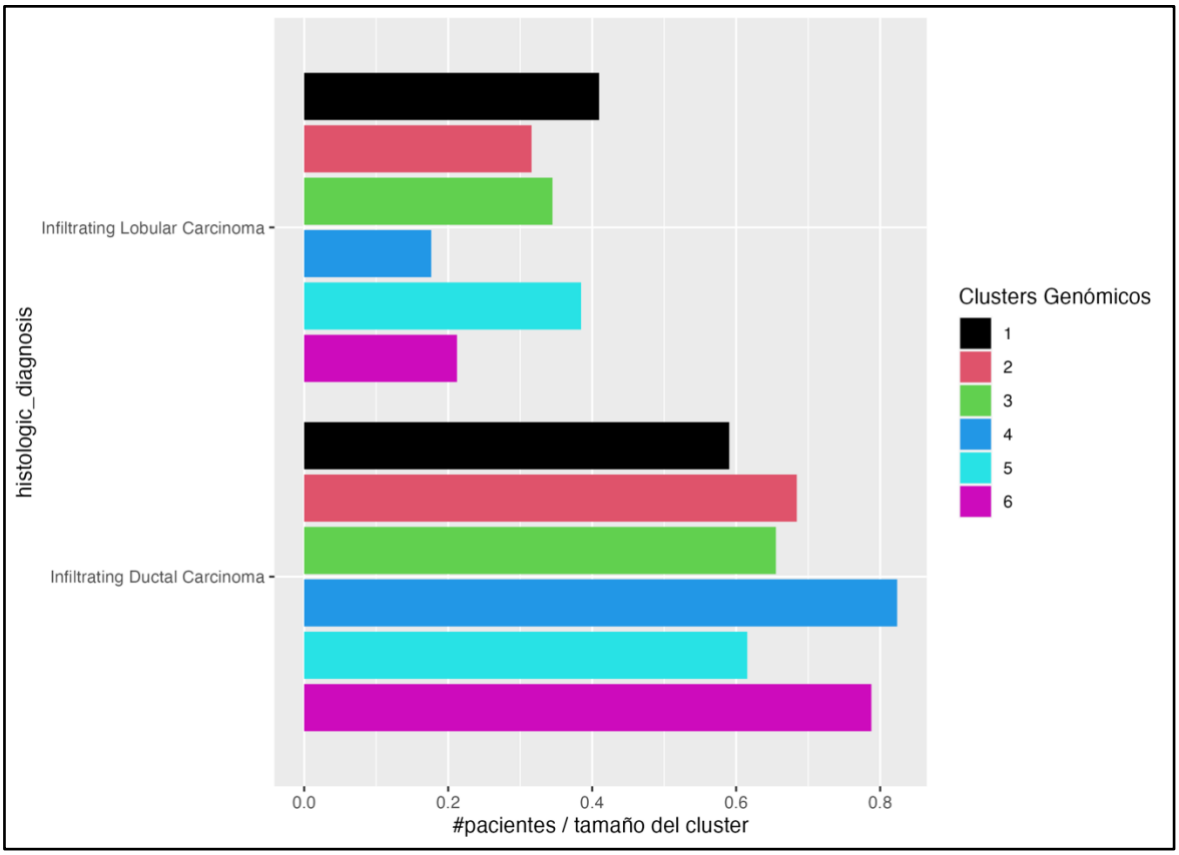


Figura 45: Comparación de la histología del tumor de los pacientes con cáncer de mama según su *cluster* genómico asignado.

3. Integración de los datos

La integración de los datos ómicos de los 913 pacientes seleccionados busca establecer una representación integrada de los datos. A continuación, se describe primero los resultados de la implementación del modelo de aprendizaje automático DGCCA, y luego se presenta la caracterización de los *clusters* de la representación integrada.

3.1. Implementación del modelo de aprendizaje automático

Se observó una disminución general del error de entrenamiento entre la primera iteración: 5.3e14, y la última iteración del algoritmo DGCCA: 3.2e13. Sin embargo, esta disminución no es lineal, el error de entrenamiento máximo fue encontrado para la iteración 54: 7.3e23 (Fig. 46).

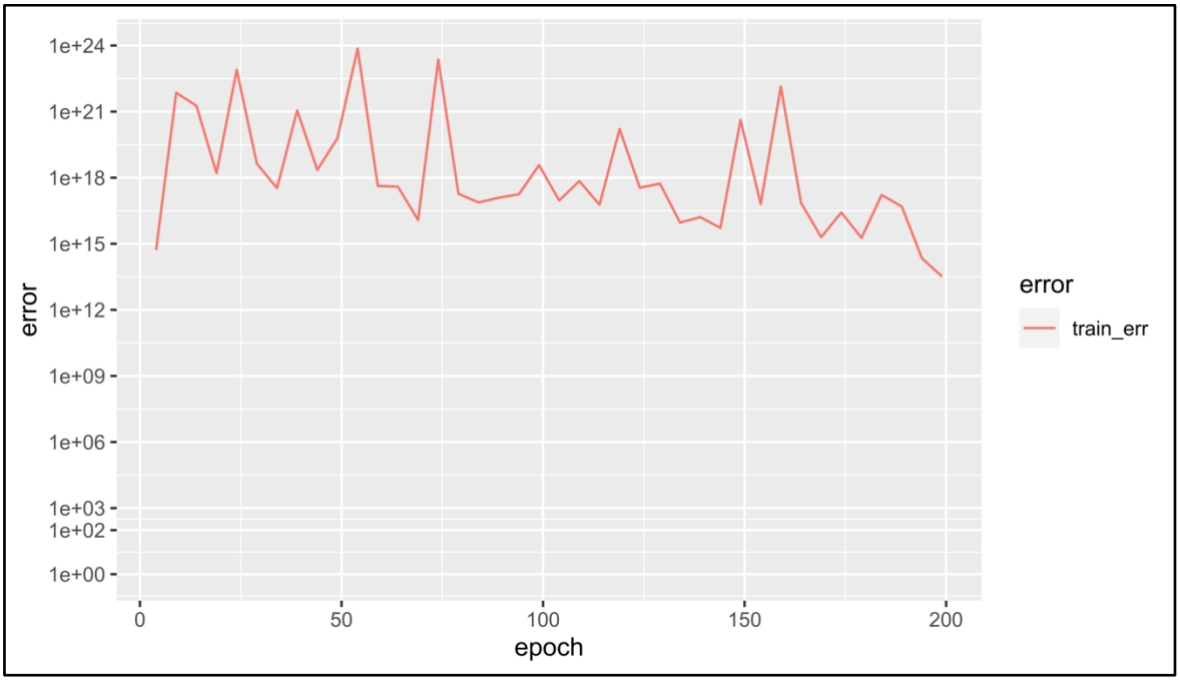


Figura 46: Evolución del error de entrenamiento o error de reconstrucción a lo largo de las iteraciones (*epochs*) del algoritmo DGCCA.

3.2. Identificación de los biomarcadores compartidos mediante el uso del modelo optimizado

El análisis de la representación integrada G de los 913 pacientes resultó en tres *clusters*. El *cluster* 2 está compuesto exclusivamente por pacientes con cáncer de mama mientras que los *clusters* 1 y 3 son mixtos: el *cluster* 1 está compuesto por 3 pacientes con cáncer de mama y 325 pacientes con cáncer de endometrio, el *cluster* 3 está compuesto por 238 pacientes con cáncer de mama y 23 pacientes con cáncer de endometrio (Fig. 47). La composición de los *clusters* de la representación integrada G fue reportada en el Anexo 8.

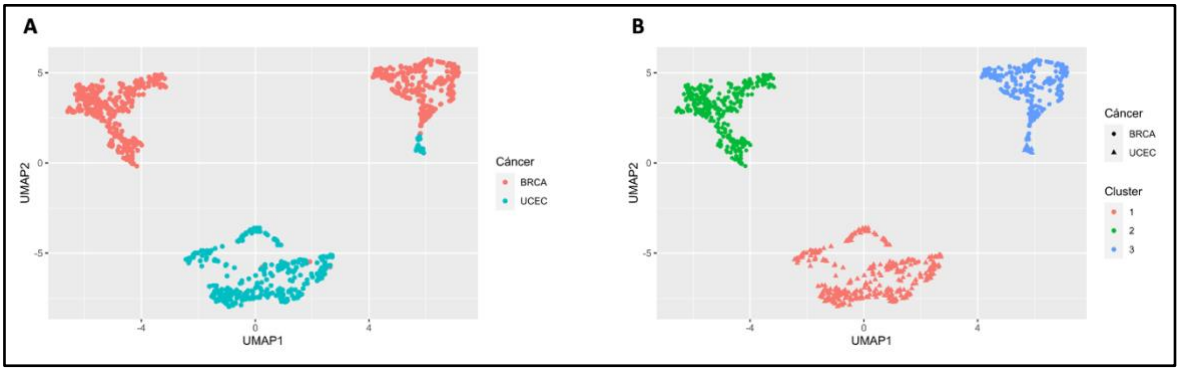


Figura 47: Visualización UMAP de los pacientes de cáncer de mama y cáncer de endometrio obtenida a partir de la representación integrada G y agrupación *KMeans*. A: Visualización de los pacientes en el espacio bidimensional UMAP1 vs. UMAP2. B: *Clustering KMeans* de los pacientes (k = 3).

Se llevó a cabo la comparación de la expresión génica promedio, de la metilación promedio de las regiones promotoras y del número de copias promedio de los 12594 genes usados para construir la representación integrada G entre los tres *clusters* obtenidos usando la prueba estadística no paramétrica de Mann-Whitney-Wilcoxon para datos pareados (Fig. 48-50). Esta prueba estadística para la expresión génica promedio reveló diferencias significativas entre los *tres clusters* (*p-value* < 2.2e-16 para todas las comparaciones) con una expresión génica promedio mayor en el *cluster* 1 (mediana de la expresión génica promedio de los 12594 genes igual a 0.052), intermedia en el *cluster* 3 (mediana igual a 0.043) y menor en el *cluster* 2 (mediana igual a 0.040). Al igual que para el análisis de los *clusters* transcriptómicos, la visualización de los resultados no permite emitir una conclusión de diferencia en la expresión génica entre *cluster*s dado que la mayoría de los genes solo presentan una expresión residual (ver Fig. 48).

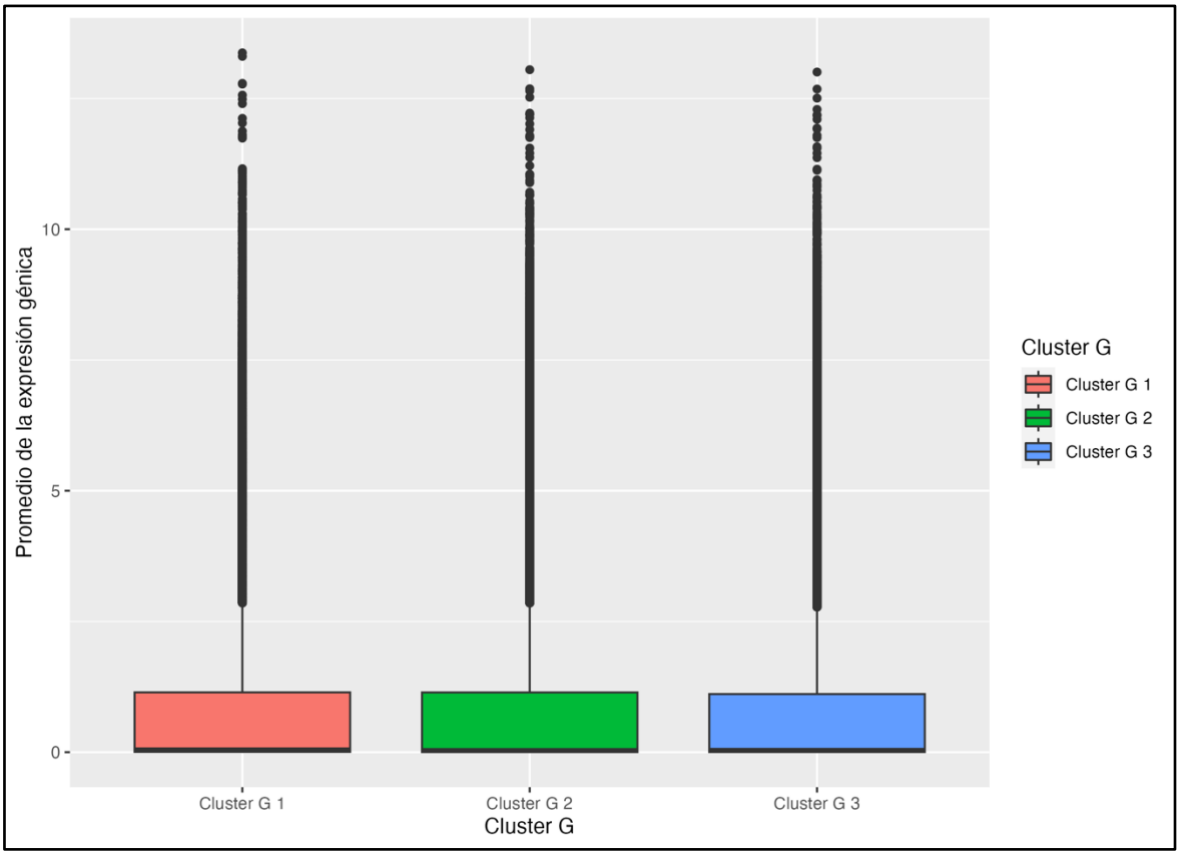


Figura 48: Comparación de la expresión génica promedio para cada gen por *cluster* de la representación integrada G.

En cuanto a la metilación promedio de las regiones promotoras, la prueba estadística delató diferencias significativas entre el *cluster 2* y los *clusters* 1 y 3 (*p-value* < 2.2e-16 para ambas comparaciones) con una metilación promedio mayor en el *cluster* 2: mediana de la metilación promedio de las regiones promotoras de los 12594 genes igual a 0.159, contra 0.133 en el *cluster* 1 y 0.159 en el *cluster* 3. El análisis visual de los resultados muestra que la diferencia de metilación promedio del ADN entre los *clusters* 2 y 3 no está clara (Fig. 49). Se puede concluir que la metilación promedio de las regiones promotoras es mayor en el *cluster* 2 a comparación del *cluster* 1.

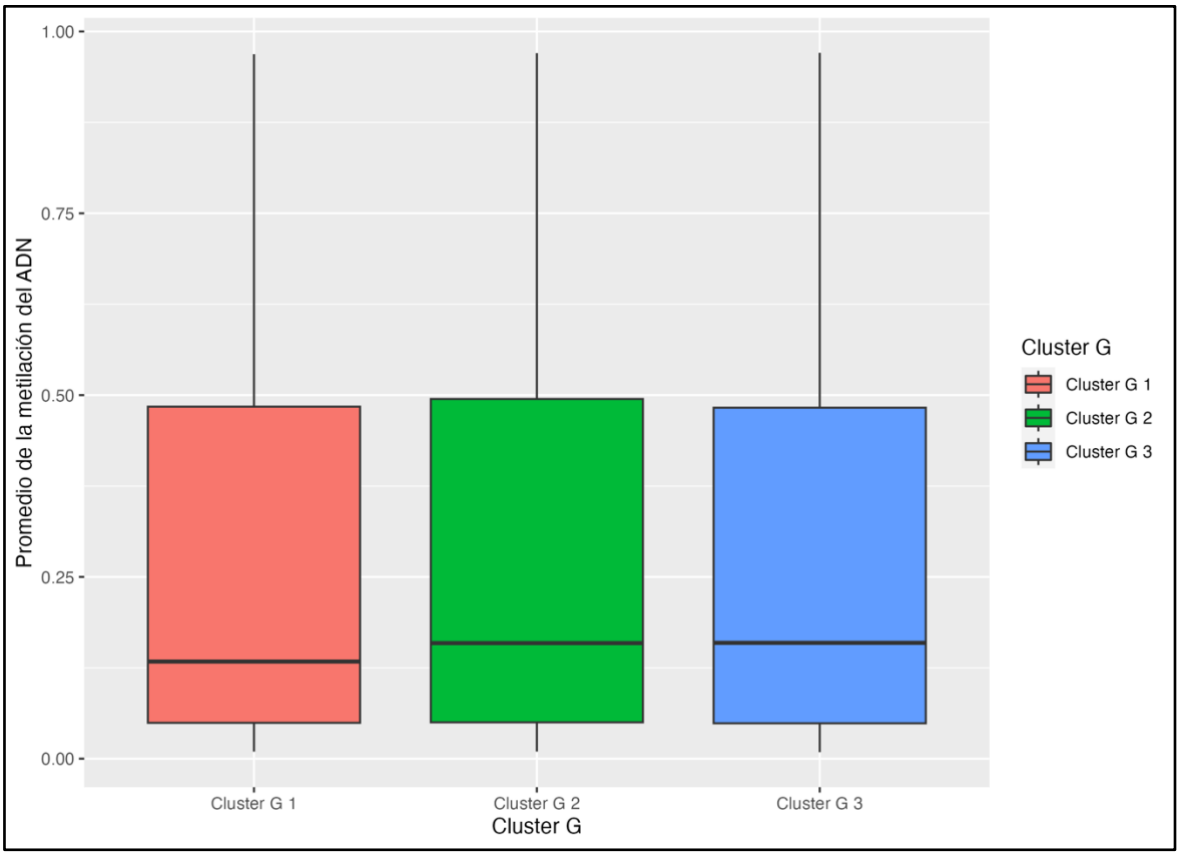


Figura 49: Comparación de la metilación promedio del ADN para cada gen por *cluster* de la representación integrada G.

Finalmente, para el análisis comparativo del número promedio de copias de los 12594 genes entre *clusters*, diferencias significativas fueron obtenidas para cada una de las comparaciones (*p-value* < 2.2e-16). Un número de copias promedio mayor fue encontrado en el *cluster* 3 con una mediana del número de copias promedio de los 12594 genes igual a 3.625. El número de copias promedio puede ser calificado de intermedio en el *cluster* 1 con una mediana igual a 2.104 y menor en el *cluster* 2 con una mediana igual a 1.988. Sin embargo, esta diferencia no es suficiente para afirmar que hay una diferencia biológica en cuanto al número de copias de los genes entre los *clusters* 1 y 2 (ver Fig. 50). Se puede concluir que el número de copias de los genes promedio es mayor en el *cluster* 3 en comparación con los demás *clusters* de la representación integrada G.

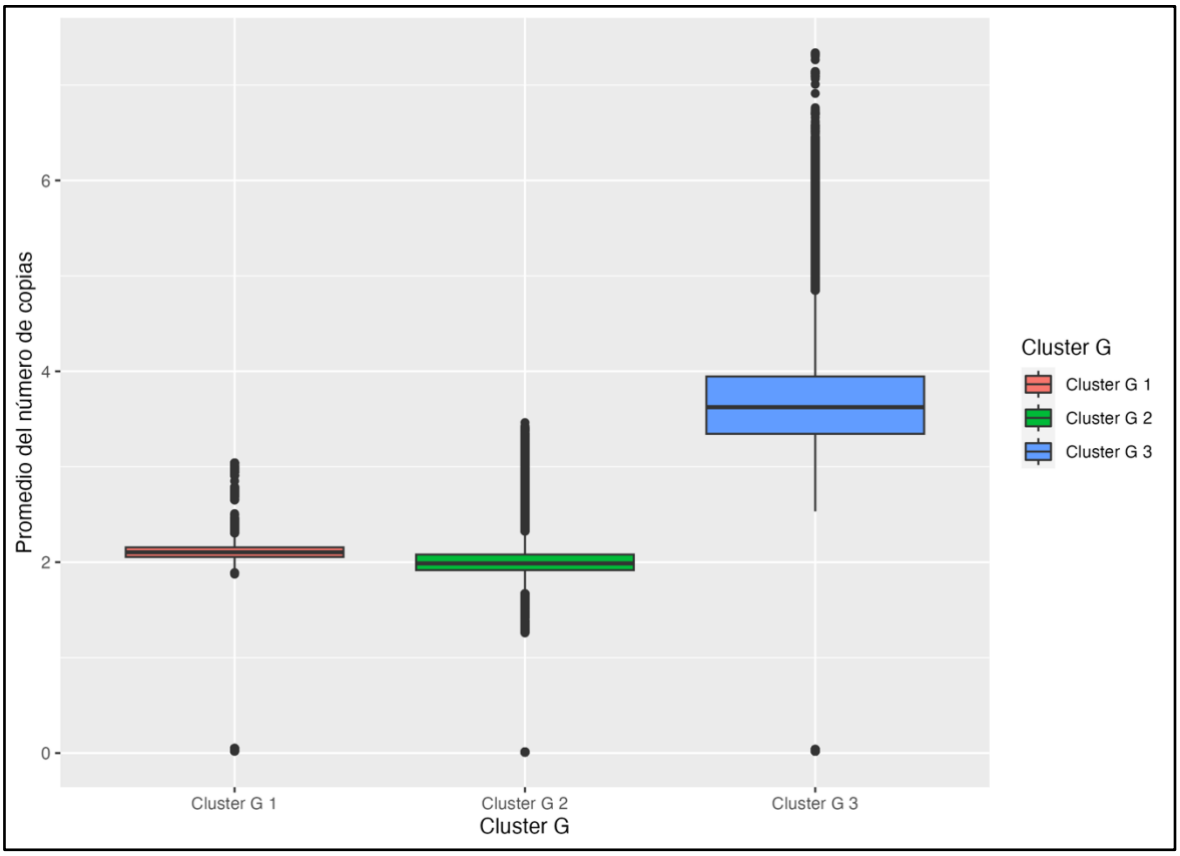


Figura 50: Comparación del número de copias promedio para cada gen por *cluster* de la representación integrada G.

Para recapitular sobre las características ómicas de los *clusters* de la representación integrada G, el *cluster* 2 presenta una metilación de las regiones promotoras promedio mayor que el *cluster* 1 mientras que el *cluster* 3 exhibe un mayor número de copias promedio que los demás *clusters*. Por lo tanto, se puede afirmar que el número de copias de los genes no está asociado con la metilación de las regiones promotoras.

100 biomarcadores fueron identificados para cada *cluster* de la representación integrada G por tener un puntaje de información mutua promedio (entre las tres vistas) alto contra las variables sintéticas UMAP1 y UMAP2. Estos fueron reportados en el Anexo 9. Es necesario mencionar que independientemente del *cluster* considerado, se obtuvieron puntajes de información mutua mayores para los genes a nivel transcriptómico y epigenómico. En efecto, los puntajes de información mutua a nivel genómico tanto para UMAP1 como UMAP2 son más bajos con un máximo igual a 0.52 contra un máximo igual a 1 a nivel epigenómico o transcriptómico (Fig. 51-53). Este resultado se debe a la

redundancia de algunos valores en la matriz de datos genómicos: muchos pacientes presentan el mismo número de copias para un gen dado.

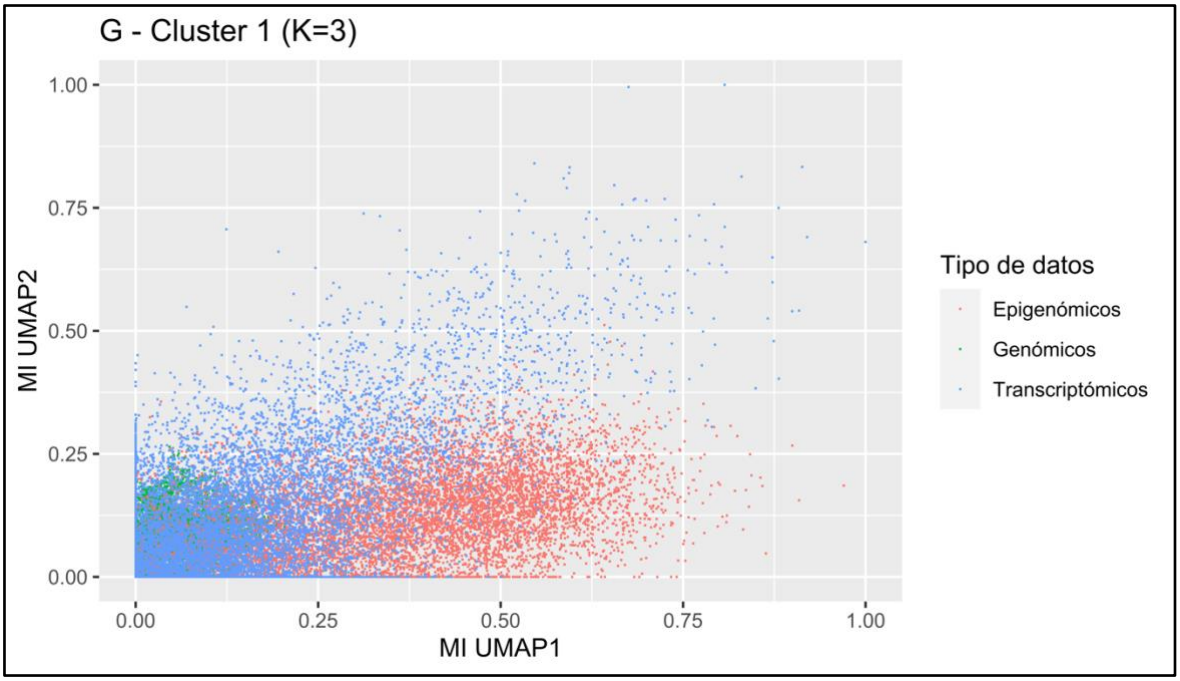


Figura 51: Visualización de las variables ómicas (transcriptómicas, epigenómicas y genómicas) según su puntaje de información mutua contra las variables sintéticas UMAP1 y UMAP2 en el *cluster* 1 de la representación integrada G.

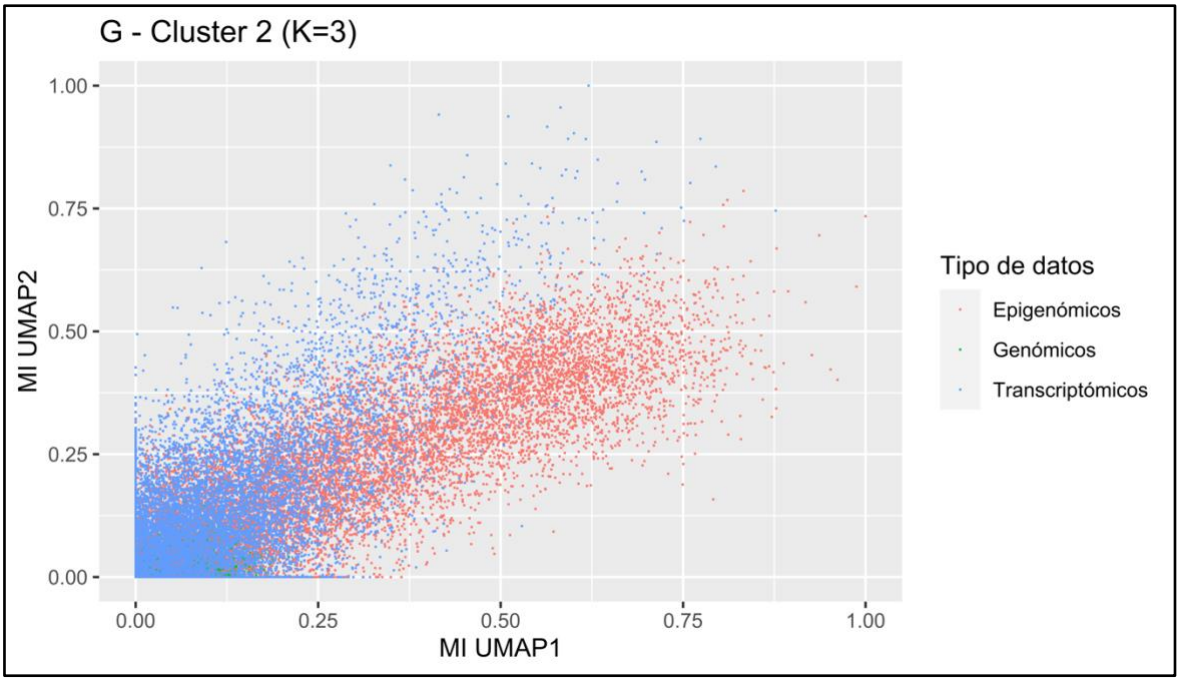


Figura 52: Visualización de las variables ómicas (transcriptómicas, epigenómicas y genómicas) según su puntaje de información mutua contra las variables sintéticas UMAP1 y UMAP2 en el *cluster* 2 de la representación integrada G.

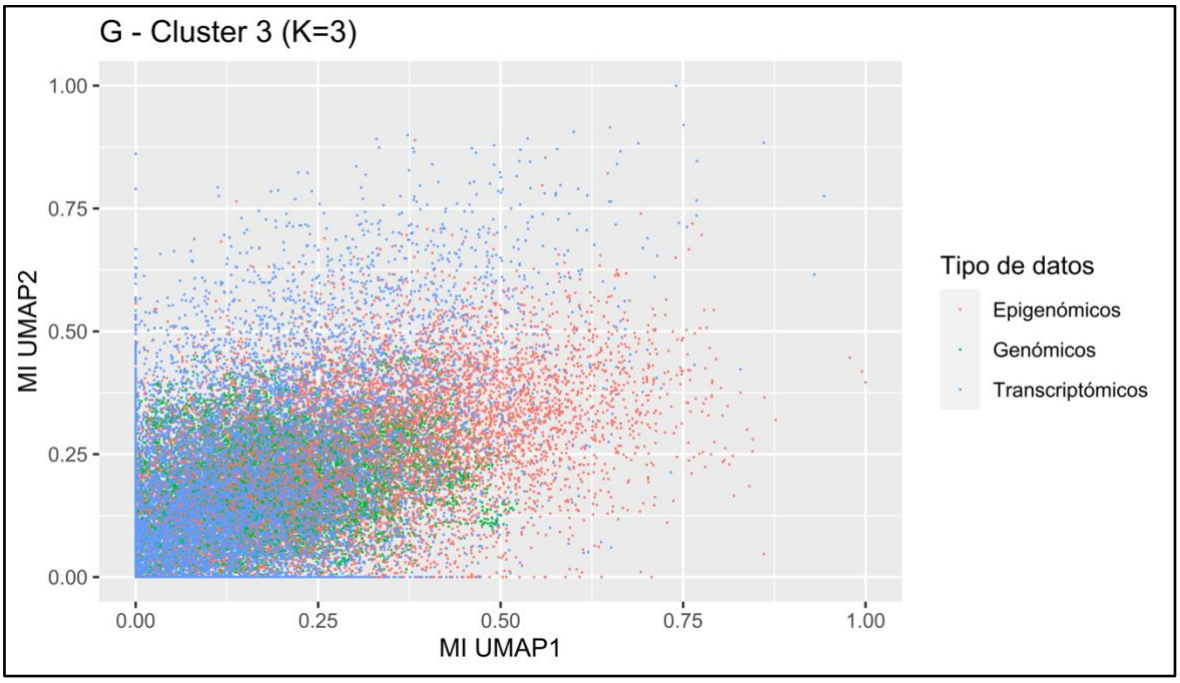


Figura 53: Visualización de las variables ómicas (transcriptómicas, epigenómicas y genómicas) según su puntaje de información mutua contra las variables sintéticas UMAP1 y UMAP2 en el *cluster* 3 de la representación integrada G.

De la misma manera que para el análisis de los *clusters* genómicos, el estudio clínico de los pacientes según su *cluster* de la representación integrada G asignado se hizo separando los pacientes con cáncer de mama de los pacientes con cáncer de endometrio en cada *cluster* para eliminar los sesgos observados entre los dos tipos de cáncer. Cabe destacar el desequilibrio entre ambos tipos de cáncer en cada *cluster* (Fig. 54). Ninguna diferencia significativa fue encontrada comparando los pacientes con cáncer de endometrio entre los tres *clusters*. Para los pacientes con cáncer de mama, diferencias significativas fueron expuestas en cuanto a la histología y a la estadificación del tumor.

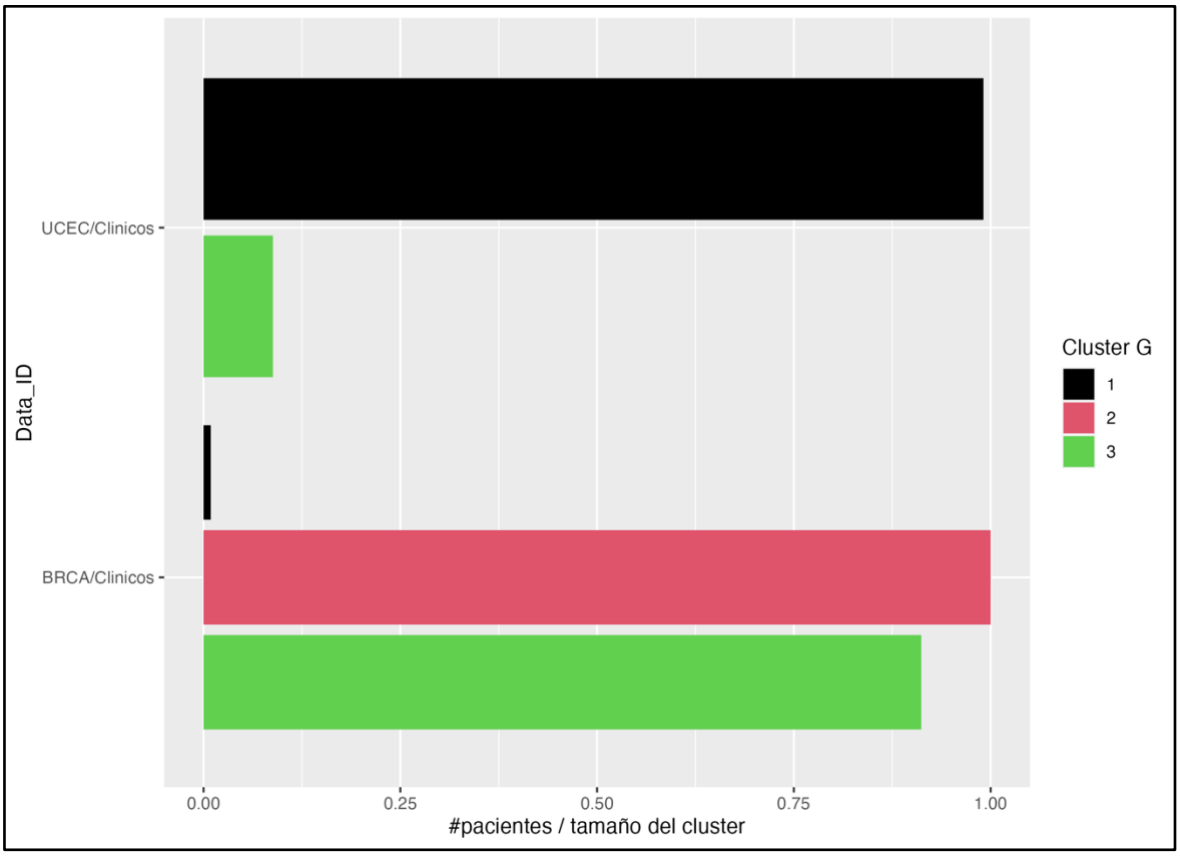


Figura 54: Comparación del tipo de cáncer de los pacientes según su *cluster* de la representación integrada G asignado.

Representación integrada G – Pacientes con cáncer de mama

El análisis de la histología del cáncer de los pacientes con cáncer de mama entre los *clusters* de la representación integrada G reveló una diferencia significativa (*p-value* = 0.001) entre *clusters*. Los 3 pacientes con cáncer de mama del *cluster* 1 tienen un carcinoma ductal infiltrante. El *cluster* 2 es el más equilibrado con 33.6% de pacientes de mama con carcinoma lobulillar infiltrante y 66.4% con carcinoma ductal infiltrante mientras que el *cluster* 3 cuenta con 20.2% de pacientes de mama con carcinoma lobulillar infiltrante y 79.8% con carcinoma ductal infiltrante (Fig. 55).

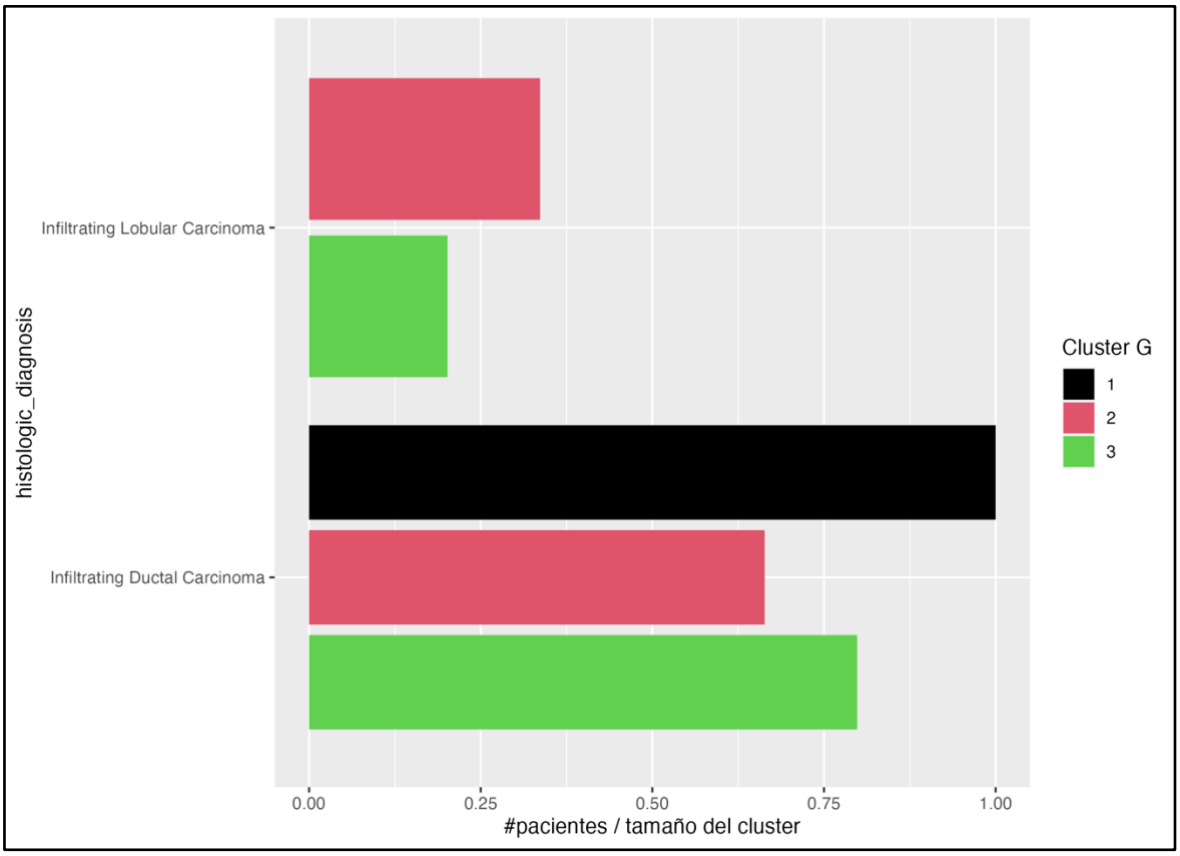


Figura 55: Comparación de la histología del tumor de los pacientes con cáncer de mama según su *cluster* de la representación integrada G asignado.

