Universitat Oberta de Catalunya

uoc.edu

Identificación de marcadores genéticos predictores de respuesta al tratamiento endocrino en pacientes con cáncer de mama luminal A.

Elena de Rafael González

Máster en bioinformática y bioestadística

Estudios genéticos de enfermedades humanas.

Consultora: Helena Brunel Montaner

Profesor responsable de la asignatura: Antoni Pérez Navarro

Enero 2022

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada 3.0 España de Creative Commons

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Identificación de marcadores genéticos predictores de respuesta al tratamiento endocrino en pacientes con cáncer de mama luminal A.*
Nombre del autor:	*Elena de Rafael González*
Nombre del consultor/a:	*Helena Brunel Montaner*
Nombre del PRA:	*Antoni Pérez navarro*
Fecha de entrega (mm/aaaa):	12/2021
Titulación:	*Máster en Bioinformática y Bioestadística*
Idioma del trabajo:	Castellano
Número de créditos:	15 créditos
Palabras clave	*Cáncer de mama luminal A, tratamiento endocrino, marcadores genéticos.*
Resumen del Trabajo.	

El cáncer de mama es la enfermedad oncológica más frecuente a nivel mundial en población femenina. Existen diferentes subtipos en función de los receptores presentes en el tejido tumoral. El luminal A es el de mejor pronóstico y menor riesgo de recaídas. Su tratamiento principalmente es endocrino.

Este trabajo se centra en estos pacientes: Cáncer de mama luminal A en tratamiento endocrino sin quimioterapia. El objetivo es buscar marcadores génicos que predigan una mejor respuesta al tratamiento.

Para ello, se valora la supervivencia por causa específica en función del nivel de expresión génica. La búsqueda de genes se realiza primero mediante un procedimiento dirigido de análisis univariante, donde solo se estudia la influencia de un gen, y multivariante, estudiando combinaciones de genes.

Posteriormente, se desarrolla una predicción de variables significativas en el modelo, mediante regresión Lasso, para obtener la combinación de genes que aporte la mayor distancia entre las curvas de supervivencia.

Finalmente, el análisis funcional de los resultados identifica vías metabólicas significativas.

Las conclusiones obtenidas muestran que sí existen genes relacionados con una mejora en la supervivencia en estos pacientes en función de su nivel de expresión: alta expresión de SCP2, MYO6, PDLIM, U2SURP y baja expresión de MRPL23, CAD, FBP1 y RBP2. Por otro lado, la combinación de MRPL23 y MTA3 con bajo nivel de expresión aporta un 99.3% de tasa de supervivencia. En

el análisis funcional, la vía del metabolismo de fructosa y manosa ha sido significativa para el conjunto de genes predictores estudiado.

Abstract:

Breast cancer is the most common cancer disease worldwide in the female population. There are different subtypes depending on the tumor tissue's receptors. Luminal A subtype has the best prognosis and the lowest risk of recurrence, and the treatment is, in many cases, endocrine treatment.

This work focuses on these patients: breast cancer luminal A, with endocrine treatment, without chemotherapy. The aim is to search for gene signatures that predict a better response to treatment.

Survival by specific cause is assessed depending on the expression level of the genes. The search for genes is carried out first, by a directed procedure of univariate analysis, where only the influence of one gene is studied, and multivariate, studying combinations of genes.

Subsequently, a prediction of significant variables in this model, using Lasso regression, to obtain the combination of genes that provides the greatest distance between the survival curves.

In addition, a functional analysis of the results is performed to identify significant metabolic pathways.

The conclusions obtained show that there are genes related to an improvement in survival in these patients depending on their level of expression: high expression of SCP2, MYO6, PDLIM, U2SURP and low expression of MRPL23, CAD, FBP1 and RBP2. On the other hand, the combination of MRPL23 and MTA3, both with low level expression contributes a 99.3% survival rate. In the functional analysis, fructose and mannose metabolism pathway has been significant for the set of predictor genes studied.

Índice

1 Resumen ... 2

2 Introducción ... 3

2.1 Contexto y justificación del Trabajo ... 3

2.2 Objetivos del Trabajo ... 5

2.3 Enfoque y método seguido ... 6

2.4 Planificación del Trabajo ... 7

2.5 Breve sumario de contribuciones y productos obtenidos ... 7

2.6 Breve descripción de los otros capítulos de la memoria ... 7

3 Metodología ... 9

3.1 Obtención y preparación de los datos ... 9

3.2 Análisis univariado: Metodología ... 10

3.3 Análisis multivariado: Metodología ... 13

3.4 Otras estrategias: Regresión Lasso. ... 13

3.5 Análisis funcional de genes ... 14

4 Resultados ... 16

4.1 Análisis desciptivo de los datos. ... 16

4.1.1 Análisis descriptivo dataset Clin_data: ... 16

4.1.2 Análisis descriptivo dataset LumA_HT2: ... 17

4.2 Análisis univariado: Resultados ... 18

4.3 Análisis multivariado: Resultados ... 20

4.4 Regresión Lasso ... 21

4.5 Análisis funcional de los genes ... 22

5 Discusión ... 27

6 Conclusiones ... 30

6.1 Conclusiones ... 30

6.2 Líneas de futuro ... 30

6.3 Seguimiento de la planificación ... 31

7 Bibliografía ... 32

8 Anexos ... 35

Lista de figuras

- **Figura 1:** Tratamiento sistémico del cáncer de mama en función de marcadores de expresión. [Figura extraída de 6]
- **Figura 2:** Diagrama de Gantt: Planificación del trabajo.
- **Figura 3:** Curva de Kaplan-Meier. [Figura adaptada de 16]
- **Figura 4:** Esquema de funcionamiento de Metascape. [Figura extraída de 24].
- **Figura 5:** Edad al diagnóstico de cáncer de mama en función de la clasificación PAM50.
- **Figura 6:** Estado vital del paciente al final del seguimiento en función de la clasificación PAM50
- **Figura 7:** Edad al diagnóstico en cáncer de mama Luminal A
- **Figura 8**: Curvas de Kaplan-Meier de los genes MRPL23 y MYO6.
- **Figura 9**: Curvas Kaplan-Meier de la combinación de genes MRPL23 + MPA3
- **Figura 10:** Principales vías metabólicas obtenidas por Metascape de los 34 genes obtenidos en el análisis univariante
- **Figura 11:** Principales rutas metabólicas de Gene Ontology
- **Figura 12:** Mapa de enriquecimiento obtenido en Metascape
- **Figura 13:** Análisis funcional de proteínas del análisis univariante obtenida en STRING

Lista de tablas

- **Tabla 1**: Clasificación molecular del cáncer de mama. [Tabla adaptada de 7].
- **Tabla 2**: Genes que mejoran la supervivencia a 5 años cuando están infraexpresados.
- **Tabla 3**: Genes que mejoran la supervivencia a 5 años cuando están sobreexpresados.
- **Tabla 4**: Supervivencia a 5 años de la combinación de genes MRPL23 + MTA3.
- **Tabla 5**: Combinación de genes obtenidos mediante regresión Lasso.
- **Tabla 6**: Resultados obtenidos en análisis univariante y regresión Lasso
- **Tabla 7:** Rutas metabólicas significativas de las proteínas del análisis univariante obtenidas mediante STRING
- **Tabla 8:** Tejidos donde se encuentran expresadas las proteínas del análisis univariante. Obtenido por STRING**.**

1 Resumen

El cáncer de mama es la enfermedad oncológica más frecuente a nivel mundial en población femenina. Existen diferentes subtipos en función de los receptores presentes en el tejido tumoral. El luminal A es el de mejor pronóstico y menor riesgo de recaídas. Su tratamiento principalmente es endocrino.

Este trabajo se centra en estos pacientes: Cáncer de mama luminal A en tratamiento endocrino sin quimioterapia. El objetivo es buscar marcadores génicos que predigan una mejor respuesta al tratamiento.

Para ello, se valora la supervivencia por causa específica en función del nivel de expresión génica. La búsqueda de genes se realiza primero mediante un procedimiento dirigido de análisis univariante, donde solo se estudia la influencia de un gen, y multivariante, estudiando combinaciones de genes.

Posteriormente, se desarrolla una predicción de variables significativas en el modelo, mediante regresión Lasso, para obtener la combinación de genes que aporte la mayor distancia entre las curvas de supervivencia.

Finalmente, el análisis funcional de los resultados identifica vías metabólicas significativas.

