Laboratorio de Estudios de Interacción
Celular en Reproducción y Cáncer

Uso de herramientas bioinformáticas, bioestadísticas y modelos celulares en la búsqueda de potenciales biomarcadores del Cáncer de Endometrio

Autoras:
María Cecilia Argibay - 54296
Luciana Montivero - 55232

Directora:
Dra. Mónica Vazquez-Levin
Co-directora:
Dra. María José Besso

Proyecto Final de Carrera

Año 2019

Resumen

El **cáncer de endometrio (CE)** es la segunda neoplasia ginecológica más frecuente y la séptima causa de muerte por cáncer en mujeres tanto a nivel mundial como nacional; en los próximos 20 años se prevén aumentos mayores al 50% en las tasas de incidencia y mortalidad de esta enfermedad. Se caracteriza por la proliferación descontrolada de las células del endometrio sobre otros tejidos y órganos del cuerpo humano.

En la práctica clínica, el diagnóstico de CE se realiza a partir de la sintomatología característica de sangrado uterino anormal y/o factores de riesgo aumentados, mediante una ecografía transvaginal y una biopsia o legrado uterino con o sin histeroscopía. El diagnóstico se completa con la extracción quirúrgica del útero, llamada **histerectomía**, y la clasificación y estadificación definitiva del cáncer basada en el estadio, grado e invasión miometrial del sistema de la Federación Internacional de Ginecología y Obstetricia. A partir de lo anterior se define la indicación del tratamiento terapéutico quirúrgico y adyuvante. A diferencia de otros tipos de cáncer, actualmente no existen pruebas de detección para el CE. En función de criterios basados en la evaluación anatómico-patológica, el 75% de los casos es diagnosticado en estadio temprano (estadio I) y presenta una **sobrevida total (OS)** a 5 años del ∼90%. El 25% restante es diagnosticado en estadios avanzados, con invasión mayor al 50% del miometrio y/o nódulos linfáticos. Estos casos presentan un peor pronóstico y tasas de OS a 5 años entre el 20 y 65%. Asimismo, del 75% de los casos inicialmente diagnosticados en estadio temprano, ∼20% sufre una reclasificación a estadios más avanzados luego de la estadificación quirúrgico-patológica. Estos casos presentan un pronóstico pobre, con un aumento en la tasa de recurrencia post-quirúrgica, mayor probabilidad de presentar metástasis extra-pélvica en el momento de la recurrencia y, consecuentemente, una disminución en la sobrevida a 5 años.

Sobre la base de lo expuesto, resulta imperativo desarrollar estudios orientados a la identificación de biomarcadores moleculares que complementen el manejo actual del CE, contribuyendo a una mejor clasificación de tumores, estimación de riesgos y disminución de efectos secundarios del tratamiento. Los biomarcadores tumorales en CE tienen el potencial de asistir el manejo preoperatorio de la enfermedad para identificar pacientes de alto riesgo, aún en casos en estadio temprano, y de ser trasladados a la práctica clínica de rutina mediante un método estandarizado, reproducible y de bajo costo. Son de especial interés los **biomarcadores diagnósticos para la categorización de la agresividad tumoral** y los **biomarcadores pronósticos de recurrencia temprana**. Considerando la relevancia que ha ganado la bioinformática en la investigación biomédica, en años recientes se han implementado diversos algoritmos que emplean abordajes bioinformáticos para la identificación de biomarcadores para diversas enfermedades, particularmente el cáncer. Esto ha aportado nuevas herramientas basadas en la evaluación de conjuntos de genes con expresión diferencial para el diagnóstico y pronóstico de cáncer de mama, colon, estómago y próstata, entre otros.

El presente estudio tuvo por objeto utilizar un algoritmo basado en una combinación de herramientas bioinformáticas de minería de texto, datos y priorización génica, así como análisis bioestadísticos para identificar genes con potencial de biomarcadores de la progresión y agresividad del CE. En primer lugar, el análisis de los resultados de un estudio de transcriptómica con microarreglos de ADN (GSE17025; plataforma GEO) de muestras de pacientes con CE permitió identificar genes diferencialmente expresados, a partir de los que se seleccionaron 39 con expresión diferencial para tres parámetros del CE: la expresión en el tumor respecto del tejido no tumoral, el subtipo histológico (CE endometrioide respecto de CE no endometrioide) y el grado (grados 1 y 2 respecto de grado 3). Este análisis contribuyó a enfocar la evaluación en aquellos genes asociados a un pronóstico más desfavorable de CE. Por otra parte, se realizó una búsqueda de genes previamente reportados como asociados a CE, utilizando la base de datos DisGeNET, y se identificaron 962 genes asociados a la enfermedad. A continuación se realizó un análisis de priorización génica (ToppGene) empleando la lista de genes seleccionados en el análisis de transcriptómica GEO como "lista de prueba" y los genes de DisGeNET como "lista de entrenamiento". Luego de desestimar 6 genes por diversos criterios, la expresión de los 33 restantes fue evaluada en un muestreo de CE de TCGA (TCGA-UCEC). Los genes fueron posteriormente evaluados empleando modelos estadísticos como el método de Kaplan-Meier, *Odds Ratio* y el modelo de riesgos proporcionales de Cox, seleccionándose 20, 16 y 6 genes, respectivamente. Finalmente, tres de los 6 genes fueron elegidos sobre la base de los datos disponibles en el repositorio *Human Protein Atlas*: **PTCH1**, **TMPRSS2** y **TPX2**.

La expresión de los transcriptos de PTCH1, TMPRSS2 y TPX2 fue evaluada por PCR estándar en dos líneas celulares (Hec-1a e Ishikawa) en un modelo de sobreexpresión del factor de transcripción ETV5 (cuya expresión se encontró aumentada en el frente tumoral del CE) (HGE e Ishikawa-ETV5) y sus controles (Hec-1a e Ishikawa). Como resultado, se encontró expresión menor o indetectable para PTCH1 y TMPRSS2 en las líneas celulares HGE e Ishikawa-ETV5 de fenotipo más agresivo, en concordancia con los resultados de los estudios bioinformáticos y bioestadísticos. El transcripto TPX2 fue detectado en las 4 líneas celulares, y su expresión aumentada en las líneas HGE e Ishikawa-ETV5 fue detectada en protocolos de PCR cuantitativa en tiempo real. Por encontrarse sobreexpresado en los modelos más agresivos, el gen TPX2 fue caracterizado además a nivel de proteína, encontrándose una expresión mayor en las lineas HGE e Ishikawa-ETV5 respecto de los controles. La expresión aumentada de TPX2 se ha asociado a la progresión de diferentes cánceres, entre ellos carcinoma esofágico de células escamosas, de vejiga, de cuello uterino, hepatocelular, de próstata y gástrico. Más aún, los resultados se encuentran en línea con dos reportes recientes que identifican a TPX2 como un gen relacionado al CE, y un estudio que propone a TPX2 como un indicador de pronóstico pobre. Asimismo, otros estudios identifican a Aurora Quinasa A como una molécula clave para el CE, la que se ha propuesto constituye una unidad funcional con TPX2 para la progresión tumoral.

En estudios futuros, se podrán diseñar estrategias para evaluar sobre muestras frescas de tejido de pacientes con CE la expresión de los transcriptos PTCH1, TMPRSS2 y TPX2 empleando PCR en tiempo real. Estos estudios contribuirán a su validación final como biomarcadores de agresividad

de CE. Teniendo en cuenta que para el estudio de biomarcadores proteicos asociados a la progresión y agresividad del cáncer se prefiere la identificación de moléculas cuya expresión aumenta con la malignidad frente a aquellas cuya expresión disminuye pues facilita los abordajes de detección (ej. en inmunohistoquímica), se podrán realizar estudios de expresión de la proteína TPX2 en muestreos de pacientes con CE. Los abordajes propuestos permitirán determinar el aporte al diagnóstico y pronóstico de la enfermedad, así como la factibilidad de su implementación en la clínica.

Agradecimientos

Agradecemos al Instituto Tecnológico de Buenos Aires por formarnos en la disciplina de la Ingeniería y brindarnos las herramientas para comenzar a transitar nuestras carreras profesionales. También al Instituto de Biología y Medicina Experimental, en particular al Laboratorio de Estudios de Interacción Celular en Reproducción y Cáncer, por la oportunidad de desarrollar nuestro proyecto final rodeadas de científicos destacados y en un ambiente de investigación académica, así como las herramientas y la financiación para llevarlo a cabo.

Agradecemos especialmente a la Dra. Mónica Vazquez-Levin por permitirnos realizar este proyecto final de carrera bajo su dirección y ofrecernos todos los recursos y el tiempo necesarios para lograrlo. Gracias a su dedicación, paciencia, generosidad, confianza y optimismo logramos llevar a cabo este proyecto, aprendiendo y creciendo como profesionales y personas. En especial, le agradecemos por abrirnos las puertas a su laboratorio y acompañarnos a lo largo de todo el camino, ya que sin ella no hubiese sido posible este trabajo. A la Dra. María José Besso, quien pacientemente nos guió en los experimentos que conforman este trabajo y nos aconsejó en los inicios del proceso. Agradecemos también a Gustavo, Rocío y Jorge por acompañar nuestras jornadas de trabajo en el laboratorio.

Al Ing. Norberto Lerendegui, quien nos orientó en los primeros pasos y nos dio la libertad para elegir este proyecto. Asimismo, agradecemos su apoyo y excelente predisposición frente a todas nuestras inquietudes. A la Dra. Carolina Cernadas, por despertar con sus clases nuestro interés por la estadística y ayudarnos a aplicarla en este trabajo. Agradecemos al Bioing. Federico Paschetta por su preocupación, esfuerzo y gestión en las últimas etapas del proyecto.

Agradezco enormemente a mi familia, por el esfuerzo, generosidad y apoyo incondicional a lo largo de todos estos años. A mi mamá por siempre demostrarme que con esfuerzo y perseverancia todo se logra. A mi papá por transmitirme su curiosidad y creatividad en todos los aspectos de la vida. A mis hermanos por su paciencia e interés en todos mis proyectos. Les agradezco de todo corazón por confiar en mi en todo momento, motivándome a nunca quedarme en mi zona de confort.

A mis amigos, que me apoyaron siempre con optimismo, en especial a Rochi, Lu y Mai, por su amistad incondicional en todos estos años. También a todas las personas que me crucé a lo largo de este camino, en especial a toda la gente que me dejó el SABF, un lugar que me brindó un espacio de introspección y crecimiento profesional y personal, además de íntimos amigos.

A Copi, mi compañero de vida, gracias por toda la paciencia, tardes de estudio, y esfuerzo compartidos. Por estar ahí, al pie del cañón, siempre con una sonrisa.

Por último, quería agradecer especialmente a Ceci, a quien la carrera me cruzó casi por casualidad, pero gracias a ella pude coronarla acompañada no sólo de una excelente profesional, sino también de una gran amiga. Gracias confiar en mi para emprender este proyecto y por las innumerables tardes de risas y catarsis.

Sin todos ustedes, esto no hubiese sido lo mismo. ¡Muchísimas gracias!

Luciana

En primer lugar quiero agradecer a toda mi familia: mis padres, abuelos, tíos, primos y Arturo, quienes me acompañaron a lo largo de toda mi carrera, en las alegrías y dificultades. Especialmente a mis abuelos por inspirarme con su valentía y optimismo. Agradezco enormemente a mi mamá por su generosidad y apoyo incondicionales, y por creer en mí incluso en los momentos en que ni yo creía. A mi papá, mi mayor inspiración para elegir una carrera en las ciencias, y quien sé que estaría orgulloso de este momento. Sus consejos me acompañan todos los días.

A mis amigas de la vida Delfi, Delfi, Vicky y Eli, y a Fer, que fue una señal en mi camino. A Ayelén y Tomás, que son primos, hermanos y amigos. A todos los amigos que conocí en mis años de estudio y que, de una manera u otra, hicieron de ésta una etapa inolvidable. A mis compañeros de trabajo por escucharme, ayudarme y hacerme reír en estos últimos meses tan intensos. Y muy especialmente a Lu, a quien me crucé por casualidad y con el tiempo supo convertirse en mi amiga y compañera de equipo en esta aventura.

A Santi, por estar siempre al lado mío. Con su alegría y paciencia me enseña todos los días a ser una mejor persona.

Y, por último, a mi gato por ser mi compañera incondicional de tardes y noches de estudio.

María Cecilia

Índice general

Índice de figuras

Índice de tablas

Introducción

1 Generalidades del cáncer

El cáncer es un conjunto de más de 100 enfermedades caracterizadas por la **proliferación descontrolada de células anormales** en algún tejido del cuerpo que eventualmente pueden diseminarse e invadir tejidos cercanos, el sistema linfático y otros órganos a través del torrente sanguíneo [1]. Si bien el cáncer puede desarrollarse prácticamente en cualquier tejido del cuerpo humano y en cada uno presenta características particulares, los procesos básicos de la tumorigénesis son similares en todos los casos: señalización de crecimiento autosuficiente, **inmortalidad replicativa**, evasión de señalización supresora de crecimiento celular, resistencia a la apoptosis, **inducción de angiogénesis** y activación de invasión y **metástasis** [2, 3].

En las últimas décadas, múltiples grupos de investigación se han abocado a dilucidar las bases moleculares de la carcinogénesis. Como uno de los resultados principales, se ha identificado a las **mutaciones genéticas** como factores clave en el proceso de expansión clonal del crecimiento tumoral [4], siendo el cáncer de mama, con los genes BRCA 1 y 2, un ejemplo de estos avances [5]. Aunque este tipo de descubrimientos han tenido un gran impacto en el entendimiento, la prevención y el tratamiento del cáncer a nivel mundial, esta enfermedad aún presenta una complejidad subyacente que los sistemas de salud no han podido resolver [6].

De acuerdo con la Organización Mundial de la Salud (OMS), en el 2018 se produjeron 9,6 millones de muertes por causas relacionadas con cáncer, el 70% de ellas en países de ingresos bajos y medio-bajos [7]. Por su parte, la Argentina es un país con mortalidad por cáncer media-alta, con 68 700 muertes en 2018, de acuerdo con las cifras del Instituto Nacional del Cáncer (INC) [8].

El mayor agravante de esta situación son las predicciones para la incidencia y mortalidad de la enfermedad: se estima un **aumento del 63,4% en la incidencia y 71,5% en la mortalidad** a nivel mundial, y del **47,8% y 53,9% en la Argentina**, respectivamente [9].

2 Cáncer de Endometrio

El útero es un órgano muscular hueco, normalmente del tamaño de un puño, situado en la pelvis menor, entre la vejiga y el recto. Es uno de los principales órganos del aparato reproductor femenino: su función es recibir al blastocisto para su implantación y, posteriormente, la nutrición y preservación del embrión y el feto en el transcurso del embarazo. La **Figura 1.1** esquematiza la anatomía del útero. Éste presenta cuatro regiones diferenciadas: el **cuerpo**, el fondo (zona donde se insertan las trompas de Falopio), el **cérvix** (o cuello uterino, conecta el interior del útero con la vagina) y el itsmo (ubicado entre el cuerpo y el cérvix). El tejido graso y conjuntivo que rodea al cuerpo uterino es el **parametrio**, y las tres capas que componen el cuerpo del útero se denominan:

- **Endometrio**: es una mucosa glandular que recubre internamente el útero; consiste en un epitelio simple altamente vascularizado. Durante el ciclo menstrual, el endometrio presenta alteraciones cíclicas en sus glándulas y vasos sanguíneos en preparación para la implantación de un blastocisto.

- **Miometrio**: es una capa muscular intermedia formada por músculo liso, ubicada entre el endometrio y la serosa, que constituye la mayor parte de la pared uterina. Es el tejido más flexible del cuerpo humano y, durante el parto, tiene la función de contraerse para la expulsión del feto por el canal cervical.

- **Serosa**: también llamada perimetrio, es una membrana epitelial que recubre externamente al útero.

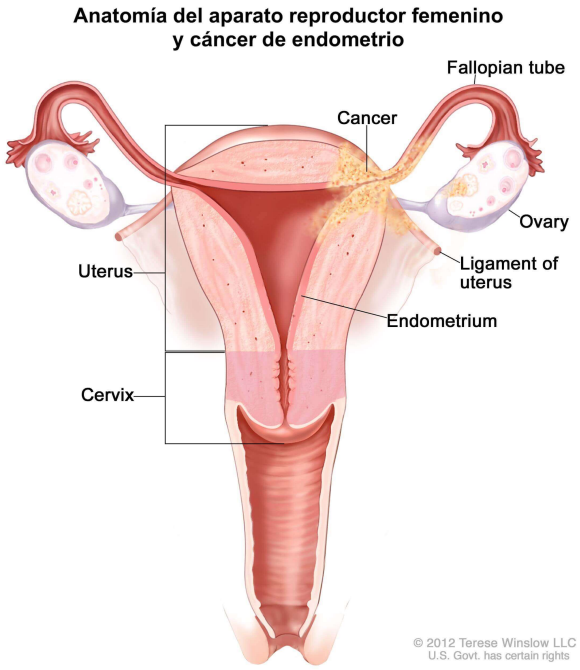


Figura 1.1. Diagrama anatómico del aparato reproductor femenino, en el que se distinguen las siguientes estructuras: ovarios (*Ovary)*, trompas de Falopio (*Fallopian tube*), cuerpo (*Uterus*), cérvix (*Cervix*), ligamentos uterinos (*Ligament of uterus*) y endometrio (*Endometrium*). Adicionalmente, la imagen retrata un ejemplo de neoplasia de endometrio en estadio avanzado (*Cancer*). La Sección 2.4 de este capítulo proporciona más información acerca de la estadificación y clasificación del CE). Fuente de la imagen: *National Cancer Institute*, EEUU (2018) Terese Winslow LLC

El Cáncer de Endometrio (CE) es el tipo de **cáncer de cuerpo uterino** más frecuente[1]. Se origina cuando algunas células del endometrio proliferan sin control sobre el resto del tejido e incluso sobre otros tejidos. El CE afecta principalmente a mujeres postmenopáusicas y la edad promedio de detección primaria es de 63 años [10].

La mayor parte de las veces, el CE se detecta tempranamente (es decir, cuando el o los tumores aún presentan bajo grado y estadio temprano) gracias a su principal síntoma: el Sangrado Uterino Anormal (SUA). Concretamente, el 80% de las neoplasias son diagnosticadas en estadio I y poseen una tasa de sobrevida a 5 años del 95%. Por el contrario, el pronóstico es especialmente desfavorable cuando existe diseminación local o metástasis de las neoplasias (sobrevida a 5 años del 68% y 17%, respectivamente) [11]. En aquellos casos con tumores agresivos y de alto grado, la enfermedad progresa rápidamente y se espera que las pacientes presenten metástasis en el transcurso

[1]El otro tipo de cáncer de cuerpo uterino es el **sarcoma uterino**, que se origina en el miometrio o el tejido conectivo del útero; los tumores de este tipo son muy infrecuentes.

de aproximadamente un año a partir de la detección [12].

2.1 Incidencia y mortalidad

El CE es el cuarto cáncer más frecuente en mujeres en EEUU [11] y el séptimo en la Argentina [8]. Además, es la primera neoplasia ginecológica en prevalencia en países desarrollados, y la segunda en países en vías de desarrollo, siguiendo al cáncer de cuello de útero [13]. En la **Figura 1.2** se representa gráficamente la situación actual (año 2018) de incidencia y mortalidad por cáncer en Argentina.

La **Tabla 1** presenta las cifras a nivel mundial y en Argentina de incidencia de cáncer según el tipo de tejido primario. Éstas corresponden al año 2018 y permiten apreciar que, en ambos casos, aproximadamente el 2% (2,11% y 1,87%, que surgen de las cifras 382 069/18 078 957 y 2 412/129 047 en el mundo y Argentina, respectivamente) de los nuevos casos detectados corresponden a neoplasias del cuerpo uterino. Por su parte, la **Tabla 2** lista la cantidad de muertes de pacientes por tipo de cáncer en el mundo y en Argentina en el mismo período. En este caso, el cáncer de cuerpo uterino representa cerca del 1% del total en ambas poblaciones (0,94% y 1,3%, respectivamente, provenientes de calcular 89 929/9 555 027 y 895/68 778).

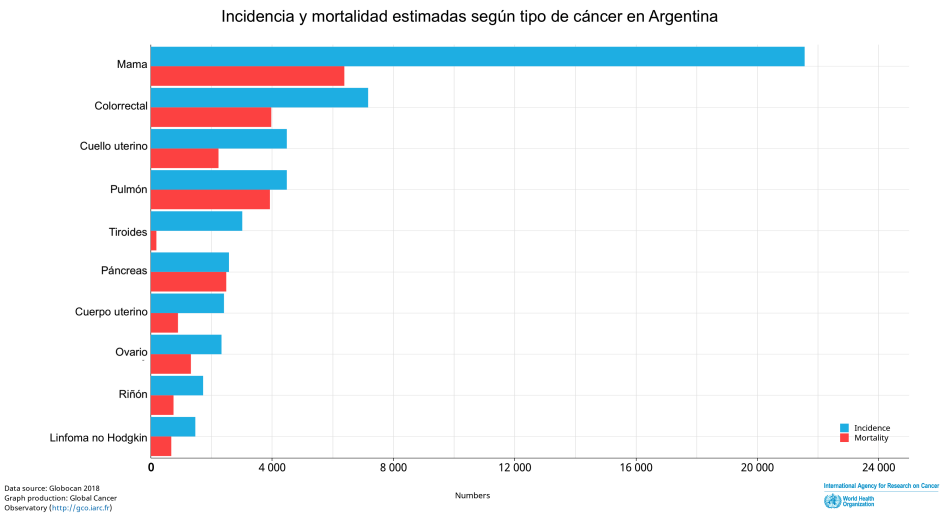


Figura 1.2. Incidencia y mortalidad por cáncer en mujeres en Argentina, año 2018. Gráfica de incidencia y mortalidad de los 10 cánceres más prevalentes en mujeres en la Argentina. El de cuerpo uterino, que incluye al CE, se encuentra en 7ª posición en incidencia, con casi 2 500 nuevos casos y aproximadamente 1 000 muertes en 2018. Fuente: *Globocan* 2018, *International Agency for Research on Cancer*.

Incidencia de cáncer

Tipo de Cáncer	Número de casos	
	Mundo	Argentina
Todos	18 078 957	129 047
Pulmón	2 093 876	11 595
Mama	2 088 849	21 558
Colorrectal	1 849 518	15 692
Próstata	1 276 106	11 600
Estómago	1 033 701	3 980
Hígado	841 080	2 343
Esófago	572 034	2 299
Cérvix	569 847	4 484
Tiroide	567 233	3 482
Vejiga	549 393	3 631
Páncreas	458 918	4 878
Riñón	403 262	4 889
Cuerpo uterino	**382 069**	**2 412**
Ovario	295 414	2 330
Testículos	71 105	1 724
Vulva	44 235	355
Pene	34 475	352
Otros	2 316 378	14 667

Tabla 1: Incidencia de cáncer en el mundo y Argentina en 2018. En la tabla se presenta una estimación de nuevos casos en el mundo y Argentina en 2018, clasificados según tipo de tumor. Se destaca la categoría cuerpo uterino, que incluye al CE. Fuente: Globocan 2018, *International Agency for Research on Cancer*.

Mortalidad por cáncer

Tipo de Cáncer	Número de casos	
	Mundo	Argentina
Todos	9 555 027	68 778
Pulmón	1 761 007	10 662
Colorrectal	880 792	8 721
Estómago	782 685	3 202
Hígado	781 631	2 113
Mama	626 679	6 38
Esófago	508 585	1 891
Páncreas	432 242	4 683
Próstata	358 989	3 974
Cérvix	311 365	2 231

Leucemia	309 006	2 098
Cerebro, SNC	241 037	1 616
Vejiga	199 922	1 599
Ovario	184 799	1 321
Riñón	175 098	2 314
Cuerpo uterino	**89 929**	**895**
Melanoma de piel	60 712	592
Tiroide	41 071	265
Vulva	15 222	159
Pene	15 138	131
Otros	1 062 437	6 315

Tabla 2: Mortalidad por cáncer en el mundo y Argentina en 2018. En la tabla se presenta una estimación de muertes por cáncer en el mundo y Argentina en 2018, clasificados según tipo de tumor. Se destaca la categoría cuerpo uterino, que incluye al CE. Fuente: *Globocan* 2018, *International Agency for Research on Cancer*.

El aumento en la incidencia del cáncer ha sido un patrón constante en las últimas décadas, tanto en la población mundial como en la Argentina. Este fenómeno se presenta como consecuencia de factores ambientales como la polución, médicos como el incremento en el porcentaje de mujeres con obesidad y enfermedades asociadas, y sociales como el mayor acceso a prestaciones médicas y el aumento en la expectativa de vida de la población [14, 15].

Actualmente, la incidencia estimada para cáncer de cuerpo uterino en el mundo [expresada en *Age-Standardized Rate* (ASR)[2]] es de 8,4 casos cada 100 000 mujeres [9]. Por su parte, la tasa de mortalidad es de 1,8 [9]. Los mapas de las **Figuras 1.3** y **1.4** presentan gráficamente los datos de incidencia (en color azul) y mortalidad (en rojo) para cáncer de cuerpo uterino de todo el mundo. Estas visualizaciones son útiles para entender mejor la epidemiología de la enfermedad: mientras que la Argentina presenta una incidencia moderada (7,6) y similar a la media mundial, los países desarrollados tienen tasas significativamente superiores. Respecto de la mortalidad, la Argentina presenta una tasa moderada a elevada (2,3) y superior a la media mundial. Las tasas más elevadas están distribuidas entre países de ingresos diversos en los cinco continentes: Bahamas, Letonia, Emiratos Árabes Unidos y Zimbabue tienen los mayores indicadores.

[2]ASR es una medida de resumen que refleja la cantidad de casos anuales cada 100 000 personas. Se calcula normalizando la población en estudio a una estructura etaria estándar.

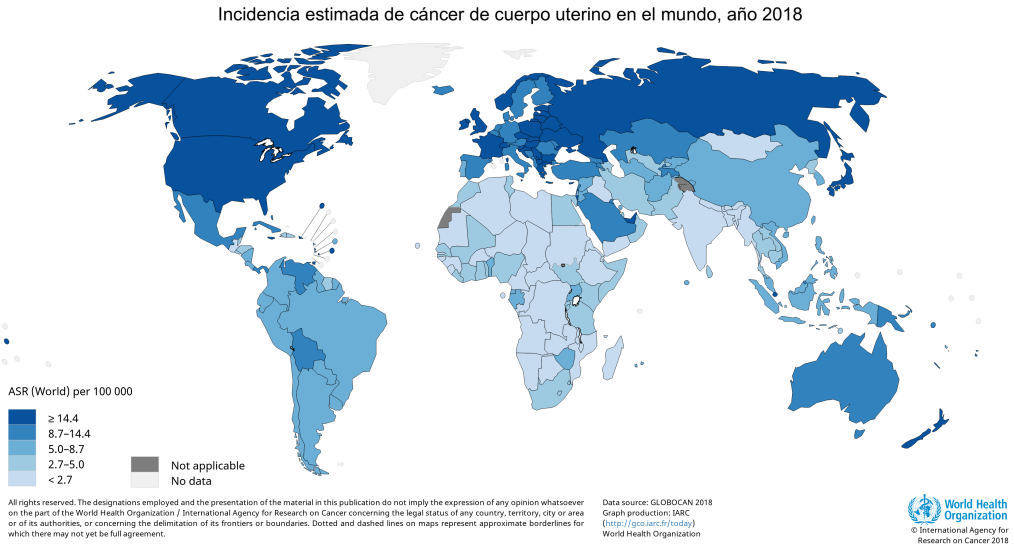


Figura 1.3. Incidencia de cáncer de cuerpo uterino a nivel mundial, año 2018. La figura representa las tasas de incidencia (ASR) de cáncer de cuerpo uterino cada 100 000 mujeres en el mundo en 2018. La Argentina presenta una tasa de 7,6 casos de cáncer de cuerpo uterino, que incluye al CE. Fuente: *Globocan* 2018, *International Agency for Research on Cancer*.

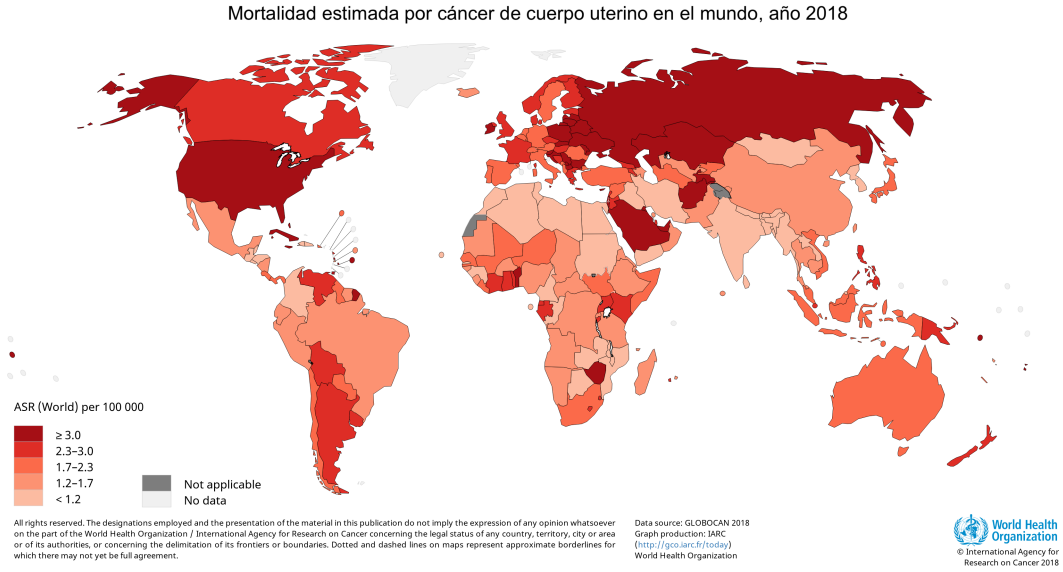


Figura 1.4. Mortalidad por cáncer de cuerpo uterino a nivel mundial, año 2018. La figura representa las tasas de mortalidad (ASR) cada 100 000 mujeres en el mundo en 2018. La Argentina presenta una tasa de 2,3 muertes por cáncer de cuerpo uterino, que incluye al CE. Fuente: *Globocan* 2018, *International Agency for Research on Cancer*.

Adicionalmente, de acuerdo con la información disponible en la página web del Observatorio Global del Cáncer (Globocan) se prevé un aumento del 52,7% en la incidencia y del 70,6% en mortalidad femenina por cáncer de cuerpo uterino en los próximos 21 años a nivel mundial. En Argentina, las proyecciones de la misma fuente estiman un aumento del 44,9% y 50,9%, respectivamente.

Las estadísticas epidemiológicas anteriores reflejan claramente la relevancia del CE en los sistemas de salud, tanto en el mundo como en la Argentina. A su vez, plantean al nivel de desarrollo e ingresos de cada país como una variable de peso en la incidencia y mortalidad del cáncer en general, y del CE en particular. Esta temática ha sido abordada por gran cantidad de autores que establecen que las desigualdades de ingresos, riqueza, educación y poder impactan en los individuos, comunidades y países, produciendo gradientes sociales en la incidencia y mortalidad del cáncer tanto dentro de un país como entre países en el mundo [16]. La gráfica de la **Figura 1.5** acompaña este análisis y refleja cómo, si bien la incidencia es significativamente mayor en países con alto Índice de Desarrollo Humano (IDH), la proporción mortalidad-incidencia es cerca de 3 veces mayor en países con bajo IDH.

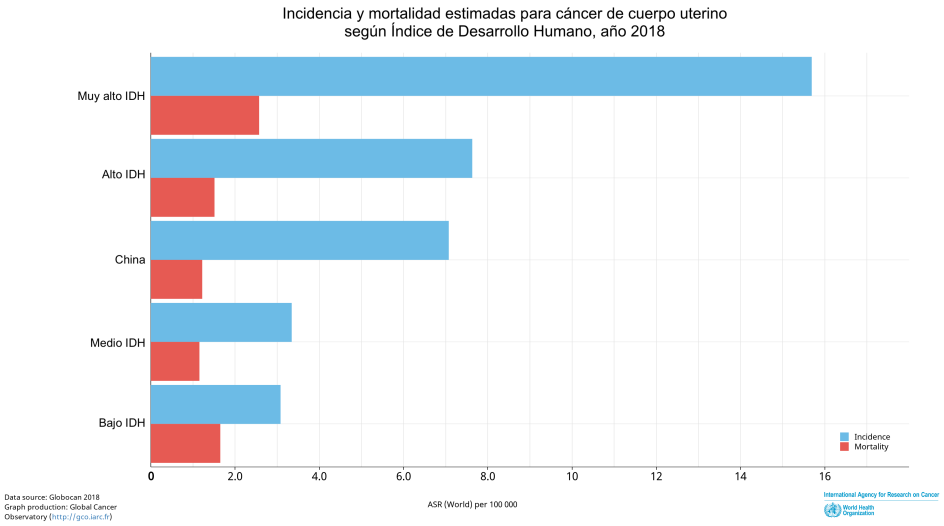


Figura 1.5. Incidencia y mortalidad por cáncer de cuerpo uterino en el mundo según Índice de Desarrollo Humano (IDH). La gráfica presenta cifras de incidencia y mortalidad por cáncer de cuerpo uterino según el IDH de las Naciones Unidas. Aunque la incidencia disminuye conforme lo hace el IDH, la mortalidad en países con bajo IDH es superior a la de países con alto IDH, y la proporción incidencia/mortalidad es significativamente mayor en los países con bajo IDH. Fuente: *Globocan* 2018, *International Agency for Research on Cancer.*

2.2 Factores de riesgo y protectores

A continuación se presentan factores cuya asociación favorable o desfavorable con el CE ha sido comprobada clínicamente.

En primer lugar, según la OMS, un **factor de riesgo** es cualquier rasgo, característica o exposición de un individuo que aumente su probabilidad de sufrir una enfermedad. Puntualmente para el CE, el Consenso Nacional Inter-Sociedades del Programa Argentino de Consensos de Enfermedades Oncológicas relaciona los factores de riesgo con la exposición aumentada a largo plazo a estrógenos, ya sea exógenos o endógenos [13]. Algunos de los más importantes son:

- **Menarca temprana y/o menopausia tardía**: implican mayor cantidad de ciclos menstruales y más años de exposición hormonal [17].
- El **estado postmenopáusico**: en esta etapa las pacientes no producen niveles compensatorios de progesterona por lo que hay una condición de hiperestrogenismo [18].
- **Nuliparidad**: el embarazo disminuye los niveles de estrógeno en sangre, por lo que las mujeres con al menos un embarazo tienen menor exposición que aquellas sin embarazos [19].
- **Anovulación**: las irregularidades en el ciclo menstrual, la ovulación infrecuente y la anovulación se asocian a mayores niveles de estrógeno y una deficiencia de progesterona [18].
- El uso de **estrógenos sin oposición progestacional**: se indican para tratamientos de largo plazo en mujeres menopáusicas y aumentan en más 6 veces el riesgo de CE [20].
- El tratamiento con **tamoxifeno**: es un fármaco estándar para algunos tipos de cáncer de mama que induce hiperestrogenismo en el útero. Incrementa 4 veces el riesgo de CE en mujeres postmenopáusicas [21]. De todos modos, se sugiere evaluar la continuidad del tratamiento en función de sus efectos beneficiosos sobre el cáncer de mama.

Adicionalmente, la bibliografía coincide en los siguientes factores de riesgo:

- Un índice de masa corporal mayor a 25 mg/m^2
- La inactividad física
- Una ingesta calórica excesiva

- Presión arterial superior a 140/90
- Altas concentraciones de glucosa en sangre
- La diabetes no insulinodependiente (tipo II)

El Instituto Nacional del Cáncer (NCI, *National Cancer Institute*) de Estados Unidos agrega como factores de riesgo en CE a los genéticos. Entre ellos destaca a los síndromes de ovario poliquístico, de Cowden y de Lynch, el cáncer de colon no poliposo hereditario y una historia familiar de CE en familiares de primer grado (madre, hermanas o hijas) [19].

No obstante, hasta un 50% de las pacientes portadoras de CE son diagnosticadas sin estos factores [13]. Asimismo, el Consenso expone que, manteniendo un peso normal y siendo físicamente activas, las mujeres pueden reducir sustancialmente su riesgo de cáncer de endometrio.

Por otra parte, los **factores protectores** en salud son características o exposiciones detectables en un individuo que favorecen el mantenimiento o la recuperación de su salud. En cáncer, los factores protectores son aquellos que se asocian con una reducción en el riesgo de incidencia, recurrencia o muerte respecto de individuos sin presencia de estos factores. En particular para CE, los factores protectores se relacionan con una baja exposición a estrógenos [17, 18]. De acuerdo con el NCI, algunos de estos son [19]:

- Menarca retrasada
- El embarazo y la lactancia, en particular la pluriparidad y los períodos de lactancia prolongados
- Anticonceptivos orales combinados de estrógeno y progesterona [20]
- La práctica habitual de actividad física

2.3 Presentación clínica y diagnóstico

El síntoma más frecuente de CE y por el que se diagnostica el 90% de los casos es el **Sangrado Uterino Anormal (SUA)** o intermitente [21]. Este síntoma característico contribuye a la detección del cáncer en estadios tempranos en un 75% de los casos, lo que mejora el pronóstico de las pacientes diagnosticadas [13, 18]. En pacientes postmenopáusicas, el SUA se presenta como

metrorragia postmenopáusica; para pacientes premenopáusicas, el SUA ocurre entre ciclos menstruales [10] o en forma de menorragia. La edad promedio de detección primaria es de 63 años [10] y el 90% de los casos se presenta en mujeres mayores a 50 años, especialmente hasta los 60 [22].

Con respecto a la **prevención**, hasta el momento no existen pruebas de detección de CE en pacientes asintomáticas. Por este motivo, el Consenso sobre CE de las Sociedades Europeas de Oncología Médica (ESMO), Radioterapia y Oncología (ESTRO) y Oncología Ginecológica (ESGO) del año 2016 recomienda la difusión y concientización sobre los síntomas y factores de riesgo en todas las pacientes, **especialmente en aquellas con riesgo aumentado o SUA** [21]. En este grupo se indica adicionalmente el siguiente **protocolo diagnóstico**, ilustrado en la **Figura 1.6**:

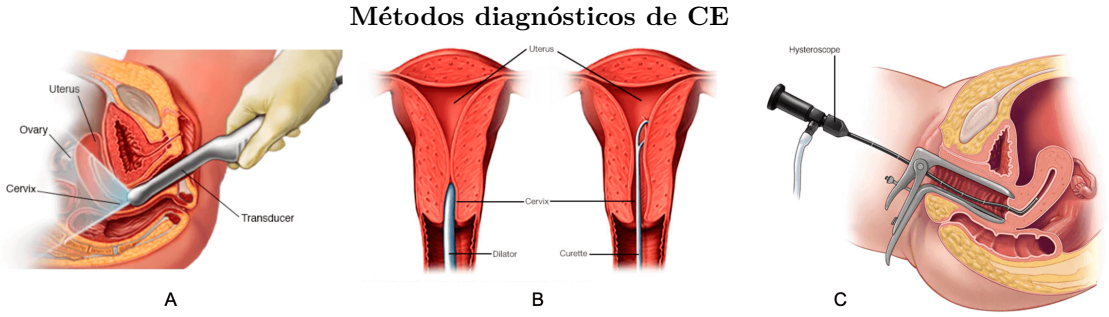


Figura 1.6. Métodos diagnósticos de CE. Esquema de los estudios a realizar en pacientes con sintomatología, riesgo y/o sospecha de CE: **A.** ecografía transvaginal, **B.** biopsia endometrial ambulatoria y **C.** legrado uterino con (o sin) histeroscopía [21]. Fuente de la imagen: *Mayo Foundation for Medical Education and Research*, 2019.

1. **Ecografía transvaginal**: es un método de alta sensibilidad y moderada especificidad diagnóstica para diferenciar espesores endometriales (por ejemplo, 90% y 54% respectivamente para el punto de corte de 5 mm) [13]. Es útil como primer estudio de detección en pacientes con sintomatología dudosa por su capacidad de discriminar aquellas que requieren estudios complementarios de las que solo necesitan controles periódicos [13]. No existe un criterio unificado sobre el espesor endometrial crítico y el procedimiento a seguir a partir de los resultados de una ecografía transvaginal. En general, en todas las pacientes con SUA a repetición se indica adicionalmente una biopsia endometrial, independientemente de los resultados de la ecografía. En pacientes postmenopáusicas con sintomatología dudosa y riesgo aumentado o factores de riesgo detectados, el punto de corte de 3 mm (también 4 o 5 mm

para algunos autores) determina la realización de la biopsia [10, 13, 23]. Por último, en pacientes postmenopáusicas con o sin factores de riesgo y asintomáticas, un espesor endometrial ecográfico mayor a 11 mm se deriva en una biopsia endometrial [21, 23].

2. **Biopsia endometrial ambulatoria**: consiste en la extracción ambulatoria de tejido endometrial por medio de la introducción de un instrumento a través del cuello uterino. Posteriormente, la muestra del tejido es examinada por un patólogo para detectar células cancerosas; si el resultado es positivo, el patólogo también procede a la determinación del tipo histológico y el grado de diferenciación tumoral [19]. La biopsia por aspiración endocavitaria es el segundo paso en la rutina de diagnóstico de CE, después de la detección de un espesor anormal en el tejido. Tiene gran utilidad en la confirmación del diagnóstico, no así en el descarte. Por este motivo, en casos de sospecha firme puede reemplazar al procedimiento quirúrgico bajo anestesia solo si el análisis patológico es positivo [13]. Actualmente, dos de los instrumentos utilizados para biopsias endometriales son las cánulas Pipelle y Vabra; estas técnicas tienen sensibilidad y especificidad de detección elevadas (99,6 y 97,1%, respectivamente [10]) pero requieren un orificio cervical interno complaciente [13] y muestras apropiadas. La obtención de muestras puede presentar dificultades, por lo que en algunos casos es necesario recurrir al procedimiento de dilatación y raspado, con o sin histeroscopía [18].

3. **Legrado uterino fraccionado con o sin histeroscopía**: este procedimiento extrae una muestra de tejido del interior de la cavidad uterina. En primer lugar, el cuello uterino se dilata por medios físicos o farmacológicos. Posteriormente se introduce un instrumento quirúrgico llamado legra que remueve una porción del revestimiento endometrial; el legrado puede realizarse tanto por aspiración como por raspado del tejido [22]. En ambos casos, la muestra obtenida es analizada por un patólogo. Adicionalmente, el procedimiento de dilatación y legrado puede ser acompañado por una **histeroscopía**, esto es, la introducción de una pieza instrumental con luz y una cámara para inspeccionar visualmente el revestimiento uterino. La histeroscopía también extrae una porción de tejido para su posterior análisis. Este método diagnóstico tiene sensibilidad y especificidad elevadas (99,2% y 86,4%, respectivamente), y es más costoso, invasivo y riesgoso que el legrado sin histeroscopía [18]. De acuerdo con el Consenso Nacional Inter-Sociedades, actualmente se considera que el legrado dirigido por

histeroscopía es el *gold standard* en la evaluación diagnóstica del carcinoma endometrial, ya que presenta como ventaja la posibilidad de determinar la localización y extensión del tumor, y seleccionar el sitio más adecuado para la toma de la muestra [13].

2.4 Clasificación y estadificación de tumores

La sección anterior describe los algoritmos de diagnóstico de CE para distintos escenarios de sintomatología, edad y factores de riesgo. Una vez diagnosticada la presencia de tejido tumoral, es necesario obtener más información sobre el cuadro, ya que el CE abarca un conjunto de neoplasias histológica, biológica y patológicamente muy diverso con tratamientos y pronósticos diferentes.

La estadificación final de la enfermedad es siempre quirúrgica [13, 22], pero existen instancias previas que contribuyen a su caracterización preoperatoria. Entre ellas, la evaluación del cuello uterino mediante un raspado endocervical o una biopsia cervical, un examen físico general de las regiones ganglionares inguinales y supraclaviculares, una tomografía computada de tórax y otra de abdomen y pelvis con contraste, y un laboratorio completo con hepatograma [13]. El objetivo de estos estudios complementarios al diagnóstico es definir, junto con la estadificación quirúrgica, el tipo histológico y grado de los tumores, el sitio anatómico y la extensión de la enfermedad (es decir, identificar la extensión local del tumor y la existencia de metástasis a distancia) [18]. A continuación se describen las principales características del CE utilizadas para su clasificación.

2.4.1 Clasificación histopatológica

Desde un punto de vista histológico existe una diferencia entre las lesiones precursoras ("precursores") y los tumores ("carcinomas") de endometrio. Los **precursores** son hiperplasias, es decir, procesos patológicos de proliferación celular exagerada. Afectan tanto a las células epiteliales como a las del estroma endometrial y son producto de la exposición estrogénica prolongada sin oposición progestacional [13]. Los **carcinomas** de endometrio son usualmente precedidos por lesiones precursoras. En el año 2014, la OMS, en conjunto con la Sociedad Internacional de Patología Ginecológica, difundió la siguiente clasificación de tumores y precursores epiteliales [24]:

- **Precursores**
  - Hiperplasia sin atipia

  - Hiperplasia atípica/neoplasia intraepitelial endometrial

- **Carcinomas endometriales**
  - Carcinoma endometrioide
    - con diferenciación escamosa
    - velloglandular
    - secretor
  - Carcinoma mucinoso
  - Carcinoma seroso intraepitelial endometrial
  - Carcinoma seroso
  - Carcinoma de células claras
  - Tumores neuroendócrinos
  - Adenocarcinoma mixto
  - Carcinoma indiferenciado
  - Carcinoma desdiferenciado

Asimismo, los carcinomas endometriales se clasifican en dos grupos a partir del perfil histopatológico del tumor [13, 24, 25]. El **Cáncer de Endometrio Endometroide (CEE)** designa a carcinomas endometroides estrógeno-dependientes o de tipo I, mientras que el **Cáncer de Endometrio No Endometrioide (CENE)** refiere a carcinomas no endometroides o de tipo II (seroso, mucinoso, de células claras, etc) [10]. El CEE se asocia a la presencia aumentada de estrógenos y se da principalmente en pacientes jóvenes, obesas o perimenopáusicas. Suelen ser tumores de bajo grado y tener antecedentes de hiperplasias [22]. Por otra parte, los tumores de tipo CENE son primariamente serosos y de alto grado, se dan principalmente en pacientes de edad avanzada, sin obesidad y postmenopáusicas [12], y se asocian a endometrio atrófico, no así a hiperestrogenismo [10]. El tipo I representa la mayor parte de los casos: aproximadamente el 80% de los diagnósticos de CE corresponden a tumores CEE [10]. El pronóstico y resultado clínico de los tumores tipo I son más favorables que los de tipo II [22]. La **Tabla 3** resume las principales diferencias entre los dos grupos.

Características principales según clasificación histopatológica

Característica	CEE (tipo I)	CENE (tipo II)
Estado menopáusico de la paciente	Pre y perimenopausia	Postmenopausia
Estímulo estrogénico	Presente	Ausente
Hiperplasia	Presente	Ausente
Grado histológico	Bajo	Alto
Invasión Miometrial	Mínima	Extensa
Subtipo histológico	Endometrioide	Seroso
	Mucinoso	Células claras
	Velloglandular	Indiferenciado
		Endometrioide Grado 3
		Carcinosarcoma

Tabla 3: Clasificación histopatológica dual de CE. Propuesta por Bokhman en 1983, esta clasificación divide a los tumores en dos grupos según sus características histológicas, moleculares y clínicas: CEE o tipo I y CENE o tipo II. Fuente: [13].

2.4.2 Grados de diferenciación

El CE admite la clasificación de los tumores por grado de diferenciación. El grado de un tumor es una descripción histopatológica del tejido, que se asocia con su agresividad y la rapidez de propagación de las células cancerosas [25]; por este motivo, la gradación histológica tiene un impacto importante en el pronóstico y es considerada un indicador sensible de la enfermedad [13, 22, 26].

El CE presenta tres grados de diferenciación. En los tumores de Grado 1 (G1) la mayor parte (hasta el 95%) del tejido tumoral es similar al tejido normal y las células cancerosas forman glándulas [11, 25]. Este tipo de tumores reciben el nombre de "**bien diferenciados**" y tienden a crecer y diseminarse lentamente. Los tumores de Grado 2 (G2) son moderadamente diferenciados: hasta el 50% del tejido tumoral es glandular [25, 26]. Por último, los tumores de Grado 3 (G3) son escasamente diferenciados o indiferenciados ya que carecen de las estructuras del tejido normal; las células presentan estructuras atípicas y se reproducen descontroladamente. Estos tumores son de crecimiento y diseminación rápidos, y tienen un pronóstico menos favorable que los de menor grado [11, 22].

El grado del tumor se relaciona con otras características clínicas del CE, como la invasión miometrial y el estadio, descriptos a continuación.

2.4.3 Estadificación

El estadio del CE describe la extensión del cáncer y el nivel de diseminación de las células cancerosas; a partir de 1988 su determinación definitiva es quirúrgica, y se basa en factores como la localización anatómica del tumor primario, su tamaño y si están afectados los ganglios linfáticos regionales y distantes.

La cirugía de estadificación de inicio es el procedimiento por el que se asocia el caso clínico de una paciente con un estadio del sistema Federación Internacional de Ginecología y Obstetricia (FIGO). El examen pélvico y las ecografías son métodos diagnósticos que permiten establecer tentativamente el estadio del CE, pero no tienen resolución suficiente para evaluar el compromiso de los ganglios linfáticos [27]. Por este motivo es necesario que, en todos los casos, el médico compruebe la extensión del cáncer mediante una cirugía y extraiga muestras de tejido a ser analizadas por un patólogo [13, 22].

Adicionalmente, el estadio se basa en la invasión histológica del carcinoma. Esto es, las células cancerosas se pueden diseminar por el tejido como un frente amplio expansivo o en forma difusa, y en todos los casos se puede medir el porcentaje de tejido afectado por el carcinoma respecto del total del endometrio para definir la **Invasión Miometrial (IM)** [28]. Este parámetro se correlaciona con el grado de diferenciación del tumor: las lesiones de G1 no presentan infiltración miometrial, mientras que en las de G3 esta suele ser profunda [13]. En las clasificaciones se utiliza el 50% como punto de corte, donde una IM menor se asocia a un mejor pronóstico [13].

Tradicionalmente, la estadificación del CE se efectúa de acuerdo con el sistema propuesto por la FIGO. Este trabajo se basa en la versión más actualizada, publicada en el año 2009 (**Tabla 4**) pero es importante notar que una parte significativa de la bibliografía sobre el tema aún responde a la clasificación anterior, del año 1988. Según el sistema FIGO, el CE puede clasificarse en cuatro estadios, denominados con números romanos. Las **Figuras 1.7** y **1.8** esquematizan la localización, IM y propagación a ganglios y tejidos de todos los estadios descriptos por el sistema FIGO del año 2009.

Es común hallar en la literatura la clasificación de los tumores en **estadio temprano y avanzado**, siendo el primeros los estadios I y II y los últimos, III y IV. Este trabajo adopta esta convención.

En la clínica, aproximadamente el 75% de los nuevos casos son detectados en estadio I, 12% en estadio II, 13% en estadio III y 3% en estadio IV [22]. En pacientes con estadios tempranos, es decir I y II, la tasa de sobrevida a 5 años es alta: 85-90% y 75-85% respectivamente [18, 24, 25, 29]. Por el contrario, para aquellas pacientes con recurrencia de la enfermedad o estadios avanzados, la tasa de sobrevida oscila entre 20 y 65%, por lo que el pronóstico clínico es más desalentador [24, 25, 29]. El estadio IV se caracteriza por la presencia de sitios metastásicos, siendo la vagina, los ovarios y los pulmones los más frecuentes [25]. Asimismo, el CE de alto riesgo engloba presentaciones clínicas heterogéneas que incluyen subtipos CEE y CENE, y estadios desde IB G3 hasta IVB [22]. Por este motivo, una correcta estadificación es esencial para el abordaje de la enfermedad y la optimización del pronóstico de las pacientes [12].

Sistema de Estadificación de CE - FIGO 2009

Estadio I	Cáncer confinado al útero y glándulas cervicales. Sin propagación fuera del cuerpo uterino: cuello uterino, ganglios linfáticos o sitios distantes.
IA	IM menor al 50%
IB	IM mayor al 50%
Estadio II	Cáncer propagado desde el endometrio hacia tejido conectivo de soporte del cuello uterino. Sin propagación fuera del útero.
Estadio III	Cáncer propagado fuera del útero, aunque no alcanza el recto o la vejiga urinaria. Sin propagación a ganglios linfáticos ni sitios distantes.
IIIA	Propagación hacia la serosa y/o las trompas de Falopio u ovarios.
IIIB	Propagación hacia la vagina y/o parametrio.
IIIC1	Propagación hacia ganglios linfáticos de la pelvis.
IIIC2	Propagación hacia ganglios linfáticos paraaórticos
Estadio IV	Cáncer propagado a la vejiga urinaria, intestinos o metástasis distantes
IVA	Propagación al revestimiento interior del recto, vejiga urinaria o ganglios linfáticos.
IVB	Propagación a ganglios linfáticos inguinales o metástasis en abdomen, epiplón u órganos distantes (pulmones, hígado, huesos, etc).

Tabla 4: Sistema de estadificación de CE de la Federación Internacional de Ginecología y Obstetricia (FIGO), 2009. Los tumores se clasifican en cuatro estadios de acuerdo con tres factores: tamaño, nódulos y metástasis. Fuente: [25].

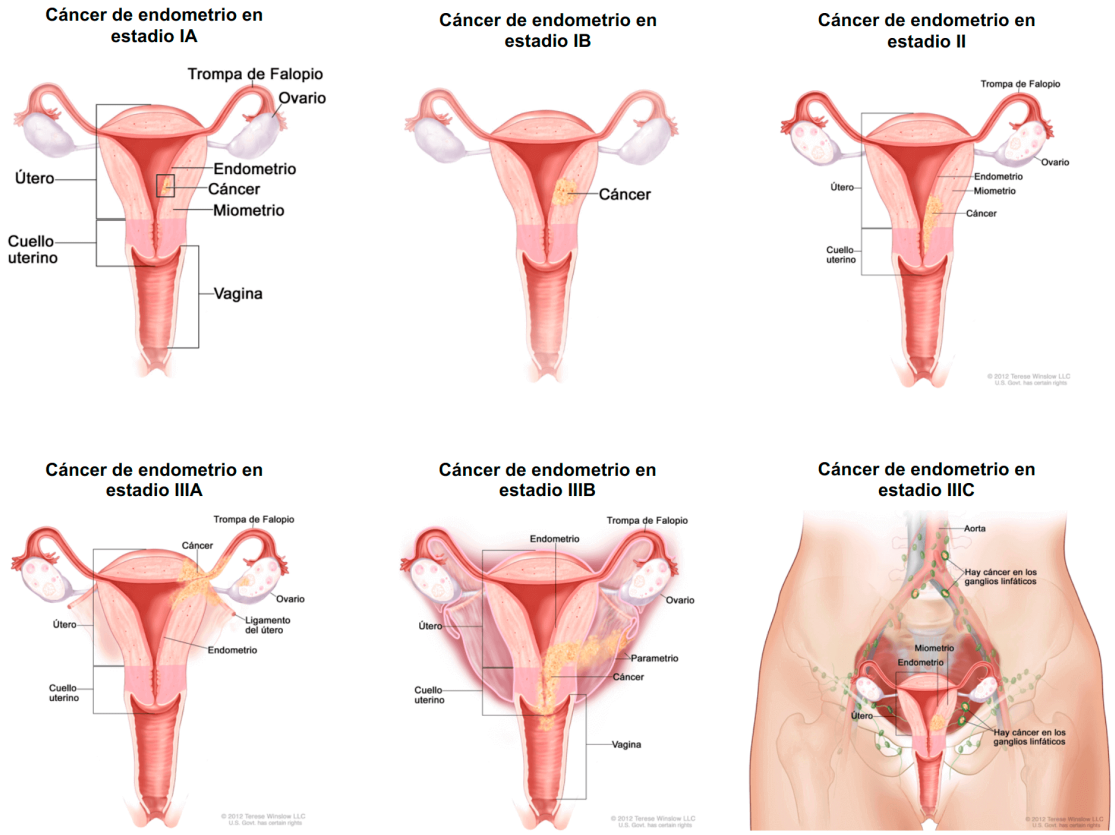


Figura 1.7. Representación del tamaño, localización y propagación del CE en estadios I, II y III según el sistema FIGO 2009. El estadio I presenta tumores confinados al cuerpo uterino, con IM variable, mientras que en el estadio II el cáncer alcanza todo el útero. El estadio III se caracteriza por la propagación de tumores a las trompas de Falopio, ovarios, vagina o ganglios linfáticos pélvicos. Fuente de la imagen: *Mayo Foundation for Medical Education and Research*, 2019.