La comparación de la estadificación del cáncer de los pacientes con cáncer de mama dio una diferencia significativa (*p-value* = 0.015) entre *clusters*. Importantemente, los 3 pacientes con cáncer de mama del *cluster* 1 son clasificados en etapa III: tienen un cáncer muy avanzado. Además, se puede observar un porcentaje mayor de pacientes clasificados en etapa II y III en el *cluster* 3 (56.7% y 26.1%, respectivamente) que en el *cluster* 2 (50.0% y 21.3%, respectivamente). Por el contrario, el porcentaje de pacientes con cáncer de mama clasificados en etapa I es menor en el *cluster* 3: 5.9% que en el *cluster* 2: 11.4%. La estadificación del cáncer no fue medida para algunos pacientes de cáncer de mama de los *clusters* 2 y 3 (ver Fig. 56). Se puede concluir de esta figura que el *cluster* 3 tiene más pacientes de cáncer de mama con un tumor medianamente avanzado (etapa II) o muy avanzado (etapa III) que el *cluster* 2.

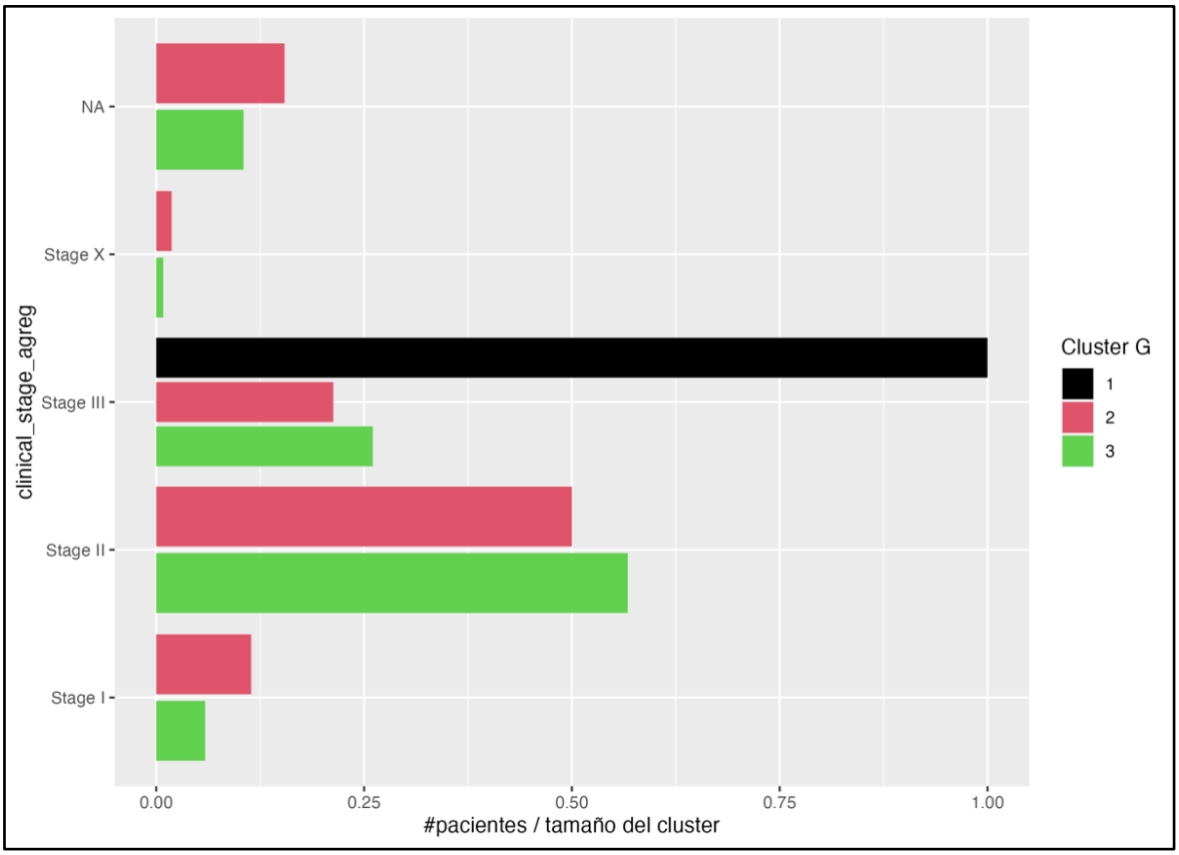


Figura 56: Comparación de la estadificación del cáncer de los pacientes con cáncer de mama según su *cluster* de la representación integrada G asignado.

Para resumir, los *clusters* obtenidos a partir de la representación integrada G tienen las siguientes características:

- El *cluster* 1 contiene la mayoría de los pacientes con cáncer de endometrio y 3 pacientes con cáncer de mama con carcinoma ductal infiltrante muy avanzado (etapa III).
- El *cluster* 2 es caracterizado por una metilación promedio de las regiones promotoras mayor que en el *cluster* 1. Está compuesto exclusivamente por pacientes con cáncer de mama con un balance relativamente equilibrado entre carcinoma ductal y carcinamo lobulillar infiltrante. Es el *cluster* que cuenta más cánceres tempranos.
- El *cluster* 3 es caracterizado por un número promedio de copias de genes mucho mayor que en los demás *clusters*. Está compuesto por una mayoría de pacientes con cáncer de mama, más específicamente, carcinoma ductal infiltrante, y 23 pacientes con cáncer de endometrio.

4. Caracterización de los resultados

El trabajo de caracterización biológica de los biomarcadores de los *clusters* genómicos, epigenómicos, transcriptómicos y de los *clusters* de la representación integrada G pretende determinar la naturaleza de estos genes, es decir los procesos biológicos en los cuales están involucrados. Esta sección se divide en dos partes: primero, el análisis cualitativo de los procesos biológicos asociados a los biomarcadores encontrados previamente, y, luego, el estudio más profundo de los biomarcadores compartidos entre el análisis de datos ómicos por separado y el análisis integrativo.

4.1. Caracterización biológica de los biomarcadores encontrados

Para cada lista de biomarcadores, los términos GO relativos a procesos biológicos fueron enriquecidos y visualizados, junto a sus biomarcadores asociados, en diagramas de red. El tamaño de los nodos de términos GO corresponde al número de biomarcadores asociados, el color representa el *p-value* del enriquecimiento del término considerado. Solo fueron representados términos GO enriquecidos con un *p-value* menor a 0.05. No se encontraron términos GO interesantes para todas las listas de biomarcadores. Por ejemplo, los términos GO asociados con los biomarcadores del *cluster* transcriptómico 2, del *cluster* epigenómico 3 y de los *clusters* genómicos 1, 3, 5 y 6 no fueron seleccionados para ser visualizados. Los términos GO de interés pueden ser clasificados en 4 categorías:

- Términos GO relativos a la regulación del ciclo celular: Regulación positiva de la proliferación de las células endoteliales, Regulación positiva de la proliferación celular, Regulación de la replicación del ADN, División Celular, Reparación del ADN, Punto de control (*checkpoint*) de daño del ADN, y, Regulación positiva de la vía de las proteínas MAP cinasas activadas por mitógenos (MAPK).
- Términos GO relativos al proceso de apoptosis: Regulación negativa del proceso de apoptosis y Regulación del proceso de apoptosis.

- Términos GO relativos al proceso de angiogénesis: Regulación positiva de la proliferación de las células endoteliales y Vía de señalización de los receptores del factor de crecimiento endotelial vascular (VEGF).
- Término GO relativo al envejecimiento celular.

Por un lado, la desregulación del balance proliferación celular/apoptosis a favor de una proliferación celular excesiva y de una apoptosis inhibida es una característica general del cáncer que promueve el crecimiento del tumor. Por otro lado, la angiogénesis es un proceso fisiológico por el cual se forman vasos sanguíneos nuevos a partir de los vasos preexistentes formado durante el desarrollo embriológico. VEGF es el mediador clave de la angiogénesis en el contexto del cáncer. Su expresión es regulada positivamente en este contexto por la expresión de oncogenes, por una variedad de factores de crecimiento y también por hipoxia. La angiogénesis es esencial para el desarrollo y el crecimiento del cáncer: antes de que un tumor pueda crecer más allá de 1-2 mm, requiere vasos sanguíneos para obtener nutrientes y oxígeno. La producción de VEGF y otros factores de crecimiento por parte del tumor da como resultado el "cambio angiogénico", donde se forma nueva vasculatura dentro y alrededor del tumor gracias a la proliferación de las células endoteliales, lo que le permite crecer exponencialmente (Carmeliet, 2005). Finalmente, el envejecimiento celular es altamente relacionado con el cáncer dado que la longitud de los telómeros se acorta con la edad. Los telómeros corresponden a las regiones de ADN no codificante ubicadas en los extremos de los cromosomas y cuya función es la estabilidad estructural de los cromosomas. El acortamiento progresivo de los telómeros conduce a la senescencia, la apoptosis – estos dos fenómenos son favorables ya que permiten evitar la aparición de tumores – o la transformación oncogénica de las células somáticas, lo que afecta la salud y la vida útil de un individuo (Shammas, 2011).

A continuación, se presentan los biomarcadores y los términos GO asociados del *cluster* transcriptómico 1, de los *clusters* epigenómicos 1 y 2, de los *clusters* genómicos 2 y 4, y, de los *clusters* de la representación integrada G 1, 2, y 3 (ver Fig. 57-65).

Datos transcriptómicos

Los biomarcadores ATM, E2F1, PTPN11 y CLOCK del *cluster* transcriptómico 1 – compuesto exclusivamente por pacientes con cáncer de mama – son asociados al término de Punto de control de daño del ADN mientras que los biomarcadores AHCTF1, CKAP5,

BRCC3, SMC1A, SMC3, NEDD1, WAPL, USP9X y UBE2S son asociados al término de División Celular (ver Fig. 57).

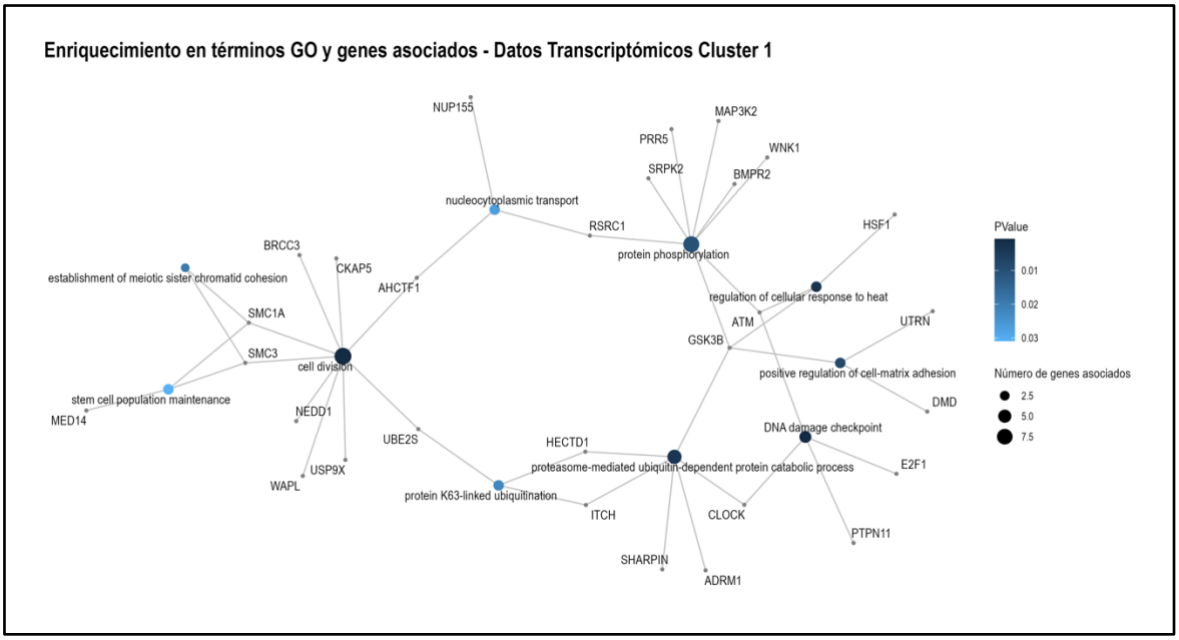


Figura 57: Visualización de los términos GO enriquecidos y los genes asociados del *cluster* transcriptómico 1.

Datos epigenómicos

Los biomarcadores SON, LTF, PRNP, WNT7A, SOX8, MT3, SNCA, PROK2, PIM1, FAB1 y CCND del *cluster* epigenómico 1 – compuesto exclusivamente por pacientes con cáncer de mama – son asociados al término de Regulación negativa del proceso de apoptosis (ver Fig. 58).

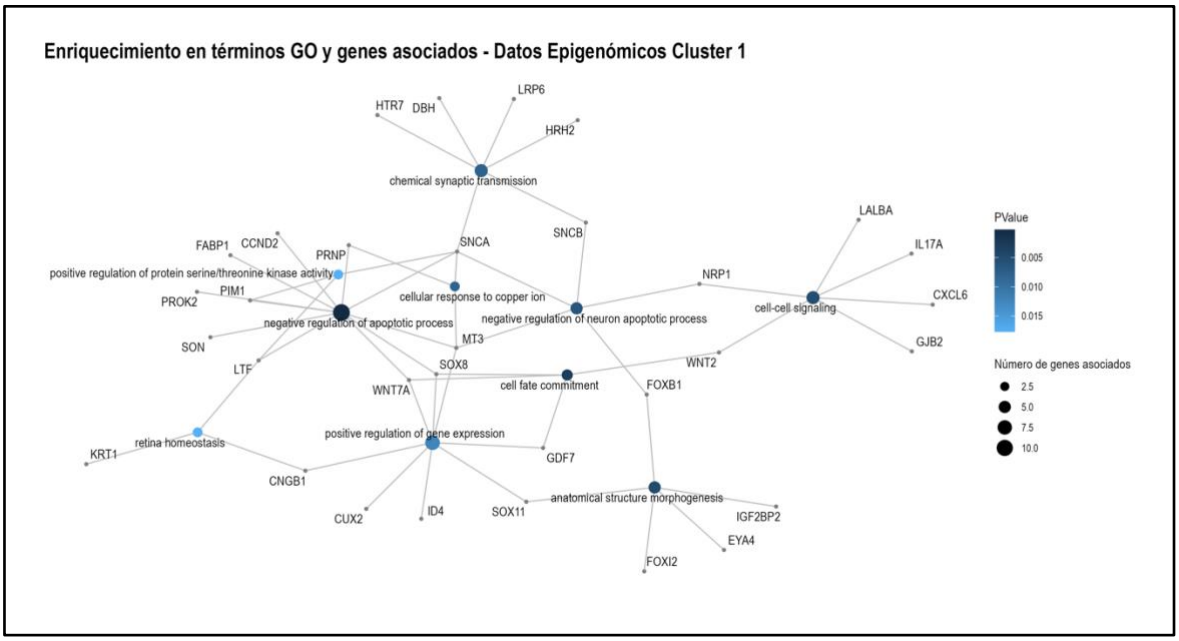


Figura 58: Visualización de los términos GO enriquecidos y los genes asociados del *cluster* epigenómico 1.

Los biomarcadores DPP4, CCND2, EDNRB, REG1A, REG3A, PKHD1, FGF6, SFRP1, WNT2, IGF2 y KIT del *cluster* epigenómico 2 – también compuesto exclusivamente por pacientes con cáncer de mama – son asociados al término de Regulación positiva de la proliferación celular. Además, los genes IGF2 y KIT también son involucrados en la regulación positiva de la vía de las proteínas MAP cinasas activadas por mitógenos (ver Fig. 59). Para concluir sobre la naturaleza de los biomarcadores de los *clusters* epigenómicos 1 y 2, varios de ellos están involucrados en el balance proliferación celular/apoptosis.

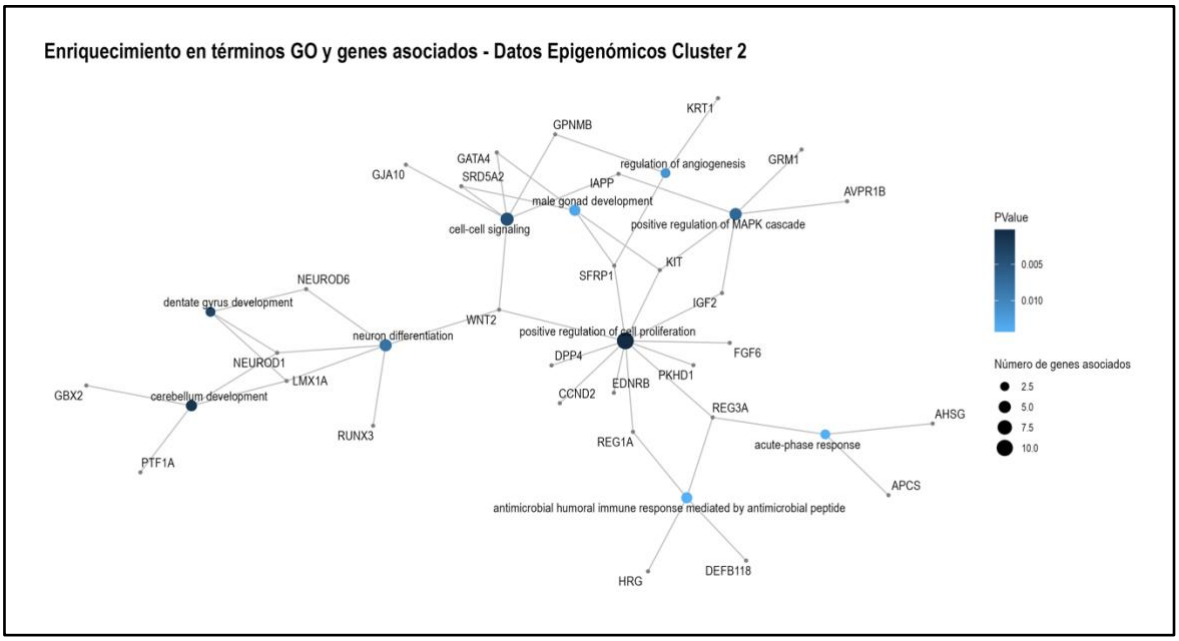


Figura 59: Visualización de los términos GO enriquecidos y los genes asociados del *cluster* epigenómico 2.

Datos genómicos

Los biomarcadores ARNT, ECM1 y SIRT1 del *cluster* genómico 2 son asociados al término de Regulación positiva de la proliferación de las células endoteliales. Además, los biomarcadores SIRT1 y CDK1 son marcadores del envejecimiento celular (ver Fig. 60).

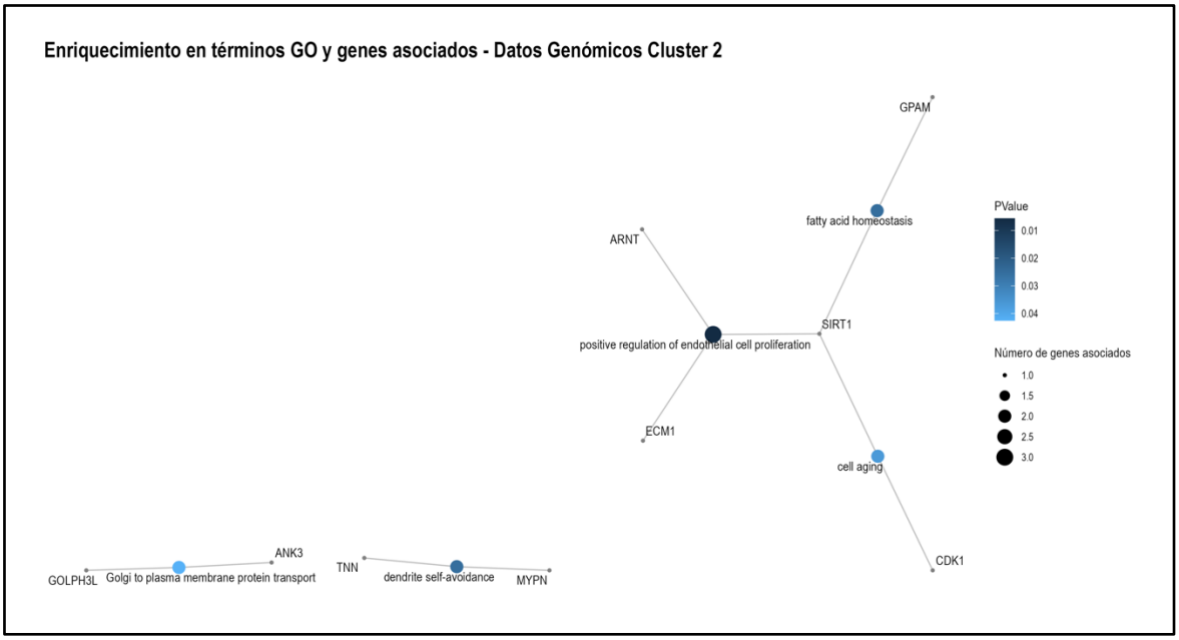


Figura 60: Visualización de los términos GO enriquecidos y los genes asociados del *cluster* genómico 2.

Los biomarcadores PTK2 y SULF1 del *cluster* genómico 4 son asociados al término de Vía de señalización de los receptores del factor de crecimiento endotelial vascular (ver Fig. 61). Para concluir sobre la naturaleza de los biomarcadores de los *clusters* genómicos 2 y 4, los cuales presentan una alta tasa de mortalidad para sus pacientes con cáncer de mama, varios de ellos están involucrados en el proceso de angiogénesis el cual contribuye al suministro de nutrientes y oxígeno por el tumor. Esta asociación entre la alta tasa de mortalidad y la presencia de biomarcadores involucrados en el proceso de angiogénesis es interesante dado que la angiogénesis favorece el crecimiento cancerígeno. Dado que los *clusters* genómicos 2 y 4 son mixtos, se puede decir que la angiogénesis parece ser un proceso común de algunos pacientes con cáncer de mama y otros con cáncer de endometrio.

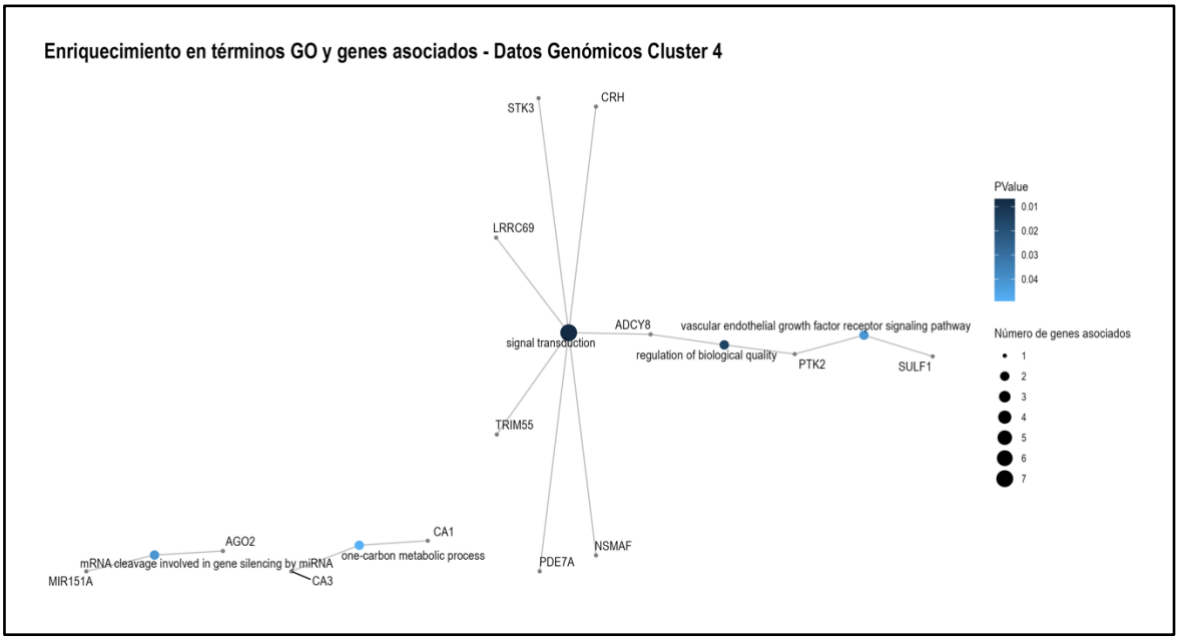


Figura 61: Visualización de los términos GO enriquecidos y los genes asociados del *cluster* genómico 4.

Representación integrada G

Los biomarcadores POLK, PARP4, RFC3, RFC1, APEX1, SMARCA5 y CDC5L del *cluster* 1 de la representación integrada G – compuesto por 3 pacientes con cáncer de mama y 325 pacientes con cáncer de endometrio – son asociados al término de Reparación del ADN (ver Fig. 62).

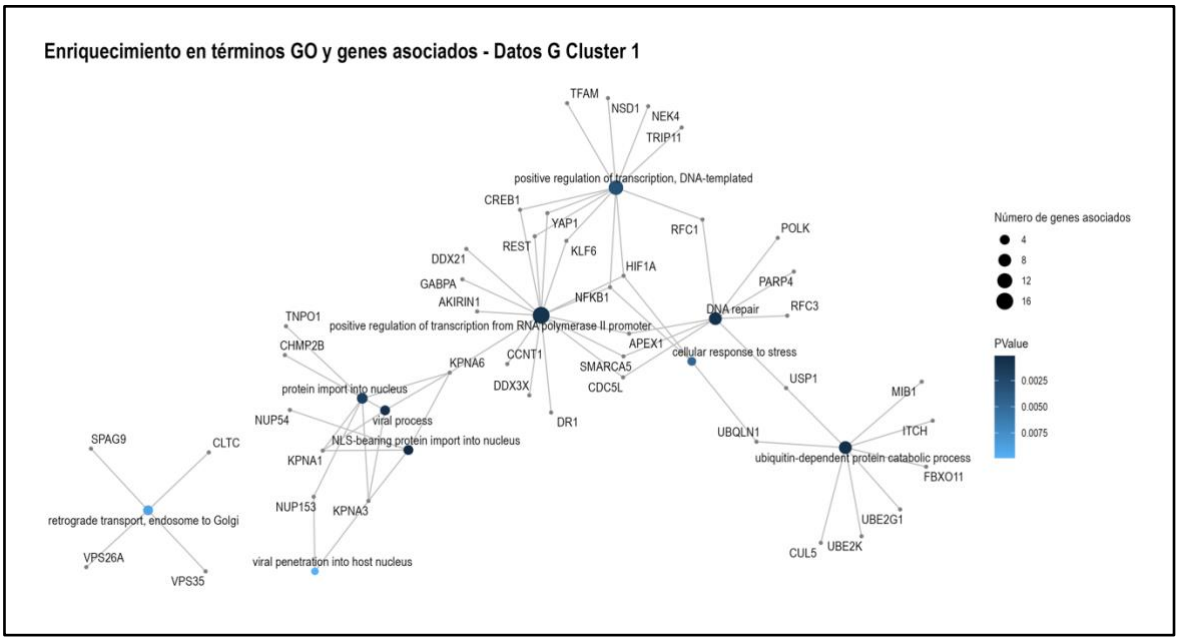


Figura 62: Visualización de los términos GO enriquecidos y los genes asociados del *cluster* 1 de la representación integrada G.

Los biomarcadores USP37, INO80E, SMARCA5, SMC3 y STAG1 del *cluster* 2 de la representación integrada G – compuesto exclusivamente por pacientes con cáncer de mama – son asociados al término de Regulación de la replicación del ADN y los biomarcadores SMC3, STAG1, CLTC, PDS5A, ANAPC1, BIRC6, KLHL9 y ZWILCH al término de División celular (ver Fig. 63).

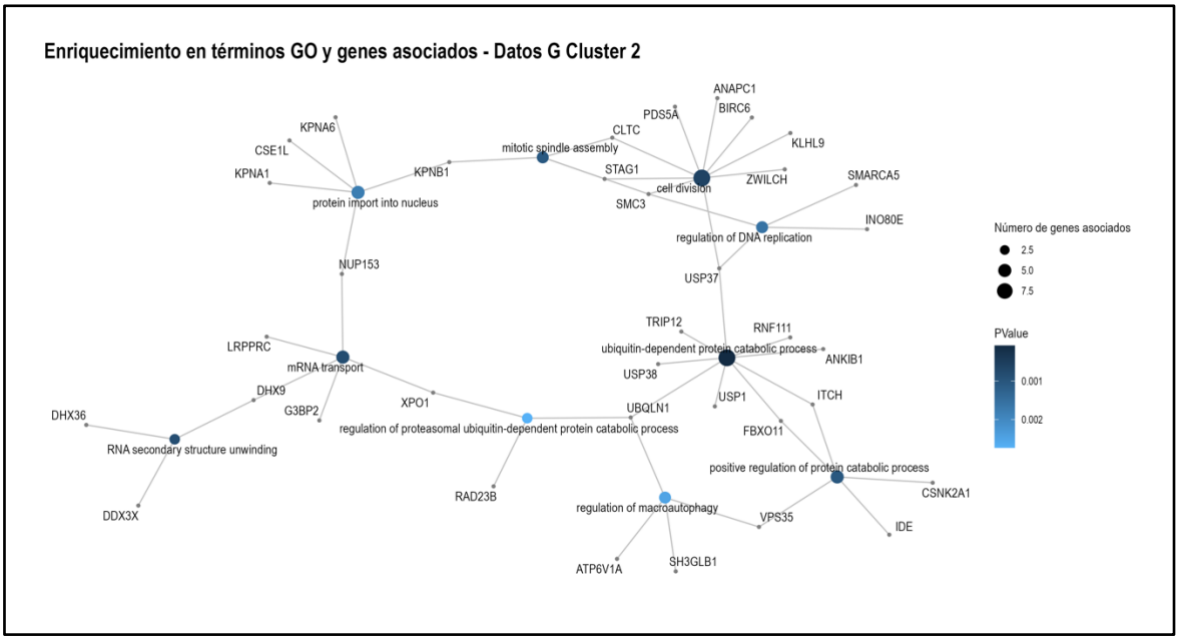


Figura 63: Visualización de los términos GO enriquecidos y los genes asociados del *cluster* 2 de la representación integrada G.

Los biomarcadores SLK, BIRC6, FBXO11, PAX8 y BCL2L12 del *cluster* 3 de la representación integrada G – compuesto por 238 pacientes con cáncer de mama y 23 pacientes con cáncer de endometrio – son asociados al término de Regulación del proceso de apoptosis (ver Fig. 64). Para concluir sobre la naturaleza de los biomarcadores de los *clusters* de la representación integrada G, muchos de ellos están involucrados en el balance proliferación celular/apoptosis como para los biomarcadores epigenómicos. Los *clusters* 1 y 3 de la representación integrada G son mixtos. Por lo tanto, se puede avanzar que tanto los pacientes con cáncer de mama y los pacientes con cáncer de endometrio presentan biomarcadores relacionados con los procesos de proliferación celular y apoptosis.

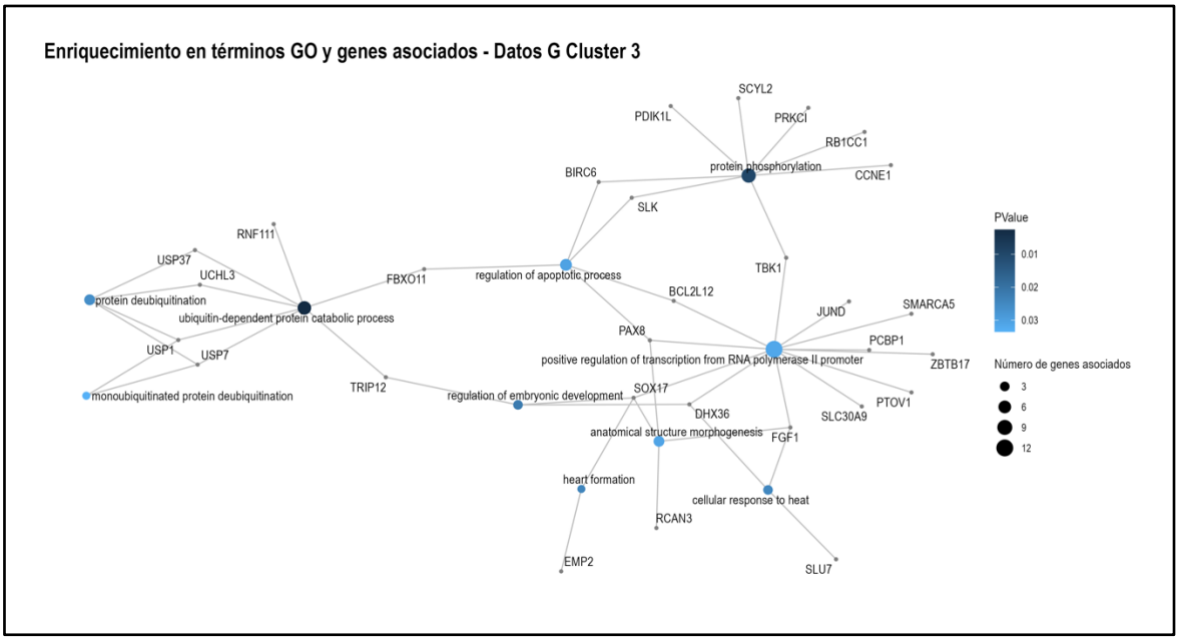


Figura 64: Visualización de los términos GO enriquecidos y los genes asociados del *cluster* 3 de la representación integrada G.

4.2. Comparación de los resultados obtenidos con las dos metodologías

Es interesante notar la existencia de varios biomarcadores compartidos entre el análisis de datos ómicos de distintos tipos y también entre el análisis de datos ómicos de un solo tipo y el estudio integrativo (ver Fig. 65).

- Los biomarcadores RRS1 y HHIPL2 son compartidos entre los *clusters* genómicos y epigenómicos. El gen RRS1 codifica para una proteína de 365 aminoácidos

involucrada en el ensamblaje de la subunidad grande del ribosoma, indispensable para el proceso de traducción. El gen HHIPL2 codifica para una proteína de 724 aminoácidos que no ha sido caracterizada a nivel funcional.

- El biomarcador RSPH1 es compartido entre los *clusters* epigenómicos y transcriptómicos. Este gen codifica para una proteína de 309 aminoácidos que hace parte del axonema, es decir, de la estructura interna axial de los cilios y flagelos celulares, y que juega un rol en la motilidad celular.
- Los biomarcadores CAPN7, ZG16B, CLTC y SON son compartidos entre los *clusters* epigenómicos y los *clusters* de la representación integrada G. Los genes CLTC y SON, asociados con la división celular y la regulación negativa del proceso de apoptosis, respectivamente, fueron seleccionados para un análisis más profundo.
- Los biomarcadores USP1, STRN, NCKAP1, ZBTB11, GMPS, DNAJB14, NAA15, UBLCP1, PHTF2, AP3M1, CCNT1, CAND1, EEA1, SCYL2, ZNF770 ITCH, GABPA, MED14, FBXO11, SMC3, NEDD1 y PTPN11 son compartidos entre los *clusters* transcriptómicos y los *clusters* de la representación integrada G. Los genes FBXO11, SMC3, NEDD1 y PTPN11 asociados con la regulación de la apoptosis para FBXO11, la división celular para SMC3 y NEDD1, y el punto de control de daño del ADN para PTPN11, también fueron extraídos para un análisis más completo.