Las conclusiones obtenidas muestran que sí existen genes relacionados con una mejora en la supervivencia en estos pacientes en función de su nivel de expresión: alta expresión de SCP2, MYO6, PDLIM, U2SURP y baja expresión de MRPL23, CAD, FBP1 y RBP2. Por otro lado, la combinación de MRPL23 y MTA3 con bajo nivel de expresión aporta un 99.3% de tasa de supervivencia. En el análisis funcional, la vía del metabolismo de fructosa y manosa ha sido significativa para el conjunto de genes predictores estudiado.

2 Introducción

2.1 Contexto y justificación del Trabajo

El cáncer de mama es la enfermedad oncológica más frecuente a nivel mundial en la población femenina. Según el Observatorio del cáncer de la AECC[1], en España, se diagnosticaron 33.875 casos durante el año 2020 en mujeres. Este dato, representa un 28,62% del total de cánceres diagnosticados en mujeres en ese mismo año.

Tradicionalmente, el cáncer de mama se ha clasificado según la histopatología del tumor[2]. Sin embargo, gracias al avance en las técnicas moleculares, se ha establecido un nuevo sistema de clasificación en función del perfil de expresión génica de las células tumorales.

Actualmente se reconoce la existencia de 4 subtipos moleculares de cáncer de mama[3,4]: luminal A, luminal B (dentro de este grupo se diferencian dos subtipos), HER2 positivo y triple negativo. Esta clasificación se basa principalmente en el estado de los receptores de estrógenos (RE), de progesterona (RP) y la sobreexpresión del receptor factor de crecimiento epidérmico humano 2 (HER2/neu) en los tejidos cancerígenos.

	RE	RP	HER2	Ki67	Grado histológico	Frecuencia	Pronóstico
Luminal A	+	≥20%	-	>20%	I y II	50%	Bueno
Luminal B	+	<20%	-	>20%	II y III	20%	Intermedio
	+	+/-	+	+/-	III		
HER2	-	-	+		III	15%	Malo
Triple negativo	-	-	-		III	15%	Malo

***Tabla 1**: Clasificación molecular del cáncer de mama. [Tabla adaptada de 7]*

Cuando los tumores de mama se clasifican usando una combinación de los datos obtenidos de la clasificación molecular e histopatológica, está demostrado que se obtiene mayor información sobre el pronóstico del paciente, el riesgo de recaída y la respuesta al tratamiento. Incluso hay sociedades, como la Sociedad Española de Oncología Médica (SEOM) o la European Society for Medical Oncology (ESMO) que, en sus guías clínicas, recomiendan tipificar el tipo de cáncer en función de la histología e inmunohistoquímica, para establecer el pronóstico y el tratamiento más adecuado para el paciente [5,6].

Los factores pronósticos más importantes al diagnóstico de un cáncer de mama en estadio temprano son: Expresión de RE, RP, HER2, marcadores de proliferación como Ki67, número de ganglios linfáticos afectados, histología del tejido tumoral, tamaño, grado histológico y presencia de invasión vascular.

Este trabajo se va a centrar en los tumores de mama subtipo luminal A. El cáncer de mama luminal A se caracteriza por presentar receptores de estrógenos, al menos un 20% de las células del tejido tumoral presenta receptores de progesterona y no sobreexpresan HER2, es decir, tiene un patrón RE+, RP≥20% y HER2-. Constituye el 50-60% de todos los cánceres luminales, que a su vez, son el 75-80% de todos los cánceres de mama. Es el subtipo con menor riesgo de recaídas y mejor respuesta al tratamiento. Además, suele diagnosticarse en estadios tempranos, son de menor grado al diagnóstico y el tejido presenta baja proliferación. La supervivencia por causa específica de este tipo de cáncer es del 91.6% a 5 años.

En el tratamiento del cáncer de mama se combinan diferentes estrategias. Se pueden emplear tratamientos locales, como cirugía o radioterapia, o terapias sistémicas que se usan de forma complementaria. Aquí se encuentra la quimioterapia, el tratamiento endocrino y terapias moleculares dirigidas. Además, en algunos casos se requieren medidas de adicionales como la protección de la función ovárica u otras medidas de soporte necesarias. Ante este amplio abanico de tratamientos disponibles, la selección del abordaje terapéutico más adecuado para cada paciente y cada momento va a depender principalmente del subtipo tumoral.

La expresión de RE, RP y HER2 son actualmente los únicos factores predictivos aceptados por las guías clínicas para la selección de pacientes candidatos a recibir tratamiento endocrino y/o terapias anti-HER2[7].

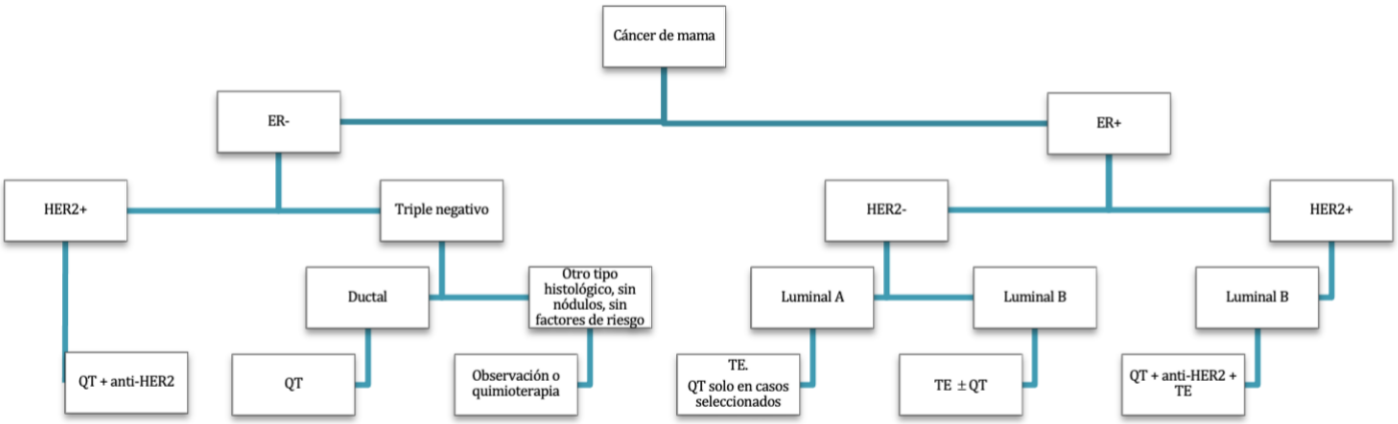


Figura 2: *Tratamiento sistémico del cáncer de mama en función de marcadores de expresión. [Figura extraída de 6]. QT, quimioterapia; TE, terapia endocrina*

En aquellos tumores RE+ se aconseja estudiar el perfil de expresión génica para valorar el beneficio de la quimioterapia adyuvante. Existen paneles de genes validados y estandarizados que aportan una puntuación del riesgo de recaída del paciente, como son Oncotype DX (estudia 21 genes), MammaPrint (70 genes) o la clasificación PAM50, entre otros[8]. La estrategia empleada va a depender también del tipo tumoral y la presencia o no de ganglios linfáticos afectados.

Una alta expresión del RE se asocia generalmente con un bajo beneficio de la quimioterapia, por lo que el tratamiento más adecuado en los tumores

Luminal A es únicamente el hormonal. La quimioterapia está indicada principalmente en tumores HER2-, donde se haya demostrado un beneficio del tratamiento.

Existen principalmente dos tipos de tratamientos hormonales: bloqueadores del receptor de estrógeno (por ejemplo, tamoxifeno) e inhibidores de la aromatasa (letrozol o anastrozol).

De forma general, en casos de cáncer de mama Luminal A, se aconseja el empleo de tamoxifeno durante 5-10 años en pacientes premenopáusicas, añadiendo o cambiándolo por un inhibidor de la aromatasa en postmenopáusicas.

Aunque, por lo general, este tipo de tumor suele responder bien al tratamiento endocrino, el riesgo de recaída a los 10 años es solo de un 10-15%, existen casos donde no se consigue una adecuada respuesta.

El objetivo de este trabajo consiste en identificar, en la base de datos pública de METABRIC (Molecular Taxonomy of Breast Cancer International Consortium), datos de expresión de pacientes de cáncer de mama subtipo luminal A en tratamiento exclusivamente hormonal, que sirvan como marcadores genéticos que se correlacionen con una buena o mala respuesta al tratamiento endocrino.

La calidad de la respuesta al tratamiento se evalúa en base a la supervivencia por causa específica a 5 años. Se estudian tanto genes aislados como combinaciones de genes que permitan predecir un incremento en la tasa de éxito o fracaso en la respuesta.