2.5 Tratamiento

2.5.1 Abordaje quirúrgico

En todos los casos el tratamiento del CE se basa en terapias quirúrgicas [21]. El tratamiento primario del carcinoma endometrial consiste en una histerectomía total o radical con salpingooforectomía bilateral.

De acuerdo con el NCI, en una **histerectomía total** se extirpan el útero y el cuello uterino, mientras que en una histerectomía total con **salpingooforectomía** se extirpan el útero, uno de los ovarios y una de las trompas de Falopio si es unilateral, o el útero, ambos ovarios y ambas trompas de Falopio si es bilateral. Este procedimiento se indica en estadios tempranos, ya que

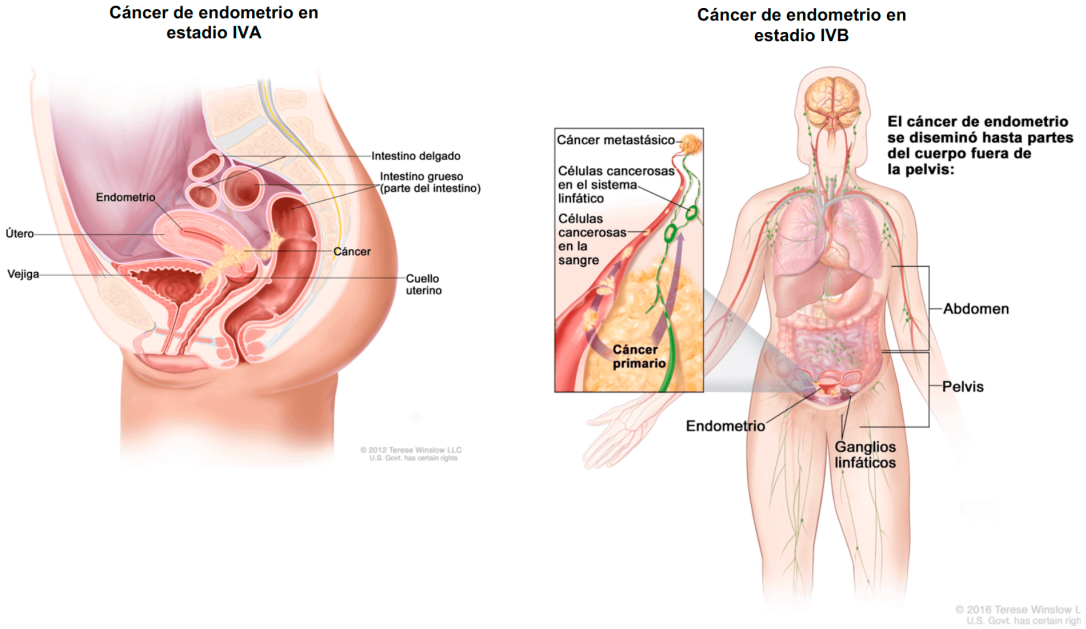


Figura 1.8. Estadios de CE IVA y IVB. En el primer caso se observa compromiso de la vejiga urinaria y recto; el estadio IVB, el más avanzado, se caracteriza por tumores metastásicos en órganos y tejidos fuera de la pelvis. Fuente de la imagen: *Mayo Foundation for Medical Education and Research*, 2019.

el cáncer está confinado al cuerpo del útero y no hay compromiso del cuello uterino, vagina ni recto. El objetivo de la salpingooforectomía bilateral es prevenir el cáncer de ovario y descartar metástasis ováricas [10]; en pacientes menores de 45 años con tumores de G1, IM menor al 50% y sin antecedentes familiares de cáncer se puede considerar la preservación de los ovarios, no así de las trompas de Falopio. Esta recomendación evita la denominada "menopausia quirúrgica" y sus efectos adversos, pero no la pérdida de fertilidad de la paciente [21].

La **histerectomía radical** consiste en extirpar el útero, el cuello uterino, ambos ovarios, ambas trompas de Falopio y tejidos circundantes, como la porción superior de la vagina. Se indica lo anterior en pacientes con CE en estadio II solamente si presentan propagación evidente hacia el parametrio; en estos casos un patólogo puede evaluar intraquirúrgicamente la histología de los tejidos resecados, hasta definir márgenes de resección libres (es decir, sin células atípicas) [10].

Ambos procedimientos se realizan mediante cirugía abierta, laparotomía o histerectomía vaginal laparoscópica [22], aunque generalmente se recomienda el abordaje mínimamente invasivo para

evitar complicaciones postoperatorias [21]. En pacientes no aptas para intervenciones quirúrgicas se evalúa la realización de una histerectomía vaginal y, de no ser posible, se recurre a tratamientos hormonales o radioterapia.

El tratamiento del CE en estadios III y IV presenta una complejidad mayor a la de los estadios I y II, por lo que se recomiendan terapias multimodales. Éstas articulan la histerectomía radical y salpingooforectomía bilateral con cirugía citorreductora.

La **cirugía citorreductora** consiste en la resección quirúrgica de la mayor porción posible de tejidos tumorales. El objetivo es reducir el volumen de enfermedad residual, lo que afecta positivamente la sobrevida de las pacientes y puede mejorar los resultados de las terapias adyuvantes [21]. Este procedimiento es importante en pacientes que presentan propagación abdominal, ya sea para el tratamiento del cáncer o el alivio de sus síntomas. La citorreducción puede ser intestinal, diafragmática, hepática, pancreática, etc.

La elección del tratamiento depende del estado general del cáncer y la paciente [10]. Aquellos casos con tumores resecables quirúrgicamente (generalmente en estadio III) son candidatos a cirugías citorreductoras completas a fin de eliminar la enfermedad macroscópica residual [30]. Por otra parte, en los casos en estadio IVB con metástasis distantes se recomiendan tratamientos quirúrgicos paliativos [30]. En el 10%-15% de las pacientes con estadio IVB la intervención quirúrgica no es posible; en estos casos el tratamiento se basa en radioterapia y braquiterapia intracavitaria vaginal [10].

La **linfadenectomía** es un procedimiento quirúrgico que se aplica en todos los estadios del CE para una estadificación precisa [21, 22, 31]. Se define como la disección de ganglios linfáticos pélvicos (sobre las venas ilíacas comunes y externas) y/o paraaórticos (sobre la vena cava inferior y la arteria aorta inferior hasta las venas renales).

La linfadenectomía pélvica y paraaórtica son, junto con la histerectomía total y la salpingooforectomía bilateral, parte del procedimiento de estadificación definitiva del CE [32]. Incluso en pacientes con estadio I y G1, la omisión de la linfadenectomía podría perjudicar la toma de decisiones respecto al tratamiento [32]. Más aún, frente a la ausencia de métodos diagnósticos no invasivos con especificidad, sensibilidad, y valor predictivo positivo y negativo adecuados, la estadificación quirúrgica es actualmente el criterio de referencia para determinar la extensión del CE [22]. Aunque existe consenso en la recomendación de una cirugía de estadiaje inicial con linfadenectomía para

pacientes de alto riesgo histológico [10, 21, 22, 31], aún hay discrepancias sobre los beneficios terapéuticos de su aplicación en grupos de bajo riesgo [25].

La **Figura 1.9** resume los conceptos expuestos anteriormente sobre el abordaje quirúrgico de CE según el estadio y grupo de riesgo histológico, que comprende tanto el grado y el subtipo como la IM. Se entiende por tumores de **bajo riesgo histológico** a aquellos de subtipo CEE, G1 o G2 e IM<50%. El **riesgo histológico medio** comprende a los tumores CEE de G1 o G2 e IM>50%, como así también los de G3 e IM<50%. Por último, los tumores de **riesgo histológico alto** son los de G3 e IM>50%, y todos los de subtipo CENE (sin importar el grado o la invasión miometrial).

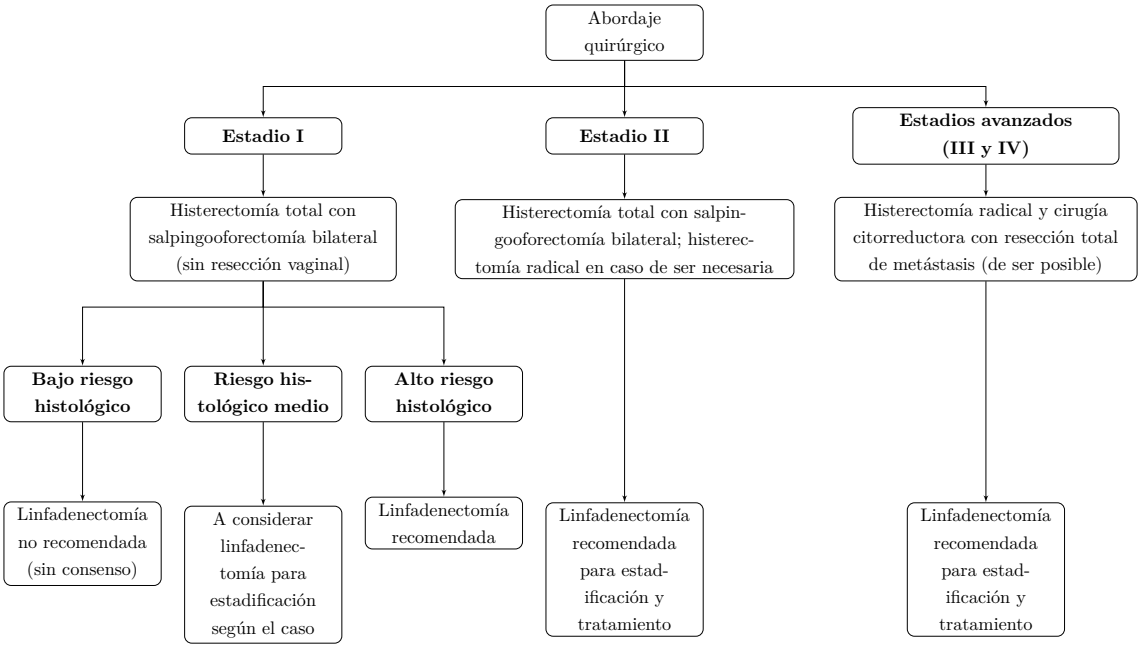


Figura 1.9. Algoritmo de abordaje quirúrgico de CE. En todos los casos el tratamiento primario es la histerectomía (total o radical), que se combina con salpingooforectomía bilateral, resección vaginal, linfadenectomía y/o cirugía citorreductora según el estadio del cáncer y riesgo histológico.
Fuente: [21, 22].

2.5.2 Terapias adyuvantes

Aproximadamente el 75% de las pacientes con diagnóstico de CE tienen tumores en estadio I que se curan luego de los procedimientos quirúrgicos de histerectomía total y salpingooforectomía bilateral [10, 22]. Aquellos casos donde el cáncer es más avanzado requieren tratamientos quirúr-

gicos adicionales (citorreducción y linfadenectomía) e incluso terapias adyuvantes[3] como radio o quimioterapia. Como sucede para el abordaje quirúrgico, la indicación de estos tratamientos depende del estadio, el grado, la IM y el tipo histológico del cáncer. En aquellos tumores o pacientes donde la intervención quirúrgica no es posible, la radio o quimioterapia representan el tratamiento primario de la enfermedad [22].

A continuación se presenta un resumen de las terapias adyuvantes aplicadas en casos de CE, de acuerdo con los algoritmos y recomendaciones del Consenso ESMO, ESGO y ESTRO del año 2016 [21, 25]. Actualmente, la indicación de terapias adyuvantes se basa en los niveles de riesgo de la enfermedad:

1. **Riesgo bajo (CEE en estadio I, G1-2 e IM nula o $< 50\%$)**: no requiere terapias adyuvantes.

2. **Riesgo intermedio (CEE en estadio I, G1-2 e IM $> 50\%$)**: se recomienda braquiterapia intracavitaria; en pacientes menores a 60 años se puede considerar la opción de no administrar terapias adyuvantes.

3. **Riesgo intermedio-alto (CEE en estadio I, G3 e IM $> 50\%$)**:

   - Si no hay propagación a ganglios linfáticos (confirmado por linfadenectomía): la indicación es similar a la de riesgo intermedio.
   - Si hay (sospecha de) propagación a ganglios linfáticos: se recomienda radioterapia externa, preferentemente junto con quimioterapia (combinada o secuencial).

4. **Riesgo alto, CEE en estadio II**: se indica radioterapia externa y refuerzo con braquiterapia vaginal; para tumores G3 y propagación linfática se recomienda acoplar con quimioterapia (combinada o secuencial).

5. **Riesgo alto, CEE en estadio III sin enfermedad residual**: recomendación de radioterapia externa y quimioterapia (combinada o secuencial).

6. **Riesgo alto, CENE**:

[3] Se define como 'terapias adyuvantes' a los tratamientos para el cáncer que se administran después del tratamiento primario para mejorar su efectividad y prevenir la progresión y/o recurrencia de la enfermedad. Fuente: Diccionario del Cáncer, NCI .

- Carcinoma seroso o de células claras: existe evidencia que respalda la indicación de quimioterapia, a combinar con braquiterapia en estadios tempranos y radioterapia externa en estadios avanzados. Sin embargo no hay un criterio unificado al respecto.
- Carcinosarcoma o tumores indiferenciados: se recomienda quimioterapia; algunos estudios sugieren beneficios al aplicar radioterapia externa.

7. **CE en estadio avanzado, no resecable o con enfermedad residual**: se indica radioterapia radical (es decir, externa y braquiterapia intrauterina) combinada con quimioterapia en pacientes con metástasis o alto riesgo de recurrencia. Adicionalmente, se puede considerar radioterapia paliativa en sitios metastásicos.

En términos generales, se desaconsejan las terapias secundarias en pacientes de bajo riesgo, especialmente en mujeres jóvenes y casos de IM menor al 50% [12]. En cuadros de riesgo intermedio con histología favorable se prefiere la braquiterapia a la radioterapia, teniendo en cuenta el compromiso entre efectividad y efectos secundarios de ambos tratamientos. La radioterapia es el criterio estándar en el tratamiento de pacientes con tumores endometrioides en estadio II y III para controlar la progresión del cáncer en la zona pélvica, especialmente en tipos endometrioides [22]. Sin embargo, no existe consenso en estadios más avanzados. Algunos autores indican que la combinación de radio y quimioterapia es el tratamiento más efectivo para maximizar las tasas de Sobrevida Libre de Recurrencia (RFS, del inglés *Recurrence-Free Survival*) (ver Sección 2.6 de este capítulo) en CE (CEE y CENE) en estadios III y IV con propagación hacia la serosa uterina y/o metástasis [25, 33], mientras que otros coinciden en la quimioterapia pero recomiendan supeditar la indicación de radioterapia al subtipo histológico [10, 21].

2.6 Tasas de sobrevida y riesgo de recurrencia

Por último, en la clínica del CE se identifican dos conceptos de gran importancia relacionados con el pronóstico de la enfermedad. Estos son **recurrencia** y **sobrevida**.

Se entiende por recurrencia a la reaparición del cáncer después de un período de tiempo durante el que no pudo ser detectado; normalmente ocurre en el transcurso de dos años después de la finalización del tratamiento del tumor primario [22]. La recurrencia del CE puede ser **regional** si ocurre en los ganglios linfáticos o tejidos del útero, vagina o colon, o **distante** si afecta órganos

alejados de la zona pélvica. En este último caso se habla de **metástasis**. Entre el 13 y el 25% de los casos de CE presentan recurrencia y metástasis [18].

Por otra parte, las tasas de sobrevida indican el porcentaje de pacientes diagnosticadas con algún tipo de CE que sobreviven durante un período de observación determinado, generalmente de 5 años. Se construyen estadísticamente a partir de estudios poblacionales y varían según el subtipo histológico, grado, estadio y profundidad de invasión del tumor, y los factores de riesgo de cada sujeto. Las tasas de sobrevida a 5 años se interpretan como la probabilidad de que, dadas las condiciones particulares del caso clínico, la paciente viva después de 5 años de la detección del CE. Esta información permite al médico analizar el pronóstico de cada paciente en función de sus características clínico-patológicas. Existen variantes de la tasa de sobrevida, siendo la Sobrevida Total (OS, del inglés *Overall Survival*) y la Sobrevida Libre de Recurrencia (RFS) dos de las más frecuentemente utilizadas. La primera corresponde a la cantidad de pacientes vivos al término del período de observación, ya sea con o sin enfermedad, mientras que la RFS es la cantidad de pacientes vivos sin recurrencia de la enfermedad en el mismo período. Por lo tanto, la RFS es una representación más específica de la progresión de la enfermedad y sus efectos clínicos que la OS [34]. En CE, la tasa de sobrevida relativa a 5 años para CE (promediando todos los estadios) es del 84% [29].

2.7 Desafíos clínicos

El principal desafío clínico en CE es que no existe un consenso para el diagnóstico y manejo quirúrgico óptimo de la enfermedad. En primer lugar, la indicación de linfadenectomía, cirugía citorreductora y terapias adyuvantes aún presenta controversias, especialmente en pacientes con estadios avanzados o CENE. Por otra parte, los estadios tempranos se tratan típicamente con histerectomía y salpingooforectomía bilateral pero aproximadamente el 20% de los casos presenta recurrencia temprana de la enfermedad [35]. Adicionalmente, el 20% de las pacientes inicialmente diagnosticadas en estadios tempranos sufre una reclasificación quirúrgica [12, 18, 22]. Por último, a pesar de la baja incidencia de CE en pacientes menores a 40 años, el diagnóstico en pacientes en edad reproductiva debe realizarse con precisión para preservar la calidad de vida (y la fertilidad cuando fuera posible) [25]. Estos puntos representan una oportunidad de mejora en el abordaje clínico del CE.

En este contexto resulta evidente la importancia de identificar con exactitud el estadio, la IM, la afectación del espacio linfovascular, el grado y el subtipo histológico del CE en las etapas iniciales del diagnóstico. Se entiende que los tumores en estadios avanzados, de grado alto (G3), invasión miometrial profunda o subtipo histológico CENE son más agresivos respecto del resto de los casos de CE por presentar bajas tasas de sobrevida y altas tasas de recurrencia temprana. Las **clasificaciones basadas en características genómicas** ofrecen una alternativa novedosa para definir la **agresividad** de los tumores de CE, lo que podría mejorar la clasificación de tumores y la estimación de riesgos, como así también disminuir los efectos secundarios gracias a la elección selectiva del tratamiento. Son de especial interés los **biomarcadores diagnósticos para la categorización de la agresividad tumoral** y los **biomarcadores pronósticos de recurrencia temprana**.

3 Biología molecular y celular

3.1 Genómica

Las células son la unidad básica de todos los seres vivos. En el cuerpo humano proveen estructura, toman y convierten nutrientes en energía y llevan a cabo funciones especializadas. Además, contienen el material hereditario del cuerpo, llamado **Ácido Desoxirribonucleico (ADN)**. La mayor parte del ADN se encuentra encerrado en el núcleo celular (ADN nuclear), pero una pequeña fracción se puede encontrar en las mitocondrias[4] (ADN mitocondrial) [36].

La doble hélice de ADN está formada por cuatro bases nucleotídicas: Adenina (A), Guanina (G), Timina (T) y Citosina (C). Las bases de ADN se aparean entre ellas, A con T y C con G, para formar unidades llamadas pares de bases (pb). Además, cada base está conectada al monosacárido desoxirribosa y unida a un grupo fosfato [37, 38]. Una base, un azúcar y un grupo fosfato unidos forman un **nucleótido**. El orden -o **secuencia**- de estas bases determina la información disponible para construir y mantener a un organismo, similarmente a como las letras del abecedario dispuestas en un cierto orden forman una palabra con significado [36, 38]. Dentro de la secuencia completa de ADN existen regiones codificantes o **exones** separadas por regiones no codificantes llamadas

[4]Las mitocondrias son estructuras celulares que convierten la energía de los nutrientes en una forma energética que la célula puede utilizar.

intrones. Tres nucleótidos consecutivos (un codón) de ADN codifican un **aminoácido** [36, 38].

Una de las propiedades elementales del ADN es su capacidad de replicarse a sí mismo mediante un complejo enzimático. Cada una de las hebras de la doble hélice de ADN funciona como molde para duplicar la secuencia de bases. Esto asegura la herencia genética en cada división celular [36, 38].

Cada región de ADN que cuenta con información genética para codificar una estructura se considera un **gen**. El tamaño medio de un gen es de 27 000 pb y en promedio cada gen tiene $10,4$ exones. Los genes están determinados por secuencias de inicio de la codificación, llamadas de iniciación, y secuencias de finalización, o secuencias de *stop*. En promedio, el ser humano cuenta con aproximadamente 25 000 genes en su ADN [38].

Sin embargo, el ADN no sintetiza proteínas directamente, sino que usa el Ácido Ribonucleico (ARN)[5] como intermediario [38]. En el proceso de **transcripción**, cada una de las hebras de la doble hélice funciona como molde para la enzima a cargo de polimerizar esta cadena. Las secuencias de iniciación y de terminación se utilizan para determinar a partir de y hasta dónde deben transcribirse las distintas secciones de la cadena de ADN [38]. En esta instancia, se transcriben a ARN tanto las secuencias codificantes como las no codificantes. Luego, el ARN se procesa en un mecanismo complejo de *splicing* que lleva a la eliminación de los intrones, y resulta así en ARN mensajero (ARNm). Por último, el ARNm se exporta al citoplasma y para ser **traducido** a proteínas. En la **Figura 1.10** se ilustra este proceso [36, 38].

[5]En el ARN, las bases nucleotídicas son iguales a las del ADN, con la diferencia de que la timina (T) se reemplaza por Uracilo (U).

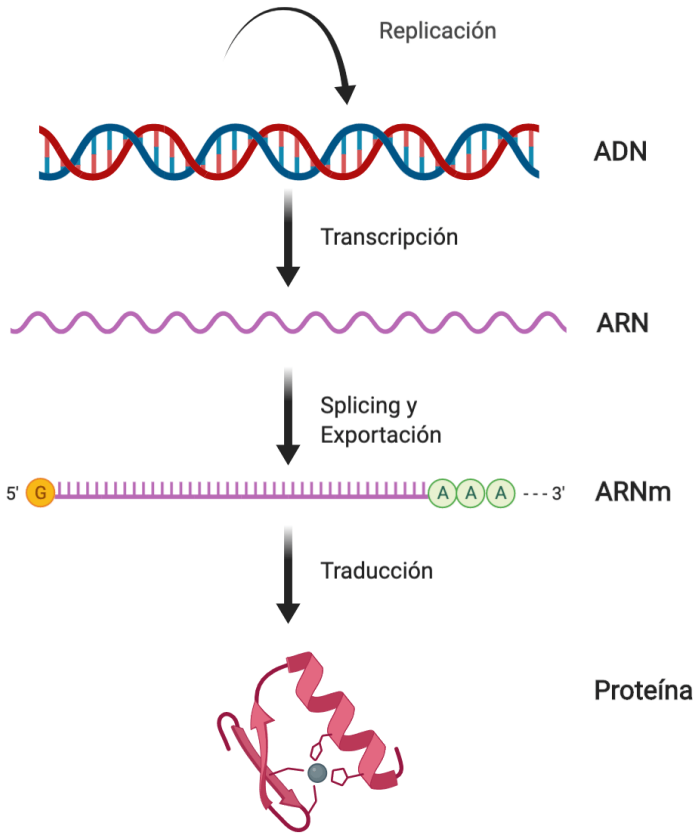


Figura 1.10. Mecanismo de traducción del ADN a proteínas. En una primera instancia, el ADN se transcribe en ARN, que luego de la transcripción sufre un proceso llamado *splicing* para eliminar las regiones no codificantes. Además, se producen cambios moleculares en los extremos del ARN y éste es exportado al citoplasma como ARNm. Por último, mediante el proceso de traducción, cada proteína codificada en ese fragmento puede ser sintetizada según las necesidades del organismo.

3.1.1 Expresión génica

Todas las células de un organismo multicelular cuentan con el mismo material genético nuclear, independientemente de su estructura o función. Durante mucho tiempo se creyó que la diferenciación celular estaba determinada por los genes que cada célula lleva en su núcleo, pero eventualmente se ha determinado que en la mayoría de los casos la diferenciación celular depende de cambios en la **expresión génica** y que los tipos celulares se vuelven distintos entre sí porque sintetizan y acumulan diferentes conjuntos de ARN y proteínas [38].

Generalmente, una célula humana expresa entre el 30% y el 60% de los genes disponibles en su núcleo. Al comparar los patrones de ARNm en distintas líneas celulares humanas, se observa que

los niveles de expresión de los genes activos varían en cada una. Los patrones de expresión de ARNm se determinan mediante diferentes abordajes que estudian genes individuales o conjuntos de genes. En este último caso, existen herramientas de evaluación global como los microarreglos de ADN y, más recientemente, la secuenciación global de ARN.

El primer reporte de **microarreglos de ADN** fue en 1995, cuando la comunidad científica enfrentaba el desafío de desarrollar un sistema para monitorear la expresión de múltiples genes en simultáneo. Los microarreglos de ADN son dispositivos de vidrio preparados mediante impresión robótica a alta velocidad de ADN complementario (ADNc) u oligonucleótidos cortos sintéticos, representativos del genoma y desarrollados para realizar mediciones cuantitativas de la expresión de los genes correspondientes a un sistema en estudio [39].

Su principal aplicación es la de obtener información relevante de los niveles de expresión génica, por ejemplo, para diferenciar subtipos de cáncer o proporcionar información pronóstica o predictiva. Por otro lado, los microarreglos de ADN se han utilizado ampliamente para caracterizar el material genético en búsqueda de mutaciones humanas, dada su facilidad para caracterizar cientos de secuencias en paralelo. Por último, permiten llevar a cabo la hibridación genómica comparativa basada en matrices, lo que brinda información de las variaciones del número de copias [40].

Por su parte, el término **RNA-Seq** engloba al conjunto de técnicas que aplican *Next Generation Sequencing* (NGS) al estudio del transcriptoma celular. Este proceso consiste en cuantificar el nivel de expresión de genes mediante el análisis de la abundancia y secuencia de transcripción del ARNm presente en una sonda. La técnica consiste en la secuenciación de tipo *high-throughput* sobre muestras de ARN y permite obtener decenas de millones de lecturas de secuencias y billones de bases individuales en pocos días (o incluso horas, dependiendo de la tecnología utilizada) [41].

Las primeras publicaciones en reportar el uso de RNA-Seq fueron [42] y [43] en 2008 y, desde entonces, su uso se ha popularizado y revolucionado el estudio del transcriptoma. Además de proveer mayor velocidad, mejor cobertura y resolución del ARN respecto de las técnicas anteriores, como la secuenciación de Sanger y los microarreglos, el RNA-Seq admite el descubrimiento de nuevos transcriptos y la identificación de variantes de *splicing* alternativo [44].

Naturalmente, los volúmenes de datos procedentes de las tecnologías de microarreglos y RNA-Seq se pueden complementar con técnicas de bioinformática y bioestadística para la interpretación de sus resultados (ver Sección 5, Introducción).

Tanto los microarreglos de ADN como la secuenciación global, permiten determinar la expresión de los genes característicos de un tipo celular, lo que se utiliza para diversas aplicaciones. Una de ellas es la tipificación de células cancerosas humanas de tejidos de orígenes desconocidos, comparándolas con los perfiles ya conocidos, como se puede observar en la **Figura 1.11**. La expresión de los genes que determinan un tipo celular puede estar regulada diferencialmente en los distintos pasos del mecanismo transcripcional previamente explicado [38]. Las alteraciones de este mecanismo conllevan una desregulación de la expresión génica, generando que ciertos productos se sub o sobreexpresen, con consecuencias importantes para la célula.

3.1.2 Mutaciones en el genoma y su relación con el cáncer

Si bien las variaciones genéticas ocasionales aumentan la probabilidad de supervivencia a largo plazo de todas las especies, para que un organismo sobreviva se requiere una cierta estabilidad genética [45]. En los distintos procesos que lleva a cabo el ADN existen mecanismos de verificación de secuencia que detectan y corrigen las alteraciones a la secuencia de nucleótidos que puedan surgir, para mitigar la variabilidad genética indeseada. Estos mecanismos tienen una baja tasa de fallos, pero cuando ocurren resultan en cambios permanentes en la secuencia de nucleótidos del ADN llamados **mutaciones**. Estas mutaciones pueden no tener efectos sobre el organismo o ser fatales si ocurren en una posición particular de la secuencia genómica [38]. Las consecuencias de una alta inestabilidad genética son evidentes en humanos, donde una pérdida en la capacidad de reparación está asociada a una mayor predisposición a ciertas enfermedades, entre ellas al cáncer [46].

Adicionalmente, las mutaciones pueden afectar a las secuencias que regulan la expresión de ciertos genes, de modo que una desregulación génica podría tener consecuencias graves para la célula o tejido. Entre ellas, enfermedades genéticas como el cáncer [47].

En condiciones normales, los genes del cuerpo humano sufren hasta 10^{10} mutaciones. Incluso en ambientes libres de agentes mutagénicos ocurren mutaciones espontáneas en una tasa estimada de 10^{-6} mutaciones por gen por división celular[6]. La evidencia indica que el desarrollo de cáncer requiere que ocurra un número sustancial de accidentes genéticos en el linaje de una célula para que esta se vuelva tumoral [38].

[6]Este valor está determinado en base a las limitaciones fundamentales de la exactitud de la replicación y reparación del ADN.

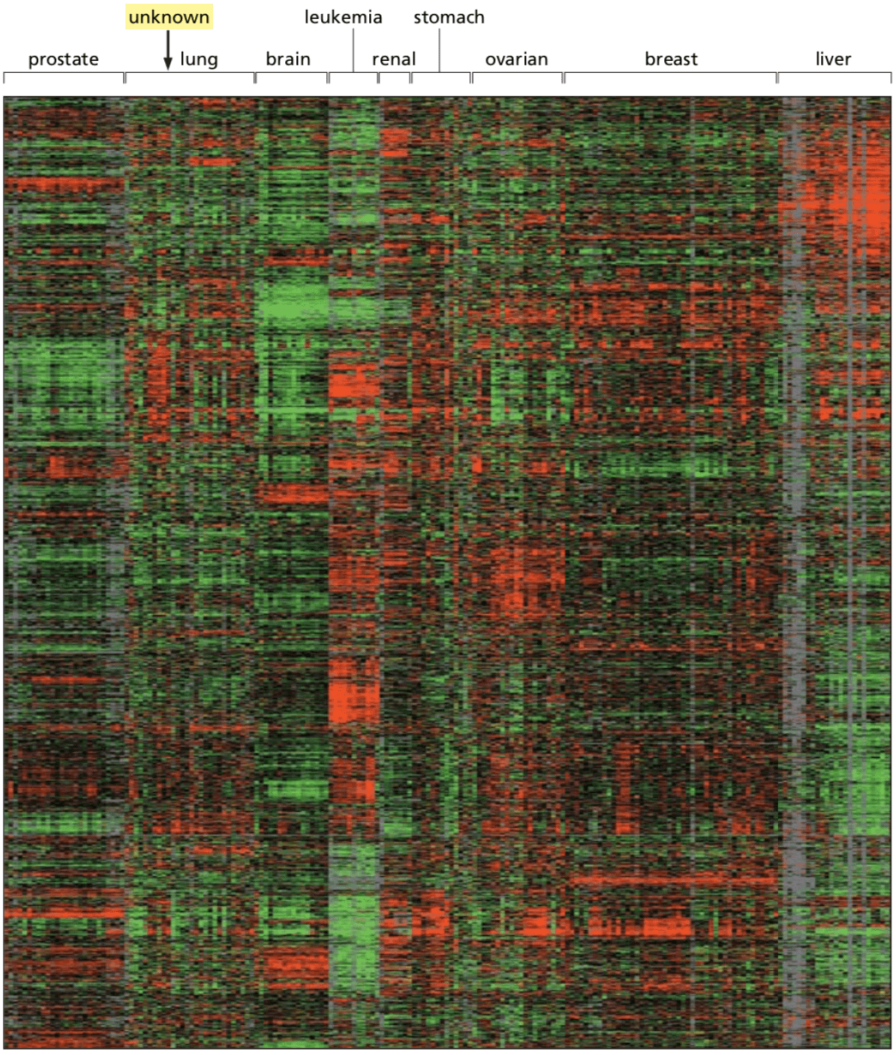


Figura 1.11. Perfiles de expresión de ARNm en distintos tipos de células cancerosas humanas. Comparación de perfiles de expresión génica de 142 líneas celulares de tumores humanos (dispuestas de derecha a izquierda), donde se determinaron los niveles de ARNm de 1 800 genes seleccionados (dispuestos de arriba hacia abajo). Las bandas rojas indican que ese gen en ese tumor se transcribe a un nivel significativamente más alto que el promedio en todas las líneas celulares, es decir que se **sobreexpresa**. Las bandas verdes indican un nivel de expresión menor que el promedio en todas las lineas celulares, es decir que el gen se **subexpresa** para ese tipo de tumor. En cuanto a las bandas negras, indican que el nivel de expresión está cerca del promedio en los diferentes tumores. Para generar estos datos se utilizan microarreglos. En cada gen analizado la expresión varía notablemente dependiendo de la línea tumoral. De esta forma, es posible definir un patrón de expresión génica característico de cada línea celular [38]. Fuente de la imagen: [38]

3.2 Técnicas de biología molecular y celular

Las **técnicas de biología molecular** continúan siendo una herramienta central en el desarrollo de trabajos de investigación en áreas afines. Estas técnicas no solo permiten manipular y analizar el ADN, sino también las estructuras y mecanismos moleculares responsables de procesos complejos como el crecimiento, la división y la diferenciación celular. De manera más general, son las que permiten manipular moléculas críticas para estos procesos y observar alteraciones en los sistemas [48].

Dentro de los diversos estudios de biología molecular, uno de los primeros pasos a abordar es la **extracción y purificación de ácidos nucleicos** de ADN y/o ARN. Estos se aíslan del resto de los componentes celulares mediante la ruptura celular mecánica, preservando la fragmentación de las cadenas. El método de extracción se elige a partir del ácido nucleico de interés (ADN, ARN, ARNm, etc.), el organismo de origen, la fuente (cultivo celular, biopsia de tejido, sangre, etc.) y la técnica en que se utilizarán los ácidos nucleicos [49]. El reactivo TRIzol (Thermo Fisher Scientific, EEUU) es uno de los principales agentes empleados en la extracción y separación de ácidos nucleicos y permite aislar fracciones de ARN, ADN y proteínas a partir de muestras de tejidos y células de diversos orígenes.

Existen múltiples técnicas de análisis de ácidos nucleicos. En primer lugar, la **digestión con enzimas específicas** o enzimas de restricción son enzimas capaces de reconocer una secuencia de nucleótidos particular en una molécula de ADN y escindirla en ese punto. Este método es de amplia aplicación para evaluar la presencia y tamaño de los fragmentos de ADN, particularmente en el abordaje de enfermedades genéticas [50]. Por su parte, la técnica de amplificación de segmentos de ADN con el procedimiento de la **Reacción en Cadena de la Polimerasa (PCR, del inglés *Polymerase Chain Reaction*)** permite amplificar, es decir copiar a gran escala, una secuencia específica de ADN. Esta técnica constituye uno de los grandes pilares de la biología molecular médica, gracias a su utilidad, y la simplicidad de su ejecución [49]. La técnica de PCR requiere, entre otros, un ADN molde, cebadores, nucleótidos y la enzima polimerasa del ADN. Los cebadores son secuencias cortas de nucleótidos complementarias al ADN molde que delimitan la región a amplificar y sirven como punto de partida para la polimerasa del ADN. Esta última sintetiza el producto del ensayo de PCR a partir de los nucleótidos individuales: A, T, C y G. El ensayo

consiste en ciclos de desnaturalización, hibridación y elongación o extensión. Los componentes se mezclan y colocan en un termociclador, que aumenta y disminuye cíclicamente la temperatura de la muestra. El aumento inicial de temperatura alcanza el umbral de desnaturalización del ADN, por lo que éste se separa en dos hebras complementarias. A continuación se disminuye la temperatura para que los cebadores se unan a los segmentos de ADN molde en la etapa de hibridación o *annealing*. Finalmente, se eleva nuevamente la temperatura y la enzima polimerasa del ADN extiende los cebadores uniendo nucleótidos complementarios a la hebra molde de ADN, en la etapa denominada elongación o extensión. De este modo, al finalizar cada ciclo se duplican las moléculas de ADN y al repetir los ciclos éstas se amplifican exponencialmente en pocas horas [51]. También es posible amplificar fragmentos de ARN, en cuyo caso es necesario primero realizar un proceso de retrotranscripción del ARN para generar una hebra de ADNc a partir de la de ARN, resultando el ADNc el molde para la amplificación.

El procedimiento de PCR permite detectar la presencia o ausencia de un producto específico de ADN [51]. El método de electroforesis es un método de separación basado en la movilidad de las biomoléculas en una fase líquida sometida a un campo eléctrico [52]. La **electroforesis en gel** es la técnica más utilizada para analizar los productos de PCR. Ésta permite analizar la presencia, el tamaño y la carga de ADN, ARN o proteínas por comparación con estándares de tamaño molecular conocido [51, 53]. Los geles utilizados en el análisis de ácidos nucleicos y proteínas son de agarosa y/o poliacrilamida, según el tamaño de las moléculas.

El método de Reacción en Cadena de la Polimerasa Cuantitativa (qPCR, del inglés *Quantitative Polymerase Chain Reaction*) es una variante de la PCR tradicional o cualitativa y sirve para amplificar y cuantificar en tiempo real el producto de PCR; es de gran utilidad para el análisis de expresión génica [54]. Al igual que en la PCR, el ADN (o ADNc) se amplifica cíclicamente en un termociclador. En este caso se adicionan a la mezcla fluoróforos para medir la tasa de generación de los productos específicos y en cada ciclo de amplificación se utilizan distintas longitudes de onda para cuantificar en el tiempo la abundancia de cada fluoróforo con sensores de fluorescencia incorporados en el termociclador [54].

La técnica de **secuenciación de Sanger** permitió conocer la primera secuencia completa de ADN de un gen codificante de proteínas [55] y su descubrimiento, junto con el de la PCR, significó un avance fundamental para la investigación genómica [56]. Esta técnica permite secuenciar

regiones de ADN, es decir determinar la secuencia de nucleótidos [53]: se amplifica una hebra de ADN con la enzima polimerasa y un cebador (similarmente a la PCR), y se realizan cuatro reacciones en paralelo. Cada reacción contiene cuatro deoxinucleótidos que componen la secuencia de ADN (dATP, dGTP, dCTP y dTTP) y uno de los dideoxinucleótidos (ddATP, ddGTP, ddCTP y ddTTP) marcados con fluoróforos que interrumpen la síntesis de la hebra en un nucleótido particular. De este modo, por cada reacción se obtienen cadenas de ADN de distintas longitudes según el dNTP final, que al ser ordenados y comparados con los resultados de las reacciones restantes permiten reconstruir la secuencia de nucleótidos de la cadena original de ADN.

3.2.1 Cultivos celulares

Las técnicas de biología molecular y celular anteriores permiten el análisis del genoma celular. A continuación se abordan técnicas de obtención de muestras células sobre las que se aplican estas técnicas.

Una **línea celular** es un conjunto de células establecidas, que proliferarán dados el medio y el ambiente adecuados [57]. Se obtiene a partir de un cultivo celular primario (conformado por células aisladas obtenidas directamente de un tejido o fluido biológico) optimizado para su subcultivo, pudiendo ser desarrollado dentro del laboratorio, adquirido de otros laboratorios o de origen comercial [58, 59]. Los **cultivos celulares** se han utilizado ampliamente como instancia preclínica para imitar las condiciones *in vivo* y desarrollar estrategias de estudio reproducibles de ciertos tejidos o células [60–64]. En el transcurso de las últimas décadas, los cultivos celulares han contribuido al estudio del cáncer mediante estudios bioquímicos, celulares, moleculares y funcionales en condiciones *in vitro* controladas [60, 65, 66].

Los requisitos técnicos para la preparación y el mantenimiento de un cultivo varían dada su necesidad de una temperatura apropiada, mezcla de gases específica y, en especial, el medio de crecimiento necesario [67]. Para cada tejido particular existen distintas líneas celulares que se comercializan, como es el caso del CE. Las líneas celulares comerciales de CE son extraídas de tumores de distinto grado o estadio [68]. Entre ellas, las más utilizadas son:

- **CE tipo I (CEE)**: AN3, ECC-1, EN, EN-1, EN-11, Hec-1a, Hec-1b, Ishikawa, KLE, MFE-280, MFE-296, MFE-319

- **CE tipo II (CENE)**: ARK1, ARK2, HEC-155/180, SPEC-2

3.2.1.1 Modelos de estudio

Como se describió anteriormente, dentro de los modelos disponibles para el estudio del CE se encuentran las líneas celulares Hec-1a e Ishikawa. Ambas provienen de adenocarcinomas de endometrio en estadio IA tipo CEE. La diferencia entre ellas es que la línea Hec-1a es derivada de un tumor G2 y la línea Ishikawa, de un tumor G1. Se ha demostrado que ambas líneas que cuentan con el factor transcripcional ETV5 presentan un aumento en la migración e invasión celular [69], lo que se asocia a una actividad tumoral mayor en el tejido endometrial [70].

ETV5 es un factor transcripcional de la familia de factores *E26 Transformation-Specific* (ETS). Este factor se encuentra **sobreexpresado en estadios tempranos del CE**, especialmente en aquellos restringidos al cuerpo uterino. Se asocia generalmente al estadio IB, donde la IM es mayor al 50% [70]. Además, se ha reportado que juega un papel fundamental durante los eventos tempranos de la tumorigénesis endometrial y podría estar asociado a una mutación inicial que desencadena la invasión miometrial [71]. Por otro lado, se han detectado mediante inmunohistoquímica mayores niveles de ETV5 en el frente invasivo de los tumores de endometrio [72].

El grupo colaborador encabezado por el Dr. J. Reventós genero transfectantes estables de ETV5 en ambas líneas celulares, caracterizándose a niveles celular, molecular y funcional. La expresión de ETV5 en ambas líneas les confirió a Hec-1a e Ishikawa un fenotipo de mayor agresividad. En este estudio se contó con las dos líneas parentales Hec-1a e Ishikawa y las transfectantes estables de ETV5 en ambas, Hec-1a-ETV5 (HGE) e Ishikawa-ETV5, las que se utilizaron en los estudios de expresión de los genes seleccionados en los estudios bioinformáticos.

4 Bioinformática

La bioinformática, surgida en la segunda mitad del siglo XX, es una disciplina relativamente joven que combina ciencias como la informática, matemática, estadística, química, biología, lingüística e ingeniería para el estudio de organismos vivos y sus redes de interacción de genes, proteínas y reacciones bioquímicas. El objetivo común de las técnicas de esta área es la solución, desde un punto de vista computacional, de **problemas biológicos** a gran escala que requieren un **uso intensivo**

de datos. En líneas generales, las estrategias bioinformáticas involucran la recolección de datos biológicos, el diseño y la construcción de un modelo computacional, y la evaluación de su desempeño [73]. Algunas aplicaciones de este tipo son el análisis de datos de secuenciación genómica, la construcción de entornos de simulación y modelado celular, la predicción de la estructura y función de proteínas, y el modelado de redes y dinámicas de regulación génica. Otras, más sencillas, consisten en el diseño de cebadores para la replicación de material genético o la predicción de la función de productos de genes [74].

En el año 2014, el Dr. Russ Altman, profesor de la Universidad de Stanford en Estados Unidos, definió a la **bioinformática traslacional** como el conjunto de métodos informáticos que relacionan entidades biológicas (genes, proteínas y moléculas) con entidades clínicas (enfermedades, síntomas y fármacos), o vice versa [75]. En este sentido, y en el contexto actual de la investigación científica en el mundo, la bioinformática cumple el rol fundamental de acortar la brecha entre la investigación básica y su aplicación en la clínica. Asimismo, propicia la transformación en la descripción y clasificación de enfermedades, incorporando los mecanismos moleculares a los métodos tradicionales, consistentes principalmente en síntomas macroscópicos [18].

Uno de los mayores hitos de la bioinformática fue el Proyecto Genoma Humano (PGH), una iniciativa internacional de financiamiento mixto para determinar la secuencia de los pares de bases del ADN humano. Más de 2 800 investigadores participaron del proyecto, compartiendo la autoría una vez publicada la versión final. El PGH representó un avance inédito para las ciencias, en particular las biológicas, y extendió notablemente las fronteras en el estudio de las enfermedades a través del genoma.

4.1 Proyecto Genoma Humano

Hasta el 2001, año de lanzamiento del primer borrador del PGH, no se conocía con exactitud la cantidad de genes humanos, y las estimaciones variaban entre 50 000 y 140 000. En el 2003, con la publicación final del proyecto, se reveló que existen cerca de 25 000 genes en el genoma humano y se dieron a conocer grandes volúmenes de información al respecto en tres grandes ejes: la secuencia de bases nucleotídicas de todo el genoma, mapas de la localización de genes en las distintas secciones cromosomales y mapas de vinculación donde se pueden rastrear los rasgos heredados de generación en generación [76, 77]. Asimismo, el informe develó que el largo promedio de un gen expresado es

de 3 000 bases y que la secuencia de dos personas es idéntica en un $99,9\%$, planteándose así nuevos interrogantes para comprender la función de la información genética [78, 79].

Por primera vez en la historia se pudo "leer" el genoma humano, dando lugar a nuevas disciplinas como la biología sintética, la medicina molecular o la farmacogenética. Esto abrió las puertas no sólo a la comprensión y el análisis de la extensa información disponible, sino también al modelado y simulación computacional de la información genética [76, 79].

Si bien se esperaba que el conocimiento del genoma humano completo resolviera muchos interrogantes de las disciplinas biológicas, estos no hicieron más que aumentar con el PGH. Si bien gracias a éste se dispone de más información y un mayor entendimiento sobre el genoma y sus mecanismos intrínsecos, también se reveló, entre otras cosas, que mas del 50% de los genes descubiertos tienen una función aún desconocida [79].

El PGH fue el primer trabajo bioinformático a gran escala e interdisciplinario en el mundo. Entre sus logros se destacan el desarrollo de nuevas tecnologías de laboratorio, la generación de mapas genéticos, transcriptómicos y físicos de los genomas de varios organismos, la incorporación de un programa de bioética, y la gratuidad y accesibilidad de los resultados para toda la comunidad científica. Sin lugar a dudas fue el puntapié inicial para el desarrollo de muchos de los proyectos y tecnologías que conviven actualmente [80–82]. Además, reforzó la idea de que pequeñas variaciones genéticas tienen diferentes incidencias que podrían generar enfermedades [79]; esto representó un cambio de paradigma para la medicina, abriendo el camino a la **medicina genómica personalizada**, donde el tratamiento, el tipo y el dosaje de medicación se podrían definir a medida para la información genética particular de cada individuo [81].

4.2 Áreas de la bioinformática

En la **era de las tecnologías "multiómicas"**, caracterizada por técnicas moleculares a gran escala que estudian sistemáticamente el genoma, transcriptoma y proteoma de los organismos, la capacidad de análisis e integración de datos de la bioinformática se convierten en herramientas centrales para explotar el potencial de esta nueva información [83]. En este contexto existen dos áreas diferenciadas dentro de la bioinformática que, en la práctica, son complementarias; éstas son la minería de texto y la minería de datos.

4.2.1 Minería de texto

La minería de texto es una disciplina dentro de la bioinformática que ha cobrado gran relevancia académica en los últimos 20 años. Consiste en la aplicación de técnicas de procesamiento del lenguaje natural para la obtención y organización de información biológica relevante y estructurada a partir de extensas colecciones de datos biológicos o biomédicos [84].

La cantidad de publicaciones de PubMed[7] del año 2018 que contienen la palabra "*cancer*" en su resumen asciende a 175 587, mientras que 1 725 publicaciones contienen "*endometrial cancer*". Es así que la minería de texto se convierte en una herramienta imprescindible para el desarrollo de trabajos de investigación exhaustivos en el área; la tarea de relevar las publicaciones científicas y bases de datos en busca de información de interés solo es posible con técnicas apropiadas.

A modo de ejemplo, una de las aplicaciones más frecuentes de la minería de datos es la creación automática o semi-automática de ontologías, es decir, vocabularios comunes para estandarizar descripciones, tipos y relaciones de entidades. Existen proyectos de referencia para genes y enfermedades, como *Gene Ontology* (GO) [85] y el árbol de Encabezados de Temas Médicos (MeSH, del inglés *Medical Subject Headings*) de la Biblioteca Nacional de Medicina (NLM, del inglés *National Library of Medicine*) [86], que sirven como base para la aplicación de minería de datos sobre conjuntos de genes, enfermedades y asociaciones de ambos.

4.2.2 Minería de datos

Un área especialmente activa y prolífera dentro de la bioinformática es la minería de datos, que se refiere a la extracción (de ahí el término "minería") de conocimiento estructurado y patrones confiables y útiles a partir de grandes volúmenes de datos [87]. En la práctica, los principales objetivos de la minería de datos aplicada a la bioinformática son la predicción de características y la descripción de entidades biológicas, mediante inteligencia artificial y tecnologías de la información. Algunos de los métodos y algoritmos de los que se vale son clasificadores, redes neuronales, regresión, reglas de asociación, *clustering* y visualizaciones [88].

La extensión de las bases de datos biológicas actuales [por ejemplo RefSeq, EMBL, Ensembl, *Protein Data Bank* (PDB), *The Human Protein Atlas* (HPA))] representan tanto un desafío como

[7] PubMed es uno de los motores de búsqueda científicos más conocidos; contiene más de 20 millones de citaciones a la base de datos MEDLINE y diversas revistas científicas.

una oportunidad para la bioinformática y, precisamente, para la minería de datos. El PGH, descripto en la Sección 3 de este capítulo, introdujo una aproximación a gran escala y colaborativa para la resolución de interrogantes biológicos. Asimismo, la invención de la técnica de RNA-Seq en 2008 (Sección 3.1.1, Introducción) optimizó la detección y cuantificación de transcriptos, potenciando así el aumento de información disponible en el área [89]. En este contexto de complejidad informática creciente, los métodos de minería de datos son uno de los recursos disponibles para abordar y explotar grandes repositorios de datos genómicos.

4.3 Biomarcadores en cáncer

Una aplicación de las técnicas bioinformáticas es la búsqueda de biomarcadores para el diagnóstico temprano, pronóstico preciso y tratamiento efectivo de enfermedades, entre ellas el cáncer. Un **biomarcador** se define como una característica objetivamente medible que describe el estado biológico normal o anormal en un organismo o tejido, ya sea mediante biomoléculas como ADN, ARN, proteínas y péptidos, o sus modificaciones químicas [90]. La OMS provee una definición más amplia del término: cualquier sustancia, estructura o proceso que puede ser medido en el cuerpo o sus productos, y que predice o influye en la incidencia de un síntoma o enfermedad [91].

Los biomarcadores pueden clasificarse de la siguiente forma de acuerdo a su uso: **biomarcadores predictivos, pronósticos, diagnósticos y blancos terapéuticos** [92]. Los biomarcadores *predictivos* caracterizan a los pacientes de acuerdo con el grado de respuesta a un tratamiento determinado y se utilizan para predecir los resultados de la intervención terapéutica. Los *pronósticos* identifican pacientes según las tasas de sobrevida o el riesgo de obtener resultados clínicos adversos, tales como la recurrencia o progresión de la enfermedad. Los biomarcadores *diagnósticos* permiten identificar una enfermedad o alguna de sus condiciones. Por último, los *blancos terapéuticos* son biomoléculas que pueden utilizarse en terapias de modulación de la expresión génica. En la práctica es posible que un biomarcador posea características predictivas, pronósticas y diagnósticas en simultáneo [93].

Este enfoque se enmarca en el concepto de **medicina de precisión**, que toma en cuenta la variabilidad interindividual para desarrollar métodos diagnósticos y tratamientos más efectivos. El estado del arte sugiere que la integración de perfiles genómicos, proteómicos y transcriptómicos con variables clínicas bien definidas conduce a un mejor desempeño diagnóstico y pronóstico [18].

Particularmente en cáncer, dada gran complejidad de los mecanismos moleculares y vías de señalización, se sugiere a la identificación de patrones de expresión característicos, conocidos como **firmas moleculares**, como una herramienta prometedora en el diagnóstico y seguimiento de la enfermedad [94]. Éstas resultan de la combinación de biomarcadores moleculares centrales en la progresión tumoral con moléculas específicas de cada tejido alterado, resultado de los cambios en la expresión génica celular que ocurren de manera dinámica durante la enfermedad. Los esfuerzos en el ámbito de la investigación están puestos en desarrollar en una primera instancia biomarcadores y luego firmas moleculares para vías aberrantes de procesos tumorigénicos y terapias inmunológicas [92, 94]. A modo de ejemplo, [95], [96] y [97] exploran biomarcadores para cáncer hepático, colorrectal y de mama, respectivamente, a partir de análisis bioinformáticos.

5 Análisis de datos

Como se planteó anteriormente, en los últimos años se ha producido un incremento en la producción y recopilación de datos a nivel mundial. Este fenómeno representó (y aún representa) tanto un desafío como una oportunidad para la sociedad, por su complejidad y utilidad inherentes. En particular la biología, medicina y sus áreas afines han experimentado importantes avances tecnológicos que resultaron en nuevas metodologías para la obtención de vastos volúmenes de información (por ejemplo, la introducción de la técnica de RNA-Seq -Sección 3 del presente capítulo-, el PGH o la implementación de historias clínicas electrónicas en los centros de salud); adicionalmente, parte de esta información es pública y accesible a cualquier investigador a través de internet. Esto implica que existe un nuevo enfoque para el estudio del cuerpo humano, su genoma y sus enfermedades, que complementa a las técnicas tradicionales y contribuye al concepto de medicina de precisión [98].

Actualmente, más allá de los costos de recolección y mantenimiento, una de las principales limitaciones en la adopción de datos a gran escala para investigación científica es su **interpretación**. Si bien la bioinformática (Sección 4, Introducción) es una herramienta clave para la extracción de datos desde servidores y repositorios y su posterior transformación y estructuración, también es necesario llevar a cabo un análisis de los datos. El objetivo es dotar de sentido a la extracción del conocimiento e **interpretar los resultados** obtenidos. Solo así es posible garantizar la **precisión, representatividad y reproducibilidad** del estudio [99].

Hadley Wickham, uno de los referentes actuales en el área de ciencia de datos, plantea **tres etapas** para el análisis de cualquier conjunto de datos: ordenarlos, comprenderlos y comunicarlos [100]. A su vez, la comprensión implica un proceso cíclico de **transformación, visualización y modelado**, como muestra la **Figura 1.12**.

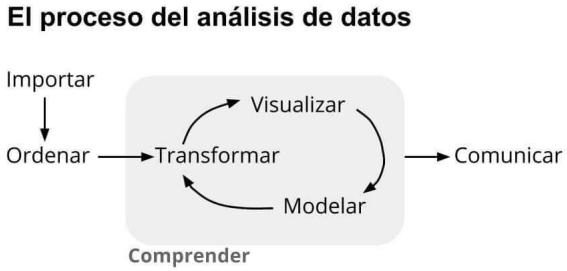


Figura 1.12. Etapas del proceso de análisis de datos de acuerdo con Hadley Wickham, referente en el área de ciencia de datos [101].

5.1 Bioestadística

La bioestadística es central en los procesos de transformación, visualización y modelado de datos. La primera parte del análisis bioestadístico consiste en el **diseño**, tanto de la base de datos como de los modelos a aplicar. En trabajos de investigación científica, esta es una actividad crucial para dar respuesta a las preguntas de investigación [102]. Diseñar una base de datos implica definir qué datos son de interés (estos reciben el nombre de "variables operativas"), el formato necesario y la relación entre ellos. La **definición de variables operativas** consiste en asignar a cada observación un tipo de dato pertinente de acuerdo con los objetivos del estudio y la pregunta de investigación; los tipos posibles son variables nominales, ordinales, categóricas, discretas y continuas. Este proceso puede requerir modificar, reducir o pre-procesar los datos genómicos crudos para ajustarlos a las definiciones anteriores.

El siguiente paso es la **estadística descriptiva** o análisis exploratorio de los datos, que tiene el propósito de explorar las variables y las relaciones entre ellas, resumir los datos y presentarlos gráficamente. Esto permite identificar tanto características sobresalientes como inesperadas en el conjunto de muestras, que podrían ser relevantes para el análisis inferencial. A su vez, este resumen brinda información sobre la calidad de los datos disponibles [103]. Algunos de los gráficos más habituales en esta instancia son los de dispersión, de barras, de cajas y bigotes e histogramas.