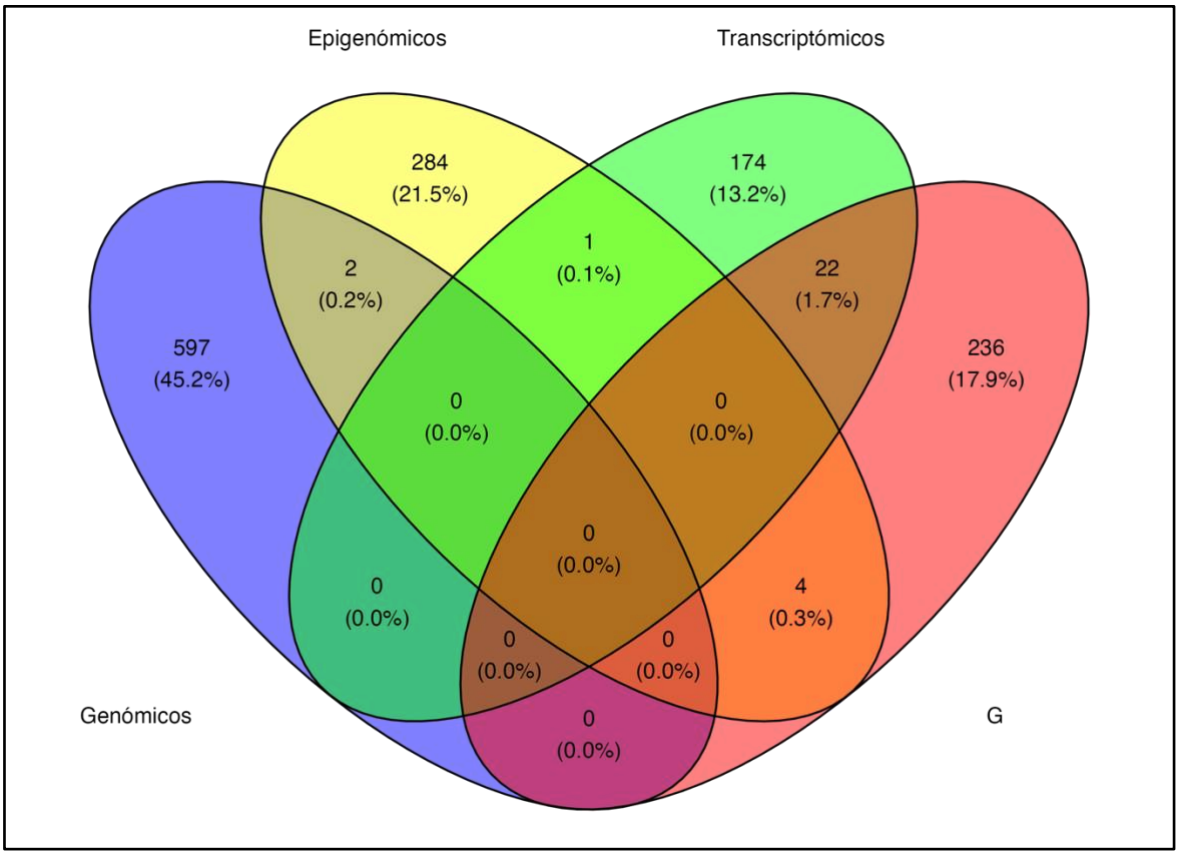


Figura 65: Comparación de los biomarcadores obtenidos mediante el análisis de las diferentes representaciones (perfil genómico, epigenómico, transcriptómico y representación integrada G). Las intersecciones en el diagrama de Venn corresponden a los biomarcadores compartidos entre varios conjuntos de datos.

Biomarcadores comunes entre el análisis de los datos epigenómicos y de la representación integrada G

CLTC es un biomarcador tanto del *cluster* epigenómico 3 (cáncer de endometrio) como de los *clusters* 1 (3 pacientes con cáncer de mama y 325 pacientes con cáncer de endometrio) y 2 (cáncer de mama) de la representación integrada G. Por la composición de los *clusters* mencionados, se puede afirmar que CLTC es un biomarcador de ciertos pacientes con cáncer de mama o cáncer de endometrio. CLTC codifica para la cadena pesada 1 de la clatrina, una proteína de 1675 aminoácidos. La clatrina se encuentra en la cara citoplasmática de los orgánulos intracelulares los cuales son involucrados en el tráfico intracelular de receptores y también en la endocitosis de macromoléculas. Importantemente, el rol de la clatrina en el proceso de la organización del huso mitótico, y, por lo tanto, en la división celular ha sido reportado (Royle, 2012).

SON es un biomarcador tanto del *cluster* epigenómico 1 (cáncer de mama) como del *cluster* 1 (3 pacientes con cáncer de mama y 325 pacientes con cáncer de endometrio) de la representación integrada G. Este gen parece ser un biomarcador de cáncer de mama y de cáncer de endometrio. El gen SON codifica para una proteína de 2426 aminoácidos que actúa como cofactor en el empalme o *splicing* de ARN. Este proceso puede ser calificado de maduración del pre-ARN mensajero en ARN mensajero mediante la eliminación de los intrones del transcrito primario y la unión de los exones (regiones codificantes). Un estudio identificó la proteína SON como una proteína inhibidora de la apoptosis (Greenhalf et al., 1999).

Tanto CLTC como SON nunca han sido asociados con cáncer de mama o cáncer de endometrio. El análisis comparativo de la metilación de las regiones promotoras de CLTC y SON entre los *clusters* epigenómicos (ver Fig. 66) y entre los *clusters* de la representación integrada G (ver Fig. 67) reveló diferencias significativas entre *clusters*. Sin embargo, los niveles de metilación de las regiones promotoras de estos dos genes son extremadamente bajos con una mediana igual a 0.038 para CLTC y una mediana igual a 0.062 para SON (ver Fig. 66-67). Esto hace la comparación entre *clusters* inútil. Para concluir, la ausencia de metilación de las regiones promotoras de los genes CLTC y SON es una característica de ciertos pacientes con cáncer de mama y cáncer de endometrio.

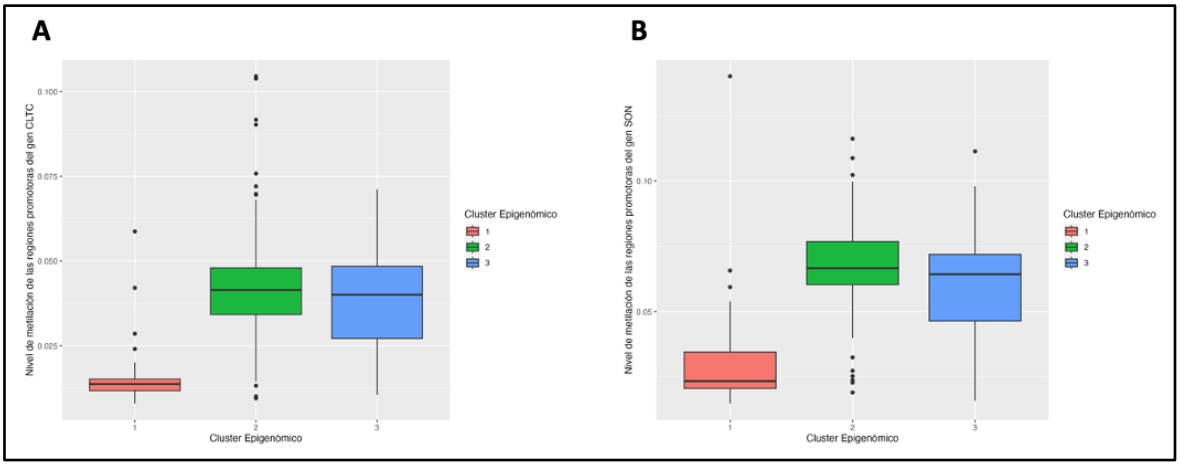


Figura 66: Comparación del nivel de metilación de las regiones promotoras de los biomarcadores CLTC y SON en el tumor de los pacientes según su *cluster* epigenómico asignado. Nivel de metilación de las regiones promotoras de CLTC (A) y SON (B).

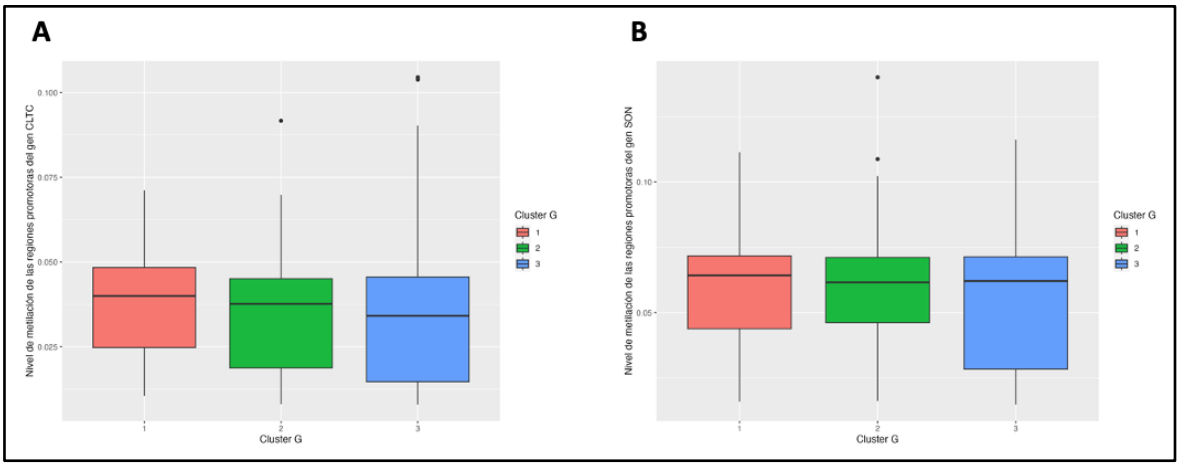


Figura 67: Comparación del nivel de metilación de las regiones promotoras de los biomarcadores CLTC y SON en el tumor de los pacientes según su *cluster* de la representación G asignado. Nivel de metilación de las regiones promotoras de CLTC (A) y SON (B).

Biomarcadores comunes entre el análisis de los datos transcriptómicos y de la representación integrada G

FBXO11 es un biomarcador tanto del *cluster* transcriptómico 2 (pacientes con cáncer de endometrio) como de los tres *clusters* de la representación integrada G. Por lo tanto, todo indica que FBXO11 es un biomarcador compartido entre cáncer de mama y cáncer de endometrio. FBXO11 codifica para una proteína de 927 aminoácidos que pertenece a la familia de las proteínas con caja F. Estas proteínas constituyen una de las cuatro subunidades del complejo de ubiquitinación SCF. SCF es un intermediario en la ubiquitinación de las proteínas marcadas para ser degradadas en el proteasoma. Algunos blancos de FBXO11 para la ubiquitinación, y en consecuencia, la degradación fueron identificados: BCL6 y Snail (Duan et al., 2012; Jin et al., 2015). Por un lado, BCL6 es el producto de un protooncogén implicado en los linfomas de células B así que su degradación es un evento favorable para evitar tal patología. Por otro lado, FBXO11 induce la degradación de las proteínas de la familia Snail y así inhibe la transición epitelio-mesénquima oponiéndose a la progresión tumoral. Sin embargo, un efecto adverso de FBXO11 ha sido reportado en pacientes con vitíligo: la expresión de este gen promueve la proliferación celular y reduce la apoptosis de los melanocitos (Xu, 2010). Dado que la proteína codificada por FBXO11 tiene varios blancos, la sobreexpresión de este gen puede tener efectos antagónicos. En el contexto del cáncer de mama, fue establecido que FBXO11 impulsa específicamente la formación de tumores a través de la inhibición de la vía de señalización p53/p21 (Bagger et al., 2018). No se ha reportado aún ninguna

asociación entre la desregulación de la expresión del gen FBXO11 y el cáncer de endometrio.

SMC3 es un biomarcador tanto del *cluster* transcriptómico 1 (pacientes con cáncer de mama) como del *cluster* 2 (pacientes con cáncer de mama) de la representación integrada G. SMC3 puede ser calificado de biomarcador de los pacientes con cáncer de mama. Este gen codifica para una proteína de 1169 aminoácidos requerida para la condensación y partición cromosómica durante el proceso de división celular. En efecto, antes de que las células se dividan, copian sus cromosomas a través de la etapa de replicación del ADN. El ADN copiado de cada cromosoma se organiza en dos estructuras idénticas, llamadas cromátidas hermanas, que se unen entre sí durante las primeras etapas de la división celular. La proteína codificada por SMC3 es parte de un grupo de proteínas llamado complejo de cohesina que mantiene unidas a las cromátidas hermanas. Por lo tanto, esta proteína tiene un rol importante en la estabilización de la información genética durante la división celular. De manera interesante, niveles elevados de subunidades de cohesina se correlacionan con un mal pronóstico y resistencia a las quimioterapias, radioterapias y terapias con hormonas en el contexto del cáncer de mama. Esto podría ser debido a una regulación transcripcional positiva de otros genes como Runx1, Runx3 y Myc. Además, la cohesina se une al receptor de estrógenos ERα en las células de cáncer de mama, lo que sugiere que puede estar involucrada en la transcripción de genes de respuesta a los estrógenos (Rhodes et al., 2011).

NEDD1 es un biomarcador tanto del *cluster* transcriptómico 1 (pacientes con cáncer de mama) como del *cluster* 3 (238 pacientes con cáncer de mama y 23 pacientes con cáncer de endometrio) de la representación integrada G. NEDD1 parece ser un biomarcador del cáncer de mama esencialmente. Codifica para una proteína de 660 aminoácidos requerida para la progresión de la mitosis. La proteína codificada por NEDD1 se ubica en el centrosoma y promueve la nucleación de microtúbulos y la organización del huso mitótico. En efecto, se une al complejo del anillo de γ-tubulina y lo dirige al centrosoma y al huso mitótico. Por lo tanto, la extinción de NEDD1 provoca la pérdida del complejo del anillo de γ-tubulina del centrosoma, y, como resultado, la falla de la nucleación de microtúbulos y del ensamblaje del huso mitótico. Se traduce en una detención del ciclo celular y en una inhibición de la proliferación celular (Manning & Kumar, 2007). Al contrario, una sobre expresión de este gen permitiría mejorar la eficiencia del proceso de división

celular y favorecer la proliferación. De hecho, el gen NEDD1 fue calificado de potencial blanco farmacológico para inducir la detención del ciclo celular en la línea celular inmortalizada MDA-MB-231 de adenocarcinoma de mama (Tillement et al., 2009).

PTPN11 es un biomarcador tanto de los dos *clusters* transcriptómicos como del *cluster* 2 (cáncer de mama) de la representación integrada G. En consecuencia, PTPN11 podría ser un biomarcador de ambos tipos de cáncer. La proteína codifica por PTPN11 mide 593 aminoacidos y es una proteína tirosina fosfatasa que, como su nombre lo indica, elimina los grupos fosfato de los residuos de tirosina fosforilados. Esta proteína actúa en la transducción de señales desde la membrana celular, y más específicamente desde los receptores tirosina-cinasas activados, hasta el núcleo, mediante una regulación positiva de la vía de señalización RAS/RAF/MAPK. Por lo tanto, la proteína codifica por PTPN11 favorece el crecimiento y la proliferación celular en un escenario de estimulación. En el contexto del síndrome de Noonan, el cual se caracteriza por afectaciones del desarrollo, fueron exhibidas mutaciones de ganancia de función para el gen PTPN11 que conducen a una transformación epitelio-mesénquima y a una proliferación celular excesiva en el corazón durante la valvulogénesis (Tartaglia et al., 2002). En el contexto del cáncer de mama, la proteína codificada por PTPN11 fue calificada de promotor del mantenimiento y de la progresión de células iniciadoras de tumores a través de la activación de factores de transcripción clave como v-Myc, c-Myc y ZEB1 y de un bucle de retroalimentación positiva (Aceto et al., 2012).

El estudio comparativo de la expresión génica de FBXO11, SMC3, NEDD1 y PTPN11 entre los *clusters* transcriptómicos reveló diferencias significativas entre los pacientes con cáncer de mama y los pacientes con cáncer de endometrio (*p-value* = 2.5e-52 para FBXO11, 7.6e-78 para SMC3, 2.2e-76 para NEDD1 y 5.8e-53 para PTPN11). Para los 4 genes, una mayor expresión fue medida en los pacientes con cáncer de mama con una mediana igual a 5.2 para FBXO11, 6.5 para SMC3, 5.0 para NEDD1 y 7.1 para PTPN11 a comparación de los pacientes con cáncer de endometrio caracterizados por una mediana de expresión génica igual a 4.4 para FBXO11, 5.5 para SMC3, 4.1 para NEDD1 y 6.4 para PTPN11 (ver Fig. 68).

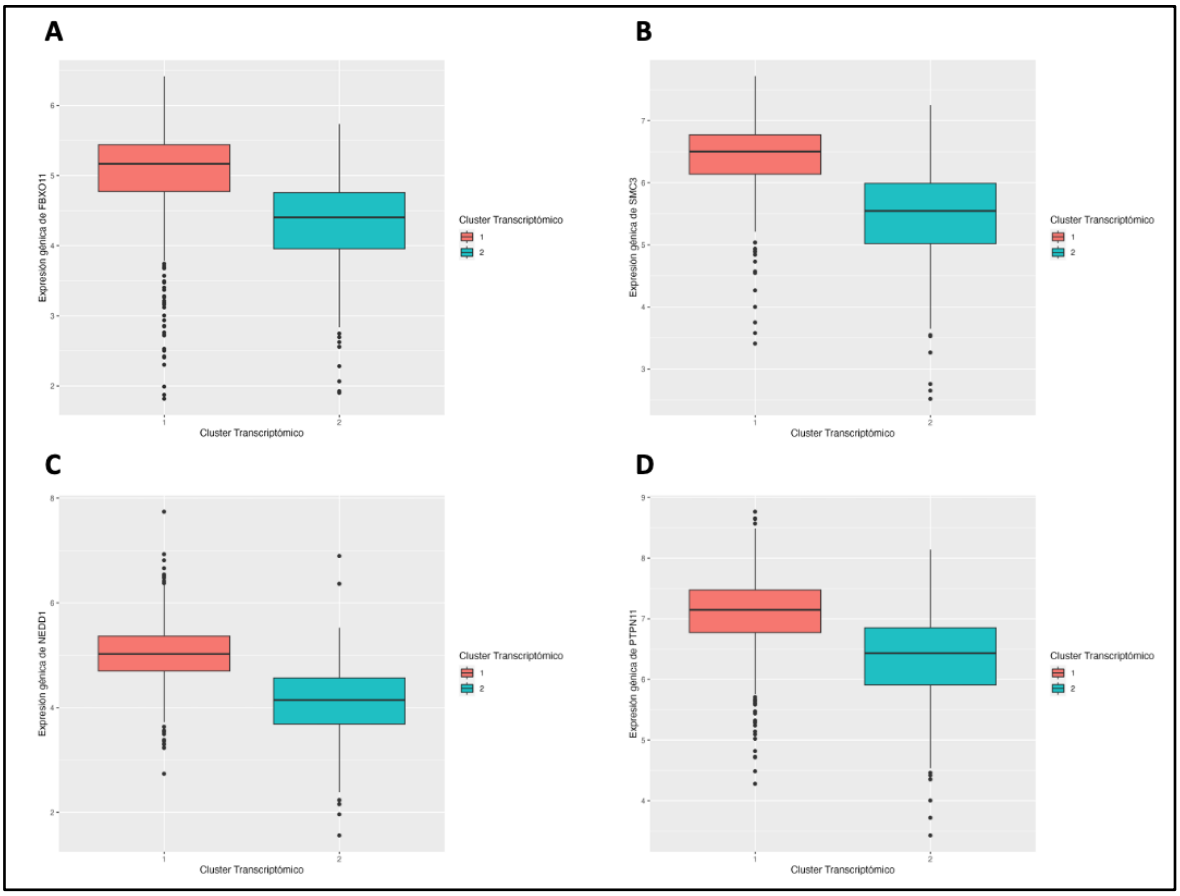


Figura 68: Comparación de la expresión génica de los biomarcadores FBXO11, SMC3, NEDD1 y PTPN11 en el tumor de los pacientes según su *cluster* transcriptómico asignado. Expresión génica de FBXO11 (A), SMC3 (B), NEDD1 (C) y PTPN11 (D).

El estudio comparativo de la expresión génica de FBXO11, SMC3, NEDD1 y PTPN11 entre los *clusters* de la representación integrada G reveló diferencias significativas entre *clusters* (*p-value* = 1.9e-50 para FBXO11, 1.8e-73 para SMC3, 1.7e-73 para NEDD1 y 5.2e-52 para PTPN11). Para los genes SMC3, NEDD1 y PTPN11, la prueba estadística *post hoc* arrojó diferencias significativas entre los tres *clusters* mientras que para el gen FOXB11, las diferencias significativas se encuentran entre los *clusters* 1 y 2 (*p-value* = 4.3e-43) y entre los *clusters* 1 y 3 (*p-value* = 2.1e-32) pero no hay una diferencia significativa entre los *clusters* 2 y 3 (*p-value* = 0.36). La expresión génica de FOXB11 es mayor en los *clusters* 2 y 3 – los cuales presentan una mayoría de pacientes con cáncer de mama: 562 de los 565 de la cohorte – con una mediana igual a 5.2 y 5.1, respectivamente, en comparación con el *cluster* 1 con una mediana igual a 4.4. Para los genes SMC3, NEDD1 y PTPN11, la expresión es menor en el *cluster* 1, intermedia en el *cluster* 2 y mayor en el *cluster* 3. Sin embargo, la diferencia de expresión es mayor entre los *clusters* 1 y 2 que entre los *clusters* 2 y 3 (ver Fig. 69).

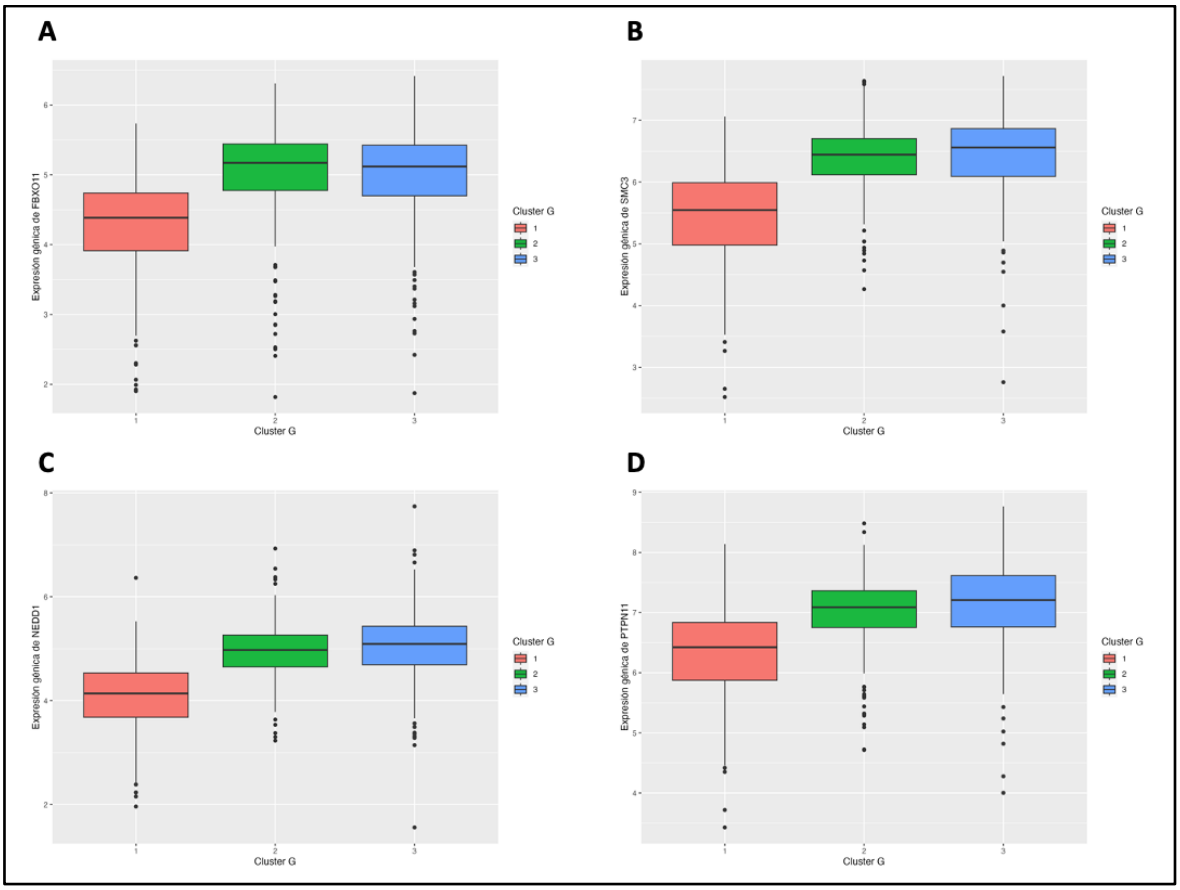


Figura 69: Comparación de la expresión génica de los biomarcadores FBXO11, SMC3, NEDD1 y PTPN11 en el tumor de los pacientes según su *cluster* de la representación integrada G asignado. Expresión génica de FBXO11 (A), SMC3 (B), NEDD1 (C) y PTPN11 (D).

Conclusiones y trabajo futuro

El cáncer de mama y el cáncer de endometrio son el primero y el cuarto cáncer más frecuentes en mujeres, respectivamente. Son enfermedades complejas con varios subtipos moleculares e histológicos. Uno de los factores de riesgo más reportado de estos dos tipos de cáncer es el desequilibrio en el balance estrógenos/progesterona. En efecto, los cánceres de mama luminales A y B se caracterizan por desregulaciones hormonales (balance estrógenos/progestágenos) asociadas con la proliferación excesiva de las células endoteliales de los conductos de leche o de las unidades lobulillares de las mamas mientras que la mayoría de los cánceres de endometrio tipo I surgen en un contexto de estimulación estrogénica excesiva de la capa endometrial del útero. Por este motivo, se planteó la hipótesis que algunos subtipos de cáncer de mama comparten biomarcadores dada su semejanza en cuanto a mecanismos celulares, moleculares, y, factores de riesgo. El objetivo general de este proyecto de investigación era entonces diseñar un modelo de aprendizaje automático de integración de datos genómicos, epigenómicos y transcriptómicos de pacientes con cáncer de endometrio y con cáncer de mama con el fin de identificar biomarcadores compartidos.

Un proceso riguroso de selección de los pacientes fue aplicado a una cohorte inicial de 1606 pacientes cuyos datos provienen de la plataforma TCGA con el objetivo de conservar los pacientes susceptibles de compartir biomarcadores además de minimizar la heterogeneidad a nivel clínico e inmunológico. Primero, los pacientes mujeres con carcinoma ductal infiltrante, carcinoma lobulillar infiltrante, o adenocarcinoma endometrioide fueron conservados. Luego, la selección a nivel hormonal conservó los individuos positivos para la expresión de los receptores ERα y PR medida por inmunohistoquímica y con una expresión intermedia o alta de los transcritos ESR1 y PGR que codifican para ERα y PR, respectivamente. Finalmente, los individuos con una alta infiltración inmunológica en su tumor fueron eliminados para reducir el sesgo constituido por la alta cantidad de leucocitos, los cuales presentan perfiles de expresión génica y de

metilación del ADN diferentes de los perfiles de las células cancerígenas. Los filtros de selección aplicados a la cohorte inicial dejaron una cohorte de 913 pacientes: 565 con cáncer de mama y 348 con cáncer de endometrio.

El análisis de los perfiles ómicos de los pacientes seleccionados reveló diferencias significativas según el tipo de datos ómicos considerado. Por un lado, el *clustering* de los pacientes según su perfil transcriptómico o epigenómico resultó en una separación clara de los pacientes con cáncer de mama y de los pacientes con cáncer de endometrio debido a la especificidad de las firmas transcriptómicas y epigenómicas. El análisis de los dos *clusters* transcriptómicos obtenidos mostró que los pacientes con cáncer de mama tienen tumores más avanzados que los pacientes con cáncer de endometrio lo que se traduce en una mayor tasa de mortalidad. Los pacientes con cáncer de mama fueron separados en dos *clusters* (*clusters* 1 y 2) en el caso del estudio epigenómico. El *cluster* epigenómico 2 tiene una metilación promedio de las regiones promotoras mayor que los demás *clusters* epigenómicos. Además, este *cluster* exhibe una edad mediana menor, asociada con más pacientes en estados premenopáusico y más ganglios linfáticos positivos para micrometástasis o residuos tumorales, lo que podría explicar la cantidad mayor de pacientes clasificados en etapa III, una etapa tardía del cáncer. Así que el perfil epigenómico de los pacientes con cáncer de mama está asociado con parámetros clínicos de avance y gravedad de la enfermedad. Por otro lado, el *clustering* de los pacientes según su perfil genómico dio seis *clusters* mixtos con un número de copias promedio de los genes mucho mayor en el *cluster* 6, una tasa de mortalidad muy alta en los *clusters* 2 y 4 y un fuerte desequilibrio en el balance histológico: carcinoma ductal infiltrante/carcinoma lobulillar infiltrante (con un porcentaje mayor de pacientes con un carcinoma ductal infiltrante) en los *clusters* 4 y 6. La asociación de estos diferentes resultados no es evidente.

El análisis correlacional entre los distintos tipos de datos ómicos reveló dos comportamientos totalmente diferentes para la asociación entre los datos genómicos y transcriptómicos y la asociación entre los datos epigenómicos y transcriptómicos. La correlación positiva obtenida entre el número de copias de los genes y su expresión génica está en adecuación con la idea generalmente aceptada que, entre más copias haya de un gen, mayor es su expresión génica. Los resultados obtenidos para la asociación entre los datos epigenómicos y transcriptómicos contrastaron con lo visto previamente: la mayoría

de los 12594 genes estudiados no presenta ninguna correlación entre la metilación de sus regiones promotoras y su expresión génica. Es importante que haya un esfuerzo de la comunidad científica para mejorar el entendimiento del impacto de la hipometilación o de la hipermetilación del ADN en la expresión génica en el contexto del cáncer.

Después del análisis de los datos ómicos por separado y de la búsqueda de relaciones matemáticas entre los distintos tipos de datos, los datos multi-ómicos fueron integrados mediante el uso del algoritmo de aprendizaje multi-vista DGCCA para descubrir nuevos biomarcadores que no aparecían en el análisis de datos ómicos de un solo tipo. Este algoritmo de aprendizaje no supervisado lleva a cabo el aprendizaje de un mapeo no lineal para cada vista que maximiza la correlación entre las vistas. Luego, una representación integrada llamada G es construida. El estudio de la representación integrada G resultó en la creación de tres *clusters*: el *cluster* 1 está compuesto por 3 pacientes con cáncer de mama y 325 pacientes con cáncer de endometrio, el *cluster* 2 está compuesto exclusivamente por pacientes con cáncer de mama mientras que el *cluster* 3 está compuesto por 238 pacientes con cáncer de mama y 23 pacientes con cáncer de endometrio.

Diferencias significativas del promedio de metilación del ADN y del promedio del número de copias de los genes fueron identificadas entres *clusters*: el *cluster* 2 presenta una metilación promedio de las regiones promotoras mayor que el *cluster* 1 y el número promedio de copias de los genes es mayor en el *cluster* 3 a comparación de los demás *clusters* de la representación integrada G. Así que el número de copias no está asociado con la metilación de las regiones promotoras. El análisis de las características clínicas de los pacientes según su *cluster* reveló que el *cluster* 2 está compuesto exclusivamente por pacientes con cáncer de mama con un balance relativamente equilibrado entre carcinoma ductal y carcinoma lobulillar infiltrante. Es el *cluster* que cuenta más cánceres tempranos. En conclusión, el *clustering* de los pacientes según la representación integrada G permitió establecer *clusters* diferenciables a nivel de la metilación del ADN y de la expresión génica pero poco diferenciables a nivel clínico.

Para cada *cluster* establecido (dos *clusters* transcriptómicos, tres *clusters* epigenómicos, seis *clusters* genómicos y tres *clusters* de la representación integrada G), 100 biomarcadores fueron extraídos y analizados a través del enriquecimiento de términos

GO. Aunque para ciertos *clusters* los términos GO asociados con los biomarcadores no parecen estar asociados con una patología de tipo cáncer, ciertos biomarcadores del *cluster* transcriptómico 1, de los *clusters* epigenómicos 1 y 2, de los *clusters* genómicos 2 y 4, y, de los *clusters* de la representación integrada G 1, 2, y 3 están asociados con términos GO interesantes clasificados en 4 categorías: regulación del ciclo celular, apoptosis, angiogénesis y envejecimiento celular. Los *clusters* de la representación integrada G se caracterizan por tener más biomarcadores asociados con términos GO de interés. Muchos de ellos están involucrados en el balance proliferación celular/apoptosis. Pocos biomarcadores están compartidos entre los análisis de datos ómicos de un solo tipo y el análisis integrativo. Como se ha reportado también en la literatura, la integración de datos multi-ómicos permite descubrir nuevos biomarcadores.

Los biomarcadores compartidos entre el análisis de datos epigenómicos o transcriptómicos y el análisis de la representación integrada G fueron estudiados de manera individual. Los genes CLTC y SON, asociados con la división celular y la regulación negativa del proceso de apoptosis, respectivamente, son biomarcadores compartidos entre cáncer de mama y cáncer de endometrio. Fueron encontrados tanto en el análisis epigenómico como en el estudio integrativo; presentan una ausencia de metilación en sus regiones promotoras en ambos contextos patológicos. Los biomarcadores FBXO11 y PTPN11 están asociados con la regulación de la apoptosis y el punto de control de daño al ADN, respectivamente, también son compartidos entre cáncer de mama y cáncer de endometrio. Provienen del análisis transcriptómico y de la integración de los datos ómicos. De la misma manera, los genes SMC3 y NEDD1, asociados con el proceso de organización del huso mitotico, y, por lo tanto, involucrados en el proceso de división celular, son biomarcadores de los pacientes con cáncer de mama esencialmente. Los biomarcadores FBXO11, PTPN11, SMC3 y NEDD1 presentan niveles de expresión muy altos en ambos contextos patológicos.

Distintos biomarcadores de cáncer de mama y de cáncer de endometrio fueron descubiertos y descritos en esta investigación. Es importante llevar a cabo una validación experimental para determinar la importancia de estos genes en la carcinogénesis, la angiogénesis y la invasión tumoral. Esta verificación se puede hacer de varias formas. Se puede empezar por silenciar los genes identificados como FBXO11, PTPN11, SMC3 y NEDD1 mediante ARN interferencia en líneas celulares inmortalizadas como MCF-7 para

el cáncer de mama o Ishikawa para el cáncer de endometrio. Luego se puede realizar varios ensayos en estas células desde estudios de proliferación, apoptosis e invasión tumoral hasta estudios transcriptómicos para determinar cuáles vías de señalización fueron impactadas por el silenciamiento de los biomarcadores.

Asimismo, se puede modelar el cáncer de mama o el cáncer de endometrio de manera más elaborada a través de modelos murinos u organoides que recapitulan los determinantes estructurales y moleculares de la enfermedad. Por un lado, la técnica PDX (*patient-derived xenograft*) para la elaboración de modelos murinos es de especial interés: corresponde a la implantación de un trozo de tumor recién aislado de un paciente en ratones inmunocomprometidos. Múltiples estudios han demostrado que los modelos PDX mantienen la heterogeneidad histológica, molecular y funcional original presente en los tumores de los pacientes. Por otro lado, los organoides son versiones simplificadas de órganos o tumores reproduciendo la anatomía y la composición fisiológica real. Los organoides son generalmente derivados de células madre embrionarias o células madre pluripotentes, pero también pueden ser establecidos a partir de células tumorales humanas. Permiten reducir el uso de animales de laboratorio. Organoides de cáncer de mama y de cáncer de endometrio ya han sido desarrollados. Dichos organoides lograron reproducir la arquitectura de los tumores originales, pudieron ser cultivados durante períodos prolongados (hasta 5 meses) y mostraron estabilidad genética y molecular (Mohan et al., 2021; Sakamoto et al., 2015; Van Nyen et al., 2018). Se podrían implementar las técnicas genéticas de inactivación génica (*knockout*) o de reducción de expresión génica (*knockdown*) en modelos murinos u organoides para evaluar la importancia de los biomarcadores identificados en un contexto más representativo de la realidad fisiopatológica.

Se podría afinar el análisis trabajando con datos ómicos de células únicas o, aún mejor, datos ómicos de células únicas con su ubicación en las tres dimensiones del tumor. A pesar del esfuerzo realizado en este proyecto para disminuir la heterogeneidad a nivel celular con la eliminación de los pacientes con una fuerte infiltración inmunológica, es probable que los tumores estudiados presenten bastante diversidad a nivel celular con células normales, células cancerígenas, leucocitos, células mesenquimales estromales y otros tipos celulares minoritarios. Este nivel de resolución permitiría eliminar totalmente el sesgo de la heterogeneidad tumoral. Además, estos análisis constituyen una gran

oportunidad para mejorar el conocimiento acerca de la interacción entre las células cancerígenas y el microambiente tumoral. La meta fundamental de las ciencias ómicas espaciales en el ámbito de la oncología es la obtención de una representación multi-ómica espacial del tumor con una resolución de célula única (X. Li & Wang, 2021).

Importantemente, se pueden emitir ciertas criticas acerca de los datos usados y del método de identificación de los biomarcadores, sobre todo en la etapa de integración de los datos multi-ómicos con el modelo DGCCA. En efecto, la metilación de las regiones promotoras de los genes no mostró ninguna correlación con la expresión génica, es decir que no se pueden relacionar los datos epigenómicos con los datos transcriptómicos. A lo mejor la inclusión de los datos epigenómicos en la integración de los datos multi-ómicos no es tan coherente dado que el modelo de aprendizaje multi-vista DGCCA lleva a cabo el aprendizaje de un mapeo no lineal para cada vista que maximiza la correlación entre las vistas. Una alternativa sería dejar los datos epigenómicos a un lado e implementar el algoritmo DCCA (*Deep Canonical Correlation Analysis*), una variante de DGCCA, para llevar a cabo la integración de los datos genómicos y transcriptómicos, los cuales presentan una correlación positiva general. DCCA es un método para aprender transformaciones no lineales complejas de dos vistas de datos de modo que las representaciones resultantes estén altamente correlacionadas linealmente. Los parámetros de ambas transformaciones se aprenden conjuntamente para maximizar la correlación total (Andrew et al., 2013). Otra ventaja de solo incluir los datos genómicos y transcriptómicos es que estos datos presentan más variables compartidas (57444 genes) a comparación del número de variables compartidas entre los tres tipos de datos ómicos (12594 genes).