Conocer estos marcadores genéticos es una manera de permitir personalizar aún más el tratamiento de las pacientes de cáncer, con el objetivo de incrementar su supervivencia y evitar el uso de fármacos que no vayan a ser del todo eficaces e incrementen tiempo, riesgos y costes en el tratamiento de manera innecesaria.

2.2 Objetivos del Trabajo

2.2.1 Objetivo general:

El objetivo principal de este trabajo consiste en identificar, en pacientes de cáncer de mama luminal A en tratamiento endocrino incluidos en repositorios de datos públicos, genes o combinaciones de genes que se correlacionen con un aumento de la supervivencia por causa específica.

2.2.2 Objetivos específicos:

- Obtener un conjunto de datos públicos que recoja información clínica a largo plazo y datos moleculares de expresión génica, de pacientes de cáncer de mama luminal A en tratamiento endocrino.
- A partir de una selección de genes cuyas proteínas resultantes se ha comprobado que interactúan con el tratamiento endocrino, identificar

genes capaces de predecir una mejor supervivencia a 5 años empleando análisis univariado.

- Análisis multivariado: Explorar combinaciones de 2 genes. En caso de que sea posible incrementar aún más la tasa de supervivencia, estudiar combinaciones de 3 o más genes que mejoren el pronóstico.
- Análisis funcional de los resultados obtenidos
- Identificar vías moleculares o biológicas, relacionadas con los genes estudiados, que permitan identificar mecanismos de resistencia al fármaco.

2.3 Enfoque y método seguido

Para el desarrollo de este trabajo se parte de un conjunto de muestras de cáncer de mama, obtenido de un repositorio público, con datos clínicos de seguimiento a largo plazo, durante un periodo de 5-10 años, y datos moleculares de expresión génica. Además de una lista de genes cuyas proteínas resultantes se ha comprobado, mediante ensayos de proteómica dirigida, que interactúan con la terapia endocrina empleada para el cáncer de mama luminal A.

A partir de estos datos, el objetivo es determinar si la expresión de los genes seleccionados, o combinaciones de los mismos, puede influir en la supervivencia de los pacientes.

Para ello, se pretende comparar la supervivencia de los pacientes con alta expresión, con la supervivencia de los pacientes con baja expresión para una determinada signatura génica.

En primer lugar, se comenzará con un análisis univariado, estudiando la influencia de un solo gen, para luego extenderlo a un análisis multivariado basado en la combinación de 2 genes.

El cáncer de mama luminal A tiene, de por sí, una elevada tasa de supervivencia, por lo que mejorar este dato es una tarea ambiciosa. Si tras observar los resultados obtenidos en el análisis univariado y de la combinación de dos genes se ve conveniente, se ampliará el estudio con la combinación de 3 o más genes, siempre que esto implique una mejora significativa en el pronóstico.

Por último, una vez se han identificado aquellos genes o combinaciones predictoras de respuesta al tratamiento, se puede realizar un análisis funcional de los resultados obtenidos, buscando qué rutas o mecanismos de resistencia están relacionadas. De esta forma, se podría entender con mayor profundidad qué aspectos pueden influir en la mejor o peor respuesta al tratamiento y abrir la posibilidad de encontrar nuevos fármacos que incrementen la supervivencia de estos pacientes.

2.4 Planificación del Trabajo

Las distintas fases en las que se va a planificar el trabajo son:

- Fase 1:
    a. Selección de la muestra de estudio y análisis descriptivo de los datos.
    b. Análisis univariado: Estudio de curvas de supervivencia con 1 gen
    c. Análisis multivariado I: Estudio de curvas de supervivencia con 2 genes
    d. Entrega PEC 2

- Fase 2:
    a. Análisis multivariado II: Estudio de curvas de supervivencia con 3 o más genes
    b. Análisis funcional de los resultados.
    c. Entrega PEC 3
- Fase 3: Redacción de la memoria del TFM.
- Fase 4: Elaboración de la presentación.
- Fase 5: Defensa pública del TFM.

A continuación se muestra un diagrama de Gantt con la planificación por semanas para la realización de las distintas fases, desde el 4-octubre-2021 al 21-enero-2022:

Diagrama de Gantt: Planificación del trabajo

04.Octubre.2021 – 21.Enero.2022

TAREA	INICIO	FIN
Fase 1		
Análisis descriptivo	4-10-21	10-10-21
Análisis univariado	11-10-21	24-10-21
Análisis multivariado I	25-10-21	7-11-21
Entrega PEC 2	4-11-21	8-11-21
Fase 2		
Análisis multivariado II	8-11-21	14-11-21
Análisis funcional	15-11-21	5-12-21
Entrega PEC 3	2-12-21	9-12-21
Fase 3		
Memoria TFM	10-12-21	24-12-21
Fase 4		
Presentación	27-12-21	3-1-22
Fase 5		
Exposición pública	13-1-22	21-1-22

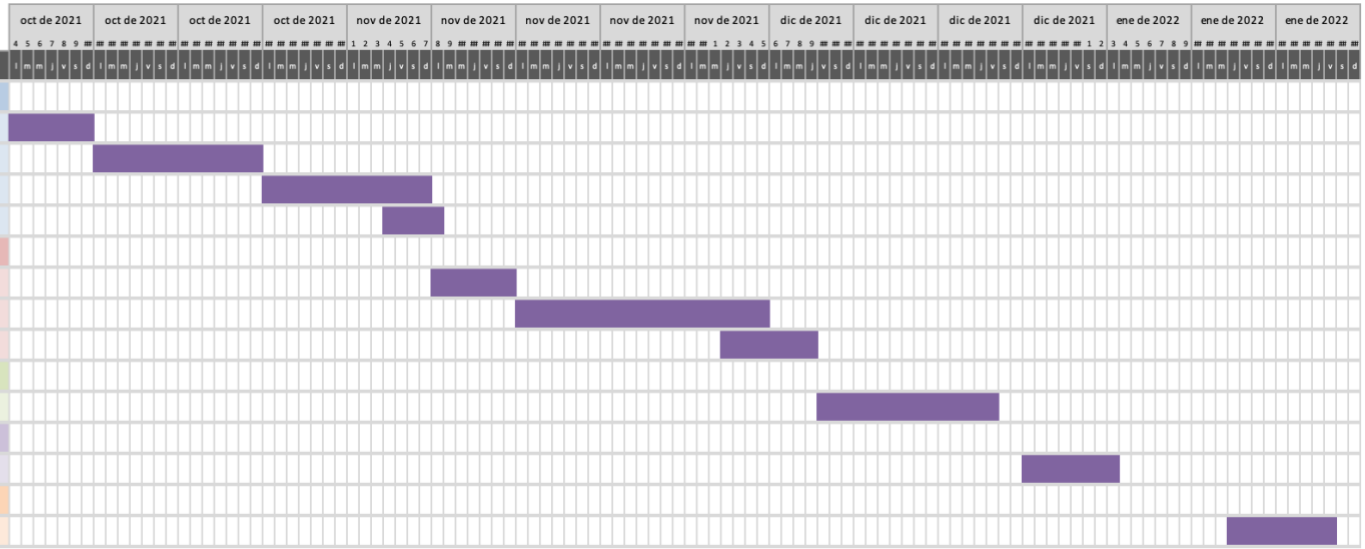


Figura 2: Diagrama de Gantt

2.5 Breve sumario de contribuciones y productos obtenidos

- Planificación del trabajo, diagrama de Gantt.
- Elección de la base de datos.
- Análisis descriptivo de los datos de estudio.
- Análisis univariado: Metodología y resultados.
- Análisis multivariado: Metodología y resultados.
- Regresión Lasso: Metodología y resultados.
- Análisis funcional de genes significativos en el análisis univariado.