5.1.1 Modelos de análisis

La definición de modelos de análisis es el primer paso en el proceso de comprensión de datos y comprende la elección de una metodología apropiada para dar respuesta a la pregunta de investigación. En función del tamaño de la muestra, el tipo y la distribución de las variables y la relación entre ellas se adoptan uno o varios modelos capaces de representar los datos. También es necesario definir la cantidad de parámetros a incluir en el modelo.

La literatura ofrece distintos enfoques para estudiar la relación entre biomarcadores moleculares y la agresividad del cáncer. Una aproximación tradicional para cuantificar el efecto de distintos biomarcadores sobre los resultados clínicos es el cálculo de *Odds Ratios* (OR) e Intervalos de Confianza (IC) del 95% [104–106]. Adicionalmente, algunos autores proponen análisis de sobrevida con curvas de Kaplan-Meier y pruebas de *logRanks* para estimar la curva de sobrevida y el valor pronóstico de los biomarcadores (con OS o RFS) seguidos de un modelo de riesgos proporcionales de Cox univariado y/o multivariado [107–110]. Las secciones que siguen presentan los fundamentos de estos modelos de análisis.

5.1.1.1 *Odds Ratios* (OR)

El índice de disparidad, razón de momios o, más comúnmente, OR es una medida de la fuerza de asociación entre dos variables nominales, aleatorias e independientes. Es una métrica muy utilizada en investigación en salud [111], donde se aplica para estudiar la asociación entre una variable de exposición y una de resultado. En otras palabras, representa la posibilidad de ocurrencia de un evento de interés (variable de respuesta) dada una exposición (variable de agrupación). Esto permite no solo determinar si una exposición representa un factor de riesgo para el resultado en estudio, sino también **comparar la medida del efecto de distintos riesgos**.

El OR es una magnitud adimensional, con un rango de 0 a ∞. Una de las ventajas de esta medida de efecto es su robustez, ya que, a diferencia del Riesgo Relativo (RR), se mantiene constante al intercambiar las variables de exposición y resultado en la tabla de contingencia.

La interpretación de los resultados es la siguiente: $\mathbf{OR = 1}$ si no hay asociación entre la presencia del factor exposición y el evento; $\mathbf{OR > 1}$ si la asociación es positiva, es decir que la presencia del factor se asocia a mayor ocurrencia del evento; por último, $\mathbf{OR < 1}$ si la asociación es negativa.

A modo de ejemplo, si el resultado de una exposición A (puede ser un tratamiento o la expresión aumentada de un gen para el caso de estudio de este trabajo) frente a una exposición B (otro tratamiento, o la expresión disminuida del mismo gen que en A) produce un OR igual a 2,8 se puede concluir que la razón entre ocurrencia versus no ocurrencia del evento es 2,8 veces mayor en pacientes con A en comparación a pacientes con B. Asimismo, un OR igual a 1 implicaría que ambas exposiciones tienen la misma probabilidad de ocurrencia del evento y que, por lo tanto, no existe una asociación entre la exposición y el evento.

Adicionalmente, se evalúa el error de este estimador mediante un IC, donde el ancho del intervalo refleja la variabilidad de la medida del OR. El IC del 95% se trabaja en escala logarítmica para cumplir con la asunción de normalidad de la distribución.

Nuevamente, un IC menor a 1 indica una magnitud negativa del efecto con exposición, mientras que un IC mayor a 1, una magnitud positiva. Si el IC del OR incluye a 1, el estadístico no es significativo.

5.1.1.2 Análisis de sobrevida, método de Kaplan-Meier y prueba del intervalo logarítmico

El **análisis de sobrevida** es el conjunto de técnicas estadísticas apropiadas para el seguimiento de pacientes hasta la ocurrencia de un evento determinado, generalmente la recurrencia de una enfermedad o muerte del sujeto. Lo que se pretende estimar, la **función de sobrevida** $S(t)$, es la probabilidad de sobrevida de un individuo más allá del tiempo t sin la ocurrencia del evento. La gráfica de *S(t)* en función del tiempo recibe el nombre de **curva de sobrevida**. Otra función relevante en el análisis de sobrevida es la de **riesgo** $h(t)$; esta mide el riesgo de ocurrencia del evento en un período de tiempo concreto.

Uno de los análisis llevados a cabo en este estudio es el **método de Kaplan-Meier**, inicialmente descripto en 1958 [112]. Éste es un método no paramétrico de estimación de la función de sobrevida que se construye a partir de los tiempos de sobrevida exactos de los individuos de una población. *S(t)* se estima como el producto entre la proporción de pacientes con y sin ocurrencia del evento inmediatamente después del instante *t*. De este modo, la curva de sobrevida (la **Figura 1.13** ejemplifica dos curvas de sobrevida: una para la expresión aumentada y otra para la expresión disminuida del gen ACSL5) se mantiene constante durante los períodos entre eventos y cambia

cuando ocurre uno. Por su parte, la función de riesgo es una medida de la probabilidad de ocurrencia del evento estudiado entre los pacientes "sanos" después del instante t. Concretamente, representa la probabilidad de que un individuo de los que todavía no han sufrido el evento lo sufra en el período de tiempo en cuestión.

A partir de dos curvas de sobrevida de Kaplan-Meier, la **prueba del intervalo logarítmico** (conocida como ***logRanks***) permite comparar las distribuciones. La hipótesis nula del análisis es que estas curvas (para el caso en estudio, de expresión aumentada y disminuida) son idénticas y se busca probar si la diferencia entre ambas es mayor a la esperable por azar.

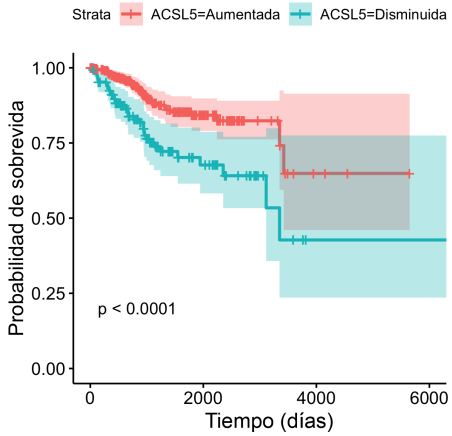


Figura 1.13. Ejemplo de curvas de sobrevida construidas mediante el método de Kaplan-Meier. En este caso se comparan dos grupos definidos por un umbral en los niveles de expresión del gen ASCL5. El eje horizontal representa el tiempo y el vertical, la probabilidad de sobrevida. Se puede notar que el grupo con expresión disminuida presenta un peor pronóstico.

5.1.1.3 Análisis de sobrevida y modelo de riesgos proporcionales de Cox

Normalmente, la herramienta más utilizada para modelar la relación entre una variable de respuesta y múltiples variables predictoras es la regresión múltiple. Sin embargo, los datos de sobrevida tienen una distribución sesgada y censuras, por lo que la regresión tradicional (ya sea linear o logística) no es una opción válida. El modelo de riesgos proporcionales o modelo de regresión de Cox, inicialmente descripto en el año 1972 [113] es un método popular para trabajar con datos de este tipo.

El modelo de Cox estudia la relación entre el tiempo de sobrevida y una o más variables independientes y permite modelar el tiempo a un evento determinado. En otras palabras, el análisis de

datos clínicos con el modelo de Cox estima la probabilidad de ocurrencia de un evento (como ser la recurrencia o muerte de pacientes) a partir de los valores de las variables independientes (llamadas **covariables**) en un tiempo determinado. El principal resultado de este modelo son los ***Hazard Ratios*** **(HR)**, que miden cuánto afecta cada covariable a la función de riesgo para el evento de interés ($H(t)$). Esta función representa la probabilidad del evento en un tiempo determinado (el riesgo o *hazard*) y se modela según la **Ecuación 1.1**. Con covariables dicotómicas como los niveles de expresión de genes, el HR expresa el aumento en la probabilidad de ocurrencia del evento cuando la variable predictora está presente respecto de cuando está ausente.

$$H(t) = H_0(t) \times \exp\left(b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_kX_k\right), \tag{1.1}$$

donde $X_1 \ldots X_k$ son las covariables, $b_1 \ldots b_k$, los parámetros o coeficientes de estas covariables, y $H_0(t)$, el **riesgo basal** en el tiempo t, es decir, el riesgo para una persona con todas las variables predictoras iguales a 0.

La estimación de los parámetros b en la **Ecuación 1.1** tiene una distribución aproximadamente normal, lo que admite distintas pruebas de contraste de hipótesis para evaluar la coherencia de afirmar cada valor b como un parámetro del modelo. Una de las más utilizadas es la **prueba de Wald**, que arroja un IC y un valor de significancia estadístico para cada coeficiente [114].

El modelo de regresión de Cox busca la relación entre los riesgos de ocurrencia de un evento entre dos individuos expuestos a factores de riesgo diferentes y parte de dos hipótesis fundamentales: la primera, que **las observaciones son independientes**; y la segunda, que **los riesgos son proporcionales**. Esta última hipótesis es central para el modelo y significa que el HR debe ser constante en función del tiempo. En otras palabras, se debe cumplir que el efecto de cada variable en el riesgo del evento entre dos individuos (para covariables dicotómicas, esto es un individuo con el factor de riesgo y otro sin él) sea constante [115]. La evaluación de la hipótesis de riesgos proporcionales puede hacerse por métodos gráficos (por ejemplo, a partir de las curvas de Kaplan-Meier) o por un estudio de la bondad del ajuste [115, 116].

Los OR, el método de Kaplan-Meier y la prueba del intervalo logarítmico (*logRanks*) de las Secciones 5.1.1.1 y 5.1.1.2 del presente capítulo son ejemplos de **análisis univariado**, ya que estudian la relación entre un resultado o evento y una variable independiente sin tener en cuenta

el impacto de otros factores. El modelo de riesgos proporcionales de Cox puede ser univariado cuando se lleva a cabo con una sola covariable o **multivariado** cuando hay dos o más covariables, ya que modela en simultáneo la influencia de cada variable independiente como factor de riesgo en el evento de interés. Este modelo es comúnmente utilizado en trabajos de investigación médica, donde existen muchos componentes clínicos con potenciales efectos sobre el pronóstico de los pacientes, para identificar y caracterizar tanto a los factores como a sus efectos.

En síntesis, el modelo de riesgos proporcionales de Cox multivariado:

- Requiere una **variable de estado** binaria (por ejemplo, recurrencia o muerte), un **tiempo asociado a la variable de estado** (tiempo hasta la recurrencia o muerte) y **covariables** continuas o dicotómicas.
- Estima la función de riesgo $H(t)$ (**Ecuación 1.1**).
- Estima un parámetro b_i para cada covariable i.
- Permite conocer el HR de cada covariable como $exp(b_i)$.
  - Si b_i es mayor a 0 (o lo que es equivalente, HR es mayor a 1), la covariable i es un **factor pronóstico desfavorable** y se asocia positivamente con la probabilidad de ocurrencia del evento y negativamente con el tiempo de sobrevida.
  - Si b_i es menor a 0 (y HR menor a 1), la covariable es un **factor pronóstico favorable**.
  - Si HR es igual a uno, no hay efecto de la covariable en el riesgo del evento.
- Mediante el estadístico de Wald, evalúa la significancia de cada HR y devuelve un IC y un valor p.
  - Como sucede en el análisis de OR, si el IC incluye al 1, el efecto no es significativo; el valor p es coherente con estos resultados.
- Asume que las observaciones son independientes y que el HR de cada variables es constante en el tiempo.
  - Esto puede evaluarse mediante las curvas de sobrevida de Kaplan-Meier de cada covariable.

5.1.1.4 Manejo de errores

En modelos de estudio multivariados cobra relevancia el **error de tipo I o alfa**. Éste es un error cometido al realizar múltiples pruebas estadísticas en simultáneo y representa el riesgo de rechazar la hipótesis nula siendo esta verdadera. En otras palabras, el error alfa refleja la tasa de falsos positivos del análisis y se relaciona con la significancia estadística (valor p) de una prueba. Tradicionalmente, el valor p utilizado en análisis univariados es de 0,05. Sin embargo, al agregar variables independientes al análisis de una misma variable dependiente, es conveniente reducir este valor para prevenir el error alfa. Una aproximación simple al problema es la corrección de Bonferroni [117, 118], que ajusta el valor de significancia de cada una según $\alpha_{Bonferroni} = \frac{\alpha}{n}$, donde n es la cantidad de variables independientes del análisis. Este enfoque es conservador ya que, cuando se tratan muchas variables independientes en simultáneo, el valor p que arroja puede ser demasiado pequeño.

La contrapartida de controlar el error alfa es el **error beta**, que afecta el **poder** de una prueba estadística. El poder es la habilidad de la prueba de rechazar la hipótesis nula cuando esta es falsa (y, por lo tanto, debe ser rechazada). En otras palabras, es la probabilidad de que la prueba detecte un efecto o asociación presente entre las variables y depende del valor p y el tamaño de la muestra, entre otros factores. Ajustar el error alfa con un valor p demasiado estricto disminuye la probabilidad de rechazar la hipótesis nula si es verdadera pero también aumenta la probabilidad de no rechazarla cuando es falsa (esto es, aumenta el error beta).

En una prueba estadística ideal ambos errores, alfa y beta, son pequeños. Sin embargo, en la práctica suele ser necesario encontrar un valor p que equilibre ambos. Algunas publicaciones científicas utilizan la corrección de Bonferroni y profundizan en el poder de la prueba estadística, otras recurren a un valor p de 0,01 buscando un equilibrio entre los errores, y la mayor parte utiliza un valor p de 0,05 por convención.

6 Hipótesis

Como ilustran las secciones anteriores, el CE es una patología compleja cuyo abordaje clínico presenta desafíos aún por resolver para aportar al diagnóstico y pronóstico de la enfermedad, así como a las alternativas terapéuticas. El marcado aumento en la incidencia y mortalidad de este tipo de cáncer en los últimos 20 años y, más aún, los pronósticos para los próximos 20 lo convierten en un tema relevante en investigaciones académicas. A esto se suman las inconsistencias y discrepancias en la categorización y estadificación de tumores, que pueden conllevar imprecisiones en el tratamiento, ya sea por deficiencia o exceso. Estas situaciones impactan en la calidad de vida de las pacientes.

Por otra parte, la relevancia de los biomarcadores en el ámbito de la investigación y la clínica ha aumentado notablemente en los últimos años habiéndose demostrado que aportan sustancialmente en el abordaje de estas enfermedades. La incorporación de marcadores moleculares en la clasificación de tumores y sus respectivos riesgos ha sido aceptada para complementar el proceso de toma de decisiones médicas y mejorar así el pronóstico y sobrevida de las pacientes. Con respecto a las dificultades inherentes a la identificación de biomarcadores de este tipo, el crecimiento exponencial de datos multiómicos y el constante desarrollo de técnicas bioinformáticas de alto rendimiento facilitan la adaptación e implementación de métodos diagnósticos y tratamientos con biomarcadores al ámbito clínico, aportando además a estudios que permiten, a partir de los datos obtenidos, ofrecer soluciones simples para el diagnostico y manejo de la enfermedad.

Sobre la base de los antecedentes presentados se propone la siguiente **hipótesis de investigación**:

Existen productos biológicos de la expresión génica que constituyen biomarcadores para la detección y clasificación temprana de tumores en pacientes con CE. Estos pueden ser identificados mediante el uso de herramientas bioinformáticas de minería de texto y datos, evaluaciones bioestadísticas y experimentales.

7 Objetivo general

Este trabajo tiene por objetivo identificar potenciales biomarcadores de la progresión tumoral en CE para contribuir a las herramientas actuales y aportar a un manejo personalizado de la enfermedad.

8 Objetivos específicos

1. Realizar un análisis bioinformático de minería de datos y de texto para identificar genes con un potencial rol de biomarcador del CE.
2. Llevar a cabo un análisis de datos sobre muestreos de pacientes para evaluar la expresión de los genes identificados en 1) y su relación con los parámetros clínico-patológicos.
3. Realizar un conjunto de verificaciones sobre estos genes candidatos en modelos de estudio de cultivos de líneas celulares de CE.

Para cumplir con los objetivos, se desarrollarán las siguientes estrategias:

8.1 Análisis bioinformático

- Desarrollar un flujo de algoritmos de minería de texto y datos que integre información de repositorios públicos para identificar moléculas con potencial como biomarcadores de CE.
- Recopilar características clínico-patológicas vinculadas con la agresividad del CE y seleccionar genes con expresión diferencial a nivel transcripcional frente a estas características.
- Integrar la información de genes con asociación reportada a CE y la selección de genes con expresión diferencial mediante algoritmos de priorización génica.

8.2 Análisis de datos

- A partir de un estudio de transcriptómica global, utilizar modelos bioestadísticos con los siguientes objetivos:
  - Describir las características de la muestra de tejidos tumorales

  - Evaluar la asociación entre la expresión de los genes previamente seleccionados y los siguientes parámetros clínico-patológicos del CE:
    * Grado del tumor (I, II, III o IV)
    * Subtipo histológico (seroso o endometrioide)
    * Estadio del tumor (I, II o III)
    * Invasión miometrial ($<$ o $>$ 50%)
    * Mortalidad
    * Recurrencia
  - Evaluar la asociación entre la expresión de los mismos genes y el tiempo de sobrevida y sobrevida libre de recurrencia de los pacientes.

- Sacar conclusiones respecto de la asociación entre la expresión de los genes candidatos y los parámetros clínico-patológicos de agresividad en CE definidos en la Sección 8.1 de este Capítulo.

- Reducir la lista de genes seleccionados sobre la base de los resultados del análisis bioestadístico.

8.3 Estudios experimentales

- Aplicar técnicas de biología molecular y celular sobre los potenciales biomarcadores identificados y analizar su expresión en modelos experimentales de cultivos celulares de CE con fenotipo molecular y funcional previamente caracterizado. En particular se emplearan las lineas celulares de CE Hec-1a e Ishikawa y las transfectantes estables del factor de transcripción ETV5 en ambas lineas celulares (Hec-1a-ETV5; e Ishikawa-ETV5, respectivamente).

Materiales y Métodos

MATERIALES

1 Estudios *in silico*: herramientas bioinformáticas

1.1 DisGeNET

La plataforma *online* DisGeNET (`www.disgenet.org`)[1] es un repositorio de Asociaciones Gen-Enfermedad (GDA, del inglés *Gene-Disease Associations*) que integra información sobre genes y variantes asociadas a enfermedades humanas mediante técnicas de minería de texto, mapeos de vocabulario de genes y enfermedades, y asociaciones ontológicas propias de la plataforma [119, 120].

La información a partir de la cual se construyen las GDA proviene de tres fuentes: **bases de datos curadas manualmente** como UniProt, *The Human Phenotype Ontology* (HPO), *The Comparative Toxicogenomics Database* (CTD) o el Catálogo *Genome-Wide Association Studies* (GWAS), **modelos animales** de ratas y ratones, y diversas **publicaciones científicas** disponibles en bases de datos mundiales, por ejemplo MEDLINE. Cada una de estas fuentes proporciona un grado de evidencia distinto para respaldar las GDA y DisGeNET las jerarquiza mediante un índice entre 0 y 1, según la cantidad y calidad de sus fuentes. DisGeNET posee muchas **funcionalidades**; una de ellas permite al usuario buscar enfermedades y, como resultado, le proporciona una lista de genes que presentan GDA con esos términos. Este listado puede descargarse en formato csv[2].

[1] Todas las páginas web utilizadas en este capítulo se encuentran detalladas en **Anexo A**.

[2] El formato "Valores separados por comas (csv, del inglés *Comma Separated Values*)" es un tipo de archivo de almacenamiento de datos en forma de columnas separadas por coma y filas definidas por saltos de línea. Es una forma habitual de importar y exportar bases de datos de baja complejidad.

La información reportada por DisGeNET sobre cada gen asociado a alguno de los términos de enfermedad (términos de búsqueda) incluye el nombre y el símbolo *HUGO Gene Nomenclature Committee* (HGNC)[3] del gen. Adicionalmente, para cada uno proporciona dos índices de interés en los análisis de este trabajo: el Índice de Especificidad de la Enfermedad (DSI, del inglés *Disease Specificity Index*) y el Índice de pleiotropía de la enfermedad (DPI, del inglés *Disease Pleitropy Index*). Ambos índices adoptan valores entre 0 y 1. El DSI es un valor inversamente proporcional a la cantidad de enfermedades - representadas por términos UMSL CUI[4] - asociadas a cada gen en particular (por ejemplo, si un gen está asociado a gran cantidad de enfermedades en DisGeNET, el DSI es bajo y si se asocia solamente a una, es alto). Por su parte, el DPI indica la cantidad de términos MeSH de enfermedades a las que está asociado un gen (por ejemplo, si un gen está asociado a múltiples clases de enfermedades, el DPI es alto pero si se asocia a pocos términos MeSH, el DPI es bajo). La versión de DisGeNET utilizada en el presente trabajo (v6.0, enero de 2019) recopila información sobre 17 549 genes, 24 166 enfermedades y 628 685 GDA resultantes, de las que 165 354 provienen de fuentes curadas.

1.2 *Gene Expression Omnibus* (GEO)

El repositorio público GEO (`www.ncbi.nlm.nih.gov/geo`) del *National Center for Biotechnology Information* (NCBI) reúne datos genómicos de estudios globales de secuenciación de *high-throughput*[5], como microarreglos de ADN y Secuenciación de próxima generación (NGS, del inglés *Next-Generation Sequencing*), remitidos por la comunidad científica. GEO proporciona herramientas para analizar y descargar experimentos y perfiles de expresión génica [121, 122], entre ellas **GEO2R**. GEO2R permite al usuario seleccionar uno de los estudios disponibles en GEO y dividir las muestras en dos grupos de interés, según sus características clínico-patológicas. A partir de esta clasificación, la herramienta analiza qué genes se expresan diferencialmente entre ambos grupos; para ello realiza un análisis univariado (prueba t de Student) con un valor p asociado para cada gen.

[3] El HGNC es un comité de la Organización del Genoma Humano que define un nombre único y con sentido para cada uno de los genes humanos. Además, asigna una abreviatura denominada símbolo. Un ejemplo es el gen *BRCA1 DNA repair associated*, cuyo símbolo es BRCA1.

[4] Los códigos UMLS CUI son identificadores únicos de términos biomédicos en el Sistema Unificado de Lenguaje Médico®(UMLS). Cada entrada de DisGeNET está vinculada a uno de estos identificadores.

[5] *High throughput*: término utilizado para describir la automatización de experimentos con el fin de alcanzar resultados a gran escala en tiempos acotados y conservando la calidad del proceso original. Se aplica a diversas disciplinas tales como la biología, la computación y el diseño de fármacos.

Una vez definido el estudio a analizar y realizada la separación en dos o más grupos, GEO2R devuelve una tabla en formato csv donde se listan los valores de expresión diferencial de los distintos genes entre los grupos (logFC), así como medidas estadísticas de interés (valor p, valor p ajustado, t y B). También provee un *script* de R para descargar el conjunto de datos del estudio.

1.3 ToppGene

Los estudios globales, también conocidos como del tipo *high throughput*, para el análisis de expresión génica son útiles para la clasificación y caracterización de genes y enfermedades, pero proveen información limitada para establecer asociaciones causales entre ambos. Más aún, resultan en listas de cientos o miles de genes, en los que resulta difícil establecer su relevancia [123]. Una de las metodologías disponibles para superar estas limitaciones es la **priorización de los genes candidatos** mediante diferentes algoritmos, entre los que se destaca la plataforma *online* ToppGene (`toppgene.cchmc.org`), que utiliza anotaciones funcionales del genoma humano para establecer criterios de valoración sobre los genes, a partir de una lista de genes de referencia.

Técnicamente, el funcionamiento de ToppGene se basa en la asunción de que características fenotípicas similares se asocian con genes con funciones similares. El programa usa anotaciones funcionales de todo el genoma para comparar una lista de **genes "de entrenamiento"** con otra de **genes "de prueba"**, ambas provistas por el usuario. En este esquema, la lista de entrenamiento debe contener genes con GDAs reportadas y, la de prueba, los genes que se desean priorizar. ToppGene compara las anotaciones semánticas de los genes de entrenamiento contra las de los genes de prueba y asigna una mejor posición en el *ranking* a aquellos genes de prueba que presentan similitudes funcionales con los de entrenamiento. El algoritmo evalúa parámetros que se consideran biológicamente relevantes en la lista de entrenamiento y se utilizan para calcular una medida de similitud para cada elemento de la lista de genes de prueba [123]. Ésta se construye utilizando medidas de correlación, fundamentos de inteligencia artificial y técnicas de meta-análisis estadístico. Asimismo, la lista resultante incluye un valor p para la medida de similitud de cada gen de prueba.

Finalmente, luego de realizado el análisis, se obtiene una tabla en formato csv con los genes de la lista de prueba ordenados según su similitud con la lista de entrenamiento. Todos los genes están acompañados por un puntaje o ***score*** que representa la medida de similitud, y una significancia estadística de ese *score* (valor p), según la relación entre los genes de ambas listas. También

proporciona información sobre los distintos parámetros biológicos representativos de los genes de la lista de prueba.

1.4 *The Cancer Genome Atlas* (TCGA)

Para obtener datos de expresión génica de muestras de pacientes con diagnóstico de CE se recurrió a la información disponible en el **repositorio de datos clínicos *The Cancer Genome Atlas* (TCGA)**, un programa conjunto del NCI y el *National Human Genome Research Institute* (NHGRI) de Estados Unidos, que recopila mapas multidimensionales de los cambios genómicos y moleculares de 11 000 pacientes con 33 tipos tumorales. TCGA es actualmente la colección de datos genómicos tumorales más importante a nivel mundial [124]. Los datos de expresión génica de los estudios corresponden a los resultados de la técnica de secuenciación por RNA-Seq aplicada a muestras de tejido tumoral y control con los secuenciadores *Illumina HiSeq* e *Illumina GA*. Los conjuntos de datos incluyen, además, información clínico-patológica de cada sujeto y pueden obtenerse a través de dos plataformas de descarga y visualización de datos: **cBioPortal** y **UCSC Xena** en formato de tablas csv.

1.5 *The Human Protein Atlas* (HPA)

El Atlas de Proteínas Humanas, en inglés *The Human Protein Atlas* (HPA) (`www.proteinatlas.org`), nació como un proyecto de **integración de tecnologías genómicas, proteómicas y metabolómicas** para mapear todas las proteínas presentes en células, tejidos y órganos humanos. La versión actual de la plataforma (v18.1, noviembre de 2018) presenta una estructura que relaciona la expresión de 17 000 genes con sus proteínas. Asimismo, describe la expresión y localización de estas proteínas y transcriptos en todos los órganos del cuerpo, en líneas celulares de origen humano y en tejidos normales y tumorales [125].

La totalidad de la información contenida en HPA proviene de un estudio de 144 sujetos sanos y 261 con patología tumoral, de quienes se obtuvieron muestras de 44 tipos de tejidos distintos normales[6] y de 20 tipos de cáncer. La información sobre expresión de transcriptos es analizada mediante la técnica de secuenciación global del ARN (RNA-Seq) así como por análisis de expresión

[6]Se entiende tejido *normal* como no neoplásico y morfológicamente normal. Dado que las muestras provienen de piezas quirúrgicas, muchos de los tejidos considerados normales presentan alteraciones debidas a la inflamación y degeneración.

de la proteína en secciones de tejido por técnicas de inmunohistoquímica con anticuerpos específicos y cuantificación de la señal.

HPA clasifica a las proteínas en 10 categorías de acuerdo con sus características: **localización** (de membrana, secretadas, intracelulares) y **función** (enzimas, transportadoras, receptores acoplados a proteína G, canales iónicos voltaje-dependientes, y factores de transcripción), **asociación a enfermedades** (relacionadas con cáncer y blancos terapéuticos aprobados por la *Food and Drug Administration* (FDA)) y la **disponibilidad de evidencia en UniProt**. A su vez, de acuerdo al análisis de los niveles de transcriptos, los genes que codifican para las proteínas listadas en el HPA se clasifican de acuerdo a su abundancia en distintos tejidos. En este caso, se describen las siguientes categorías: enriquecido en un tejido, enriquecido en un grupo, presente en todos los tejidos, mixto y no detectado. Estas clasificaciones aportan a la caracterización de los genes y sus funciones.

2 Estudios experimentales

2.1 Reactivos generales de laboratorio

Durante el desarrollo de este trabajo se utilizaron **reactivos de laboratorio** de grado analítico, calidad cultivo celular o calidad biología molecular según su uso específico. Las sales y otros reactivos generales fueron adquiridos de la firma Sigma-Aldrich Corporation (EEUU) o, de lo contrario, a firmas que se mencionan a lo largo del texto. Los **reactivos de electroforesis** fueron productos de BioRad Laboratories (EEUU) y de Amersham Biosciences Corporation (EEUU), y los de **biología molecular** fueron adquiridos de las firmas Qiagen (Alemania), Promega Corporation (EEUU) y Life Techonolgies-Thermo Fisher Scientific (EEUU), a menos que se especifique lo contrario.

2.2 Anticuerpos

Para los estudios de inmunodetección de TPX2 se emplearon los siguientes anticuerpos:

Anticuerpos primarios:

- **anti-TPX2**: el anticuerpo anti-TPX2 (B-5, sc-376812, Santa Cruz Biotechnology, EEUU)

es un anticuerpo monoclonal desarrollado en ratón y dirigido contra los residuos 635-675 cercanos al extremo carboxilo de TPX2 de origen humano. Se utilizó en ensayos de inmunocitoquímica de fluorescencia y de *Western immunoblotting*.

- **anti-β-tubulina**: el anti-β-tubulina (D66, Sigma-Aldrich) es un anticuerpo monoclonal desarrollado en ratón, que reconoce un epítope localizado en la región del carboxilo terminal la proteína. Se utilizó como control de carga en el ensayo de *Western immunoblotting*.

Anticuerpos secundarios:

- **Para los ensayos de inmunocitoquímica**: anti Inmunoglobulina G (IgG) de ratón acoplada a Cy3 (Sigma-Aldrich).
- **Para los ensayos de *Western immunoblotting***: anti-IgG de ratón acoplada a la enzima peroxidasa de rábano picante (Vector Laboratories Inc.).

Las concentraciones de trabajo empleadas para los anticuerpos primarios y secundarios se especifican según corresponde a lo largo del trabajo. En todos los ensayos se incluyó IgG de ratón purificada del suero normal de animales no inmunizados (Sigma-Aldrich, EEUU) como controles negativo, que fue agregada en reemplazo del primer anticuerpo y a la misma concentración que éste.

2.3 Líneas celulares

Se emplearon las líneas celulares de CE humano Hec-1a, Ishikawa y sus transfectantes estables de ETV5. Las primeras fueron transfectadas de forma estable con la secuencia del factor de transcripción ETV5 humano fusionado a la secuencia codificante de la Proteína fluorescente verde (GFP, del inglés *Green Fluorescent Protein*) (Hec-1a-GFP-ETV5 [HGE], e Ishikawa-ETV5). Las líneas transfectantes fueron generadas por el grupo del Dr. J. Reventós (Hospital Vall d'Hebron de Barcelona, España), quien gentilmente cedió tanto las líneas parentales como las transfectantes al grupo de la Dra. Vazquez-Levin, como parte de un proyecto colaborativo entre ambos grupos de investigación (Proyecto Prot-BioFluids; Programa 7mo Marco, Marie Curie Actions, Unión Europea). Las características generales de las líneas celulares empleadas y sus condiciones de cultivo se encuentran listadas en el **Anexo B**.

2.4 Cebadores

En el diseño de cebadores específicos para los protocolos de PCR se utilizó la herramienta Primer-Blast, disponible en la Plataforma del NCBI (`www.ncbi.nlm.nih.gov/tools/primer-blast`). A partir de los códigos ***refseq***[7] del ARN del gen de interés (o alguna de sus variantes transcripcionales), Primer-Blast genera una lista con todas las opciones de cebadores. El programa Oligo Analyzer 3.1 (`www.idtdna.com/pages/tools/oligoanalyzer`) permite analizar las características de cada cebador diseñado por Primer-Blast. A partir de éstas se pueden definir criterios para la selección de los cebadores más apropiados.

Para el diseño de los cebadores, se consideraron las siguientes especificaciones:

- Los cebadores deben amplificar un fragmento del ADN molde que comprenda secuencias de al menos, y preferentemente, dos exones; se optó por elegir cebadores que intercepten la unión entre dos exones del ARNm. Así, se puede distinguir si el fragmento amplificado o el amplicón provienen de un molde de ARN o ADN genómico.
- Los cebadores no deben formar homo o hétero-dimeros ni estructuras secundarias de alta estabilidad ($\Delta G > -10$).
- Se debe evitar la presencia de dos o más bases Guaninas o Citosinas en la región 3' de la secuencia del cebador diseñado.
- Se busca tener una temperatura de hibridización o anillado (*annealing*) del cebador al ADN molde entre 58 y 60^{o}C.
- Los cebadores diseñados deben tener preferentemente un tamaño mínimo de 80 y máximo de 180 pb.
- El cebador debe ser lo más específico posible: solamente debería unirse a la secuencia del ADN molde y a sus variantes, y no tener ninguna interacción con fragmentos de otros genes.

Las secuencias de cebadores utilizados y el tamaño esperado para el amplicón correspondiente a cada uno de los genes en los que se evaluó los niveles de expresión del ARN se listan en el **Anexo C**.

[7] Los códigos *refseq* son registros de secuencias únicos del genoma, transcriptoma y proteoma de la base de datos del NCBI.

MÉTODOS

3 Estudios *in silico*

3.1 Relevamiento de genes asociados a CE

La herramienta DisGeNET fue empleada con el objetivo de relevar genes asociados a CE. Para ello, se consultaron e identificaron todos aquellos términos de enfermedades neoplásicas femeninas reportados en la plataforma relacionadas al CE, de modo de identificar la mayor cantidad posible de genes asociados a la enfermedad.

3.2 Análisis de expresión diferencial

Se utilizó la librería *limma* [126] del paquete *Bioconductor*[8] (v 3.8) en R para analizar el estudio de expresión génica del CE **GSE17025** denominado **Gene Expression Analysis of Stage I Endometrial Cancers**, descargado desde la plataforma GEO [127]. Este estudio reúne los resultados del análisis de expresión génica global empleando tecnología de microarreglos de ADN de muestras de CE en estadios tempranos, de distintos grados y subtipos histológicos [128]. El estudio consta de 92 muestras de CE, de las cuales 80 corresponden a tumores endometrioides y 12 a serosos o mixtos. Asimismo, incluye 12 muestras de endometrio atrófico de mujeres postmenopáusicas (muestras control).

Sobre este muestreo de datos, se llevaron a cabo tres análisis de expresión diferencial consecutivos. En cada caso, se escogió un parámetro clínico-patológico a comparar y se obtuvo una tabla de genes diferencialmente expresados, ordenados según su significancia estadística. Los grupos comparados fueron:

1. Muestras tumorales versus no tumorales
2. Tumores de tipo CEE versus CENE
3. Tumores con histología G1-2 versus G3

[8] *Bioconductor* es un conjunto de herramientas informáticas desarrolladas en el lenguaje R para aplicaciones de análisis y comprensión de datos genómicos de tipo *high-throughput*.

A partir de estas comparaciones se obtuvieron tres tablas de genes diferencialmente expresados, de los que se seleccionó aquellos en los que las diferencias fueron significativas (valor p $\leq 0,05$) y se encontraron en los tres grupos (**Figura 2.1**).

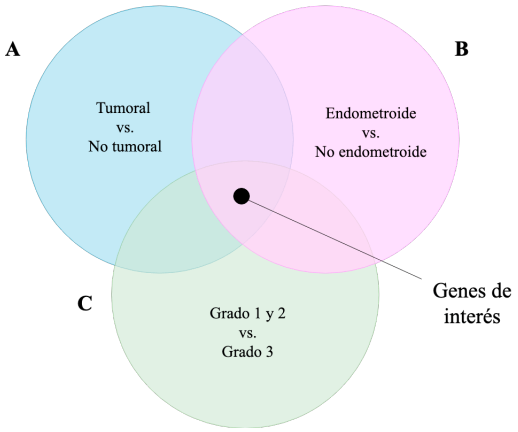


Figura 2.1. Metodología adoptada para la selección de genes candidatos a partir del análisis de expresión diferencial. Diagrama de Venn representando la metodología adoptada para la selección de genes candidatos a partir de tres listas de genes diferencialmente expresados en las muestras del estudio GSE17025. El conjunto **A** contiene genes diferencialmente expresados en tejido tumoral versus normal; **B**, tumores de tipo CEE versus CENE y **C**, tumores de G1-2 versus G3.

3.3 Priorización génica

Para el análisis y priorización de genes según asociaciones reportadas en bases de datos se empleó la herramienta ToppGene. La **lista de entrenamiento** incluyó a los genes surgidos a partir de la búsqueda en la base de datos DisGeNET (Sección 1.1, Materiales y Métodos) y la **lista de prueba** fue la de los genes comunes a los tres grupos de expresión génica global diferencial del estudio GSE17025 (Sección 3.2, Materiales y Métodos).

4 Análisis de datos

Seguidamente, se utilizó un estudio de transcriptómica global de TCGA para evaluar la expresión de los genes seleccionados a partir de GEO y el análisis de priorización con ToppGene. El estudio utilizado fue TCGA-*Uterine Corpus Endometrial Cancer* (TCGA-UCEC), que reúne datos clínicos y genómicos de un conjunto independiente de pacientes con CE.

Se importó, ordenó y visualizó la información de los pacientes para posteriormente modelar,

mediante técnicas estadísticas, la información genómica y evidencia clínica disponibles, siempre con el objetivo de propiciar el análisis de la asociación entre ambas. Las secciones que siguen describen los pasos involucrados en este proceso.

4.1 Diseño y preparación de la base de datos

El diseño de la base de datos consistió en la selección y acondicionamiento de los datos disponibles en TCGA-UCEC. Este proceso, al igual que los subsiguientes, fue llevado a cabo con el *software* estadístico **R (versión 3.5.3)**, a través del entorno de desarrollo integrado **RStudio**.

El tamaño del estudio descargado para el análisis estadístico fue de **580 registros independientes**. Los pacientes del estudio fueron enrolados bajo los siguientes **criterios de inclusión**, reportados previamente [12]:

- Pacientes con adenocarcinomas endometrioides o carcinomas serosos según el consentimiento de la institución que los reporta.
- Pacientes con indicación médica de resección quirúrgica y sin tratamiento previo (quimioterapia, radioterapia, etc) para CE.
- Pacientes con estadificación certera según la FIGO.
- Muestras con suficiente material genómico de alta calidad.
- Muestras con un mínimo del 80% correspondiente al núcleo del tumor (y menos del 20% al tejido necrótico).
- Muestras con tejido normal acompañante proveniente de a) sangre o componentes sanguíneos, b) tejido normal adyacente al tumor (con una separación mayor a 2 cm), o c) ambos.

En el análisis estadístico se desestimaron aquellas muestras sin información sobre expresión de los genes seleccionados en la Sección 3.2 del presente capítulo, como así también aquellas sin información acerca de los parámetros clínicos RFS y OS del sujeto.

En primer lugar se procedió a construir una **base de datos** propia, a partir de la información descargada del estudio TCGA-UCEC. Este proceso fue fundamental dado que el estudio incluye información no relevante para el trabajo de investigación (como, por ejemplo, datos de expresión

de microarreglos, metilación, *copy number*, expresión *Reverse Phase Protein Array* e información clínica que no es de interés para el presente estudio), y también porque la información útil no está estructurada adecuadamente para los modelos a utilizar. A continuación se describe la metodología empleada para la preparación de la base de datos.

4.1.1 Definición de variables

4.1.1.1 Variables de respuesta

En el diseño de este trabajo, las variables de respuesta elegidas fueron los parámetros clínico-patológicos que se utilizan de rutina en la clasificación del CE: estadio, grado, subtipo histológico, e IM. Estos permiten determinar la agresividad de la patología, según lo expuesto en la Sección 2.4 de la Introducción.

La **Tabla 5** representa el diseño teórico de las variables de respuesta del modelo de análisis estadístico, que son los parámetros clínico-patológicos de la enfermedad. Sin embargo, como se mencionó anteriormente, no todas las observaciones del estudio TCGA-UCEC siguen esta estructura. Por lo tanto, la primera parte del trabajo estadístico consistió en adaptar la información disponible al formato requerido para los modelos.

Diseño de variables de respuesta

Variable	Tipo	Valores
Edad del paciente (al momento del diagnóstico)	Discreta	Años de edad del paciente
Sobrevida libre de recurrencia (RFS) (estado)	Categórica dicotómica	0: paciente sin evento de recurrencia 1: paciente con evento de recurrencia
Sobrevida libre de recurrencia (RFS)(tiempo)	Discreta	Días hasta el evento de recurrencia
Sobrevida total (OS)(estado)	Categórica dicotómica	0: paciente vivo 1: paciente fallecido
Sobrevida total (OS) (tiempo)	Discreta	Días hasta el deceso del paciente
Grado histológico del tumor	Categórica dicotómica	0: G1 o 2 1: G3
Subtipo histológico del tumor	Categórica dicotómica	0: endometrioide 1: no endometrioide
Estadio del tumor	Categórica dicotómica	0: estadio I o II 1: estadio III o IV

Invasión miometrial	Categórica dicotómica	0: <50% 1: ≥50%

Tabla 5: Diseño de variables de respuesta, al que debe adaptarse la información de la base de datos del estudio UCEC disponible en TCGA (TCGA-UCEC). Cada parámetro clínico-patológico corresponde a una variable, cuyo tipo y valores fueron asignados en función del objetivo del análisis estadístico.

Los parámetros de estado del paciente (vivo/muerto) y tiempo de OS y RFS no fueron modificados, ya que su formato original cumple con la estructura deseable para la base de datos de este trabajo. Las variables grado histológico, subtipo histológico y estadio del tumor son originalmente categóricas no dicotómicas y fueron convertidas siguiendo los criterios de agrupación de la Sección 2.4 de la Introducción. Se eligieron los números 0 y 1 arbitrariamente para facilitar la posterior manipulación de la base de datos; y en todos los casos el 1 fue asignado a las características más agresivas del parámetro clínico, según lo descripto en la Sección 2.4 de la Introducción.

- **Grado histológico del tumor**: las entradas originales fueron reemplazadas por "0" (G1-2) y "1" (G3).
- **Subtipo histológico del tumor**: las entradas de la variable original fueron reducidas a "0" (subtipos endometrioides) y "1" (subtipos no endometrioides: serosos y mixtos).
- **Estadio del tumor**: toda la estadificación FIGO (Sección 2.4.3, Introducción) se redujo a "0" (estadios I, IA, IB, II) y "1" (estadios III, IIIA, IIIB, IIIC, IV, IVA y IVB), según la agresividad del tumor.
- **Invasión miometrial**: esta información no está disponible en TCGA, por lo que se determinó a partir del estadio. Los tumores con estadio FIGO IA y IB tienen invasión miometrial "$< 50\%$" y los demás, "$\geq 50\%$". A estos valores corresponden "0" y "1", respectivamente.

4.1.1.2 Variables de agrupación

De acuerdo con el diseño del trabajo, las variables de agrupación de los modelos estadísticos son las mediciones de los **niveles de expresión de los transcriptos** de los genes seleccionados, donde cada gen corresponde a una variable.

En función de los modelos de análisis disponibles, se optó por simplificar esta información y **categorizar la expresión génica en dos niveles: aumentada (1) y disminuida (0)**. Para

determinar el valor de corte en cada caso se empleó el paquete de R *maxstat* (*Maximally Selected Rank Statistics*), que evalúa todos los puntos de corte posibles de la variable continua a categorizar y utiliza la **prueba de *logRanks*** para comparar las dos distribuciones definidas por cada punto de corte y hallar el que es estadísticamente más significativo. De este modo, se obtiene un valor de significancia estadística (valor p) para cada punto de corte tentativo y se define como **punto de corte óptimo** de la variable a aquel que ha arrojado un menor valor p en la prueba. A partir del punto de corte se define en qué muestras la expresión de los transcriptos está aumentada (es decir, es mayor al valor hallado con *logRanks*) o disminuida (es menor). Este enfoque fue presentado por la herramienta *online* Cutoff Finder (`molpath.charite.de/cutoff`) [129] y adoptado por grupos de investigación en diversas aplicaciones biológicas relacionadas con cáncer [130–132] y especialmente en estudios de expresión génica tumoral [133].

En síntesis, con la información de expresión en TCGA-UCEC de todos los genes identificados mediante los estudios bioinformáticos anteriores se evaluaron todos los puntos de corte con la prueba de *logRanks* y se conservó aquel con mejor valor p respecto de la variable **estado de RFS**. De este modo se obtuvo una lista con los genes seleccionados y un valor p para cada uno. A partir de lo anterior se descartaron aquellos genes con valor $p > 0{,}05$. El **método de Kaplan-Meier** fue utilizado para estimar y comparar las curvas de sobrevida asociadas a la expresión aumentada y disminuida de cada gen con las variables de estado RFS y OS. Por cada gen se obtuvieron dos curvas de sobrevida con el evento recurrencia y dos con el evento muerte; cada par de curvas fue graficado en una figura indicando el valor p del punto de corte.

4.2 Estadística descriptiva

Se llevó a cabo un análisis exploratorio de la base de datos diseñada en la Sección 4.1, Materiales y Métodos.

Se realizó un histograma con la edad de diagnóstico para comprender la distribución de la variable. Esto provee información sobre la composición y estructura de la población del estudio. Por otra parte, para las variables categóricas se realizaron gráficos de barras y circulares, que permiten evaluar si las muestras están uniformemente distribuidas; lo contrario (es decir, que exista alguna categoría con muy poca incidencia o un desequilibrio muy grande entre categorías) podría afectar los modelos y sus resultados. En particular se graficaron los valores de RFS y OS para visualizar

la tendencia global de mortalidad y recurrencia en la población en estudio.

4.3 Modelos de análisis

En las secciones que siguen los términos *evento*, *punto final*, *resultado* y *variable de respuesta* se utilizan indistintamente para designar a la información contenida en las características clínico-patológicas del estudio TCGA. Por otra parte, *factores de riesgo*, *variables predictoras* o *de agrupación* y *exposición* hacen referencia a los niveles de expresión de los transcriptos en las muestras de TCGA. Los modelos de análisis seleccionados son OR y el modelo de riesgos proporcionales de Cox para la estimación de curvas de sobrevida.

4.3.1 *Odds Ratios* (OR)

Se efectuaron análisis de OR e IC con la expresión de cada gen candidato y todas las variables de respuesta clínicas categóricas dicotómicas. En todos los casos, se consideró "exposición" a la expresión aumentada del gen en estudio y "no exposición" a la expresión disminuida del mismo gen. Se midió el efecto de la exposición sobre el "resultado" de un parámetro clínico-patológico, es decir, la ocurrencia o no del evento más agresivo de la variable.

Siguiendo un orden asignado a las variables clínico-patológicas, se calculó para cada una los OR e IC correspondientes a todas las variables de agrupación. Los resultados de la primera ronda de análisis, es decir todas las exposiciones comparadas con la ocurrencia del evento de la primera variable de respuesta, sirvieron para filtrar las variables de agrupación antes de la segunda ronda. Así, sucesivamente, se consiguió un esquema de "cascada" que permitió seleccionar solo aquellos genes con magnitudes de efecto significativas para todas las características clínico-patológicas. El orden de las variables de respuesta en la cascada fue: RFS, OS, subtipo histológico, grado, estadio e invasión miometrial.

A lo largo del análisis se consideraron como magnitudes de efecto significativas a aquellos casos con OR perteneciente al intervalo $(0,1) \cup (1,\infty)$[9], y con IC dentro del mismo intervalo; esto es equivalente a aquellos casos con valor p $\leq$ 0,05. Aquellos genes cuyos niveles de expresión de transcriptos arrojaron magnitudes de efecto no significativas con algún parámetro clínico-patológico fueron descartados del análisis porque, con un 95% de probabilidad, el OR puede adoptar valores

[9] En el conjunto de simbología matemática, $\cup$ representa la unión de dos intervalos

iguales a 1 y, por lo tanto, no existir asociación (o efecto) entre la variable predictora y la de respuesta. La **Tabla 6** resume lo anterior.

Interpretación de *Odds Ratios*

Resultados	**Interpretación**
$OR \in (1, \infty)$ $IC \subset (1, \infty)$ valor $p \leq 0{,}05$	Asociación estadísticamente significativa entre el nivel de expresión de los transcriptos del gen y la característica clínico-patológica. La ocurrencia de la categoría más grave de la variable clínico-patológica es OR (entre IC^- e IC^+) veces mayor en pacientes con expresión alta en comparación con pacientes con expresión baja. La expresión elevada del gen se asocia a peor pronóstico.
$OR \in (0, 1)$ $IC \subset (0, 1)$ valor $p \leq 0{,}05$	Asociación estadísticamente significativa entre el nivel de expresión de los transcriptos del gen y la característica clínico-patológica. La ocurrencia de la categoría más grave de la variable clínico-patológica es 1/OR (entre IC^- e IC^+) veces mayor en pacientes con expresión baja en comparación con pacientes con expresión alta. La expresión elevada del gen se asocia a mejor pronóstico.
$1 \in IC$ valor $p > 0{,}05$	Asociación estadísticamente no significativa entre el nivel de expresión de los transcriptos del gen. En el 95% de los casos en un estudio, el OR podría ser igual a 1 y no existir efecto o asociación entre la categoría de la variable clínico-patológica y el nivel de expresión del gen. No se puede sacar conclusiones sobre la asociación entre el gen y la característica clínico-patológica.

Tabla 6: Interpretación de resultados de *Odds Ratio*, donde $\in$ significa que un valor pertenece a un conjunto y $\subset$ indica que se trata de un subconjunto.

4.3.2 Modelo de riesgos proporcionales de Cox

Se llevaron a cabo dos modelos de riesgos proporcionales de Cox:

1. El primero, con el evento de **recurrencia** y las siguientes especificaciones:
    - Variable de estado: RFS (estado)
    - Tiempo al evento: RFS (tiempo)
    - Covariables: expresión categorizada de los genes resultantes del análisis de OR (Sección 4.3.1 del presente capítulo) que cumplen con la hipótesis de riesgos proporcionales
    - Valor de significancia alfa: 0,05
2. El segundo, con el evento de **muerte** y las siguientes especificaciones:
    - Variable de estado: OS (estado)
    - Tiempo al evento: OS (tiempo)

- Covariables: expresión categorizada de los genes resultantes del análisis de OR (Sección 4.3.1 del presente capítulo) que cumplen con la hipótesis de riesgos proporcionales
- Valor de significancia alfa: 0,05

Se utilizó paquete de R *survival* para desarrollar ambos modelos.

Antes de construir los modelos se analizaron las curvas de sobrevida de Kaplan-Meier obtenidas en la Sección 4.1.1.2 de Materiales y Métodos para probar el cumplimiento de la **hipótesis de riesgos proporcionales**. Fueron descartados aquellos genes cuyas curvas de sobrevida de individuos con presencia y ausencia del factor de riesgo, es decir con expresión aumentada y disminuida del gen, se cruzan o no se comportan de manera similar en el tiempo. Dado que las curvas de Kaplan-Meier para RFS y OS son distintas, el descarte de genes por no cumplir la hipótesis se realizó independientemente para ambos modelos. Por lo tanto, la cantidad de covariables de los modelos no fue necesariamente la misma.

Se trabajaron modelos con **eliminación hacia atrás**. Esta metodología comienza con un modelo de todas las covariables y consiste en realizar iteraciones eliminando automáticamente una covariable en cada iteración de acuerdo con algún criterio. En este trabajo se adoptó el criterio del efecto menos significativo; esto es, eliminar del modelo aquella covariable con mayor valor p. El análisis termina cuando todas las covariables alcanzan el nivel de significancia alfa definido previamente.

4.3.3 Manejo de errores

Tanto en los análisis univariados (GEO, *logRanks* y OR) como en el modelo de riesgos proporcionales de Cox se utilizó un **valor de significancia estadística de 0,05**. Por lo tanto, aquellos resultados con valor p menor a 0,05 fueron considerados estadísticamente significativos.

4.4 Rastreo de genes candidatos

En el marco de la investigación, se consideró que la información de HPA podría aportar al análisis de los genes candidatos. La selección de genes obtenidos del análisis bioestadístico (Sección 4, Materiales y Métodos) fue analizada con esta herramienta (para comprender el proceso de selección de los genes candidatos, remitirse a la Sección 4.3, Materiales y Métodos) para describir los patrones

de expresión de los genes seleccionados, identificar aquellos relacionados con cáncer o el sistema reproductor y excluir genes sin suficiente información reportada.

A continuación se describen los principales aspectos tenidos en cuenta sobre el estado del arte de los genes seleccionados.

- **Registros sobre el gen, el ARNm y la proteína, y herramientas para el estudio de la proteína**: fueron de especial interés aquellos genes en los que se encontrara información sobre su localización cromosomal, presencia de variantes de *splicing* del transcripto, niveles de expresión del transcripto en diversos tejidos del cuerpo humano, datos sobre la expresión y funciones de la proteína, y alteraciones tanto en el transcripto como la proteína en diversas enfermedades.
- **Asociación del gen con enfermedades**: fueron de interés aquellos genes que se relacionaran con cáncer, especialmente con neoplasias en órganos del aparato reproductor.
- **Expresión del gen en órganos reproductores**: se buscaron aquellos genes de los que hubiera información sobre su expresión en tejidos de órganos reproductores, específicamente en cuerpo uterino.
- **Estudios de expresión de ARN y proteínas**: fueron de especial interés los genes con datos de expresión de ARN y proteínas, tanto en tejidos normales como tumorales.

5 Estudios experimentales

5.1 Cultivo celular

Se realizaron cultivos en monocapa de células epiteliales tumorales de endometrio humano:

- **Ishikawa**: células de adenocarcinoma de endometrio estadio IA, G1.
- **Ishikawa-ETV5**: células Ishikawa, transfectadas de forma estable con el factor de transcripción ETV5.
- **Hec-1a**: células de adenocarcinoma de endometrio estadio IA, G2.

- **Hec1a-GFP-ETV5 (HGE)**: células Hec-1a transfectadas de forma estable con el factor de transcripción factor ETV5.

Las líneas celulares fueron criopreservadas con Dimetil Sulfóxido (Sigma) como agente criopreservante en medio de cultivo y almacenadas en tanques de nitrógeno líquido. Una vez descongeladas siguiendo un protocolo estándar de incubación en baño de agua a 37°C por 1 minuto, se colocaron en un tubo estéril de 15 ml conteniendo medio de cultivo D-MEM (Nunc-Thermo Scientific, EEUU) suplementado con 10% Suero Fetal Bovino (SFB) y 1% de penicilina/estreptomicina; a continuación, las células se centrifugaron durante 5 minutos a 1200 rpm, descartando el sobrenadante con el agente criopreservante. Las células se resuspendieron en el mismo medio de cultivo, y se colocaron en una botella de cultivo estéril y luego en una estufa de cultivo (Forma Scientific, EEUU) a 37°C en una atmósfera húmeda con 5% de CO_2 en aire. Al alcanzar una confluencia del 80%, las células fueron sub-cultivadas utilizando una solución estéril de Tripsina 0,25% y Ácido etilendiaminotetraacético (EDTA) 0,025% en *Buffer* Fosfato Salino (PBS) (Tripsina-EDTA). Al finalizar este proceso, los cultivos fueron procesados para la preparación de extractos totales de proteínas y ácidos nucleicos.

5.2 Protocolo de extracción y análisis de ARN

5.2.1 Extracción y cuantificación del ARN total

Se realizó la extracción del ARN total para su posterior retrotranscripción y ensayo de PCR. El ARN fue extraído a partir de las líneas celulares Ishikawa, Ishikawa-ETV5, Hec-1a y HGE, utilizando el reactivo TRIzol® (Life Technologies-Thermo Fisher Scientific, EEUU) y siguiendo las instrucciones del fabricante.

Los pellets de ARN total obtenidos fueron resuspendidos en agua libre de ARNasas y cuantificados con un espectrofotómetro "NanoDrop® ND1000" (Life Technologies-Thermo Fisher Scientific, EEUU). Este dispositivo mide la absorbancia de la muestra a una longitud de onda de 260 nm. A partir del valor de absorbancia calculado por el espectrofotómetro, se determinó la concentración de ARN total de cada una de sus muestras, verificando también la relación de absorbancias 260:280 con el objetivo de determinar la pureza de las preparaciones. Aquellas con relación de absorbancia menor a $1,6$ se descartan por estar contaminadas de ADN genómico. Las muestras no contaminadas

fueron almacenadas a -70°C hasta su uso.