En cuanto al método de identificación de los biomarcadores de los *clusters* de la representación integrada G, podría ser calificado de satisfactorio dado que varios de ellos están asociados con procesos desregulados en el ámbito del cáncer. Sin embargo, cabe resaltar que, desde el conjunto inicial que contiene el número de copias, la metilación de las regiones promotoras y la expresión génica de los 12594 genes compartidos hasta la obtención de las variables sintéticas UMAP1 y UMAP2, ocurren cuatro transformaciones sucesivas de los datos: el aprendizaje del mapeo no lineal para cada vista mediante la red neuronal profunda, la transformación lineal de las salidas de las redes neuronales de cada vista, la creación de la representación integrada G y la reducción de la dimensionalidad

mediante el algoritmo UMAP. Esto podría constituir un sesgo en el cálculo de la información mutua entre las variables genómicas, epigenómicas y transcriptómicas iniciales y las variables sintéticas UMAP1 y UMAP2, y, por consiguiente, en la identificación de los biomarcadores. En el caso del estudio de los datos ómicos de un solo tipo, este problema no se presenta ya que las variables sintéticas UMAP1 y UMAP2 fueron establecidas directamente a partir de los datos iniciales. Sería interesante usar otro método para la extracción de los biomarcadores en el caso de la implementación del algoritmo DGCCA. Se puede pensar en la explicación aditiva de Shapley (SHAP) para una extracción de los biomarcadores más rigurosa. Los valores SHAP corresponden a las contribuciones de cada variable a la diferencia entre la predicción real y la predicción esperada del modelo. Los valores SHAP muestran cuanto contribuye cada variable, ya sea positivamente o negativamente, a las predicciones individuales. Sin embargo, habría que repensar el problema bioinformático dado que el método SHAP solo puede ser usado en un contexto de aprendizaje supervisado (Hwang et al., 2022).

Para concluir sobre este trabajo de investigación, aparte de las conclusiones biológicas que fueron reportadas anteriormente en esta discusión, se inscribe como una prueba de concepto de integración de distintos tipos de datos provenientes de diferentes contextos patológicos en el campo de la oncología. La estrategia de integración multi-ómica permitió descubrir biomarcadores que no aparecen en el análisis de datos ómicos de un solo tipo. La hipótesis de la existencia de biomarcadores compartidos entre cáncer de mama y cáncer de endometrio se reveló cierta; varios biomarcadores de ambos tipos de cáncer con un interés a nivel funcional fueron exhibidos. Para trabajos futuros, es importante resaltar la necesidad de proveer una selección rigurosa de los pacientes y un análisis exploratorio de los datos ómicos por separado antes de llevar a cabo la integración multi-ómica.

Bibliografía

Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., & Saeys, Y. (2010). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, *26*(3), 392–398. https://doi.org/10.1093/bioinformatics/btp630

Aceto, N., Sausgruber, N., Brinkhaus, H., Gaidatzis, D., Martiny-Baron, G., Mazzarol, G., Confalonieri, S., Quarto, M., Hu, G., Balwierz, P. J., Pachkov, M., Elledge, S. J., Van Nimwegen, E., Stadler, M. B., & Bentires-Alj, M. (2012). Tyrosine phosphatase SHP2 promotes breast cancer progression and maintains tumor-initiating cells via activation of key transcription factors and a positive feedback signaling loop. *Nature Medicine*, *18*(4), 529–537. https://doi.org/10.1038/nm.2645

Andrew, G., Arora, R., Bilmes, J., & Livescu, K. (2013). Deep Canonical Correlation Analysis. In S. Dasgupta & D. McAllester (Eds.), *Proceedings of the 30th International Conference on Machine Learning* (Vol. 28, Issue 3, pp. 1247–1255). PMLR. https://proceedings.mlr.press/v28/andrew13.html

Bagger, S. O., Hopkinson, B. M., Pandey, D. P., Bak, M., Brydholm, A. V., Villadsen, R., Helin, K., Rønnov-Jessen, L., Petersen, O. W., & Kim, J. (2018). Aggressiveness of non-EMT breast cancer cells relies on FBXO11 activity. *Molecular Cancer*, *17*(1), 171. https://doi.org/10.1186/s12943-018-0918-6

Baltimore, D. (1970). Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of RNA Tumour Viruses. *Nature*, *226*(5252), 1209–1211. https://doi.org/10.1038/2261209a0

Banno, K., Kisu, I., Yanokura, M., Tsuji, K., Masuda, K., Ueki, A., Kobayashi, Y., Yamagami, W., Nomura, H., Tominaga, E., Susumu, N., & Aoki, D. (2012). Biomarkers in endometrial cancer: Possible clinical applications (Review). *Oncology Letters*, *3*(6), 1175–1180. https://doi.org/10.3892/ol.2012.654

Benton, A., Khayrallah, H., Gujral, B., Reisinger, D. A., Zhang, S., & Arora, R. (2019). Deep Generalized Canonical Correlation Analysis. *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, 1–6. https://doi.org/10.18653/v1/W19-4301

Berger, A. C., Korkut, A., Kanchi, R. S., Hegde, A. M., Lenoir, W., Liu, W., Liu, Y., Fan, H., Shen, H., Ravikumar, V., Rao, A., Schultz, A., Li, X., Sumazin, P., Williams, C., Mestdagh, P., Gunaratne, P. H., Yau, C., Bowlby, R., … Mariamidze, A. (2018). A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell*, *33*(4), 690-705.e9. https://doi.org/10.1016/j.ccell.2018.03.014

Bian, J., Xu, Y., Wu, F., Pan, Q., & Liu, Y. (2020). Identification of a five-gene signature for predicting the progression and prognosis of stage I endometrial carcinoma. *Oncology Letters*, *20*(3), 2396–2410. https://doi.org/10.3892/ol.2020.11798

Bray, F., Ferlay, J., Laversanne, M., Brewster, D. H., Gombe Mbalawa, C., Kohler, B., Piñeros, M., Steliarova-Foucher, E., Swaminathan, R., Antoni, S., Soerjomataram, I., & Forman, D. (2015). Cancer Incidence in Five Continents: Inclusion criteria, highlights from Volume X and the global status of cancer registration: Cancer Incidence in Five Continents Volume X. *International Journal of Cancer*, *137*(9), 2060–2071. https://doi.org/10.1002/ijc.29670

Breast Cancer Overview: Causes, Symptoms, Signs, Stages & Types. (n.d.). Cleveland Clinic. Retrieved March 25, 2022, from https://my.clevelandclinic.org/health/diseases/3986-breast-cancer

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. https://doi.org/10.1007/s10549-017-4325-2

Cancer of the Endometrium—Cancer Stat Facts. (n.d.). SEER. Retrieved March 25, 2022, from https://seer.cancer.gov/statfacts/html/corp.html

Carmeliet, P. (2005). VEGF as a Key Mediator of Angiogenesis in Cancer. *Oncology*, *69*(Suppl. 3), 4–10. https://doi.org/10.1159/000088478

Chatterjee, S., Gupta, D., Caputo, T. A., & Holcomb, K. (2016). Disparities in Gynecological Malignancies. *Frontiers in Oncology*, *6*. https://doi.org/10.3389/fonc.2016.00036

Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M., & Alizadeh, A. A. (2018). Profiling Tumor Infiltrating Immune Cells with CIBERSORT. In L. von Stechow (Ed.), *Cancer Systems Biology* (Vol. 1711, pp. 243–259). Springer New York. https://doi.org/10.1007/978-1-4939-7493-1_12

Colombo, N., Creutzberg, C., Amant, F., Bosse, T., González-Martín, A., Ledermann, J., Marth, C., Nout, R., Querleu, D., Mirza, M. R., Sessa, C., Abal, M., Altundag, O., Amant, F., van Leeuwenhoek, A., Banerjee, S., Bosse, T., Casado, A., de Agustín, L. C., … Zeimet, A. G. (2016). ESMO-ESGO-ESTRO Consensus Conference on Endometrial Cancer: Diagnosis, treatment and follow-up. *Annals of Oncology*, *27*(1), 16–41. https://doi.org/10.1093/annonc/mdv484

Crick, F. (1970). Central Dogma of Molecular Biology. *Nature*, *227*(5258), 561–563. https://doi.org/10.1038/227561a0

Danaei, G., Vander Hoorn, S., Lopez, A. D., Murray, C. J., & Ezzati, M. (2005). Causes of cancer in the world: Comparative risk assessment of nine behavioural and environmental risk factors. *The Lancet*, *366*(9499), 1784–1793. https://doi.org/10.1016/S0140-6736(05)67725-2

Das, T., Andrieux, G., Ahmed, M., & Chakraborty, S. (2020). Integration of Online Omics-Data Resources for Cancer Research. *Frontiers in Genetics*, *11*, 578345–578345. https://doi.org/10.3389/fgene.2020.578345

Duan, S., Cermak, L., Pagan, J. K., Rossi, M., Martinengo, C., Di Celle, P. F., Chapuy, B., Shipp, M., Chiarle, R., & Pagano, M. (2012). FBXO11 targets BCL6 for degradation and is inactivated in diffuse large B-cell lymphomas. *Nature*, *481*(7379), 90–93. https://doi.org/10.1038/nature10688

Eisenberg, E., & Levanon, E. Y. (2013). Human housekeeping genes, revisited. *Trends in Genetics*, *29*(10), 569–574. https://doi.org/10.1016/j.tig.2013.05.010

Female Breast Cancer—Cancer Stat Facts. (n.d.). Retrieved March 25, 2022, from https://seer.cancer.gov/statfacts/html/breast.html

Gao, Y., Lin, P., Lydon, J. P., & Li, Q. (2017). Conditional abrogation of transforming growth factor-β receptor 1 in PTEN-inactivated endometrium promotes endometrial cancer progression in mice: TGFBR1 in endometrial cancer progression. *The Journal of Pathology*, *243*(1), 89–99. https://doi.org/10.1002/path.4930

García Ayala, E., Cárdenas Mastrascusa, L., Sandoval Martínez, D., & Mayorga Anaya, H. (2010). HIPERPLASIA ENDOMETRIAL: ANÁLISIS DE SERIE DE CASOS DIAGNOSTICADOS EN BIOPSIA ENDOMETRIAL. *Revista Chilena de Obstetricia y Ginecología*, *75*(3). https://doi.org/10.4067/S0717-75262010000300002

Greenhalf, W., Lee, J., & Chaudhuri, B. (1999). A selection system for human apoptosis inhibitors using yeast. *Yeast (Chichester, England)*, *15*(13), 1307–1321. https://doi.org/10.1002/(SICI)1097-0061(19990930)15:13<1307::AID-YEA455>3.0.CO;2-3

GTEx Consortium. (2017). Genetic effects on gene expression across human tissues. *Nature*, *550*(7675), 204–213. https://doi.org/10.1038/nature24277

Gu, J., Wang, Z., Wang, B. O., & Ma, X. (2023). ImmuneScore of eight-gene signature predicts prognosis and survival in patients with endometrial cancer. *Frontiers in Oncology*, *13*, 1097015. https://doi.org/10.3389/fonc.2023.1097015

Harbeck, N., Penault-Llorca, F., Cortes, J., Gnant, M., Houssami, N., Poortmans, P., Ruddy, K., Tsang, J., & Cardoso, F. (2019). Breast cancer. *Nature Reviews Disease Primers*, *5*(1), 66. https://doi.org/10.1038/s41572-019-0111-2

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology*, *18*(1), 83. https://doi.org/10.1186/s13059-017-1215-1

Heo, H., Kim, J.-H., Lim, H. J., Kim, J.-H., Kim, M., Koh, J., Im, J.-Y., Kim, B.-K., Won, M., Park, J.-H., Shin, Y.-J., Yun, M. R., Cho, B. C., Kim, Y. S., Kim, S.-Y., & Kim, M. (2022). DNA methylome and single-cell transcriptome analyses reveal CDA as a potential druggable target for ALK inhibitor–resistant lung cancer therapy. *Experimental & Molecular Medicine*, *54*(8), 1236–1249. https://doi.org/10.1038/s12276-022-00836-7

Hess, R. A. (2003). Estrogen in the adult male reproductive tract: A review. *Reproductive Biology and Endocrinology*, *1*(1), 52. https://doi.org/10.1186/1477-7827-1-52

Huang, S., Murphy, L., & Xu, W. (2018). Genes and functions from breast cancer signatures. *BMC Cancer*, *18*(1), 473. https://doi.org/10.1186/s12885-018-4388-4

Huang, S., Pang, L., & Wei, C. (2021). Identification of a Four-Gene Signature With Prognostic Significance in Endometrial Cancer Using Weighted-Gene Correlation Network Analysis. *Frontiers in Genetics*, *12*, 678780. https://doi.org/10.3389/fgene.2021.678780

Huen, M. S. Y., Sy, S. M. H., & Chen, J. (2010). BRCA1 and its toolbox for the maintenance of genome integrity. *Nature Reviews Molecular Cell Biology*, *11*(2), 138–148. https://doi.org/10.1038/nrm2831

Hwang, J., Moon, S., & Lee, H. (2022). *SDGCCA: Supervised Deep Generalized Canonical Correlation Analysis for Multi-omics Integration*. https://doi.org/10.48550/ARXIV.2204.09045

Jin, Y., Shenoy, A. K., Doernberg, S., Chen, H., Luo, H., Shen, H., Lin, T., Tarrash, M., Cai, Q., Hu, X., Fiske, R., Chen, T., Wu, L., Mohammed, K. A., Rottiers, V., Lee, S. S., & Lu, J. (2015). FBXO11 promotes ubiquitination of the Snail family of transcription factors in cancer progression and epidermal development. *Cancer Letters*, *362*(1), 70–82. https://doi.org/10.1016/j.canlet.2015.03.037

Kotliar, D., Veres, A., Nagy, M. A., Tabrizi, S., Hodis, E., Melton, D. A., & Sabeti, P. C. (2019). Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *ELife*, *8*, e43803. https://doi.org/10.7554/eLife.43803

Li, P., Piao, Y., Shon, H. S., & Ryu, K. H. (2015). Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. *BMC Bioinformatics*, *16*(1), 347. https://doi.org/10.1186/s12859-015-0778-7

Li, X., & Wang, C.-Y. (2021). From bulk, single-cell to spatial RNA sequencing. *International Journal of Oral Science*, *13*(1), 36. https://doi.org/10.1038/s41368-021-00146-0

Makker, V., MacKay, H., Ray-Coquard, I., Levine, D. A., Westin, S. N., Aoki, D., & Oaknin, A. (2021). Endometrial cancer. *Nature Reviews Disease Primers*, *7*(1), 88. https://doi.org/10.1038/s41572-021-00324-8

Manning, J., & Kumar, S. (2007). NEDD1: Function in microtubule nucleation, spindle assembly and beyond. *The International Journal of Biochemistry & Cell Biology*, *39*(1), 7–11. https://doi.org/10.1016/j.biocel.2006.08.012

Martinez-Ledesma, E., Verhaak, R. G. W., & Treviño, V. (2015). Identification of a multi-cancer gene expression biomarker for cancer clinical outcomes using a network-based algorithm. *Scientific Reports*, *5*(1), 11966. https://doi.org/10.1038/srep11966

Mohan, S. C., Lee, T.-Y., Giuliano, A. E., & Cui, X. (2021). Current Status of Breast Organoid Models. *Frontiers in Bioengineering and Biotechnology*, *9*, 745943. https://doi.org/10.3389/fbioe.2021.745943

Mørch, L. S., Skovlund, C. W., Hannaford, P. C., Iversen, L., Fielding, S., & Lidegaard, Ø. (2017). Contemporary Hormonal Contraception and the Risk of Breast Cancer. *New*

England Journal of Medicine, *377*(23), 2228–2239. https://doi.org/10.1056/NEJMoa1700732

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M., & Alizadeh, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, *12*(5), 453–457. https://doi.org/10.1038/nmeth.3337

Nguyen, T., Tagett, R., Diaz, D., & Draghici, S. (2017). A novel approach for data integration and disease subtyping. *Genome Research*, *27*(12), 2025–2039. https://doi.org/10.1101/gr.215129.116

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L. B., Martin, S., Wedge, D. C., Van Loo, P., Ju, Y. S., Smid, M., Brinkman, A. B., Morganella, S., Aure, M. R., Lingjærde, O. C., Langerød, A., Ringnér, M., … Stratton, M. R. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, *534*(7605), 47–54. https://doi.org/10.1038/nature17676

Phuong, T. K., Thuan, L. D., Thao, D. T. P., & Thuy, L. H. A. (2015). DNA Hypermethylation Signatures for Detection of Breast Cancer in Vietnamese Population. In V. V. Toi & T. H. Lien Phuong (Eds.), *5th International Conference on Biomedical Engineering in Vietnam* (Vol. 46, pp. 219–222). Springer International Publishing. https://doi.org/10.1007/978-3-319-11776-8_53

Rappoport, N., & Shamir, R. (2018). Multi-omic and multi-view clustering algorithms: Review and cancer benchmark. *Nucleic Acids Research*, *46*(20), 10546–10562. https://doi.org/10.1093/nar/gky889

Razin, A., & Cedar, H. (1991). DNA methylation and gene expression. *Microbiological Reviews*, *55*(3), 451–458. https://doi.org/10.1128/mr.55.3.451-458.1991

Reel, P. S., Reel, S., Pearson, E., Trucco, E., & Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology Advances*, *49*, 107739. https://doi.org/10.1016/j.biotechadv.2021.107739

Rhodes, J. M., McEwan, M., & Horsfield, J. A. (2011). Gene Regulation by Cohesin in Cancer: Is the Ring an Unexpected Party to Proliferation? *Molecular Cancer Research*, *9*(12), 1587–1607. https://doi.org/10.1158/1541-7786.MCR-11-0382

Ring, K. L., Bruegl, A. S., Allen, B. A., Elkin, E. P., Singh, N., Hartman, A.-R., Daniels, M. S., & Broaddus, R. R. (2016). Germline multi-gene hereditary cancer panel testing in an unselected endometrial cancer cohort. *Modern Pathology*, *29*(11), 1381–1389. https://doi.org/10.1038/modpathol.2016.135

Rodriguez, A. C., Blanchard, Z., Maurer, K. A., & Gertz, J. (2019). Estrogen Signaling in Endometrial Cancer: A Key Oncogenic Pathway with Several Open Questions. *Hormones and Cancer*, *10*(2–3), 51–63. https://doi.org/10.1007/s12672-019-0358-9

Royle, S. J. (2012). The role of clathrin in mitotic spindle organisation. *Journal of Cell Science*, *125*(1), 19–28. https://doi.org/10.1242/jcs.094607

Ryan, N. A. J., Glaire, M. A., Blake, D., Cabrera-Dandy, M., Evans, D. G., & Crosbie, E. J. (2019). The proportion of endometrial cancers associated with Lynch syndrome: A systematic review of the literature and meta-analysis. *Genetics in Medicine*, *21*(10), 2167–2180. https://doi.org/10.1038/s41436-019-0536-8

Sakamoto, K., Schmidt, J. W., & Wagner, K.-U. (2015). Mouse Models of Breast Cancer. In R. Eferl & E. Casanova (Eds.), *Mouse Models of Cancer* (Vol. 1267, pp. 47–71). Springer New York. https://doi.org/10.1007/978-1-4939-2297-0_3

Shammas, M. A. (2011). Telomeres, lifestyle, cancer, and aging: *Current Opinion in Clinical Nutrition and Metabolic Care*, *14*(1), 28–34. https://doi.org/10.1097/MCO.0b013e32834121b1

Shao, X., Lv, N., Liao, J., Long, J., Xue, R., Ai, N., Xu, D., & Fan, X. (2019). Copy number variation is highly correlated with differential gene expression: A pan-cancer study. *BMC Medical Genetics*, *20*(1), 175. https://doi.org/10.1186/s12881-019-0909-5

Shen, F., Gao, Y., Ding, J., & Chen, Q. (2017). Is the positivity of estrogen receptor or progesterone receptor different between type 1 and type 2 endometrial cancer? *Oncotarget*, *8*(1), 506–511. https://doi.org/10.18632/oncotarget.13471

Spainhour, J. C., Lim, H. S., Yi, S. V., & Qiu, P. (2019). Correlation Patterns Between DNA Methylation and Gene Expression in The Cancer Genome Atlas. *Cancer Informatics*, *18*, 117693511982877. https://doi.org/10.1177/1176935119828776

Speicher, N. K., & Pfeifer, N. (2015). Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*, *31*(12), i268–i275. https://doi.org/10.1093/bioinformatics/btv244

Steele, C. D., Abbasi, A., Islam, S. M. A., Bowes, A. L., Khandekar, A., Haase, K., Hames-Fathi, S., Ajayi, D., Verfaillie, A., Dhami, P., McLatchie, A., Lechner, M., Light, N., Shlien, A., Malkin, D., Feber, A., Proszek, P., Lesluyes, T., Mertens, F., … Pillay, N. (2022). Signatures of copy number alterations in human cancer. *Nature*, *606*(7916), 984–991. https://doi.org/10.1038/s41586-022-04738-6

Suryo Rahmanto, Y., Shen, W., Shi, X., Chen, X., Yu, Y., Yu, Z.-C., Miyamoto, T., Lee, M.-H., Singh, V., Asaka, R., Shimberg, G., Vitolo, M. I., Martin, S. S., Wirtz, D., Drapkin, R., Xuan, J., Wang, T.-L., & Shih, I.-M. (2020). Inactivation of Arid1a in the endometrium is associated with endometrioid tumorigenesis through transcriptional reprogramming. *Nature Communications*, *11*(1), 2717. https://doi.org/10.1038/s41467-020-16416-0

Tao, M. H., & Freudenheim, J. L. (2010). DNA methylation in endometrial cancer. *Epigenetics*, *5*(6), 491–498. https://doi.org/10.4161/epi.5.6.12431

Tartaglia, M., Kalidas, K., Shaw, A., Song, X., Musat, D. L., Van Der Burgt, I., Brunner, H. G., Bertola, D. R., Crosby, A., Ion, A., Kucherlapati, R. S., Jeffery, S., Patton, M. A., & Gelb, B. D. (2002). PTPN11 Mutations in Noonan Syndrome: Molecular Spectrum, Genotype-Phenotype Correlation, and Phenotypic Heterogeneity. *The American Journal of Human Genetics*, *70*(6), 1555–1563. https://doi.org/10.1086/340847

Temin, H. M., & Mizutani, S. (1970). Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of Rous Sarcoma Virus. *Nature*, *226*(5252), 1211–1213. https://doi.org/10.1038/2261211a0

The Cancer Genome Atlas Research Network, & Levine, D. A. (2013). Integrated genomic characterization of endometrial carcinoma. *Nature*, *497*(7447), 67–73. https://doi.org/10.1038/nature12113

Tillement, V., Haren, L., Roullet, N., Etievant, C., & Merdes, A. (2009). The centrosome protein NEDD1 as a potential pharmacological target to induce cell cycle arrest. *Molecular Cancer*, *8*(1), 10. https://doi.org/10.1186/1476-4598-8-10

Van Nyen, T., Moiola, C. P., Colas, E., Annibali, D., & Amant, F. (2018). Modeling Endometrial Cancer: Past, Present, and Future. *International Journal of Molecular Sciences*, *19*(8), 2348. https://doi.org/10.3390/ijms19082348

Wang, Y., Ren, F., Chen, P., Liu, S., Song, Z., & Ma, X. (2018). Identification of a six-gene signature with prognostic value for patients with endometrial carcinoma. *Cancer Medicine*, *7*(11), 5632–5642. https://doi.org/10.1002/cam4.1806

Xu. (2010). The role of VIT1/FBXO11 in the regulation of apoptosis and tyrosinase export from endoplasmic reticulum in cultured melanocytes. *International Journal of Molecular Medicine*, *26*(1). https://doi.org/10.3892/ijmm_00000435

Yaghmaie, F., Saeed, O., Garan, S. A., Freitag, W., Timiras, P. S., & Sternberg, H. (2005). Caloric restriction reduces cell loss and maintains estrogen receptor-alpha immunoreactivity in the pre-optic hypothalamus of female B6D2F1 mice. *Neuro Endocrinology Letters*, *26*(3), 197–203.

Anexos

Anexo 1: Listado de los pacientes seleccionados.

TCGA-2E-A9G8; TCGA-3C-AAAU; TCGA-3C-AALJ; TCGA-3C-AALK; TCGA-4E-A92E; TCGA-4H-AAAK; TCGA-5B-A90C; TCGA-5L-AAT0; TCGA-5L-AAT1; TCGA-5S-A9Q8; TCGA-A1-A0SD; TCGA-A1-A0SF; TCGA-A1-A0SI; TCGA-A1-A0SJ; TCGA-A1-A0SN; TCGA-A2-A04N; TCGA-A2-A04R; TCGA-A2-A04V; TCGA-A2-A04X; TCGA-A2-A04Y; TCGA-A2-A0CK; TCGA-A2-A0CO; TCGA-A2-A0CP; TCGA-A2-A0CQ; TCGA-A2-A0CS; TCGA-A2-A0CU; TCGA-A2-A0CV; TCGA-A2-A0CW; TCGA-A2-A0D3; TCGA-A2-A0D4; TCGA-A2-A0EM; TCGA-A2-A0EN; TCGA-A2-A0EO; TCGA-A2-A0ER; TCGA-A2-A0ES; TCGA-A2-A0ET; TCGA-A2-A0EU; TCGA-A2-A0EV; TCGA-A2-A0EW; TCGA-A2-A0EX; TCGA-A2-A0SU; TCGA-A2-A0SV; TCGA-A2-A0SY; TCGA-A2-A0T3; TCGA-A2-A0T4; TCGA-A2-A0T5; TCGA-A2-A0T6; TCGA-A2-A0T7; TCGA-A2-A0YC; TCGA-A2-A0YD; TCGA-A2-A0YG; TCGA-A2-A0YH; TCGA-A2-A0YI; TCGA-A2-A0YL; TCGA-A2-A1FV; TCGA-A2-A1FX; TCGA-A2-A1FZ; TCGA-A2-A1G4; TCGA-A2-A259; TCGA-A2-A25A; TCGA-A2-A25B; TCGA-A2-A25C; TCGA-A2-A25E; TCGA-A2-A3KC; TCGA-A2-A4RW; TCGA-A2-A4RY; TCGA-A2-A4S2; TCGA-A2-A4S3; TCGA-A5-A0G9; TCGA-A5-A0GA; TCGA-A5-A0GB; TCGA-A5-A0GD; TCGA-A5-A0GE; TCGA-A5-A0GG; TCGA-A5-A0GJ; TCGA-A5-A0GM; TCGA-A5-A0GN; TCGA-A5-A0GP; TCGA-A5-A0GQ; TCGA-A5-A0GU; TCGA-A5-A0GV; TCGA-A5-A0GX; TCGA-A5-A0R7; TCGA-A5-A0R8; TCGA-A5-A0R9; TCGA-A5-A0RA; TCGA-A5-A0VP; TCGA-A5-A0VQ; TCGA-A5-A1OJ; TCGA-A5-A1OK; TCGA-A5-A2K5; TCGA-A5-A2K7; TCGA-A5-A3LO; TCGA-A5-A7WJ; TCGA-A5-AB3J; TCGA-A7-A0CD; TCGA-A7-A0CJ; TCGA-A7-A13F; TCGA-A7-A13H; TCGA-A7-A2KD; TCGA-A7-A3J1; TCGA-A7-A425; TCGA-A7-A426; TCGA-A7-A4SB; TCGA-A7-A56D; TCGA-A7-A5ZW; TCGA-A7-A5ZX; TCGA-A7-A6VX; TCGA-A8-A06O; TCGA-A8-A06P; TCGA-A8-A06Q; TCGA-A8-A06T; TCGA-A8-A06U; TCGA-A8-A06Y; TCGA-A8-A06Z; TCGA-A8-A076; TCGA-A8-A079; TCGA-A8-A07B; TCGA-A8-A07E; TCGA-A8-A07F; TCGA-A8-A07G; TCGA-A8-A07J; TCGA-A8-A07L; TCGA-A8-A07P; TCGA-A8-A07W; TCGA-A8-A07Z; TCGA-A8-A081; TCGA-A8-A082; TCGA-A8-A083; TCGA-A8-A086; TCGA-A8-A08A; TCGA-A8-A08C; TCGA-A8-A08F; TCGA-A8-A08G; TCGA-A8-A08H; TCGA-A8-A08I; TCGA-A8-A08O; TCGA-A8-A08P; TCGA-A8-A08S; TCGA-A8-A08T; TCGA-A8-A08Z; TCGA-A8-A090; TCGA-A8-A092; TCGA-A8-A093; TCGA-A8-A095; TCGA-A8-A096; TCGA-A8-A097; TCGA-A8-A099; TCGA-A8-A09A; TCGA-A8-A09B; TCGA-A8-A09C; TCGA-A8-A09D; TCGA-A8-A09E; TCGA-A8-A09I; TCGA-A8-A09K; TCGA-A8-A09M; TCGA-A8-A09N; TCGA-A8-A09Q; TCGA-A8-A09R; TCGA-A8-A09T; TCGA-A8-A09V; TCGA-A8-A09W; TCGA-A8-A0A1; TCGA-A8-A0A2; TCGA-A8-A0A4; TCGA-A8-A0A6; TCGA-A8-A0A9; TCGA-A8-A0AB; TCGA-AC-A23C; TCGA-AC-A23E; TCGA-AC-A23G; TCGA-AC-A2B8; TCGA-AC-A2BM; TCGA-AC-A2FB; TCGA-AC-A2FE; TCGA-AC-A2FF; TCGA-AC-A2FK; TCGA-AC-A3BB; TCGA-AC-A3HN; TCGA-AC-A3QP; TCGA-AC-A3QQ; TCGA-AC-A3TM; TCGA-AC-A3TN; TCGA-AC-A3W6; TCGA-AC-A3YI; TCGA-AC-A4ZE; TCGA-AC-A5XS; TCGA-AC-A5XU; TCGA-AC-A62Y; TCGA-AC-A6IV; TCGA-AC-A6IX; TCGA-AC-A6NO; TCGA-AC-A8OP; TCGA-AC-A8OS; TCGA-AJ-A2QK; TCGA-AJ-A2QL; TCGA-AJ-A2QN; TCGA-AJ-A2QO; TCGA-AJ-A3BH; TCGA-AJ-A3BI; TCGA-AJ-A3BK; TCGA-AJ-A3EK; TCGA-AJ-A3EL; TCGA-AJ-A3EM; TCGA-AJ-A3I9; TCGA-AJ-A3NC; TCGA-AJ-A3NE; TCGA-AJ-A3OJ; TCGA-AJ-A3OL; TCGA-AJ-A5DV; TCGA-AJ-A5DW; TCGA-AJ-A6NU; TCGA-AJ-A8CT; TCGA-AJ-A8CV; TCGA-AJ-A8CW;

TCGA-AN-A03X; TCGA-AN-A03Y; TCGA-AN-A046; TCGA-AN-A049; TCGA-AN-A04A; TCGA-AN-A0AJ; TCGA-AN-A0FD; TCGA-AN-A0FF; TCGA-AN-A0FK; TCGA-AN-A0FN; TCGA-AN-A0FW; TCGA-AN-A0FY; TCGA-AN-A0XL; TCGA-AN-A0XP; TCGA-AN-A0XV; TCGA-AN-A0XW; TCGA-AO-A03L; TCGA-AO-A03M; TCGA-AO-A03N; TCGA-AO-A03O; TCGA-AO-A03P; TCGA-AO-A03R; TCGA-AO-A03T; TCGA-AO-A03V; TCGA-AO-A0J8; TCGA-AO-A0J9; TCGA-AO-A0JA; TCGA-AO-A0JB; TCGA-AO-A0JD; TCGA-AO-A0JF; TCGA-AO-A0JJ; TCGA-AO-A0JM; TCGA-AO-A126; TCGA-AO-A12A; TCGA-AO-A12B; TCGA-AO-A12E; TCGA-AO-A1KO; TCGA-AO-A1KP; TCGA-AO-A1KS; TCGA-AO-A1KT; TCGA-AP-A051; TCGA-AP-A053; TCGA-AP-A056; TCGA-AP-A05N; TCGA-AP-A05O; TCGA-AP-A05P; TCGA-AP-A0LE; TCGA-AP-A0LG; TCGA-AP-A0LJ; TCGA-AP-A0LL; TCGA-AP-A0LM; TCGA-AP-A0LN; TCGA-AP-A0LO; TCGA-AP-A0LP; TCGA-AP-A0LS; TCGA-AP-A0LT; TCGA-AP-A0LV; TCGA-AP-A1DH; TCGA-AP-A1DK; TCGA-AP-A1DM; TCGA-AP-A1DO; TCGA-AP-A1DP; TCGA-AP-A1DR; TCGA-AP-A1DV; TCGA-AP-A1E0; TCGA-AP-A1E1; TCGA-AP-A1E3; TCGA-AQ-A04H; TCGA-AQ-A0Y5; TCGA-AQ-A1H2; TCGA-AQ-A1H3; TCGA-AR-A0TR; TCGA-AR-A0TV; TCGA-AR-A0TW; TCGA-AR-A0TZ; TCGA-AR-A0U2; TCGA-AR-A0U3; TCGA-AR-A1AK; TCGA-AR-A1AL; TCGA-AR-A1AM; TCGA-AR-A1AN; TCGA-AR-A1AP; TCGA-AR-A1AS; TCGA-AR-A1AU; TCGA-AR-A1AX; TCGA-AR-A24H; TCGA-AR-A24K; TCGA-AR-A24L; TCGA-AR-A24N; TCGA-AR-A24P; TCGA-AR-A24R; TCGA-AR-A24S; TCGA-AR-A24T; TCGA-AR-A24V; TCGA-AR-A24W; TCGA-AR-A24Z; TCGA-AR-A254; TCGA-AR-A255; TCGA-AR-A2LJ; TCGA-AR-A2LK; TCGA-AR-A2LL; TCGA-AR-A2LM; TCGA-AR-A2LN; TCGA-AR-A2LO; TCGA-AR-A2LQ; TCGA-AR-A5QM; TCGA-AR-A5QN; TCGA-AR-A5QP; TCGA-AW-A1PO; TCGA-AX-A05S; TCGA-AX-A05T; TCGA-AX-A05U; TCGA-AX-A05W; TCGA-AX-A05Y; TCGA-AX-A05Z; TCGA-AX-A060; TCGA-AX-A062; TCGA-AX-A06B; TCGA-AX-A06J; TCGA-AX-A06L; TCGA-AX-A0IS; TCGA-AX-A0J0; TCGA-AX-A0J1; TCGA-AX-A1C4; TCGA-AX-A1C5; TCGA-AX-A1C9; TCGA-AX-A1CE; TCGA-AX-A1CF; TCGA-AX-A1CI; TCGA-AX-A1CJ; TCGA-AX-A1CK; TCGA-AX-A2H7; TCGA-AX-A2H8; TCGA-AX-A2HA; TCGA-AX-A2HC; TCGA-AX-A2HD; TCGA-AX-A2HJ; TCGA-AX-A2HK; TCGA-AX-A3FW; TCGA-AX-A3FX; TCGA-AX-A3G8; TCGA-AX-A3G9; TCGA-AX-A3GB; TCGA-B5-A0JR; TCGA-B5-A0JS; TCGA-B5-A0JT; TCGA-B5-A0JU; TCGA-B5-A0JV; TCGA-B5-A0JX; TCGA-B5-A0JY; TCGA-B5-A0JZ; TCGA-B5-A0K0; TCGA-B5-A0K1; TCGA-B5-A0K2; TCGA-B5-A0K3; TCGA-B5-A0K4; TCGA-B5-A0K6; TCGA-B5-A0K7; TCGA-B5-A0K9; TCGA-B5-A0KB; TCGA-B5-A11E; TCGA-B5-A11F; TCGA-B5-A11G; TCGA-B5-A11I; TCGA-B5-A11J; TCGA-B5-A11M; TCGA-B5-A11N; TCGA-B5-A11O; TCGA-B5-A11P; TCGA-B5-A11Q; TCGA-B5-A11R; TCGA-B5-A11S; TCGA-B5-A11U; TCGA-B5-A11V; TCGA-B5-A11W; TCGA-B5-A11Y; TCGA-B5-A11Z; TCGA-B5-A121; TCGA-B5-A1MV; TCGA-B5-A1MW; TCGA-B5-A1MX; TCGA-B5-A1MZ; TCGA-B5-A3F9; TCGA-B5-A3FB; TCGA-B5-A3FD; TCGA-B5-A3FH; TCGA-B5-A5OC; TCGA-B6-A0I5; TCGA-B6-A0I9; TCGA-B6-A0IB; TCGA-B6-A0IG; TCGA-B6-A0IO; TCGA-B6-A0IP; TCGA-B6-A0RH; TCGA-B6-A0RI; TCGA-B6-A0RL; TCGA-B6-A0RM; TCGA-B6-A0RO; TCGA-B6-A0RQ; TCGA-B6-A0WS; TCGA-B6-A0WT; TCGA-B6-A0WV; TCGA-B6-A0WW; TCGA-B6-A0WZ; TCGA-B6-A0X4; TCGA-B6-A0X5; TCGA-B6-A1KI; TCGA-B6-A2IU; TCGA-B6-A401; TCGA-B6-A40B; TCGA-B6-A40C; TCGA-BG-A0LW; TCGA-BG-A0LX; TCGA-BG-A0M0; TCGA-BG-A0M2; TCGA-BG-A0M3; TCGA-BG-A0M4; TCGA-BG-A0M7; TCGA-BG-A0M9; TCGA-BG-A0MA; TCGA-BG-A0MC; TCGA-BG-A0MG; TCGA-BG-A0MI; TCGA-BG-A0MK; TCGA-BG-A0MO; TCGA-BG-A0MQ; TCGA-BG-A0MS; TCGA-BG-A0MT; TCGA-BG-A0MU; TCGA-BG-A0RY; TCGA-BG-A0VT; TCGA-BG-A0VV; TCGA-BG-A0VW; TCGA-BG-A0VX; TCGA-BG-A0VZ; TCGA-BG-A0W1; TCGA-BG-A0W2; TCGA-BG-A0YU; TCGA-BG-A186; TCGA-BG-A187; TCGA-BG-A18A; TCGA-BG-A18B; TCGA-BG-A18C; TCGA-BG-A220; TCGA-BG-A222; TCGA-BG-A2AD; TCGA-BG-A2AE; TCGA-BG-A2L7; TCGA-BG-A3EW; TCGA-BH-A0AU; TCGA-BH-A0AY; TCGA-BH-A0AZ; TCGA-BH-A0B0;