2.6 Breve descripción de los otros capítulos de la memoria

- Capítulo 3: Metodología. Se explican los procesos seguidos en cada uno de los puntos del trabajo:
  1. Obtención y preparación de los datos: Elección de la base de datos y selección de los pacientes a incluir en el estudio
  2. Análisis univariado: Metodología empleada. Explicación de las curvas de supervivencia.
  3. Análisis multivariado. Metodología para el análisis de la combinación de dos o más genes.
  4. Otras estrategias: Regresión Lasso. Justificación y metodología de este análisis.
  5. Análisis funcional de genes con herramientas web
- Capítulo 4: Resultados:
  1. Análisis descriptivo de los datos:
     - Dataset Clin_data
     - Dataset LumA_HT2
  2. Análisis univariado: listado de genes significativos obtenidos en el análisis univariado.
  3. Análisis bivariado: Resultado de la combinación de dos genes.
  4. Regresión Lasso: Genes seleccionados mediante regresión Lasso.
  5. Análisis funcional de los genes y proteínas: Empleo de las herramientas web Metascape y STRING.
- Capítulo 5: Discusión
- Capítulo 6: Conclusiones
- Capítulo 7 Bibliografía

- Capítulo 8: Anexos:
    1. **Anexo 1**: Script para la obtención de los dataset de trabajo
    2. **Anexo 2**: Listado de genes análisis univariado
    3. **Anexo 3**: Script para el análisis univariante
    4. **Anexo 4**: Script para el análisis bivariante
    5. **Anexo 5:** Script Regresión Lasso

3 Metodología

3.1 Obtención y preparación de los datos

El primer objetivo planteado en el trabajo es la búsqueda de una base de datos pública que permita obtener tanto datos clínicos de pacientes de cáncer de mama como datos de expresión génica de estos mismos pacientes.

Respecto a los datos clínicos, para llevar a cabo el estudio posterior, es imprescindible que la base de datos aporte, como mínimo, la siguiente información:

- Clasificación del tipo de tumor. La clasificación PAM50[13] permite extraer el subtipo biológico intrínseco del tumor, además de dar información del riesgo de recurrencia. Clasifica el tumor en base al análisis de 50 genes en los siguientes subtipos: Luminal A, Luminal B, Her2, basal-like y normal-like.
- Tratamiento: diferenciando si el paciente ha recibido tratamiento hormonal, quimioterapia, ambos o ninguno, para poder extraer el grupo más adecuado al estudio.
- Supervivencia del paciente a largo plazo. Al querer analizar la influencia de la expresión génica en la supervivencia será necesario que el seguimiento de los pacientes incluidos en la base de datos sea lo más prolongada en el tiempo posible.
- Estado del paciente al finalizar el seguimiento: Es importante diferenciar, en caso de muerte, si ha sido debido al cáncer de mama o por otra causa.

Para comenzar con la elección se tuvieron en cuenta las bases de datos públicas TCGA[14] y METABRIC[15]. En ambos repositorios encontramos los datos clínicos anteriormente mencionados, sin embargo, se observa una gran diferencia en cuanto al seguimiento de los pacientes. En la base de datos TCGA la media del seguimiento de los pacientes con cáncer de mama es de 27,7 meses[10], mientras que en METABRIC es de 9,77 años, llegando a un seguimiento de 14 años en aquellos pacientes que siguen vivos pasado ese tiempo.

Teniendo en cuenta que el principal objetivo de este trabajo es valorar el efecto predictor de ciertos genes o combinaciones respecto a la supervivencia, un incremento tan notable en el seguimiento de los pacientes cuando se elige la base de datos de METABRIC, nos va a permitir detectar mejor los cambios. Además, según las guías clínicas, el tratamiento hormonal debe tener una duración entre 5 y 10 años[6], por lo que es recomendable un seguimiento mínimo de 5 años para valorar la efectividad del tratamiento. Por estos motivos principalmente, se va a seleccionar la base de datos de METABRIC[12] para la obtención de datos clínicos y de expresión génica.

La base de datos seleccionada contiene datos clínicos de 3240 pacientes recogidos entre 1977 y 2005 en 5 bancos de tumores de Reino Unido y Canadá,

sin embargo, para realizar el trabajo se va a escoger una muestra de 1980 pacientes. Este dataset más reducido se encuentra definido en el artículo de Rueda et al.[9] en la tabla 6 de la *Supplementary information*.

Respecto a los datos de expresión génica, en el dataset brca_metabric se encuentra un archivo denominado *data_expression_median.txt*, que incluye la expresión de mRNA de 24,368 genes para cada paciente, obtenido mediante la plataforma Illumina Human v3 microarray y representado como el logaritmo del nivel de intensidad de la expresión.

Una vez ya se han extraído los datos clínicos (*clin_data*) y los datos de expresión (*expr_data*), hay que seleccionar aquellos pacientes que cumplan los requisitos del estudio: cáncer de mama subtipo luminal A, en tratamiento hormonal, que no hayan recibido quimioterapia y que tengan datos de expresión génica. Todo el tratamiento de los datos y el análisis de los mismos se realiza con el entorno de trabajo Rstudio (ver. 4.1.1).

Finalmente, se obtiene una selección de 423 pacientes que cumplen los requisitos establecidos y datos de expresión de 24368 genes. Esta información se guarda en el dataset denominado *LumA_HT2*.

En el **Anexo 1** se incluye el código de R para la obtención de los diferentes datasets empleados en el trabajo:

- *Clin_data*: Datos clínicos originales descargados de METABRIC.
- *Expr_data:* Datos de expresión originales descargados de METABRIC.
- *LumA_HT2*: Datos clínicos de pacientes luminal A en hormonoterapia, sin quimioterapia, con datos de expresión.
- *m*: Datos de expresión de los pacientes del dataset LumA_HT2.

3.2 Análisis univariado: Metodología

Tras seleccionar la muestra de estudio se comienza con el análisis univariado. El objetivo es estudiar la influencia de un solo gen en la supervivencia del paciente por causa específica a 5 años.

Dado que la cantidad de genes a testar es muy grande (se tienen datos de expresión de 24.368 genes, y resultaría imposible encontrar resultados estadísticamente significativos), para este análisis se ha utilizado como conjunto de partida una lista de genes que codifican proteínas que interactúan con el tamoxifeno, el fármaco más empleado como terapia endocrina.

Este conjunto, se obtuvo mediante un experimento previo de proteómica dirigida. Para ello, se identificaron mediante espectrometría de masas proteínas capaces de unirse e interactuar con un complejo de tamoxifeno marcado con biotina.

Tras asociar estas proteínas con su correspondiente gen, se obtiene un listado de 808 genes de los 24368 incluidos en la base de datos original. El listado de estos genes se encuentra en el **Anexo 2** y constituye la lista de genes con la que se va a trabajar en los siguientes pasos.

Para estudiar qué influencia tiene cada uno de estos genes en la respuesta al tratamiento del paciente, se realiza un análisis de supervivencia, clasificando a los pacientes en función del grado de expresión que tienen de dicho gen.

El análisis de supervivencia es el conjunto de técnicas estadísticas que se emplean para analizar el tiempo que tarda en ocurrir un suceso, este tiempo se conoce como supervivencia [16, 17].

Existen diferentes tipos de supervivencia en función del evento que se considere. En los estudios de cáncer, el evento puede ser, entre otros:

1. Fallecimiento del paciente por cualquier causa: se hablaría de supervivencia absoluta.
2. Fallecimiento debido al cáncer: supervivencia por causa específica.
3. Reaparición del cáncer: supervivencia libre de recaída.

Para estudiar los datos de supervivencia se emplea la función de supervivencia (S[t]). Se calcula como el número de pacientes que sobreviven en un tiempo determinado entre el número total de pacientes en riesgo[18].

$$S_t = \frac{Sujetos\ al\ inicio - Sujetos\ muertos\ a\ tiempo\ t}{Sujetos\ al\ inicio}$$

Representa la probabilidad de que a un individuo no le ocurra el evento desde el inicio del seguimiento hasta un tiempo determinado (t). Toma valores de 0 a 1, correspondiendo el 1 con el inicio del estudio, ya que a ninguno de los pacientes les ha ocurrido el evento a ese tiempo (t=0).

Para calcular esta función de supervivencia se pueden usar métodos paramétricos o no paramétricos en función de la distribución de los datos. Generalmente, en los estudios médicos, no se conoce la distribución de los datos, por lo que se suelen usar métodos no paramétricos. Uno de los más usados es el estimador de Kaplan-Meier, que calcula la proporción acumulada de pacientes que sobrevive para cada tiempo.

Los datos de supervivencia se pueden representar mediante las curvas de Kaplan-Meier, donde el eje de abscisas representa el tiempo de seguimiento (t), y el eje de ordenadas, la función de supervivencia (*S(t)*).