5.2.2 Ensayo de retrotranscripción del ARN

Para obtener el ADNc del ARN, se realizó un ensayo de retrotranscripción del ARN. Para su síntesis se utilizaron 2 μg de ARN total de cada una de las líneas celulares Ishikawa, Ishikawa-ETV5, Hec-1a, HGE con la enzima transcriptasa reversa "SuperScript™ III" (Life Technologies-Thermo Fisher Scientific, EEUU). En la mezcla de la reacción se incluyó: ARN molde, oligodT15, dNTPs [mezcla de Desoxiadenosina trifosfato (dATP), Desoxitimidina trifosfato (dTTP), Desoxicitosina trifosfato (dCTP) y Desoxiguanosina trifosfato (dGTP); Life Technologies-Thermo Fisher Scientific, EEUU] y agua destilada estéril, cuyo volumen fue ajustado según el resto de los componentes para completar el volumen final de 20 μl. Luego de preparada la mezcla, se calentó 5 minutos a 65°C en un termociclador Thermo (Life Technologies-Thermo Fisher Scientific), y se incubó al menos un minuto en hielo. Se agregó el *buffer* "First Strand" (pH= 8,3), ditiotreitol (DTT) e inhibidor de ARNasas (RNAseOUT) (Life Technologies-Thermo Fisher Scientific, EEUU). Por último, se agregó la enzima retrotranscriptasa "SuperScript™ III" y se incubó durante 50 minutos a 50°C, seguido de 15 minutos a 70°C para inactivar la reacción enzimática. En todos los ensayos se incluyeron dos controles negativos por reacción: omisión del ARN y omisión de la enzima retrotranscriptasa.

Para comprobar el correcto desarrollo del protocolo de retrotranscripción, se amplificó un fragmento correspondiente al ARNm de la enzima GAPDH (gen endógeno), empleando un protocolo de amplificación en cadena de la polimerasa PCR a punto final.

5.2.3 Ensayo de PCR a punto final

Para el ensayo de PCR a punto final se utilizó la enzima ADN polimerasa *TaqI* (Life Technologies-Thermo Fisher Scientific) en un termociclador Thermo (Life Technologies-Thermo Fisher Scientific, EEUU). Para la reacción se utilizaron 20 μl de volumen final, 300 nM de cada uno de los nucleótidos libres (dATP, dTTP, dCTP y dGTP; Life Technologies-Thermo Fisher Scientific), una concentración de cada cebador de 1 μM y $0,75$U[10] de la enzima ADN polimerasa *TaqI*. El protocolo de amplificación utilizado se describe a continuación (**Tabla 7**).

[10]Unidad de actividad enzimática. Se define como la actividad catalítica responsable de la transformación de un μmol de sustrato por minuto en condiciones óptimas de la enzima.

Protocolo de PCR a punto final	
Estado	**T°x tiempo**
1 ciclo de:	
Desnaturalización inicial	94°C x 5 minutos
40 ciclos de tres pasos de:	
1. Desnaturalización	94°C x 30 segundos
2. Hibridación o *annealing*	60°C x 1 minuto
3. Extensión / elongación	72°C x 1 minuto
1 ciclo de:	
Extensión final	72°C x 5 minutos

Tabla 7: Protocolo general de amplificación de fragmentos específicos de los diferentes ARNm evaluados en este trabajo mediante la técnica de PCR a punto final.

En todos los ensayos realizados se incluyeron dos controles: un control negativo (omisión de ADNc) y un control positivo de PCR (fuente de ADNc en la que se comprobó previamente la presencia del fragmento de amplificación de interés). Una vez finalizado el procedimiento las muestras fueron almacenadas a -20°C hasta su análisis.

5.2.4 Electroforesis en geles de agarosa

Luego de completado el ensayo de amplificación por PCR a punto final, se sometieron los productos obtenidos a electroforesis en geles de agarosa al 2-2,5%, de acuerdo al tamaño de los productos a visualizar. Para lograr la concentración deseada, se disolvió la agarosa en *buffer* de corrida TBE [Tris 0,09 M, pH= 8,0; ácido bórico 0,09 M; EDTA 0,002 M] y se suplementó con Bromuro de Etidio a una concentración final de 0,5 μg/ml. Se prepararon las muestras con *buffer* TBE conteniendo el colorante xileno-cianol y un agente densificador, como glicerol. En todas las corridas se incluyeron mezclas estándares de peso molecular (PB-L Productos Bio-Lógicos, Argentina) para una correcta estimación del tamaño molecular aparente de los productos amplificados. Las corridas electroforéticas se realizaron a un voltaje constante de 75V en una cuba electroforética con *buffer* TBE. Al finalizar la corrida, se visualizaron las señales correspondientes a los amplicones, en un sistema de transiluminación con detección en el rango UV gracias a la fluorescencia del Bromuro de Etidio, agente intercalante del ADN. Se realizó el registro fotográfico con una cámara digital.

5.2.5 Ensayo de PCR cuantitativa

Se realizó la evaluación cuantitativa de los niveles de ARNm empleando un protocolo de PCR en tiempo real con la unidad CFX96 Touch™ (BioRad). Para las reacciones, se utilizó la mezcla preformada "SYBR Green® PCR Master Mix" (Life Technologies-Thermo Fisher Scientific) conformada por: stock 2X de una mezcla optimizada del colorante fluorescente "SYBR Green I", ADN polimerasa "AmpliTaqGold®", dNTPs con desoxiuridina trifosfato (dUTP), colorante de referencia pasiva ROX y un *buffer* optimizado para la actividad de la ADN polimerasa. Se agregaron 300 nM de los cebadores de interés (listados en el **Anexo C**), 50 ng de ADNc del molde y agua para completar un volumen final de 25 μl. En todos los casos se utilizó un control negativo de PCR por reacción (omisión del ADN molde). El protocolo de amplificación utilizado se describe en la **Tabla 8**.

Protocolo de PCR en tiempo real

Estado	**T°x tiempo**
1 ciclo de:	
Activación	95°C x 10 minutos
40 ciclos de dos pasos de:	
1. Desnaturalización	94°C x 15 segundos
2. Hibridación + Extensión/elongación	72°C x 1 minuto

Tabla 8: Protocolo general de amplificación de fragmentos específicos de los diferentes genes evaluados en este trabajo mediante la técnica de PCR cuantitativa en tiempo real.

Al final de cada paso de hibridación/elongación de cada ciclo de amplificación se realizó la correspondiente medición de fluorescencia. La fluorescencia es causada por la unión del colorante "SYBR Green I" al ADN doble cadena.

Se monitoreó en el equipo de PCR el avance en la detección del producto amplificado, midiendo la variación de la fluorescencia en cada ciclo, determinando el nivel de fluorescencia umbral dentro de la fase exponencial de la reacción o Ct (del inglés *Cycle Threshold*) correspondiente al número de ciclos que requiere cada muestra para alcanzar dicho umbral. El valor Ct está directamente relacionado con la cantidad inicial de molde presente en la reacción de PCR, permitiendo estimar cuantitativamente su cantidad relativa a una muestra control.

Una vez finalizado el ensayo se realizaron curvas de disociación (curvas de *melting*), dado que

siempre que haya moléculas de ADN doble cadena se obtiene una señal, para determinar cuan específica es la señal e identificar la presencia de señales provenientes de falsos positivos debidos a productos de amplificación de ADN espúreos provenientes de amplicones inespecíficos y/o dímeros de cebadores, entre otros.

Para calcular los **niveles de expresión** de los transcriptos de todos los genes evaluados se utilizó la **ecuación 5.2.5**. Se utilizó GAPDH como gen endógeno.

$$\text{Exp} = 2^{-\Delta C_t} \text{ donde } \Delta C_t = C_{t \text{ gen estudio}} - C_{t \text{ gen endógeno}} \tag{2.1}$$

Para determinar la **expresión relativa** de ciertos transcriptos respecto de una muestra de referencia se utilizó la **ecuación 5.2.5**.

$$\text{Exp relativa} = 2^{-\Delta\Delta C_t} \text{ donde } \Delta\Delta C_t = \Delta C_{t \text{ muestra}} - \Delta C_{t \text{ referencia}} \tag{2.2}$$

5.3 Protocolo de extracción y análisis de proteínas celulares

5.3.1 Preparación de extractos proteicos totales

Se lavaron las monocapas celulares dos veces con PBS a 4°C para remover el medio de cultivo y los suplementos, y las células se recogieron de las placas de cultivo con un *cell scrapper* y se transfirieron a un tubo de ensayo, a los que se agregó una solución tamponada para la lisar las células y extraer las proteínas. Todos los lisados celulares se prepararon con *buffer* de extracción *Radio Immunoprecipitation Assay* (RIPA) (de composición Tris-HCL 20 mM pH = 7,7, NaCl 150 mM, NP-40 1%, Deoxicolato de sodio 1%), suplementado con un cóctel de inhibidores de proteasas (p-aminobenzamidina 2 mM, fenil metil sulfonil fluoruro PMFS 1 mM, aprotinina 10 μg/ml). La lisis celular se completó mediante ruptura física con jeringa con aguja 22G en hielo. Los homogenatos proteicos obtenidos se centrifugaron a 13 000 xg durante 30 minutos a 4°C, luego de lo cual se recuperó la fracción de sobrenadante; se separo una alícuota para determinar el contenido de proteínas celulares totales y el resto de la preparación se conservó a -70°C hasta su posterior

análisis.

5.3.1.1 Determinación de la concentración de proteínas totales

Para determinar el contenido proteico de los extractos celulares, se llevó a cabo un **ensayo de cuantificación de proteínas con el método de Bradford** con un reactivo comercial (BioRad), siguiendo las indicaciones brindadas por el fabricante.

El ensayo de Bradford es un método espectrofotométrico que permite cuantificar la concentración de proteínas en el volumen de una muestra. Según este método, la muestra con proteínas es tratada con el **colorante Comassie Blue G-250**, que existe en tres formas: aniónica (coloración azul), neutra (verde) y catiónica (roja). En condiciones ácidas, la forma catiónica se convierte en aniónica que se une a las proteínas presentes en la muestra en estudio. En caso de que no haya proteínas presentes, la solución conserva un color amarronado. De este modo, la intensidad de luz con longitud de onda de 595 nm (azul) absorbida por la muestra es proporcional a la concentración del complejo proteína-colorante. Las principales ventajas de este ensayo son su sencillez, rapidez y baja sensibilidad a interferencias de otros compuestos químicos.

El primer paso del procedimiento consistió en construir una curva de calibración a partir de una solución de Albúmina Sérica Bovina (BSA, del inglés *Bovine Serum Albumin*) 0,1 mg/ml (0,1 μg/μl) con concentraciones conocidas y reactivo de Bradford. Se analizó la absorbancia a 595 nm de seis diluciones de BSA dentro del rango de determinación de proteína propio del método (entre 1 y 10 μg/μl) con el lector de placas Thermo Scientific™ Multiskan™ FC Microplate. La **Tabla 9** muestra la composición de las muestras de la curva de calibración. Cada muestra se sembró en tres pocillos de la placa de microtitulación para promediar los resultados de las mediciones.

Tabla de diluciones para la curva de calibración

BSA (μg)	**BSA** (μl)	**H_2O** (μl)	**Total** (μl)
0	0	100	100
2	20	80	100
4	40	60	100
6	60	40	100
8	80	20	100
10	100	0	100

Tabla 9: Tabla de diluciones de BSA 0,1 μg/μl para la construcción de la curva de calibración del ensayo de Bradford, según las instrucciones provistas por el fabricante.

Se aplicó un modelo de regresión lineal sobre la absorbancia media medida de cada muestra incógnita junto con la concentración conocida de BSA para obtener la ecuación de la **curva de calibración**.

Se prepararon 9 muestras incógnitas: dos por cada línea celular (Ishikawa, Hec-1a, y ambas transfectadas con el plásmido ETV5) y un control positivo (extracto proteico de células de cáncer de ovario humano SKOV3). Para cada una se creó una solución con reactivo de Bradford en una placa de microtitulación de 96 pocillos. A partir de la ecuación de la curva de calibración, se calculó la concentración de proteínas en cada muestra de interés.

5.3.2 Obtención de perfiles proteicos e identificación de TPX2

Luego de la cuantificación de extractos proteicos se realizó la técnica de electroforesis en geles de poliacrilamida seguido de electrotransferencia a membranas de nitrocelulosa, seguido de incubación con anticuerpos específicos para detectar la presencia de las(s) forma(s) proteica(s) de TPX2 en las muestras de extractos proteicos de células de endometrio humano. A continuación se describen los pasos realizados para completar este proceso.

5.3.2.1 Preparación de muestras

A partir de los extractos proteicos de las 4 líneas celulares de CE de interés obtenidos según se indica en la Sección 5.1 del presente Capítulo, el primer paso consistió en la **cuantificación de la concentración proteica** de cada extracto empleando el ensayo de cuantificación de proteínas con la técnica de Bradford (Sección 5.3.1.1 de los Materiales y Métodos). De esta manera se puede determinar la cantidad exacta en volumen del extracto de cada línea celular a colocar en cada carril del gel.

Se utilizaron volúmenes de muestras correspondientes a 30 μg de proteína total del extracto celular, según lo calculado a partir del ensayo de Bradford (Capítulo 2, Sección 5.3.1.1).

La muestra a sembrar fue suplementada con *buffer* muestra Laemmli [*buffer* Tris-HCL 60 mM (pH = 6,8), con Dodecilsulfato sódico (SDS, del inglés *Sodium Dodecyl Sulfate*) 2%, glicerol 10%, β-mercaptoetanol 5% y azul de bromofenol 0,01%] hasta llegar al volumen final de 25 μl. Todas las mezclas proteicas de siembra se incubaron durante 5 minutos a 100 °C para su desnaturalización.

Todas las corridas electroforéticas incluyeron mezclas proteicas de estándares de peso molecular

de origen comercial, de los que se utilizaron 10 μl por carril.

5.3.2.2 Ensayo de electroforesis en geles de poliacrilamida

La electroforesis en gel es una técnica utilizada para **separar distintas proteínas** en una muestra según su peso molecular, dado que las proteínas se mezclan con un detergente que al unirse les otorga una carga promedio negativa. Se utilizan geles porosos de poliacrilamida (el tamaño del poro determina el nivel de separación final de las proteínas; a mayor poro mayor resolución de proteína de alto peso molecular) y se definen "carriles" para las distintas mezclas de proteínas a resolver; se aplican corrientes eléctricas constantes para que las proteínas migren al electrodo positivo del dispositivo. Como resultado, luego de las corridas electroforéticas, las proteínas se distribuyen en la matriz de poliacrilamida, siendo las de mayor movilidad (más cercanas al frente de corrida) las de menor tamaño molecular aparente.

A partir de las muestras de proteínas preparadas anteriormente, se llevó a cabo en el laboratorio la electroforesis en geles de poliacrilamida en condiciones desnaturalizantes. Se utilizó un gel al 10% ya que es el porcentaje recomendado para detectar el peso molecular de la proteína de interés TPX2 (86 kDA). La electroforesis se llevó a cabo en el sistema de minigeles Mighty Small SE 250 (Hoefer Inc., EEUU) a 25 mA por gel utilizando *buffer* de electroforesis [*buffer* Tris-HCl 25 mM (pH= 8,3), conteniendo glicina 192 mM y SDS 0,1%].

En la corrida se incluyó como estándar de peso molecular la mezcla de proteínas: miosina (200 kDa), β-galactosidasa (116 kDa), fosforilasa b (97 kDa), BSA (66 kDa), ovoalbúmina (45 kDa), anhidrasa carbónica (31 kDa), inhibidor de tripsina de soja (21,5 kDa), lisozima (14,4 kDa), aprotinina (6,5 kDa) (BioRad Protein Broad Range, BioRad, EEUU). Los marcadores de peso molecular son una mezcla purificada de proteínas cuyo peso molecular es conocido y se utilizaron para verificar si la proteína buscada se encontraba dentro del rango de tamaño adecuado.

En todas las corridas se incluyó el análisis de un extracto de células de cáncer de ovario humano SKOV3 (datos no mostrados). La misma se empleó como control positivo de detección de TPX2, dado que en la literatura se ha descripto la expresión de dicha proteína en esa línea celular. La presencia de TPX2 fue evaluada en extractos proteicos de las líneas Ishikawa, Ishikawa-ETV5, Hec-1a, HGE. Sobre los perfiles proteicos en los que se evaluó TPX2 también se evaluó la presencia de β-tubulina (tamaño molecular aparente: 51 kDa), proteína que se utilizó como control interno

de carga proteica, esperándose una señal similar en la líneas celulares parentales y transfectantes de ETV5.

5.3.2.3 *Western immunoblotting*

Una vez finalizada la electroforesis en geles de poliacrilamida, las proteínas fueron transferidas a membranas de nitrocelulosa (Amersham Hybond ECL, GE Healthcare, EEUU) mediante electrotransferencia a voltaje constante de 100 V durante una hora, empleando *buffer* Tobwin [Tris 25 mM, glicina 192 mM, metanol 20%], con el dispositivo sumergido en agua con hielo para mantener la temperatura de la transferencia a 0°C. La membrana sirve como soporte sólido para inmovilizar a las proteínas, facilitando su detección por parte del anticuerpo.

Las membranas fueron teñidas de manera reversible con solución de ácido acético 0,5% (v/v) complementada con 0,2% de colorante Rojo Ponceau S (p/v) para visualizar los perfiles proteicos ya transferidos e identificar la ubicación de la siembra, el frente de la corrida y los estándares proteicos para la correcta estimación del peso molecular.

Para detectar la proteína de interés, se realizaron **ensayos de inmunorrevelado** de las membranas de nitrocelulosa con anticuerpos específicos, utilizando protocolos estandarizados. En primer lugar, las membranas se sumergieron en solución de bloqueo de sitios inespecíficos, empleando *buffer* PBS con Tween-20 0,02% (v/v) (PBS-T 0,02%) suplementado con leche descremada al 10% (p/v) durante 45 minutos a temperatura ambiente en agitación constante. El objetivo de la incubación con la solución de bloqueo es minimizar la unión de la proteína a sitios de unión inespecíficos de la membrana. Luego fueron incubadas durante una hora a 37°C en presencia del anticuerpo monoclonal específico diluido en solución de bloqueo (anti TPX2 2 μg/ml). Como segundo anticuerpo, se utilizó el anti-IgG de ratón (Vector Laboratories Inc.) conjugado con peroxidasa de rábano picante (0,4 μg/ml) en solución de bloqueo durante una hora a temperatura ambiente, en agitación constante. Entre estos pasos, las membranas fueron lavadas tres veces consecutivas durante 5 minutos cada una con PBS-Tween 0,02%.

Para el revelado de los anticuerpos se utilizó el sistema de quimioluminiscencia "ECL Western Blotting Detection kit" (*ECL, Enhanced Chemiluminescence*; GE Healthcare, EEUU), siguiendo las instrucciones del fabricante. Se obtuvieron réplicas de tres experimentos.

Las bandas obtenidas en las placas de revelado fueron digitalizadas y la densidad de píxeles fue

determinada con el programa "Image J" (Wright Cell Imaging Facility, UNHR, Canadá) siguiendo un protocolo para la cuantificación de la señal (`http://www.yorku.ca/yisheng/Internal/Protocols/ImageJ.pdf`). Las densidades de las señales obtenidas para las diferentes proteínas analizadas se normalizaron empleando las densidades de las señales correspondientes a β-tubulina reveladas en la misma membrana. El anticuerpo anti β-tubulina fue usado a concentración 0,5 μg/ml.

5.3.3 Ensayos de inmunocitoquímica de fluorescencia

Las células a evaluar crecieron en monocapas sobre cubreobjetos de vidrio en placas de cultivo de 6 pocillos (Nunc-Thermo Scientific, EEEU) hasta un 80% de confluencia, como explica la Sección 5.1 de este Capítulo. Se retiró el medio de cultivo y se realizaron dos lavados con PBS 1x duramte 5 minutos cada uno. Las células fueron fijadas con PBS 1x suplementado con 4% de formaldehído (v/v). Dado que la proteína en estudio posee localización perinuclear, las células fueron permeabilizadas con PBS suplementado con 0,1% de Tritón X-10 (v/v), seguido por el bloqueo de sitios no específicos con PBS suplementado con 4% de BSA (p/v) por 45 minutos. La incubación con el anticuerpo primario o IgG control diluido en PBS, suplementado con 4% de BSA (p/v), se realizó durante una hora a temperatura ambiente. Se continuó con dos lavados de 5 minutos en solución PBS-T 0,02% para remover el exceso de anticuerpo primario. Luego, se incubó con el anticuerpo secundario conjugado a fluorocromos por una hora a temperatura ambiente y se realizaron dos lavados de 10 minutos en solución PBS-T 0.02%.

El anticuerpo primario utilizado en esta instancia fue **anti-TPX2** a 2 μg/ml. Como controles negativos se utilizaron IgG de ratón purificadas y agregadas a igual concentración que el anticuerpo primario. Se incubó a las células con 1 μg/ml de Hoechst 33342 (Sigma-Aldrich) en PBS por 5 minutos para la tinción de los núcleos celulares. Luego de un lavado con PBS, los cubreobjetos se montaron con solución Vectashield (Vector Laboratories Inc., EEUU) formulada para prevenir la caída de la señal fluorescente.

Se analizaron todos los preparados en un microscopio confocal láser Nikon C2 (filtros de excitación: 488 y 544 nm; filtros de emisión: 523-530 y 570-LP nm) o en un microscopio Nikon equipado para fluorescencia con cámara acoplada a un programa de captura y procesamiento de imágenes (IPLab Scientific Image Processing versión 3.06, Scanalytics Inc., EEUU). Las imá-

genes fueron luego analizadas con el *software* "Image J" (Wright Cell Imaging Facility, UNHCR, Canadá), siguiendo el protocolo sugerido para este programa (`https://sciencetechblog.files.wordpress.com/2011/05/measuring-cell-fluorescence-using-imagej.pdf`).

Resultados

Búsqueda e identificación de potenciales biomarcadores de CE

1 Resultados *in silico*

Con el objetivo de identificar potenciales biomarcadores de CE, se empleó un algoritmo que combina herramientas bioinformáticas de minería de texto y datos. A continuación se listan los pasos seguidos y los resultados obtenidos en cada caso.

1. Se utilizó un estudio de transcriptómica con microarreglos de ADN (GEO) para identificar genes diferencialmente expresados en CE. Se seleccionaron 39 genes con expresión diferencial para tres parámetros clínico-patológicos asociados a la agresividad del CE.

2. Se realizó una búsqueda de genes previamente reportados como asociados a CE (DisGeNET). Se identificaron 962 genes asociados a CE y dos listas relacionadas a los índices DPI y DSI (386 y 417 genes, respectivamente).

3. Se realizó un análisis de priorización génica (ToppGene) empleando la lista de genes seleccionados en el análisis de transcriptómica GEO como "lista de prueba" y las dos listas de genes de DisGeNET como "listas de entrenamiento". En este punto se desestimaron dos genes por figurar tanto en la lista de prueba como en las de entrenamiento. Como resultado se obtuvieron dos priorizaciones con 37 genes, cada una correspondiente a una lista de

entrenamiento (DPI y DSI).

4. Los 37 genes restantes fueron rastreados en un estudio de expresión génica global (TCGA) y aquellos sin información disponible fueron descartados. El resultado fue un listado de 33 genes, que fueron posteriormente evaluados y seleccionados empleando modelos estadísticos (método de Kaplan-Meier [resultado: 20 genes], OR [resultado: 16 genes] y modelo de riesgos proporcionales de Cox [resultado: 6 genes]) hasta llegar a una lista de 6 genes.

5. Los genes resultantes del análisis estadístico fueron identificados en la priorización génica. Tres de los 5 genes se encontraron priorizados entre los 10 primeros, por lo que fueron seleccionados nuevamente; los dos restantes se descartaron a partir de un análisis de rastreo realizado en el HPA, por no contar con información suficiente sobre su expresión proteica y de transcritos.

Por tanto, se seleccionó un conjunto de **tres genes** con expresión diferencial en CE que fueron evaluados en modelos experimentales de CE como última etapa del trabajo.

En la **Figura 3.1** se presenta un diagrama de flujo del algoritmo empleado para su identificación. Las secciones siguientes describen el análisis realizado en cada paso del proceso.

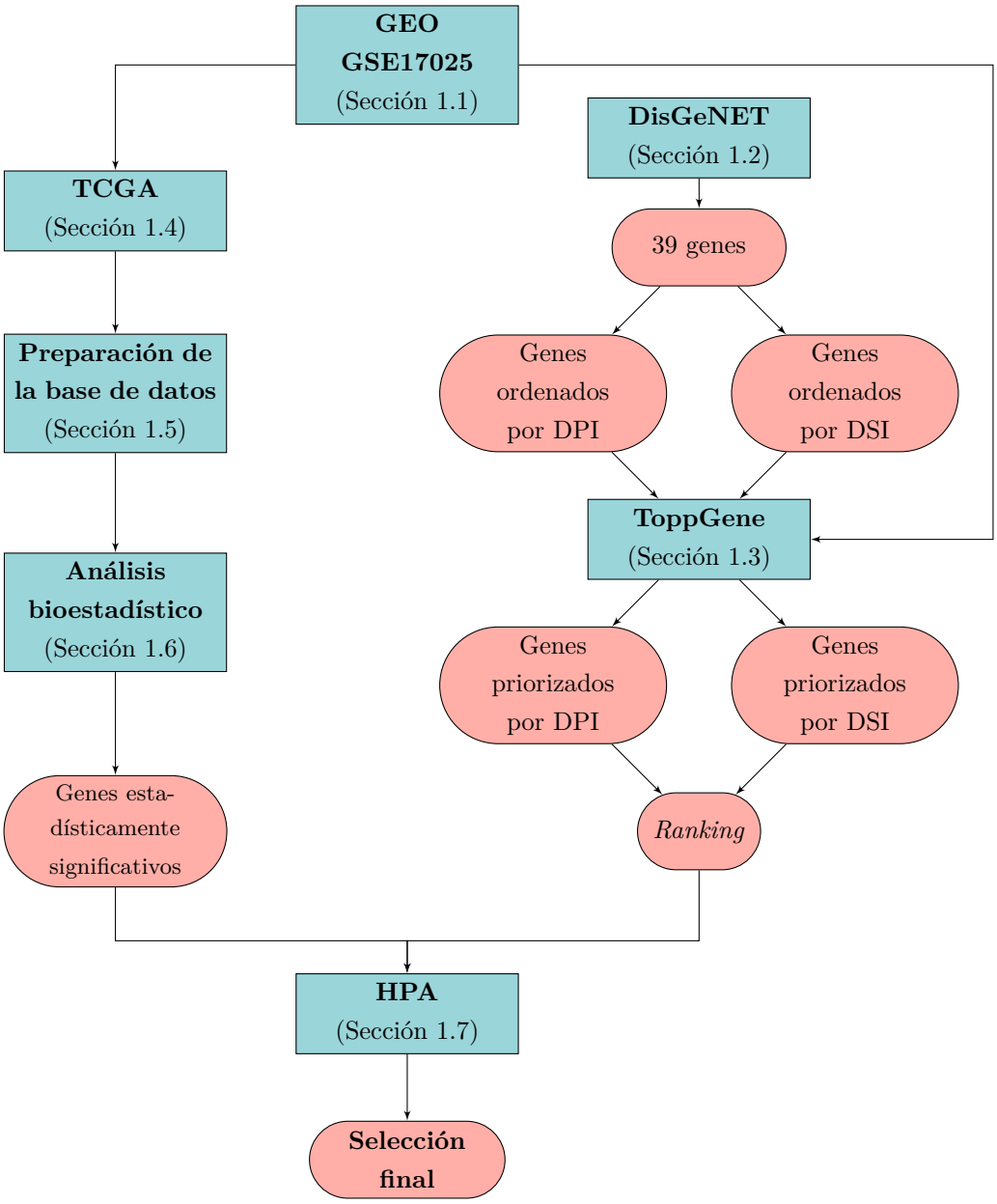


Figura 3.1. Diagrama de flujo de la metodología empleada en la búsqueda e identificación de potenciales de biomarcadores de CE. A partir de los genes expresados diferencialmente en tres características clínico-patológicas en el estudio GSE17025 (GEO) se realizó a) un análisis estadístico y b) un análisis de priorización (ToppGene) con genes cuya asociación con CE fue reportada en la literatura (DisGeNET). Los resultados de ambos se combinaron y contrastaron con la información disponible en HPA para definir la selección final de genes.

1.1 Análisis de expresión diferencial

Inicialmente, se realizó un análisis de expresión génica diferencial empleando el estudio GSE17025 disponible en el repositorio GEO. En la **Figura 3.2** se ilustra la metodología empleada en esta sección. Las características clínico-patológicas de las muestras analizadas se resumen en la **Tabla 10**.

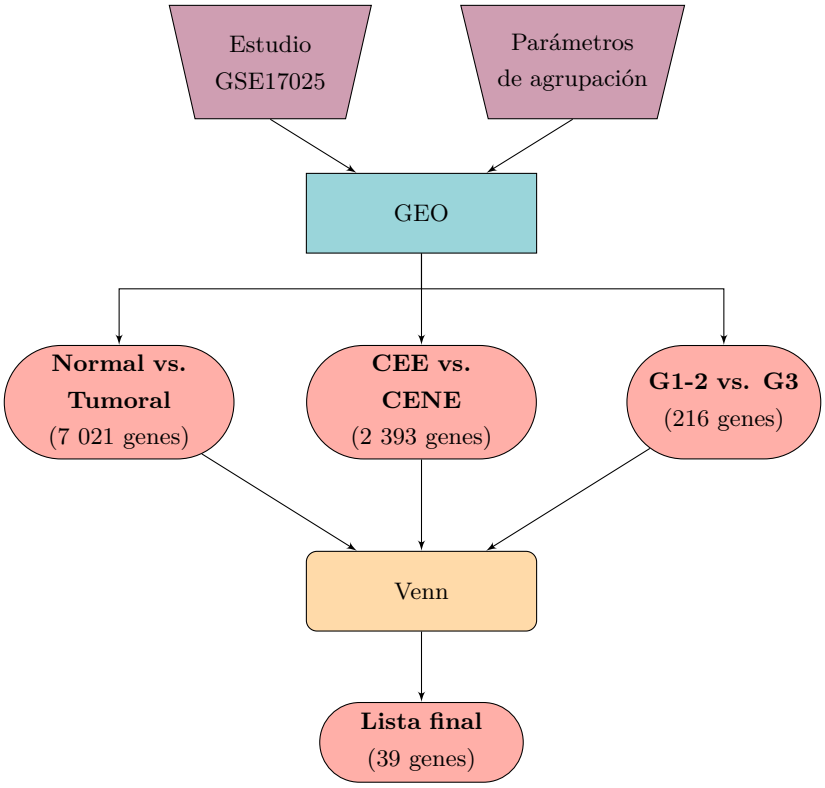


Figura 3.2. Diagrama de flujo del análisis de expresión diferencial. Se utilizó la información de expresión de microarreglos del estudio GSE17025 (disponible a través de la plataforma GEO) para evaluar la expresión diferencial en tres grupos: normal versus tumoral, endometrioide versus no endometrioide y G1-2 versus G3.

Características clínico-patológicas		
Muestras no tumorales		
	N	**%**
Endometrio atrófico	12	100%
Muestras tumorales		
Estadio	**N**	**%**
I	91	100%
Histología		
Endometrioide	79	86,81%
Seroso	12	13,19%
Grado histológico		
G1	30	32,97%
G2	36	39,56%
G3	25	27,47%
Invasión miometrial		
Superficial (<50%)	67	74,44%
Profunda (>50%)	23	25,56%

Tabla 10: Resumen de las características clínico-patológicas de las muestras del estudio GSE17025, disponible en la plataforma GEO.

Como resultado del análisis de expresión diferencial se obtuvieron tres listas (Materiales y Métodos, Sección 1.2). Los genes de cada una de ellas, luego de ser filtrados por valor p ($< 0,05$) y eliminar aquellos repetidos, son:

1. **Genes diferencialmente expresados entre muestras tumorales versus no tumorales**: 7 021 genes
2. **Genes diferencialmente expresados entre muestras tumorales de tipo CEE versus CENE**: 2 393 genes
3. **Genes diferencialmente expresados entre muestras tumorales de G1-2 versus G3**: 216 genes

A partir de estas listas, se seleccionaron los genes comunes a todos los grupos mediante un diagrama de Venn (**Figura 3.3**), resultando en una lista de candidatos con un total de **39 genes** (**Anexo D**). Este análisis contribuyó a enfocar la evaluación en aquellos genes asociados un pronóstico más desfavorable de CE.

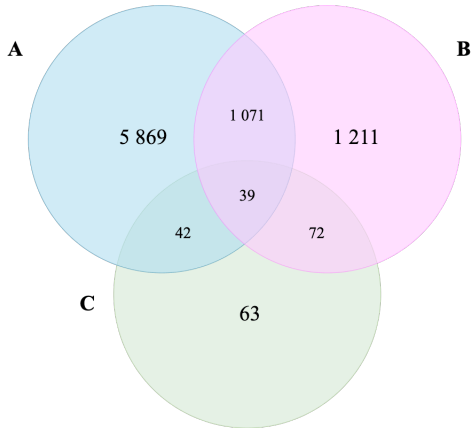


Figura 3.3. Diagrama de Venn representando la cantidad de genes diferencialmente expresados en cada característica clínica del estudio GSE17025. La intersección representa los genes comunes a los tres grupos, siendo **A** genes expresados diferencialmente en tejido tumoral respecto de no tumoral, **B** tumores de tipo CEE respecto de CENE y **C** tumores G1-2 versus G3.

Además, se registró la expresión de los 39 genes en el estudio GSE17025 de la plataforma GEO (**Tabla 11**) para los tres parámetros clínicos: tejido tumoral, subtipo y grado histológico. En cada caso se determinó la expresión aumentada o disminuida respecto de la media de los niveles de expresión de cada gen en todo el estudio. La **Tabla 11** muestra los resultados obtenidos.

Expresión diferencial de genes candidatos en GSE17025

	Tumoral	CENE	G3
ACSL5	Disminuida	Disminuida	Disminuida
ANAPC4	Disminuida	Disminuida	Disminuida
ARSD	Disminuida	Disminuida	Disminuida
ATAD2	Aumentada	Aumentada	Aumentada
CCDC160	Aumentada	Disminuida	Disminuida
CDC20B	Aumentada	Disminuida	Disminuida
CEP83	Disminuida	Disminuida	Disminuida
CREB3L4	Disminuida	Disminuida	Disminuida
DLGAP1-AS1	Disminuida	Disminuida	Disminuida
DLGAP1-AS2	Disminuida	Disminuida	Disminuida
FAM189A2	Disminuida	Disminuida	Disminuida
FOXA2	Disminuida	Disminuida	Disminuida
GSTP1	Disminuida	Disminuida	Disminuida
HAUS8	Aumentada	Aumentada	Aumentada
HES6	Aumentada	Disminuida	Disminuida
KIAA1324	Disminuida	Disminuida	Disminuida

KIF7	Disminuida	Aumentada	Aumentada
LINC00261	Disminuida	Disminuida	Disminuida
LINC00261	Disminuida	Disminuida	Disminuida
LPCAT2	Disminuida	Disminuida	Disminuida
PALMD	Disminuida	Disminuida	Disminuida
PCDH7	Disminuida	Disminuida	Disminuida
PDZRN3	Disminuida	Disminuida	Disminuida
PGR	Disminuida	Disminuida	Disminuida
PLEKHH1	Disminuida	Disminuida	Disminuida
PPM1H	Disminuida	Disminuida	Disminuida
PTCH1	Disminuida	Disminuida	Disminuida
SLC25A35	Disminuida	Disminuida	Disminuida
SLC47A1	Disminuida	Disminuida	Disminuida
SOAT1	Aumentada	Disminuida	Disminuida
SORBS2	Disminuida	Disminuida	Disminuida
SPATA6	Disminuida	Disminuida	Disminuida
TAB2	Aumentada	Aumentada	Aumentada
TBCEL	Aumentada	Disminuida	Disminuida
TMEM132A	Aumentada	Disminuida	Disminuida
TMPRSS2	Aumentada	Disminuida	Disminuida
TPX2	Aumentada	Aumentada	Aumentada
TRAF3IP2	Disminuida	Disminuida	Disminuida
ZDHHC2	Disminuida	Disminuida	Disminuida

Tabla 11: Expresión diferencial de los 39 genes seleccionados del estudio GSE17025 para tres características clínico-patológicas. Se presenta como se comporta la expresión de los genes para cada una de los grupos estudiados, respecto de su característica menos agresiva: tejido tumoral respecto de normal (**Tumoral**), tumores tipo CENE respecto de CEE (**CENE**) y tumores G3 respecto tumores G1-2 (**G3**), respecto de la expresión media del gen en todo el estudio.

1.2 Relevamiento de genes asociados a CE

Con el objetivo de relevar genes asociados a CE se empleó la herramienta DisGeNET. En la **Figura 3.4** se muestra el diagrama de flujo general de la sección. En primer lugar se relevaron los términos de búsqueda de enfermedades neoplásicas relacionadas a CE. Los términos hallados, junto con sus respectivos códigos UMLS CUI, se listan en la **Tabla 12**.

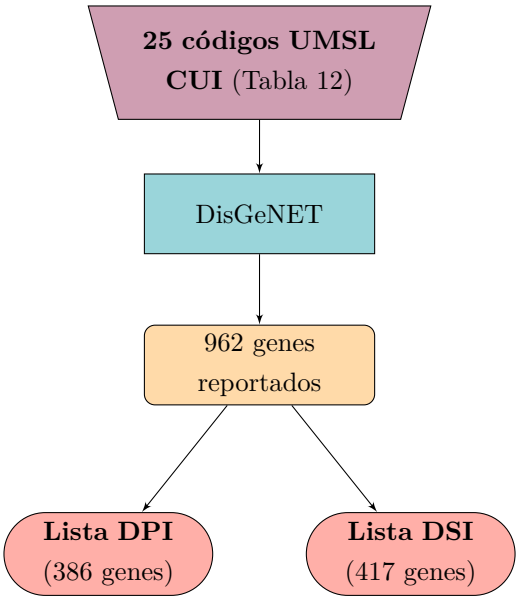


Figura 3.4. Diagrama de flujo del relevamiento de genes asociados a CE. A partir de 25 términos de búsqueda relacionados con enfermedades neoplásicas de CE se obtuvo un listado de 962 genes con GDA reportadas. Este fue filtrado mediante los índices DPI y DSI, resultando en dos listas de 386 y 417 genes, respectivamente.

Términos de búsqueda utilizados en DisGeNET

Término de búsqueda	CUI
Carcinoma In Situ of Endometrium	C0346191
Endometrial Adenocarcinoma	C1153706
Endometrial Carcinoma	C0476089
Endometrial Clear Cell Adenocarcinoma	C0279765
Endometrial Endometrioid Adenocarcinoma	C1336905
Endometrial Intraepithelial Neoplasia	C1333394
Endometrial Neoplasms	C0014170
Endometrial Neoplasm Malignant Metastatic	C0278801
Endometrial Neoplasm Malignant Stage I	C0278798
Endometrial Sarcoma	C2959547
Endometrial Serous Adenocarcinoma	C1336921
Endometrial Squamous Cell Carcinoma	C1333396
Endometrial Stromal Sarcoma	C0206630
Endometrial Stromal Sarcoma, High Grade	C2239246
Endometrial Stromal Tumors	C0334695
Endometrial Undifferentiated Carcinoma	C1516865
Low Grade Endometrial Stromal Sarcoma	C0334486
Malignant Neoplasm of Endometrium	C0007103

Metastatic endometrial carcinoma	C0813148
Recurrent Endometrial Cancer	C0278802
Serous Endometrial Intraepithelial Carcinoma	C1516857
Stage, Endometrial Cancer	C0280255
Stage I Endometrial Carcinoma	C0813147
Type I Endometrial Adenocarcinoma	C1519719
Type II Endometrial Adenocarcinoma	C1519714

Tabla 12: Términos de búsqueda de DisGeNET, con sus respectivos códigos UMLS CUI. Estos términos fueron seleccionados por representar todas las enfermedades neoplásicas relativas a CE de la plataforma.

Todos los términos de la plataforma identificados en la literatura en asociación a CE fueron empleados en el análisis de asociación gen-enfermedad. Esto tuvo por interés completar un análisis amplio y robusto y obtener un listado completo de los genes asociados a la enfermedad. Como resultado del análisis se identificó un total de **962 genes**.

Tal como se mencionó en la Sección 1.1 de Materiales y Métodos, DisGeNET proporciona dos índices que proveen información sobre la especificidad y la cantidad de enfermedades a las que un gen está relacionado: DSI y DPI, respectivamente. Estos índices adoptan valores entre 0 y 1, determinándose un DSI bajo si un gen se asocia a una gran cantidad de enfermedades (definidas por términos UMSL CUI) y un DPI bajo si se encuentra asociado a pocas clases de enfermedades (definidas por los términos MeSH). Seguidamente al relevamiento de genes en DisGeNET, se realizó el filtrado de la lista de 962 genes identificados según estos índices. La media aritmética para los valores de DPI y DSI en la lista fue de 0,624 y 0,547, respectivamente. Dichas medias fueron utilizadas independientemente como puntos de corte para filtrar la lista preliminar de genes, lo que resultó en dos listas de 386 y 417 elementos, respectivamente, disponibles en el **Anexo E**.

1.3 Priorización génica y análisis de enriquecimiento funcional

La lista de 39 genes diferencialmente expresados en el estudio GSE17025 representa una primera aproximación a la selección de biomarcadores tumorales de CE. A continuación se empleó la herramienta ToppGene (Sección 1.3 de Materiales y Métodos) para la priorización génica y el análisis de las vías de señalización y procesos biológicos representativos de los genes en la lista a fin de determinar de qué modo los genes de expresión diferencial seleccionados a partir del muestreo GEO

se relacionan con los ya reportados en la literatura e identificados con la herramienta DisGeNET. En la **Figura 3.5** se resume el procedimiento.

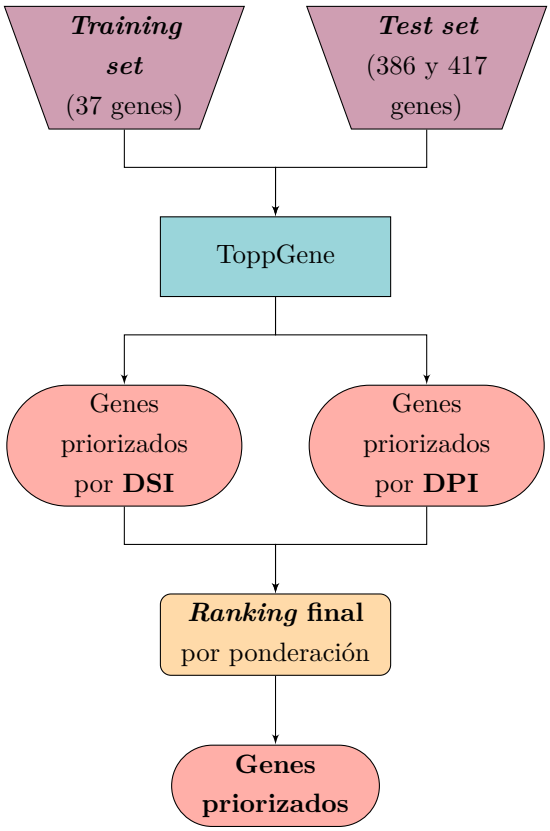


Figura 3.5. Diagrama de flujo de la metodología empleada en la priorización génica con ToppGene. Se utilizó como *training set* el listado de 39 genes obtenido luego del análisis de expresión diferencial, al que luego se le eliminaron los genes ATAD2 y KIAA1324 por ya estar reportados en ambos *test sets*. Dado que se realizaron dos priorizaciones, se utilizaron como test sets las listas obtenidas luego del relevamiento de genes asociados a CE: la lista filtrada por DPI (386 genes) y la filtrada por DSI (417 genes). Como resultado, se obtuvo una lista de genes ordenados según un *ranking* final que surge como el promedio de cada *ranking* particular.

La comparación de los genes de la Sección 1.1 de este capítulo (como **lista de prueba** a priorizar) con los de las dos listas filtradas por DPI y DSI de la Sección 1.2 del mismo capítulo (como **listas de entrenamiento**) dio como resultado dos priorizaciones distintas. En otras palabras, por cada gen de los 39 inicialmente listados en el **Anexo D** se obtuvieron dos posiciones en la priorización, una según la similitud con los genes de la lista de DPI y otra según DSI (**Tabla 13**). En ambos casos, la posición de cada gen está dada por la significación estadística de la comparación efectuada por ToppGene. En este punto es importante aclarar que la lista final contiene un número menor

de genes que los inicialmente contrastados, dado que se identificaron genes en ambas listas (DSI y DPI) repetidos en la lista de entrenamiento (DisGeNET). De este modo, los genes ATAD2 y KIAA1324, ya reportados como asociados a CE, fueron excluidos. Como muestra la **Tabla 13**, se calculó el promedio de estos resultados para el *ranking* final.

Resultados de la priorización génica en ToppGene

Gen	Posición DPI	Posición DSI	Promedio	Posición Final
PGR	2	1	1,5	1
PTCH1	1	3	2	2
FOXA2	3	2	2,5	3
SOAT1	5	6	5,5	4
TPX2	6	5	5,5	4
TMPRSS2	7	4	5,5	4
HES6	4	8	6	5
SORBS2	8	7	7,5	6
PCDH7	11	9	10	7
TAB2	10	11	10,5	8
ACSL5	9	15	12	9
GSPT1	12	12	12	9
PALMD	15	10	12,5	10
TRAF3IP2	13	16	14,5	11
CREB3L4	17	14	15,5	12
TMEM132A	19	13	16	13
SPATA6	16	20	18	14
ARSD	14	24	19	15
PPM1H	20	18	19	15
PDZRN3	18	21	19,5	16
ZDHHC2	23	17	20	17
ANAPC4	22	19	20,5	18
KIF7	21	23	22	19
LPCAT2	25	22	23,5	20
HAUS8	26	27	26,5	21
TBCEL	28	26	27	22
SLC25A35	24	31	27,5	23
CEP83	27	28	27,5	23
FAM189A2	31	25	28	24
PLEKHH1	30	30	30	25
SLC47A1	29	32	30,5	26

CDC20B	32	29	30,5	26
CCDC160	33	33	33	27
CCDC160	34	34	34	28
DLGAP1-AS2	35	36	35,5	29
DLGAP1-AS1	36	35	35,5	29
LOC100129098	37	37	37	30

Tabla 13: Resultados de la priorización génica con ToppGene. En la segunda columna se observa la posición de cada gen al ser priorizado con la lista de DisGeNET filtrada por DPI. En la tercera, al ser priorizado con el listado filtrado por DSI. A continuación se muestra el resultado del promedio de ambas posiciones. Por último, la columna 'Posición Final' enumera los 37 genes de acuerdo con el promedio anteriormente calculado (los genes con igual promedio se encuentran en la misma posición). Dos de los genes fueron excluidos durante el análisis de priorización por encontrarse en la lista de genes asociados a CE.

Respecto del enriquecimiento funcional, el análisis con ToppGene demostró que los genes priorizados se asocian a 1419 términos GO. En las **Tablas 14** y **15** se describen los 15 términos GO más representativos para el conjunto de genes previamente priorizados con cada una de las listas de entrenamientos (por DPI y DSI), así como el valor p y FDR B&H[1] asociados.

GO: Función Molecular				
	ID	**Nombre**	**valor p**	**FDR B&H**
1	GO:0003682	*chromatin binding*	3,88e-08	4,07e-05
2	GO:0003690	*double-stranded DNA binding*	2,18e-06	1,15e-03
3	GO:0005102	*signaling receptor binding*	7,54e-06	2,64e-03
4	GO:0044212	*transcription regulatory region DNA binding*	2,37e-05	6,00e-03
5	GO:0001067	*regulatory region nucleic acid binding*	2,86e-05	6,00e-03
GO: Proceso biológico				
	ID	**Nombre**	**valor p**	**FDR B&H**
1	GO:0048608	*reproductive structure development*	6,11e-11	1,24e-07
2	GO:0042127	*regulation of cell proliferation*	6,53e-11	1,24e-07
3	GO:0061458	*reproductive system development*	7,88e-11	1,24e-07
4	GO:0022414	*reproductive process*	1,18e-10	1,24e-07
5	GO:0000003	*reproduction*	1,26e-10	1,24e-07
GO: Componente celular				
	ID	**Name**	**valor p**	**FDR B&H**
1	GO:0005667	*transcription factor complex*	9,43e-02	4,00e+01
2	GO:0044427	*chromosomal part*	1,82e-01	4,00e+01
3	GO:0016327	*apicolateral plasma membrane*	2,92e-01	4,00e+01

[1] *False Discovery Rate* (FDR, por sus siglas en inglés) es un método de conceptualización del error de tipo I para hipótesis nulas de múltiples comparaciones, calculado con el procedimiento Benjamini-Hochberg (B&H)

4	GO:0043513	*inhibin B complex*	3,12e-01	4,00e+01
5	GO:0000785	*chromatin*	4,02e-01	4,06e+01

Tabla 14: Principales términos GO (según función molecular, proceso biológico y componente celular) asociados a la lista de genes priorizados por DPI.

GO: Función molecular				
	ID	**Nombre**	**valor p**	**FDR B&H**
1	GO:0003682	*chromatin binding*	3,76e-07	4,04e-04
2	GO:0018748	*iprodione amidohydrolase activity*	9,07e-06	5,13e-04
3	GO:0034882	*cis-aconitamide amidase activity*	9,07e-06	5,13e-04
4	GO:0043747	N2-*acetyl-L-lysine deacetylase activity*	9,07e-06	5,13e-04
5	GO:0034573	*didemethylisoproturon amidohydrolase activity*	9,07e-06	5,13e-04
GO: Proceso biológico				
	ID	**Nombre**	**valor p**	**FDR B&H**
1	GO:0009967	*positive regulation of signal transduction*	1,11e-09	3,47e-06
2	GO:0071495	*cellular response to endogenous stimulus*	2,54e-09	3,47e-06
3	GO:0023056	*positive regulation of signaling*	3,22e-09	3,47e-06
4	GO:0042127	*regulation of cell proliferation*	3,62e-09	3,47e-06
5	GO:0010557	*positive regulation of macromolecule biosynthetic process*	4,05e-09	3,47e-06
GO: Componente Celular				
	ID	**Nombre**	**valor p**	**FDR B&H**
1	GO:0000785	*chromatin*	7,36e-04	4,01e-01
2	GO:0044427	*chromosomal part*	1,81e-03	4,93e-01
3	GO:0005694	*chromosome*	4,18e-03	7,59e-01
4	GO:0043511	*inhibin complex*	6,93e-03	9,44e-01
5	GO:0016327	*apicolateral plasma membrane*	2,24e-02	2,45e+00

Tabla 15: Principales términos GO (según función molecular, proceso biológico y componente celular) asociados a la lista de genes priorizados por DSI.

En las **Tablas 16** y **17** se enumeran las vías de señalización más representadas en el conjunto de genes priorizados (para las listas de DPI y DSI), junto con el valor p, FDR B&H y los genes correspondientes a cada vía.

Vías de señalización					
	ID	**Nombre**	**valor p**	**FDR B&H**	**Genes**
1	82940	*Steroid hormone biosynthesis*	6,45e-06	8,78e-03	58
2	1268754	*Glycoprotein hormones*	1,20e-04	8,14e-02	12

3	1268753	*Peptide hormone biosynthesis*	3,74e-04	1,55e-01	14
4	852705	*MicroRNAs in cancer*	4,55e-04	1,55e-01	299
5	790011	*Ovarian steroidogenesis*	1,83e-03	4,97e-01	50
6	1270046	*Metabolism of steroid hormones*	6,75e-03	1,53e+00	32
7	413396	*Steroid hormone biosynthesis, cholesterol =>prognenolone =>progesterone*	1,26e-02	2,44e+00	3
8	413376	*Chondroitin sulfate degradation*	1,83e-02	3,10e+00	8
9	1427841	TFAP2 (AP-2) *family regulates transcription of growth factors and their receptors*	2,27e-02	3,43e+00	16
10	1269973	*Hyaluronan metabolism*	3,15e-02	3,99e+00	17

Tabla 16: Vías de señalización más representativas de los genes priorizados con la lista DPI. La columna **Genes** ofrece un hipervínculo al listado de genes del *training set* asociados a cada vía de señalización. Fuentes: KEGG (*Kyoto Encyclopedia of Genes and Genomes*), REACTOME y *Pathway Interaction Database*.

Vías de señalización					
	ID	**Nombre**	**valor p**	**FDR B&H**	**Genes**
1	1268754	*Glycoprotein hormones*	7,15e-08	1,10e-04	12
2	1268753	*Peptide hormone biosynthesis*	4,15e-07	3,18e-04	14
3	82940	*Steroid hormone biosynthesis*	1,86e-04	9,51e-02	58
4	1427841	TFAP2 (AP-2) *family regulates transcription of growth factors and their receptors*	1,65e-03	6,31e-01	16
5	1427839	*Transcriptional regulation by the AP-2* (TFAP2) *family of transcription factors*	5,97e-03	1,83e+00	40
6	852705	*MicroRNAs in cancer*	7,77e-03	1,99e+00	299
7	413376	*Chondroitin sulfate degradation*	2,63e-02	5,76e+00	8
8	137997	Signaling events mediated by HDAC *Class I*	4,05e-02	7,53e+00	66
9	413375	*Dermatan sulfate degradation*	4,64e-02	7,53e+00	9
10	1269973	*Hyaluronan metabolism*	4,91e-02	7,53e+00	17

Tabla 17: Vías de señalización más representativas de los genes priorizados con la lista DSI. La columna **Genes** ofrece un hipervínculo al listado de genes del *training set* asociados a cada vía de señalización. Fuentes: KEGG (*Kyoto Encyclopedia of Genes and Genomes*), REACTOME y *Pathway Interaction Database*.

1.4 Rastreo de genes en TCGA

El siguiente paso fue la realización de un análisis estadístico, que involucró la evaluación de los genes previamente identificados con información clínico-patológica proveniente del estudio de expresión génica global *Uterine Corpus Endometrioid Cancer* de TCGA (TCGA-UCEC). Del mismo modo en que en la priorización génica dos genes fueron excluidos del análisis por haberse identificado que habían sido reportados en la literatura, se eliminaron de la lista de los genes diferencialmente expresados de GEO aquellos sin información disponible en el estudio de secuenciación global de TCGA. Los genes eliminados del listado por este motivo fueron **DLGAP1-AS1, DLGAP1-AS2, LINC00261 y LOC100129098**.

Por tanto, de los 39 genes inicialmente identificados en la Sección 1.1 del presente capítulo, el listado se redujo a **33 genes**. La **Tabla 18** presenta la lista de los genes resultantes.

Genes candidatos

Gen	Nombre del gen
ACSL5	*acyl-CoA synthetase long chain family member 5*
ANAPC4	*anaphase promoting complex subunit 4*
ARSD	*arylsulfatase D*
CCDC160	*coiled-coil domain containing 160*
CDC20B	*cell division cycle 20B*
CEP83	*centrosomal protein 83*
CREB3L4	*cAMP responsive element binding protein 3-like 4*
FAM189A2	*family with sequence similarity 189 member A2*
FOXA2	*forkhead box A2*
GSPT1	*G1 to S phase transition 1*
HAUS8	*HAUS augmin like complex subunit 8*
HES6	*hes family bHLH transcription factor 6*
KIF7	*kinesin family member 7*
LPCAT2	*lysophosphatidylcholine acyltransferase 2*
PALMD	*palmdelphin*
PCDH7	*protocadherin 7*
PDZRN3	*PDZ domain containing ring finger 3*
PGR	*progesterone receptor*
PLEKHH1	*pleckstrin homology, MyTH4 and FERM domain containing H1*
PPM1H	*protein phosphatase, Mg2+/Mn2+ dependent 1H*
PTCH1	*patched 1*
SLC25A35	*solute carrier family 25 member 35*
SLC47A1	*solute carrier family 47 member 1*
SOAT1	*sterol O-acyltransferase 1*
SORBS2	*sorbin and SH3 domain containing 2*
SPATA6	*spermatogenesis associated 6*
TAB2	*TGF-beta activated kinase 1 (MAP3K7) binding protein 2*
TBCEL	*tubulin folding cofactor E like*
TMEM132A	*transmembrane protein 132A*
TMPRSS2	*transmembrane serine protease 2*
TPX2	*TPX2 microtubule nucleation facto*
TRAF3IP2	*TRAF3 interacting protein 2*
ZDHHC2	*zinc finger DHHC-type containing 2*

Tabla 18: 33 genes candidatos. Fueron obtenidos a partir de la comparación de las tres listas de expresión diferencial de GEO, eliminando aquellos que ya se encontraban reportados en la literatura (ATAD2 y KIAA1324) y aquellos sin información en el estudio de secuenciación global de TCGA (DLGAP1-AS1, DLGAP1- AS2, LINC00261 y LOC100129098).