TCGA-BH-A0B5; TCGA-BH-A0B6; TCGA-BH-A0BC; TCGA-BH-A0BD; TCGA-BH-A0BJ; TCGA-BH-A0BO; TCGA-BH-A0BP; TCGA-BH-A0BQ; TCGA-BH-A0BR; TCGA-BH-A0BS; TCGA-BH-A0BT; TCGA-BH-A0BV; TCGA-BH-A0BZ; TCGA-BH-A0C0; TCGA-BH-A0C1; TCGA-BH-A0DE; TCGA-BH-A0DH; TCGA-BH-A0DI; TCGA-BH-A0DK; TCGA-BH-A0DO; TCGA-BH-A0DP; TCGA-BH-A0DQ; TCGA-BH-A0DS; TCGA-BH-A0DT; TCGA-BH-A0DV; TCGA-BH-A0DX; TCGA-BH-A0DZ; TCGA-BH-A0E1; TCGA-BH-A0E2; TCGA-BH-A0E7; TCGA-BH-A0E9; TCGA-BH-A0EA; TCGA-BH-A0EB; TCGA-BH-A0EI; TCGA-BH-A0GY; TCGA-BH-A0GZ; TCGA-BH-A0H0; TCGA-BH-A0H3; TCGA-BH-A0H6; TCGA-BH-A0H7; TCGA-BH-A0H9; TCGA-BH-A0HA; TCGA-BH-A0HB; TCGA-BH-A0HI; TCGA-BH-A0HO; TCGA-BH-A0HQ; TCGA-BH-A0HU; TCGA-BH-A0HX; TCGA-BH-A0W3; TCGA-BH-A0W4; TCGA-BH-A0W5; TCGA-BH-A0W7; TCGA-BH-A18F; TCGA-BH-A18I; TCGA-BH-A18J; TCGA-BH-A18K; TCGA-BH-A18L; TCGA-BH-A18M; TCGA-BH-A18N; TCGA-BH-A1ES; TCGA-BH-A1ET; TCGA-BH-A1EU; TCGA-BH-A1EV; TCGA-BH-A1EX; TCGA-BH-A1EY; TCGA-BH-A1F2; TCGA-BH-A1F5; TCGA-BH-A1F8; TCGA-BH-A1FB; TCGA-BH-A1FD; TCGA-BH-A1FE; TCGA-BH-A1FG; TCGA-BH-A1FL; TCGA-BH-A1FN; TCGA-BH-A201; TCGA-BH-A202; TCGA-BH-A28O; TCGA-BH-A28Q; TCGA-BH-A2L8; TCGA-BH-A42T; TCGA-BH-A42V; TCGA-BH-A5J0; TCGA-BH-A6R8; TCGA-BH-A8FY; TCGA-BH-A8FZ; TCGA-BH-A8G0; TCGA-BH-AB28; TCGA-BK-A0C9; TCGA-BK-A0CB; TCGA-BK-A13B; TCGA-BK-A13C; TCGA-BK-A4ZD; TCGA-BK-A56F; TCGA-BK-A6W3; TCGA-BK-A6W4; TCGA-BS-A0TA; TCGA-BS-A0TC; TCGA-BS-A0TD; TCGA-BS-A0TI; TCGA-BS-A0TJ; TCGA-BS-A0U5; TCGA-BS-A0U7; TCGA-BS-A0U8; TCGA-BS-A0UA; TCGA-BS-A0UJ; TCGA-BS-A0UL; TCGA-BS-A0UM; TCGA-BS-A0UT; TCGA-BS-A0UV; TCGA-BS-A0V4; TCGA-BS-A0V6; TCGA-BS-A0V7; TCGA-BS-A0V8; TCGA-BS-A0VI; TCGA-BS-A0WQ; TCGA-C8-A12N; TCGA-C8-A12O; TCGA-C8-A12T; TCGA-C8-A12U; TCGA-C8-A12W; TCGA-C8-A130; TCGA-C8-A132; TCGA-C8-A1HE; TCGA-C8-A1HI; TCGA-C8-A1HM; TCGA-C8-A1HN; TCGA-C8-A1HO; TCGA-C8-A26V; TCGA-C8-A26W; TCGA-C8-A26Z; TCGA-C8-A273; TCGA-C8-A274; TCGA-C8-A27A; TCGA-C8-A3M8; TCGA-D1-A0ZN; TCGA-D1-A0ZO; TCGA-D1-A0ZQ; TCGA-D1-A0ZR; TCGA-D1-A0ZU; TCGA-D1-A0ZV; TCGA-D1-A101; TCGA-D1-A102; TCGA-D1-A103; TCGA-D1-A15W; TCGA-D1-A15Z; TCGA-D1-A160; TCGA-D1-A161; TCGA-D1-A162; TCGA-D1-A163; TCGA-D1-A165; TCGA-D1-A167; TCGA-D1-A169; TCGA-D1-A16B; TCGA-D1-A16D; TCGA-D1-A16E; TCGA-D1-A16N; TCGA-D1-A16O; TCGA-D1-A16Q; TCGA-D1-A16R; TCGA-D1-A16V; TCGA-D1-A16X; TCGA-D1-A174; TCGA-D1-A175; TCGA-D1-A176; TCGA-D1-A177; TCGA-D1-A17A; TCGA-D1-A17B; TCGA-D1-A17C; TCGA-D1-A17D; TCGA-D1-A17F; TCGA-D1-A17H; TCGA-D1-A17K; TCGA-D1-A17L; TCGA-D1-A17M; TCGA-D1-A17N; TCGA-D1-A17Q; TCGA-D1-A17R; TCGA-D1-A17S; TCGA-D1-A17T; TCGA-D1-A17U; TCGA-D1-A1NS; TCGA-D1-A1NY; TCGA-D1-A1NZ; TCGA-D1-A1O0; TCGA-D1-A1O5; TCGA-D1-A1O7; TCGA-D1-A2G5; TCGA-D1-A3DA; TCGA-D1-A3DG; TCGA-D8-A13Y; TCGA-D8-A140; TCGA-D8-A141; TCGA-D8-A146; TCGA-D8-A1J8; TCGA-D8-A1JB; TCGA-D8-A1JC; TCGA-D8-A1JD; TCGA-D8-A1JE; TCGA-D8-A1JH; TCGA-D8-A1JI; TCGA-D8-A1JJ; TCGA-D8-A1JN; TCGA-D8-A1JP; TCGA-D8-A1JU; TCGA-D8-A1X5; TCGA-D8-A1X6; TCGA-D8-A1X8; TCGA-D8-A1X9; TCGA-D8-A1XA; TCGA-D8-A1XB; TCGA-D8-A1XD; TCGA-D8-A1XF; TCGA-D8-A1XL; TCGA-D8-A1XM; TCGA-D8-A1XO; TCGA-D8-A1XR; TCGA-D8-A1XU; TCGA-D8-A1XY; TCGA-D8-A1Y0; TCGA-D8-A1Y1; TCGA-D8-A1Y2; TCGA-D8-A1Y3; TCGA-D8-A27G; TCGA-D8-A27I; TCGA-D8-A27K; TCGA-D8-A27L; TCGA-D8-A27N; TCGA-D8-A27P; TCGA-D8-A27R; TCGA-D8-A27T; TCGA-D8-A27V; TCGA-D8-A3Z5; TCGA-D8-A3Z6; TCGA-D8-A4Z1; TCGA-DF-A2KN; TCGA-DF-A2KR; TCGA-DF-A2KS; TCGA-DF-A2KV; TCGA-DF-A2KZ; TCGA-DF-A2L0; TCGA-DI-A0WH; TCGA-DI-A1BY; TCGA-DI-A1C3; TCGA-DI-A1NO; TCGA-DI-A2QU; TCGA-E2-A105; TCGA-E2-A108; TCGA-E2-A10A; TCGA-E2-A10B; TCGA-E2-A10C; TCGA-E2-A10E; TCGA-E2-A10F; TCGA-E2-A14O; TCGA-E2-A14Q; TCGA-E2-

A14S; TCGA-E2-A14T; TCGA-E2-A14U; TCGA-E2-A14V; TCGA-E2-A14Z; TCGA-E2-A153; TCGA-E2-A154; TCGA-E2-A156; TCGA-E2-A15A; TCGA-E2-A15C; TCGA-E2-A15D; TCGA-E2-A15E; TCGA-E2-A15F; TCGA-E2-A15H; TCGA-E2-A15I; TCGA-E2-A15J; TCGA-E2-A15K; TCGA-E2-A15L; TCGA-E2-A15M; TCGA-E2-A15O; TCGA-E2-A15P; TCGA-E2-A15R; TCGA-E2-A15T; TCGA-E2-A1B1; TCGA-E2-A1B4; TCGA-E2-A1B5; TCGA-E2-A1BC; TCGA-E2-A1BD; TCGA-E2-A1IE; TCGA-E2-A1IF; TCGA-E2-A1IG; TCGA-E2-A1IH; TCGA-E2-A1IJ; TCGA-E2-A1IK; TCGA-E2-A1IL; TCGA-E2-A1IN; TCGA-E2-A1IO; TCGA-E2-A1IU; TCGA-E2-A1L6; TCGA-E2-A1L8; TCGA-E2-A1L9; TCGA-E2-A2P5; TCGA-E2-A2P6; TCGA-E2-A3DX; TCGA-E2-A56Z; TCGA-E2-A570; TCGA-E2-A576; TCGA-E6-A1LX; TCGA-E6-A1M0; TCGA-E6-A2P9; TCGA-E9-A1N6; TCGA-E9-A1NE; TCGA-E9-A1NG; TCGA-E9-A1NH; TCGA-E9-A227; TCGA-E9-A22D; TCGA-E9-A22E; TCGA-E9-A22H; TCGA-E9-A295; TCGA-E9-A3Q9; TCGA-E9-A3X8; TCGA-E9-A54Y; TCGA-E9-A5FK; TCGA-E9-A6HE; TCGA-EC-A1NJ; TCGA-EC-A1QX; TCGA-EC-A24G; TCGA-EO-A1Y7; TCGA-EO-A22R; TCGA-EO-A22S; TCGA-EO-A22T; TCGA-EO-A22U; TCGA-EO-A22X; TCGA-EO-A22Y; TCGA-EO-A3AS; TCGA-EO-A3AU; TCGA-EO-A3AY; TCGA-EO-A3B0; TCGA-EO-A3KX; TCGA-EO-A3L0; TCGA-EW-A1IW; TCGA-EW-A1IY; TCGA-EW-A1J1; TCGA-EW-A1J2; TCGA-EW-A1J3; TCGA-EW-A1J5; TCGA-EW-A1J6; TCGA-EW-A1OY; TCGA-EW-A1P3; TCGA-EW-A1P5; TCGA-EW-A1P6; TCGA-EW-A1PA; TCGA-EW-A1PC; TCGA-EW-A1PE; TCGA-EW-A1PF; TCGA-EW-A1PG; TCGA-EW-A423; TCGA-EW-A6S9; TCGA-EW-A6SC; TCGA-EY-A1G8; TCGA-EY-A1GD; TCGA-EY-A1GE; TCGA-EY-A1GF; TCGA-EY-A1GH; TCGA-EY-A1GI; TCGA-EY-A1GK; TCGA-EY-A1GL; TCGA-EY-A1GQ; TCGA-EY-A1GR; TCGA-EY-A1GT; TCGA-EY-A1GU; TCGA-EY-A1GW; TCGA-EY-A1GX; TCGA-EY-A1H0; TCGA-EY-A214; TCGA-EY-A215; TCGA-EY-A2OM; TCGA-EY-A2OP; TCGA-EY-A2OQ; TCGA-EY-A547; TCGA-EY-A548; TCGA-EY-A549; TCGA-EY-A54A; TCGA-EY-A5W2; TCGA-EY-A72D; TCGA-FI-A2D0; TCGA-FI-A2D4; TCGA-FI-A2D5; TCGA-FI-A2F4; TCGA-GI-A2C8; TCGA-GM-A2D9; TCGA-GM-A2DA; TCGA-GM-A2DC; TCGA-GM-A2DL; TCGA-GM-A2DM; TCGA-GM-A3NY; TCGA-GM-A3XG; TCGA-GM-A3XN; TCGA-GM-A4E0; TCGA-GM-A5PV; TCGA-GM-A5PX; TCGA-H5-A2HR; TCGA-HN-A2OB; TCGA-JL-A3YX; TCGA-LD-A66U; TCGA-LD-A7W5; TCGA-LL-A440; TCGA-LL-A50Y; TCGA-LL-A5YM; TCGA-LL-A5YN; TCGA-LL-A6FQ; TCGA-LL-A73Z; TCGA-LL-A7T0; TCGA-LL-A9Q3; TCGA-LQ-A4E4; TCGA-OK-A5Q2; TCGA-OL-A5D8; TCGA-OL-A5DA; TCGA-OL-A5RV; TCGA-OL-A5RX; TCGA-OL-A66J; TCGA-OL-A66K; TCGA-OL-A66L; TCGA-OL-A66O; TCGA-OL-A6VQ; TCGA-OL-A6VR; TCGA-PE-A5DC; TCGA-PE-A5DE; TCGA-PG-A916; TCGA-PG-A917; TCGA-QF-A5YT; TCGA-QS-A5YQ; TCGA-QS-A744; TCGA-S3-A6ZF; TCGA-S3-A6ZH; TCGA-S3-AA0Z; TCGA-S3-AA11; TCGA-S3-AA14; TCGA-S3-AA17; TCGA-SJ-A6ZI; TCGA-SJ-A6ZJ; TCGA-SL-A6JA; TCGA-UL-AAZ6; TCGA-V7-A7HQ; TCGA-W8-A86G; TCGA-WT-AB41; TCGA-WT-AB44; TCGA-XX-A899; TCGA-XX-A89A; TCGA-Z7-A8R5; TCGA-Z7-A8R6.

Anexo 2: Listado de los pacientes en cada *cluster* transcriptómico.

Cluster transcriptómico 1

TCGA-3C-AAAU; TCGA-3C-AALJ; TCGA-3C-AALK; TCGA-4H-AAAK; TCGA-5L-AAT0; TCGA-5L-AAT1; TCGA-A1-A0SD; TCGA-A1-A0SF; TCGA-A1-A0SI; TCGA-A1-A0SJ; TCGA-A1-A0SN; TCGA-A2-A04N; TCGA-A2-A04R; TCGA-A2-A04V; TCGA-A2-A04X; TCGA-A2-A04Y; TCGA-A2-A0CK; TCGA-A2-A0CO; TCGA-A2-A0CP; TCGA-A2-A0CQ; TCGA-A2-A0CS; TCGA-A2-A0CU; TCGA-A2-A0CV; TCGA-A2-A0CW; TCGA-A2-A0D3; TCGA-A2-A0D4; TCGA-A2-A0EM; TCGA-A2-A0EN; TCGA-A2-A0EO; TCGA-A2-A0ER;

TCGA-A2-A0ES; TCGA-A2-A0ET; TCGA-A2-A0EU; TCGA-A2-A0EV; TCGA-A2-A0EW; TCGA-A2-A0EX; TCGA-A2-A0SU; TCGA-A2-A0SV; TCGA-A2-A0SY; TCGA-A2-A0T3; TCGA-A2-A0T4; TCGA-A2-A0T5; TCGA-A2-A0T6; TCGA-A2-A0T7; TCGA-A2-A0YC; TCGA-A2-A0YD; TCGA-A2-A0YG; TCGA-A2-A0YH; TCGA-A2-A0YI; TCGA-A2-A0YL; TCGA-A2-A1FV; TCGA-A2-A1FX; TCGA-A2-A1FZ; TCGA-A2-A1G4; TCGA-A2-A259; TCGA-A2-A25A; TCGA-A2-A25B; TCGA-A2-A25C; TCGA-A2-A25E; TCGA-A2-A3KC; TCGA-A2-A4RW; TCGA-A2-A4RY; TCGA-A2-A4S2; TCGA-A2-A4S3; TCGA-A7-A0CD; TCGA-A7-A0CJ; TCGA-A7-A13F; TCGA-A7-A13H; TCGA-A7-A2KD; TCGA-A7-A3J1; TCGA-A7-A425; TCGA-A7-A426; TCGA-A7-A4SB; TCGA-A7-A56D; TCGA-A7-A5ZW; TCGA-A7-A5ZX; TCGA-A7-A6VX; TCGA-A8-A06O; TCGA-A8-A06P; TCGA-A8-A06Q; TCGA-A8-A06T; TCGA-A8-A06U; TCGA-A8-A06Y; TCGA-A8-A06Z; TCGA-A8-A076; TCGA-A8-A079; TCGA-A8-A07B; TCGA-A8-A07E; TCGA-A8-A07F; TCGA-A8-A07G; TCGA-A8-A07J; TCGA-A8-A07L; TCGA-A8-A07P; TCGA-A8-A07W; TCGA-A8-A07Z; TCGA-A8-A081; TCGA-A8-A082; TCGA-A8-A083; TCGA-A8-A086; TCGA-A8-A08A; TCGA-A8-A08C; TCGA-A8-A08F; TCGA-A8-A08G; TCGA-A8-A08H; TCGA-A8-A08I; TCGA-A8-A08O; TCGA-A8-A08P; TCGA-A8-A08S; TCGA-A8-A08T; TCGA-A8-A08Z; TCGA-A8-A090; TCGA-A8-A092; TCGA-A8-A093; TCGA-A8-A095; TCGA-A8-A096; TCGA-A8-A097; TCGA-A8-A099; TCGA-A8-A09A; TCGA-A8-A09B; TCGA-A8-A09C; TCGA-A8-A09D; TCGA-A8-A09E; TCGA-A8-A09I; TCGA-A8-A09K; TCGA-A8-A09M; TCGA-A8-A09N; TCGA-A8-A09Q; TCGA-A8-A09R; TCGA-A8-A09T; TCGA-A8-A09V; TCGA-A8-A09W; TCGA-A8-A0A1; TCGA-A8-A0A2; TCGA-A8-A0A4; TCGA-A8-A0A6; TCGA-A8-A0A9; TCGA-A8-A0AB; TCGA-AC-A23C; TCGA-AC-A23E; TCGA-AC-A23G; TCGA-AC-A2B8; TCGA-AC-A2BM; TCGA-AC-A2FB; TCGA-AC-A2FE; TCGA-AC-A2FF; TCGA-AC-A2FK; TCGA-AC-A3BB; TCGA-AC-A3HN; TCGA-AC-A3QP; TCGA-AC-A3QQ; TCGA-AC-A3TM; TCGA-AC-A3TN; TCGA-AC-A3W6; TCGA-AC-A3YI; TCGA-AC-A4ZE; TCGA-AC-A5XS; TCGA-AC-A5XU; TCGA-AC-A62Y; TCGA-AC-A6IV; TCGA-AC-A6IX; TCGA-AC-A6NO; TCGA-AC-A8OP; TCGA-AC-A8OS; TCGA-AN-A03X; TCGA-AN-A03Y; TCGA-AN-A046; TCGA-AN-A049; TCGA-AN-A04A; TCGA-AN-A0AJ; TCGA-AN-A0FD; TCGA-AN-A0FF; TCGA-AN-A0FK; TCGA-AN-A0FN; TCGA-AN-A0FW; TCGA-AN-A0FY; TCGA-AN-A0XL; TCGA-AN-A0XP; TCGA-AN-A0XV; TCGA-AN-A0XW; TCGA-AO-A03L; TCGA-AO-A03M; TCGA-AO-A03N; TCGA-AO-A03O; TCGA-AO-A03P; TCGA-AO-A03R; TCGA-AO-A03T; TCGA-AO-A03V; TCGA-AO-A0J8; TCGA-AO-A0J9; TCGA-AO-A0JA; TCGA-AO-A0JB; TCGA-AO-A0JD; TCGA-AO-A0JF; TCGA-AO-A0JJ; TCGA-AO-A0JM; TCGA-AO-A126; TCGA-AO-A12A; TCGA-AO-A12B; TCGA-AO-A12E; TCGA-AO-A1KO; TCGA-AO-A1KP; TCGA-AO-A1KS; TCGA-AO-A1KT; TCGA-AQ-A04H; TCGA-AQ-A0Y5; TCGA-AQ-A1H2; TCGA-AQ-A1H3; TCGA-AR-A0TR; TCGA-AR-A0TV; TCGA-AR-A0TW; TCGA-AR-A0TZ; TCGA-AR-A0U2; TCGA-AR-A0U3; TCGA-AR-A1AK; TCGA-AR-A1AL; TCGA-AR-A1AM; TCGA-AR-A1AN; TCGA-AR-A1AP; TCGA-AR-A1AS; TCGA-AR-A1AU; TCGA-AR-A1AX; TCGA-AR-A24H; TCGA-AR-A24K; TCGA-AR-A24L; TCGA-AR-A24N; TCGA-AR-A24P; TCGA-AR-A24R; TCGA-AR-A24S; TCGA-AR-A24T; TCGA-AR-A24V; TCGA-AR-A24W; TCGA-AR-A24Z; TCGA-AR-A254; TCGA-AR-A255; TCGA-AR-A2LJ; TCGA-AR-A2LK; TCGA-AR-A2LL; TCGA-AR-A2LM; TCGA-AR-A2LN; TCGA-AR-A2LO; TCGA-AR-A2LQ; TCGA-AR-A5QM; TCGA-AR-A5QN; TCGA-AR-A5QP; TCGA-B6-A0I5; TCGA-B6-A0I9; TCGA-B6-A0IB; TCGA-B6-A0IG; TCGA-B6-A0IO; TCGA-B6-A0IP; TCGA-B6-A0RH; TCGA-B6-A0RI; TCGA-B6-A0RL; TCGA-B6-A0RM; TCGA-B6-A0RO; TCGA-B6-A0RQ; TCGA-B6-A0WS; TCGA-B6-A0WT; TCGA-B6-A0WV; TCGA-B6-A0WW; TCGA-B6-A0WZ; TCGA-B6-A0X4; TCGA-B6-A0X5; TCGA-B6-A1KI; TCGA-B6-A2IU; TCGA-B6-A401; TCGA-B6-A40B; TCGA-B6-A40C; TCGA-BH-A0AU; TCGA-BH-A0AY; TCGA-BH-A0AZ; TCGA-BH-A0B0; TCGA-BH-A0B5; TCGA-BH-A0B6; TCGA-BH-A0BC; TCGA-BH-A0BD; TCGA-BH-A0BJ; TCGA-BH-A0BO; TCGA-BH-A0BP; TCGA-BH-A0BQ; TCGA-BH-A0BR; TCGA-BH-A0BS; TCGA-BH-A0BT; TCGA-BH-A0BV; TCGA-BH-

A0BZ; TCGA-BH-A0C0; TCGA-BH-A0C1; TCGA-BH-A0DE; TCGA-BH-A0DH; TCGA-BH-A0DI; TCGA-BH-A0DK; TCGA-BH-A0DO; TCGA-BH-A0DP; TCGA-BH-A0DQ; TCGA-BH-A0DS; TCGA-BH-A0DT; TCGA-BH-A0DV; TCGA-BH-A0DX; TCGA-BH-A0DZ; TCGA-BH-A0E1; TCGA-BH-A0E2; TCGA-BH-A0E7; TCGA-BH-A0E9; TCGA-BH-A0EA; TCGA-BH-A0EB; TCGA-BH-A0EI; TCGA-BH-A0GY; TCGA-BH-A0GZ; TCGA-BH-A0H0; TCGA-BH-A0H3; TCGA-BH-A0H6; TCGA-BH-A0H7; TCGA-BH-A0H9; TCGA-BH-A0HA; TCGA-BH-A0HB; TCGA-BH-A0HI; TCGA-BH-A0HO; TCGA-BH-A0HQ; TCGA-BH-A0HU; TCGA-BH-A0HX; TCGA-BH-A0W3; TCGA-BH-A0W4; TCGA-BH-A0W5; TCGA-BH-A0W7; TCGA-BH-A18F; TCGA-BH-A18I; TCGA-BH-A18J; TCGA-BH-A18K; TCGA-BH-A18L; TCGA-BH-A18M; TCGA-BH-A18N; TCGA-BH-A1ES; TCGA-BH-A1ET; TCGA-BH-A1EU; TCGA-BH-A1EV; TCGA-BH-A1EX; TCGA-BH-A1EY; TCGA-BH-A1F2; TCGA-BH-A1F5; TCGA-BH-A1F8; TCGA-BH-A1FB; TCGA-BH-A1FD; TCGA-BH-A1FE; TCGA-BH-A1FG; TCGA-BH-A1FL; TCGA-BH-A1FN; TCGA-BH-A201; TCGA-BH-A202; TCGA-BH-A28O; TCGA-BH-A28Q; TCGA-BH-A2L8; TCGA-BH-A42T; TCGA-BH-A42V; TCGA-BH-A5J0; TCGA-BH-A6R8; TCGA-BH-A8FY; TCGA-BH-A8FZ; TCGA-BH-A8G0; TCGA-BH-AB28; TCGA-C8-A12N; TCGA-C8-A12O; TCGA-C8-A12T; TCGA-C8-A12U; TCGA-C8-A12W; TCGA-C8-A130; TCGA-C8-A132; TCGA-C8-A1HE; TCGA-C8-A1HI; TCGA-C8-A1HM; TCGA-C8-A1HN; TCGA-C8-A1HO; TCGA-C8-A26V; TCGA-C8-A26W; TCGA-C8-A26Z; TCGA-C8-A273; TCGA-C8-A274; TCGA-C8-A27A; TCGA-C8-A3M8; TCGA-D8-A13Y; TCGA-D8-A140; TCGA-D8-A141; TCGA-D8-A146; TCGA-D8-A1J8; TCGA-D8-A1JB; TCGA-D8-A1JC; TCGA-D8-A1JD; TCGA-D8-A1JE; TCGA-D8-A1JH; TCGA-D8-A1JI; TCGA-D8-A1JJ; TCGA-D8-A1JN; TCGA-D8-A1JP; TCGA-D8-A1JU; TCGA-D8-A1X5; TCGA-D8-A1X6; TCGA-D8-A1X8; TCGA-D8-A1X9; TCGA-D8-A1XA; TCGA-D8-A1XB; TCGA-D8-A1XD; TCGA-D8-A1XF; TCGA-D8-A1XL; TCGA-D8-A1XM; TCGA-D8-A1XO; TCGA-D8-A1XR; TCGA-D8-A1XU; TCGA-D8-A1XY; TCGA-D8-A1Y0; TCGA-D8-A1Y1; TCGA-D8-A1Y2; TCGA-D8-A1Y3; TCGA-D8-A27G; TCGA-D8-A27I; TCGA-D8-A27K; TCGA-D8-A27L; TCGA-D8-A27N; TCGA-D8-A27P; TCGA-D8-A27R; TCGA-D8-A27T; TCGA-D8-A27V; TCGA-D8-A3Z5; TCGA-D8-A3Z6; TCGA-D8-A4Z1; TCGA-E2-A105; TCGA-E2-A108; TCGA-E2-A10A; TCGA-E2-A10B; TCGA-E2-A10C; TCGA-E2-A10E; TCGA-E2-A10F; TCGA-E2-A14O; TCGA-E2-A14Q; TCGA-E2-A14S; TCGA-E2-A14T; TCGA-E2-A14U; TCGA-E2-A14V; TCGA-E2-A14Z; TCGA-E2-A153; TCGA-E2-A154; TCGA-E2-A156; TCGA-E2-A15A; TCGA-E2-A15C; TCGA-E2-A15D; TCGA-E2-A15E; TCGA-E2-A15F; TCGA-E2-A15H; TCGA-E2-A15I; TCGA-E2-A15J; TCGA-E2-A15K; TCGA-E2-A15L; TCGA-E2-A15M; TCGA-E2-A15O; TCGA-E2-A15P; TCGA-E2-A15R; TCGA-E2-A15T; TCGA-E2-A1B1; TCGA-E2-A1B4; TCGA-E2-A1B5; TCGA-E2-A1BC; TCGA-E2-A1BD; TCGA-E2-A1IE; TCGA-E2-A1IF; TCGA-E2-A1IG; TCGA-E2-A1IH; TCGA-E2-A1IJ; TCGA-E2-A1IK; TCGA-E2-A1IL; TCGA-E2-A1IN; TCGA-E2-A1IO; TCGA-E2-A1IU; TCGA-E2-A1L6; TCGA-E2-A1L8; TCGA-E2-A1L9; TCGA-E2-A2P5; TCGA-E2-A2P6; TCGA-E2-A3DX; TCGA-E2-A56Z; TCGA-E2-A570; TCGA-E2-A576; TCGA-E9-A1N6; TCGA-E9-A1NE; TCGA-E9-A1NG; TCGA-E9-A1NH; TCGA-E9-A227; TCGA-E9-A22D; TCGA-E9-A22E; TCGA-E9-A22H; TCGA-E9-A295; TCGA-E9-A3Q9; TCGA-E9-A3X8; TCGA-E9-A54Y; TCGA-E9-A5FK; TCGA-E9-A6HE; TCGA-EW-A1IW; TCGA-EW-A1IY; TCGA-EW-A1J1; TCGA-EW-A1J2; TCGA-EW-A1J3; TCGA-EW-A1J5; TCGA-EW-A1J6; TCGA-EW-A1OY; TCGA-EW-A1P3; TCGA-EW-A1P5; TCGA-EW-A1P6; TCGA-EW-A1PA; TCGA-EW-A1PC; TCGA-EW-A1PE; TCGA-EW-A1PF; TCGA-EW-A1PG; TCGA-EW-A423; TCGA-EW-A6S9; TCGA-EW-A6SC; TCGA-GI-A2C8; TCGA-GM-A2D9; TCGA-GM-A2DA; TCGA-GM-A2DC; TCGA-GM-A2DL; TCGA-GM-A2DM; TCGA-GM-A3NY; TCGA-GM-A3XG; TCGA-GM-A3XN; TCGA-GM-A4E0; TCGA-GM-A5PV; TCGA-GM-A5PX; TCGA-HN-A2OB; TCGA-JL-A3YX; TCGA-LD-A66U; TCGA-LD-A7W5; TCGA-LL-A440; TCGA-LL-A50Y; TCGA-LL-A5YM; TCGA-LL-A5YN; TCGA-LL-A6FQ; TCGA-LL-A73Z; TCGA-LL-A7T0; TCGA-LL-A9Q3; TCGA-LQ-A4E4; TCGA-OK-A5Q2; TCGA-OL-

A5D8; TCGA-OL-A5DA; TCGA-OL-A5RV; TCGA-OL-A5RX; TCGA-OL-A66J; TCGA-OL-A66K; TCGA-OL-A66L; TCGA-OL-A66O; TCGA-OL-A6VQ; TCGA-OL-A6VR; TCGA-PE-A5DC; TCGA-PE-A5DE; TCGA-S3-A6ZF; TCGA-S3-A6ZH; TCGA-S3-AA0Z; TCGA-S3-AA11; TCGA-S3-AA14; TCGA-S3-AA17; TCGA-UL-AAZ6; TCGA-V7-A7HQ; TCGA-W8-A86G; TCGA-WT-AB41; TCGA-WT-AB44; TCGA-XX-A899; TCGA-XX-A89A; TCGA-Z7-A8R5; TCGA-Z7-A8R6.