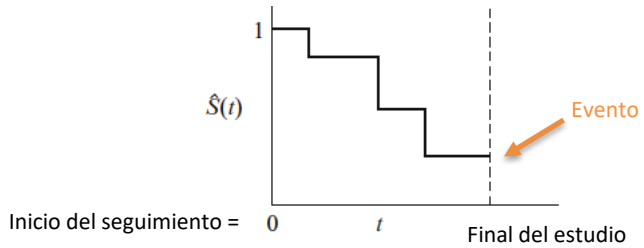


Figura 3: Curva de Kaplan-Meier. [Figura adaptada de 16] S(t)=Función de supervivencia

Los análisis de supervivencia permiten comparar si hay diferencias en la supervivencia entre dos o más poblaciones cuando se modifica alguna variable. Para establecer si dos poblaciones distintas tienen funciones de supervivencia estadísticamente equivalentes o no, se pueden emplear diversas pruebas estadísticas de contraste de hipótesis, como el *test log-rank*. Este test se basa

en una prueba chi-cuadrado, mediante la que realiza un contraste de hipótesis para ver si existen diferencias entre ambas curvas.

En este análisis, el suceso que se quiere analizar es el tiempo hasta la muerte por cáncer de mama, y no por otra causa, y la variable que se va a modificar es el nivel de expresión de un gen. Es decir, se va a comparar la supervivencia por causa específica en función del grado de expresión de un gen.

Con el fin de clasificar a los pacientes en función de sus niveles de expresión, se calculan los terciles (o quantiles 33,3%) de los niveles de expresión de los pacientes para cada uno de los genes. De esta manera, se obtienen tres subgrupos de pacientes para cada gen:

- El primer tercil o tercil superior, corresponde a los pacientes con alta expresión del gen.
- En el segundo tercil se localizan los pacientes con expresión media.
- El tercer tercil o tercil inferior, corresponde a los pacientes con baja expresión del gen.

En el análisis de supervivencia se incluyen los pacientes del primer y tercer tercil, es decir los de alta y baja expresión del gen. El segundo tercil o grupo de expresión media, se descarta con el fin de obtener mayor separación entre los grupos.

Para cada grupo y cada gen, se obtiene una función de supervivencia. Estos datos se representan en curvas de Kaplan-Meier y se realiza el *test log-rank* para evaluar si son estadísticamente equivalente.

En este análisis, se establece un valor $p<0.05$ para identificar diferencias estadísticamente significativas entre las funciones. Así, en aquellos genes donde se obtenga en el *test log-rank* un valor $p<0.05$, se podrá decir que hay diferencias estadísticamente significativas, y por tanto, deducir que el nivel de expresión del gen influye en la supervivencia.

Idealmente, al ejecutar este análisis, se debería aplicar una corrección por *multiple testing*, pero dado el carácter exploratorio del análisis y la alta supervivencia intrínseca de los cáncer luminal A (91.6% a 5 años), que dificultaría la observación de diferencias, se decide no aplicarlo para este trabajo.

El modo en el que el nivel de expresión influye sobre la supervivencia, disminuyéndola o incrementándola, se obtiene mediante la diferencia en el ratio de supervivencia (*Diff5*) en pacientes con alta expresión menos el ratio en los de baja expresión. Un resultado positivo en *Diff5*, indicará un incremento de la supervivencia cuando el gen está altamente expresado, mientras que un resultado negativo indicará un descenso de la misma.

Para generar las curvas de supervivencia en R y hacer el análisis de las mismas, se han empleado los paquetes *Survival* [19], *rms*[20] y *survminer* [21], incluidos en la librería CRAN. En el **Anexo 3** se encuentra el script ejecutado en R para realizar el análisis univariado.

3.3 Análisis multivariado: Metodología

Tras obtener los genes cuya expresión tiene influencia de forma individual sobre la supervivencia de los pacientes, se puede estudiar si existen combinaciones de dos o más genes que mejoren dicha supervivencia.

Se considera que hay una mejora en la supervivencia cuando el valor de la supervivencia a los 5 años obtenido en la combinación, supera en un 5% a la mayor de las supervivencias de los genes de forma individual.

Primero, se realiza un análisis bivariado, donde se estudian combinaciones de cada uno de los genes que han mostrado ser influyentes en el análisis univariado, con la lista original de 808 genes que interaccionan con el tamoxifeno.

Para cada una de estas combinaciones, se calculan las funciones de supervivencia correspondientes y se realiza el *test log-rank* ($p<0,05$).

Cuando se estudia la combinación de genes, existen dos escenarios. Su expresión puede ir en el mismo sentido, es decir, en el caso de dos genes, ambos pueden estar sobreexpresados o infraexpresados. En este caso, el cálculo de la expresión se hace sumando la expresión de ambos genes. Sin embargo, cuando un gen está infraexpresado, y otro sobreexpresado, el valor de la combinación se hace sumando. En cada caso, el cálculo de la expresión de la combinación se debe hacer tanto con la suma como con la resta.

El script empleado para realizar el análisis bivariado se detalla en el **Anexo 4**, empleando los mismos paquetes de R que en el análisis univariado. El cálculo de la expresión (I) se debe ajustar para hacer la suma y la resta de las expresiones individuales.

En aquellas combinaciones que muestren ser significativas, se obtiene el valor de la supervivencia a 5 años y se calcula el porcentaje de variación respecto al de cada uno de los genes por separado.

Si tras realizar el análisis bivariado se obtienen combinaciones de dos genes que mejoren la supervivencia se puede realizar un análisis de 3 genes para buscar un incremento aún mayor de dicha supervivencia.

En este análisis, hay que estudiar la combinación de cada par de genes obtenidos con los 806 genes restantes que se están incluyendo en el estudio (todos los genes menos los dos que forman parte del par).

3.4 Otras estrategias: Regresión Lasso.

Además de buscar la mayor de las supervivencias posibles en un punto concreto del tiempo, en este caso a 5 años, existen otras estrategias que permiten obtener una combinación de marcadores génicos que prediga una buena respuesta al tratamiento.

Se puede estudiar la distancia entre las curvas de supervivencia. Aquella signatura génica que permita obtener una curva de supervivencia a mayor distancia del resto hacia la derecha, es decir, con mayor valor de $S(t)$, estará indicando que la respuesta al tratamiento es mejor.

La expresión de todos los genes estudiados tiene cierta influencia en el modelo, sin embargo, no todos lo hacen con la misma fuerza. Para encontrar la combinación de genes que de verdad influye en el modelo y mejore la supervivencia desplazando la curva a mayor distancia, se puede emplear la regresión Lasso.

Este análisis permite obtener, a partir de un modelo multivariante, aquellos predictores que tienen influencia real en el modelo. Es considerada una técnica de selección de variables. Se basa en un modelo de regresión que fuerza que los coeficientes de las variables tiendan a 0, por lo que consigue excluir del modelo aquellos menos relevantes.

Es un análisis útil cuando creemos que se van a obtener pocos predictores, como es el caso del análisis de expresión génica, donde generalmente se prevé que pocos genes tengan una influencia real en el modelo[22].

En este análisis, se incluye la expresión de los 808 genes del modelo de todos los pacientes, y se obtiene una lista de genes cuyo valor de coeficiente es distinto a 0, por lo que se pueden considerar predictores relevantes en el modelo.

Para interpretar cómo afecta la expresión del gen a la supervivencia hay que fijarse en el signo y valor del coeficiente:

- Coeficiente>0: se observa un incremento en la supervivencia cuando el gen está infraexpresado.
- Coeficiente<0: dicho gen incrementará la supervivencia cuando está sobreexpresado.
- Coeficiente=0: El gen no tiene influencia real en el modelo.

El script empleado en R para realizar la regresión Lasso se encuentra en el **Anexo 5**. Se ha empleado el paquete *glmnet* [23] incluido en la librería CRAN.

3.5 Análisis funcional de genes con herramientas web

El análisis funcional de genes permite conocer la función que ejerce cada uno de los genes y las proteínas que codifican.

Actualmente existen numerosas herramientas web, públicas y gratuitas, que permiten hacer el análisis funcional de genes y proteínas empleando numerosas bases de datos.

Para hacer este análisis se van a emplear las herramientas web Metascape y STRING.

Metascape[24] compara los datos introducidos con diferentes bases de datos y sigue un proceso conocido como “CAME” (figura 3), donde los pasos principales son:

1. ID Conversion: Transforma el ID introducido, en este caso el símbolo (SYMBOL) del gen en el ID de Entrez.
2. Annotation: Extrae la anotación de los genes (descripción, funciones, proteínas…)

3. Membership: Marca genes de interés respecto a términos GO que describen procesos biológicos.
4. Enrichment analysis: Mediante el enriquecimiento se obtienen las vías metabólicas estadísticamente significativas para el conjunto de genes introducidos.