1.5 Preparación y análisis exploratorio de la base de datos

Previo al análisis estadístico de la información genómica sobre los 33 genes seleccionados según se detalla en las secciones anteriores, fue necesario llevar a cabo el diseño de la base de datos (ver Materiales y Métodos, Sección 4.1). Se utilizó la información disponible en el repositorio TCGA del estudio TCGA-UCEC a través de la plataforma Xena, por contar con una interfaz óptima para realizar la selección de la información. La descarga de la información de TCGA-UCEC y el diseño estadístico resultaron en una base de datos de **514 registros** (correspondientes a muestras de pacientes con CE), 8 variables de respuesta (estado de RFS, tiempo de RFS, estados de OS, tiempo de OS, grado y subtipo histológico, estadio del tumor e invasión miometrial) y 33 variables de agrupación correspondientes a los niveles de expresión de transcritos de cada gen en estudio.

El siguiente paso fue el análisis exploratorio de la base de datos mediante estadística descriptiva. La **Tabla 19** resume las características clínico-patológicas de las pacientes del muestreo analizado.

Características clínico-patológicas		
	N	**%**
Estadio		
I	325	63,23%
II	54	10,51%
III	117	22,76%
IV	18	3,50%
Subtipo histológico		
Endometrioide (CEE)	391	76,07%
Seroso	103	20,04%
Mixto	20	3,89%
Grado histológico		
G1	94	18,29%
G2	118	22,96%
G3	293	57%
Grado alto	9	1,75%
Invasión miometrial		
<50%	299	58,17%
>50%	215	41,82%
Edad (años)	$63,73$	$\pm 10,91$

Tabla 19: Características clínico-patológicas del estudio TCGA-UCEC, con la cantidad de muestras de cada categoría ('N') y la proporción de esas muestras sobre el total ('%').

El análisis reveló que la **edad al momento del diagnóstico** presenta una distribución normal, con una media de $63,7$ años y un desvío estándar de $10,9$ años. La edad de detección mínima es 31 años y la máxima, 90. La mediana es 63 años y su similitud a la media refuerza la normalidad de la distribución. La **Figura 3.6** muestra un histograma de edades al momento del diagnóstico.

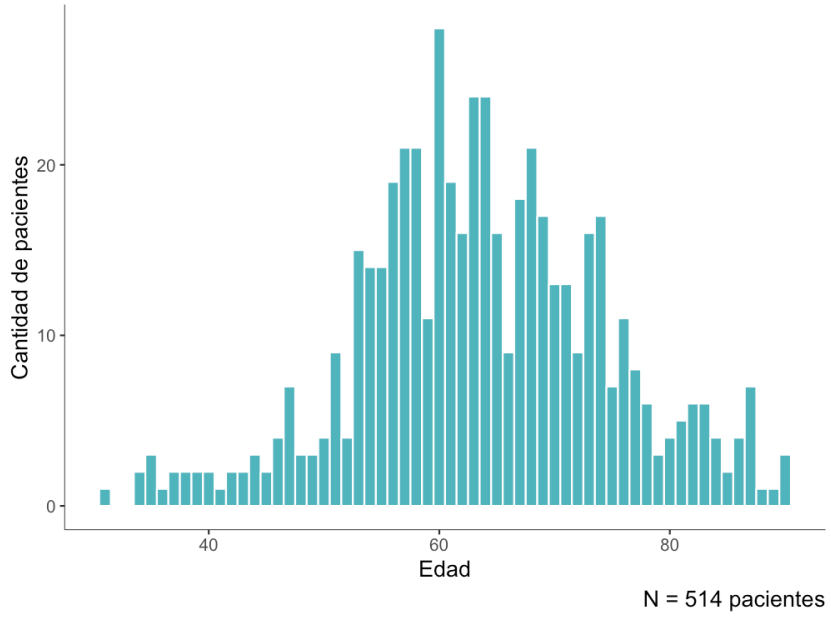


Figura 3.6. Histograma de edades al momento del diagnóstico para pacientes del estudio TCGA-UCEC (514 observaciones). Se observa mayor incidencia de diagnóstico en mujeres entre 55 y 70 años de edad; esto es coherente con la mayor incidencia de CE en mujeres postmenopáusicas.

La variable edad guarda relación con el **estado menopáusico** de las pacientes. Las muestras del estudio representan 31 casos en estado premenopáusico (menos de 6 meses desde el último período menstrual, sin ooforectomía previa y sin reemplazo de estrógeno), 19 en estado perimenopáusico (entre 6 y 12 meses desde el último período menstrual), 417 en estado postmenopáusico (más de 12 meses desde el último período menstrual sin histerectomía previa o con ooforectomía bilateral previa) y 47 de estado indeterminado, es decir, sin caracterizar. La **Figura 3.7** representa gráficamente lo anterior y deja en evidencia un sesgo en el estudio ya que una fracción mayor al 80% de las muestras proviene de pacientes postmenopáusicas.

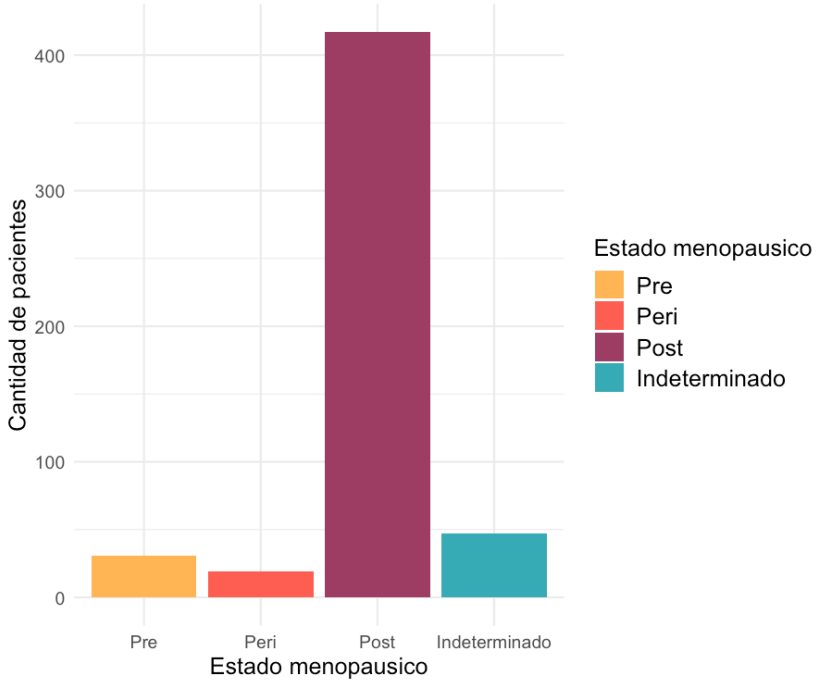


Figura 3.7. Representación de estado menopáusico al momento del diagnóstico para pacientes del estudio TCGA-UCEC. De 514 muestras, más de 400 corresponden a pacientes postemopáusicas (∼80%), resaltando a esta característica como factor de riesgo de la enfermedad.

En TCGA-UCEC, la variable **estadio** presenta 12 niveles que se corresponden con los de la **Tabla 4**. Estos fueron agrupados en cuatro categorías (estadio I, II, III y IV) para facilitar la interpretación de la información. La **Figura 3.8** presenta un gráfico circular con el porcentaje correspondiente a cada una. A simple vista se nota mayor presencia de estadios tempranos, que alcanzan cerca del 75% del total de muestras. Del porcentaje restante, el estadio IV es el menos representado en el estudio (4%).

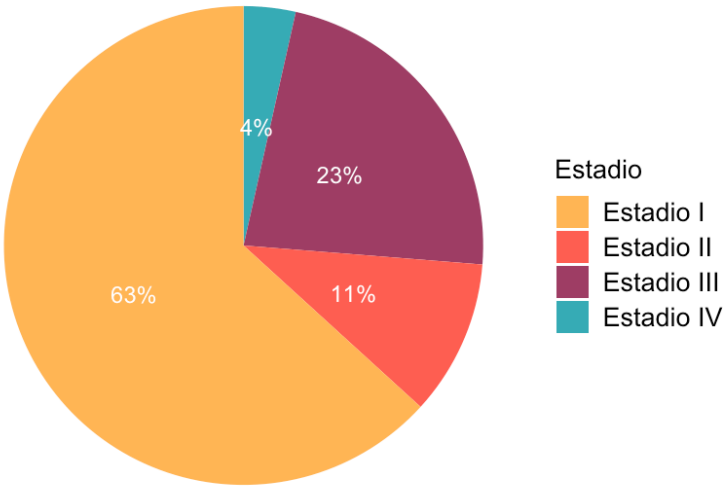


Figura 3.8. Representación de estadios de CE en el estudio TCGA-UCEC. Cerca de dos tercios de las muestras tumorales analizadas en el estudio corresponden a estadio I, un cuarto a estadio III, 10% a estadio II y menos del 5% a estadio IV. Esto indica un sesgo en el estudio hacia estadios tempranos.

De acuerdo al **subtipo histológico** y según lo expuesto en la Sección 2.4.1 de la Introducción, el CE puede clasificarse en CEE (tipo I) y CENE (tipo II). Entre los tumores CENE se distinguen, a su vez, los subtipos serosos, mucinosos, de células claras y mixtos. El estudio TCGA-UCEC describe solamente muestras de CEE, CENE mixto y CENE seroso y en las proporciones que presenta la **Figura 3.9**. Nuevamente queda evidenciado el marcado desbalance en el estudio, con mayor cantidad de casos de subtipo CEE, el menos agresivo de los tres. El subtipo CENE seroso representa un quinto de todas las muestras, mientras que el mixto menos del 5%.

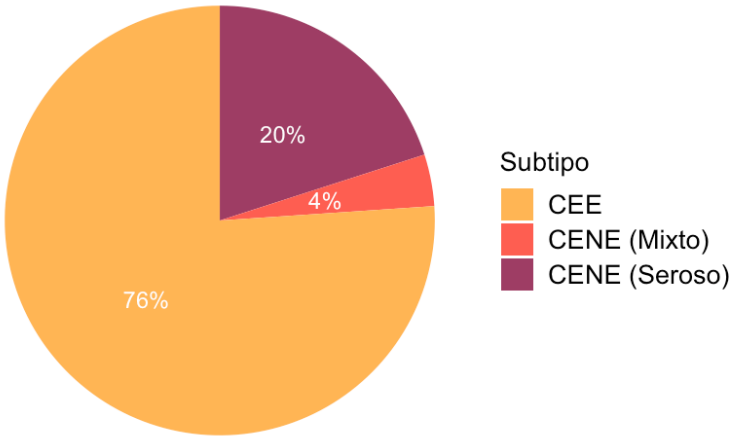


Figura 3.9. Representación de subtipos histológicos de CE en el estudio TCGA-UCEC. Más del 75% de las muestras analizadas en el estudio son de CEE, y las restantes, de tumores CENE, principalmente serosos. Nuevamente se observa un marcado desbalance en el estudio, en este caso con preponderancia de la característica menos agresiva.

Como ilustra la **Figura 3.10**, la representación de cada **grado histológico** de las pacientes con CE en el estudio clínico es: 59% de tumores G3, 23% de tumores G2 y 18% de tumores G1. A diferencia de las características clínico-patológicas anteriores, donde los rasgos menos agresivos presentan mayor incidencia, en este caso se observa un predominio de muestras de G3. Para continuar el análisis, la **Figura 3.11** presenta una gráfica de barras de los estadios en estudio (nuevamente, agrupados en estadio I, II, III y IV) donde los colores de la leyenda indican la cantidad de muestras de cada grado. De este modo se puede notar que la mayor parte de los casos de bajo grado (G1 y G2) pertenecen al estadio I, mientras que las muestras en estadios II, III y IV son predominantemente tumores G3. En otras palabras, los estadios más avanzados corresponden mayoritariamente a tumores de grado alto y, por lo tanto, muy agresivos. Por otra parte, en las muestras de estadio I se distribuyen más equitativamente los grados bajos (G1 y G2) y alto (G3).

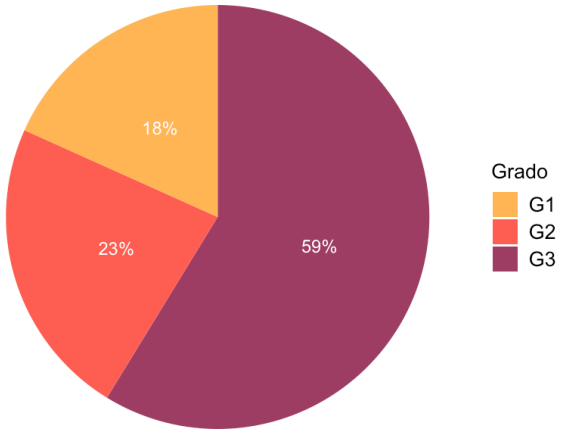


Figura 3.10. Representación de grados histológicos de CE en el estudio TCGA-UCEC. Más de la mitad de las muestras corresponden a tumores G3, aproximadamente un cuarto a tumores G2 y las restantes, a tumores G1. De este modo se puede notar que el estudio presenta un sesgo hacia los grados más agresivos.

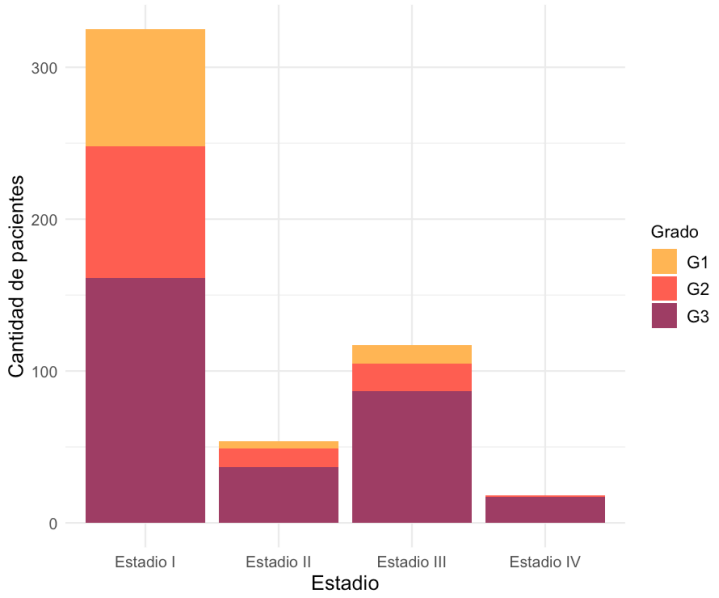


Figura 3.11. Cantidad de muestras por estadio y grado histológico en el estudio TCGA-UCEC. Cantidad de pacientes con G1, G2 y G3 para cada estadio (I, II, III y IV) en el estudio.

De los casos analizados, 76 fallecieron en el período de seguimiento del estudio sin información disponible al respecto (el motivo de la muerte es desconocido). Por otro lado, 85 pacientes

presentaron recurrencia durante el estudio, siendo el tiempo medio de recurrencia de 826 días.

1.5.1 Categorización de las variables de agrupación

Para cada gen en estudio se definió un punto de corte en los niveles de expresión de transcritos, siguiendo la metodología expuesta en Materiales y Métodos, Sección 4.1.1.2. De este modo, a partir de los niveles de expresión de cada gen en las muestras del estudio TCGA-UCEC y la variable **estado de RFS** se aplicó la prueba de *logRanks* sobre cada posible punto de corte para encontrar aquel que maximiza la diferencia entre las curvas de sobrevida para expresión aumentada y disminuida del gen. De esta forma, se obtuvo un valor p para el punto de corte óptimo de cada gen. La **Tabla 20** muestra la definición del valor de corte de cada gen y el valor p asociado.

Una vez encontrados los puntos de corte, los niveles de expresión de cada gen en cada muestra del estudio se categorizaron en expresión "aumentada" o "disminuida" de acuerdo a si los valores de expresión son superiores o inferiores a los puntos de corte de la **Tabla 20**, respectivamente. A continuación se descartaron aquellos genes con valor p mayor a 0,05 dado que la categorización no se considera estadísticamente significativa. De esta manera se seleccionaron **18 genes**. El **Anexo F** muestra las curvas de sobrevida obtenidas mediante Kaplan-Meier para la expresión aumentada y disminuida de estos genes con la variable RFS (36 curvas en 18 figuras) y OS (36 curvas en 18 figuras).

Puntos de corte de cada gen		
Gen	**Punto de corte**	**Valor p**
PGR	10,22	1,04e-06
TPX2	11,06	3,27e-06
FAM189A2	6,593	5,41e-06
TMPRSS2	8,57	5,96e-05
TRAF3IP2	9,54	6,45e-05
SLC47A1	8,90	1,06e-04
PTCH1	8,12	1,41e-04
SLC25A35	7,63	2,52e-04
ARSD	11,250	3,84e-04
ANAPC4	9,17	4,413e-04
SORBS2	9,82	2,66e-03
PLEKHH1	9,505	3,72e-03
LPCAT2	7,96	6,75e-03
ACSL5	10,590	7,96e-03
PALMD	8,831	1,08e-02
TAB2	10,4	1,48e-02
GSPT1	11,04	2,2e-02
PDZRN3	7,93	2,42e-02

Tabla 20: Puntos de corte en los niveles de expresión de transcritos de los 18 genes seleccionados. Se utilizó el método de *logRanks* y la información de TCGA-UCEC sobre expresión génica y estado de RFS de todas las pacientes. Los genes con puntos de corte no significativos fueron excluidos del análisis.

1.6 Análisis estadístico

A continuación se detallan los resultados del análisis estadístico realizado sobre la base de datos del estudio TCGA-UCEC, con el objetivo de identificar los genes más prometedores como biomarcadores de CE. La **Figura 3.12** esquematiza la metodología y los resultados de esta etapa.

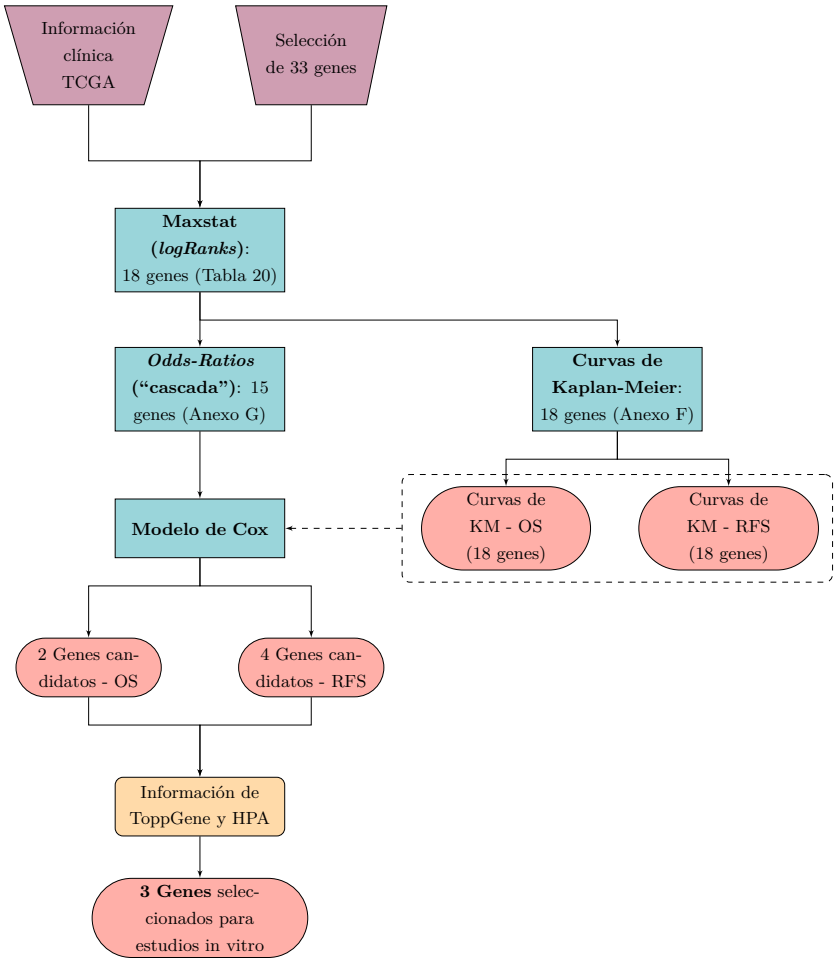


Figura 3.12. Diagrama de flujo de la preparación de la base de datos de TCGA-UCEC y el análisis estadístico. Se tomó la lista de 33 genes obtenida a partir del análisis de expresión diferencial (sin los cuatro genes reportados en TCGA) en conjunto con la información clínica del estudio TCGA-UCEC. Se categorizó la base de datos con el paquete `maxstat` de R (prueba de *logRanks*) para luego medir la fuerza de asociación de cada gen con las características clínico-patológicas de interés con los OR. Por último, se realizaron dos modelos de regresión de Cox, uno con la variable de respuesta OS y otro con la variable de respuesta RFS. Los genes resultantes de cada análisis se cotejaron con la priorización génica de ToppGene y la información disponible en HPA para terminar de definir el listado de genes para los estudios *in vitro*.

1.6.1 *Odds Ratios* (OR)

Para medir la "fuerza" de asociación de cada gen con las características clínico-patológicas de interés, se realizó el **cálculo en "cascada" de *Odds Ratios* (OR)** para las siguientes características clínico-patológicas: RFS, OS, histología, grado, estadio e IM. Los detalles de cómo se calcula esta medida de asociación se encuentran en la Sección 4.3.1 de Materiales y Métodos.

Fueron descartados del análisis aquellos genes sin asociación significativa (valor p o IC no significativos, ver **Tabla 6**) con al menos un parámetro clínico-patológico. De este modo, los genes PDZRN3, GSPT1 y TAB2 quedaron excluidos del estudio por no asociarse significativamente con estadio, grado y RFS, resultando en la selección de un total de **15 genes** con medidas de efecto significativas sobre las características clínico-patológicas estudiadas (**Figura 3.13**). En el **Anexo G** se listan los valores obtenidos para cada gen y todas las características clínico-patológicas.

Además de analizar si el efecto de la variación en la expresión de los genes sobre los parámetros clínicos es significativo, el cálculo de OR permite determinar si es la expresión aumentada (**sobreexpresión**) o disminuida (**subexpresión**) de cada gen se asocia a tumores con características agresivas. En otras palabras, en algunos casos una menor expresión de los transcritos de interés se asocia con condiciones favorables como ausencia de recurrencia, subtipo histológico CEE, bajo grado, estadio temprano e IM superficial, mientras que en otros casos sucede lo contrario. A partir del análisis de OR se determina que los genes que se subexpresan (es decir que para la población en estudio y de acuerdo a cómo se definieron las variables de respuesta, la menor expresión de estos transcritos se asocia a un tumor más agresivo) son: **ACSL5, ANAPC4, ARSD, FAM189A2, LPCAT2, PALMD, PGR, PLEKHH1, PTCH1, SLC25A35, SLC47A1, SORBS2, TMPRSS2** y **TRAF3IP2**. **TPX2** es el único gen que se sobreexpresa en todas las etapas de la cascada, es decir que la expresión aumentada del transcrito está asociada a características clínico-patológicas más agresivas.

Figura 3.13. Representación de los resultados del análisis de OR para la selección de 18 genes. Se evaluaron las características clínico-patológicas RFS, OS, subtipo ('Histología') y grado histológico, estadio e IM para cada gen. El esquema representa en **verde** aquellos genes con medidas de efecto significativas en todas las instancias. En **rojo** se presentan aquellos genes que no arrojaron resultados estadísticamente significativos para la primera característica comparada (GSPT1 y TAB2) y en **amarillo** aquel gen que presento resultados estadísticamente significativos para todos los parámetros clínico-patológicos excepto el último (PDZRN3). Se ilustra también en qué etapa de la cascada fueron descartados para los análisis subsiguientes.

1.6.2 Modelo de riesgos proporcionales de Cox

Por último, se realizaron dos análisis multivariados de regresión de Cox para identificar variables predictoras independientes de pronóstico en CE: uno para la variable de estado RFS y otro para OS. En ambos casos se partió de los niveles de expresión categorizados de los 15 genes resultantes del análisis de OR como variables independientes. Para definir las covariables de cada modelo se probó la **asunción de riesgos proporcionales**. Con este fin se compararon las curvas de sobrevida de expresión aumentada y disminuida de los genes; fueron descartados como covariables del modelo aquellos genes cuyas curvas se cruzan o no mantienen un comportamiento similar en el tiempo. De este modo, para el modelo de Cox con RFS todos los genes fueron considerados como covariables en el primer paso, mientras que para el modelo de Cox con OS se consideró que 11 de 15 genes cumplen la asunción. Las curvas de Kaplan-Meier de ANAPC4, ARSD, LPCAT2 y PLEKHH1 no garantizan que el riesgo de la exposición (ya sea aumentada o disminuida) sea constante en el tiempo (**Figura 3.14**), por lo que estos genes no fueron incluidos como covariables del modelo.

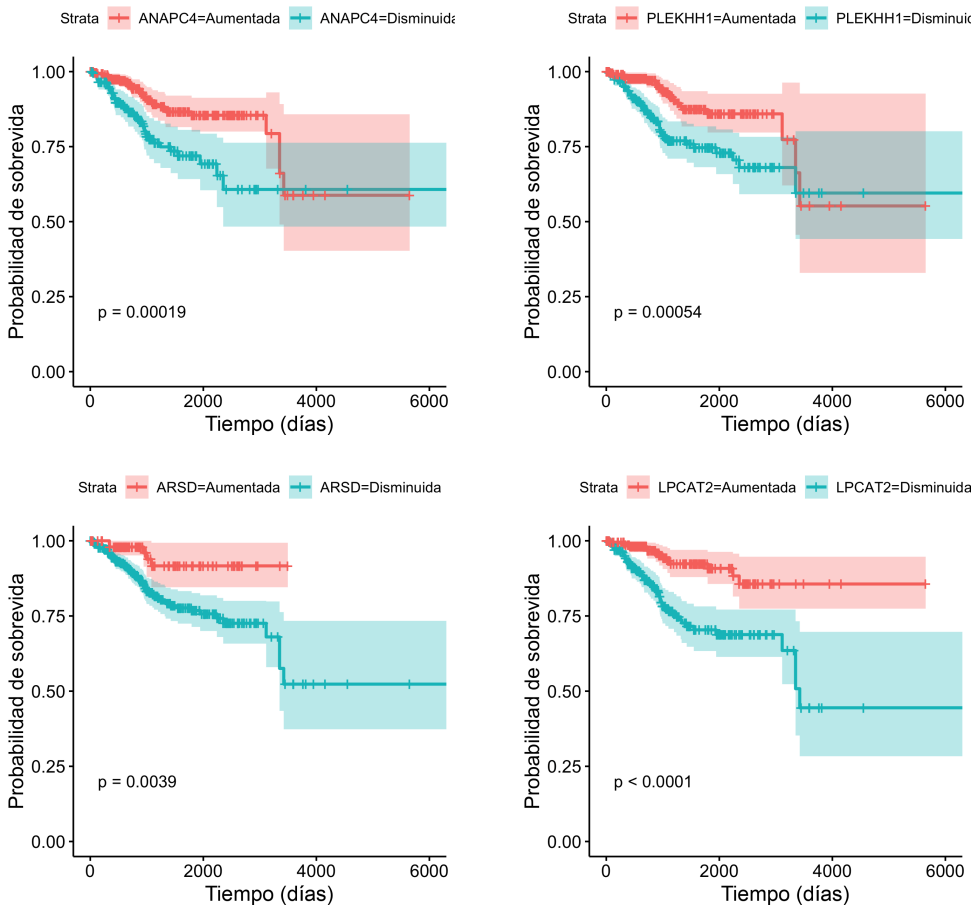


Figura 3.14. Curvas de sobrevida de los genes que no cumplen la hipótesis de riesgos proporcionales del modelo de Cox. Las curvas de expresión aumentada (**rosa**) y disminuida (**celeste**) de ANAPC4 (arriba izq.) y PLEKHH1 (arriba der.) se cruzan, por lo que la proporción entre ambas curvas invierte su signo. En el caso de ARSD (**abajo izq.**) y LPCAT2 (**abajo der.**) el par de curvas no mantiene un comportamiento similar en el tiempo, por lo que la proporción no es constante.

Se construyeron dos modelos de riesgos proporcionales de Cox con eliminación hacia atrás como describe la Sección 4.3.2, Materiales y Métodos. Como resultado se obtuvieron dos grupos de genes predictores a partir de varios pasos en cada modelo. Todos los pasos para ambas variables clínico-patológicas se exponen en el **Anexo H**.

Tras el análisis de riesgos proporcionales de Cox con la **variable de estado RFS** se obtuvieron los siguientes resultados:

- **TPX2**: b = 0,62, HR = 1,86, valor p = 0,015

- **PTCH1**: b = -0,6, HR = $0,55$, valor p = $0,013$
- **TMPRSS2**: b = -0,66, HR = $0,52$, valor p = $0,005$
- **SLC25A35**: b = -0,58, HR = $0,56$, valor p = $0,017$

Por otra parte, el análisis multivariado con la variable **estado de OS** arrojó los siguientes resultados:

- **TPX2**: b = 0,69, HR = $1,99$, valor p = $0,008$
- **SLC47A1**: b = -0,98, HR = $0,38$, valor p = $0,001$

A partir de estos resultados, se definió como genes candidatos a todos aquellos predictores de las variables pronósticas RFS y OS en CE. Por lo tanto, de la lista inicial de 962 genes candidatos asociados a CE, el análisis *in silico* realizado combinando diversas herramientas y criterios basados en el análisis de parámetros moleculares y clínico-patológicos condujo a la selección de un conjunto de **5 genes candidatos: PTCH1, SLC25A35, SLC47A1, TMPRSS2 y TPX2** como potenciales biomarcadores de CE cuya asociación con la enfermedad aún no ha sido reportada por la literatura.

1.6.3 Curvas de sobrevida de genes candidatos

Se ilustran en la **Figura 3.15** las curvas de sobrevida obtenidas para los genes resultantes del modelo de Cox con la variable pronóstica **estado de RFS**. Estas curvas fueron construidas en la Sección 1.5.1 del presente capítulo con el método de Kaplan-Meier a partir de la información de expresión génica dicotomizada y clínico-patológica del estudio TCGA-UCEC.

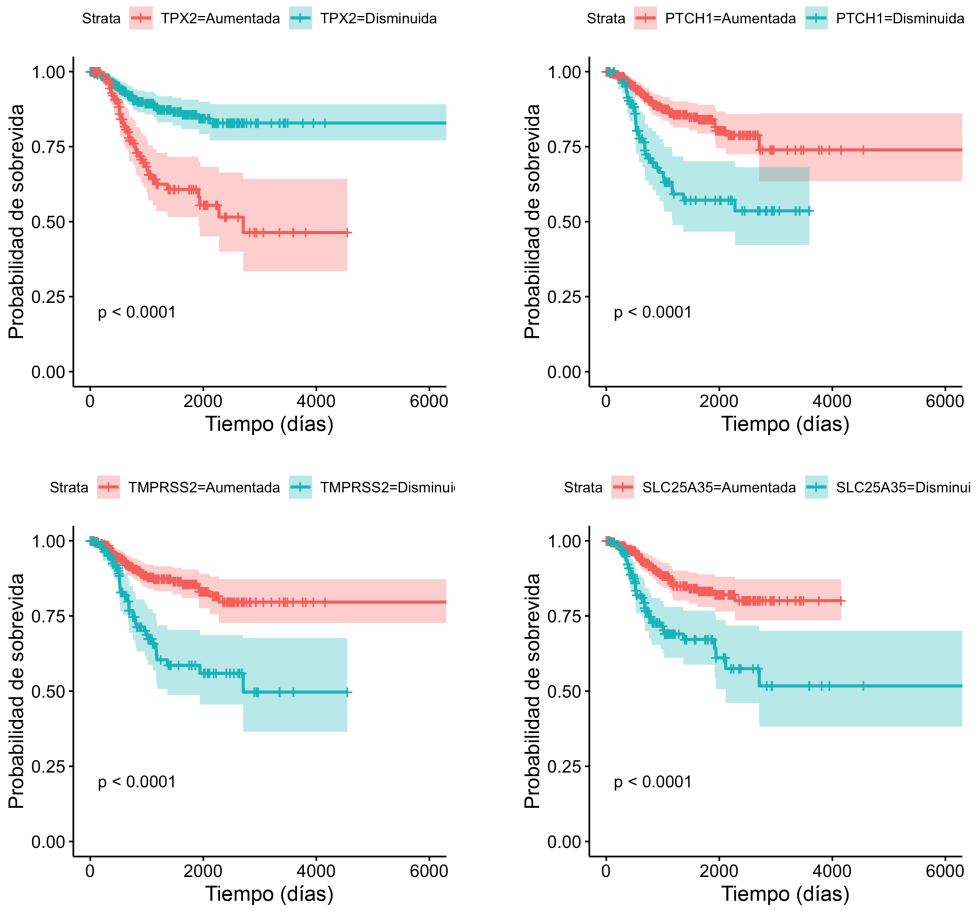


Figura 3.15. Curvas de sobrevida de los genes candidatos estimadas mediante el modelo de Kaplan-Meier para la variable pronóstica estado de RFS. Los genes representados son: TPX2, PTCH1, TMPRSS2 y SLC25A35. Para cada gen, la curva de sobrevida para la expresión aumentada (rosa) se compara con la de expresión disminuida (celeste).

La **Figura 3.16** muestra las curvas de sobrevida de los genes resultantes del modelo de riesgos proporcionales de Cox con el parámetro clínico estado de OS.

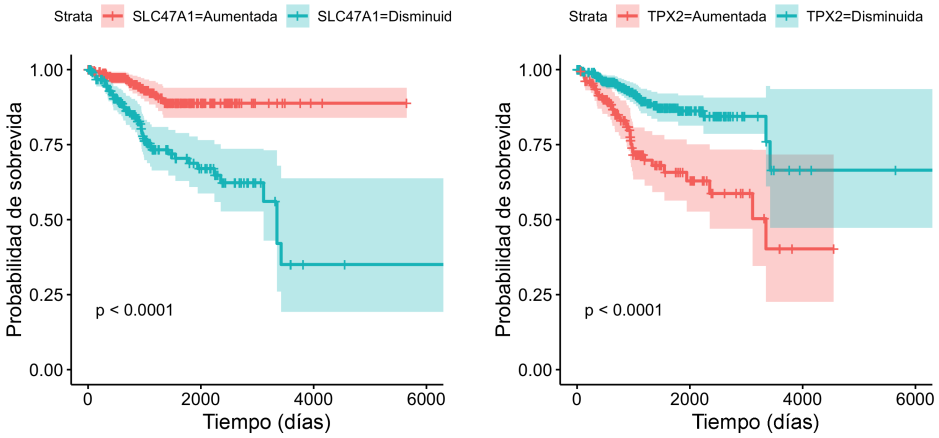


Figura 3.16. Curvas de sobrevida de los genes candidatos estimadas mediante el modelo de Kaplan-Meier para la variable pronóstica estado de OS. Los genes representados son: SLC47A1 y TPX2. Para cada gen, la curva de sobrevida para la expresión aumentada (rosa) se compara con la de expresión disminuida (celeste).

1.7 Rastreo de genes candidatos

La última instancia de evaluación teórica de los 5 genes seleccionados consistió en rastrear y resumir la información disponible de cada uno sobre su localización en el genoma, fenotipo, patrones de expresión del transcrito y la proteína, vías metabólicas relacionadas y asociación con cáncer.

A continuación se expone un resumen de la información identificada en HPA:

1. **PTCH1: *Patched 1***

   - Gen localizado en el cromosoma 9q22.32 humano.
   - Gen codificante de 15 isoformas de la proteína transmembrana *patched-1*.
   - Proteína expresada en todos los tejidos del cuerpo humano (**Figura 3.17**).
   - La proteína *patched-1* actúa como receptor de la proteína *Sonic Hedgehog* (SHh), que está implicada en el desarrollo temprano, la formación de estructuras embrionarias, el crecimiento y la especialización celular y la tumorigénesis. En ausencia del ligando SHh, *patched-1* previene el crecimiento y la proliferación celular. Todos los componentes de la vía *Hedhog* (Hh), incluidas SHh y *patched-1*, han sido propuestos como potenciales biomarcadores.

- Expresión elevada en tejidos normales de cuello y cuerpo uterino. No hay información sobre alteraciones en su expresión en tejidos con CE (**Figura 3.18**).

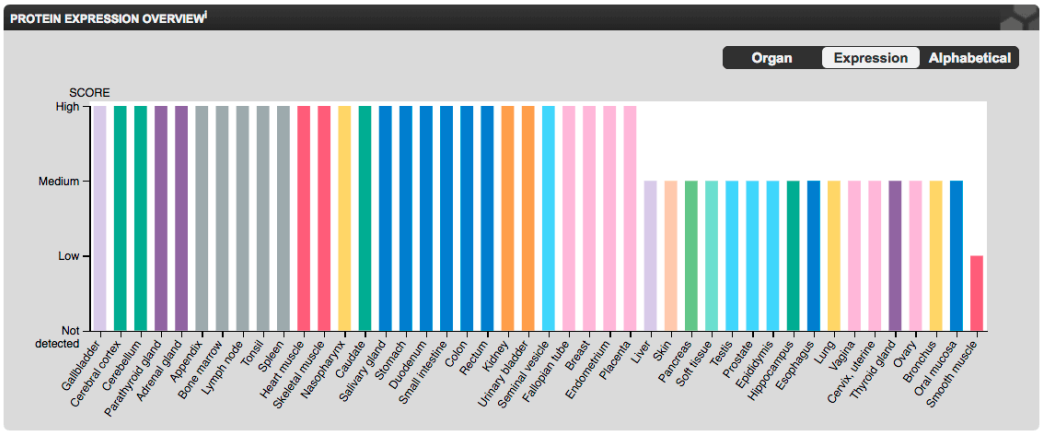


Figura 3.17. Expresión de la proteína *patched-1* en tejidos normales de 44 órganos del cuerpo humano. Fuente: *Human Protein Atlas*, 2019.

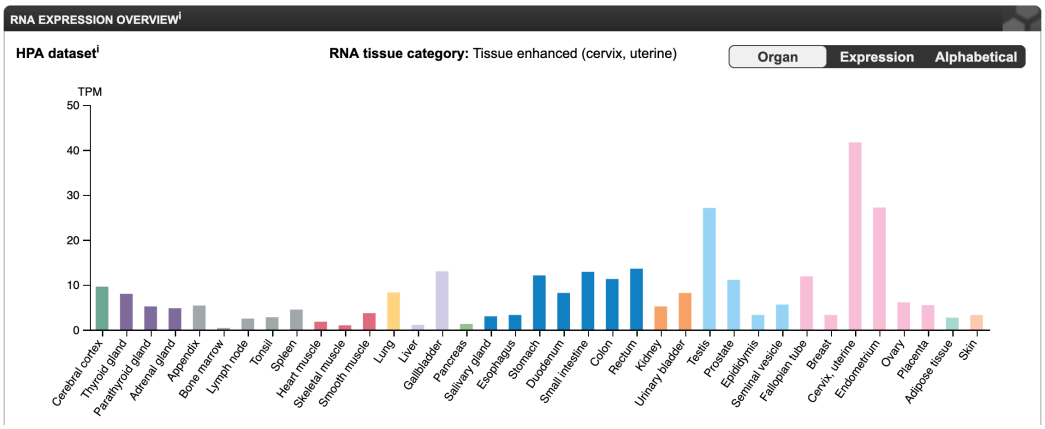


Figura 3.18. Expresión de PTCH1 (medida en transcritos por millón; TPM) para muestras de 44 tejidos normales del cuerpo humano. Fuente: *Human Protein Atlas*, 2019.

2. **SLC25A35:** ***Solute carrier family 25 member 35***

- Gen localizado en el cromosoma 17p13.1 humano.
- Gen codificante de una proteína *solute carrier*, que cataliza el transporte de grupos fosfato desde el citoplasma celular a la matriz mitocondrial.
- Gen con 5 variantes de *splicing* alternativo.

- Expresión del transcrito en todos los tejidos, elevada en tejidos femeninos, especialmente del útero y la placenta, y masculinos, principalmente de los testículos (**Figura 3.19**).
- **La expresión aumentada del transcrito de SLC25A35 es un marcador pronóstico favorable en cáncer renal, de vejiga y de endometrio**.

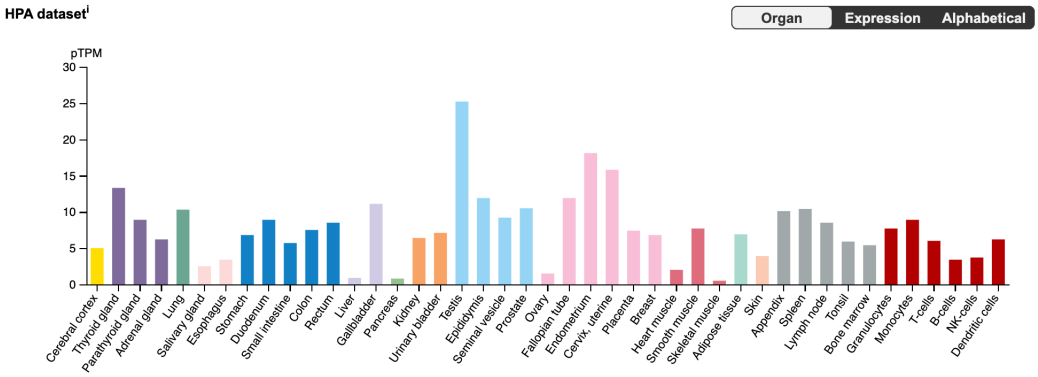


Figura 3.19. Expresión de SLC25A35 (medida en transcritos por millón; TPM) para muestras de 44 tejidos normales del cuerpo humano. Fuente: *Human Protein Atlas*, 2019.

Nota: Las anotaciones disponibles en HPA sobre expresión de la proteína no son concluyentes.

3. **SLC47A1:** ***Solute carrier family 47 member 1***

- Gen localizado en el cromosoma 17p11.2 humano.
- Gen codificante de una de las proteínas transmembrana que regulan el transporte de sustancias a través de la membrana celular, también conocido como *Multidrug and Toxin Extrusion Transporter* (MATE1), fue descubierto hace relativamente pocos años y ha cobrado gran relevancia. La familia de proteínas MATE tiene la función de excretar electrolitos tóxicos, tanto endógenos como exógenos, mediante la bilis y orina.
- Proteína expresada especialmente en el citoplasma y membranas celulares de tejidos de los túbulos renales y las glándulas adrenales (**Figura 3.20**).
- **SLC47A1 es un marcador pronóstico favorable en cáncer renal, de pulmón y de endometrio** y su expresión es superior a la media en tejidos normales de riñón, glándulas adrenales, cuerpo y cuello uterino y hepatocitos (**Figura 3.21**).

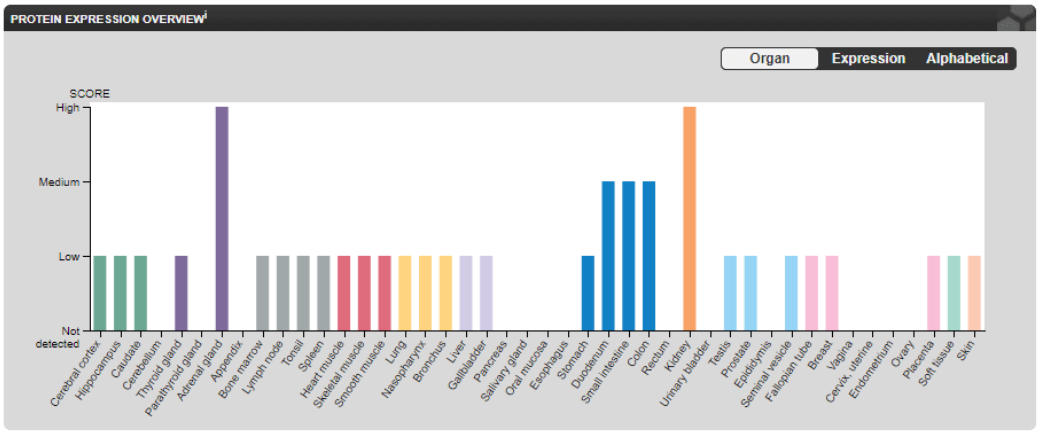


Figura 3.20. Expresión de la proteína codificada por SLC47A1 en tejidos normales de 44 órganos del cuerpo humano. Fuente: *Human Protein Atlas*, 2019.

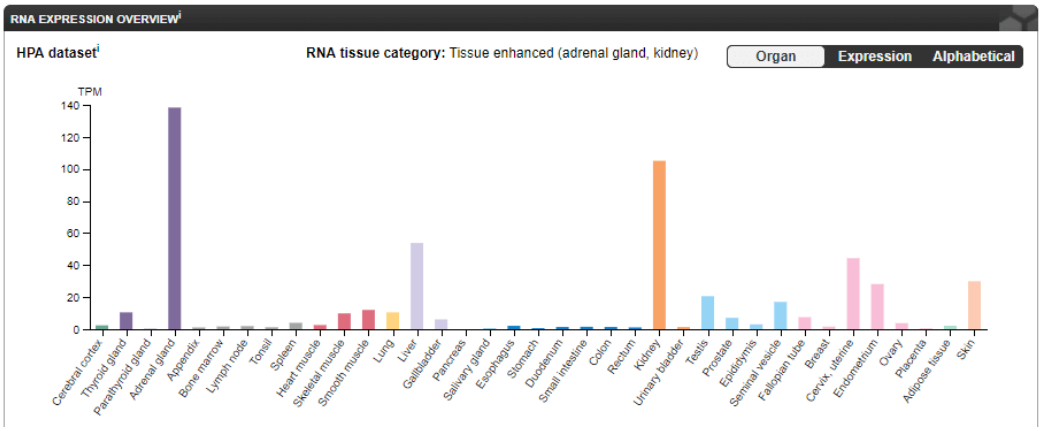


Figura 3.21. Expresión de SLC47A1 (medida en transcritos por millón; TPM) para muestras de 44 tejidos normales del cuerpo humano. Fuente: *Human Protein Atlas*, 2019.

4. **TMPRSS2:** ***Transmembrane protease, serine 2***

- Gen localizado en el cromosoma 21q22.3 humano.
- Gen con hasta 6 variantes de *splicing* (dos variantes anotadas, variantes 1 y 2).
- Gen codificante de una proteína de la familia de proteasas de serina, detectada en uniones intercelulares y, presuntamente, en el nucleoplasma. Se asocia a diversos procesos fisiológicos y patológicos del cuerpo humano.
- Proteína expresada en tejido normal de próstata, tracto gastrointestinal, riñón y páncreas, entre otros (**Figura 3.22**).

- Expresión del transcrito elevada en próstata y otros tejidos (**Figura 3.23**).
- **TMPRSS2 es un marcador pronóstico favorable tanto de cáncer renal como de endometrio.**

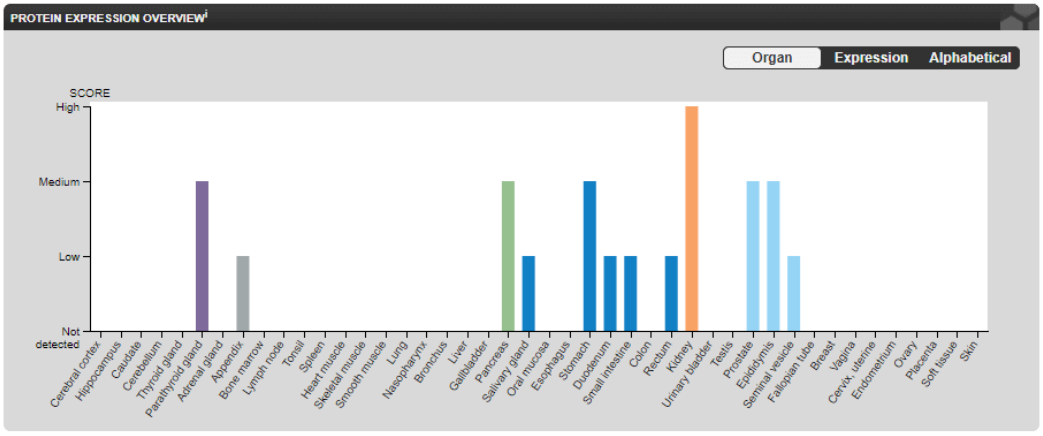


Figura 3.22. Expresión de la proteína codificada por TMPRSS2 en tejidos normales de 44 órganos del cuerpo humano. Fuente: *Human Protein Atlas* 2019.

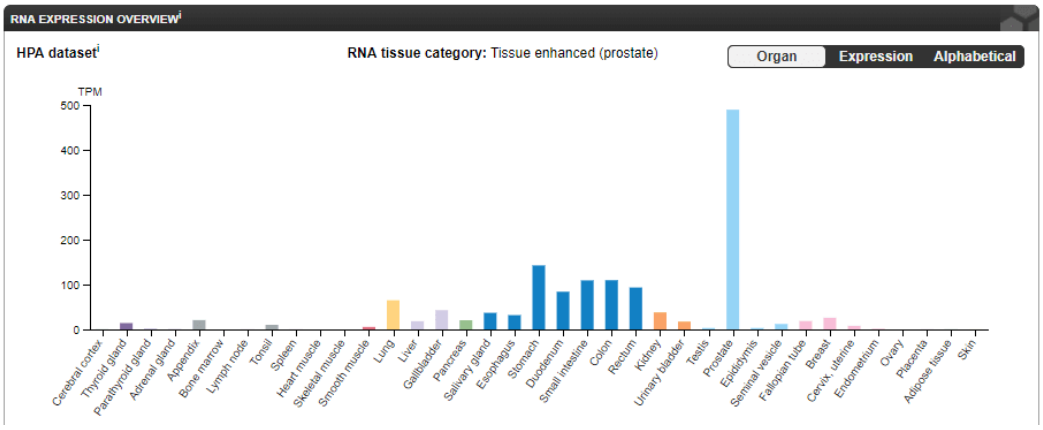


Figura 3.23. Expresión de TMPRSS2 (medida en transcritos por millón; TPM) para muestras de 44 tejidos normales del cuerpo humano. Fuente: *Human Protein Atlas* 2019.

5. **TPX2:** ***Targeting protein for Xklp2***

- Gen localizado en el cromosoma 20.q11.21 humano.
- Gen codificante de dos isoformas de la proteína *targeting protein for Xenopus kinesin-like protein 2*, expresada en la región perinuclear.

- Proteína requerida durante el ensamblaje del huso acromático en la fase de reproducción celular y también durante la reorganización de microtúbulos en el proceso de apoptosis.
- Expresión proteica elevada en tejido de testículos, estómago, piel y médula ósea (**Figura 3.24**).
- Expresión del transcrito elevada en tejido testicular (**Figura 3.25**).
- **TPX2 es un marcador pronóstico desfavorable en cáncer renal, hepático, pancreático y de endometrio**.

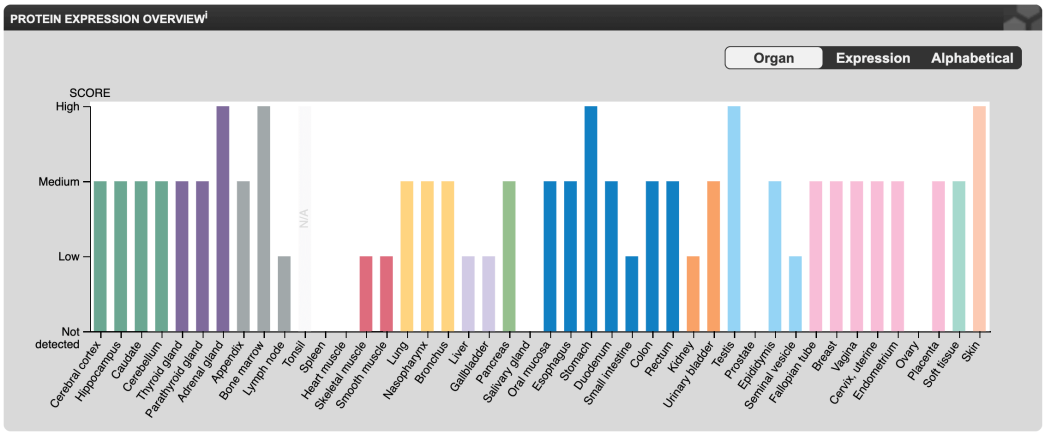


Figura 3.24. Expresión de la proteína codificada por TPX2 en tejidos normales de 44 órganos del cuerpo humano. Fuente: *Human Protein Atlas* 2019.

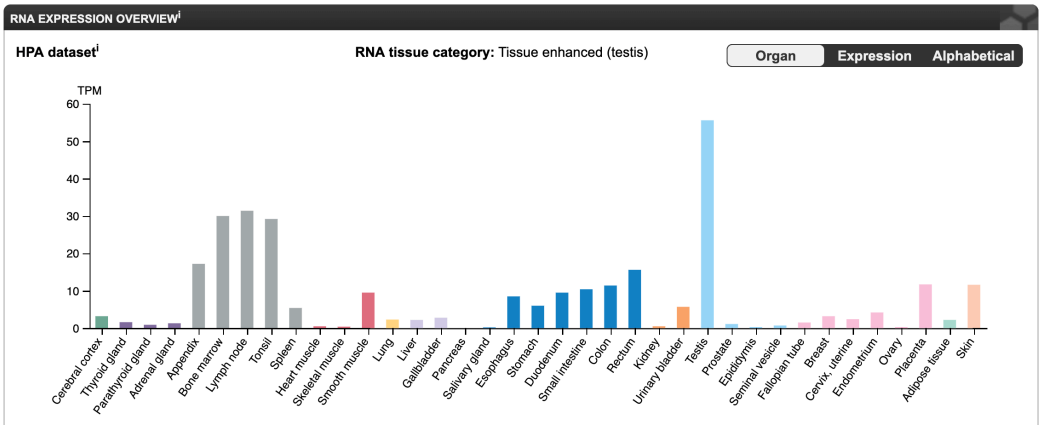


Figura 3.25. Expresión de TPX2 (medida en transcritos por millón; TPM) para muestras de 44 tejidos normales del cuerpo humano. Fuente: *Human Protein Atlas* 2019.

1.8 Selección de potenciales biomarcadores de CE

A partir de los resultados del análisis estadístico (Sección 1.6 del presente capítulo) se identificaron 5 genes como potenciales biomarcadores de CE. Estos son PTCH1, SLC25A35, SLC47A1, TMPRSS2 y TPX2. A continuación se recurrió a los resultados de la Sección 1.3, que proveen una lista de 37 genes ordenados según el análisis de priorización génica. La posición resultante de la priorización para los 5 genes candidatos se expone en la **Tabla 21**. Esta información fue utilizada para reducir la selección de genes: se decidió conservar a PTCH1, TMPRSS2 y TPX2 porque, al encontrarse entre las primeras cinco posiciones de la **Tabla 21**, estos genes tienen mayor similitud funcional con los reportados por la literatura (DisGeNET). Lo anterior fue validado con los resultados expuestos en la Sección 1.7 del presente capítulo: **PTCH1, TMPRSS2 y TPX2** se comportan como biomarcadores pronósticos en distintos tipos de cáncer y sus funciones se relacionan con vías de señalización tumorigénicas.

Los dos genes restantes, **SLC25A35 y SLC47A1**, se encuentran en el tercio inferior del *ranking* de priorización de la **Tabla 21** (posiciones 23 y 26, respectivamente), por lo que no fueron seleccionados para los estudios *in vitro*. Esta decisión fue validada con los resultados del rastreo de genes, que probaron que la información disponible sobre ambos genes es aún limitada, especialmente en el caso de SCL25A35.

Por lo anterior, **PTCH1, TMPRSS2 y TPX2** fueron los genes escogidos para continuar con los estudios *in vitro*.

***Ranking* de genes candidatos**	
Gen	**Posición Final**
PTCH1	2
TPX2	4
TMPRSS2	4
SLC25A35	23
SLC47A1	26

Tabla 21: *Ranking* de genes candidatos obtenido luego de la priorización con ToppGene (Sección 1.3, Resultados). Se muestran resaltados los genes seleccionados para el análisis *in vitro*.

Evaluación de potenciales biomarcadores de CE

2 Resultados experimentales

2.1 Descripción del modelo celular

Para los estudios experimentales se emplearon dos líneas celulares establecidas de CE de origen comercial, Hec-1a e Ishikawa. Ambas líneas fueron obtenidas a partir de adenocarcinoma de endometrio humano en estadios tempranos y con alto grado de diferenciación (ver **Anexo B**). Además, se utilizaron células de ambas líneas transfectadas de forma estable con el factor de transcripción ETV5 (células HGE e Ishikawa-ETV5, respectivamente). La expresión del factor ETV5 se ha encontrado aumentada en CE estadio IB así como en el frente invasivo de los tumores [18] y la sobreexpresión de ETV5 en líneas celulares de CE se ha asociado con la adquisición de un fenotipo celular agresivo, caracterizado por la expresión de marcadores moleculares típicos y capacidad migratoria e invasiva aumentadas [69].