***Cluster* transcriptómico 2**

TCGA-2E-A9G8; TCGA-4E-A92E; TCGA-5B-A90C; TCGA-5S-A9Q8; TCGA-A5-A0G9; TCGA-A5-A0GA; TCGA-A5-A0GB; TCGA-A5-A0GD; TCGA-A5-A0GE; TCGA-A5-A0GG; TCGA-A5-A0GJ; TCGA-A5-A0GM; TCGA-A5-A0GN; TCGA-A5-A0GP; TCGA-A5-A0GQ; TCGA-A5-A0GU; TCGA-A5-A0GV; TCGA-A5-A0GX; TCGA-A5-A0R7; TCGA-A5-A0R8; TCGA-A5-A0R9; TCGA-A5-A0RA; TCGA-A5-A0VP; TCGA-A5-A0VQ; TCGA-A5-A1OJ; TCGA-A5-A1OK; TCGA-A5-A2K5; TCGA-A5-A2K7; TCGA-A5-A3LO; TCGA-A5-A7WJ; TCGA-A5-AB3J; TCGA-AJ-A2QK; TCGA-AJ-A2QL; TCGA-AJ-A2QN; TCGA-AJ-A2QO; TCGA-AJ-A3BH; TCGA-AJ-A3BI; TCGA-AJ-A3BK; TCGA-AJ-A3EK; TCGA-AJ-A3EL; TCGA-AJ-A3EM; TCGA-AJ-A3I9; TCGA-AJ-A3NC; TCGA-AJ-A3NE; TCGA-AJ-A3OJ; TCGA-AJ-A3OL; TCGA-AJ-A5DV; TCGA-AJ-A5DW; TCGA-AJ-A6NU; TCGA-AJ-A8CT; TCGA-AJ-A8CV; TCGA-AJ-A8CW; TCGA-AP-A051; TCGA-AP-A053; TCGA-AP-A056; TCGA-AP-A05N; TCGA-AP-A05O; TCGA-AP-A05P; TCGA-AP-A0LE; TCGA-AP-A0LG; TCGA-AP-A0LJ; TCGA-AP-A0LL; TCGA-AP-A0LM; TCGA-AP-A0LN; TCGA-AP-A0LO; TCGA-AP-A0LP; TCGA-AP-A0LS; TCGA-AP-A0LT; TCGA-AP-A0LV; TCGA-AP-A1DH; TCGA-AP-A1DK; TCGA-AP-A1DM; TCGA-AP-A1DO; TCGA-AP-A1DP; TCGA-AP-A1DR; TCGA-AP-A1DV; TCGA-AP-A1E0; TCGA-AP-A1E1; TCGA-AP-A1E3; TCGA-AW-A1PO; TCGA-AX-A05S; TCGA-AX-A05T; TCGA-AX-A05U; TCGA-AX-A05W; TCGA-AX-A05Y; TCGA-AX-A05Z; TCGA-AX-A060; TCGA-AX-A062; TCGA-AX-A06B; TCGA-AX-A06J; TCGA-AX-A06L; TCGA-AX-A0IS; TCGA-AX-A0J0; TCGA-AX-A0J1; TCGA-AX-A1C4; TCGA-AX-A1C5; TCGA-AX-A1C9; TCGA-AX-A1CE; TCGA-AX-A1CF; TCGA-AX-A1CI; TCGA-AX-A1CJ; TCGA-AX-A1CK; TCGA-AX-A2H7; TCGA-AX-A2H8; TCGA-AX-A2HA; TCGA-AX-A2HC; TCGA-AX-A2HD; TCGA-AX-A2HJ; TCGA-AX-A2HK; TCGA-AX-A3FW; TCGA-AX-A3FX; TCGA-AX-A3G8; TCGA-AX-A3G9; TCGA-AX-A3GB; TCGA-B5-A0JR; TCGA-B5-A0JS; TCGA-B5-A0JT; TCGA-B5-A0JU; TCGA-B5-A0JV; TCGA-B5-A0JX; TCGA-B5-A0JY; TCGA-B5-A0JZ; TCGA-B5-A0K0; TCGA-B5-A0K1; TCGA-B5-A0K2; TCGA-B5-A0K3; TCGA-B5-A0K4; TCGA-B5-A0K6; TCGA-B5-A0K7; TCGA-B5-A0K9; TCGA-B5-A0KB; TCGA-B5-A11E; TCGA-B5-A11F; TCGA-B5-A11G; TCGA-B5-A11I; TCGA-B5-A11J; TCGA-B5-A11M; TCGA-B5-A11N; TCGA-B5-A11O; TCGA-B5-A11P; TCGA-B5-A11Q; TCGA-B5-A11R; TCGA-B5-A11S; TCGA-B5-A11U; TCGA-B5-A11V; TCGA-B5-A11W; TCGA-B5-A11Y; TCGA-B5-A11Z; TCGA-B5-A121; TCGA-B5-A1MV; TCGA-B5-A1MW; TCGA-B5-A1MX; TCGA-B5-A1MZ; TCGA-B5-A3F9; TCGA-B5-A3FB; TCGA-B5-A3FD; TCGA-B5-A3FH; TCGA-B5-A5OC; TCGA-BG-A0LW; TCGA-BG-A0LX; TCGA-BG-A0M0; TCGA-BG-A0M2; TCGA-BG-A0M3; TCGA-BG-A0M4; TCGA-BG-A0M7; TCGA-BG-A0M9; TCGA-BG-A0MA; TCGA-BG-A0MC; TCGA-BG-A0MG; TCGA-BG-A0MI; TCGA-BG-A0MK; TCGA-BG-A0MO; TCGA-BG-A0MQ; TCGA-BG-A0MS; TCGA-BG-A0MT; TCGA-BG-A0MU; TCGA-BG-A0RY; TCGA-BG-A0VT; TCGA-BG-A0VV; TCGA-BG-A0VW; TCGA-BG-A0VX; TCGA-BG-A0VZ; TCGA-BG-A0W1; TCGA-BG-A0W2; TCGA-BG-A0YU; TCGA-BG-A186; TCGA-BG-A187; TCGA-BG-A18A; TCGA-BG-A18B; TCGA-BG-A18C; TCGA-BG-A220; TCGA-BG-A222; TCGA-BG-A2AD; TCGA-BG-A2AE; TCGA-BG-A2L7; TCGA-BG-A3EW; TCGA-BK-A0C9; TCGA-BK-A0CB; TCGA-BK-A13B; TCGA-BK-A13C; TCGA-BK-A4ZD; TCGA-BK-A56F; TCGA-BK-A6W3; TCGA-BK-A6W4; TCGA-BS-A0TA; TCGA-BS-A0TC; TCGA-BS-A0TD; TCGA-BS-A0TI; TCGA-BS-

A0TJ; TCGA-BS-A0U5; TCGA-BS-A0U7; TCGA-BS-A0U8; TCGA-BS-A0UA; TCGA-BS-A0UJ; TCGA-BS-A0UL; TCGA-BS-A0UM; TCGA-BS-A0UT; TCGA-BS-A0UV; TCGA-BS-A0V4; TCGA-BS-A0V6; TCGA-BS-A0V7; TCGA-BS-A0V8; TCGA-BS-A0VI; TCGA-BS-A0WQ; TCGA-D1-A0ZN; TCGA-D1-A0ZO; TCGA-D1-A0ZQ; TCGA-D1-A0ZR; TCGA-D1-A0ZU; TCGA-D1-A0ZV; TCGA-D1-A101; TCGA-D1-A102; TCGA-D1-A103; TCGA-D1-A15W; TCGA-D1-A15Z; TCGA-D1-A160; TCGA-D1-A161; TCGA-D1-A162; TCGA-D1-A163; TCGA-D1-A165; TCGA-D1-A167; TCGA-D1-A169; TCGA-D1-A16B; TCGA-D1-A16D; TCGA-D1-A16E; TCGA-D1-A16N; TCGA-D1-A16O; TCGA-D1-A16Q; TCGA-D1-A16R; TCGA-D1-A16V; TCGA-D1-A16X; TCGA-D1-A174; TCGA-D1-A175; TCGA-D1-A176; TCGA-D1-A177; TCGA-D1-A17A; TCGA-D1-A17B; TCGA-D1-A17C; TCGA-D1-A17D; TCGA-D1-A17F; TCGA-D1-A17H; TCGA-D1-A17K; TCGA-D1-A17L; TCGA-D1-A17M; TCGA-D1-A17N; TCGA-D1-A17Q; TCGA-D1-A17R; TCGA-D1-A17S; TCGA-D1-A17T; TCGA-D1-A17U; TCGA-D1-A1NS; TCGA-D1-A1NY; TCGA-D1-A1NZ; TCGA-D1-A1O0; TCGA-D1-A1O5; TCGA-D1-A1O7; TCGA-D1-A2G5; TCGA-D1-A3DA; TCGA-D1-A3DG; TCGA-DF-A2KN; TCGA-DF-A2KR; TCGA-DF-A2KS; TCGA-DF-A2KV; TCGA-DF-A2KZ; TCGA-DF-A2L0; TCGA-DI-A0WH; TCGA-DI-A1BY; TCGA-DI-A1C3; TCGA-DI-A1NO; TCGA-DI-A2QU; TCGA-E6-A1LX; TCGA-E6-A1M0; TCGA-E6-A2P9; TCGA-EC-A1NJ; TCGA-EC-A1QX; TCGA-EC-A24G; TCGA-EO-A1Y7; TCGA-EO-A22R; TCGA-EO-A22S; TCGA-EO-A22T; TCGA-EO-A22U; TCGA-EO-A22X; TCGA-EO-A22Y; TCGA-EO-A3AS; TCGA-EO-A3AU; TCGA-EO-A3AY; TCGA-EO-A3B0; TCGA-EO-A3KX; TCGA-EO-A3L0; TCGA-EY-A1G8; TCGA-EY-A1GD; TCGA-EY-A1GE; TCGA-EY-A1GF; TCGA-EY-A1GH; TCGA-EY-A1GI; TCGA-EY-A1GK; TCGA-EY-A1GL; TCGA-EY-A1GQ; TCGA-EY-A1GR; TCGA-EY-A1GT; TCGA-EY-A1GU; TCGA-EY-A1GW; TCGA-EY-A1GX; TCGA-EY-A1H0; TCGA-EY-A214; TCGA-EY-A215; TCGA-EY-A2OM; TCGA-EY-A2OP; TCGA-EY-A2OQ; TCGA-EY-A547; TCGA-EY-A548; TCGA-EY-A549; TCGA-EY-A54A; TCGA-EY-A5W2; TCGA-EY-A72D; TCGA-FI-A2D0; TCGA-FI-A2D4; TCGA-FI-A2D5; TCGA-FI-A2F4; TCGA-H5-A2HR; TCGA-PG-A916; TCGA-PG-A917; TCGA-QF-A5YT; TCGA-QS-A5YQ; TCGA-QS-A744; TCGA-SJ-A6ZI; TCGA-SJ-A6ZJ; TCGA-SL-A6JA.

Anexo 3: Listado de los biomarcadores de los *clusters* transcriptómicos.

Cluster transcriptómico 1

CFLAR; IBTK; RPUSD1; BAZ1B; MBTPS2; PREX2; MPHOSPH9; ZFR; WNK1; NCKAP1; WAPL; ZBTB11; EXOC5; SMC1A; MRPS34; SNRPA; ITCH; SENP1; CRYBG3; GSK3B; MRPL28; HECTD1; EMC9; ADNP; E2F1; PYCR3; SMC3; UBE2S; CAND1; NUP155; STAU1; USP9X; TIMM17B; PTPRB; PUS7L; NSUN5; ADRM1; CLOCK; SRPK2; TBRG4; SCAF11; NEDD1; SBNO1; RGL1; ATM; DIXDC1; DST; UTRN; AHCTF1; ROBO3; ABI3BP; APOOL; RHPN1; VPS28; TONSL; RECQL4; PPP1R16A; TAMALIN; TEDC2; OXER1; GMPS; ZNF148; NAA15; TMEM184C; PPP1R14A; MAP3K2; CERS6; MCRIP2; RSRC1; CKAP5; EXOSC4; CYC1; PTPN11; SHARPIN; MED14; FBXL6; TSHZ2; CEP97; PROS1; XPOT; AP3M1; HSF1; BRCC3; SLC52A2; PRR5; CYHR1; ARID2; FAT4; HELZ; NCOA6; RASSF9; OPA1; DMD; BMPR2; IPO7; ZNF611; LCAT; AL035458.1; EIF6; BOP 1,00.

Cluster transcriptómico 2

PHTF2; EFCAB1; DNAH5; DCUN1D1; HLTF; AFF4; PCNP; ERGIC2; RFX2; OSBPL8; SPEF1; EEA1; KRR1; RSPH4A; MAK; RNGTT; UBE3A; STRN; CEBPZ; MFN2; RO60; CFAP94; BBOF1; PANK3; UFM1; DNAI1; NCOA3; GLO1; CCNT1; PRKAA1; SCYL2; WASHC4; WDR38; FBXO11; FHAD1; SDE2; LNPK; HAUS6; NPAT; C2orf50; SRFBP1;

USP12; CAPSL; FAM81B; GABPA; DCK; PHF6; DZIP1L; TPPP3; RSPH1; LARP4; ZYG11B; TTLL10; USP1; BROX; TEKT4; CIP2A; HPS3; DNAJB14; ABCE1; ZNF474; LMBRD2; UBLCP1; C7orf57; RAD21; C9orf24; TMX3; VWA3B; ZBBX; TEFM; MANEA; SP3; AGR3; PIFO; FBXO45; WDR49; DNAH12; RSRC1; LYSMD3; MAP3K19; ANO6; PTPN11; KLHL28; ZDHHC20; AIDA; AKAP14; ZFP91; CFAP54; ZNF675; ZNF770; CHAMP1; OPA1; MZT1; C5orf51; FNIP1; CROCC2; ANKRD44-AS1; FAM133B; TUBA4B; AC007906.2.

Anexo 4: Listado de los pacientes en cada *cluster* epigenómico.

Cluster epigenómico 1

TCGA-A1-A0SD; TCGA-A2-A04N; TCGA-A2-A04V; TCGA-A2-A04X; TCGA-A2-A04Y; TCGA-A2-A0CP; TCGA-A2-A0CQ; TCGA-A2-A0CS; TCGA-A2-A0CU; TCGA-A2-A0CV; TCGA-A2-A0CW; TCGA-A2-A0D3; TCGA-A2-A0D4; TCGA-A2-A0EM; TCGA-A2-A0EO; TCGA-A2-A0ER; TCGA-A2-A0ES; TCGA-A2-A0ET; TCGA-A2-A0EV; TCGA-A2-A0EW; TCGA-A2-A0EX; TCGA-A7-A0CD; TCGA-A7-A0CJ; TCGA-A8-A06P; TCGA-A8-A06Q; TCGA-A8-A06T; TCGA-A8-A06U; TCGA-A8-A06Y; TCGA-A8-A06Z; TCGA-A8-A076; TCGA-A8-A079; TCGA-A8-A07B; TCGA-A8-A07E; TCGA-A8-A07F; TCGA-A8-A07G; TCGA-A8-A07J; TCGA-A8-A07L; TCGA-A8-A07P; TCGA-A8-A07W; TCGA-A8-A07Z; TCGA-A8-A081; TCGA-A8-A083; TCGA-A8-A086; TCGA-A8-A08A; TCGA-A8-A08C; TCGA-A8-A08F; TCGA-A8-A08G; TCGA-A8-A08H; TCGA-A8-A08I; TCGA-A8-A08P; TCGA-A8-A08S; TCGA-A8-A08T; TCGA-A8-A08Z; TCGA-A8-A090; TCGA-A8-A092; TCGA-A8-A093; TCGA-A8-A095; TCGA-A8-A096; TCGA-A8-A097; TCGA-A8-A099; TCGA-A8-A09A; TCGA-A8-A09B; TCGA-A8-A09C; TCGA-A8-A09D; TCGA-A8-A09E; TCGA-A8-A09I; TCGA-A8-A09K; TCGA-A8-A09M; TCGA-A8-A09N; TCGA-A8-A09Q; TCGA-A8-A09R; TCGA-A8-A09T; TCGA-A8-A09V; TCGA-A8-A09W; TCGA-A8-A0A1; TCGA-A8-A0A2; TCGA-A8-A0A4; TCGA-A8-A0A9; TCGA-A8-A0AB; TCGA-AN-A03Y; TCGA-AN-A046; TCGA-AN-A049; TCGA-AN-A04A; TCGA-AN-A0AJ; TCGA-AN-A0FD; TCGA-AN-A0FK; TCGA-AN-A0FN; TCGA-AN-A0FW; TCGA-AN-A0FY; TCGA-AO-A03O; TCGA-AO-A03P; TCGA-AO-A03R; TCGA-AO-A03T; TCGA-AO-A03V; TCGA-AO-A0J8; TCGA-AO-A0J9; TCGA-AO-A12A; TCGA-B6-A0I5; TCGA-B6-A0I9; TCGA-B6-A0IB; TCGA-B6-A0IG; TCGA-B6-A0IO; TCGA-B6-A0IP; TCGA-B6-A0RH; TCGA-B6-A0RQ; TCGA-B6-A0WS; TCGA-B6-A0X5; TCGA-BH-A0AY; TCGA-BH-A0B0; TCGA-BH-A0BD; TCGA-BH-A0BO; TCGA-BH-A0BP; TCGA-BH-A0BQ; TCGA-BH-A0BR; TCGA-BH-A0BV; TCGA-BH-A0C1; TCGA-BH-A0DE; TCGA-BH-A0DO; TCGA-BH-A0DT; TCGA-BH-A0DX; TCGA-BH-A0DZ; TCGA-BH-A0E7; TCGA-BH-A0E9; TCGA-BH-A0EA; TCGA-BH-A0EB; TCGA-BH-A0EI; TCGA-BH-A0HO; TCGA-BH-A0HQ; TCGA-BH-A0HU; TCGA-BH-A0W7; TCGA-BH-A18F; TCGA-BH-A18I; TCGA-BH-A18J; TCGA-BH-A18K; TCGA-BH-A18L; TCGA-BH-A18M; TCGA-BH-A18N; TCGA-C8-A12N; TCGA-C8-A12O; TCGA-C8-A12T; TCGA-C8-A12U; TCGA-C8-A12W; TCGA-C8-A130; TCGA-C8-A132; TCGA-D8-A13Y; TCGA-D8-A140; TCGA-D8-A141; TCGA-D8-A146; TCGA-E2-A10A; TCGA-E2-A14O; TCGA-E2-A14Q; TCGA-E2-A14S; TCGA-E2-A14T; TCGA-E2-A14V; TCGA-E2-A14Z; TCGA-E2-A153; TCGA-E2-A154; TCGA-E2-A156; TCGA-E2-A15A; TCGA-E2-A15C; TCGA-E2-A15D; TCGA-E2-A15E; TCGA-E2-A15F; TCGA-E2-A15H; TCGA-E2-A15L; TCGA-E2-A15M; TCGA-E2-A15O; TCGA-E2-A15P; TCGA-E2-A15T.

Cluster epigenómico 2

TCGA-3C-AAAU; TCGA-3C-AALJ; TCGA-3C-AALK; TCGA-4H-AAAK; TCGA-5L-AAT0; TCGA-5L-AAT1; TCGA-A1-A0SF; TCGA-A1-A0SI; TCGA-A1-A0SJ; TCGA-A1-A0SN; TCGA-A2-A04R; TCGA-A2-A0CK; TCGA-A2-A0CO; TCGA-A2-A0EN; TCGA-A2-A0EU; TCGA-A2-A0SU; TCGA-A2-A0SV; TCGA-A2-A0SY; TCGA-A2-A0T3; TCGA-A2-A0T4; TCGA-A2-A0T5; TCGA-A2-A0T6; TCGA-A2-A0T7; TCGA-A2-A0YC; TCGA-A2-A0YD; TCGA-A2-A0YG; TCGA-A2-A0YH; TCGA-A2-A0YI; TCGA-A2-A0YL; TCGA-A2-A1FV; TCGA-A2-A1FX; TCGA-A2-A1FZ; TCGA-A2-A1G4; TCGA-A2-A259; TCGA-A2-A25A; TCGA-A2-A25B; TCGA-A2-A25C; TCGA-A2-A25E; TCGA-A2-A3KC; TCGA-A2-A4RW; TCGA-A2-A4RY; TCGA-A2-A4S2; TCGA-A2-A4S3; TCGA-A7-A13F; TCGA-A7-A13H; TCGA-A7-A2KD; TCGA-A7-A3J1; TCGA-A7-A425; TCGA-A7-A426; TCGA-A7-A4SB; TCGA-A7-A56D; TCGA-A7-A5ZW; TCGA-A7-A5ZX; TCGA-A7-A6VX; TCGA-A8-A06O; TCGA-A8-A082; TCGA-A8-A08O; TCGA-A8-A0A6; TCGA-AC-A23C; TCGA-AC-A23E; TCGA-AC-A23G; TCGA-AC-A2B8; TCGA-AC-A2BM; TCGA-AC-A2FB; TCGA-AC-A2FE; TCGA-AC-A2FF; TCGA-AC-A2FK; TCGA-AC-A3BB; TCGA-AC-A3HN; TCGA-AC-A3QP; TCGA-AC-A3QQ; TCGA-AC-A3TM; TCGA-AC-A3TN; TCGA-AC-A3W6; TCGA-AC-A3YI; TCGA-AC-A4ZE; TCGA-AC-A5XS; TCGA-AC-A5XU; TCGA-AC-A62Y; TCGA-AC-A6IV; TCGA-AC-A6IX; TCGA-AC-A6NO; TCGA-AC-A8OP; TCGA-AC-A8OS; TCGA-AN-A03X; TCGA-AN-A0FF; TCGA-AN-A0XL; TCGA-AN-A0XP; TCGA-AN-A0XV; TCGA-AN-A0XW; TCGA-AO-A03L; TCGA-AO-A03M; TCGA-AO-A03N; TCGA-AO-A0JA; TCGA-AO-A0JB; TCGA-AO-A0JD; TCGA-AO-A0JF; TCGA-AO-A0JJ; TCGA-AO-A0JM; TCGA-AO-A126; TCGA-AO-A12B; TCGA-AO-A12E; TCGA-AO-A1KO; TCGA-AO-A1KP; TCGA-AO-A1KS; TCGA-AO-A1KT; TCGA-AQ-A04H; TCGA-AQ-A0Y5; TCGA-AQ-A1H2; TCGA-AQ-A1H3; TCGA-AR-A0TR; TCGA-AR-A0TV; TCGA-AR-A0TW; TCGA-AR-A0TZ; TCGA-AR-A0U2; TCGA-AR-A0U3; TCGA-AR-A1AK; TCGA-AR-A1AL; TCGA-AR-A1AM; TCGA-AR-A1AN; TCGA-AR-A1AP; TCGA-AR-A1AS; TCGA-AR-A1AU; TCGA-AR-A1AX; TCGA-AR-A24H; TCGA-AR-A24K; TCGA-AR-A24L; TCGA-AR-A24N; TCGA-AR-A24P; TCGA-AR-A24R; TCGA-AR-A24S; TCGA-AR-A24T; TCGA-AR-A24V; TCGA-AR-A24W; TCGA-AR-A24Z; TCGA-AR-A254; TCGA-AR-A255; TCGA-AR-A2LJ; TCGA-AR-A2LK; TCGA-AR-A2LL; TCGA-AR-A2LM; TCGA-AR-A2LN; TCGA-AR-A2LO; TCGA-AR-A2LQ; TCGA-AR-A5QM; TCGA-AR-A5QN; TCGA-AR-A5QP; TCGA-B6-A0RI; TCGA-B6-A0RL; TCGA-B6-A0RM; TCGA-B6-A0RO; TCGA-B6-A0WT; TCGA-B6-A0WV; TCGA-B6-A0WW; TCGA-B6-A0WZ; TCGA-B6-A0X4; TCGA-B6-A1KI; TCGA-B6-A2IU; TCGA-B6-A401; TCGA-B6-A40B; TCGA-B6-A40C; TCGA-BH-A0AU; TCGA-BH-A0AZ; TCGA-BH-A0B5; TCGA-BH-A0B6; TCGA-BH-A0BC; TCGA-BH-A0BJ; TCGA-BH-A0BS; TCGA-BH-A0BT; TCGA-BH-A0BZ; TCGA-BH-A0C0; TCGA-BH-A0DH; TCGA-BH-A0DI; TCGA-BH-A0DK; TCGA-BH-A0DP; TCGA-BH-A0DQ; TCGA-BH-A0DS; TCGA-BH-A0DV; TCGA-BH-A0E1; TCGA-BH-A0E2; TCGA-BH-A0GY; TCGA-BH-A0GZ; TCGA-BH-A0H0; TCGA-BH-A0H3; TCGA-BH-A0H6; TCGA-BH-A0H7; TCGA-BH-A0H9; TCGA-BH-A0HA; TCGA-BH-A0HB; TCGA-BH-A0HI; TCGA-BH-A0HX; TCGA-BH-A0W3; TCGA-BH-A0W4; TCGA-BH-A0W5; TCGA-BH-A1ES; TCGA-BH-A1ET; TCGA-BH-A1EU; TCGA-BH-A1EV; TCGA-BH-A1EX; TCGA-BH-A1EY; TCGA-BH-A1F2; TCGA-BH-A1F5; TCGA-BH-A1F8; TCGA-BH-A1FB; TCGA-BH-A1FD; TCGA-BH-A1FE; TCGA-BH-A1FG; TCGA-BH-A1FL; TCGA-BH-A1FN; TCGA-BH-A201; TCGA-BH-A202; TCGA-BH-A28O; TCGA-BH-A28Q; TCGA-BH-A2L8; TCGA-BH-A42T; TCGA-BH-A42V; TCGA-BH-A5J0; TCGA-BH-A6R8; TCGA-BH-A8FY; TCGA-BH-A8FZ; TCGA-BH-A8G0; TCGA-BH-AB28; TCGA-C8-A1HE; TCGA-C8-A1HI; TCGA-C8-A1HM; TCGA-C8-A1HN; TCGA-C8-A1HO; TCGA-C8-A26V; TCGA-C8-A26W; TCGA-C8-A26Z; TCGA-C8-A273; TCGA-C8-A274; TCGA-C8-A27A; TCGA-C8-A3M8; TCGA-D8-A1J8; TCGA-D8-A1JB; TCGA-D8-A1JC; TCGA-D8-A1JD; TCGA-D8-A1JE; TCGA-D8-A1JH; TCGA-D8-A1JI; TCGA-D8-A1JJ; TCGA-D8-A1JN; TCGA-D8-A1JP; TCGA-D8-A1JU; TCGA-D8-A1X5; TCGA-D8-A1X6; TCGA-D8-A1X8; TCGA-D8-A1X9; TCGA-D8-A1XA; TCGA-D8-A1XB; TCGA-D8-A1XD; TCGA-D8-A1XF; TCGA-D8-A1XL; TCGA-D8-A1XM;

TCGA-D8-A1XO; TCGA-D8-A1XR; TCGA-D8-A1XU; TCGA-D8-A1XY; TCGA-D8-A1Y0; TCGA-D8-A1Y1; TCGA-D8-A1Y2; TCGA-D8-A1Y3; TCGA-D8-A27G; TCGA-D8-A27I; TCGA-D8-A27K; TCGA-D8-A27L; TCGA-D8-A27N; TCGA-D8-A27P; TCGA-D8-A27R; TCGA-D8-A27T; TCGA-D8-A27V; TCGA-D8-A3Z5; TCGA-D8-A3Z6; TCGA-D8-A4Z1; TCGA-E2-A105; TCGA-E2-A108; TCGA-E2-A10B; TCGA-E2-A10C; TCGA-E2-A10E; TCGA-E2-A10F; TCGA-E2-A14U; TCGA-E2-A15I; TCGA-E2-A15J; TCGA-E2-A15K; TCGA-E2-A15R; TCGA-E2-A1B1; TCGA-E2-A1B4; TCGA-E2-A1B5; TCGA-E2-A1BC; TCGA-E2-A1BD; TCGA-E2-A1IE; TCGA-E2-A1IF; TCGA-E2-A1IG; TCGA-E2-A1IH; TCGA-E2-A1IJ; TCGA-E2-A1IK; TCGA-E2-A1IL; TCGA-E2-A1IN; TCGA-E2-A1IO; TCGA-E2-A1IU; TCGA-E2-A1L6; TCGA-E2-A1L8; TCGA-E2-A1L9; TCGA-E2-A2P5; TCGA-E2-A2P6; TCGA-E2-A3DX; TCGA-E2-A56Z; TCGA-E2-A570; TCGA-E2-A576; TCGA-E9-A1N6; TCGA-E9-A1NE; TCGA-E9-A1NG; TCGA-E9-A1NH; TCGA-E9-A227; TCGA-E9-A22D; TCGA-E9-A22E; TCGA-E9-A22H; TCGA-E9-A295; TCGA-E9-A3Q9; TCGA-E9-A3X8; TCGA-E9-A54Y; TCGA-E9-A5FK; TCGA-E9-A6HE; TCGA-EW-A1IW; TCGA-EW-A1IY; TCGA-EW-A1J1; TCGA-EW-A1J2; TCGA-EW-A1J3; TCGA-EW-A1J5; TCGA-EW-A1J6; TCGA-EW-A1OY; TCGA-EW-A1P3; TCGA-EW-A1P5; TCGA-EW-A1P6; TCGA-EW-A1PA; TCGA-EW-A1PC; TCGA-EW-A1PE; TCGA-EW-A1PF; TCGA-EW-A1PG; TCGA-EW-A423; TCGA-EW-A6S9; TCGA-EW-A6SC; TCGA-GI-A2C8; TCGA-GM-A2D9; TCGA-GM-A2DA; TCGA-GM-A2DC; TCGA-GM-A2DL; TCGA-GM-A2DM; TCGA-GM-A3NY; TCGA-GM-A3XG; TCGA-GM-A3XN; TCGA-GM-A4E0; TCGA-GM-A5PV; TCGA-GM-A5PX; TCGA-HN-A2OB; TCGA-JL-A3YX; TCGA-LD-A66U; TCGA-LD-A7W5; TCGA-LL-A440; TCGA-LL-A50Y; TCGA-LL-A5YM; TCGA-LL-A5YN; TCGA-LL-A6FQ; TCGA-LL-A73Z; TCGA-LL-A7T0; TCGA-LL-A9Q3; TCGA-LQ-A4E4; TCGA-OK-A5Q2; TCGA-OL-A5D8; TCGA-OL-A5DA; TCGA-OL-A5RV; TCGA-OL-A5RX; TCGA-OL-A66J; TCGA-OL-A66K; TCGA-OL-A66L; TCGA-OL-A66O; TCGA-OL-A6VQ; TCGA-OL-A6VR; TCGA-PE-A5DC; TCGA-PE-A5DE; TCGA-S3-A6ZF; TCGA-S3-A6ZH; TCGA-S3-AA0Z; TCGA-S3-AA11; TCGA-S3-AA14; TCGA-S3-AA17; TCGA-UL-AAZ6; TCGA-V7-A7HQ; TCGA-W8-A86G; TCGA-WT-AB41; TCGA-WT-AB44; TCGA-XX-A899; TCGA-XX-A89A; TCGA-Z7-A8R5; TCGA-Z7-A8R6.

***Cluster* epigenómico 3**

TCGA-2E-A9G8; TCGA-4E-A92E; TCGA-5B-A90C; TCGA-5S-A9Q8; TCGA-A5-A0G9; TCGA-A5-A0GA; TCGA-A5-A0GB; TCGA-A5-A0GD; TCGA-A5-A0GE; TCGA-A5-A0GG; TCGA-A5-A0GJ; TCGA-A5-A0GM; TCGA-A5-A0GN; TCGA-A5-A0GP; TCGA-A5-A0GQ; TCGA-A5-A0GU; TCGA-A5-A0GV; TCGA-A5-A0GX; TCGA-A5-A0R7; TCGA-A5-A0R8; TCGA-A5-A0R9; TCGA-A5-A0RA; TCGA-A5-A0VP; TCGA-A5-A0VQ; TCGA-A5-A1OJ; TCGA-A5-A1OK; TCGA-A5-A2K5; TCGA-A5-A2K7; TCGA-A5-A3LO; TCGA-A5-A7WJ; TCGA-A5-AB3J; TCGA-AJ-A2QK; TCGA-AJ-A2QL; TCGA-AJ-A2QN; TCGA-AJ-A2QO; TCGA-AJ-A3BH; TCGA-AJ-A3BI; TCGA-AJ-A3BK; TCGA-AJ-A3EK; TCGA-AJ-A3EL; TCGA-AJ-A3EM; TCGA-AJ-A3I9; TCGA-AJ-A3NC; TCGA-AJ-A3NE; TCGA-AJ-A3OJ; TCGA-AJ-A3OL; TCGA-AJ-A5DV; TCGA-AJ-A5DW; TCGA-AJ-A6NU; TCGA-AJ-A8CT; TCGA-AJ-A8CV; TCGA-AJ-A8CW; TCGA-AP-A051; TCGA-AP-A053; TCGA-AP-A056; TCGA-AP-A05N; TCGA-AP-A05O; TCGA-AP-A05P; TCGA-AP-A0LE; TCGA-AP-A0LG; TCGA-AP-A0LJ; TCGA-AP-A0LL; TCGA-AP-A0LM; TCGA-AP-A0LN; TCGA-AP-A0LO; TCGA-AP-A0LP; TCGA-AP-A0LS; TCGA-AP-A0LT; TCGA-AP-A0LV; TCGA-AP-A1DH; TCGA-AP-A1DK; TCGA-AP-A1DM; TCGA-AP-A1DO; TCGA-AP-A1DP; TCGA-AP-A1DR; TCGA-AP-A1DV; TCGA-AP-A1E0; TCGA-AP-A1E1; TCGA-AP-A1E3; TCGA-AW-A1PO; TCGA-AX-A05S; TCGA-AX-A05T; TCGA-AX-A05U; TCGA-AX-A05W; TCGA-AX-A05Y; TCGA-AX-A05Z; TCGA-AX-A060; TCGA-AX-A062; TCGA-AX-A06B; TCGA-AX-A06J; TCGA-AX-A06L; TCGA-AX-A0IS; TCGA-AX-A0J0; TCGA-AX-A0J1; TCGA-AX-A1C4;

TCGA-AX-A1C5; TCGA-AX-A1C9; TCGA-AX-A1CE; TCGA-AX-A1CF; TCGA-AX-A1CI; TCGA-AX-A1CJ; TCGA-AX-A1CK; TCGA-AX-A2H7; TCGA-AX-A2H8; TCGA-AX-A2HA; TCGA-AX-A2HC; TCGA-AX-A2HD; TCGA-AX-A2HJ; TCGA-AX-A2HK; TCGA-AX-A3FW; TCGA-AX-A3FX; TCGA-AX-A3G8; TCGA-AX-A3G9; TCGA-AX-A3GB; TCGA-B5-A0JR; TCGA-B5-A0JS; TCGA-B5-A0JT; TCGA-B5-A0JU; TCGA-B5-A0JV; TCGA-B5-A0JX; TCGA-B5-A0JY; TCGA-B5-A0JZ; TCGA-B5-A0K0; TCGA-B5-A0K1; TCGA-B5-A0K2; TCGA-B5-A0K3; TCGA-B5-A0K4; TCGA-B5-A0K6; TCGA-B5-A0K7; TCGA-B5-A0K9; TCGA-B5-A0KB; TCGA-B5-A11E; TCGA-B5-A11F; TCGA-B5-A11G; TCGA-B5-A11I; TCGA-B5-A11J; TCGA-B5-A11M; TCGA-B5-A11N; TCGA-B5-A11O; TCGA-B5-A11P; TCGA-B5-A11Q; TCGA-B5-A11R; TCGA-B5-A11S; TCGA-B5-A11U; TCGA-B5-A11V; TCGA-B5-A11W; TCGA-B5-A11Y; TCGA-B5-A11Z; TCGA-B5-A121; TCGA-B5-A1MV; TCGA-B5-A1MW; TCGA-B5-A1MX; TCGA-B5-A1MZ; TCGA-B5-A3F9; TCGA-B5-A3FB; TCGA-B5-A3FD; TCGA-B5-A3FH; TCGA-B5-A5OC; TCGA-BG-A0LW; TCGA-BG-A0LX; TCGA-BG-A0M0; TCGA-BG-A0M2; TCGA-BG-A0M3; TCGA-BG-A0M4; TCGA-BG-A0M7; TCGA-BG-A0M9; TCGA-BG-A0MA; TCGA-BG-A0MC; TCGA-BG-A0MG; TCGA-BG-A0MI; TCGA-BG-A0MK; TCGA-BG-A0MO; TCGA-BG-A0MQ; TCGA-BG-A0MS; TCGA-BG-A0MT; TCGA-BG-A0MU; TCGA-BG-A0RY; TCGA-BG-A0VT; TCGA-BG-A0VV; TCGA-BG-A0VW; TCGA-BG-A0VX; TCGA-BG-A0VZ; TCGA-BG-A0W1; TCGA-BG-A0W2; TCGA-BG-A0YU; TCGA-BG-A186; TCGA-BG-A187; TCGA-BG-A18A; TCGA-BG-A18B; TCGA-BG-A18C; TCGA-BG-A220; TCGA-BG-A222; TCGA-BG-A2AD; TCGA-BG-A2AE; TCGA-BG-A2L7; TCGA-BG-A3EW; TCGA-BK-A0C9; TCGA-BK-A0CB; TCGA-BK-A13B; TCGA-BK-A13C; TCGA-BK-A4ZD; TCGA-BK-A56F; TCGA-BK-A6W3; TCGA-BK-A6W4; TCGA-BS-A0TA; TCGA-BS-A0TC; TCGA-BS-A0TD; TCGA-BS-A0TI; TCGA-BS-A0TJ; TCGA-BS-A0U5; TCGA-BS-A0U7; TCGA-BS-A0U8; TCGA-BS-A0UA; TCGA-BS-A0UJ; TCGA-BS-A0UL; TCGA-BS-A0UM; TCGA-BS-A0UT; TCGA-BS-A0UV; TCGA-BS-A0V4; TCGA-BS-A0V6; TCGA-BS-A0V7; TCGA-BS-A0V8; TCGA-BS-A0VI; TCGA-BS-A0WQ; TCGA-D1-A0ZN; TCGA-D1-A0ZO; TCGA-D1-A0ZQ; TCGA-D1-A0ZR; TCGA-D1-A0ZU; TCGA-D1-A0ZV; TCGA-D1-A101; TCGA-D1-A102; TCGA-D1-A103; TCGA-D1-A15W; TCGA-D1-A15Z; TCGA-D1-A160; TCGA-D1-A161; TCGA-D1-A162; TCGA-D1-A163; TCGA-D1-A165; TCGA-D1-A167; TCGA-D1-A169; TCGA-D1-A16B; TCGA-D1-A16D; TCGA-D1-A16E; TCGA-D1-A16N; TCGA-D1-A16O; TCGA-D1-A16Q; TCGA-D1-A16R; TCGA-D1-A16V; TCGA-D1-A16X; TCGA-D1-A174; TCGA-D1-A175; TCGA-D1-A176; TCGA-D1-A177; TCGA-D1-A17A; TCGA-D1-A17B; TCGA-D1-A17C; TCGA-D1-A17D; TCGA-D1-A17F; TCGA-D1-A17H; TCGA-D1-A17K; TCGA-D1-A17L; TCGA-D1-A17M; TCGA-D1-A17N; TCGA-D1-A17Q; TCGA-D1-A17R; TCGA-D1-A17S; TCGA-D1-A17T; TCGA-D1-A17U; TCGA-D1-A1NS; TCGA-D1-A1NY; TCGA-D1-A1NZ; TCGA-D1-A1O0; TCGA-D1-A1O5; TCGA-D1-A1O7; TCGA-D1-A2G5; TCGA-D1-A3DA; TCGA-D1-A3DG; TCGA-DF-A2KN; TCGA-DF-A2KR; TCGA-DF-A2KS; TCGA-DF-A2KV; TCGA-DF-A2KZ; TCGA-DF-A2L0; TCGA-DI-A0WH; TCGA-DI-A1BY; TCGA-DI-A1C3; TCGA-DI-A1NO; TCGA-DI-A2QU; TCGA-E6-A1LX; TCGA-E6-A1M0; TCGA-E6-A2P9; TCGA-EC-A1NJ; TCGA-EC-A1QX; TCGA-EC-A24G; TCGA-EO-A1Y7; TCGA-EO-A22R; TCGA-EO-A22S; TCGA-EO-A22T; TCGA-EO-A22U; TCGA-EO-A22X; TCGA-EO-A22Y; TCGA-EO-A3AS; TCGA-EO-A3AU; TCGA-EO-A3AY; TCGA-EO-A3B0; TCGA-EO-A3KX; TCGA-EO-A3L0; TCGA-EY-A1G8; TCGA-EY-A1GD; TCGA-EY-A1GE; TCGA-EY-A1GF; TCGA-EY-A1GH; TCGA-EY-A1GI; TCGA-EY-A1GK; TCGA-EY-A1GL; TCGA-EY-A1GQ; TCGA-EY-A1GR; TCGA-EY-A1GT; TCGA-EY-A1GU; TCGA-EY-A1GW; TCGA-EY-A1GX; TCGA-EY-A1H0; TCGA-EY-A214; TCGA-EY-A215; TCGA-EY-A2OM; TCGA-EY-A2OP; TCGA-EY-A2OQ; TCGA-EY-A547; TCGA-EY-A548; TCGA-EY-A549; TCGA-EY-A54A; TCGA-EY-A5W2; TCGA-EY-A72D; TCGA-FI-A2D0; TCGA-FI-A2D4; TCGA-FI-A2D5; TCGA-FI-A2F4; TCGA-H5-A2HR; TCGA-PG-A916; TCGA-PG-A917; TCGA-QF-A5YT; TCGA-QS-A5YQ; TCGA-QS-A744; TCGA-SJ-A6ZI; TCGA-SJ-A6ZJ; TCGA-SL-A6JA.

Anexo 5: Listado de los biomarcadores de los *clusters* epigenómicos.