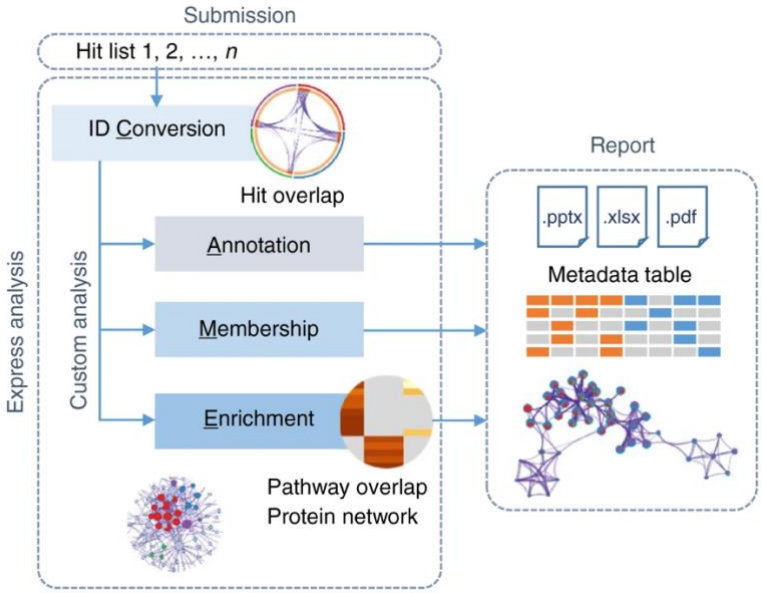


Figura 4: Esquema de funcionamiento de Metascape. Imagen extraída de 24.

Es una herramienta fácil de utilizar, sin embargo los resultados obtenidos no tienen mucha estabilidad. Al estar en constante actualización, los resultados pueden sufrir modificaciones en función de las actualizaciones que se hayan ido produciendo.

En el caso de STRING (*Search Tool for the Retrieval of Interacting Genes/Proteins*)[25], es una base de datos de interacciones entre proteínas conocidas y predichas. Permite conocer las interacciones directas o físicas e indirectas o funcionales entre las diferentes proteínas introducidas. Emplea bases de datos primarias (BioGRID, DIP, IntAct, MINT…) y de datos predichos.

4 Resultados

4.1 Análisis desciptivo de los datos.

Los datos con los que se ha realizado el trabajo están recogidos en el dataset *LumA_HT2,* que es una selección de pacientes a partir del dataset original *(Clin_data).*

Para comenzar a entender estos datos, la información que aportan y su distribución, se realiza un análisis descriptivo de los mismos.

4.1.1 Análisis descriptivo dataset Clin_data:

El conjunto de datos del que se ha partido recoge información clínica de 1980 pacientes de cáncer de mama, todas ellas mujeres.

Si se atiende a la clasificación PAM50, un 61,09% de los pacientes corresponden a cáncer de mama tipo luminal (36,37% luminal A y 24,72% luminal B). El 16,67% de los pacientes se clasificaron como basal-like, 12,16% Her2 y 10,08% normal-like.

Respecto a la edad al diagnóstico, la media de edad para todos los pacientes fue de 61,80 años (rango 21,93-96,29). Diferenciando por subtipos, los luminales son los de mayor edad media: 67,39 para el luminal A (26,36-90,23) y 66,36 para el luminal B (28,62-92,14). La edad media para los basales es de 54,22 años (26,72-96,29), para los Her2 59,20 (21,93-90,08) y 58,90 para los normales (31,87-83,92).

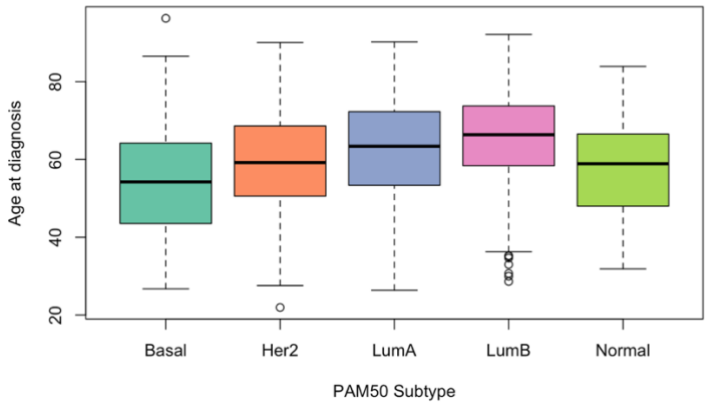


Figura 5: Edad al diagnóstico de cáncer de mama en función del subtipo PAM50

Una vez se ha realizado el seguimiento de las pacientes, en la Figura 4 se muestra el estado vital en función del subtipo de tumor. La muerte debida al cáncer de mama ha sido de un 32,64% en todas las pacientes incluidas en la base de datos. Este dato llega a ser de un 47,09% en pacientes con subtipo Her2 y se reduce hasta un 21,51% en cáncer de mama luminal A, el cual, se considera el subtipo de mejor pronóstico.

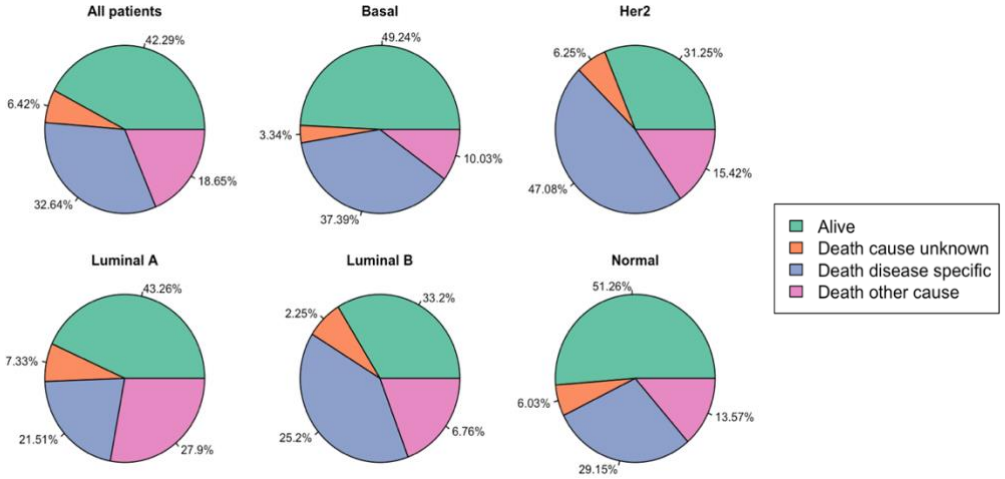


Figura 6: Estado vital del paciente al final del seguimiento en función de la clasificación PAM50

4.1.2 Análisis descriptivo dataset LumA_HT2:

El dataset denominado LumA_HT2, recoge la muestra de estudio seleccionada a partir de *Clin_data*. Contiene 423 pacientes de cáncer de mama luminal A, en tratamiento con terapia hormonal, sin quimioterapia. Todas las pacientes son mujeres, con una media de edad al diagnóstico de 67,39 años (26,36-90,23 años). En la Figura 5 se observa cómo la frecuencia del diagnóstico de este tipo de tumor se incrementa entre los 60 y 75 años.

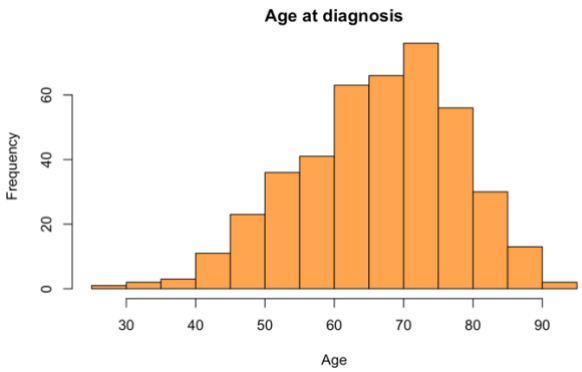


Figura 7: Edad al diagnóstico en cáncer de mama Luminal A

Respecto a la lateralidad del tumor, hay casi la misma proporción de tumores en la mama derecha que izquierda, siendo en esta última ligeramente más frecuente: el 51,88% de los pacientes presentaron el tumor en la mama izquierda.

Atendiendo a la histología del tejido, el tumor más frecuente es el carcinoma ductal invasivo con un 75,41% de los casos, seguido del carcinoma lobular invasivo con solo un 9,92%. El subtipo menos frecuente es el medular (0,23%).

Según el estadio del tumor en el momento del diagnóstico, el 65,90% de las pacientes estaba en estadio 2, y un 27,21% en estadio 1.