Hec-1a e Ishikawa presentan morfología cuboide, típica de células epiteliales, mientras que las células transfectantes de ETV5 adquieren morfología de tipo fibroblástica (**Figura 3.26**). Por otro lado, las células HGE presentan una organización en forma de cúmulos celulares, con crecimiento en tres direcciones (hacia arriba y sobre el plano).

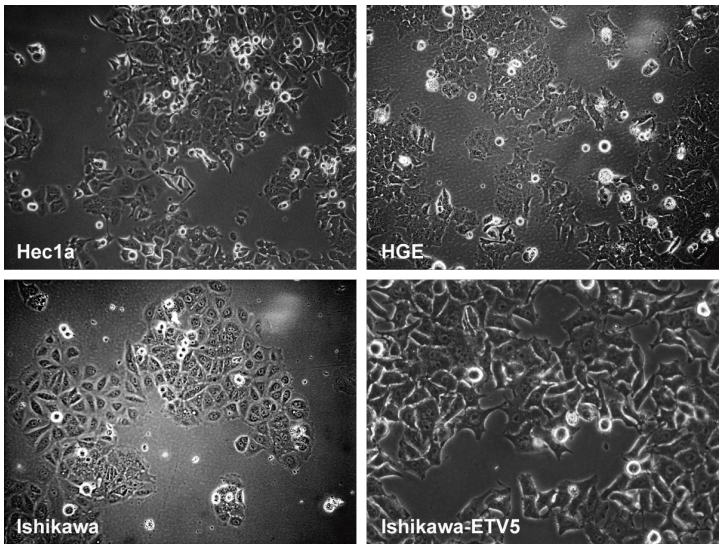


Figura 3.26. Líneas celulares del estudio *in vitro*. Imágenes en campo claro tomadas en un microscopio invertido con objetivo 20X. **Arriba**: líneas celulares Hec-1a (izq.) y HGE (der.). **Abajo**: líneas celulares Ishikawa (izq.) e Ishikawa ETV5 (der.).

2.2 Evaluación de la expresión de transcriptos

Como se explicó en la Sección 5.2.3 de Materiales y Métodos, se realizó la técnica **PCR a punto final** para evaluar la expresión de los transcriptos de interés identificados como resultado de los estudios bioinformáticos y estadísticos sobre muestras de ADNc obtenido mediante retrotranscripción de ARN de las líneas Hec-1a, Ishikawa, HGE e Ishikawa-ETV5.

Una vez completados los procedimientos de preparación de ARN y cuantificación, seguidas de retrotranscripción para la obtención de ADNc a ser usado como molde en los procedimientos de PCR, se realizó un protocolo de amplificación de un gen endógeno cuya expresión se ha demostrado que no varía con la sobreexpresión del factor de ETV5 y, por lo tanto, se puede tomar como trazador de los procedimientos de extracción de ARN y retrotranscripción. El gen seleccionado es **GAPDH**, que codifica para la gliceraldehído 3-fosfato deshidrogenasa, enzima de 37 kDa que cataliza el sexto paso de la glucólisis, como parte del mecanismo para descomponer la glucosa de las moléculas de energía y carbono. El fragmento esperado, según el diseño de los cebadores, es de 88 pb. Como resultado, el procedimiento de PCR condujo a la obtención de un fragmento del tamaño esperado y señal con intensidad comparable en las cuatro líneas celulares y ausente en el control (sin molde), confirmando de esta manera la especificidad de la señal detectada (**Figura 3.27**).

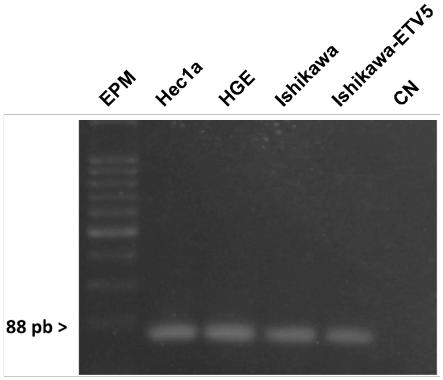


Figura 3.27. PCR a punto final del gen endógeno GAPDH. Se puede notar que el procedimiento de PCR condujo a la obtención de un fragmento del tamaño esperado, con señal comparable en las 4 líneas. A la derecha se incluye la calle correspondiente al control negativo.

A continuación se procedió a realizar sobre estas muestras de ADNc el**análisis de expresión** de los genes seleccionados: PTCH1, TMPRSS2 y TPX2. Teniendo en cuenta que la sobreexpresión de ETV5 se ha asociado a un peor pronóstico en CE y que las células HGE e Ishikawa-ETV5 presentan un comportamiento más agresivo que las células parentales Ishikawa y Hec-1a, se espera encontrar una menor expresión de PTCH1 y TMPRSS2 en las líneas menos agresivas respecto de las más agresivas y una mayor expresión de TPX2 en las líneas más agresivas respecto de las menos agresivas. Los resultados obtenidos al realizar PCR a punto final se ilustran en la **Figura 3.28 (A-C)**.

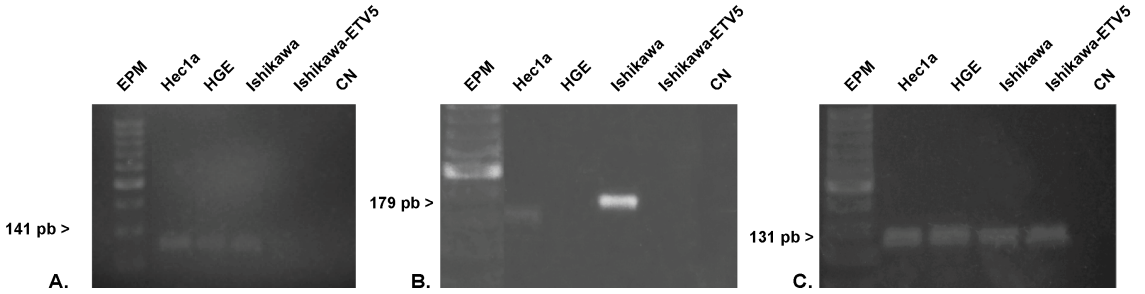


Figura 3.28. PCR a punto final de los genes A. PTCH1, B. TMPRSS2 y C. TPX2 con el ARN obtenido de las líneas celulares Hec-1a, HGE, Ishikawa e Ishikawa-ETV5. A la izquierda se incluyen los perfiles de los estándares de peso molecular (EPM), correspondientes a una escalera de 50 pb (se resalta el fragmento de 250 pb). A la derecha se ubica la calle correspondiente al control negativo de PCR.

En todos los casos se determinó la presencia de una amplificación específica del fragmento de ADN de interés con el protocolo de amplificación empleado, evidenciado por la ausencia de producto

en los controles de ensayo en los que se omitió el molde correspondiente al ADNc de las líneas celulares para todos los *sets* de cebadores. Asimismo, los productos amplificados presentaron tamaños moleculares esperados según el diseño de los cebadores (ver **Anexo C**). En cuanto a los resultados obtenidos para cada uno de los genes:

En el caso de **PTCH1**, el protocolo de amplificación condujo a la obtención de un producto con señal de baja intensidad para los productos de amplificación de ambas líneas parentales, detectándose una menor intensidad en el correspondiente a la línea HGE respecto de Hec-1a y señal indetectable en las células Ishikawa-ETV5. La menor o ausente expresión en las células con fenotipo más agresivo concuerda con los resultados esperados.

Los resultados de **TMPRSS2** revelan la presencia de un fragmento amplificado en ambas líneas celulares parentales, y ausencia de señal detectable en las transfectantes de ETV5 de las dos líneas de CE, en concordancia con lo esperado. La señal del fragmento es de mayor intensidad en la línea Ishikawa-ETV5 comparado con la de la línea HGE. Asimismo, el análisis de los resultados revela una diferencia en el tamaño esperado de los fragmentos correspondientes a ambas líneas celulares parentales. El tamaño molecular de los fragmentos no fue calculado, si bien la banda correspondiente al fragmento generado en la línea Hec-1a aparece más próximo al marcador de 150 pb, por lo que en principio presentaría un tamaño menor al esperado para el fragmento.

Por último, para **TPX2** se detectó expresión en las cuatro líneas celulares, con señales comparables. Para cuantificar los niveles de expresión del transcripto, se realizó PCR en tiempo real para este gen (**Figura 3.29**). En concordancia con los resultados esperados, las células HGE e Ishikawa-ETV5 mostraron mayores niveles de expresión del transcripto TPX2 en comparación a los detectados en las células Hec-1a e Ishikawa, respectivamente.

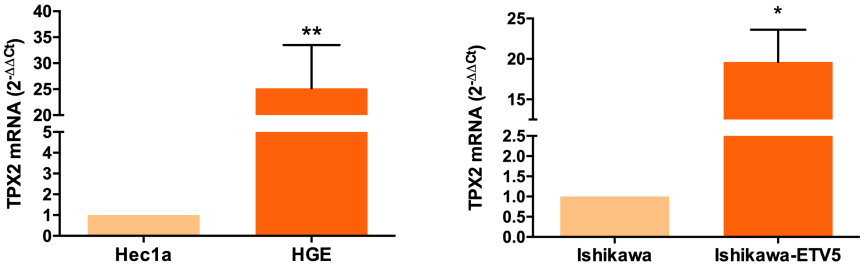


Figura 3.29. Expresión del transcripto TPX2 en las células Hec-1a, HGE, Ishikawa e Ishikawa-ETV5 evaluada con PCR en tiempo real. En ambos casos se observan los niveles de expresión relativa de ARNm de TPX2, mediante PCR en tiempo real, de acuerdo a la expresión $2^{-\Delta\Delta C_T}$ empleando GAPDH como gen endógeno. Los asteriscos determinan la significancia estadística de los resultados: * $p < 0,05$ *one sample t−test*, ** $p < 0,01$ *one sample t−test*

2.3 Evaluación de expresión de la proteína TPX2

Se procedió a caracterizar la expresión de la proteína TPX2 en ambos modelos celulares, empleando estrategias de detección de niveles de expresión de la(s) forma(s) proteica(s) por *Western immunoblotting* y su localización celular por inmunocitoquímica de fluorescencia. La elección de esta proteína para su caracterización se basó en los datos de HPA, que lo proponen como un factor desfavorable para CE y por la posibilidad de su seguimiento en tumores a través de la detección de un incremento en su expresión.

2.3.1 Electroforesis en geles de poliacrilamida y *Western immunoblotting*

Las células de las cuatro líneas celulares de CE fueron cultivadas en monocapas, cosechadas y procesadas para la extracción de proteínas totales, siguiendo el protocolo detallado en la Sección 5.3.2 de Materiales y Métodos. Una vez cuantificado el contenido proteico de cada uno de los extractos, los componentes de las mezclas fueron separados en matrices de poliacrilamida en el procedimiento de electroforesis. Las matrices fueron luego sometidas al procedimiento de electrotransferencia a membranas de nitrocelulosa e inmunodetección de TPX2 con un anticuerpo de origen comercial; el tamaño molecular aparente esperado para TPX2 es de 100 kDa. Como control de carga se empleó β-tubulina, con un tamaño molecular aparente esperado de 51 kDa. Los detalles del procedimiento de electrotransferencia fueron los presentados en la Sección 5.3.2 de Materiales y Métodos. Los resultados obtenidos al realizar esta técnica se presentan en la **Figura 3.30**.

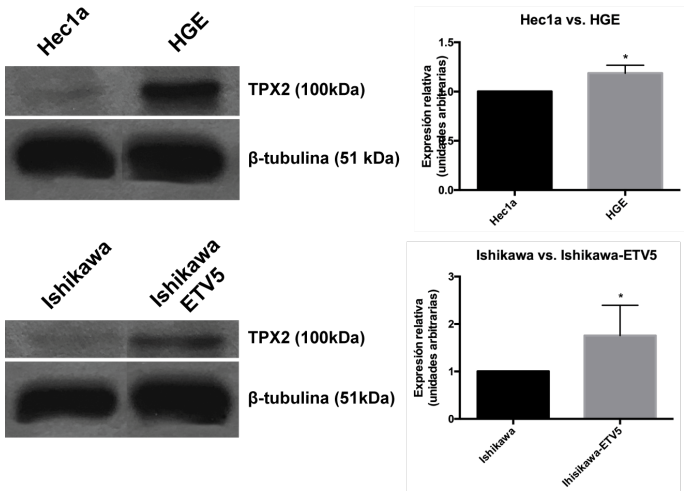


Figura 3.30. Evaluación de la expresión de TPX2 empleando electroforesis en geles de poliacrilamida y *Western immunoblotting*. **Izquierda**: Análisis por *Western immunoblotting* de la expresión de TPX2 en las líneas celulares Hec-1a, HGE, Ishikawa e Ishikawa-ETV5. Las membranas fueron teñidas con anti TPX2 (2 μg/ml); como control de carga se empleó β-tubulina (5 μg/ml). **Derecha**: Cuantificación de la expresión de TPX2 en las líneas celulares. Representación gráfica de los resultados del análisis densitométrico de la señal para TPX2 en las líneas celulares, normalizado con respecto al control de carga, empleando el *software* ImageJ. Se tomó como referencia la expresión de las líneas parentales. $^{*}p < 0,0001$ *one sample t−test*
.

Según se observa en las imágenes de la izquierda, en las cuatro líneas se detectó una señal intensa para el control de carga en el tamaño molecular esperado. En relación a TPX2, se observó una banda correspondiente al tamaño molecular esperado con una intensidad mayor en los extractos celulares de las líneas transfectantes de ETV5 (**Figura 3.30** (izq.)), en concordancia con los resultados del análisis de la expresión del transcripto en las líneas celulares (**Figura 3.29**). El análisis de cuantificación de la señal observada para TPX2 en las líneas parentales y transfectantes de ETV5 reveló un aumento en ambos casos en las células con fenotipo agresivo (**Figura 3.30** (der.)).

2.3.2 Inmunocitoquímica

Conjuntamente con el análisis de expresión de la proteína total TPX2 en los modelos celulares, se procedió a estudiar su expresión y localización en extendidos de células cultivadas en monocapas, fijadas y evaluadas empleando un protocolo de inmunocitoquímica de fluorescencia, según detalla la Sección 5.3.3 de Materiales y Métodos. Los núcleos celulares fueron visualizados empleando una

tinción con Hoescht 33342, un colorante fluorescente que se une al ADN.

Los resultados obtenidos al hacer la inmunocitoquímica de fluorescencia se muestran en la **Figura 3.31**.

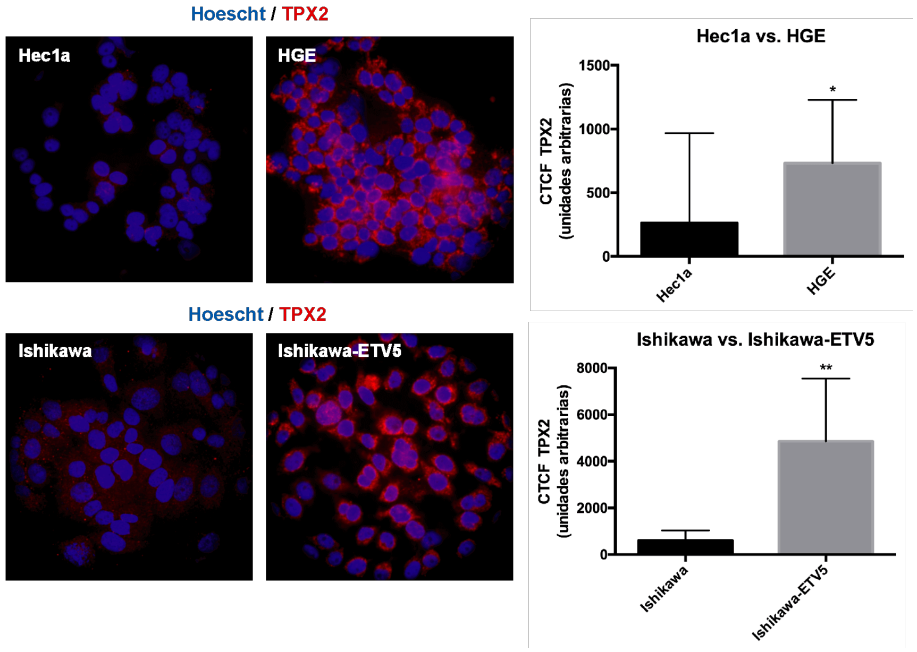


Figura 3.31. Evaluación de la expresión y localización de TPX2 mediante inmunocitoquímica de fluorescencia en las células Hec-1a, HGE, Ishikawa e Ishikawa-ETV5. **Izquierda**: Imágenes del ensayo de inmunocitoquímica de fluorescencia realizado utilizando el anticuerpo anti-TPX2 (2 μg/ml). Los núcleos celulares se visualizaron mediante tinción Hoechst 33342 (1 μg/ml). Las imágenes fueron registradas en un microscopio de fluorescencia con una magnificación 600X. **Derecha**: Cuantificación de la fluorescencia celular total, corregida empleando el *software* ImageJ. Se representa el *Corrected Total Cell Fluorescence* (CTCF), que se calculó empleando la siguiente fórmula: CTCF = Densidad Integrada (Área de la célula x fluorescencia promedio de las lecturas correspondientes al positivo) - (Área de la célula x fluorescencia promedio de las lecturas correspondientes al control negativo). $^{*}p < 0,01$ *one sample t−test*, $^{**}p < 0,001$ *one sample t−test*

A partir de la **Figura 3.31** se concluye que en las cuatro líneas celulares, Hec-1a, HGE, Ishikawa e Ishikawa-ETV5, se inmunodetectó la presencia de la proteína TPX2 mediante una señal fluorescente roja. Se puede observar que en las líneas Ishikawa-ETV5 y HGE se observa una señal de mayor intensidad respecto de las líneas Ishikawa y Hec-1a, en concordancia con los resultados del análisis de la expresión de proteína por *Western immunoblotting*. También se puede apreciar que en el caso de las líneas Ishikawa e Ishikawa-ETV5, la señal es más intensa que en el caso de Hec-1a y HGE. Se puede inferir una mayor expresión de la proteína en las líneas celulares más agresivas, lo que concuerda con la predicción inicial de que TPX2 está asociado a una sobreexpresión

en los tejidos tumorales más agresivos. Asimismo, un análisis de la localización subcelular de la proteína revelo una señal intensa perinuclear, esperada para esta proteína según sus características funcionales y la descripción de la hoja de producto del anticuerpo (**Figura 3.32**).

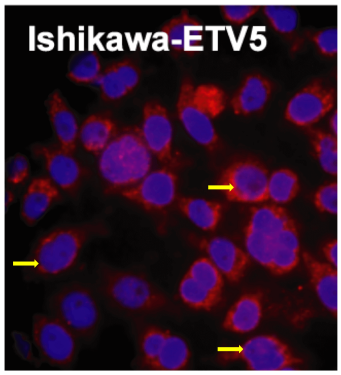


Figura 3.32. Localización del transcripto TPX2 mediante el ensayo de inmunocitoquímica. Se confirma la localización perinuclear del transcripto, tal y como indican las flechas amarillas.

Discusión

El CE es la segunda neoplasia ginecológica más frecuente en mujeres tanto a nivel mundial como nacional. Además, es la séptima causa de muerte por cáncer en mujeres a nivel mundial, la sexta en nuestro país y a futuro se estima un marcado incremento en la incidencia de esta enfermedad. Si bien aproximadamente el 80% de los casos de CE se diagnostica tempranamente, el diagnóstico preoperatorio, que combina técnicas de imágenes y análisis histopatológico de una biopsia endometrial, presenta limitaciones para una clasificación certera. Se han reportado discordancias en hasta un 40% de los casos entre el diagnóstico en la biopsia preoperatoria y el definitivo en la pieza quirúrgica, por sub-estimación en la biopsia de la invasión miometrial y grado histológico [134–137]. Por otra parte, alrededor del 20% de los casos diagnosticados clínicamente como estadio temprano (estadio I) son diagnosticados definitivamente en la cirugía como CE en estadio avanzado (estadios III y IV), cuando el tumor invade más del 50% del miometrio y puede diseminarse fuera de la cavidad uterina, lo que se refleja en una mayor tasa de recurrencia post-quirúrgica y una disminución significativa de la sobrevida de las pacientes. Además, alrededor del 13 al 25% de las pacientes con CE subtipo endometrioide (CEE) en estadio temprano presentan recurrencia y metástasis de la enfermedad, lo que se asocia a una disminución en la respuesta a la quimioterapia convencional y la sobrevida [138, 139].

Por todo lo expuesto, numerosos estudios se han enfocado en profundizar el estudio de esta patología a través del uso de diferentes estrategias de investigación tanto básica como clínica, con el interés de aportar nuevas herramientas para su diagnóstico y terapéutica. En un esfuerzo por comprender más exhaustivamente la enfermedad y su manejo, el consorcio TCGA ha reportado una caracterización genómica integral de una cohorte de aproximadamente 400 casos de pacientes

afectadas con CE, y que condujo a la identificación de cuatro subtipos moleculares de esta patología [12]. Sin embargo, a pesar de su relevancia, este sistema de estratificación presenta dificultades para ser aplicado en la práctica clínica, debido a la complejidad de las tecnologías empleadas y los costos elevados para su realización. Es por ello que muchos grupos de investigación se han enfocado en identificar biomarcadores característicos de cada grupo de la clasificación molecular para ser evaluados mediante el uso de inmunohistoquímica y así facilitar su implementación [140]. Por su parte, otros han abordado estudios moleculares para identificar en diversos muestreos de pacientes la presencia de genes con expresión diferencial según ciertos parámetros clínico-patológicos, a pesar de que en muchos casos los muestreos son de bajos números de pacientes (detección temprana: [141], CE recurrente: [142, 143]). Resulta imperativo continuar realizando investigaciones que contribuyan a identificar biomarcadores que puedan asistir en el manejo del CE para identificar pacientes de alto riesgo, aún en casos de CE en estadio temprano, y que estos biomarcadores puedan ser trasladados a la práctica clínica de rutina mediante un método estandarizado, reproducible y de bajo costo.

Considerando la relevancia que ha ganado la bioinformática en la investigación biomédica, en años recientes se han implementado diversos algoritmos que emplean abordajes bioinformáticos con el objetivo de identificar biomarcadores con valor pronóstico que puedan ser empleados en el manejo de diversos cánceres. En particular, el análisis de perfiles de expresión génica a partir de datos genómicos y clínicos almacenados en repositorios públicos ha permitido identificar y proponer biomarcadores con potencial valor pronóstico en diversos tumores como cáncer de próstata [144–146], de mama [147–149], y gástrico [150, 151], entre otros. Específicamente en lo que concierne al CE, en los últimos años se han reportado varios estudios que han empleado diversas estrategias bioinformáticas. De este modo se han presentado informes sobre el rastreo bioinformático para definir una firma mutacional para la detección temprana del CE [152], así como la identificación de regiones de riesgo genético de sufrir CE a partir del análisis bioinformático de estudios globales de secuenciación génica en poblaciones europeas [153]. En lo que respecta a análisis de expresión génica, varios estudios recientes han reportado el uso de abordajes de análisis bioinformático para identificar genes y vías de señalización [148, 154] y en particular factores de transcripción [155] cuya expresión es determinante en el CE. Otros estudios han comparado los resultados de los repositorios de datos de expresión de tumores de pacientes con parámetros histopatológicos característicos,

pudiendo identificar genes con expresión diferencial luego de comparar tumores de endometrio diferenciados y pobremente diferenciados [156] y tumores endometrioides y no endometrioides [157], entre otros.

En el presente estudio se empleó un conjunto de herramientas de minería de texto y de datos, en combinación con análisis estadísticos y evaluaciones en modelos celulares, a fin de identificar potenciales biomarcadores de agresividad del CE. El primer paso en la selección del presente estudio consistió en el análisis de expresión diferencial en el estudio GSE17025 del repositorio GEO del NCBI, una herramienta ampliamente empleada en más de 500 reportes listados en NCBI desde el año 2001. A partir de todos los genes evaluados y tres características histopatológicas altamente relevantes, como son la presencia (Tumoral vs No Tumoral), subtipo histológico (CEE vs CENE) y grado (G1-2 vs G3) del tumor, este análisis permitió identificar 39 potenciales biomarcadores diagnósticos de CE (**Figura 3.3**). Esta selección representó un avance significativo en el trabajo, a pesar de las limitaciones del estudio sobre el que se basó, en particular el número de muestras analizado, constituido por 103 muestras de tejido de endometrio, 91 de ellas tumorales y 12 de tejido de endometrio atrófico; si bien el endometrio atrófico no evidencia histopatología de CE, no es el control ideal para mujeres jóvenes cuyo endometrio está sometido a la regulación hormonal del tejido de una mujer que aún no ha ingresado en la menopausia. Dada la baja disponibilidad de muestras de endometrio sano, especialmente en pacientes jóvenes, en la literatura se utiliza habitualmente este tipo de muestras como control no tumoral. Una alternativa menos frecuente es analizar el segmento no tumoral de una muestra de CE; sin embargo, pocos estudios cuentan con este tipo de datos. Otra limitación encontrada fue la representatividad de los diferentes tipos de tumores en el muestreo utilizado. Específicamente, el muestreo corresponde a tumores en estadio I, el 87% de los tumores fue clasificado como CEE y el 75% de las muestras presenta IM menor al 50%. Si bien esto no representa todos los tipos de tumores, sí es representativo de la incidencia de la enfermedad que, como se indicara en la introducción, presenta una baja proporción de tumores no endometrioides y una alta proporción en estadios tempranos con baja proporción de muestras con IM mayor al 50%.

Al analizar la expresión diferencial de cada uno de los 39 genes en las tres características clínico-patológicas (**Tabla 11**), se identificaron cuatro genes ATAD2, HAUS8, TAB2 y TPX2 con una expresión mayor (upregulación) en las muestras con características más agresivas del cáncer (mues-

tras tumorales, CENE y G3) respecto de las menos agresivas (muestras no tumorales, CEE y G1-2, respectivamente). Para el resto de los genes en los que se encontró una expresión mayor en las muestras tumorales que en las no tumorales (CCDC160, CDC20B, HES6, SOAT1, TBCEL, TMEM132A y TMPRSS2), no se encontró una consistencia en los otros dos parámetros evaluados (CENE vs CEE; G3 vs G1-2). Se ha reportado en la literatura que los transcritos con este comportamiento tienen gran potencial como biomarcadores, dado que en la práctica suele ser más sencillo diseñar una estrategia de evaluación para detectar un incremento en la expresión que una disminución. Por su parte, el transcrito KIF7 mostró una expresión menor en las muestras tumorales respecto de las no tumorales, pero mayor en los otros dos parámetros asociados a agresividad, por lo que su comportamiento tampoco fue consistente. El resto de los transcritos presentaron una menor expresión en los tumores respecto de los controles y, en concordancia con este comportamiento, menor expresión en los tumores más agresivos (CENE) y de mayor grado.

El relevamiento de genes asociados a CE en la herramienta de minería de texto DisGeNET permitió identificar un total de 962 genes. A través del uso de los parámetros DPI y DSI, los genes con DPI alto (asociados a muchos términos MeSH) y DSI bajo (asociados a gran cantidad de enfermedades) fueron descartados del análisis, al filtrar la lista inicial con la media en cada caso; de esta forma aumentó la especificidad de los conjuntos de genes de entrenamiento ("training set") en el análisis de priorización. La combinación de los resultados de DisGeNET con la herramienta de priorización génica ToppGene permitió ponderar los potenciales candidatos obtenidos con GEO a través de las listas de genes previamente reportados en la literatura. En un principio esto brindó información descriptiva que luego sirvió para caracterizar la afinidad de los cinco genes predictores de pronóstico (resultado del análisis de sobrevida de Cox) con la agresividad de la enfermedad. Debe destacarse que dos genes, ATAD2 y KIAA1324, fueron excluidos del análisis porque se encontraron repetidos en la lista de prueba (proveniente de GEO) y las de entrenamiento (proveniente de DisGeNET). Por su parte, las tablas con términos GO asociados a los genes en estudio y el análisis de enriquecimiento funcional producto de la priorización (**Tablas 14-17**) incluyeron términos asociados a cáncer ("*MicroRNAs in cancer*"), como así también relacionados con la regulación hormonal del aparato reproductor ("*Steroid hormone biosynthesis*", "*Ovarian steroidogenesis*", resultados consistentes con la selección de genes para las listas de entrenamiento del análisis de priorización.

Tras reducir el universo de genes en estudio a un total de 39, descartar aquellos ya reportados en la literatura, priorizar los restantes a partir de su similitud funcional y biológica con genes asociados a CE y rastrear los 33 resultantes en un estudio de transcriptómica global, se procedió al análisis de la información de expresión de los genes en muestras de anatomía patológica de CE. Para el rastreo de genes se utilizó el estudio TCGA-UCEC de TCGA descargado desde la plataforma Xena. El estudio elegido cuenta con 580 registros de pacientes, de los cuales 66 debieron ser descartados del trabajo por no proporcionar información sobre una o más características clínico-patológicas de interés. Adicionalmente, TCGA-UCEC no proporciona información de expresión de cuatro de los 37 genes del listado resultante de GEO y ToppGene: DLGAP1-AS1, DLGAP1-AS2, LINC00261 y LOC100129098. Este hecho motivó su exclusión del trabajo, aunque no se profundizó en la(s) razón(es) por la(s) cual(es) estos genes no fueron listados en el estudio, quedando abierta la posibilidad de su estudio en caso de que sean incluidos en futuras actualizaciones del muestreo de resultados de TCGA-UCEC u otro repositorio. De este modo, el tamaño de la muestra del análisis bioestadístico se redujo a 514 registros. Pese a ser recomendable, para ninguno de los modelos se llevó a cabo un cálculo del tamaño muestral.

Para organizar la información, en esta instancia se recurrió a la construcción y análisis de una base de datos a partir de los datos proporcionados por TCGA. Esta estrategia permitió disponer de la información de manera completa y correctamente organizada y, por lo tanto, facilitó un manejo óptimo de los datos de acuerdo con las necesidades del trabajo. En este punto se optó por no analizar la información que excedía los objetivos de este trabajo. De este modo, se incluyeron en la base de datos construida solamente los niveles de expresión de los 33 genes finalmente seleccionados, la edad de las pacientes, su estado clínico (menopáusico), las variables pronósticas de recurrencia (RFS) y sobrevida total (OS) y las características clínico-patológicas relacionadas con la agresividad del tumor, según se expone en la Sección 2.4 de la Introducción. En este punto es importante destacar que no se contó con muestras control en el estudio, situación que representa una limitación en el análisis ya que no permite rastrear la asociación entre los genes y la presencia/ausencia de neoplasias, especialmente teniendo en cuenta que ésta fue una de las tres comparaciones efectuadas en el análisis de expresión diferencial de GEO. Por tanto, la expresión de los 33 genes seleccionados no pudo ser analizada en muestras no tumorales en los análisis estadísticos subsiguientes.

A partir de la base de datos construida se llevó a cabo una serie de análisis descriptivos de la

información recopilada. La distribución de la edad de las pacientes cuyos datos fueron incorporados en el estudio es normal (**Figura 3.6**), con valores correspondientes al promedio y desvío estándar siguiendo la tendencia mundial (Sección 2, Introducción) y los factores de riesgo del CE (Sección 2.2, Introducción). Asimismo, el estadio postmenopáusico y la anovulación se asocian a una mayor predisposición a desarrollar neoplasias de endometrio, y la muestra del estudio es representativa de este hecho (**Figura 3.7**). Por otra parte, los gráficos circulares de las **Figuras 3.8**, **3.9** y **3.10** revelan sesgos en el estudio TCGA-UCEC: en primer lugar, el subtipo histológico tiene un desbalance hacia el CEE, que es la característica menos agresiva; algo equivalente sucede con el estadio tumoral, observándose que el estadio I representa el doble de los estadios más avanzados combinados y, en particular, se observa una muy baja representación muestral del estadio IV. Estos datos reflejan la epidemiología clínica del CE, donde es más frecuente el diagnóstico de neoplasias en estadios tempranos y de tipo CEE. Lo anterior no impide que, a causa del desbalance, los resultados de este trabajo tengan una representatividad limitada para casos de CE con estadio y subtipo agresivos. Lo contrario sucede al examinar la distribución del grado histológico de las muestras, dado que más de la mitad de las muestras son de G3, el más agresivo.

A fin de acondicionar la información de la base de datos para los análisis bioestadísticos subsiguientes, se categorizaron las variables de agrupación, es decir, los niveles de expresión de los 33 genes en estudio. Esto responde a la potencial aplicación clínica de los resultados: en la práctica habitual del diagnóstico y pronóstico de cáncer, es útil disponer de un punto de corte para clasificar las muestras con diferentes niveles de expresión sin necesidad de informar los niveles absolutos de expresión de los ARNm. De esta forma, se definió un valor de corte para cada gen. La elección del punto de corte óptimo para la dicotomización de estas variables es un problema complejo sin una solución única. Las estrategias disponibles en la literatura incluyen la mediana de la expresión [158], que representa una aproximación simple y poco robusta al problema, la intersección entre dos curvas de distribuciones normales ajustadas a la expresión aumentada y disminuida [129, 158], que requiere biomarcadores con distribución bimodal, la optimización de la sensibilidad y especificidad y la regresión logística con prueba de Fischer [129, 158]. La solución con la prueba de *logRanks* implementada en este trabajo fue difundida por la herramienta Cutoff Finder y representa un método flexible y robusto para estas variables y aplicación clínica. Para preservar la validez del método se decidió no analizar los 15 genes cuya dicotomización arrojó un valor p no significativo.

Esta decisión implicó dejar fuera del análisis cerca de la mitad de los genes no en relación a su asociación con CE, sino simplemente por sus curvas de sobrevida con los datos de TCGA-UCEC. En futuros análisis podría tener sentido analizarlos con otras muestras clínico-patológicas. En la categorización se utilizó como variable de respuesta pronóstica a RFS, que representa la recurrencia o no del cáncer en el tiempo de seguimiento. Si bien la variable OS también constituye una opción técnicamente válida para indicar los eventos en la categorización, puede proveer información errónea para el análisis debido a que las muertes registradas de las pacientes no necesariamente están asociadas al cuadro clínico. En este punto cabe destacar que los puntos de corte óptimos de la **Tabla 20** no brindan información sobre la sobre o subexpresión de los transcritos, sino solamente con qué valor se diferencia mejor el evento de recurrencia de CE.

A partir de los resultados de la categorización se obtuvieron 18 genes, de los que se presentan las curvas de sobrevida por el método de Kaplan-Meier en el **Anexo F**. Éstas fueron estimadas tanto con RFS como con OS porque ambas variables de estado se utilizan en el análisis de riesgos proporcionales de Cox. Como resultado, la mayor parte de los genes tienen mayor probabilidad del evento con sobreexpresión, es decir, aquellos donde la exposición a la expresión aumentada del gen produce características más agresivas. GSPT1, TAB2 y TPX2 presentan la situación contraria. Estos resultados no son consistentes con los de expresión diferencial en el estudio GSE17025 de GEO, donde, solo TPX2 presentó expresión aumentada en los tumores más agresivos. Por otro lado, también resulta evidente a partir de las curvas que algunos cambian la proporción de riesgo entre muestras con y sin exposición en el tiempo, ya que sus curvas se cruzan en algún punto. Finalmente, las comparaciones de sobrevida de OS son más irregulares que las de RFS, un hecho que puede justificarse teniendo en cuenta que las muertes registradas en la variable sobrevida total no siempre se vinculan al CE.

El análisis de Odds Ratios se llevó a cabo de forma univariada para cada gen. Permitió cuantificar, con un IC del 95%, el efecto de la exposición a niveles de expresión elevados del gen en cuestión con respecto a la exposición a niveles de expresión bajos para una característica clínico-patológica. Como la proporción de pacientes con expresión aumentada y la de pacientes con expresión disminuida son comparables entre sí para todos los genes, la elección de OR respecto de RR es justificada; a medida que aumenta la incidencia o prevalencia del evento, la diferencia entre OR y RR es mayor. Adicionalmente, en la literatura se observa una tendencia creciente a

utilizar OR en forma general [159]. Los resultados arrojados por esta sección del análisis estadístico revelaron la ausencia de una asociación significativa entre la expresión de PDZRN3 y el estadio (categorizado en temprano -estadios I y II- y avanzados -estadios III y IV-). Similarmente, no se encontró una asociación significativa entre la expresión de GSPT1 y TAB2 y RFS y el **Anexo G** muestra los resultados para los 15 genes restantes; esto es, si la expresión aumentada implica un riesgo mayor o menor de ocurrencia de la característica más agresiva. En todos los genes, excepto por TPX2, ambos límites de todos los IC (RFS, OS, subtipo histológico, grado, estadio e IM) resultaron ser menores a 1. Por lo tanto, la expresión aumentada de cada uno de estos genes se asocia a un menor riesgo de recurrencia, muerte, CENE, G3, estadio avanzado e IM profunda respecto de aquellas muestras con expresión disminuida. Contrastando con estos hallazgos, los resultados de TPX2 revelan que el riesgo de neoplasias agresivas es mayor cuando la expresión del transcrito está aumentada para las muestras analizadas. A modo de ejemplo, según los OR de TPX2 con RFS, el riesgo de recurrencia es entre 2 y 6 veces mayor cuando la paciente presenta expresión aumentada de los transcritos de este gen. Este comportamiento es consistente con los resultados del análisis de expresión diferencial del estudio GSE17025 (GEO) y las curvas de sobrevida de Kaplan-Meier, donde se identificó a TPX2 como un gen con expresión aumentada en los tumores. En GEO también se identificó a TMPRSS2 como un gen con expresión aumentada en algunas características clínico-patológicas y disminuida en otras. Sin embargo, con los resultados de OR del **Anexo G** se comprueba que para el punto de corte definido mediante *logRanks*, la expresión aumentada de transcritos de TMPRSS2 tiene un efecto de reducción en la agresividad tumoral para todas las características clínicas.

Para el análisis de riesgos proporcionales de Cox se optó por desarrollar dos modelos en paralelo para analizar tanto la RFS como la OS. En cada caso se construyó un modelo multivariado con eliminación hacia atrás. Un paso fundamental en el análisis fue la evaluación de la hipótesis de riesgos proporcionales para cada covariable mediante las curvas de sobrevida de Kaplan-Meier expuestas en el **Anexo F**. En este paso cuatro genes fueron descartados, previniendo la aplicación del modelo incorrectamente. En cada modelo se realizaron varias iteraciones con eliminación hacia atrás, cada una con menos covariables que la anterior hasta alcanzar un conjunto reducido de covariables con coeficientes estadísticamente significativos. Si bien es utilizada habitualmente, la mayor desventaja de esta aproximación al problema es la posibilidad de pasar por alto genes

importantes para el modelo por la necesidad de eliminar necesariamente uno en cada paso. En este contexto, se eligió un valor de significancia estadística alfa de 0,05 pese a no estar ajustado por el testeo múltiple y, por lo tanto, acarrear el riesgo de incurrir en un mayor error alfa. Para el caso de la variable de estado RFS, cuatro genes (PTCH1, SLC25A35, TMPRSS2 y TPX2) resultaron en el modelo final. Para OS, los genes fueron dos: SLC47A1 y TPX2. Si bien el valor alfa se fijó en 0,05 para el análisis, cabe destacar que los valores p de todos los genes se aproximan a 0,01, con un máximo de 0,017. Esto podría atribuirse a la eliminación hacia atrás, que tiende a favorecer los valores p resultantes.

El rastreo de los genes seleccionados en el repositorio HPA reveló diferencias entre los genes respecto de las funciones celulares en las que intervienen, sus niveles de expresión proteica y de transcritos en condiciones normales y patológicas, y el volumen y la calidad de la información disponible al respecto. La asociación entre la proteína PTCH1 y la vía de señalización implicada en el desarrollo tumorigénico Hh indica gran potencial de PTCH1 como biomarcador diagnóstico de tumores. Asimismo, la expresión de transcritos de ARN elevada en tejidos como el de cuello uterino, endometrio y testículos plantea una relación con la función reproductiva. En esta línea, la literatura sugiere que un incremento en la actividad de la vía Hh se asocia a un peor pronóstico en cáncer de próstata [160] y de ovario [161]. Adicionalmente, HPA y el repositorio PubMed no contienen información que relacione la expresión de PTCH1 con tejidos tumorales de CE, lo que significa que, de ser un marcador de esta enfermedad, sería un hallazgo novedoso. SLC25A35 presentó, al igual que PTCH1, expresión elevada de ARN en tejidos de órganos reproductores y HPA reporta que la expresión aumentada de este gen es un marcador favorable de CE de acuerdo con las muestras del repositorio. Sin embargo, no hay reportes en la literatura que respalden el hallazgo o relacionen a SLC25A35 con algún tipo de cáncer. Adicionalmente, las anotaciones de HPA sobre expresión de la proteína no son concluyentes y, dado que encontrar expresión de transcritos no asegura que una proteína se exprese, es importante estudiar también la expresión de dicha proteína. Lo anterior convierte a SCL25A35 en un gen aún poco estudiado y una elección arriesgada para los estudios experimentales de este trabajo. Por su parte, la proteína SLC47A1 pertenece a la familia excretora de fármacos y toxinas MATE, resultado coherente con la expresión elevada tanto de la proteína como del transcrito en el tejido renal. A diferencia de SLC25A35, la relevancia de SLC47A1 ha aumentado notablemente en los últimos años y hay publicaciones recientes que

estudian su asociación con cáncer renal [162]. Asimismo, a partir de las muestras estudiadas, HPA define a SLC47A1 como un marcador pronóstico favorable en distintos tipo de cáncer, entre ellos el de endometrio, lo que lo convierte en un buen candidato para el estudio experimental. Si bien la función metabólica de TMPRSS2 aún presenta interrogantes, HPA plantea evidencia que asocia a la expresión diferencial del gen con tumores renales y de endometrio. También, la literatura sugiere que en neoplasias de próstata, TMPRSS2 tiene expresión diferencial: se sobreexpresa en células prostáticas con hormonas androgénicas y se subexpresa en tejido prostático no androgénico [163]. Por lo anterior, como sucede con PTCH1, TMPRSS2 tiene potencial como un biomarcador novel de CE. Finalmente, TPX2 codifica una proteína homónima relacionada con los microtúbulos del huso acromático en el proceso de mitosis celular. Varios estudios recientes revelan sobreexpresión de TPX2 en distintos tipos de cáncer, entre ellos cervical, de mama y próstata, y sugieren la asociación del gen con agresividad neoplásica aumentada y progresión tumoral [164–167]. Estos antecedentes plantean la relevancia de TPX2 como potencial biomarcador tumoral, particularmente en neoplasias del aparato reproductor femenino y masculino. Paralelamente, los análisis de las muestras del repositorio HPA arrojan resultados concluyentes para considerar a TPX2 como un marcador pronóstico desfavorable en CE. Lo anterior justifica la realización de estudios experimentales de expresión de transcritos y proteica de TPX2.

La selección de potenciales biomarcadores de CE se basó tanto en los resultados de la priorización génica con ToppGene como en el rastreo de genes en HPA. En primer lugar, la elección de PTCH1, TPX2 y TMPRSS2 fue justificada principalmente por el ranking de la priorización, dada la gran diferencia de posiciones entre este conjunto de genes y los restantes. La información disponible en HPA sobre SLC47A1 dejó en evidencia la asociación de este gen con el cáncer renal, pero no existen trabajos que respalden su potencial como biomarcador de otras neoplasias. La expresión de transcritos de SLC25A35 es elevada en órganos reproductores pero HPA no contiene información sobre la expresión de la proteína SLC25A35 y tampoco hay publicaciones científicas que estudien la expresión de este gen o la proteína en relación a ninguna patología. Si bien los resultados arrojados por el análisis de las muestras de HPA para estos genes constituyeron un primer paso para el descubrimiento de nuevas aplicaciones clínicas de SLC47A1 y SLC25A35 en CE, en el marco del presente trabajo no fueron suficientes para justificar su inclusión en el análisis in vitro.

Por todo lo anterior, los genes en estudio fueron reducidos desde un conjunto inicial de 39 genes

hasta la cantidad de tres: PTCH1, TMPRSS2 y TPX2. Es pertinente notar que los tres genes formaron parte del modelo final de riesgos proporcionales de Cox con RFS. TPX2 fue, además, parte del modelo de Cox de OS. En un principio, esto indicaría que TPX2 es un potencial predictor independiente de recurrencia y muerte en CE.

Los resultados de los estudios *in vitro* de TMPRSS2 revelan la presencia de un fragmento amplificado en ambas líneas celulares parentales y una ausencia de señal detectable en las transfectantes de ETV5 de las dos líneas de CE, en concordancia con lo esperado. La señal del fragmento es de mayor intensidad en la línea Ishikawa-ETV5 comparado con la de la línea HGE. Asimismo, el análisis de los resultados revela una diferencia en el tamaño esperado de los fragmentos correspondientes a ambas líneas celulares parentales. El tamaño molecular de los fragmentos de PCR identificado no fue calculado, si bien la banda correspondiente al fragmento generado en la línea Hec-1a aparece más próximo al marcador de 150 pb, por lo que en principio presentaría un tamaño menor al esperado para el fragmento. Se han descripto al menos dos variantes de *splicing* para TMPRSS2, con lo que la diferencia se podría atribuir al resultado de la detección de una de estas variantes reportadas en la línea Hec-1a y la otra en la línea celular Ishikawa, dado que los cebadores fueron diseñados en un segmento del transcrito común a ambas variantes (**Anexo C**). La información correspondiente a las variantes ya caracterizadas muestra que las deleciones/inserciones y consecuentes cambios en el marco de lectura ocurren río arriba de la región elegida para el diseño de los cebadores (`https://www.ensembl.org/Homo_sapiens/Transcript/Exons?db=core;g=ENSG00000184012;r=21:41464305-41508158;t=ENST00000332149`). La presencia de una deleción en el transcrito de TMPRSS2 de las células Hec-1a en la secuencia comprendida por los cebadores, o la ocurrencia de un evento de *splicing* alternativo novel asociado al tumor del que derivó la línea podría explicar la diferencia de tamaño de ambos amplicones; en función del diseño de los cebadores (exones 12/13 y 14), se estima que la deleción podría ocurrir en el exón 13 para justificar el menor tamaño, comprometiendo o no en parte la secuencia de los cebadores. Al respecto de la presencia de deleciones, en la base de mutaciones somáticas de COSMIC se reporta la evaluación de 7 muestras de adenocarcinoma de endometrio, en las que no se identifican mutaciones para este gen (`https://cancer.sanger.ac.uk/cosmic/browse/tissue?wgs=off&sn=endometrium&ss=NS&hn=carcinoma&sh=adenocarcinoma&in=t&src=tissue&all_data=n`)
y 930 muestras de tumores en los que no se detectan mutaciones en este gen (`https:`

//cancer.sanger.ac.uk/cosmic/browse/tissue?wgs=off&sn=endometrium&ss=all&hn=all&sh=all&in=t&src=tissue&all_data=n). El análisis de secuenciación del amplicón de ambas líneas celulares permitirá determinar las causales de las diferencias en el tamaño de los fragmentos encontradas.

De los 3 genes seleccionados, el gen TPX2 fue caracterizado además a nivel de proteína, por encontrarse sobreexpresado en los modelos más agresivos. TPX2 (*targeting protein for Xklp2*) es un factor clave para el ensamblaje del huso mitótico y para el ensamblaje normal de microtúbulos durante la apoptosis. La expresión aumentada de TPX2 se ha asociado a la progresión de diferentes cánceres, entre ellos: carcinoma esofágico de células escamosas [168], cáncer de vejiga [169], cáncer de cuello uterino [170] y carcinoma hepatocelular [171], cáncer de próstata [165], colangiocarcinoma [172] y cáncer gástrico [173]. Asimismo, los resultados del presente trabajo se encuentran en línea con lo reportado recientemente por otros autores en relación a la expresión de TPX2 en CE: mientras que dos estudios bioinformáticos han identificado a TPX2 como un gen relacionado al CE [154, 173], otro estudio ha encontrado a TPX2 como uno de los genes con expresión significativamente aumentada en muestras de CE respecto de muestras de endometrio control, y lo proponen como un indicador de pronóstico pobre [174]. Asimismo, este último estudio propone a TPX2 como un blanco de miR-29a-5p, mecanismo que regularía la proliferación e invasión celular y la apoptosis en CE. Se ha demostrado que TPX2 activa a la Aurora Quinasa A durante la mitosis y dirige su actividad al huso mitótico, desempeñando un rol importante en la mitosis; en años recientes además se ha demostrado que ambas proteínas conformarían una unidad funcional con propiedades oncogénicas [175]. En línea con estos resultados, otro estudio identifica al gen de la Aurora Quinasa A como uno de los 13 genes centrales principales diferencialmente expresados en y determinantes de la carcinogénesis en tumores de endometrio pobremente diferenciados [157].

En resumen, la combinación de estrategias bioinformáticas, bioestadísticas y estudios celulares, bioquímicos y moleculares han permitido identificar tres posibles nuevos biomarcadores de la agresividad del CE. En estudios futuros, se podrán diseñar estrategias para evaluar la expresión de los tres transcriptos sobre muestras frescas de tejido empleando PCR en tiempo real. Estudios de validación posteriores en muestreos provenientes de diversos centros del país y, eventualmente, de otras comunidades, contribuirán a la validación final de las moléculas identificadas como biomarcadores del CE y la factibilidad de su incorporación en la práctica clínica de rutina. Ésta se implementará

a través de ensayos sencillos, reproducibles y de bajo costo que permitan el alcance a todas las pacientes que puedan beneficiarse de su uso, para contribuir a un mejor manejo de la patología. Teniendo en cuenta que para el estudio de biomarcadores proteicos asociados a la progresión y agresividad del cáncer se prefiere la identificación de moléculas cuya expresión aumenta con la malignidad frente a aquellas cuya expresión disminuye pues facilita los abordajes de detección (por ejemplo, en inmunohistoquímica), se podrán realizar estudios de expresión de la proteína TPX2 en muestreos de pacientes con CE. Asimismo, se podrán diseñar estudios futuros para la identificación de firmas moleculares a partir de los tres potenciales biomarcadores descriptos, en conjunto con otros genes ya reportados y/o descartados en este trabajo.

Conclusiones

- El cáncer de endometrio (CE) es la segunda neoplasia ginecológica más frecuente y la séptima causa de muerte por cáncer en mujeres tanto a nivel mundial como nacional.
- En los próximos 20 años se prevén aumentos mayores al 50% en las tasas de incidencia y mortalidad de esta enfermedad.
- Actualmente la clínica no cuenta con **biomarcadores** moleculares establecidos del CE.
- Se desarrolló un proyecto para identificar genes con potencial como biomarcadores de la progresión y agresividad del CE.
- Se utilizó un algoritmo basado en una combinación de **herramientas bioinformáticas** de minería de texto, datos y priorización génica, así como análisis bioestadísticos y estudios bioquímicos y moleculares.

Como resultado,

- el análisis de resultados de un estudio de transcriptómica con microarreglos de ADN (GSE 17025; plataforma GEO) de muestras de pacientes con CE permitió identificar genes diferencialmente expresados, a partir de los que se seleccionaron 39 con expresión diferencial para tres parámetros del CE asociados a su agresividad: la expresión en el tumor respecto del tejido no tumoral, la expresión en tumores de endometrio no endometrioides respecto de endometrioides y la expresión en tumores de alto grado (G3) respecto de bajo grado (G1-2);
- una búsqueda de genes previamente reportados como asociados a CE, utilizando la base de datos DisGeNET, permitió identificar 962 genes asociados a la enfermedad. Con esta

misma herramienta se determinaron para cada gen sus índices de especificidad y cantidad de enfermedades a la que están relacionados (DSI y DPI, respectivamente);

- un análisis de priorización génica (ToppGene) empleando la lista de genes seleccionados en el análisis de transcriptómica GEO como "lista de prueba" y los genes de DisGeNET como "lista de entrenamiento" permitió seleccionar 33 genes.

- Evaluaciones posteriores empleando modelos estadísticos (Método de *Kaplan-Meier*, *Odds Ratio* y Modelo de riesgos proporcionales de Cox), permitieron seleccionar 20, 16 y 6 genes, respectivamente. Tres de los 6 genes fueron finalmente seleccionados sobre la base de la priorización de ToppGene y los datos de Human Protein Atlas (HPA): **PTCH1**, **TMPRSS2** y **TPX2**;

- la expresión de los transcriptos de PTCH1, TMPRSS2 y TPX2 fue evaluada por PCR estándar en dos líneas celulares de CE (Hec-1a e Ishikawa), eligiéndose un modelo de sobreexpresión de un factor de transcripción ETV5 (con expresión aumentada en el frente tumoral del CE; Hec-1a-ETV5 ó HGE, e Ishikawa-ETV5) y sus controles (Hec-1a e Ishikawa). Se encontró expresión menor o indetectable para PTCH1 y TMPRSS2 en las líneas celulares HGE e Ishikawa-ETV5, ambas de fenotipo más agresivo, en concordancia con los resultados de los estudios bioinformáticos. El transcripto TPX2 fue detectado en las 4 líneas celulares, y su expresión aumentada en las líneas HGE e Ishikawa-ETV5 fue detectada en protocolos de PCR cuantitativa en tiempo real;

- en estudios futuros, se podrán diseñar estrategias para evaluar sobre muestras frescas de tejido de pacientes con CE la expresión de los transcriptos PTCH1, TMPRSS2 y TPX2 empleando PCR en tiempo real. Estos estudios contribuirán a su validación final como biomarcadores de agresividad de CE.;

- teniendo en cuenta que para el estudio de biomarcadores proteicos asociados a la progresión y agresividad del cáncer se prefiere la identificación de moléculas cuya expresión aumenta con la malignidad frente a aquellas cuya expresión disminuye, pues facilita los abordajes de detección (por ejemplo, en inmunohistoquímica), se podrán realizar estudios de expresión de la proteína TPX2 en muestreos de pacientes con CE.

- estos hallazgos permitirán, en estudios futuros, analizar potenciales firmas moleculares a partir de PTCH1, TMPRSS2 y TPX2 en conjunto con otros genes ya reportados y/o descartados en este trabajo.

En conclusión, la combinación de estrategias bioinformáticas y bioestadísticas con estudios celulares, bioquímicos y moleculares ha contribuido al avance en la búsqueda de biomarcadores moleculares de agresividad y pronóstico del CE.

Abreviaturas

A Adenina.

ADN Ácido Desoxirribonucleico.

ADNc ADN complementario.

ARN Ácido Ribonucleico.

ARNm ARN mensajero.

ASR *Age-Standardized Rate*.

BSA Albúmina Sérica Bovina (del inglés *Bovine Serum Albumin*).

C Citosina.

CE Cáncer de Endometrio.

CEE Cáncer de Endometrio Endometroide.

CENE Cáncer de Endometrio No Endometrioide.

csv Valores separados por comas (del inglés *Comma Separated Values*).

CTCF *Corrected Total Cell Fluorescence*.

CTD *The Comparative Toxicogenomics Database*.

dATP Desoxiadenosina trifosfato.

dCTP Desoxicitosina trifosfato.

dGTP Desoxiguanosina trifosfato.

DPI Índice de pleiotropía de la enfermedad (del inglés *Disease Pleitropy Index*).

DSI Índice de Especificidad de la Enfermedad (del inglés *Disease Specificity Index*).

dTTP Desoxitimidina trifosfato.

dUTP desoxiuridina trifosfato.

EDTA Ácido etilendiaminotetraacético.

ETS *E26 Transformation-Specific.*

FDA *Food and Drug Administration.*

FIGO Federación Internacional de Ginecología y Obstetricia.

G Guanina.

G1 Grado 1.

G2 Grado 2.

G3 Grado 3.

GDA Asociaciones Gen-Enfermedad (del inglés *Gene-Disease Associations*).

GEO *Gene Expression Omnibus.*

GFP Proteína fluorescente verde (del inglés *Green Fluorescent Protein*).

Globocan Observatorio Global del Cáncer.

GO *Gene Ontology.*

GWAS *Genome-Wide Association Studies.*

HGE Hec1a-GFP-ETV5.

HGNC *HUGO Gene Nomenclature Committee.*

HPA *The Human Protein Atlas.*

HPO *The Human Phenotype Ontology.*

HR *Hazard Ratios.*

IC Intervalos de Confianza.

IDH Índice de Desarrollo Humano.

IgG Inmunoglobulina G.

IM Invasión Miometrial.

INC Instituto Nacional del Cáncer.

MATE1 *Multidrug and Toxin Extrusion Transporter*.

MeSH Encabezados de Temas Médicos (del inglés *Medical Subject Headings*).

NCBI *National Center for Biotechnology Information.*

NCI Instituto Nacional del Cáncer (*National Cancer Institute*).

NGS Secuenciación de próxima generación (del inglés *Next-Generation Sequencing*).