Cluster epigenómico 1

SOX8; SCN4A; LTF; BCAT1; SNCAIP; NUP133; LRP6; CNGB1; IGF2BP2; SNCB; ADD2; DLGAP4; MT3; NRP1; NEFH; PSMC6; DNTTIP1; ST8SIA5; ARHGEF7; GML; WNT2; GHRHR; FAM20A; CUX2; IL17A; EYA4; RASGRF2; HRH2; HRG; IL18R1; CCND2; RGSL1; CD244; DBH; WFDC3; CXCL6; CSN1S1; PRDM7; KIF1A; NPAS1; PKDREJ; RAB11FIP4; GPR12; TRAFD1; TM6SF1; DMRT1; PIM1; HCN4; SERBP1; FAM163A; GDF7; ALDH1L1; SLIT2; EPHA5; SNCA; CPLX2; A1CF; HTR7; TAF5; KCNA6; WNT7A; CLDN17; SORCS3; LYZL4; GART; SON; GFI1; NLRP3; PROK2; FABP1; SAMD3; GJB2; PRKCB; LPO; LALBA; KRT1; ZNF232; P2RY1; GTSF1; OR2K2; PRNP; FOXB1; ID4; ACOT12; PNLIP; EIF4E1B; B3GNT5; SOX11; GPR150; OR3A1; MFSD5; OR2V2; SLC35F3; SIX6; MIXL1; FOXI2; LILRB4; NCCRP1; COL15A1; MAGEL2.

Cluster epigenómico 2

RUNX3; KCNQ1; SLC9A3; CHAT; SPP2; CD5L; CACNG5; SLCO1A2; REM1; FETUB; APOH; MYL9; BRS3; SFRP1; FSD1; WNT2; CRHR2; MYH1; FGF6; IL17F; EYA4; OR12D3; SLC22A2; HRG; REG1A; CHRNB4; CCND2; IAPP; CD244; SIRPD; INS-IGF2; DPP6; DEFB118; MGAT1; TMEM204; APCS; SYT6; GJA10; EDNRB; GPNMB; GATA4; MAPK4; STAC2; KCNA10; FAM163A; HHIPL2; FLG; ACTA1; AHSG; A1CF; PTPRO; GRM1; GPR26; C1orf158; KIT; XDH; CD1B; CD1E; ZG16B; LMX1A; NEUROD1; DPPA2; NEUROD6; TMCO5A; GLB1L3; IGF2; KRT1; PTF1A; GBX2; KLK7; MUC17; SLC23A1; PKHD1; NMUR1; REG3A; ZNF80; FMR1NB; FCER1A; DEFB119; FAM89A; KCNIP1; PCBP3; GPR1; EMILIN3; ANKRD45; IQCF2; OR6A2; DLK1; GABRA5; SLC18A3; PPP3R2; OSTN; ADARB1; DPP4; AVPR1B; LTC4S; OR5V1; SIGLEC12; MYH4; SRD5A2.

Cluster epigenómico 3

ITGA2B; SLC4A8; RASGRF1; CCAR1; WNT8A; DDX20; ROGDI; RPL31; SLC35C2; MYBPC2; BLNK; FBXL19; CABP7; NAMPT; SFRP4; STIM2; ITFG2; LDHB; ADAM23; SLC35A3; RHOQ; RCL1; SCPEP1; ESX1; TTPAL; SOX21; ARMC7; TTF1; ALKBH7; VASP; UROD; HECTD3; PRDM12; ZNF141; CAPN7; BEX2; TSPAN2; MYCN; EPHA7; B4GALNT1; BIN1; DCUN1D5; SENP7; GSTCD; INTS12; TMEM19; CLTC; RXRG; OSR1; GULP1; MYO10; TCTE1; RNF144A; FLI1; KCNMA1; CACNA1D; ACAN; DPYSL5; PAFAH2; SV2A; CBR3; CCDC28B; RSPH1; DMKN; CADM3; CCDC138; ALPI; CRYBA2; ZMYM6; MYOZ3; FZD6; NDUFB6; ENDOG; CACNB3; RAB31; TM4SF4; CANT1; PRNP; PHF8; TRIB1; NMNAT1; STAT5B; TMEM81; CCDC85B; CTNNBIP1; RRS1; TPPP2; UBE2E2; TSKU; SPATA13; GBP6; GRIN2A; CMTM4; LIN9; PEG3; GPX1; MYCNOS; TRIM26; ASPRV1; DHRS11.

Anexo 6: Listado de los pacientes en cada *cluster* genómico.

Cluster genómico 1

TCGA-3C-AALK; TCGA-5L-AAT0; TCGA-5L-AAT1; TCGA-A1-A0SD; TCGA-A2-A04R; TCGA-A2-A04X; TCGA-A2-A0CO; TCGA-A2-A0EN; TCGA-A2-A0SY; TCGA-A2-A0T4; TCGA-A2-A0T6; TCGA-A2-A0YC; TCGA-A2-A0YL; TCGA-A2-A1FV; TCGA-A2-A25C;

TCGA-A2-A25E; TCGA-A2-A4RW; TCGA-A7-A3J1; TCGA-A7-A5ZX; TCGA-A8-A07E; TCGA-A8-A083; TCGA-A8-A09W; TCGA-A8-A0A1; TCGA-AC-A3HN; TCGA-AC-A3TN; TCGA-AC-A3W6; TCGA-AC-A6IV; TCGA-AC-A6IX; TCGA-AN-A046; TCGA-AO-A0JA; TCGA-AO-A0JB; TCGA-AO-A0JF; TCGA-AO-A0JJ; TCGA-AR-A1AM; TCGA-AR-A1AN; TCGA-AR-A1AP; TCGA-AR-A24L; TCGA-AR-A2LN; TCGA-B6-A0IP; TCGA-B6-A0RI; TCGA-B6-A0RL; TCGA-B6-A0RQ; TCGA-B6-A0WZ; TCGA-B6-A0X4; TCGA-B6-A40B; TCGA-BH-A0B0; TCGA-BH-A0BC; TCGA-BH-A0BO; TCGA-BH-A0BS; TCGA-BH-A0DK; TCGA-BH-A0DX; TCGA-BH-A0GZ; TCGA-BH-A0HI; TCGA-BH-A0HO; TCGA-BH-A0HU; TCGA-BH-A18J; TCGA-BH-A18M; TCGA-BH-A1EY; TCGA-BH-A201; TCGA-BH-A42T; TCGA-BH-A8FY; TCGA-BH-A8G0; TCGA-C8-A1HE; TCGA-C8-A1HI; TCGA-C8-A273; TCGA-D8-A1J8; TCGA-D8-A1JB; TCGA-D8-A1JN; TCGA-D8-A1JP; TCGA-D8-A1JU; TCGA-D8-A1X8; TCGA-D8-A1XM; TCGA-D8-A1XO; TCGA-D8-A1XU; TCGA-D8-A27I; TCGA-D8-A27N; TCGA-D8-A27P; TCGA-D8-A27T; TCGA-E2-A15L; TCGA-E2-A15T; TCGA-E2-A1IG; TCGA-E2-A1IL; TCGA-E9-A1N6; TCGA-E9-A227; TCGA-E9-A22H; TCGA-EW-A6SC; TCGA-GM-A4E0; TCGA-HN-A2OB; TCGA-LL-A5YM; TCGA-LQ-A4E4; TCGA-OL-A66L; TCGA-S3-AA11; TCGA-XX-A899; TCGA-4E-A92E; TCGA-5S-A9Q8; TCGA-A5-A0GB; TCGA-A5-A0GG; TCGA-A5-A0GJ; TCGA-A5-A0GU; TCGA-A5-A0GV; TCGA-A5-A0GX; TCGA-AJ-A3OL; TCGA-AJ-A6NU; TCGA-AP-A05O; TCGA-AP-A05P; TCGA-AP-A0LE; TCGA-AP-A0LO; TCGA-AP-A1DM; TCGA-AX-A05S; TCGA-AX-A05W; TCGA-AX-A1CI; TCGA-AX-A1CJ; TCGA-AX-A2H7; TCGA-B5-A0JT; TCGA-B5-A0JV; TCGA-B5-A0K0; TCGA-B5-A0K4; TCGA-B5-A0K6; TCGA-B5-A11P; TCGA-BG-A0LX; TCGA-BG-A0M2; TCGA-BG-A0M3; TCGA-BG-A0M4; TCGA-BG-A0MO; TCGA-BG-A0MQ; TCGA-BG-A0MS; TCGA-BG-A0MU; TCGA-BG-A0RY; TCGA-BG-A0VT; TCGA-BG-A0VV; TCGA-BG-A186; TCGA-BG-A222; TCGA-BK-A0C9; TCGA-BS-A0U7; TCGA-BS-A0UM; TCGA-BS-A0V4; TCGA-BS-A0VI; TCGA-D1-A0ZN; TCGA-D1-A0ZQ; TCGA-D1-A102; TCGA-D1-A16O; TCGA-D1-A16V; TCGA-D1-A177; TCGA-D1-A17R; TCGA-DF-A2KZ; TCGA-EC-A1QX; TCGA-EY-A1GF; TCGA-EY-A1GH; TCGA-EY-A1GK; TCGA-EY-A1GX; TCGA-EY-A2OP; TCGA-EY-A549; TCGA-EY-A5W2; TCGA-EY-A72D; TCGA-PG-A917; TCGA-QS-A744; TCGA-SL-A6JA.

***Cluster* genómico 2**

TCGA-A8-A07J; TCGA-A8-A08H; TCGA-A8-A0A4; TCGA-AC-A2FB; TCGA-AO-A1KO; TCGA-AR-A0TZ; TCGA-BH-A0BR; TCGA-BH-A0DE; TCGA-BH-A0E2; TCGA-BH-A0W4; TCGA-BH-A28O; TCGA-C8-A26V; TCGA-D8-A13Y; TCGA-D8-A1JH; TCGA-D8-A1X9; TCGA-E2-A14U; TCGA-E2-A1B4; TCGA-E2-A1B5; TCGA-E2-A1BC; TCGA-E2-A1IU; TCGA-E9-A1NG; TCGA-E9-A3Q9; TCGA-E9-A5FK; TCGA-EW-A1PF; TCGA-EW-A1PG; TCGA-GM-A2DA; TCGA-GM-A2DL; TCGA-GM-A2DM; TCGA-GM-A3XG; TCGA-GM-A5PV; TCGA-S3-AA14; TCGA-XX-A89A; TCGA-2E-A9G8; TCGA-A5-A0GD; TCGA-A5-A0GM; TCGA-A5-A0GQ; TCGA-A5-A0VQ; TCGA-A5-A2K5; TCGA-A5-A2K7; TCGA-AJ-A3BH; TCGA-AJ-A3I9; TCGA-AJ-A5DV; TCGA-AJ-A5DW; TCGA-AJ-A8CT; TCGA-AJ-A8CV; TCGA-AP-A053; TCGA-AP-A05N; TCGA-AP-A0LG; TCGA-AP-A0LM; TCGA-AP-A0LV; TCGA-AP-A1DK; TCGA-AP-A1DR; TCGA-AP-A1E0; TCGA-AX-A05Z; TCGA-AX-A062; TCGA-AX-A06B; TCGA-AX-A06L; TCGA-AX-A1C9; TCGA-AX-A2HD; TCGA-AX-A2HJ; TCGA-B5-A0JU; TCGA-B5-A0JY; TCGA-B5-A0K1; TCGA-B5-A0K3; TCGA-B5-A0KB; TCGA-B5-A11J; TCGA-B5-A11M; TCGA-B5-A11N; TCGA-B5-A11O; TCGA-B5-A11U; TCGA-B5-A1MX; TCGA-B5-A3FB; TCGA-BG-A0LW; TCGA-BG-A0MC; TCGA-BG-A0MI; TCGA-BG-A0VX; TCGA-BG-A18B; TCGA-BG-A18C; TCGA-BG-A220; TCGA-BG-A2AD; TCGA-BG-A2L7; TCGA-BK-A56F; TCGA-BK-A6W4; TCGA-BS-A0UA; TCGA-BS-A0UJ; TCGA-BS-A0UL; TCGA-D1-A0ZO; TCGA-D1-A0ZV; TCGA-D1-A101; TCGA-D1-A15W; TCGA-D1-A163; TCGA-D1-A167; TCGA-D1-A16N; TCGA-D1-A16Q; TCGA-D1-

A174; TCGA-D1-A175; TCGA-D1-A17D; TCGA-D1-A17F; TCGA-D1-A17H; TCGA-D1-A17K; TCGA-D1-A17Q; TCGA-D1-A17U; TCGA-D1-A1NS; TCGA-D1-A1O5; TCGA-D1-A3DG; TCGA-DF-A2KV; TCGA-DI-A1BY; TCGA-EC-A24G; TCGA-EO-A22R; TCGA-EO-A22X; TCGA-EO-A3AU; TCGA-EO-A3KX; TCGA-EY-A1GI; TCGA-EY-A1GL; TCGA-EY-A548; TCGA-FI-A2D0; TCGA-PG-A916; TCGA-QF-A5YT; TCGA-SJ-A6ZI.

***Cluster* genómico 3**

TCGA-3C-AAAU; TCGA-3C-AALJ; TCGA-4H-AAAK; TCGA-A1-A0SF; TCGA-A1-A0SJ; TCGA-A1-A0SN; TCGA-A2-A0CK; TCGA-A2-A0CS; TCGA-A2-A0CU; TCGA-A2-A0CV; TCGA-A2-A0CW; TCGA-A2-A0D3; TCGA-A2-A0D4; TCGA-A2-A0EO; TCGA-A2-A0ER; TCGA-A2-A0ET; TCGA-A2-A0EU; TCGA-A2-A0EV; TCGA-A2-A0EW; TCGA-A2-A0SU; TCGA-A2-A0T3; TCGA-A2-A0YG; TCGA-A2-A0YH; TCGA-A2-A0YI; TCGA-A2-A25A; TCGA-A2-A25B; TCGA-A7-A0CD; TCGA-A7-A0CJ; TCGA-A7-A13F; TCGA-A7-A2KD; TCGA-A7-A425; TCGA-A7-A426; TCGA-A7-A4SB; TCGA-A7-A56D; TCGA-A7-A6VX; TCGA-A8-A06P; TCGA-A8-A06Q; TCGA-A8-A06T; TCGA-A8-A06U; TCGA-A8-A06Z; TCGA-A8-A076; TCGA-A8-A079; TCGA-A8-A07B; TCGA-A8-A07F; TCGA-A8-A07W; TCGA-A8-A07Z; TCGA-A8-A081; TCGA-A8-A08F; TCGA-A8-A08G; TCGA-A8-A08I; TCGA-A8-A08O; TCGA-A8-A090; TCGA-A8-A092; TCGA-A8-A093; TCGA-A8-A095; TCGA-A8-A096; TCGA-A8-A097; TCGA-A8-A09A; TCGA-A8-A09B; TCGA-A8-A09C; TCGA-A8-A09E; TCGA-A8-A09I; TCGA-A8-A09K; TCGA-A8-A09M; TCGA-A8-A09N; TCGA-A8-A09Q; TCGA-A8-A09R; TCGA-A8-A09T; TCGA-A8-A0A2; TCGA-A8-A0A9; TCGA-A8-A0AB; TCGA-AC-A23E; TCGA-AC-A2B8; TCGA-AC-A2BM; TCGA-AC-A2FE; TCGA-AC-A3BB; TCGA-AC-A4ZE; TCGA-AC-A5XS; TCGA-AC-A5XU; TCGA-AC-A62Y; TCGA-AC-A8OP; TCGA-AC-A8OS; TCGA-AN-A03Y; TCGA-AN-A04A; TCGA-AN-A0AJ; TCGA-AN-A0FK; TCGA-AN-A0FW; TCGA-AN-A0FY; TCGA-AO-A03L; TCGA-AO-A03N; TCGA-AO-A03O; TCGA-AO-A03P; TCGA-AO-A03V; TCGA-AO-A0J9; TCGA-AO-A0JD; TCGA-AO-A0JM; TCGA-AO-A126; TCGA-AO-A12A; TCGA-AO-A12B; TCGA-AO-A12E; TCGA-AO-A1KS; TCGA-AO-A1KT; TCGA-AQ-A04H; TCGA-AQ-A0Y5; TCGA-AR-A0TR; TCGA-AR-A0TV; TCGA-AR-A0TW; TCGA-AR-A0U2; TCGA-AR-A24H; TCGA-AR-A24K; TCGA-AR-A24R; TCGA-AR-A24S; TCGA-AR-A24T; TCGA-AR-A24V; TCGA-AR-A24W; TCGA-AR-A24Z; TCGA-AR-A254; TCGA-AR-A255; TCGA-AR-A2LJ; TCGA-AR-A2LK; TCGA-AR-A2LL; TCGA-AR-A2LO; TCGA-AR-A5QM; TCGA-AR-A5QP; TCGA-B6-A0I9; TCGA-B6-A0IB; TCGA-B6-A0IO; TCGA-B6-A0RO; TCGA-B6-A0WV; TCGA-B6-A1KI; TCGA-B6-A401; TCGA-BH-A0AU; TCGA-BH-A0AY; TCGA-BH-A0B5; TCGA-BH-A0B6; TCGA-BH-A0BJ; TCGA-BH-A0BP; TCGA-BH-A0BT; TCGA-BH-A0BV; TCGA-BH-A0BZ; TCGA-BH-A0C0; TCGA-BH-A0C1; TCGA-BH-A0DP; TCGA-BH-A0DS; TCGA-BH-A0DZ; TCGA-BH-A0E9; TCGA-BH-A0EB; TCGA-BH-A0GY; TCGA-BH-A0H7; TCGA-BH-A0HX; TCGA-BH-A0W3; TCGA-BH-A0W5; TCGA-BH-A18I; TCGA-BH-A18K; TCGA-BH-A18L; TCGA-BH-A18N; TCGA-BH-A1ES; TCGA-BH-A1EU; TCGA-BH-A1EV; TCGA-BH-A1EX; TCGA-BH-A1F2; TCGA-BH-A1FN; TCGA-BH-A202; TCGA-BH-A2L8; TCGA-BH-A5J0; TCGA-BH-A8FZ; TCGA-C8-A12O; TCGA-C8-A12U; TCGA-C8-A12W; TCGA-C8-A130; TCGA-C8-A1HM; TCGA-C8-A1HN; TCGA-C8-A26Z; TCGA-C8-A274; TCGA-C8-A27A; TCGA-D8-A146; TCGA-D8-A1JD; TCGA-D8-A1JE; TCGA-D8-A1JI; TCGA-D8-A1JJ; TCGA-D8-A1X5; TCGA-D8-A1X6; TCGA-D8-A1XA; TCGA-D8-A1XF; TCGA-D8-A1XL; TCGA-D8-A1XR; TCGA-D8-A1XY; TCGA-D8-A1Y1; TCGA-D8-A1Y2; TCGA-D8-A1Y3; TCGA-D8-A27G; TCGA-D8-A27R; TCGA-D8-A27V; TCGA-E2-A10A; TCGA-E2-A10B; TCGA-E2-A10C; TCGA-E2-A10E; TCGA-E2-A14O; TCGA-E2-A14T; TCGA-E2-A14V; TCGA-E2-A154; TCGA-E2-A156; TCGA-E2-A15A; TCGA-E2-A15E; TCGA-E2-A15H; TCGA-E2-A15M; TCGA-E2-A1BD; TCGA-E2-A1IF; TCGA-E2-A1IH; TCGA-E2-A1IN; TCGA-E2-A1L6; TCGA-E2-A1L8; TCGA-E2-A2P5; TCGA-E2-A56Z; TCGA-E2-A570;

TCGA-E9-A22D; TCGA-E9-A22E; TCGA-E9-A295; TCGA-E9-A3X8; TCGA-E9-A54Y; TCGA-E9-A6HE; TCGA-EW-A1IW; TCGA-EW-A1J1; TCGA-EW-A1J6; TCGA-EW-A1OY; TCGA-EW-A1P5; TCGA-EW-A1PC; TCGA-EW-A1PE; TCGA-EW-A423; TCGA-EW-A6S9; TCGA-GI-A2C8; TCGA-JL-A3YX; TCGA-LL-A5YN; TCGA-LL-A6FQ; TCGA-LL-A7T0; TCGA-OL-A5DA; TCGA-OL-A5RX; TCGA-OL-A66K; TCGA-OL-A66O; TCGA-PE-A5DE; TCGA-S3-A6ZF; TCGA-S3-A6ZH; TCGA-S3-AA17; TCGA-UL-AAZ6; TCGA-Z7-A8R5; TCGA-5B-A90C; TCGA-A5-A0GE; TCGA-A5-A3LO; TCGA-AJ-A2QK; TCGA-AJ-A2QN; TCGA-AJ-A3BI; TCGA-AJ-A3EM; TCGA-AX-A05Y; TCGA-AX-A060; TCGA-AX-A06J; TCGA-AX-A1CF; TCGA-AX-A2H8; TCGA-AX-A2HK; TCGA-AX-A3FX; TCGA-B5-A11E; TCGA-B5-A11I; TCGA-B5-A11R; TCGA-B5-A11Y; TCGA-B5-A121; TCGA-B5-A3F9; TCGA-B5-A3FD; TCGA-BK-A4ZD; TCGA-BS-A0TI; TCGA-BS-A0V7; TCGA-BS-A0V8; TCGA-D1-A17B; TCGA-D1-A17M; TCGA-D1-A1O0; TCGA-DF-A2KR; TCGA-DF-A2KS; TCGA-DF-A2L0; TCGA-DI-A1NO; TCGA-EO-A1Y7; TCGA-EO-A22T; TCGA-EY-A1GR; TCGA-EY-A1GW; TCGA-EY-A214; TCGA-EY-A2OQ; TCGA-EY-A54A; TCGA-H5-A2HR.

Cluster genómico 4

TCGA-A2-A04V; TCGA-A2-A0ES; TCGA-A2-A1G4; TCGA-A2-A4RY; TCGA-A8-A07P; TCGA-A8-A082; TCGA-AC-A2FF; TCGA-AC-A2FK; TCGA-AC-A3YI; TCGA-AN-A0FD; TCGA-AN-A0FN; TCGA-AQ-A1H2; TCGA-AR-A0U3; TCGA-AR-A5QN; TCGA-B6-A0IG; TCGA-B6-A0WT; TCGA-BH-A0DI; TCGA-BH-A0DQ; TCGA-BH-A0EI; TCGA-BH-A0HA; TCGA-BH-A1ET; TCGA-BH-A1F8; TCGA-C8-A132; TCGA-C8-A1HO; TCGA-C8-A26W; TCGA-D8-A1XD; TCGA-D8-A4Z1; TCGA-E2-A105; TCGA-E2-A15F; TCGA-E2-A15I; TCGA-E2-A15P; TCGA-E2-A15R; TCGA-E2-A1B1; TCGA-E2-A1IE; TCGA-E2-A1IK; TCGA-E2-A1IO; TCGA-EW-A1PA; TCGA-LL-A440; TCGA-OL-A66J; TCGA-WT-AB41; TCGA-A5-A0GP; TCGA-A5-A0R8; TCGA-A5-A0RA; TCGA-A5-A0VP; TCGA-AJ-A2QL; TCGA-AJ-A2QO; TCGA-AJ-A3EK; TCGA-AJ-A3EL; TCGA-AJ-A3NE; TCGA-AJ-A3OJ; TCGA-AP-A056; TCGA-AP-A0LP; TCGA-AP-A0LS; TCGA-AP-A0LT; TCGA-AP-A1DH; TCGA-AP-A1DV; TCGA-AX-A0J0; TCGA-AX-A1C5; TCGA-AX-A1CE; TCGA-AX-A3FW; TCGA-AX-A3G9; TCGA-B5-A0JS; TCGA-B5-A0JZ; TCGA-B5-A0K2; TCGA-B5-A11G; TCGA-B5-A11W; TCGA-BG-A0M7; TCGA-BG-A0MK; TCGA-BG-A0VW; TCGA-BG-A0VZ; TCGA-BG-A0YU; TCGA-BK-A6W3; TCGA-BS-A0TA; TCGA-BS-A0TC; TCGA-BS-A0TD; TCGA-BS-A0TJ; TCGA-BS-A0U5; TCGA-BS-A0U8; TCGA-BS-A0V6; TCGA-D1-A103; TCGA-D1-A169; TCGA-D1-A16B; TCGA-D1-A16D; TCGA-D1-A16E; TCGA-D1-A16X; TCGA-D1-A17C; TCGA-D1-A17N; TCGA-D1-A17S; TCGA-D1-A1NY; TCGA-D1-A1O7; TCGA-DF-A2KN; TCGA-DI-A1C3; TCGA-E6-A1LX; TCGA-EC-A1NJ; TCGA-EO-A22S; TCGA-EO-A22U; TCGA-EO-A3AY; TCGA-EY-A1GE; TCGA-EY-A215; TCGA-FI-A2D5; TCGA-QS-A5YQ.

***Cluster* genómico 5**

TCGA-A2-A04N; TCGA-A2-A04Y; TCGA-A2-A0CP; TCGA-A2-A0EX; TCGA-A2-A0T7; TCGA-A2-A1FX; TCGA-A2-A259; TCGA-A2-A4S3; TCGA-A8-A06O; TCGA-A8-A07L; TCGA-A8-A086; TCGA-A8-A08S; TCGA-A8-A09D; TCGA-AC-A3QP; TCGA-AN-A049; TCGA-AN-A0XV; TCGA-AO-A03M; TCGA-AO-A03R; TCGA-AO-A03T; TCGA-AO-A1KP; TCGA-AR-A1AU; TCGA-AR-A1AX; TCGA-B6-A0RM; TCGA-B6-A0WS; TCGA-B6-A0X5; TCGA-BH-A0DH; TCGA-BH-A0E7; TCGA-BH-A0H6; TCGA-BH-A0HB; TCGA-BH-A0W7; TCGA-BH-A1F5; TCGA-BH-A1FB; TCGA-BH-A1FD; TCGA-BH-A1FE; TCGA-BH-A1FG; TCGA-BH-A42V; TCGA-C8-A12N; TCGA-C8-A12T; TCGA-D8-A1JC; TCGA-D8-A27L; TCGA-E2-A108; TCGA-E2-A14S; TCGA-E2-A14Z; TCGA-E2-A15C; TCGA-E2-A15O; TCGA-E2-A576; TCGA-EW-A1P6; TCGA-LD-A7W5; TCGA-OL-A6VR; TCGA-PE-A5DC;

TCGA-V7-A7HQ; TCGA-Z7-A8R6; TCGA-A5-A0R7; TCGA-A5-A0R9; TCGA-AJ-A3BK; TCGA-AP-A0LJ; TCGA-AP-A0LN; TCGA-AP-A1E3; TCGA-AW-A1PO; TCGA-AX-A05T; TCGA-AX-A05U; TCGA-AX-A1C4; TCGA-AX-A2HC; TCGA-B5-A0JX; TCGA-B5-A11F; TCGA-B5-A11V; TCGA-B5-A1MV; TCGA-B5-A1MW; TCGA-BG-A0M0; TCGA-BG-A0M9; TCGA-BG-A0MA; TCGA-BG-A0MG; TCGA-BG-A0W2; TCGA-BG-A187; TCGA-BG-A18A; TCGA-BG-A2AE; TCGA-BK-A0CB; TCGA-BS-A0UV; TCGA-BS-A0WQ; TCGA-D1-A0ZU; TCGA-D1-A15Z; TCGA-D1-A161; TCGA-D1-A165; TCGA-D1-A16R; TCGA-D1-A17T; TCGA-DI-A2QU; TCGA-E6-A1M0; TCGA-EO-A3AS; TCGA-EO-A3B0; TCGA-EO-A3L0; TCGA-EY-A1G8; TCGA-EY-A1H0; TCGA-EY-A2OM; TCGA-EY-A547; TCGA-FI-A2D4; TCGA-SJ-A6ZJ.

***Cluster* genómico 6**

TCGA-A1-A0SI; TCGA-A2-A0CQ; TCGA-A2-A0EM; TCGA-A2-A0SV; TCGA-A2-A0T5; TCGA-A2-A0YD; TCGA-A2-A1FZ; TCGA-A2-A3KC; TCGA-A2-A4S2; TCGA-A7-A13H; TCGA-A7-A5ZW; TCGA-A8-A06Y; TCGA-A8-A07G; TCGA-A8-A08A; TCGA-A8-A08C; TCGA-A8-A08P; TCGA-A8-A08T; TCGA-A8-A08Z; TCGA-A8-A099; TCGA-A8-A09V; TCGA-A8-A0A6; TCGA-AC-A23C; TCGA-AC-A23G; TCGA-AC-A3QQ; TCGA-AC-A3TM; TCGA-AC-A6NO; TCGA-AN-A03X; TCGA-AN-A0FF; TCGA-AN-A0XL; TCGA-AN-A0XP; TCGA-AN-A0XW; TCGA-AO-A0J8; TCGA-AQ-A1H3; TCGA-AR-A1AK; TCGA-AR-A1AL; TCGA-AR-A1AS; TCGA-AR-A24N; TCGA-AR-A24P; TCGA-AR-A2LM; TCGA-AR-A2LQ; TCGA-B6-A0I5; TCGA-B6-A0RH; TCGA-B6-A0WW; TCGA-B6-A2IU; TCGA-B6-A40C; TCGA-BH-A0AZ; TCGA-BH-A0BD; TCGA-BH-A0BQ; TCGA-BH-A0DO; TCGA-BH-A0DT; TCGA-BH-A0DV; TCGA-BH-A0E1; TCGA-BH-A0EA; TCGA-BH-A0H0; TCGA-BH-A0H3; TCGA-BH-A0H9; TCGA-BH-A0HQ; TCGA-BH-A18F; TCGA-BH-A1FL; TCGA-BH-A28Q; TCGA-BH-A6R8; TCGA-BH-AB28; TCGA-C8-A3M8; TCGA-D8-A140; TCGA-D8-A141; TCGA-D8-A1XB; TCGA-D8-A1Y0; TCGA-D8-A27K; TCGA-D8-A3Z5; TCGA-D8-A3Z6; TCGA-E2-A10F; TCGA-E2-A14Q; TCGA-E2-A153; TCGA-E2-A15D; TCGA-E2-A15J; TCGA-E2-A15K; TCGA-E2-A1IJ; TCGA-E2-A1L9; TCGA-E2-A2P6; TCGA-E2-A3DX; TCGA-E9-A1NE; TCGA-E9-A1NH; TCGA-EW-A1IY; TCGA-EW-A1J2; TCGA-EW-A1J3; TCGA-EW-A1J5; TCGA-EW-A1P3; TCGA-GM-A2D9; TCGA-GM-A2DC; TCGA-GM-A3NY; TCGA-GM-A3XN; TCGA-GM-A5PX; TCGA-LD-A66U; TCGA-LL-A50Y; TCGA-LL-A73Z; TCGA-LL-A9Q3; TCGA-OK-A5Q2; TCGA-OL-A5D8; TCGA-OL-A5RV; TCGA-OL-A6VQ; TCGA-S3-AA0Z; TCGA-W8-A86G; TCGA-WT-AB44; TCGA-A5-A0G9; TCGA-A5-A0GA; TCGA-A5-A0GN; TCGA-A5-A1OJ; TCGA-A5-A1OK; TCGA-A5-A7WJ; TCGA-A5-AB3J; TCGA-AJ-A3NC; TCGA-AJ-A8CW; TCGA-AP-A051; TCGA-AP-A0LL; TCGA-AP-A1DO; TCGA-AP-A1DP; TCGA-AP-A1E1; TCGA-AX-A0IS; TCGA-AX-A0J1; TCGA-AX-A1CK; TCGA-AX-A2HA; TCGA-AX-A3G8; TCGA-AX-A3GB; TCGA-B5-A0JR; TCGA-B5-A0K7; TCGA-B5-A0K9; TCGA-B5-A11Q; TCGA-B5-A11S; TCGA-B5-A11Z; TCGA-B5-A1MZ; TCGA-B5-A3FH; TCGA-B5-A5OC; TCGA-BG-A0MT; TCGA-BG-A0W1; TCGA-BG-A3EW; TCGA-BK-A13B; TCGA-BK-A13C; TCGA-BS-A0UT; TCGA-D1-A0ZR; TCGA-D1-A160; TCGA-D1-A162; TCGA-D1-A176; TCGA-D1-A17A; TCGA-D1-A17L; TCGA-D1-A1NZ; TCGA-D1-A2G5; TCGA-D1-A3DA; TCGA-DI-A0WH; TCGA-E6-A2P9; TCGA-EO-A22Y; TCGA-EY-A1GD; TCGA-EY-A1GQ; TCGA-EY-A1GT; TCGA-EY-A1GU; TCGA-FI-A2F4.

Anexo 7: Listado de los biomarcadores de los *clusters* genómicos.

***Cluster* genómico 1**

ITGA3; GSC2; DLX3; DGCR2; MED15; SERPIND1; CRKL; ESS2; RGS4; SLC35B1; DGCR6L; GGT2; TOB1; NECTIN4; F11R; NDUFS2; TOMM40L; AC007663.1; USP41; TUBA3FP; PEX19; PHB; NKIRAS2; BCRP2; JUP; TAC4; FLJ40194; KLHL11; NXPH3; AIFM3; ZNF74; LCE1E; AC091180.1; ARHGAP30; FAM230G; LCE1C; ZNF652; RNU6-1313P; TTC25; FAM230H; P2RX6P; TSSK2; Y_RNA; Y_RNA; KRTAP3-2; AC004471.1; SMPD4P1; PPIAP37; LINC02089; AP000552.1; E2F6P3; LINC02876; AL139011.1; ABHD17AP4; POM121L8P; EIF4EP2; TSSK1A; AP000550.2; SRP14P3; AC007731.3; KRT18P5; AC004461.1; AP000552.2; LINC01637; DGCR5; SLC9A3P2; PICART1; AC091180.2; AC027801.2; AC027801.3; TUBA3GP; AC091180.3; AC002401.2; AC015795.1; AC091180.4; AP000552.3; AC027801.4; AC125257.1; LINC01311; AC091180.5; LINC01982; LINC02073; AC000095.1; AC007326.1; AC004471.2; AC002470.1; DGCR5; AC007663.3; DGCR11; RIMBP3B; AC002401.3; AC243571.1; Metazoa_SRP; IGLL4P; YWHAEP7; AC004461.2; AC000095.2; AC007326.3; AC002472.2; FAM246A.

***Cluster* genómico 2**

SIRT1; TNR; EDEM3; PRG4; CAMSAP2; TECTB; GPAM; TNN; MRPS14; RNF2; ARL4AP1; NIBAN1; MYPN; ECM1; TARS2; ADAMTSL4; MCL1; CTSK; CERS2; ENSA; ARNT; HORMAD1; GOLPH3L; ARID5B; ANK3; TNNI1; CSRP1; CDK1; ATOH7; ENTR1P2; BRINP2; TSEN15; RNU6-1309P; RNA5SP72; Y_RNA; ADAMTSL4-AS1; Y_RNA; RNU6-1042P; Y_RNA; Y_RNA; RPL5P5; BX322639.1; RN7SKP202; RNU2-72P; AL359081.1; RPS10P7; AL645474.1; RPS27AP6; FDPSP1; RPS3AP38; HNRNPA1P46; FALEC; KSR1P1; RPS6P15; AL590133.1; AL445228.1; AL078645.2; AL096803.2; AL356273.1; LINC02640; LINC02625; RPS29P4; AL645504.1; TNR-IT1; KIAA0040; POU5F1P5; SLC4A1APP2; AL356292.1; LINC01351; BTBD7P2; ADAMTSL4-AS2; AL139135.2; RN7SL654P; Y_RNA; AL356356.1; ANK3-DT; AL590133.2; LINC01633; Z94057.1; MIR4257; MIR5191; RN7SL220P; MCRIP2P2; NTAN1P1; AL713866.2; AL445228.2; AL031601.1; AC097065.2; PACERR; AL031601.2; RN7SL600P; U4; RN7SL473P; AC119427.1; CTXND2; AL117339.4; MIR6878; AL158011.1; AL133553.2; AL049198.1.

***Cluster* genómico 3**

OR4K2; OR4K5; OR4K3; OR4N2; OR4M1; OR4H12P; OR4Q3; POTEG; OR11H13P; IGKV7-3; RNU6-458P; IGKC; IGKJ5; IGKJ4; IGKJ3; IGKJ2; IGKJ1; IGKV4-1; IGKV5-2; RNU6-1268P; RNU6-1239P; CR383658.1; POTEM; PGBD4P5; DUXAP9; BMS1P17; OR4K6P; RPL22P20; IGKV1-6; AC244205.1; IGKV1-8; IGKV3-11; IGKV3-7; IGKV1-12; IGKV1-5; DUXAP10; IGKV2-4; IGKV1-13; OR11H12; AL512310.1; MED15P6; CR383658.2; CR383656.1; CR383656.2; AL589743.1; BNIP3P6; AL512310.2; AL512310.3; AL512310.4; LINC02297; CR383656.3; CR383656.4; CR383656.5; CR383656.6; CR383656.7; CR383656.8; AL512310.5; CR383656.9; AL512310.6; MED15P1; AL512310.7; AL929601.1; CR383656.10; NF1P4; CR383656.11; ARHGAP42P4; ARHGAP42P5; AL929601.2; AL589182.1; AL512310.8; CR383656.12; CR383656.13; OR11K2P; OR11H2; OR4K4P; OR4N1P; CDRT15P13; AL391156.1; AL512310.9; AL589743.2; AL512310.10; AL512310.11; GRAMD4P4; AL512624.1; LINC01297; AL929601.3; AL512624.2; AL589743.3; TOMM40P1; AL589743.4; NEK2P1; NF1P10; NBEAP5; NF1P7; NF1P11; GRAMD4P3; AL589182.2; NBEAP6; AL589182.3; AL929602.1.