Respecto al tratamiento local, un 60,43% de los pacientes se sometió a mastectomía. Entre los tratamientos endocrinos disponibles, el más empleado ha sido el Tamoxifeno, en un 89,2% de las pacientes, seguido de los inhibidores de la aromatasa en un 4,25%. Otros fármacos utilizados aunque menos frecuentes han sido, la combinación de Tamoxifeno y Goserelina, en un 0,47% o los análogos a la hormona liberadora de gonadotropina (GnRHA) empleados en un 1,18% de los pacientes.

Como se ha comentado en la evaluación global de los datos, la mayoría de los pacientes de la muestra, un 43,26%, seguían vivos pasado el periodo de seguimiento, lo que confirma el buen pronóstico que suelen tener este tipo de tumores. El 21,51% fallecieron por causa específica, es decir, a consecuencia del cáncer de mama (figura 4).

4.2 Análisis univariado: Resultados

Una vez realizado el análisis univariado sobre la lista de genes seleccionados, se obtienen 34 genes con valor $p<0.05$. En estos genes, se puede asumir que una diferencia en el nivel de expresión del gen influye de manera estadísticamente significativa en la supervivencia a 5 años.

En los genes significativos, hay 18 genes que implican un incremento de la supervivencia por encima del 91.6% cuando están infraexpresados (*Tabla 2*), mientras que en otros 16, el incremento se produce cuando están sobreexpresados (*Tabla 3*).

Gen	Valor p	Supervivencia	diff5
Supervivencia Luminal A		0,9160	
MYH14	0.042	0,9169	-0,0120
NUMA1	0.023	0,9188	-0,0008
SORD	0.032	0,9336	-0,0392
HNRNPUL2	0.015	0,9395	-0,0509
FAU	0.014	0,9417	-0,0409
MRPL23	0.044	0,9421	-0,0547
FBP1	0.042	0,9425	-0,0234
NCOA3	0.015	0,9467	-0,0413
KCNK1	0.038	0,9471	-0,0268
EXOC6	0.031	0,9486	-0,0616
HNRNPF	0.009	0,9491	-0,0626
MYO1B	0.047	0,9555	-0,0520
RBP2	0.012	0,9560	-0,0363
CAD	0.035	0,9615	-0,0720
MRPL11	0.015	0,9638	-0,0702
MAGED2	0.024	0,9705	-0,0818
EMG1	0.015	0,9707	-0,0981
RICTOR	0.026	0,9784	-0,0885

Tabla 2: Genes que mejoran la supervivencia a 5 años cuando están infraexpresados.

Gen	Valor p	Supervivencia	diff5
Supervivencia Luminal A		0,9160	
PA2G4	0.026	0,9320	0,0051
PKP3	0.030	0,9337	0,0064
ABCF1	0.041	0,9337	0,0447
SRSF10	0.036	0,9394	0,0423
PRMT1	0.032	0,9487	0,0076
U2SURP	0.034	0,9492	0,0484
MYO6	0.018	0,9495	0,0913
SCP2	0.012	0,9545	0,0500
STRN	0.001	0,9556	0,0541
LDB1	0.007	0,9563	0,0449
STAT3	0.013	0,9635	0,0598
RAI1	0.001	0,9701	0,0898
CTNNA1	0.030	0,9707	0,0879
PDLIM5	0.003	0,9778	0,0827
PFKP	0.041	0,9785	0,0676
TUBGCP3	0.012	0,9852	0,1145

Tabla 3: Genes que mejoran la supervivencia a 5 años cuando están sobreexpresados.

Se muestran como ejemplo las curvas de supervivencia obtenidas para los genes *MRPL23* y *MYO6* (figura 7). La línea azul representa la supervivencia del grupo de pacientes con alta expresión del gen, y la línea roja los de baja expresión. El tiempo se expresa en meses.

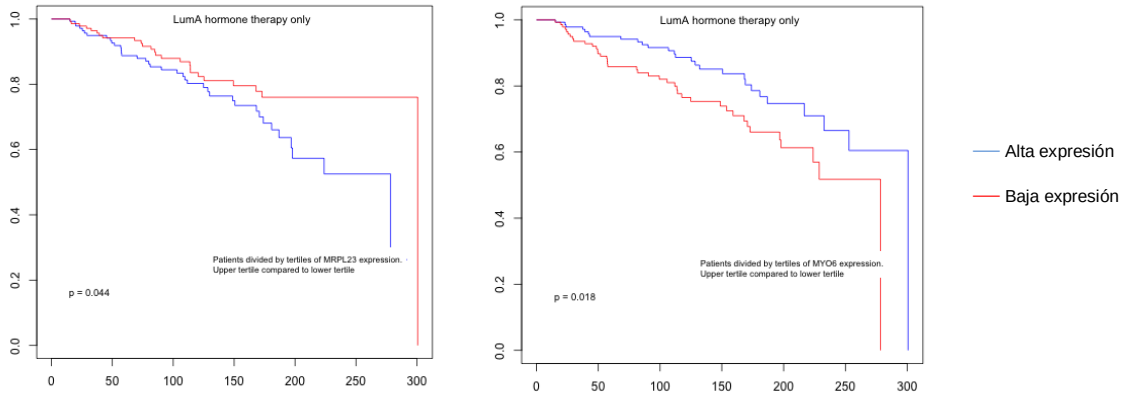


Figura 8: Curvas de supervivencia de los genes MRPL23 y MYO6

4.3 Análisis multivariado: Resultados

En este análisis se ha comenzado con un análisis bivariado, donde se han combinado cada uno de los 34 genes obtenidos en el análisis univariado, con los 808 genes originales, generándose más de 20.000 combinaciones.

Tras realizar el estudio, solo se ha obtenido una combinación que mejore la supervivencia obtenida por los genes de forma individual en, al menos, un 5%.

En la combinación de genes MRPL23 + MTA3, cuando ambos están infraexpresados, se obtiene una tasa de supervivencia del 99.28%. De forma individual, la supervivencia de MRPL23 infraexpresado es del 94.21%, y de MTA3 infraexpresado de 93.30%, por tanto, la combinación de ambos genes, supera la supervivencia de MRPL23 (que es la mayor de las supervivencias individuales) en un 5.38%.

Gen 1	Supervivencia Gen 1	Gen 2	Supervivencia Gen 2	Combinación	Supervivencia combinación	Valor p
MRPL23	*0,9421*	*MTA3*	*0,9330*	*MRPL23+MTA3*	*0,9928*	0,0007

Tabla 4: Supervivencia a 5 años de la combinación de genes MRPL23 + MTA3

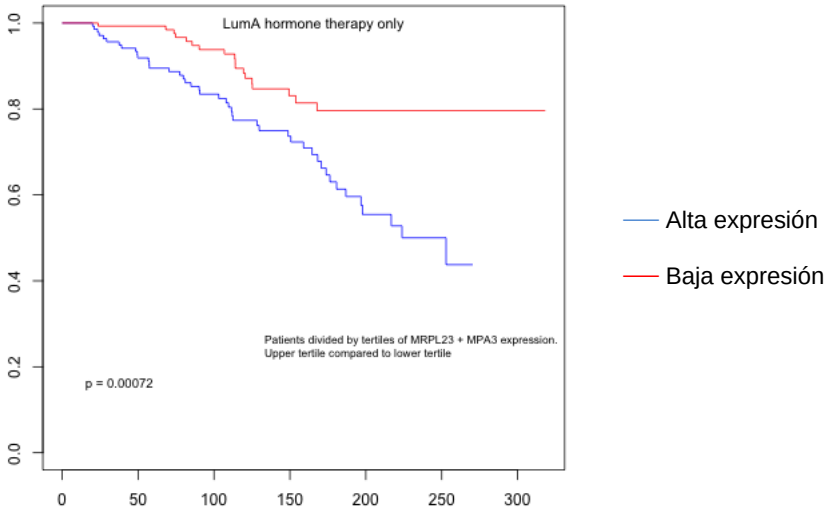


Figura 9: Curvas Kaplan-Meier de la combinación de genes MRPL23 + MPA3

Con el análisis bivariado ya se ha obtenido una supervivencia a 5 años del 99,28%, lo que hace imposible que se pueda incrementar en un 5% al combinar 3 genes. Por este motivo, no se continúa con el análisis multivariado.

4.4 Regresión Lasso

Tras ejecutar la selección sobre un modelo lineal, mediante regresión Lasso, se obtiene una lista de 13 genes, de los 808 incluidos en el modelo, que tienen influencia sobre la supervivencia (*tabla 5*).