NHGRI *National Human Genome Research Institute.*

NLM Biblioteca Nacional de Medicina (del inglés *National Library of Medicine*).

OMS Organización Mundial de la Salud.

OR *Odds Ratios.*

OS Sobrevida Total (del inglés *Overall Survival*).

pb pares de bases.

PBS *Buffer* Fosfato Salino.

PCR Reacción en Cadena de la Polimerasa (del inglés *Polymerase Chain Reaction*).

PDB *Protein Data Bank*.

PGH Proyecto Genoma Humano.

qPCR Reacción en Cadena de la Polimerasa Cuantitativa (del inglés *Quantitative Polymerase Chain Reaction*).

RFS Sobrevida Libre de Recurrencia (del inglés *Recurrence-Free Survival*).

RIPA *Radio Immunoprecipitation Assay*.

RR Riesgo Relativo.

SDS Dodecilsulfato sódico (del inglés *Sodium Dodecyl Sulfate*).

SFB Suero Fetal Bovino.

SHh *Sonic Hedgehog*.

SUA Sangrado Uterino Anormal.

T Timina.

TCGA *The Cancer Genome Atlas*.

TCGA-UCEC TCGA-*Uterine Corpus Endometrial Cancer*.

U Uracilo.

Anexo A

Plataformas bioinformáticas consultadas

cBioPortal: http://www.cbioportal.org

Cutoff Finder: http://molpath.charite.de/cutoff

DisGeNET: www.disgenet.org

GEO: www.ncbi.nlm.nih.gov/geo

HPA: www.proteinatlas.org

Oligo Analyzer: www.idtdna.com/calc/analyzer

Primer-Blast: www.ncbi.nlm.nih.gov/primer-blast

PubMed: www.ncbi.nlm.nih.gov/pubmed

TCGA: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

ToppGene: https://toppgene.cchmc.org

Xena: https://xena.ucsc.edu

Anexo B

Características generales de las líneas celulares utilizadas

Línea celular	Tipo tumoral	Fuente	Medio de cultivo
HEC-1A	Adenocarcinoma de endometrio	Tumor primario, CEE, estadío IA, G2	Medio de cultivo McCoy's 5A, suplementado con 10% de Suero Fetal Bovino (SFB) y 1% de primario, Penicilina- Estreptomicina
Ishikawa	Adenocarcinoma de endometrio	Tumor primario, CEE, estadío IA, G1	Dulbecco's Modified Eagle's Medium-Ham's F12 (DMEM-F12), suplementado, con 10% de SFB y 1% de Penicilina-Estreptomicina

Tabla 22: Características generales de las líneas celulares utilizadas a lo largo del trabajo, junto con información de qué tipo tumoral provienen, cuáles son las características clínico-patológicas de la linea celular, y cuál es el medio de cultivo idóneo para dicha línea celular.

Anexo C

Cebadores

1 Secuencias utilizadas

Gen	Foward 5'-3'	Reverse 3'-5'	pb
GAPDH	TGCACCACCAACTGCTTAGC	GGCATGGACTGTGGTCATGAG	88
TMPRSS2	ACGTCGATTCTTGCCAGGGT	GCCGTCTGCCCTCATTTGTC	179
TPX2	GCGCTCTGATTGGTGCATTC	GCGCTCTGATTGGTGCATTC	131
PTCH1	TCACGTTGCTTTGGCCTTTCTG	AACTCAGATCCCGCCAGCAT	141

Tabla 23: Secuencias de cebadores diseñados para la realización de PCR estándar.

2 Información adicional sobre TMPRSS2

Gen de 14 exones

Foward: `ACGTCGATTCTTGCCAGGGT`

Reverse: `GACAAATGAGGGCAGACGGC`

- **NM_001135099.1** ***Homo sapiens transmembrane serine protease 2 (TMPRSS2), transcript variant 1, mRNA***:

  Se intercalan los colores rojo y negro para establecer los exones 1 a 14. Según se observa, los cebadores fueron diseñados sobre secuencias localizadas en los exones 12 y 13 (cebador *forward*) y 13-14 (cebador *reverse*).

ACCAGGGTCCCGGCTCGGGGTCCGGGCTGGGGAGGGGAACCTGGGCGCCTGGGACCCGCCGATGCCCCCTGCCCCG
CCCGGAGGTGAAAGCGGGTGTGAGGAGCGCGGCGCGGCAGGTCATATTGAACATTCCAGATACCTATCATTACTCG
ATGCTGTTGATAACAGCAAGATGGCTTTGAACTCAGGGTCACCACCAGCTATTGGACCTTACTATGAAAACCATGG
ATACCAACCGGAAAACCCCTATCCCGCACAGCCCACTGTGGTCCCCACTGTCTACGAGGTGCATCCGGCTCAGTAC
TACCCGTCCCCCGTGCCCCAGTACGCCCCGAGGGTCCTGACGCAGGCTTCCAACCCCGTCGTCTGCACGCAGCCCA
AATCCCCATCCGGGACAGTGTGCACCTCAAAGACTAAGAAAGCACTGTGCATCACCTTGACCCTGGGGACCTTCCT
CGTGGGAGCTGCGCTGGCCGCTGGCCTACTCTGGAAGTTCATGGGCAGCAAGTGCTCCAACTCTGGGATAGAGTGC
GACTCCTCAGGTACCTGCATCAACCCCTCTAACTGGTGTGATGGCGTGTCACACTGCCCCGGCGGGGAGGACGAGA
ATCGGTGTGTTCGCCTCTACGGACCAAACTTCATCCTTCAGGTGTACTCATCTCAGAGGAAGTCCTGGCACCCTGT
GTGCCAAGACGACTGGAACGAGAACTACGGGCGGGCGGCCTGCAGGGACATGGGCTATAAGAATAATTTTTACTCT
AGCCAAGGAATAGTGGATGACAGCGGATCCACCAGCTTTATGAAACTGAACACAAGTGCCGGCAATGTCGATATCT
ATAAAAAACTGTACCACAGTGATGCCTGTTCTTCAAAAGCAGTGGTTTCTTTACGCTGTATAGCCTGCGGGGTCAA
CTTGAACTCAAGCCGCCAGAGCAGGATTGTGGGCGGCGAGAGCGCGCTCCCGGGGGCCTGGCCCTGGCAGGTCAGC
CTGCACGTCCAGAACGTCCACGTGTGCGGAGGCTCCATCATCACCCCCGAGTGGATCGTGACAGCCGCCCACTGCG
TGGAAAAACCTCTTAACAATCCATGGCATTGGACGGCATTTGCGGGGATTTTGAGACAATCTTTCATGTTCTATGG
AGCCGGATACCAAGTAGAAAAAGTGATTTCTCATCCAAATTATGACTCCAAGACCAAGAACAATGACATTGCGCTG
ATGAAGCTGCAGAAGCCTCTGACTTTCAACGACCTAGTGAAACCAGTGTGTCTGCCCAACCCAGGCATGATGCTGC
AGCCAGAACAGCTCTGCTGGATTTCCGGGTGGGGGGCCACCGAGGAGAAAGGGAAGACCTCAGAAGTGCTGAACGC
TGCCAAGGTGCTTCTCATTGAGACACAGAGATGCAACAGCAGATATGTCTATGACAACCTGATCACACCAGCCATG
ATCTGTGCCGGCTTCCTGCAGGGGAACGTCGATTCTTGCCAGGGTGACAGTGGAGGGCCTCTGGTCACTTCGAAGA
ACAATATCTGGTGGCTGATAGGGGATACAAGCTGGGGTTCTGGCTGTGCCAAAGCTTACAGACCAGGAGTGTACGG
GAATGTGATGGTATTCACGGACTGGATTTATCGACAAATGAGGGCAGACGGCTAATCCACATGGTCTTCGTCCTTG
ACGTCGTTTTACAAGAAAACAATGGGGCTGGTTTTGCTTCCCCGTGCATGATTTACTCTTAGAGATGATTCAGAGG
TCACTTCATTTTTATTAAACAGTGAACTTGTCTGGCTTTGGCACTCTCTGCCATTCTGTGCAGGCTGCAGTGGCTC
CCCTGCCCAGCCTGCTCTCCCTAACCCCTTGTCCGCAAGGGGTGATGGCCGGCTGGTTGTGGGCACTGGCGGTCAA
GTGTGGAGGAGAGGGGTGGAGGCTGCCCCATTGAGATCTTCCTGCTGAGTCCTTTCCAGGGGCCAATTTTGGATGA
GCATGGAGCTGTCACCTCTCAGCTGCTGGATGACTTGAGATGAAAAAGGAGAGACATGGAAAGGGAGACAGCCAGG
TGGCACCTGCAGCGGCTGCCCTCTGGGGCCACTTGGTAGTGTCCCCAGCCTACCTCTCCACAAGGGGATTTTGCTG
ATGGGTTCTTAGAGCCTTAGCAGCCCTGGATGGTGGCCAGAAATAAAGGGACCAGCCCTTCATGGGTGGTGACGTG
GTAGTCACTTGTAAGGGGAACAGAAACATTTTTGTTCTTATGGGGTGAGAATATAGACAGTGCCCTTGGTGCGAGG
GAAGCAATTGAAAAGGAACTTGCCCTGAGCACTCCTGGTGCAGGTCTCCACCTGCACATTGGGTGGGGCTCCTGGG
AGGGAGACTCAGCCTTCCTCCTCATCCTCCCTGACCCTGCTCCTAGCACCCTGGAGAGTGCACATGCCCCTTGGTC
CTGGCAGGGCGCCAAGTCTGGCACCATGTTGGCCTCTTCAGGCCTGCTAGTCACTGGAAATTGAGGTCCATGGGGG
AAATCAAGGATGCTCAGTTTAAGGTACACTGTTTCCATGTTATGTTTCTACACATTGCTACCTCAGTGCTCCTGGA
AACTTAGCTTTTGATGTCTCCAAGTAGTCCACCTTCATTTAACTCTTTGAAACTGTATCATCTTTGCCAAGTAAGA
GTGGTGGCCTATTTCAGCTGCTTTGACAAAATGACTGGCTCCTGACTTAACGTTCTATAAATGAATGTGCTGAAGC
AAAGTGCCCATGGTGGCGGCGAAGAAGAGAAAGATGTGTTTTGTTTTGGACTCTCTGTGGTCCCTTCCAATGCTGT
GGGTTTCCAACCAGGGGAAGGGTCCCTTTTGCATTGCCAAGTGCCATAACCATGAGCACTACTCTACCATGGTTCT
GCCTCCTGGCCAAGCAGGCTGGTTTGCAAGAATGAAATGAATGATTCTACAGCTAGGACTTAACCTTGAAATGGAA
AGTCATGCAATCCCATTTGCAGGATCTGTCTGTGCACATGCCTCTGTAGAGAGCAGCATTCCCAGGGACCTTGGAA
ACAGTTGGCACTGTAAGGTGCTTGCTCCCCAAGACACATCCTAAAAGGTGTTGTAATGGTGAAAACGTCTTCCTTC
TTTATTGCCCCTTCTTATTTATGTGAACAACTGTTTGTCTTTTTTTGTATCTTTTTTAAACTGTAAAGTTCAATTG
TGAAAATGAATATCATGCAAATAAATTATGCAATTTTTTTTTCAAAGTAAAAAAAAAA

AMPLICON 179 pb

ACGTCGATTCTTGCCAGGGTGACAGTGGAGGGCCTCTGGTCACTTCGAAGAACAATATCTGGTGGCTGATAGGGGA
TACAAGCTGGGGTTCTGGCTGTGCCAAAGCTTACAGACCAGGAGTGTACGGGAATGTGATGGTATTCACGGACTGG
ATTTATCGACAAATGAGGGCAGACGGC

- **NM_005656.4** ***Homo sapiens transmembrane serine protease 2 (TMPRSS2), transcript variant 2, mRNA*****:**

GAGTAGGCGCGAGCTAAGCAGGAGGCGGAGGCGGAGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCTGGAGCGCGGC
AGGTCATATTGAACATTCCAGATACCTATCATTACTCGATGCTGTTGATAACAGCAAGATGGCTTTGAACTCAGGG
TCACCACCAGCTATTGGACCTTACTATGAAAACCATGGATACCAACCGGAAAACCCCTATCCCGCACAGCCCACTG
TGGTCCCCACTGTCTACGAGGTGCATCCGGCTCAGTACTACCCGTCCCCCGTGCCCCAGTACGCCCCGAGGGTCCT
GACGCAGGCTTCCAACCCCGTCGTCTGCACGCAGCCCAAATCCCCATCCGGGACAGTGTGCACCTCAAAGACTAAG
AAAGCACTGTGCATCACCTTGACCCTGGGGACCTTCCTCGTGGGAGCTGCGCTGGCCGCTGGCCTACTCTGGAAGT
TCATGGGCAGCAAGTGCTCCAACTCTGGGATAGAGTGCGACTCCTCAGGTACCTGCATCAACCCCTCTAACTGGTG
TGATGGCGTGTCACACTGCCCCGGCGGGGAGGACGAGAATCGGTGTGTTCGCCTCTACGGACCAAACTTCATCCTT
CAGGTGTACTCATCTCAGAGGAAGTCCTGGCACCCTGTGTGCCAAGACGACTGGAACGAGAACTACGGGCGGGCGG
CCTGCAGGGACATGGGCTATAAGAATAATTTTTACTCTAGCCAAGGAATAGTGGATGACAGCGGATCCACCAGCTT
TATGAAACTGAACACAAGTGCCGGCAATGTCGATATCTATAAAAAACTGTACCACAGTGATGCCTGTTCTTCAAAA
GCAGTGGTTTCTTTACGCTGTATAGCCTGCGGGGTCAACTTGAACTCAAGCCGCCAGAGCAGGATTGTGGGCGGCG
AGAGCGCGCTCCCGGGGGCCTGGCCCTGGCAGGTCAGCCTGCACGTCCAGAACGTCCACGTGTGCGGAGGCTCCAT
CATCACCCCCGAGTGGATCGTGACAGCCGCCCACTGCGTGGAAAAACCTCTTAACAATCCATGGCATTGGACGGCA
TTTGCGGGGATTTTGAGACAATCTTTCATGTTCTATGGAGCCGGATACCAAGTAGAAAAAGTGATTTCTCATCCAA
ATTATGACTCCAAGACCAAGAACAATGACATTGCGCTGATGAAGCTGCAGAAGCCTCTGACTTTCAACGACCTAGT
GAAACCAGTGTGTCTGCCCAACCCAGGCATGATGCTGCAGCCAGAACAGCTCTGCTGGATTTCCGGGTGGGGGGCC
ACCGAGGAGAAAGGGAAGACCTCAGAAGTGCTGAACGCTGCCAAGGTGCTTCTCATTGAGACACAGAGATGCAACA
GCAGATATGTCTATGACAACCTGATCACACCAGCCATGATCTGTGCCGGCTTCCTGCAGGGGAACGTCGATTCTTG
CCAGGGTGACAGTGGAGGGCCTCTGGTCACTTCGAAGAACAATATCTGGTGGCTGATAGGGGATACAAGCTGGGGT
TCTGGCTGTGCCAAAGCTTACAGACCAGGAGTGTACGGGAATGTGATGGTATTCACGGACTGGATTTATCACGTCG
ATTCTTGCCAGGGTTAATCCACATGGTCTTCGTCCTTGACGTCGTTTTACAAGAAAACAATGGGGCTGGTTTTGCT
TCCCCGTGCATGATTTACTCTTAGAGATGATTCAGAGGTCACTTCATTTTTATTAAACAGTGAACTTGTCTGGCTT
TGGCACTCTCTGCCATTCTGTGCAGGCTGCAGTGGCTCCCCTGCCCAGCCTGCTCTCCCTAACCCCTTGTCCGCAA
GGGGTGATGGCCGGCTGGTTGTGGGCACTGGCGGTCAAGTGTGGAGGAGAGGGGTGGAGGCTGCCCCATTGAGATC
TTCCTGCTGAGTCCTTTCCAGGGGCCAATTTTGGATGAGCATGGAGCTGTCACCTCTCAGCTGCTGGATGACTTGA
GATGAAAAAGGAGAGACATGGAAAGGGAGACAGCCAGGTGGCACCTGCAGCGGCTGCCCTCTGGGGCCACTTGGTA
GTGTCCCCAGCCTACCTCTCCACAAGGGGATTTTGCTGATGGGTTCTTAGAGCCTTAGCAGCCCTGGATGGTGGCC
AGAAATAAAGGGACCAGCCCTTCATGGGTGGTGACGTGGTAGTCACTTGTAAGGGGAACAGAAACATTTTTGTTCT
TATGGGGTGAGAATATAGACAGTGCCCTTGGTGCGAGGGAAGCAATTGAAAAGGAACTTGCCCTGAGCACTCCTGG
TGCAGGTCTCCACCTGCACATTGGGTGGGG

AMPLICON 179 pb

ACGTCGATTCTTGCCAGGGTGACAGTGGAGGGCCTCTGGTCACTTCGAAGAACAATATCTGGTGGCTGATAGGGGA
TACAAGCTGGGGTTCTGGCTGTGCCAAAGCTTACAGACCAGGAGTGTACGGGAATGTGATGGTATTCACGGACTGG
ATTTATCGACAAATGAGGGCAGACGGC

Anexo D

Resultados: Selección de genes con expresión diferencial en CE (GEO)

Listado de 39 genes con expresión diferencial en CE identificados a partir del estudio GSE17025 de la plataforma GEO.

Genes con expresión diferencial en CE		
ACSL5	HAUS8	PTCH1
ANAPC4	HES6	SLC25A35
ARSD	KIAA1324	SLC47A1
ATAD2	KIF7	SOAT1
CCDC160	LINC00261	SORBS2
CDC20B	LOC100129098	SPATA6
CEP83	LPCAT2	TAB2
CREB3L4	PALMD	TBCEL
DLGAP1-AS1	PCDH7	TMEM132A
DLGAP1-AS2	PDZRN3	TMPRSS2
FAM189A2	PGR	TPX2
FOXA2	PLEKHH1	TRAF3IP2
GSPT1	PPM1H	ZDHHC2

Tabla 24: Genes con expresión diferencial en CE. Listado de 39 genes que se encuentran en la intersección de las tres listas obtenidas mediante el análisis de expresión diferencial del estudio GSE17025 la plataforma GEO (Sección 1.1, Resultados).

Anexo E

Resultados: DisGeNET (DSI y DPI)

Listado de genes obtenidos luego de la búsqueda de genes asociados a CE, resultado de unificar los términos existentes de enfermedad relacionados al CE de todas las bases de datos de DisGeNET. Listas filtradas por DSI y luego por DPI

Gen	DSI	Gen	DSI	Gen	DSI
MEAF6	1	ENOSF1	0,668	SIRT2	0,594
ZC3H7A	1	RARRES1	0,668	FOXA1	0,594
ERVV-2	1	FES	0,668	P2RY2	0,594
MIG7	1	STOML2	0,668	LXN	0,594
NUTM2A	0,93	RNF43	0,664	ZIC2	0,594
NUTM2B	0,93	TNK2	0,664	CDC25C	0,594
GPCPD1	0,93	LAMC1	0,664	NR5A2	0,593
TOX4	0,93	HAND2	0,664	SRA1	0,593
C19orf33	0,889	HSPE1	0,661	ELAVL2	0,593
RDH16	0,889	GZMA	0,661	RPL10	0,593
ZC3H7B	0,889	MYBL2	0,661	SERPINB5	0,591
COPZ2	0,889	RELB	0,661	SOCS2	0,591
SNHG12	0,889	KDM4A	0,661	NCOA2	0,591
C10orf67	0,86	TSPYL2	0,657	ESRRB	0,591
MBTD1	0,86	SPOP	0,657	GGTLC5P	0,591
JPH4	0,86	PSME3	0,657	GGTLC3	0,591

YTHDC1	0,86	TRAM1	0,657	GGT2	0,591
MRPL19	0,86	FOSL2	0,657	GGTLC4P	0,591
OVGP1	0,838	CALB1	0,657	KISS1R	0,591
CXorf67	0,838	HYAL2	0,657	MLH3	0,589
CGRRF1	0,838	MIR129-2	0,654	ETV5	0,589
GORASP2	0,838	LOC102724023	0,654	GPER1	0,589
IPO8	0,819	CIRBP	0,654	DDX3X	0,589
ETF1	0,819	SOX7	0,654	EGLN1	0,589
THUMPD1	0,804	CGB8	0,651	RCAN1	0,589
MIR618	0,804	MIR181A2	0,651	CCNE1	0,587
HYAL3	0,804	MED15	0,651	FGF4	0,587
KCNH6	0,804	UGT2B17	0,651	PWAR4	0,587
CASC2	0,79	C21orf33	0,651	MAL	0,587
APBB3	0,79	HAS1	0,648	MSI1	0,586
CAPN9	0,79	BHLHE41	0,648	HSD17B7	0,586
CIDEA	0,79	LINC-ROR	0,648	CALD1	0,586
RBMXP1	0,79	TUBA1B	0,648	DAB2	0,586
RXFP3	0,79	KLK8	0,648	HOXA10	0,584
DENND1A	0,79	MT1E	0,648	SIRT3	0,584
SRSF10	0,779	SEMA3B	0,645	INHA	0,584
HAAO	0,779	UTS2R	0,645	HOXA11	0,582
TEAD4	0,779	MIR490	0,645	CDKN2B-AS1	0,581
CHD4	0,768	MIR503	0,645	SALL4	0,581
SULT2B1	0,768	GAS5	0,645	SNCG	0,579
EPC1	0,768	RPL11	0,645	UHRF1	0,579
CDC42SE2	0,768	CRABP1	0,645	NOV	0,579
THEMIS2	0,768	CCL28	0,645	SUZ12	0,578
SLC7A10	0,768	CGB5	0,642	JUP	0,578
FAS-AS1	0,768	PRAP1	0,642	TAM	0,578
ABCF2	0,768	RHOC	0,642	LNPEP	0,578
TRIM22	0,768	JAZF1	0,639	PER1	0,578
DYNLL2	0,768	PELP1	0,639	ADIPOR2	0,578
MACROD1	0,768	KDM4B	0,639	PCLAF	0,578
ARMC3	0,758	TNFSF12-TNFSF13	0,639	ITGA4	0,578
PPP4C	0,758	SIRT7	0,639	ZMYND10	0,578
TCEAL7	0,758	NRIP1	0,636	ERAP1	0,578
SPAM1	0,758	LPP	0,636	RAB40B	0,578
ELOF1	0,758	SERPINF2	0,636	CASP7	0,576
TMEM54	0,758	PROK1	0,633	NEURL1	0,576
PLXNA3	0,75	ZFHX3	0,633	FGF9	0,576

CSNK2B	0,75	BST2	0,633	ADAMTS1	0,576
PPME1	0,75	HSD17B2	0,631	MIR148A	0,575
CELP	0,75	HOXB13	0,631	NBR1	0,575
RAB32	0,75	MMP26	0,631	PCBP4	0,575
DGKA	0,75	RPL36A	0,631	RBP1	0,575
GNRH2	0,75	PGRMC1	0,631	CEACAM1	0,573
ZFP36L1	0,75	PRMT1	0,631	MIR204	0,573
MAP3K4	0,741	MIR99A	0,631	BANF1	0,573
AATF	0,741	LGR6	0,631	RAD50	0,573
PLXNA1	0,741	PLEK	0,628	CTNND1	0,573
RXFP1	0,741	TRIM27	0,628	BAD	0,573
MIR124-2	0,734	NUPR1	0,628	NTN1	0,573
TRIM25	0,727	FOXD3	0,628	PDIK1L	0,572
WFDC2	0,727	RBMX	0,628	NOTCH4	0,572
RPL22	0,727	HTRA2	0,628	QPCT	0,57
THADA	0,727	SEMA3F	0,626	MSN	0,57
PIK3C2A	0,721	PAK4	0,626	SLC25A20	0,57
ELF1	0,721	ADH7	0,626	TNFSF12	0,57
MUC20	0,721	C1QBP	0,626	AQP2	0,569
SIRT5	0,721	PR@	0,626	MIR106B	0,569
SIRT4	0,721	CDC25B	0,623	BIRC2	0,568
RPS6KA6	0,721	BARD1	0,623	SAG	0,568
OTUB1	0,721	HSD3B1	0,623	STAR	0,566
MIR505	0,721	TNFRSF10C	0,623	AKR1B10	0,566
TSNAX	0,721	CHFR	0,621	CCND3	0,566
KIAA1324	0,714	FABP5	0,621	KLK4	0,566
HTRA3	0,714	CGB3	0,621	GLYAT	0,566
PHF1	0,714	MIR152	0,621	PRDM2	0,565
TFAP2C	0,714	NDC80	0,619	TSHZ1	0,565
C1GALT1C1	0,714	AMHR2	0,619	CXCL2	0,565
ST6GALNAC1	0,714	HPR	0,619	F2RL3	0,565
SLC22A16	0,709	NEDD4	0,616	MTA1	0,565
EEC1	0,709	MIR130B	0,616	PAK3	0,564
TNKS2	0,709	PCDH10	0,616	EPHB4	0,564
SYTL2	0,709	HHEX	0,614	KCNMA1	0,564
EMP2	0,703	CTCFL	0,614	CXCL11	0,564
GSTZ1	0,703	IGFBP6	0,614	CLDN4	0,562
PTENP1	0,703	ARRB1	0,614	SMARCA1	0,562
INHBB	0,703	UBE2N	0,614	MIR200B	0,562
PTGES2	0,703	PTPRA	0,612	PPP2R1A	0,562

LRG1	0,698	RBL2	0,612	TERF2	0,562
SCRIB	0,698	HMMR	0,61	MUC16	0,562
MIR26A1	0,698	BAAT	0,61	ERVK-18	0,561
ASCL2	0,698	PRMT5	0,61	DDX53	0,561
CTNNBIP1	0,693	OLFM4	0,61	PPP1R2P9	0,56
ATAD2	0,693	ESRRA	0,61	SMARCE1	0,56
CTBP2	0,693	MIR134	0,61	NCOA3	0,558
MIR337	0,693	ERRFI1	0,608	MIR100	0,558
PDE7A	0,693	NCOA1	0,608	MIR34B	0,558
SOX1	0,693	HYAL1	0,608	NOD1	0,557
CABLES1	0,693	MIR34C	0,608	CYP4F3	0,557
HMGN5	0,693	PIGR	0,608	IGFBP5	0,557
IFITM1	0,688	CRY1	0,606	CEACAM7	0,556
BRD7	0,688	GPBAR1	0,606	LTF	0,555
NID1	0,688	DACH1	0,606	ADIPOR1	0,555
BHD	0,688	RNU1-1	0,606	EPHA8	0,555
CYP3A7	0,684	PGPEP1	0,606	EFEMP1	0,555
DDT	0,684	PRB2	0,606	MIR222	0,555
EFNA2	0,684	AKR1C1	0,604	DCTN4	0,555
MIR372	0,684	RNR1	0,604	PIK3R2	0,555
KLLN	0,68	WNT10B	0,604	PAG1	0,555
SLC14A2	0,68	DROSHA	0,604	SUMO1	0,555
SLC29A2	0,676	SULT2A1	0,604	STC1	0,553
KLF9	0,676	PDGFD	0,604	THBS2	0,553
ERV3-1	0,676	SERPINA6	0,604	SRD5A2	0,552
MIR199B	0,676	EBAG9	0,602	PTGER4	0,552
MIR98	0,676	INTS2	0,602	PPIA	0,551
STARD13	0,676	VEGFB	0,602	MIR182	0,551
LPXN	0,676	INHBE	0,602	SREBF1	0,551
KCNN4	0,672	IGF2BP3	0,6	HCRT	0,551
LLGL1	0,672	INHBA	0,6	SOX4	0,55
CDK8	0,672	PSMD10	0,6	CEACAM3	0,55
DAB1	0,672	GPRC5A	0,6	BBS2	0,55
GGN	0,672	POLE	0,598	AKR1C3	0,549
MARCH8	0,672	SIRT6	0,598	PAWR	0,549
AFDN	0,672	WNT2	0,598	CGA	0,549
SPA17	0,672	FEN1	0,596	AGPAT2	0,549
PIWIL2	0,672	GHRHR	0,596	MAP3K1	0,549
PKN1	0,672	MSR1	0,596	PSG2	0,549
CCDC54	0,672	HSD17B1	0,594	SAI1	0,549

CACUL1	0,672	FSTL1	0,594	CLDN7	0,549
SCGB2A1	0,668	CLDN3	0,594	MIR200C	0,548
ND3	0,668	UGT2B7	0,594	LOC100128922	0,548

Tabla 25: Genes filtrados por DSI. Listado de genes obtenidos luego de la búsqueda de genes asociados a CE, junto con su DSI asociado, filtrados por el valor medio del DSI.

Gen	**DPI**	**Gen**	**DPI**	**Gen**	**DPI**
MEAF6	0,071	NEURL1	0,357	CDC25B	0,536
ZC3H7A	0,071	PLXNA3	0,393	MIR152	0,536
ERVV-2	0,071	THADA	0,393	NDC80	0,536
MIG7	0,071	PIK3C2A	0,393	RBL2	0,536
NUTM2A	0,071	MUC20	0,393	PRB2	0,536
NUTM2B	0,071	SIRT4	0,393	AKR1C1	0,536
COPZ2	0,071	CTBP2	0,393	SULT2A1	0,536
SNHG12	0,071	SOX1	0,393	VEGFB	0,536
JPH4	0,071	CYP3A7	0,393	POLE	0,536
KCNH6	0,071	EFNA2	0,393	FOXA1	0,536
ABCF2	0,071	DAB1	0,393	NR5A2	0,536
GPCPD1	0,107	AFDN	0,393	ETV5	0,536
TOX4	0,107	CCDC54	0,393	DDX3X	0,536
C19orf33	0,107	ND3	0,393	CCNE1	0,536
RDH16	0,107	KDM4A	0,393	FGF4	0,536
MBTD1	0,143	MIR129-2	0,393	HSD17B7	0,536
CXorf67	0,143	MED15	0,393	HOXA10	0,536
ETF1	0,143	KLK8	0,393	HOXA11	0,536
MIR618	0,143	RHOC	0,393	SNCG	0,536
HYAL3	0,143	PELP1	0,393	MIR204	0,536
CASC2	0,143	KDM4B	0,393	BANF1	0,536
RBMXP1	0,143	SIRT7	0,393	PDIK1L	0,536
DYNLL2	0,143	MIR99A	0,393	PAK3	0,536
MACROD1	0,143	CGB3	0,393	PPP2R1A	0,536
PPME1	0,143	AMHR2	0,393	MUC16	0,536
ASCL2	0,143	HPR	0,393	BBS2	0,536
ZC3H7B	0,179	HMMR	0,393	SAI1	0,536
C10orf67	0,179	ERRFI1	0,393	WNT4	0,536
YTHDC1	0,179	HYAL1	0,393	MIR205	0,536
OVGP1	0,179	SUZ12	0,393	HSD3B2	0,536
GORASP2	0,179	PHF1	0,429	AMH	0,536
IPO8	0,179	PTGES2	0,429	SFRP4	0,536

CDC42SE2	0,179	MIR337	0,429	GNRH1	0,536
SPAM1	0,179	NID1	0,429	MARCH8	0,571
DGKA	0,179	SLC29A2	0,429	LAMC1	0,571
RXFP1	0,179	STARD13	0,429	RELB	0,571
MIR124-2	0,179	ENOSF1	0,429	PSME3	0,571
TSNAX	0,179	FES	0,429	PGRMC1	0,571
MIR372	0,179	MYBL2	0,429	ADH7	0,571
CGRRF1	0,214	LOC102724023	0,429	PR@	0,571
THUMPD1	0,214	CIRBP	0,429	FABP5	0,571
CIDEA	0,214	C21orf33	0,429	HHEX	0,571
RXFP3	0,214	SEMA3B	0,429	PRMT5	0,571
SRSF10	0,214	RPL11	0,429	CRY1	0,571
CHD4	0,214	PRAP1	0,429	GPBAR1	0,571
SULT2B1	0,214	PROK1	0,429	DROSHA	0,571
PPP4C	0,214	ZFHX3	0,429	PDGFD	0,571
TMEM54	0,214	BARD1	0,429	HSD17B1	0,571
GNRH2	0,214	EBAG9	0,429	FSTL1	0,571
ZFP36L1	0,214	IGF2BP3	0,429	NCOA2	0,571
AATF	0,214	PSMD10	0,429	DAB2	0,571
MIR505	0,214	GPRC5A	0,429	PCLAF	0,571
KIAA1324	0,214	EGLN1	0,429	RAB40B	0,571
EEC1	0,214	JUP	0,429	NBR1	0,571
HMGN5	0,214	ZMYND10	0,429	CEACAM1	0,571
RNF43	0,214	HOTAIR	0,429	MTA1	0,571
MRPL19	0,25	RAB32	0,464	KCNMA1	0,571
DENND1A	0,25	PDE7A	0,464	STC1	0,571
HAAO	0,25	SPA17	0,464	SREBF1	0,571
EPC1	0,25	CACUL1	0,464	SOX4	0,571
THEMIS2	0,25	FOSL2	0,464	AGPAT2	0,571
ELOF1	0,25	HAS1	0,464	MIR200C	0,571
CELP	0,25	BHLHE41	0,464	LHCGR	0,571
PLXNA1	0,25	MIR503	0,464	LEF1	0,571
TRIM25	0,25	JAZF1	0,464	CD82	0,571
C1GALT1C1	0,25	HSD17B2	0,464	FGF7	0,571
TNKS2	0,25	MMP26	0,464	TSPYL2	0,607
SYTL2	0,25	FOXD3	0,464	MT1E	0,607
INHBB	0,25	SEMA3F	0,464	CCL28	0,607
CTNNBIP1	0,25	PCDH10	0,464	BST2	0,607
CABLES1	0,25	UBE2N	0,464	TNFRSF10C	0,607
BHD	0,25	OLFM4	0,464	MIR130B	0,607

KLLN	0,25	MIR134	0,464	RNU1-1	0,607
LLGL1	0,25	WNT10B	0,464	WNT2	0,607
RARRES1	0,25	GHRHR	0,464	FEN1	0,607
TEAD4	0,286	CLDN3	0,464	P2RY2	0,607
FAS-AS1	0,286	KISS1R	0,464	LXN	0,607
ARMC3	0,286	MLH3	0,464	CDC25C	0,607
TCEAL7	0,286	INHA	0,464	SRA1	0,607
SIRT5	0,286	CTNND1	0,464	SOCS2	0,607
OTUB1	0,286	CLDN4	0,464	ESRRB	0,607
HTRA3	0,286	HTR1B	0,464	CALD1	0,607
LRG1	0,286	SATB2	0,464	SIRT3	0,607
IFITM1	0,286	ABCC9	0,464	FGF9	0,607
KLF9	0,286	GSTZ1	0,5	ADAMTS1	0,607
STOML2	0,286	SLC14A2	0,5	MIR148A	0,607
TNK2	0,286	ERV3-1	0,5	NTN1	0,607
HYAL2	0,286	HSPE1	0,5	STAR	0,607
MIR490	0,286	TRAM1	0,5	CCND3	0,607
HOXB13	0,286	UGT2B17	0,5	CXCL2	0,607
CHFR	0,286	LINC-ROR	0,5	EPHB4	0,607
APBB3	0,321	UTS2R	0,5	MIR200B	0,607
CAPN9	0,321	GAS5	0,5	SMARCE1	0,607
SLC7A10	0,321	CRABP1	0,5	MIR100	0,607
TRIM22	0,321	LPP	0,5	EPHA8	0,607
CSNK2B	0,321	LGR6	0,5	EFEMP1	0,607
WFDC2	0,321	RBMX	0,5	MIR222	0,607
RPL22	0,321	HTRA2	0,5	THBS2	0,607
TFAP2C	0,321	HSD3B1	0,5	AKR1C3	0,607
ST6GALNAC1	0,321	CTCFL	0,5	MAP3K1	0,607
SLC22A16	0,321	IGFBP6	0,5	LOC100128922	0,607
SCRIB	0,321	BAAT	0,5	MTDH	0,607
MIR26A1	0,321	ESRRA	0,5	SH2B3	0,607
ATAD2	0,321	NCOA1	0,5	PRSS55	0,607
MIR199B	0,321	MIR34C	0,5	IRS2	0,607
LPXN	0,321	DACH1	0,5	GHRH	0,607
GGN	0,321	INTS2	0,5	SPZ1	0,607
SCGB2A1	0,321	INHBE	0,5	CRISP2	0,607
SPOP	0,321	ZIC2	0,5	MIR214	0,607
CGB8	0,321	GPER1	0,5	MALAT1	0,607
MIR181A2	0,321	MSI1	0,5	DKK3	0,607
CGB5	0,321	NOV	0,5	MIR143	0,607

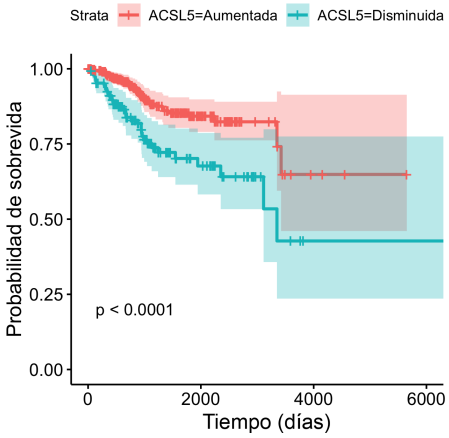
SERPINB5	0,321	AQP2	0,5	FUT4	0,607
MAP3K4	0,357	KLK4	0,5	CYP11A1	0,607
ELF1	0,357	NCOA3	0,5	DDR1	0,607
RPS6KA6	0,357	SRD5A2	0,5	HES1	0,607
EMP2	0,357	CLDN7	0,5	GRP	0,607
PTENP1	0,357	GNRHR	0,5	FSD1L	0,607
BRD7	0,357	AURKB	0,5	SKP2	0,607
DDT	0,357	KRT19	0,5	KISS1	0,607
MIR98	0,357	PKN1	0,536	FSD1	0,607
KCNN4	0,357	GZMA	0,536	NTRK2	0,607
CDK8	0,357	CALB1	0,536	PMS2	0,607
PIWIL2	0,357	TNFSF12-TNFSF13	0,536	ZEB1	0,607
HAND2	0,357	NRIP1	0,536	TRNL1	0,607
SOX7	0,357	SERPINF2	0,536	MSH6	0,607
RPL36A	0,357	PRMT1	0,536	TERC	0,607
NUPR1	0,357	TRIM27	0,536	S100A4	0,607
PAK4	0,357	C1QBP	0,536		

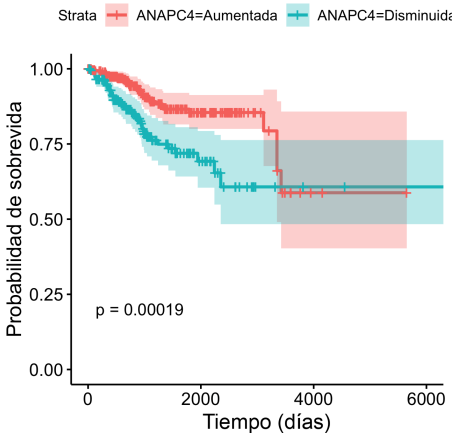
Tabla 26: Genes filtrados por DPI. Listado de genes obtenidos luego de la búsqueda de genes asociados a CE, junto con su DPI asociado, filtrados por el valor medio del DPI.

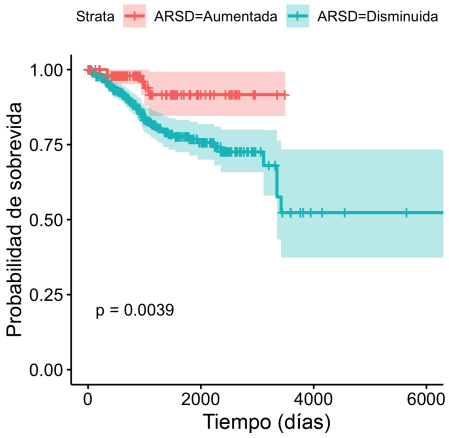
Anexo F

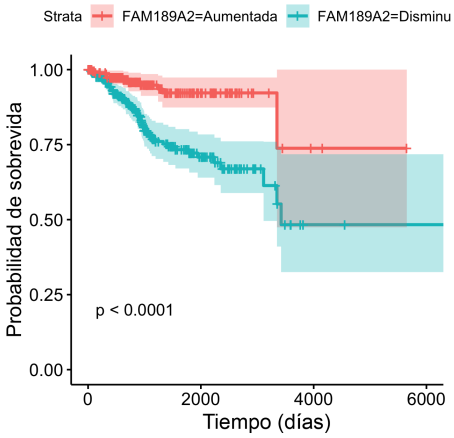
Resultados: curvas de sobrevida

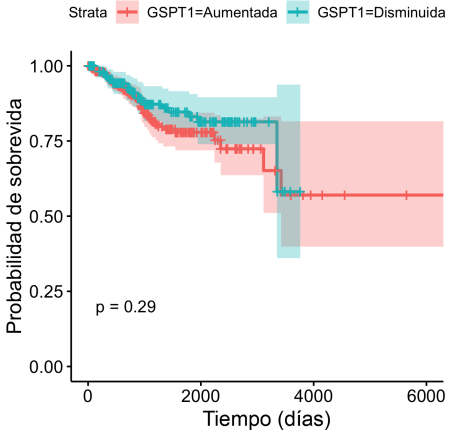
1 Sobrevida total (OS)

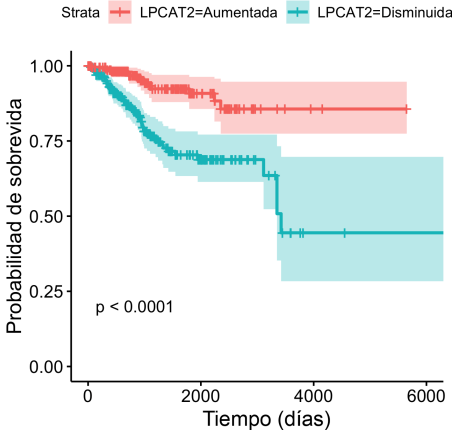


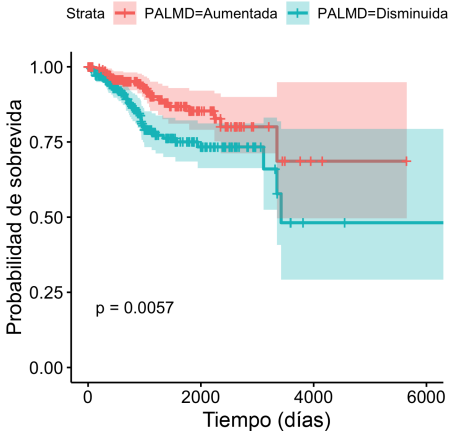


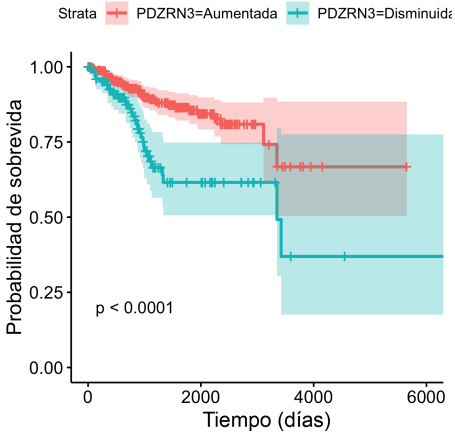

Strata ARSD=Aumentada ARSD=Disminuida
Probabilidad de sobrevida
1.00
0.75
0.50
0.25
0.00
p = 0.0039
0
2000
4000
6000
Tiempo (días)

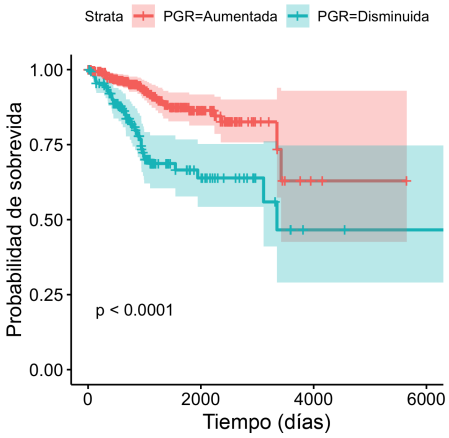

Strata FAM189A2=Aumentada FAM189A2=Disminu
Probabilidad de sobrevida
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0
2000
4000
6000
Tiempo (días)

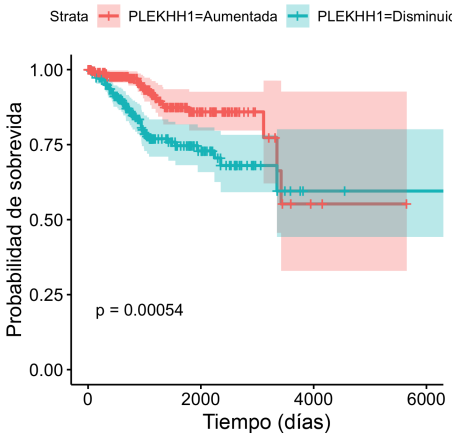

Strata GSPT1=Aumentada GSPT1=Disminuida
Probabilidad de sobrevida
1.00
0.75
0.50
0.25
0.00
p = 0.29
0
2000
4000
6000
Tiempo (días)

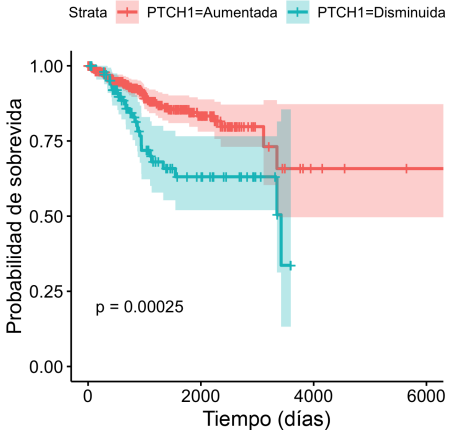

Strata LPCAT2=Aumentada LPCAT2=Disminuida
Probabilidad de sobrevida
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0
2000
4000
6000
Tiempo (días)

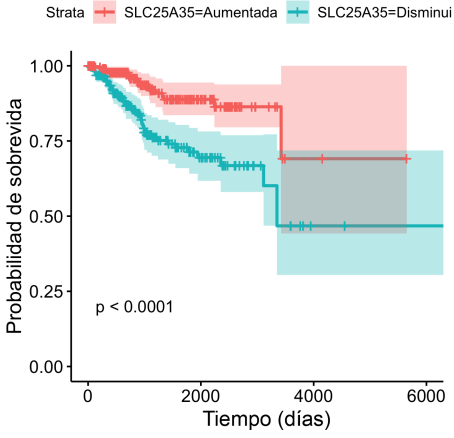

Strata PALMD=Aumentada PALMD=Disminuida
Probabilidad de sobrevida
1.00
0.75
0.50
0.25
0.00
p = 0.0057
0
2000
4000
6000
Tiempo (días)

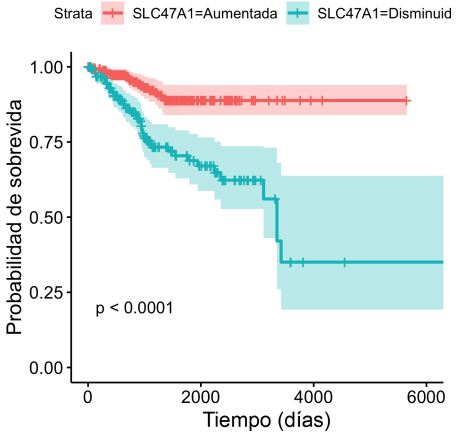

Strata PDZRN3=Aumentada PDZRN3=Disminuid
Probabilidad de sobrevida
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0
2000
4000
6000
Tiempo (días)

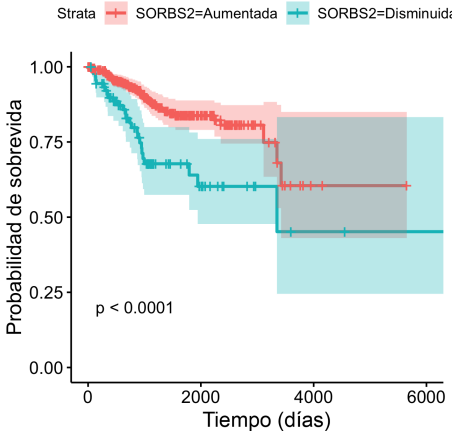

Strata PGR=Aumentada PGR=Disminuida
Probabilidad de sobrevida
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0
2000
4000
6000
Tiempo (días)


Strata PLEKHH1=Aumentada PLEKHH1=Disminuic
Probabilidad de sobrevida
1.00
0.75
0.50
0.25
0.00
p = 0.00054
0
2000
4000
6000
Tiempo (días)


Strata PTCH1=Aumentada PTCH1=Disminuida
Probabilidad de sobrevida
1.00
0.75
0.50
0.25
0.00
p = 0.00025
0
2000
4000
6000
Tiempo (días)


Strata SLC25A35=Aumentada SLC25A35=Disminui
Probabilidad de sobrevida
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0
2000
4000
6000
Tiempo (días)


Strata SLC47A1=Aumentada SLC47A1=Disminuid
Probabilidad de sobrevida
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0
2000
4000
6000
Tiempo (días)


Strata SORBS2=Aumentada SORBS2=Disminuida
Probabilidad de sobrevida
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0
2000
4000
6000
Tiempo (días)

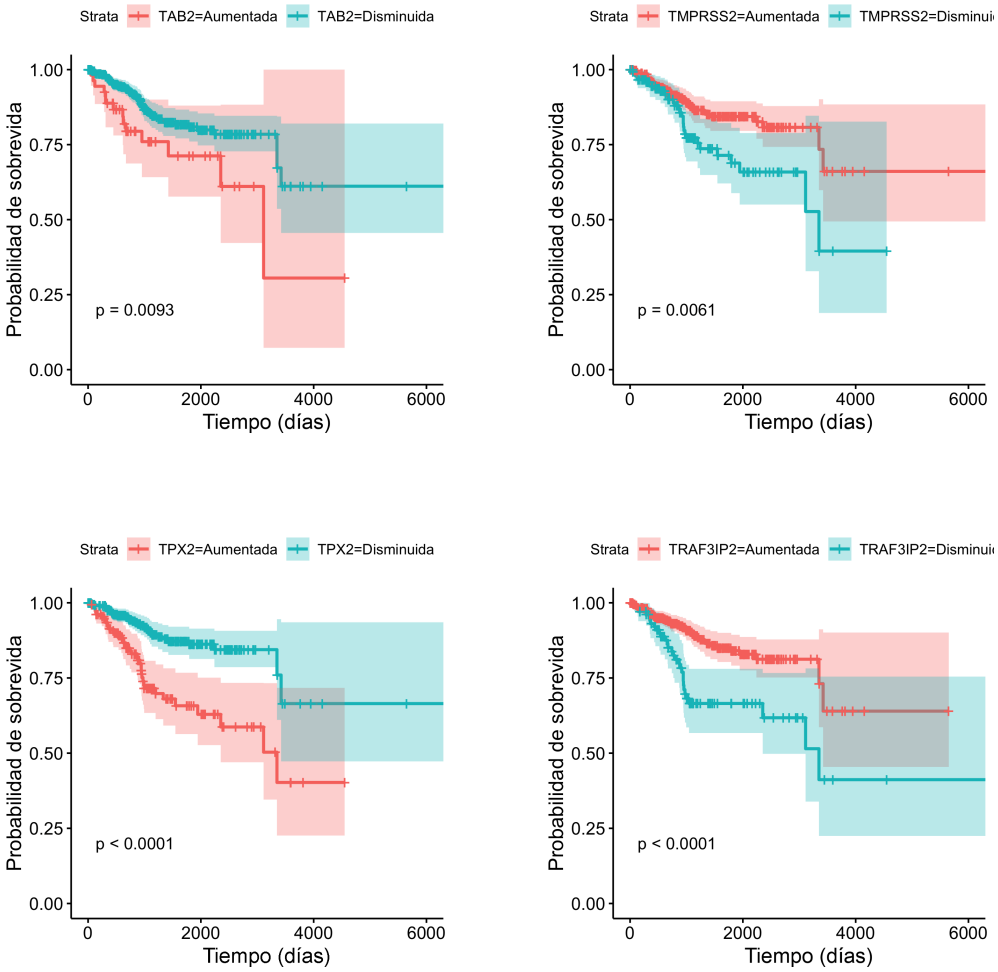
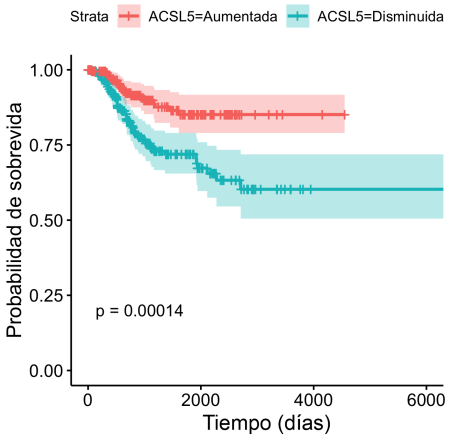
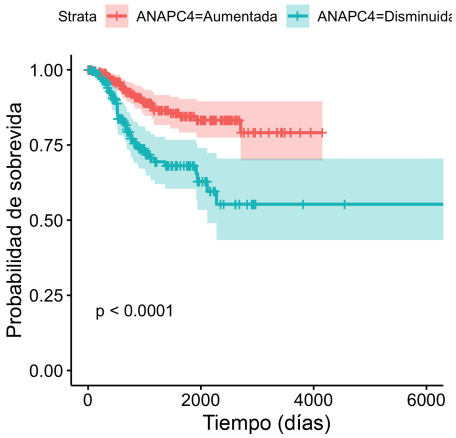
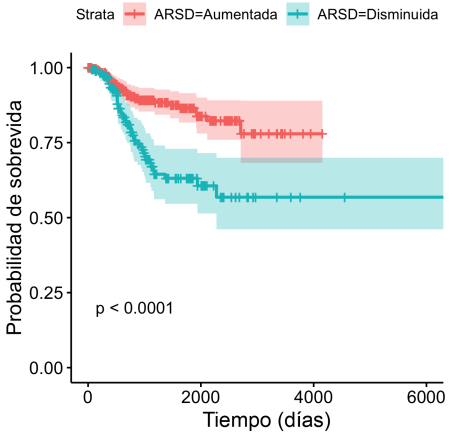


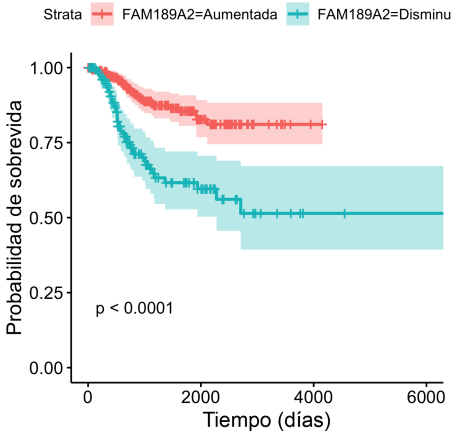
Figura F.1. Curvas de sobrevida total (OS) de los 18 genes obtenidos luego del cálculo de ORs. Los genes son: ACSL5, ANAPC4, ARSD, FAM189A2, GSPT1, LPACT2, PALMD, PDZRN3, PGR, PLEKHH1, PTCH1, SLC25A35, SLC47A1, SORBS2, TAB2, TMPRSS2, TPX2 y TRAF3IP2.