***Cluster* genómico 4**

NSMAF; MTFR1; STK3; ARMC1; AGO2; KHDRBS3; CA1; ENPP2; SULF1; SDCBP; DNAJC5B; CRH; TRIM55; FAM135B; OTUD6B; ADCY8; MMP16; KCNS2; CA3; OSR2; CYP7A1; AC090152.1; PTK2; COL22A1; ABRA; RRS1; TOX; Y_RNA; RNU4-50P; PDE7A; RNU6-144P; AC087698.1; RNU6-748P; PPIAP86; LRRC69; RN7SKP85; AC016877.1; RPL19P14; RPL26P26; AC068522.1; RRS1-AS1; AC090568.1; AF117829.1; RNA5SP272; RNA5SP278; LINC00967; AC104986.1; AC084082.1; AC103726.1; AC046195.1; AC012400.1; AC087341.1; AC016877.2; AP003355.1; CYCSP23; CPP; CA3-AS1; AC090578.1; RPSAP74; AC090987.1; LINC01592; AC105177.1; AC093331.1; AP003467.1; MAPRE1P1; VPS13B-DT; AC079015.1; AC103726.2; AC105150.1; AC110053.1; LINC01299; LINC02055; PPIAP85; AP003467.2; TPM3P3; AC015522.1; AC087354.1; MRPL57P7; MIR151A; AC083967.1; AC046195.2; AC107375.1; AC090578.2; AC016877.3; AC100814.1; AC055822.1; AC100812.1; AP003355.2; AP003789.1; AP003355.3; AC067931.1; AC105235.1; ERICD; AC067931.2; AC105213.1; AC139019.1; AC090578.3; AC009879.4; AC090994.1; AC018953.2.

***Cluster* genómico 5**

NDUFAB1; EEF1AKNMT; PALB2; TNRC6A; UBFD1; EARS2; GGA2; CPPED1; BFAR; CLUHP3; ERN 2,00; LHX9; NEK7; DCTN5; PLK1; CHP2; COG7; SCNN1B; VN1R3; ZNF267; MRTFB; ZNF720; AC034105.1; MIR365A; AC092324.1; Y_RNA; C1orf53; AC140658.1; SNORA75; AL031864.1; AC142381.1; LINC01567; AL031864.2; AC130464.1; PRR13P1; DNM3-IT1; VN1R66P; SHISA9; RPL35AP34; RN7SL274P; AC005774.1; AL135931.1; AC136428.1; AC008915.1; AC002519.1; AC130456.1; AC034105.2; AC034105.3; LINC02130; IGHV1OR16-3; AC008915.2; AC034105.4; SUB1P4; AC093515.1; AC136428.2; AC109597.1; AC074050.1; AC008870.1; BMS1P8; AC074050.2; AC142381.3; RBM22P12; AC007598.1; AC008870.2; AC127459.2; AC136428.3; AC109597.2; AC005774.2; CLEC19A; AC008870.3; RBM22P13; AC002519.2; CCNYL7; ENPP7P13; AC012317.1; AC106730.1; AC008731.1; IGHV1OR16-1; AC008870.4; AC142381.4; AC074050.3; LINC02186; AC003009.1; LINC02185; AC009134.1; TVP23CP2; AC130650.1; U91319.1; AC040173.1; MIR4718; IGHV3OR16-12; IGHV3OR16-9; IGHV3OR16-13; AC130650.2; AC008870.5; AC012317.2; RN7SKP23; AC136428.4; AC099482.2; BX248415.1.

***Cluster* genómico 6**

NCDN; TRAPPC3; PKN2; PLXNA2; TEKT2; EDEM3; RPE65; TNNI3K; TFAP2E; ADPRS; MAP7D1; CR2; RCOR3; NEK2; C4BPA; PSMB2; AGO3; CHRM3; TAF1A; HHIPL2; FMN2; LRRIQ3; LRRC53; SLC30A1; LPAR3; COL8A2; NEGR1; ERICH3; FAM177B; COLGALT2; RNA5SP50; Y_RNA; AURKAP1; AL356361.1; RPS7P5; RN7SKP229; RNU6-791P; LINC02238; ADH5P3; C4BPAP1; TXNP2; AL358453.1; AL445493.1; AL606753.1; RPS26P13; AL359918.1; NEGR1-IT1; PIN1P1; ZRANB2-AS2; AC093578.1; UBE2V2P4; AL356361.2; AL354949.1; LINC01717; CHRM3-AS2; RPS3AP8; AL138789.1; AL359918.2; AC105271.1; ERICH3-AS1; CHRM3-AS1; AC004865.1; AL596218.1; AC093158.1; PKN2-AS1; LRRC7-AS1; RPL36AP10; AC099063.2; RN7SL583P; AL445493.2; RNU4ATAC8P; RNA5SP70; FPGT; FPGT-TNNI3K; LINC02769; AC105275.2; LINC01735; RN7SL131P; AL592148.2; AL138787.2; AL359821.1; AL161734.1; AL513314.2; AC138393.3; AL391597.1; LINC02767; AL035412.1; AC104169.1; AL033530.1; AL591463.1; AL713852.1; AL606753.2; AL713852.2; AL596214.1; AL445438.1; AL356361.3; AL583825.1; AL590138.1; AL392172.1; AL161734.2.

Anexo 8: Listado de los pacientes en cada *cluster* de la representación integrada G.

Cluster G 1

TCGA-AO-A0JB; TCGA-LL-A5YM; TCGA-V7-A7HQ; TCGA-2E-A9G8; TCGA-4E-A92E; TCGA-5S-A9Q8; TCGA-A5-A0G9; TCGA-A5-A0GA; TCGA-A5-A0GB; TCGA-A5-A0GD; TCGA-A5-A0GG; TCGA-A5-A0GJ; TCGA-A5-A0GM; TCGA-A5-A0GN; TCGA-A5-A0GP; TCGA-A5-A0GQ; TCGA-A5-A0GU; TCGA-A5-A0GV; TCGA-A5-A0GX; TCGA-A5-A0R7; TCGA-A5-A0R8; TCGA-A5-A0R9; TCGA-A5-A0RA; TCGA-A5-A0VP; TCGA-A5-A0VQ; TCGA-A5-A1OJ; TCGA-A5-A1OK; TCGA-A5-A2K5; TCGA-A5-A2K7; TCGA-A5-A7WJ; TCGA-A5-AB3J; TCGA-AJ-A2QL; TCGA-AJ-A2QO; TCGA-AJ-A3BH; TCGA-AJ-A3BI; TCGA-AJ-A3BK; TCGA-AJ-A3EK; TCGA-AJ-A3EL; TCGA-AJ-A3I9; TCGA-AJ-A3NC; TCGA-AJ-A3NE; TCGA-AJ-A3OJ; TCGA-AJ-A3OL; TCGA-AJ-A5DV; TCGA-AJ-A5DW; TCGA-AJ-A6NU; TCGA-AJ-A8CT; TCGA-AJ-A8CV; TCGA-AJ-A8CW; TCGA-AP-A051; TCGA-AP-A053; TCGA-AP-A056; TCGA-AP-A05N; TCGA-AP-A05O; TCGA-AP-A05P; TCGA-AP-A0LE; TCGA-AP-A0LG; TCGA-AP-A0LJ; TCGA-AP-A0LL; TCGA-AP-A0LM; TCGA-AP-A0LN; TCGA-AP-A0LO; TCGA-AP-A0LP; TCGA-AP-A0LS; TCGA-AP-A0LT; TCGA-AP-A0LV; TCGA-AP-A1DH; TCGA-AP-A1DK; TCGA-AP-A1DM; TCGA-AP-A1DO; TCGA-AP-A1DP; TCGA-AP-A1DR; TCGA-AP-A1DV; TCGA-AP-A1E0; TCGA-AP-A1E1; TCGA-AP-A1E3; TCGA-AW-A1PO; TCGA-AX-A05S; TCGA-AX-A05T; TCGA-AX-A05U; TCGA-AX-A05W; TCGA-AX-A05Y; TCGA-AX-A05Z; TCGA-AX-A062; TCGA-AX-A06B; TCGA-AX-A06L; TCGA-AX-A0IS; TCGA-AX-A0J0; TCGA-AX-A0J1; TCGA-AX-A1C4; TCGA-AX-A1C5; TCGA-AX-A1C9; TCGA-AX-A1CE; TCGA-AX-A1CI; TCGA-AX-A1CJ; TCGA-AX-A1CK; TCGA-AX-A2H7; TCGA-AX-A2HA; TCGA-AX-A2HC; TCGA-AX-A2HD; TCGA-AX-A2HJ; TCGA-AX-A3FW; TCGA-AX-A3G8; TCGA-AX-A3G9; TCGA-AX-A3GB; TCGA-B5-A0JR; TCGA-B5-A0JS; TCGA-B5-A0JT; TCGA-B5-A0JU; TCGA-B5-A0JV; TCGA-B5-A0JX; TCGA-B5-A0JY; TCGA-B5-A0JZ; TCGA-B5-A0K0; TCGA-B5-A0K1; TCGA-B5-A0K2; TCGA-B5-A0K3; TCGA-B5-A0K4; TCGA-B5-A0K6; TCGA-B5-A0K7; TCGA-B5-A0K9; TCGA-B5-A0KB; TCGA-B5-A11F; TCGA-B5-A11G; TCGA-B5-A11J; TCGA-B5-A11M; TCGA-B5-A11N; TCGA-B5-A11O; TCGA-B5-A11P; TCGA-B5-A11Q; TCGA-B5-A11S; TCGA-B5-A11U; TCGA-B5-A11V; TCGA-B5-A11W; TCGA-B5-A11Z; TCGA-B5-A1MV; TCGA-B5-A1MW; TCGA-B5-A1MX; TCGA-B5-A1MZ; TCGA-B5-A3FB; TCGA-B5-A3FH; TCGA-B5-A5OC; TCGA-BG-A0LW; TCGA-BG-A0LX; TCGA-BG-A0M0; TCGA-BG-A0M2; TCGA-BG-A0M3; TCGA-BG-A0M4; TCGA-BG-A0M7; TCGA-BG-A0M9; TCGA-BG-A0MA; TCGA-BG-A0MC; TCGA-BG-A0MG; TCGA-BG-A0MI; TCGA-BG-A0MK; TCGA-BG-A0MO; TCGA-BG-A0MQ; TCGA-BG-A0MS; TCGA-BG-A0MT; TCGA-BG-A0MU; TCGA-BG-A0RY; TCGA-BG-A0VT; TCGA-BG-A0VV; TCGA-BG-A0VW; TCGA-BG-A0VX; TCGA-BG-A0VZ; TCGA-BG-A0W1; TCGA-BG-A0W2; TCGA-BG-A0YU; TCGA-BG-A186; TCGA-BG-A187; TCGA-BG-A18A; TCGA-BG-A18B; TCGA-BG-A18C; TCGA-BG-A220; TCGA-BG-A222; TCGA-BG-A2AD; TCGA-BG-A2AE; TCGA-BG-A2L7; TCGA-BG-A3EW; TCGA-BK-A0C9; TCGA-BK-A0CB; TCGA-BK-A13B; TCGA-BK-A13C; TCGA-BK-A56F; TCGA-BK-A6W3; TCGA-BK-A6W4; TCGA-BS-A0TA; TCGA-BS-A0TC; TCGA-BS-A0TD; TCGA-BS-A0TJ; TCGA-BS-A0U5; TCGA-BS-A0U7; TCGA-BS-A0U8; TCGA-BS-A0UA; TCGA-BS-A0UJ; TCGA-BS-A0UL; TCGA-BS-A0UM; TCGA-BS-A0UT; TCGA-BS-A0UV; TCGA-BS-A0V4; TCGA-BS-A0V6; TCGA-BS-A0V8; TCGA-BS-A0VI; TCGA-BS-A0WQ; TCGA-D1-A0ZN; TCGA-D1-A0ZO; TCGA-D1-A0ZQ; TCGA-D1-A0ZR; TCGA-D1-A0ZU; TCGA-D1-A0ZV; TCGA-D1-A101; TCGA-D1-A102; TCGA-D1-A103; TCGA-D1-A15W; TCGA-D1-A15Z; TCGA-D1-A160; TCGA-D1-A161; TCGA-D1-A162; TCGA-D1-A163; TCGA-D1-A165; TCGA-D1-A167; TCGA-D1-A169; TCGA-D1-A16B;

TCGA-D1-A16D; TCGA-D1-A16E; TCGA-D1-A16N; TCGA-D1-A16O; TCGA-D1-A16Q; TCGA-D1-A16R; TCGA-D1-A16V; TCGA-D1-A16X; TCGA-D1-A174; TCGA-D1-A175; TCGA-D1-A176; TCGA-D1-A177; TCGA-D1-A17A; TCGA-D1-A17B; TCGA-D1-A17C; TCGA-D1-A17D; TCGA-D1-A17F; TCGA-D1-A17H; TCGA-D1-A17K; TCGA-D1-A17L; TCGA-D1-A17N; TCGA-D1-A17Q; TCGA-D1-A17R; TCGA-D1-A17S; TCGA-D1-A17T; TCGA-D1-A17U; TCGA-D1-A1NS; TCGA-D1-A1NY; TCGA-D1-A1NZ; TCGA-D1-A1O0; TCGA-D1-A1O5; TCGA-D1-A1O7; TCGA-D1-A2G5; TCGA-D1-A3DA; TCGA-D1-A3DG; TCGA-DF-A2KN; TCGA-DF-A2KR; TCGA-DF-A2KS; TCGA-DF-A2KV; TCGA-DF-A2KZ; TCGA-DF-A2L0; TCGA-DI-A0WH; TCGA-DI-A1BY; TCGA-DI-A1C3; TCGA-DI-A1NO; TCGA-DI-A2QU; TCGA-E6-A1LX; TCGA-E6-A1M0; TCGA-E6-A2P9; TCGA-EC-A1NJ; TCGA-EC-A1QX; TCGA-EC-A24G; TCGA-EO-A1Y7; TCGA-EO-A22R; TCGA-EO-A22S; TCGA-EO-A22T; TCGA-EO-A22U; TCGA-EO-A22X; TCGA-EO-A22Y; TCGA-EO-A3AS; TCGA-EO-A3AU; TCGA-EO-A3AY; TCGA-EO-A3B0; TCGA-EO-A3KX; TCGA-EO-A3L0; TCGA-EY-A1G8; TCGA-EY-A1GD; TCGA-EY-A1GE; TCGA-EY-A1GF; TCGA-EY-A1GH; TCGA-EY-A1GI; TCGA-EY-A1GK; TCGA-EY-A1GL; TCGA-EY-A1GQ; TCGA-EY-A1GR; TCGA-EY-A1GT; TCGA-EY-A1GU; TCGA-EY-A1GW; TCGA-EY-A1GX; TCGA-EY-A1H0; TCGA-EY-A214; TCGA-EY-A215; TCGA-EY-A2OM; TCGA-EY-A2OP; TCGA-EY-A2OQ; TCGA-EY-A547; TCGA-EY-A548; TCGA-EY-A549; TCGA-EY-A54A; TCGA-EY-A5W2; TCGA-EY-A72D; TCGA-FI-A2D0; TCGA-FI-A2D4; TCGA-FI-A2D5; TCGA-FI-A2F4; TCGA-H5-A2HR; TCGA-PG-A916; TCGA-PG-A917; TCGA-QF-A5YT; TCGA-QS-A5YQ; TCGA-QS-A744; TCGA-SJ-A6ZI; TCGA-SJ-A6ZJ; TCGA-SL-A6JA.

***Cluster* G 2**

TCGA-3C-AAAU; TCGA-3C-AALK; TCGA-5L-AAT0; TCGA-5L-AAT1; TCGA-A1-A0SD; TCGA-A1-A0SI; TCGA-A2-A04N; TCGA-A2-A04R; TCGA-A2-A04V; TCGA-A2-A04X; TCGA-A2-A04Y; TCGA-A2-A0CO; TCGA-A2-A0CP; TCGA-A2-A0CQ; TCGA-A2-A0EM; TCGA-A2-A0EN; TCGA-A2-A0ES; TCGA-A2-A0EX; TCGA-A2-A0SV; TCGA-A2-A0SY; TCGA-A2-A0T4; TCGA-A2-A0T5; TCGA-A2-A0T6; TCGA-A2-A0T7; TCGA-A2-A0YC; TCGA-A2-A0YD; TCGA-A2-A0YL; TCGA-A2-A1FV; TCGA-A2-A1FX; TCGA-A2-A1FZ; TCGA-A2-A1G4; TCGA-A2-A259; TCGA-A2-A25C; TCGA-A2-A25E; TCGA-A2-A3KC; TCGA-A2-A4RW; TCGA-A2-A4RY; TCGA-A2-A4S2; TCGA-A2-A4S3; TCGA-A7-A13H; TCGA-A7-A3J1; TCGA-A7-A5ZW; TCGA-A7-A5ZX; TCGA-A8-A06O; TCGA-A8-A06Y; TCGA-A8-A07E; TCGA-A8-A07G; TCGA-A8-A07J; TCGA-A8-A07L; TCGA-A8-A07P; TCGA-A8-A082; TCGA-A8-A083; TCGA-A8-A086; TCGA-A8-A08A; TCGA-A8-A08C; TCGA-A8-A08H; TCGA-A8-A08P; TCGA-A8-A08S; TCGA-A8-A08T; TCGA-A8-A08Z; TCGA-A8-A099; TCGA-A8-A09D; TCGA-A8-A09V; TCGA-A8-A09W; TCGA-A8-A0A1; TCGA-A8-A0A4; TCGA-A8-A0A6; TCGA-AC-A23C; TCGA-AC-A23G; TCGA-AC-A2FB; TCGA-AC-A2FF; TCGA-AC-A2FK; TCGA-AC-A3HN; TCGA-AC-A3QP; TCGA-AC-A3QQ; TCGA-AC-A3TM; TCGA-AC-A3TN; TCGA-AC-A3W6; TCGA-AC-A3YI; TCGA-AC-A6IV; TCGA-AC-A6IX; TCGA-AC-A6NO; TCGA-AN-A03X; TCGA-AN-A046; TCGA-AN-A049; TCGA-AN-A0FD; TCGA-AN-A0FF; TCGA-AN-A0FN; TCGA-AN-A0XL; TCGA-AN-A0XP; TCGA-AN-A0XV; TCGA-AN-A0XW; TCGA-AO-A03M; TCGA-AO-A03R; TCGA-AO-A03T; TCGA-AO-A0J8; TCGA-AO-A0J9; TCGA-AO-A0JA; TCGA-AO-A0JF; TCGA-AO-A0JJ; TCGA-AO-A1KO; TCGA-AO-A1KP; TCGA-AQ-A1H2; TCGA-AQ-A1H3; TCGA-AR-A0TZ; TCGA-AR-A0U3; TCGA-AR-A1AK; TCGA-AR-A1AL; TCGA-AR-A1AM; TCGA-AR-A1AN; TCGA-AR-A1AP; TCGA-AR-A1AS; TCGA-AR-A1AU; TCGA-AR-A1AX; TCGA-AR-A24L; TCGA-AR-A24N; TCGA-AR-A24P; TCGA-AR-A2LM; TCGA-AR-A2LN; TCGA-AR-A2LQ; TCGA-AR-A5QN; TCGA-B6-A0I5; TCGA-B6-A0IG; TCGA-B6-A0IP; TCGA-B6-A0RH; TCGA-B6-A0RI; TCGA-B6-A0RL; TCGA-B6-A0RM; TCGA-B6-A0RQ; TCGA-B6-A0WS; TCGA-B6-A0WT; TCGA-B6-A0WW; TCGA-B6-A0WZ; TCGA-B6-A0X4; TCGA-B6-A0X5;

TCGA-B6-A2IU; TCGA-B6-A40B; TCGA-B6-A40C; TCGA-BH-A0AZ; TCGA-BH-A0B0; TCGA-BH-A0BC; TCGA-BH-A0BD; TCGA-BH-A0BO; TCGA-BH-A0BQ; TCGA-BH-A0BR; TCGA-BH-A0BS; TCGA-BH-A0DE; TCGA-BH-A0DH; TCGA-BH-A0DI; TCGA-BH-A0DK; TCGA-BH-A0DO; TCGA-BH-A0DQ; TCGA-BH-A0DS; TCGA-BH-A0DT; TCGA-BH-A0DV; TCGA-BH-A0DX; TCGA-BH-A0E1; TCGA-BH-A0E2; TCGA-BH-A0E7; TCGA-BH-A0EA; TCGA-BH-A0EI; TCGA-BH-A0GZ; TCGA-BH-A0H0; TCGA-BH-A0H3; TCGA-BH-A0H6; TCGA-BH-A0H9; TCGA-BH-A0HA; TCGA-BH-A0HB; TCGA-BH-A0HI; TCGA-BH-A0HO; TCGA-BH-A0HQ; TCGA-BH-A0HU; TCGA-BH-A0W4; TCGA-BH-A0W7; TCGA-BH-A18F; TCGA-BH-A18J; TCGA-BH-A18M; TCGA-BH-A1ET; TCGA-BH-A1EY; TCGA-BH-A1F5; TCGA-BH-A1F8; TCGA-BH-A1FB; TCGA-BH-A1FD; TCGA-BH-A1FE; TCGA-BH-A1FG; TCGA-BH-A1FL; TCGA-BH-A201; TCGA-BH-A28O; TCGA-BH-A28Q; TCGA-BH-A42T; TCGA-BH-A42V; TCGA-BH-A6R8; TCGA-BH-A8FY; TCGA-BH-A8G0; TCGA-BH-AB28; TCGA-C8-A12N; TCGA-C8-A12T; TCGA-C8-A132; TCGA-C8-A1HE; TCGA-C8-A1HI; TCGA-C8-A1HO; TCGA-C8-A26V; TCGA-C8-A26W; TCGA-C8-A273; TCGA-C8-A3M8; TCGA-D8-A13Y; TCGA-D8-A140; TCGA-D8-A141; TCGA-D8-A1J8; TCGA-D8-A1JB; TCGA-D8-A1JC; TCGA-D8-A1JE; TCGA-D8-A1JH; TCGA-D8-A1JN; TCGA-D8-A1JP; TCGA-D8-A1JU; TCGA-D8-A1X8; TCGA-D8-A1X9; TCGA-D8-A1XB; TCGA-D8-A1XD; TCGA-D8-A1XM; TCGA-D8-A1XO; TCGA-D8-A1XU; TCGA-D8-A1Y0; TCGA-D8-A27I; TCGA-D8-A27K; TCGA-D8-A27L; TCGA-D8-A27N; TCGA-D8-A27P; TCGA-D8-A27T; TCGA-D8-A3Z5; TCGA-D8-A3Z6; TCGA-D8-A4Z1; TCGA-E2-A105; TCGA-E2-A108; TCGA-E2-A10B; TCGA-E2-A10F; TCGA-E2-A14Q; TCGA-E2-A14S; TCGA-E2-A14U; TCGA-E2-A14Z; TCGA-E2-A153; TCGA-E2-A15C; TCGA-E2-A15D; TCGA-E2-A15F; TCGA-E2-A15I; TCGA-E2-A15J; TCGA-E2-A15K; TCGA-E2-A15L; TCGA-E2-A15O; TCGA-E2-A15P; TCGA-E2-A15R; TCGA-E2-A15T; TCGA-E2-A1B1; TCGA-E2-A1B4; TCGA-E2-A1B5; TCGA-E2-A1BC; TCGA-E2-A1IE; TCGA-E2-A1IG; TCGA-E2-A1IJ; TCGA-E2-A1IK; TCGA-E2-A1IL; TCGA-E2-A1IO; TCGA-E2-A1IU; TCGA-E2-A1L8; TCGA-E2-A1L9; TCGA-E2-A2P6; TCGA-E2-A3DX; TCGA-E2-A576; TCGA-E9-A1N6; TCGA-E9-A1NE; TCGA-E9-A1NG; TCGA-E9-A1NH; TCGA-E9-A227; TCGA-E9-A22H; TCGA-E9-A3Q9; TCGA-E9-A5FK; TCGA-EW-A1IY; TCGA-EW-A1J2; TCGA-EW-A1J3; TCGA-EW-A1J5; TCGA-EW-A1P3; TCGA-EW-A1P6; TCGA-EW-A1PA; TCGA-EW-A1PF; TCGA-EW-A1PG; TCGA-EW-A6SC; TCGA-GM-A2D9; TCGA-GM-A2DA; TCGA-GM-A2DC; TCGA-GM-A2DL; TCGA-GM-A2DM; TCGA-GM-A3NY; TCGA-GM-A3XG; TCGA-GM-A3XN; TCGA-GM-A4E0; TCGA-GM-A5PV; TCGA-GM-A5PX; TCGA-HN-A2OB; TCGA-LD-A66U; TCGA-LD-A7W5; TCGA-LL-A440; TCGA-LL-A50Y; TCGA-LL-A73Z; TCGA-LL-A9Q3; TCGA-LQ-A4E4; TCGA-OK-A5Q2; TCGA-OL-A5D8; TCGA-OL-A5RV; TCGA-OL-A66J; TCGA-OL-A66L; TCGA-OL-A6VQ; TCGA-OL-A6VR; TCGA-PE-A5DC; TCGA-S3-AA0Z; TCGA-S3-AA11; TCGA-S3-AA14; TCGA-W8-A86G; TCGA-WT-AB41; TCGA-WT-AB44; TCGA-XX-A899; TCGA-XX-A89A; TCGA-Z7-A8R5; TCGA-Z7-A8R6.

***Cluster* G 3**

TCGA-3C-AALJ; TCGA-4H-AAAK; TCGA-A1-A0SF; TCGA-A1-A0SJ; TCGA-A1-A0SN; TCGA-A2-A0CK; TCGA-A2-A0CS; TCGA-A2-A0CU; TCGA-A2-A0CV; TCGA-A2-A0CW; TCGA-A2-A0D3; TCGA-A2-A0D4; TCGA-A2-A0EO; TCGA-A2-A0ER; TCGA-A2-A0ET; TCGA-A2-A0EU; TCGA-A2-A0EV; TCGA-A2-A0EW; TCGA-A2-A0SU; TCGA-A2-A0T3; TCGA-A2-A0YG; TCGA-A2-A0YH; TCGA-A2-A0YI; TCGA-A2-A25A; TCGA-A2-A25B; TCGA-A7-A0CD; TCGA-A7-A0CJ; TCGA-A7-A13F; TCGA-A7-A2KD; TCGA-A7-A425; TCGA-A7-A426; TCGA-A7-A4SB; TCGA-A7-A56D; TCGA-A7-A6VX; TCGA-A8-A06P; TCGA-A8-A06Q; TCGA-A8-A06T; TCGA-A8-A06U; TCGA-A8-A06Z; TCGA-A8-A076; TCGA-A8-A079; TCGA-A8-A07B; TCGA-A8-A07F; TCGA-A8-A07W; TCGA-A8-A07Z; TCGA-A8-A081; TCGA-A8-A08F; TCGA-A8-A08G; TCGA-A8-A08I; TCGA-A8-A08O;

TCGA-A8-A090; TCGA-A8-A092; TCGA-A8-A093; TCGA-A8-A095; TCGA-A8-A096; TCGA-A8-A097; TCGA-A8-A09A; TCGA-A8-A09B; TCGA-A8-A09C; TCGA-A8-A09E; TCGA-A8-A09I; TCGA-A8-A09K; TCGA-A8-A09M; TCGA-A8-A09N; TCGA-A8-A09Q; TCGA-A8-A09R; TCGA-A8-A09T; TCGA-A8-A0A2; TCGA-A8-A0A9; TCGA-A8-A0AB; TCGA-AC-A23E; TCGA-AC-A2B8; TCGA-AC-A2BM; TCGA-AC-A2FE; TCGA-AC-A3BB; TCGA-AC-A4ZE; TCGA-AC-A5XS; TCGA-AC-A5XU; TCGA-AC-A62Y; TCGA-AC-A8OP; TCGA-AC-A8OS; TCGA-AN-A03Y; TCGA-AN-A04A; TCGA-AN-A0AJ; TCGA-AN-A0FK; TCGA-AN-A0FW; TCGA-AN-A0FY; TCGA-AO-A03L; TCGA-AO-A03N; TCGA-AO-A03O; TCGA-AO-A03P; TCGA-AO-A03V; TCGA-AO-A0JD; TCGA-AO-A0JM; TCGA-AO-A126; TCGA-AO-A12A; TCGA-AO-A12B; TCGA-AO-A12E; TCGA-AO-A1KS; TCGA-AO-A1KT; TCGA-AQ-A04H; TCGA-AQ-A0Y5; TCGA-AR-A0TR; TCGA-AR-A0TV; TCGA-AR-A0TW; TCGA-AR-A0U2; TCGA-AR-A24H; TCGA-AR-A24K; TCGA-AR-A24R; TCGA-AR-A24S; TCGA-AR-A24T; TCGA-AR-A24V; TCGA-AR-A24W; TCGA-AR-A24Z; TCGA-AR-A254; TCGA-AR-A255; TCGA-AR-A2LJ; TCGA-AR-A2LK; TCGA-AR-A2LL; TCGA-AR-A2LO; TCGA-AR-A5QM; TCGA-AR-A5QP; TCGA-B6-A0I9; TCGA-B6-A0IB; TCGA-B6-A0IO; TCGA-B6-A0RO; TCGA-B6-A0WV; TCGA-B6-A1KI; TCGA-B6-A401; TCGA-BH-A0AU; TCGA-BH-A0AY; TCGA-BH-A0B5; TCGA-BH-A0B6; TCGA-BH-A0BJ; TCGA-BH-A0BP; TCGA-BH-A0BT; TCGA-BH-A0BV; TCGA-BH-A0BZ; TCGA-BH-A0C0; TCGA-BH-A0C1; TCGA-BH-A0DP; TCGA-BH-A0DZ; TCGA-BH-A0E9; TCGA-BH-A0EB; TCGA-BH-A0GY; TCGA-BH-A0H7; TCGA-BH-A0HX; TCGA-BH-A0W3; TCGA-BH-A0W5; TCGA-BH-A18I; TCGA-BH-A18K; TCGA-BH-A18L; TCGA-BH-A18N; TCGA-BH-A1ES; TCGA-BH-A1EU; TCGA-BH-A1EV; TCGA-BH-A1EX; TCGA-BH-A1F2; TCGA-BH-A1FN; TCGA-BH-A202; TCGA-BH-A2L8; TCGA-BH-A5J0; TCGA-BH-A8FZ; TCGA-C8-A12O; TCGA-C8-A12U; TCGA-C8-A12W; TCGA-C8-A130; TCGA-C8-A1HM; TCGA-C8-A1HN; TCGA-C8-A26Z; TCGA-C8-A274; TCGA-C8-A27A; TCGA-D8-A146; TCGA-D8-A1JD; TCGA-D8-A1JI; TCGA-D8-A1JJ; TCGA-D8-A1X5; TCGA-D8-A1X6; TCGA-D8-A1XA; TCGA-D8-A1XF; TCGA-D8-A1XL; TCGA-D8-A1XR; TCGA-D8-A1XY; TCGA-D8-A1Y1; TCGA-D8-A1Y2; TCGA-D8-A1Y3; TCGA-D8-A27G; TCGA-D8-A27R; TCGA-D8-A27V; TCGA-E2-A10A; TCGA-E2-A10C; TCGA-E2-A10E; TCGA-E2-A14O; TCGA-E2-A14T; TCGA-E2-A14V; TCGA-E2-A154; TCGA-E2-A156; TCGA-E2-A15A; TCGA-E2-A15E; TCGA-E2-A15H; TCGA-E2-A15M; TCGA-E2-A1BD; TCGA-E2-A1IF; TCGA-E2-A1IH; TCGA-E2-A1IN; TCGA-E2-A1L6; TCGA-E2-A2P5; TCGA-E2-A56Z; TCGA-E2-A570; TCGA-E9-A22D; TCGA-E9-A22E; TCGA-E9-A295; TCGA-E9-A3X8; TCGA-E9-A54Y; TCGA-E9-A6HE; TCGA-EW-A1IW; TCGA-EW-A1J1; TCGA-EW-A1J6; TCGA-EW-A1OY; TCGA-EW-A1P5; TCGA-EW-A1PC; TCGA-EW-A1PE; TCGA-EW-A423; TCGA-EW-A6S9; TCGA-GI-A2C8; TCGA-JL-A3YX; TCGA-LL-A5YN; TCGA-LL-A6FQ; TCGA-LL-A7T0; TCGA-OL-A5DA; TCGA-OL-A5RX; TCGA-OL-A66K; TCGA-OL-A66O; TCGA-PE-A5DE; TCGA-S3-A6ZF; TCGA-S3-A6ZH; TCGA-S3-AA17; TCGA-UL-AAZ6; TCGA-5B-A90C; TCGA-A5-A0GE; TCGA-A5-A3LO; TCGA-AJ-A2QK; TCGA-AJ-A2QN; TCGA-AJ-A3EM; TCGA-AX-A060; TCGA-AX-A06J; TCGA-AX-A1CF; TCGA-AX-A2H8; TCGA-AX-A2HK; TCGA-AX-A3FX; TCGA-B5-A11E; TCGA-B5-A11I; TCGA-B5-A11R; TCGA-B5-A11Y; TCGA-B5-A121; TCGA-B5-A3F9; TCGA-B5-A3FD; TCGA-BK-A4ZD; TCGA-BS-A0TI; TCGA-BS-A0V7; TCGA-D1-A17M.

Anexo 9: Listado de los biomarcadores de los *clusters* de la representación integrada G.

***Cluster* G 1**

SPAG9; UTP18; STRAP; KPNA6; RFC1; ZFYVE16; ARFGEF1; KLF6; VPS35; MPP5; TOP2B; UBE2K; ITCH; TNPO1; CHMP2B; REST; SLC25A24; ZW10; XRN2; CDC5L; SOS 2; HIF1A; TRIP11; APEX1; SRP54; SEC23A; MIB1; PARP4; KPNA3; TFAM; NFKB1; C11orf58; CAND1; TBCCD1; KPNA1; KLHL24; NEK4; PCGF1; SLC35D1; DR1; CREB1; CCNI; RAB14; POLK; VPS26A; ACSL3; NUP153; HNRNPR; KTN1; CCNT1; UBE2G1; RFC3; EDEM1; UBQLN1; NAA35; ISCA1; CAPRIN1; SUCLA2; IREB2; STAM; YME1L1; SMC2; YAP1; RDX; HAUS2; DBT; FBXO11; NUP54; TMX1; IQGAP1; CLTC; YIPF5; ZMYM4; SMARCA5; PPP2R5E; GABPA; SON; USP1; TFB2M; SLC25A46; UBLCP1; OXR1; STRBP; PRPF18; NSD1; DDX21; CUL5; ANKS3; MAPK1IP1L; CDKN2AIP; PGM2; RALGAPB; PAIP1; AKIRIN1; NDUFA11; C16orf72; AP3M1; GMFB; TOP 1,00; DDX3X.

***Cluster* G 2**

ANKIB1; PHTF2; MAP4K5; SLC30A9; KPNA6; NUP160; SIKE1; NCKAP1; BTBD1; ZBTB11; DHX8; VPS35; AP3M2; ZNF37A; ITCH; XPO1; RAB10; CPNE3; DHPS; SH3GLB1; CSNK2A1; SNRNP70; TRMT1; CCDC130; AVL9; MTPN; SMC3; KPNB1; GNPTAB; KCTD20; HMGCR; KPNA1; ACAP2; ATP6V1A; PAPOLG; BIRC6; STAG1; ANKRD13C; RAD23B; AFTPH; IDE; PDS5A; RASL11A; FYTTD1; CSE1L; NUP153; PPP1R12C; CAPN7; UBQLN1; DHX9; CEP350; USP37; SCYL2; UGGT1; STAM; C11orf1; ACTR2; FBXO11; LRPPRC; G3BP2; TSSK4; CLTC; TMEM91; CAPN10; TOR1AIP1; OSBPL11; TRPT1; UEVLD; NBAS; ANAPC1; SMARCA5; TRIP12; MIER3; RNF111; USP1; AZI2; GMPS; NAA15; CPSF2; ZNF143; ANKS3; INO80E; ZNF747; USP38; RALGAPB; KIF5B; AGFG1; ZWILCH; DHX36; SEC24C; C3orf38; PTPN11; MED14; AP3M1; ZBTB6; LIN54; LCOR; ZNF770; KLHL9; DDX3X.

***Cluster* G 3**

SLC30A9; RB1CC1; MFAP3; AIFM2; LIMA1; PTGER3; SLK; ZBTB11; DHX29; DRD4; ZNF638; BZW1; TNPO1; SCAMP1; DDX18; C3orf18; EEA1; XYLT1; TRPS1; NFKBIB; SNRNP70; CCDC130; PTOV1; CCNE1; COPE; FKBP8; LZTS2; HBS1L; FGF1; BBX; BIRC6; STRN; ZBTB17; RCAN3; UBN1; UCHL3; AFTPH; TMEM39B; FYTTD1; PPP1R12C; CHCHD5; PAX8; HNRNPR; TMEM53; BCL2L12; JUND; ZNF428; SERINC3; BLOC1S1; USP37; SCYL2; DENND4C; PREPL; FBXO11; RAP1GDS1; NEDD1; TMEM91; SYTL1; LYPLAL1; SCRN3; SETD7; ITGB1; SCOC; SMARCA5; TRIP12; HKDC1; AASDH; RNF111; EIF5B; ZG16B; USP1; IL20; CCNYL1; PRKCI; DNAJB14; PRRC1; SFXN1; SLU7; SOX17; NSD1; ANKS3; STXBP6; PCBP1; IRX2; ATF7IP; SRP72; NDUFA11; DHX36; PDIK1L; RPLP2; ZNF366; C16orf72; TBK1; C5orf38; USP7; PLEKHN1; DPYD; EMP2; MRPL23; ZNF579.