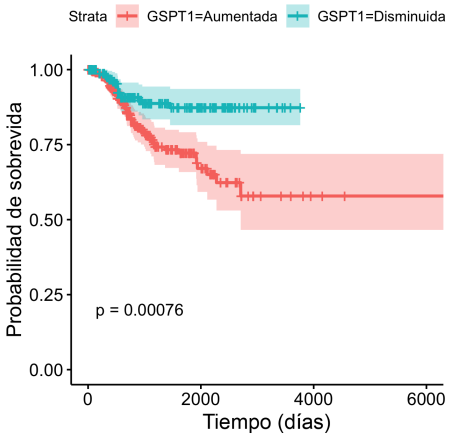
2 Sobrevida libre de recurrencia (RFS)

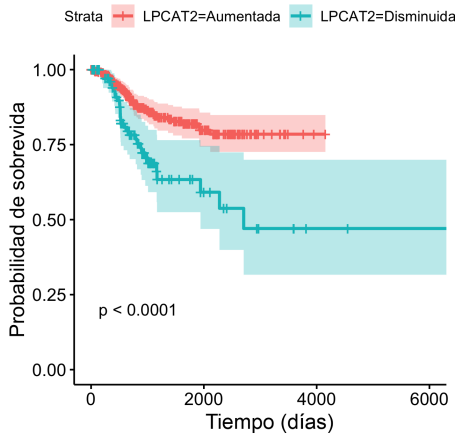


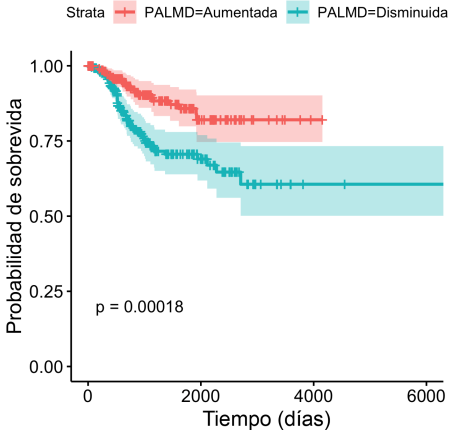


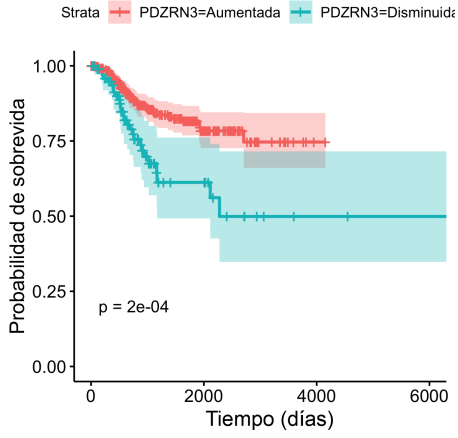


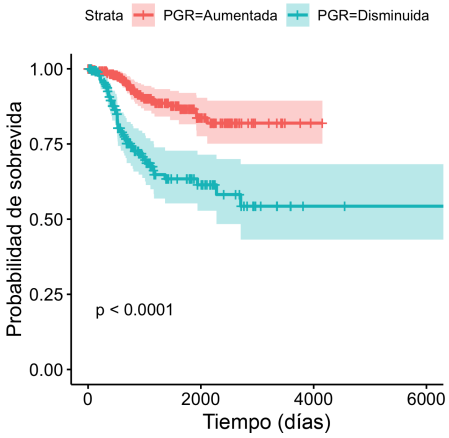


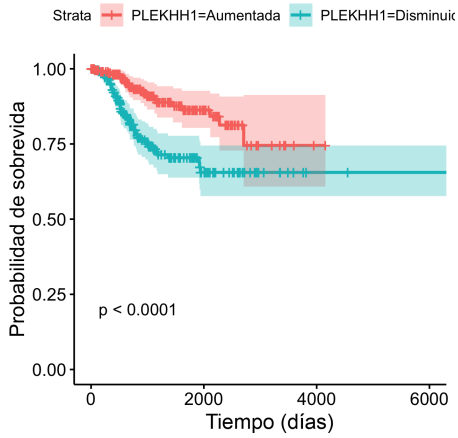

Strata GSPT1=Aumentada GSPT1=Disminuida
Probabilidad de sobrevida
1.00
0.75
0.50
0.25
0.00
p = 0.00076
0
2000
4000
6000
Tiempo (días)

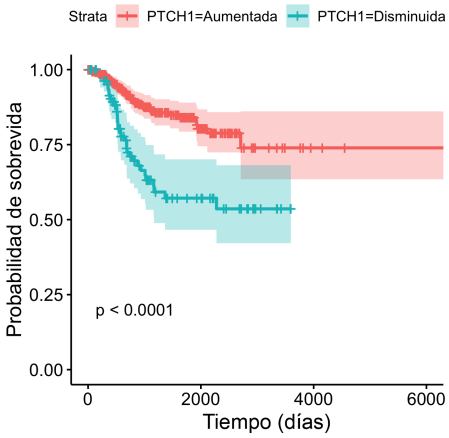

Strata LPCAT2=Aumentada LPCAT2=Disminuida
Probabilidad de sobrevida
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0
2000
4000
6000
Tiempo (días)

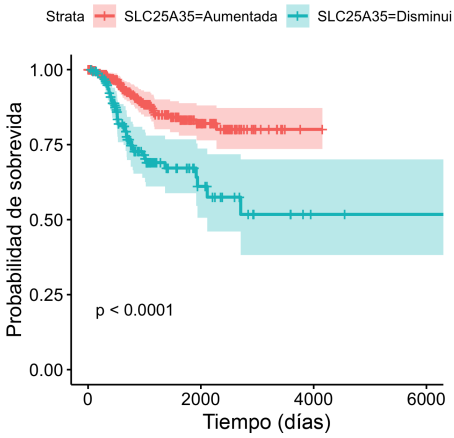

Strata PALMD=Aumentada PALMD=Disminuida
Probabilidad de sobrevida
1.00
0.75
0.50
0.25
0.00
p = 0.00018
0
2000
4000
6000
Tiempo (días)

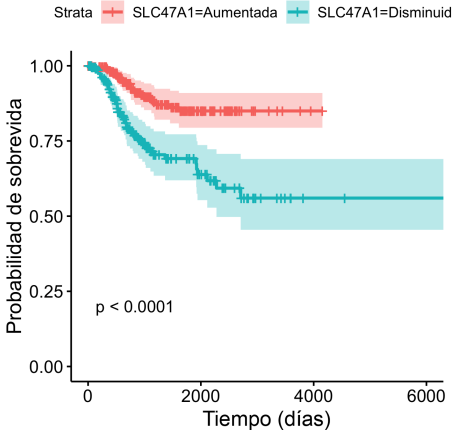

Strata PDZRN3=Aumentada PDZRN3=Disminuid
Probabilidad de sobrevida
1.00
0.75
0.50
0.25
0.00
p = 2e-04
0
2000
4000
6000
Tiempo (días)

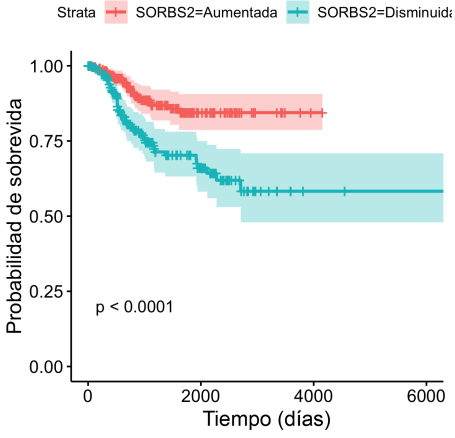

Strata PGR=Aumentada PGR=Disminuida
Probabilidad de sobrevida
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0
2000
4000
6000
Tiempo (días)

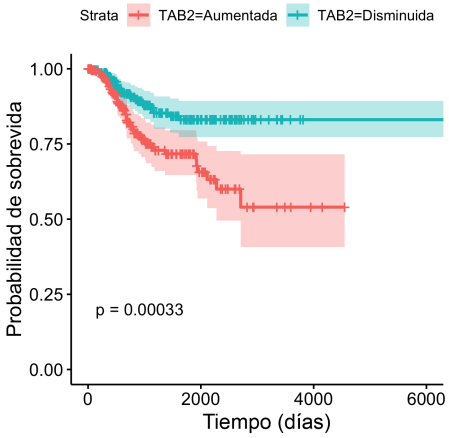

Strata PLEKHH1=Aumentada PLEKHH1=Disminuic
Probabilidad de sobrevida
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0
2000
4000
6000
Tiempo (días)

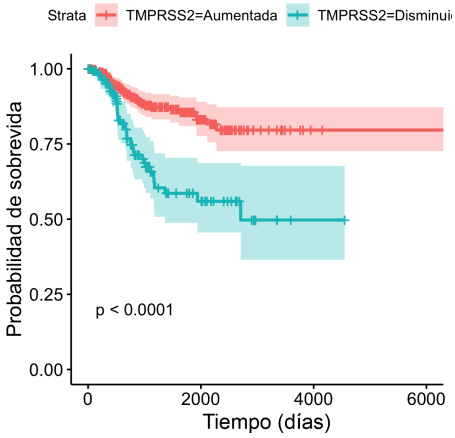

Strata PTCH1=Aumentada PTCH1=Disminuida
Probabilidad de sobrevida
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0
2000
4000
6000
Tiempo (días)


Strata SLC25A35=Aumentada SLC25A35=Disminui
Probabilidad de sobrevida
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0
2000
4000
6000
Tiempo (días)


Strata SLC47A1=Aumentada SLC47A1=Disminuid
Probabilidad de sobrevida
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0
2000
4000
6000
Tiempo (días)


Strata SORBS2=Aumentada SORBS2=Disminuida
Probabilidad de sobrevida
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0
2000
4000
6000
Tiempo (días)


Strata TAB2=Aumentada TAB2=Disminuida
Probabilidad de sobrevida
1.00
0.75
0.50
0.25
0.00
p = 0.00033
0
2000
4000
6000
Tiempo (días)


Strata TMPRSS2=Aumentada TMPRSS2=Disminui
Probabilidad de sobrevida
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0
2000
4000
6000
Tiempo (días)

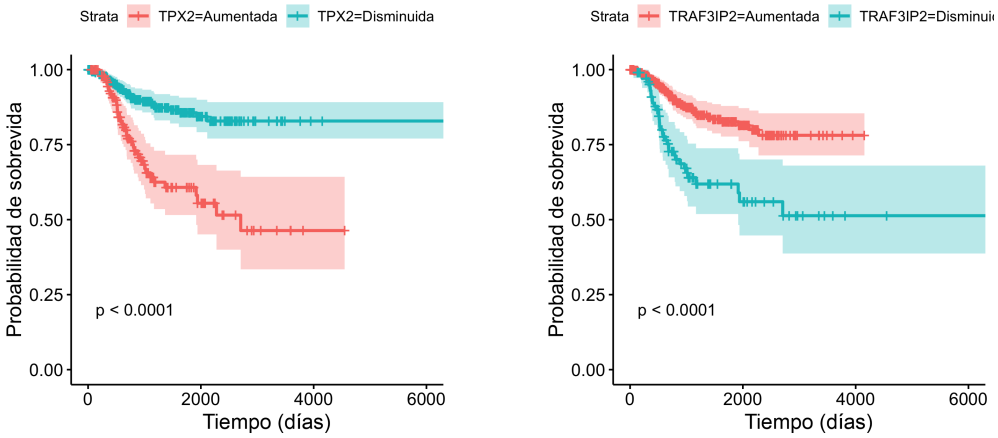


Figura F.2. Curvas de sobrevida libre de recurrencia (RFS) de los 18 genes obtenidos luego del cálculo de ORs. Los genes son: ACSL5, ANAPC4, ARSD, FAM189A2, GSPT1, LPACT2, PALMD, PDZRN3, PGR, PLEKHH1, PTCH1, SLC25A35, SLC47A1, SORBS2, TAB2, TMPRSS2, TPX2 y TRAF3IP2.

Anexo G

Resultados: *Odds Ratios* (OR) e Intervalos de Confianza (IC)

	ACSL5	**ANAPC4**	**ARSD**	**FAM189A2**
RFS *upper*	0,611	0,554	0,542	0,453
RFS *middle*	0,360	0,340	0,334	0,278
RFS *lower*	0,212	0,209	0,206	0,171
OS *upper*	0,811	0,727	0,949	0,749
OS *middle*	0,476	0,440	0,575	0,453
OS *lower*	0,280	0,266	0,348	0,273
Hist *upper*	0,301	0,341	0,358	0,225
Hist *middle*	0,181	0,220	0,232	0,144
Hist *lower*	0,109	0,142	0,151	0,092
Grado *upper*	0,293	0,227	0,234	0,225
Grado *middle*	0,200	0,147	0,148	0,139
Grado *lower*	0,136	0,095	0,094	0,086
Estadío *upper*	0,722	0,507	0,591	0,500
Estadío *middle*	0,475	0,336	0,392	0,331
Estadío *lower*	0,312	0,223	0,261	0,218
IM *upper*	0,806	0,655	0,716	0,546
IM *middle*	0,561	0,454	0,494	0,372
IM *lower*	0,390	0,314	0,340	0,253
	LPCAT2	**PALMD**	**PGR**	**PLEKHH1**
RFS *upper*	0,644	0,671	0,411	0,611

RFS *middle*	0,388	0,400	0,249	0,362
RFS *lower*	0,233	0,239	0,151	0,215
OS *upper*	0,822	0,896	0,571	0,709
OS *middle*	0,481	0,531	0,343	0,414
OS *lower*	0,281	0,315	0,206	0,242
Hist *upper*	0,222	0,364	0,126	0,437
Hist *middle*	0,139	0,225	0,075	0,275
Hist *lower*	0,087	0,139	0,045	0,173
Grado *upper*	0,251	0,478	0,153	0,352
Grado *middle*	0,139	0,331	0,094	0,241
Grado *lower*	0,077	0,229	0,057	0,165
Estadío *upper*	0,655	0,822	0,462	0,586
Estadío *middle*	0,419	0,545	0,306	0,382
Estadío *lower*	0,268	0,361	0,202	0,250
IM *upper*	0,889	0,817	0,551	0,639
IM *middle*	0,583	0,568	0,379	0,443
IM *lower*	0,382	0,396	0,261	0,307
	PTCH1	**SLC25A35**	**SLC47A1**	**SORBS2**
RFS *upper*	0,490	0,543	0,493	0,547
RFS *middle*	0,297	0,335	0,297	0,331
RFS *lower*	0,180	0,206	0,179	0,200
OS *upper*	0,652	0,917	0,413	0,793
OS *middle*	0,385	0,553	0,237	0,477
OS *lower*	0,227	0,333	0,136	0,287
Hist *upper*	0,447	0,193	0,070	0,132
Hist *middle*	0,284	0,122	0,035	0,075
Hist *lower*	0,180	0,077	0,017	0,043
Grado *upper*	0,290	0,179	0,216	0,222
Grado *middle*	0,166	0,107	0,143	0,147
Grado *lower*	0,095	0,064	0,094	0,097
Estadío *upper*	0,691	0,479	0,502	0,708
Estadío *middle*	0,442	0,316	0,331	0,471
Estadío *lower*	0,282	0,209	0,218	0,314
IM *upper*	0,849	0,634	0,610	0,790
IM *middle*	0,556	0,433	0,424	0,551
IM *lower*	0,364	0,296	0,294	0,384
	TMPRSS2	**TPX2**	**TRAF3IP2**	
RFS *upper*	0,519	6,076	0,439	
RFS *middle*	0,319	3,726	0,264	
RFS *lower*	0,196	2,285	0,159	

OS *upper*	0,902	5,642	0,506
OS *middle*	0,541	3,391	0,298
OS *lower*	0,324	2,038	0,176
Hist *upper*	0,462	1,671	0,125
Hist *middle*	0,301	1,045	0,076
Hist *lower*	0,196	6,527	0,046
Grado *upper*	0,274	2,665	0,181
Grado *middle*	0,169	1,464	0,090
Grado *lower*	0,105	8,033	0,045
Estadío *upper*	0,708	3,583	0,518
Estadío *middle*	0,466	2,363	0,329
Estadío *lower*	0,307	1,560	0,209
IM *upper*	0,724	3,380	0,626
IM *middle*	0,491	2,297	0,404
IM *lower*	0,333	1,561	0,261

Tabla 27: Resultados del análisis de OR para todas las características clínico-patológicas en estudio. La tabla representa los extremos del IC ('*lower*' y '*upper*') y la medida de OR ('*middle*'). La variable 'Hist' refiere al subtipo histológico.

Anexo H

Resultados: modelo de riesgos proporcionales de Cox

1 Variable de estado: RFS

Variable de estado: recurrencia (RFS)						
	Covariable	**b**	**HR**	**CI (inf.)**	**CI (sup.)**	**valor p**
Paso 1	TPX2	0,35	1,41	0,81	2,48	0,226
	LPCAT2	-0,03	0,97	0,58	1,62	0,901
	PTCH1	-0,46	0,63	0,38	1,05	0,079
	ACSL5	-0,26	0,77	0,44	1,34	0,361
	PALMD	0,04	1,04	0,56	1,92	0,906
	ANAPC4	-0,2	0,82	0,48	1,39	0,457
	TMPRSS2	-0,44	0,65	0,4	1,05	0,078
	ARSD	-0,13	0,87	0,53	1,45	0,601
	SLC25A35	-0,3	0,74	0,44	1,24	0,255
	SORBS2	0,02	1,02	0,57	1,83	0,934
	TRAF3IP2	-0,1	0,91	0,53	1,55	0,72
	FAM189A2	-0,34	0,71	0,4	1,25	0,236
	PGR	-0,04	0,96	0,48	1,94	0,91
	SLC47A1	-0,11	0,9	0,47	1,72	0,744
	PLEKHH1	-0,38	0,69	0,4	1,18	0,172
Paso 2	TPX2	0,34	1,41	0,81	2,47	0,227

	LPCAT2	-0,03	0,97	0,58	1,62	0,908
	PTCH1	-0,45	0,64	0,38	1,05	0,076
	ACSL5	-0,26	0,77	0,44	1,34	0,362
	PALMD	0,04	1,04	0,56	1,92	0,904
	ANAPC4	-0,2	0,82	0,48	1,39	0,46
	TMPRSS2	-0,44	0,65	0,4	1,05	0,078
	ARSD	-0,13	0,88	0,53	1,45	0,603
	SLC25A35	-0,3	0,74	0,44	1,24	0,254
	TRAF3IP2	-0,09	0,91	0,54	1,55	0,726
	FAM189A2	-0,34	0,71	0,4	1,25	0,237
	PGR	-0,04	0,96	0,48	1,94	0,915
	SLC47A1	-0,11	0,9	0,47	1,72	0,748
	PLEKHH1	-0,38	0,69	0,4	1,18	0,173
Paso 3	TPX2	0,35	1,42	0,81	2,47	0,221
	LPCAT2	-0,03	0,97	0,58	1,61	0,896
	PTCH1	-0,46	0,63	0,39	1,04	0,07
	ACSL5	-0,26	0,77	0,44	1,34	0,354
	PALMD	0,03	1,03	0,56	1,9	0,915
	ANAPC4	-0,2	0,82	0,48	1,38	0,45
	TMPRSS2	-0,44	0,64	0,4	1,04	0,074
	ARSD	-0,13	0,88	0,53	1,45	0,603
	SLC25A35	-0,3	0,74	0,44	1,24	0,253
	TRAF3IP2	-0,1	0,91	0,54	1,54	0,716
	FAM189A2	-0,35	0,7	0,41	1,22	0,209
	SLC47A1	-0,11	0,89	0,48	1,67	0,722
	PLEKHH1	-0,38	0,68	0,4	1,18	0,17
Paso 4	TPX2	0,34	1,41	0,81	2,44	0,221
	LPCAT2	-0,03	0,97	0,59	1,6	0,907
	PTCH1	-0,45	0,64	0,39	1,03	0,067
	ACSL5	-0,26	0,77	0,44	1,33	0,351
	ANAPC4	-0,2	0,82	0,49	1,38	0,454
	TMPRSS2	-0,44	0,65	0,4	1,04	0,074
	ARSD	-0,13	0,88	0,53	1,45	0,604
	SLC25A35	-0,3	0,74	0,44	1,24	0,253
	TRAF3IP2	-0,1	0,91	0,54	1,54	0,717
	FAM189A2	-0,34	0,71	0,42	1,2	0,203
	SLC47A1	-0,11	0,89	0,48	1,67	0,725
	PLEKHH1	-0,38	0,68	0,4	1,16	0,16
Paso 5	TPX2	0,34	1,41	0,82	2,44	0,218
	PTCH1	-0,45	0,64	0,39	1,03	0,067

	ACSL5	-0,26	0,77	0,44	1,33	0,349
	ANAPC4	-0,2	0,82	0,49	1,37	0,442
	TMPRSS2	-0,44	0,65	0,4	1,04	0,074
	ARSD	-0,13	0,87	0,53	1,44	0,599
	SLC25A35	-0,3	0,74	0,44	1,24	0,253
	TRAF3IP2	-0,1	0,9	0,54	1,52	0,7
	FAM189A2	-0,34	0,71	0,42	1,2	0,201
	SLC47A1	-0,12	0,89	0,48	1,66	0,712
	PLEKHH1	-0,39	0,68	0,4	1,16	0,157
Paso 6	TPX2	0,36	1,43	0,83	2,47	0,193
	PTCH1	-0,46	0,63	0,39	1,02	0,06
	ACSL5	-0,29	0,75	0,44	1,28	0,289
	ANAPC4	-0,21	0,81	0,49	1,36	0,424
	TMPRSS2	-0,43	0,65	0,4	1,04	0,075
	ARSD	-0,13	0,88	0,53	1,44	0,603
	SLC25A35	-0,33	0,72	0,44	1,19	0,198
	TRAF3IP2	-0,13	0,88	0,53	1,46	0,627
	FAM189A2	-0,36	0,7	0,42	1,18	0,177
	PLEKHH1	-0,38	0,68	0,4	1,16	0,158
Paso 7	TPX2	0,39	1,47	0,86	2,51	0,155
	PTCH1	-0,48	0,62	0,38	0,99	0,045
	ACSL5	-0,3	0,74	0,44	1,26	0,272
	ANAPC4	-0,21	0,81	0,48	1,35	0,414
	TMPRSS2	-0,45	0,64	0,4	1,03	0,064
	ARSD	-0,13	0,88	0,53	1,45	0,608
	SLC25A35	-0,34	0,72	0,44	1,17	0,184
	FAM189A2	-0,37	0,69	0,41	1,15	0,155
	PLEKHH1	-0,39	0,68	0,4	1,15	0,151
Paso 8	TPX2	0,42	1,52	0,9	2,56	0,117
	PTCH1	-0,49	0,61	0,38	0,98	0,043
	ACSL5	-0,31	0,73	0,43	1,24	0,245
	ANAPC4	-0,22	0,8	0,48	1,34	0,394
	TMPRSS2	-0,47	0,63	0,39	1	0,05
	SLC25A35	-0,34	0,71	0,43	1,17	0,175
	FAM189A2	-0,39	0,68	0,41	1,14	0,141
	PLEKHH1	-0,39	0,67	0,4	1,15	0,147
Paso 9	TPX2	0,45	1,57	0,93	2,63	0,09
	PTCH1	-0,49	0,61	0,38	0,98	0,042
	ACSL5	-0,34	0,71	0,42	1,21	0,207
	TMPRSS2	-0,48	0,62	0,39	0,99	0,047

	SLC25A35	-0,39	0,68	0,42	1,1	0,116
	FAM189A2	-0,4	0,67	0,4	1,12	0,13
	PLEKHH1	-0,45	0,64	0,38	1,08	0,092
Paso 10	TPX2	0,5	1,64	0,98	2,76	0,06
	PTCH1	-0,51	0,6	0,37	0,96	0,033
	TMPRSS2	-0,49	0,61	0,38	0,98	0,042
	SLC25A35	-0,45	0,64	0,39	1,04	0,07
	FAM189A2	-0,41	0,66	0,39	1,11	0,119
	PLEKHH1	-0,44	0,64	0,38	1,08	0,094
Paso 11	TPX2	0,62	1,86	1,13	3,06	0,014
	PTCH1	-0,56	0,57	0,36	0,91	0,02
	TMPRSS2	-0,58	0,56	0,35	0,89	0,013
	SLC25A35	-0,48	0,62	0,38	1	0,049
	PLEKHH1	-0,48	0,62	0,37	1,04	0,068
Paso 12	TPX2	0,62	1,86	1,13	3,07	0,015
	PTCH1	-0,6	0,55	0,34	0,88	0,013
	TMPRSS2	-0,66	0,52	0,33	0,82	0,005
	SLC25A35	-0,58	0,56	0,35	0,9	0,017

Tabla 28: Detalle de cada paso del análisis multivariado de Cox con la variable de estado RFS. Las **covariables** son los genes obtenidos luego del cálculo de OR (Sección 1.6.1, Resultados); la columna **b** indica los coeficientes de la Ecuación 1.1; **HR**, los *hazard ratios*, que son iguales a exp(b); **IC**, los intervalos de confianza definidos por el límite inferior y superior; por último, la columna **valor p** muestra la significancia estadística de los resultados de b y HR.

2 Variable de estado: OS

Variable de estado: sobrevida (OS)						
	Covariable	**b**	**HR**	**IC (inf.)**	**IC (sup.)**	**valor p**
Paso 1	TPX2	0,66	1,93	1,04	3,6	0,037
	PTCH1	-0,27	0,77	0,44	1,34	0,351
	ACSL5	0,05	1,05	0,59	1,86	0,872
	PALMD	0,09	1,09	0,58	2,04	0,785
	TMPRSS2	-0,21	0,81	0,47	1,39	0,438
	SLC25A35	0,12	1,13	0,66	1,92	0,666
	SORBS2	0,24	1,26	0,69	2,32	0,449
	TRAF3IP2	-0,18	0,84	0,47	1,51	0,557
	FAM189A2	0,04	1,04	0,58	1,88	0,893
	PGR	-0,06	0,95	0,45	1,97	0,882
	SLC47A1	-0,98	0,38	0,19	0,76	0,006

Paso 2	TPX2	0,66	1,93	1,04	3,56	0,037
	PTCH1	-0,27	0,76	0,44	1,34	0,345
	ACSL5	0,05	1,05	0,59	1,85	0,874
	PALMD	0,09	1,1	0,59	2,04	0,767
	TMPRSS2	-0,21	0,81	0,47	1,39	0,445
	SLC25A35	0,12	1,13	0,66	1,93	0,66
	SORBS2	0,24	1,27	0,69	2,32	0,448
	TRAF3IP2	-0,17	0,84	0,47	1,51	0,562
	PGR	-0,04	0,96	0,47	1,95	0,906
	SLC47A1	-0,98	0,38	0,19	0,76	0,006
Paso 3	TPX2	0,66	1,94	1,06	3,55	0,032
	PTCH1	-0,27	0,76	0,44	1,32	0,331
	ACSL5	0,04	1,05	0,59	1,85	0,878
	PALMD	0,09	1,09	0,59	2	0,779
	TMPRSS2	-0,22	0,81	0,48	1,36	0,42
	SLC25A35	0,12	1,12	0,66	1,91	0,668
	SORBS2	0,23	1,26	0,69	2,28	0,452
	TRAF3IP2	-0,18	0,84	0,47	1,5	0,554
	SLC47A1	-0,99	0,37	0,19	0,73	0,004
Paso 4	TPX2	0,66	1,93	1,06	3,54	0,032
	PTCH1	-0,27	0,76	0,44	1,32	0,335
	PALMD	0,09	1,09	0,59	2	0,783
	TMPRSS2	-0,21	0,81	0,48	1,36	0,423
	SLC25A35	0,12	1,13	0,66	1,91	0,663
	SORBS2	0,23	1,26	0,7	2,28	0,443
	TRAF3IP2	-0,18	0,84	0,47	1,5	0,552
	SLC47A1	-0,97	0,38	0,2	0,72	0,003
Paso 5	TPX2	0,64	1,89	1,06	3,36	0,031
	PTCH1	-0,25	0,78	0,45	1,33	0,353
	TMPRSS2	-0,21	0,81	0,48	1,37	0,439
	SLC25A35	0,12	1,13	0,67	1,92	0,644
	SORBS2	0,24	1,27	0,7	2,29	0,429
	TRAF3IP2	-0,17	0,84	0,47	1,51	0,561
	SLC47A1	-0,97	0,38	0,2	0,72	0,003
Paso 6	TPX2	0,61	1,85	1,05	3,27	0,035
	PTCH1	-0,28	0,76	0,45	1,28	0,302
	TMPRSS2	-0,19	0,83	0,49	1,38	0,469
	SORBS2	0,25	1,28	0,71	2,31	0,412
	TRAF3IP2	-0,17	0,85	0,47	1,51	0,574

	SLC47A1	-0,92	0,4	0,22	0,73	0,003
Paso 7	TPX2	0,64	1,9	1,09	3,32	0,024
	PTCH1	-0,3	0,74	0,44	1,25	0,261
	TMPRSS2	-0,22	0,8	0,49	1,33	0,395
	SORBS2	0,21	1,24	0,7	2,19	0,47
	SLC47A1	-0,95	0,39	0,21	0,7	0,002
Paso 8	TPX2	0,58	1,79	1,05	3,06	0,032
	PTCH1	-0,27	0,77	0,46	1,28	0,31
	TMPRSS2	-0,19	0,83	0,51	1,36	0,458
	SLC47A1	-0,91	0,4	0,22	0,73	0,003
Paso 9	TPX2	0,62	1,86	1,1	3,15	0,021
	PTCH1	-0,29	0,75	0,45	1,25	0,268
	SLC47A1	-0,93	0,4	0,22	0,71	0,002
Paso 10	TPX2	0,69	1,99	1,19	3,32	0,008
	SLC47A1	-0,98	0,38	0,21	0,67	0,001

Tabla 29: Detalle de cada paso del análisis multivariado de Cox con la variable de estado OS. Las **covariables** son los genes obtenidos luego del cálculo de OR (Sección 1.6.1, Resultados); la columna **b** indica los coeficientes de la Ecuación 1.1; **HR**, los *hazard ratios*, que son iguales a exp(b); **IC**, los intervalos de confianza definidos por el límite inferior y superior; por último, la columna **valor p** muestra la significancia estadística de los resultados de b y HR.

Bibliografía

1. Geoffrey M. Cooper, R. E. H. *The cell: A molecular approach* (2013).
2. National Institutes of Health (EEUU). *Understanding cancer.* `https://www.ncbi.nlm.nih.gov/books/NBK20362` (2007).
3. Hanahan, D. & Weinberg, R. Hallmarks of cancer: The next generation. *Cell* **144,** 646–674 (2011).
4. Coleman, B, W., Tsongalis & J, G. Molecular mechanisms of human carcinogenesis. *EXS,* 321–349 (2006).
5. Sheikh, A. *et al.* The spectrum of genetic mutations in breast cancer. *Asian Pacific Journal of Cancer Prevention* **16,** 2177–2185 (2015).
6. Biemar, F. & Foti, M. Global progress against cancer-challenges and opportunities. *Cancer biology & medicine* **10,** 183–186 (2013).
7. *Data query, Organización Mundial de la Salud, Naciones Unidas.* Último acceso: 19/4/2019. `https://www.who.int/cancer`.
8. *El cáncer en números, Instituto Nacional de Cáncer.* Último acceso: 19/4/2019. `https://www.argentina.gob.ar/salud/inc`.
9. *Cancer today, Global cancer observatory (Globocan) 2018, International Agency for Research on Cancer 2019.* Último acceso: 22/4/2019. `https://gco.iarc.fr/`.
10. Colombo, N., Preti, E., Landoni, F., S. Carinelli, A. C. & Sessa, C. M. C̃. Endometrial cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Annals of Oncology* **24** (2013).

11. *Physicians data query (PDQ), National Cancer Institute.* Último acceso: 22/4/2019. `https://www.cancer.gov/espanol/publicaciones/pdq`.

12. Levine, D. A. & TCGA. Integrated genomic characterization of endometrial carcinoma. *Nature* **497,** 67–73 (2013).

13. Consenso nacional inter-sociedades sobre cáncer de endometrio. *Programa Nacional de Consensos Inter-Sociedades* (2016).

14. You, W. & Henneberg, M. Cancer incidence increasing globally: The role of relaxed natural selection. *Evolutionary Applications* **11,** 140–152 (2017).

15. Ponce, J., Torrejón, R. & Barahoma, M. Oncoguía SEGO: Cáncer de endometrio. Guías de práctica clínica en cáncer ginecológico y mamario. *Obstetricia SedGy, Publicaciones SEGO.* (2010).

16. Vaccarella, S. *et al.* Reducing social inequalities in cancer: Setting priorities for research. *CA: A Cancer Journal for Clinicians* **68,** 324–326 (2018).

17. Cramer, D. W. The epidemiology of endometrial and ovarian cancer. *Hematology/Oncology Clinics of North America* **26,** 1–12 (2012).

18. Besso, M. J. *Uso de herramientas bioinformáticas, moleculares y funcionales en el estudio de las bases moleculares de la progresión tumoral y en la identificación de biomarcadores del cáncer de endometrio.* PhD thesis (2018).

19. PDQ Screening and Prevention editorial Board. *Endometrial cancer prevention.* Published online. (2019).

20. Weiderpass, E. *et al.* Risk of endometrial cancer following estrogen replacement with and without progestins. *JNCI: Journal of the National Cancer Institute* **91,** 1131–1137 (1999).

21. Colombo, N. & Creutzberg, C. ESMO-ESGO-ESTRO consensus conference on endometrial cancer diagnosis, treatment and follow-up. *International Journal of Gynecological Cancer* **24** (2016).

22. Sorosky, J. I. Endometrial cancer. *Obstetrics & Gynecology* **120,** 383–397 (2012).

23. Raby, T. *et al.* Capacidad diagnóstica de la ecografía para detectar cáncer de endometrio en mujeres posmenopaúsicas sintomáticas y asintomáticas. *Revista chilena de obstetricia y ginecología* (2014).

24. Talhouk, A. & McAlpine, J. N. New classification of endometrial cancers: The development and potential applications of genomic-based classification in research and clinical care. *Gynecologic Oncology Research and Practice* **3** (2016).

25. Amant, F., Mirza, M. R., Koskas, M. & Creutzberg, C. L. Cancer of the corpus uteri. *International Journal of Gynecology & Obstetrics* **143,** 37–50 (2018).

26. Scholten, A. N., Smit, V. T. H. B. M., Beerman, H., van Putten, W. L. J. & Creutzberg, C. L. Prognostic significance and interobserver variability of histologic grading systems for endometrial carcinoma. *Cancer* **100,** 764–772 (2004).

27. Kang, S. *et al.* Preoperative assessment of lymph node metastasis in endometrial cancer: A Korean Gynecologic Oncology Group study. *Cancer* **123,** 263–272 (2016).

28. Sánchez, M. *et al.* Diagnóstico preoperatorio de invasión miometrial con resonancia magnética y estudio intraoperatorio por congelación en pacientes con cáncer de endometrio. *Servicios de Anatomía Patológica, Ginecología Oncológica y Diagnóstico por Imágenes. Hospital Italiano de Buenos Aires* (2014).

29. Sasada, S. *et al.* Baseline risk of recurrence in stage I-II endometrial carcinoma. *Journal of Gynecologic Oncology* **29** (2018).

30. Van der Steen-Banasik, E. Primary brachytherapy as a radical treatment for endometrial carcinoma. *Journal of Contemporary Brachytherapy* **1,** 106–112 (2014).

31. Rungruang, B. & Olawaiye, A. B. Comprehensive surgical staging for endometrial cancer. *Reviews in obstetrics & gynecology* **5,** 28–34 (2012).

32. Frost, J. A., Webster, K. E., Bryant, A. & Morrison, J. Lymphadenectomy for the management of endometrial cancer. *Cochrane Database of Systematic Reviews* (2017).

33. Meyer, L. A., Bohlke, K. & Wright, A. A. Postoperative radiation therapy for endometrial cancer: American society of clinical oncology clinical practice guideline endorsement of

the american society for radiation oncology evidence-based guideline. *Journal of Oncology Practice* **12,** 182–185 (2016).

34. Trujillo, C. Conceptos básicos de oncología. *Medwave* **3** (2003).

35. Werner, H. M. J. *Clinical and molecular markers in endometrial cancer* MA thesis (2014).

36. Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, National Institutes of Health & Department of Health & Human Services. *Help me understand genetics: How genes work.* Genetics Home Reference. 2019.

37. Goldman, L. & Schafer, A. I. *Goldman-Cecil medicine* (2016).

38. Alberts, B. *et al. Molecular biology of the cell* (2008).

39. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270,** 467–470 (1995).

40. Marzancola, M. G., Sedighi, A. & Li, P. C. H. DNA microarray-based diagnostics. *Methods in Molecular Biology* **1368,** 161–178 (2016).

41. Chu, Y. & Corey, D. R. RNA Sequencing: Platform selection, experimental design, and data interpretation. *Nucleic Acid Therapeutics* **22,** 271–274 (2012).

42. Wilhelm, B. T. *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453,** 1239–1243 (2008).

43. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320,** 1344–1349 (2008).

44. Kukurba, K. R. & Montgomery, S. B. RNA sequencing and analysis. *Cold Spring Harbor Protocols* **2015** (2015).

45. López-Maury, L., Marguerat, S. & Bähler, J. Tuning gene expression to changing environments: From rapid responses to evolutionary adaptation. *Nature Reviews Genetics* **9,** 583–593 (2008).

46. Kim, N. & Jinks-Robertson, S. Transcription as a source of genome instability. *Nature Reviews Genetics* **13,** 204–214 (2012).

47. Ko, J. Y., Oh, S. & Yoo, K. H. Functional enhancers as master regulators of tissue-specific gene regulation and cancer development. *Molecules and cells* **40,** 169–177 (2017).

48. Rapley, R. *Basic techniques in molecular biology* 1–12 (2005).

49. Walker, J. M. & Rapley, R. *Molecular biomethods handbook* (Humana Press, 2008).

50. Jenkins, G. J. S. *et al.* Restriction enzymes in the analysis of genetic alterations responsible for cancer progression. *British Journal of Surgery* **89,** 8–20 (Jan. 2002).

51. Garibyan, L. & Avashia, N. Polymerase Chain Reaction. *Journal of Investigative Dermatology* **133,** 1–4 (2013).

52. Castro, J. M. P., Ramírez, O. G. & Figueroa, B. E. B. Los métodos experimentales que permiten el estudio de las macromoléculas de la vida: historia, fundamentos y perspectivas. *Educación química* **24** (2013).

53. Smith, A. *Oxford dictionary of biochemistry and molecular biology* (2000).

54. Maddocks, S. & Jenkins, R. in *Understanding PCR,* 45–52 (2017).

55. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107,** 1–8 (2016).

56. Metzker, M. L. Emerging technologies in DNA sequencing. *Genome Research* **15,** 1767–1776 (2005).

57. Ulrich, A. B. & Pour, P. M. *Cell Lines* 310–311 (2001).

58. Uysal, O., Sevimli, T., Sevimli, M., Gunes, S. & Sariboyaci, A. E. *Cell and tissue culture* 391–429 (2018).

59. Geraghty, R. J. *et al.* Guidelines for the use of cell lines in biomedical research. *British Journal of Cancer* **111,** 1021–1046 (2014).

60. Maru, Y., Tanaka, N., Itami, M. & Hippo, Y. Efficient use of patient-derived organoids as a preclinical model for gynecologic tumors. *Gynecologic Oncology* (2019).

61. Casbas-Hernandez, P., Fleming, J. M. & Troester, M. A. Gene expression analysis of in vitro cocultures to study interactions between breast epithelium and stroma. *Journal of Biomedicine and Biotechnology* **2011,** 1–12 (2011).

62. Ya, Z., Hailemichael, Y., Overwijk, W. & Restifo, N. P. Mouse model for pre-clinical study of human cancer immunotherapy. *Current protocols in immunology* **108,** 20.1.1–20.143 (2015).

63. Recuenco, S., Warnock, E., Osinubi, M. O. V. & Rupprecht, C. E. A single center, open label study of intradermal administration of an inactivated purified chick embryo cell culture rabies virus vaccine in adults. *Vaccine* **35,** 4315–4320 (2017).

64. Schulze, S., Wehrum, D., Dieter, P. & Hempel, U. A supplement-free osteoclast-osteoblast co-culture for pre-clinical application. *Journal of cellular physiology* **233,** 4391–4400 (2018).

65. Zhang, S. Cell isolation and culture. *WormBook,* 1–39 (2013).

66. Baust, J. M. *et al.* Best practices in cell culture: An overview. *In Vitro Cellular & Developmental Biology - Animal* **53,** 669–672 (2017).

67. Honegger, P. Overview of cell and tissue culture techniques. *Current protocols in pharmacology* **4,** 12.1.1–12.1.12 (1999).

68. Zhou, X., Wang, Z., Ying, Z., Podratz, K. & Jiang, S. Characterization of sixteen endometrial cancer cell lines. *Cancer Research* **67,** 3870–3870 (2007).

69. Colas, E. *et al.* ETV5 cooperates with LPP as a sensor of extracellular signals and promotes EMT in endometrial carcinomas. *Oncogene* **31,** 4778–4788 (2012).

70. Planagumà, J. *et al.* Up-regulation of ERM/ETV5 correlates with the degree of myometrial infiltration in endometrioid endometrial carcinoma. *The Journal of Pathology* **207,** 422–429 (2005).

71. Doll, A. *et al.* Novel molecular profiles of endometrial cancer-new light through old windows. *The Journal of Steroid Biochemistry and Molecular Biology* **108,** 221–229 (2008).

72. Planagumà, J. *et al.* Matrix metalloproteinase-2 and matrix metalloproteinase-9 codistribute with transcription factors RUNX1/AML1 and ETV5/ERM at the invasive front of endometrial and ovarian carcinoma. *Human Pathology* **42,** 57–67 (2011).

73. Can, T. *Introduction to bioinformatics,* 51–71 (2013).

74. Bayat, A. Science, medicine, and the future: Bioinformatics. *BMJ* **324,** 1018–1022 (2002).

75. Tenenbaum, J. D. Translational bioinformatics: Past, present, and future. *Genomics, proteomics & bioinformatics* **14,** 31–41 (2016).

76. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409** (2001).

77. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431** (2004).

78. Fox, J. The human genome project: The impact of genome sequencing technology on human health. *The Science Creative Quarterly* (2006).

79. U.S. Department of Energy. Genomics and its impact on science and society: A 2008 primer. *Human Genome Program* (2008).

80. Collins, F. S., Morgan, M. & Patrinos, A. The human genome project: Lessons from large-scale biology. *Building on the DNA revolution* **300,** 286–290 (2003).

81. Ayday, E., Cristofaro, E. D., Hubaux, J.-P. & Tsudik, G. The chills and thrills of whole genome sequencing. *IEEE Computer Magazine* (2013).

82. Gazal, S., Sahbatou, M., Babron, M.-C., Génin, E. & Leutenegger, A.-L. High level of inbreeding in final phase of 1000 Genomes Project. *Scientific Reports* **5** (2015).

83. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biology* **18** (2017).

84. Clegg, A. & Shepherd, A. Text mining. *Methods Mol. Bio.* 471–491 (2008).

85. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* **32,** 258D–261 (2004).

86. Lipscomb, C. E. Medical Subject Headings (MeSH). *Journal of the Medical Library Association* (2000).

87. Sumathi, S. & Sivanandam, S. N. *Introduction to data mining and its applications* (2006).

88. Hashemi, F. S. G. *et al.* Intelligent mining of large-scale bio-data: Bioinformatics applications. *Biotechnology & Biotechnological Equipment* **32,** 10–29 (2017).

89. Salzberg, S. L. Open questions: How many genes do we have? *BMC Biology* **16** (2018).

90. Committee on the Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials *et al.* Evolution of translational omics: Lessons learned and the path forward. *Washington, DC: The National Academies Press* (2012).

91. World Health Organization & International Programme on Chemical Safety. *Biomarkers in risk assessment: Validity and validation* (2001).

92. Goossens, N., Nakagawa, S., Sun, X. & Hoshida, Y. Cancer biomarker discovery and validation. *Translational cancer research* **4,** 256–269 (2015).

93. Gosho, M., Nagashima, K. & Sato, Y. Study designs and statistical analyses for biomarker research. *Sensors* **12,** 8966–8986 (2012).

94. Mamatjan, Y. *et al.* Molecular Signatures for Tumor Classification. *An Analysis of The Cancer Genome Atlas Data* **19,** 881–891 (2017).

95. Wang, Z. *et al.* A six-gene-based prognostic signature for hepatocellular carcinoma overall survival prediction. *Life Sciences* **203,** 83–91 (2018).

96. Han, B. *et al.* Identification and interaction analysis of molecular markers in colorectal cancer by integrated bioinformatics analysis. *Medical Science Monitor* **24,** 6059–6069 (2018).

97. Mihály, Z. *et al.* A meta-analysis of gene expression-based biomarkers predicting outcome after tamoxifen treatment in breast cancer. *Breast Cancer Research and Treatment* **140,** 219–232 (2013).

98. Fröhlich, H. *et al.* From hype to reality: Data science enabling personalized medicine. *BMC Medicine* **16** (2018).

99. Kass, R. E. *et al.* Ten Simple Rules for Effective Statistical Practice. *PLOS Computational Biology* **12,** e1004961 (2016).

100. Horton, N. J., Baumer, B. S. & Wickham, H. Setting the stage for data science: Integration of data management skills in introductory and second courses in statistics. *Chance* (2015).

101. Brust, A. V. *Ciencia de Datos: Una introducción a la exploración, análisis y visualización de datos* (2019).

102. Choi, K. R., Ryu, J. Y. & Lee, S. Y. Revisiting statistical design and analysis in scientific research. *Small* **14,** 1802604 (2018).

103. Pagano, M. & Gauvreau, K. *Principles of biostatistics* (2018).

104. Touvier, M. *et al.* Association between prediagnostic biomarkers of inflammation and endothelial function and cancer risk: A nested case-control study. *American Journal of Epidemiology* **177,** 3–13 (2012).

105. Reeves, K. W. *et al.* Urinary phthalate biomarker concentrations and postmenopausal breast cancer risk. *JNCI: Journal of the National Cancer Institute* (2019).

106. Pepe, M. S. Limitations of the Odds Ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology* **159,** 882–890 (2004).

107. Park, J. H., Watt, D. G., Roxburgh, C. S. D., Horgan, P. G. & McMillan, D. C. Colorectal cancer, systemic inflammation, and outcome. *Annals of Surgery* **263,** 326–336 (2016).

108. Oue, N. *et al.* Signal peptidase complex 18, encoded by SEC11A, contributes to progression via TGF-α secretion in gastric cancer. *Oncogene* **33,** 3918–3926 (2013).

109. Mongre, R. *et al.* Prognostic and clinicopathological significance of SERTAD1 in various types of cancer risk: A systematic review and retrospective analysis. *Cancers* **11,** 337 (2019).

110. Townsend, M. H. *et al.* Potential new biomarkers for endometrial cancer. *Cancer Cell International* **19** (2019).

111. Norton, E. C., Dowd, B. E. & Maciejewski, M. L. Odds Ratios-current best practice and use. *JAMA* **320,** 84 (2018).

112. Kaplan, E. L. & Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53,** 457–481 (1958).

113. Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society* **34,** 187–220 (1972).

114. Mendolia, F., Klein, J. P., Petersdorf, E. W., Malkki, M. & Wang, T. Comparison of statistics in association tests of genetic markers for survival outcomes. *Statistics in Medicine* **33,** 828–844 (2013).

115. Del Val, E. B. *El modelo de regresión de Cox* (2017).

116. Álvarez, P. V. *Modelo de regresión de Cox y sus aplicaciones biosanitarias* (2015).

117. Feise, R. Do multiple outcome measures require p-value adjustment? *BMC Medical Research Methodology* **2** (2002).

118. Lee, S. & Lee, D. K. What is the proper way to apply the multiple comparison test? *Korean Journal of Anesthesiology* **71,** 353–360 (2018).

119. Piñero, J. *et al.* DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research* **45,** D833–D839 (2016).

120. Pinero, J. *et al.* DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes. *Database* **2015,** bav028–bav028 (2015).

121. Edgar, R., Domrachev, M. & Lash, A. E. L. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* **30,** 207–210 (2002).

122. Barrett, T. *et al.* NCBI GEO: Archive for functional genomics data sets-update. *Nucleic Acids Research* **41,** D991–D995 (2012).

123. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research* **37,** W305–W311 (2009).

124. Tomczak, K., Czerwinska, P. & Wiznerowicz, M. Review the Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Wspolczesna Onkologia* **1A,** 68–77 (2015).

125. Pontén, F., Jirström, K. & Uhlen, M. The Human protein Atlas (HPA)-a tool for pathology. *The Journal of Pathology* **216,** 387–393 (2008).

126. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43,** e47–e47 (2015).

127. Smyth, G. *et al. limma* 2017.

128. Day, R. S. *et al.* Identifier mapping performance for integrating transcriptomics and proteomics experimental results. *BMC bioinformatics* **12,** 213 (2011).

129. Budczies, J. *et al.* Cutoff Finder: A comprehensive and straightforward web application enabling rapid biomarker cutoff optimization. *PLoS ONE* **7,** e51862 (2012).

130. Tomita, M., Ayabe, T., Maeda, R. & Nakamura, K. The inflammatory prognostic index predicts cancer-specific outcomes of patients with resected non-small cell lung cancer. *Asian Pacific Journal of Cancer Prevention* **19** (2018).

131. Dupre, A. *et al.* Preoperative leucocyte-based inflammatory scores in patients with colorectal liver metastases: Can we count on them? *World Journal of Surgery* **43,** 1351–1359 (2019).

132. Holländer, N., Sauerbrei, W. & Schumacher, M. Confidence intervals for the effect of a prognostic factor after selection of an 'optimal' cutpoint. *Statistics in Medicine* **23,** 1701–1713 (2004).

133. Camacho-Urkaray, E. *et al.* Establishing cut-off points with clinical relevance for bcl-2, cyclin D1, p16, p21, p27, p53, Sox11 and WT1 expression in glioblastoma - a short report. *Cellular Oncology* **41,** 213–221 (2017).

134. Frumovitz, M. *et al.* Predictors of final histology in patients with endometrial cancer. *Gynecologic Oncology* **95,** 463–468 (2004).

135. Neubauer, N. L. *et al.* The role of lymphadenectomy in the management of preoperative grade 1 endometrial carcinoma. *Gynecologic Oncology* **112,** 511–516 (2009).

136. Seracchioli, R. *et al.* Controversies in surgical staging of endometrial cancer. *Obstetrics and Gynecology International* **2010,** 1–8 (2010).

137. Body, N. *et al.* Are preoperative histology and MRI useful for classification of endometrial cancer risk? *BMC Cancer* **16** (2016).

138. Carmen, M. G. D., Boruta, D. M. & Schorge, J. O. Recurrent endometrial cancer. *Clinical Obstetrics and Gynecology* **54,** 266–277 (2011).

139. Jeppesen, M. M., Jensen, P. T., Hansen, D. G., Iachina, M. & Mogensen, O. The nature of early-stage endometrial cancer recurrence-A national cohort study. *European Journal of Cancer* **69,** 51–60 (2016).

140. Akhtar, M., Hyassat, S. A., Elaiwy, O., Rashid, S. & Nabet, A. D. M. H. A. Classification of endometrial carcinoma. *Advances In Anatomic Pathology* **26,** 421–427 (2019).

141. Lupini, L. *et al.* Molecular biomarkers predicting early development of endometrial carcinoma: A pilot study. *European Journal of Cancer Care* **28** (2019).

142. Moroney, M. R. *et al.* Molecular markers in recurrent stage I, grade 1 endometrioid endometrial cancers. *Gynecologic Oncology* **153,** 517–520 (2019).

143. Prendergast, E. N. *et al.* Comprehensive genomic profiling of recurrent endometrial cancer: Implications for selection of systemic therapy. *Gynecologic Oncology* **154,** 461–466 (2019).

144. Yoshie, H. *et al.* A bioinformatics-to-clinic sequential approach to analysis of prostate cancer biomarkers using TCGA datasets and clinical samples: A new method for precision oncology? *Oncotarget* **8** (2017).

145. He, Z., Duan, X. & Zeng, G. Identification of potential biomarkers and pivotal biological pathways for prostate cancer using bioinformatics analysis methods. *PeerJ* **7,** e7872 (2019).

146. Song, Z. *et al.* The identification of potential biomarkers and biological pathways in prostate cancer. *Journal of Cancer* **10,** 1398–1408 (2019).

147. Deng, J.-L., Xu, Y.-h. & Wang, G. Identification of potential crucial genes and key pathways in breast cancer using bioinformatic analysis. *Frontiers in Genetics* **10** (2019).

148. Liu, F. *et al.* Identification of core genes and potential molecular mechanisms in breast cancer using bioinformatics analysis. *Pathology - Research and Practice* **215,** 152436 (2019).

149. Zhai, Q., Li, H., Sun, L., Yuan, Y. & Wang, X. Identification of differentially expressed genes between triple and non-triple-negative breast cancer using bioinformatics analysis. *Breast Cancer* **26,** 784–791 (2019).

150. Cao, L. *et al.* Identification of hub genes and potential molecular mechanisms in gastric cancer by integrated bioinformatics analysis. *PeerJ* **6,** e5180 (2018).

151. Wu, Q. *et al.* Integrated bioinformatics analysis reveals novel key biomarkers and potential candidate small molecule drugs in gastric cancer. *Pathology - Research and Practice* **215,** 1038–1048 (2019).

152. Costas, L. *et al.* Defining a mutational signature for endometrial cancer screening and early detection. *Cancer Epidemiology* **61,** 129–132 (2019).

153. O'Mara, T. A., Glubb, D. M., Kho, P. F., Thompson, D. J. & Spurdle, A. B. Genome-wide association studies of endometrial cancer: Latest developments and future directions. *Cancer Epidemiology Biomarkers & Prevention* **28,** 1095–1102 (2019).

154. Shen, L., Liu, M., Liu, W., Cui, J. & Li, C. Bioinformatics analysis of RNA sequencing data reveals multiple key genes in uterine corpus endometrial carcinoma. *Oncology Letters* (2018).

155. Song, Y., Chen, Q.-T. & He, Q.-Q. Identification of key transcription factors in endometrial cancer by systems bioinformatics analysis. *Journal of Cellular Biochemistry* **120,** 15443–15454 (2019).

156. Zang, Y. *et al.* Bioinformatics analysis of key differentially expressed genes in well and poorly differentiated endometrial carcinoma. *Molecular Medicine Reports* (2018).

157. Zhang, K. *et al.* Identification of key genes and pathways between type I and type II endometrial cancer using bioinformatics analysis. *Oncology Letters* (2019).

158. Woo, S. & Henderson, D. Dichotomization of continuous biomarkers. *Axio research, Whitepaper* (2015).

159. Thierer, J. *Medidas de asociación y efecto: Nociones de análisis de regresión, univariado y multivariado* 2012.

160. Gonnissen, A. *et al.* Patched 1 expression correlates with biochemical relapse in high-risk prostate cancer patients. *The American Journal of Pathology* **188,** 795–804 (2018).

161. Szkandera, J., Kiesslich, T., Haybaeck, J., Gerger, A. & Pichler, M. Hedgehog signaling pathway in ovarian cancer. *International Journal of Molecular Sciences* **14,** 1179–1196 (2013).

162. Fangning, W. *et al.* Identification and validation of soluble carrier family expression signature for predicting poor outcome of renal cell carcinoma. *Journal of Cancer* **8,** 2010–2017 (2017).

163. Clinckemalie, L. *et al.* Androgen regulation of the TMPRSS2 gene and the effect of a SNP in an androgen response element. *Molecular Endocrinology* **27,** 2028–2040 (2013).

164. Jiang, Y., Liu, Y., Tan, X., Yu, S. & Luo, J. TPX2 as a novel prognostic indicator and promising therapeutic target in triple-negative breast cancer. *Clinical Breast Cancer* (2019).

165. Zou, J. *et al.* Overexpression of TPX2 is associated with progression and prognosis of prostate cancer. *Oncology Letters* (2018).

166. Cai, Y. *et al.* Identification of five hub genes as monitoring biomarkers for breast cancer metastasis in silico. *Hereditas.* **156** (2019).

167. Cheng, J. The TPX2 gene is a promising diagnostic and therapeutic target for cervical cancer. *Oncology Reports* (2012).

168. Hsu, P.-K. *et al.* TPX2 expression is associated with cell proliferation and patient outcome in esophageal squamous cell carcinoma. *Journal of Gastroenterology* **49,** 1231–1240 (2013).

169. Yan, L. *et al.* Target protein for Xklp2 (TPX2), a microtubule-related protein, contributes to malignant phenotype in bladder carcinoma. *Tumor Biology* **34,** 4089–4100 (2013).

170. Jiang, P. *et al.* TPX2 regulates tumor growth in human cervical carcinoma cells. *Molecular Medicine Reports* **9,** 2347–2351 (2014).

171. Huang, Y., Guo, W. & Kan, H. TPX2 is a prognostic marker and contributes to growth and metastasis of human hepatocellular carcinoma. *International Journal of Molecular Sciences* **15,** 18148–18161 (2014).

172. Zou, Z. *et al.* TPX2 level correlates with cholangiocarcinoma cell proliferation, apoptosis, and EMT. *Biomedicine & Pharmacotherapy* **107,** 1286–1293 (2018).

173. Liu, Y., Hua, T., Chi, S. & Wang, H. Identification of key pathways and genes in endometrial cancer using bioinformatics analyses. *Oncology Letters* (2019).

174. Jiang, T. *et al.* MiR-29a-5p inhibits proliferation and invasion and induces apoptosis in endometrial carcinoma via targeting TPX2. *Cell Cycle* **17,** 1268–1278 (2018).

175. Asteriti, I. A., Rensen, W. M., Lindon, C., Lavia, P. & Guarguaglini, G. The Aurora-A/TPX2 complex: A novel oncogenic holoenzyme? *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* **1806,** 230–239 (2010).