Gen	***Coeficiente***	***Interpretación***
U2SURP	*-0.103*	*Incremento de la supervivencia cuando está sobreexpresado*
MYO6	*-0.102*	
SCP2	*-0.044*	
PDLIM5	*-0.029*	
FIP1L1	*-0.013*	
FBP1	*0.001*	*Incremento de la supervivencia cuando está infraexpresado*
PRPF19	*0.005*	
ASH2L	*0.040*	
CSTF2	*0.048*	
MRPL23	*0.051*	
AGR2	*0.070*	
CAD	*0.073*	
RBP2	*0.103*	

Tabla 5: Combinación de genes obtenidos mediante regresión Lasso

Estos genes, sin embargo, tienen un coeficiente muy próximo a 0, por lo que no se puede asumir que tengan mucha fuerza sobre el modelo. Además, los

valores están muy próximos entre sí, por lo que no hay un gen claramente influyente sobre los demás de la lista obtenida. Los resultados obtenidos muestran tendencias, por lo que se deben interpretar con prudencia.

Hay algunos genes obtenidos en este apartado que coinciden con los obtenidos en el análisis univariante (*tabla 6*). En ambos análisis, además, coincide cómo tiene que estar la expresión del gen (incrementada o disminuida) para que se produzca un aumento en la supervivencia, lo que confirma la influencia de estos genes en la respuesta al tratamiento.

Incremento de la supervivencia con

Baja expresión	*Alta expresión*
MRPL23	*SCP2*
CAD	*MYO6*
FBP1	*PDLIM5*
RBP2	*U2SURP*

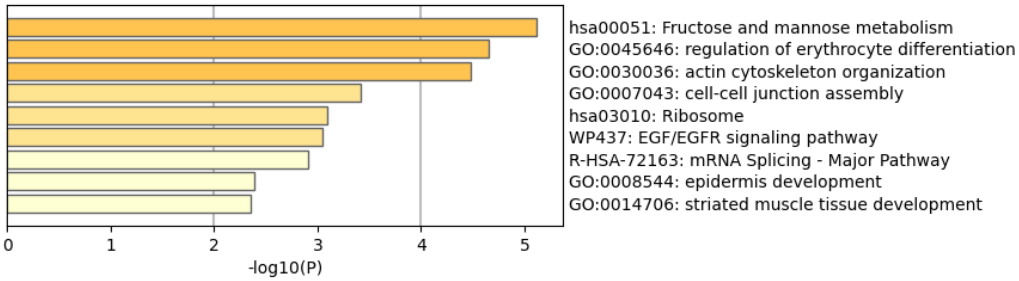
Tabla 6: Resultados obtenidos en análisis univariante y regresión Lasso

4.5 Análisis funcional de genes

Para realizar el análisis funcional de los genes y conocer en qué rutas metabólicas participan, se ha empleado la herramienta web Metascape.

El análisis funcional se ha hecho sobre la lista de genes significativos obtenidos en el análisis univariante (Tabla 2 y 3).

Las principales rutas metabólicas obtenidas se muestran en la figura 9. Los resultados se muestran también en un gráfico de barras en función del valor -*log10(P).*



GO	Category	Description	Count	%	Log10(P)	Log10(q)
hsa00051	KEGG Pathway	Fructose and mannose metabolism	3	8.82	-5.16	-0.81
GO:0045646	GO Biological Processes	regulation of erythrocyte differentiation	3	8.82	-4.69	-0.69
GO:0030036	GO Biological Processes	actin cytoskeleton organization	6	17.65	-4.55	-0.69
GO:0007043	GO Biological Processes	cell-cell junction assembly	3	8.82	-3.45	-0.06
hsa03010	KEGG Pathway	Ribosome	3	8.82	-3.13	0.00
WP437	WikiPathways	EGF/EGFR signaling pathway	3	8.82	-3.09	0.00
R-HSA-72163	Reactome Gene Sets	mRNA Splicing - Major Pathway	3	8.82	-2.94	0.00
GO:0008544	GO Biological Processes	epidermis development	3	8.82	-2.42	0.00
GO:0014706	GO Biological Processes	striated muscle tissue development	3	8.82	-2.39	0.00

Figura 10: Principales vías metabólicas obtenidas por Metascape de los 34 genes obtenidos en el análisis univariante

Principalmente, estas rutas están involucradas procesos inmunológicos y procesos celulares como el metabolismo, la respuesta a estímulos o el desarrollo (figura 10).

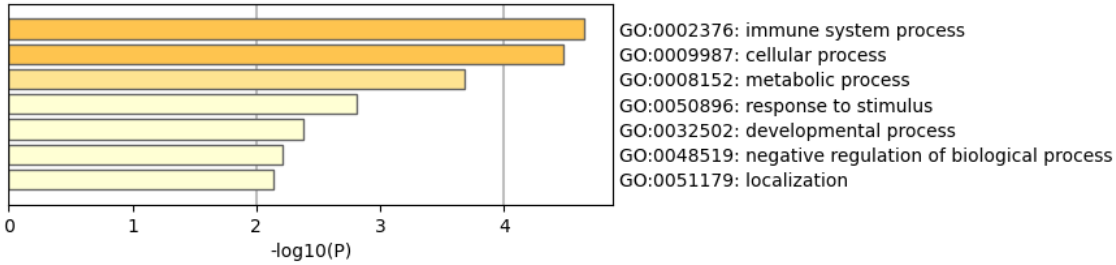


Figura 11: Principales rutas metabólicas de Gene Ontology

A continuación, se muestra de forma gráfica el mapa de enriquecimiento (figura 11), donde se conectan aquellos términos enriquecidos con una similitud mayor a 0,3.En este tipo de gráficos, los conjuntos de genes que se superponen tienden a agruparse, lo que facilita la identificación de módulos funcionales.

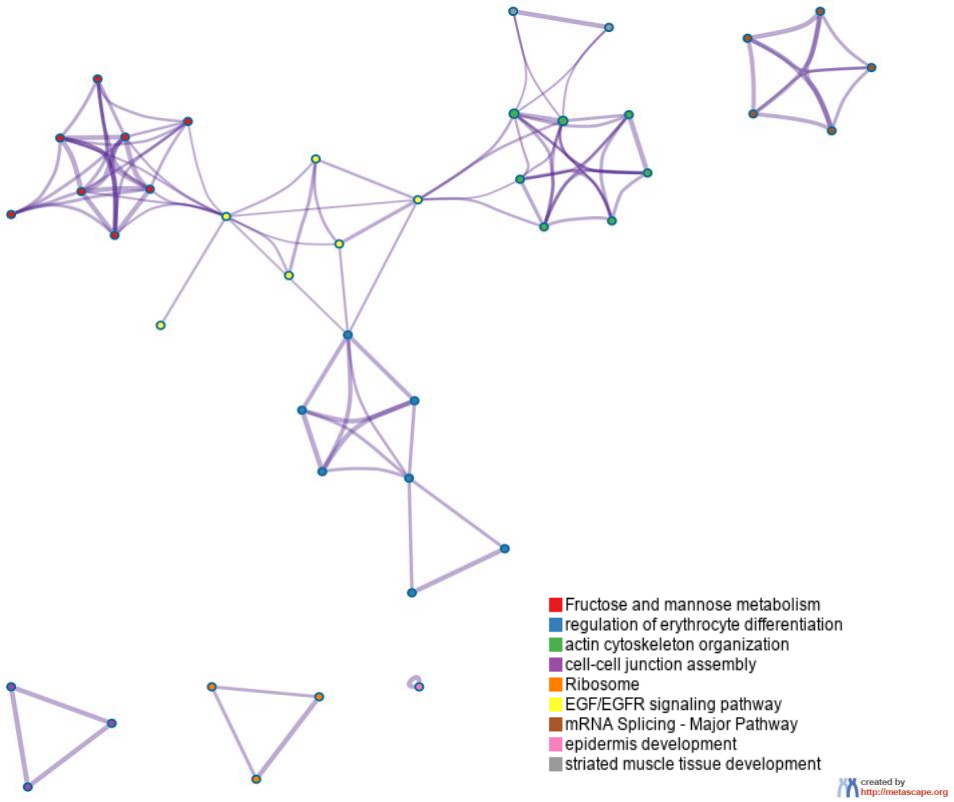


Figura 12: Mapa de enriquecimiento obtenido en Metascape

En el gráfico se muestra la relación entre aquellas rutas involucradas en el metabolismo de azúcares, la diferenciación de los eritrocitos, la organización del citoesqueleto de actina, el desarrollo del músculo estriado y la ruta de señalización mediada por el factor de crecimiento epidérmico (EGF/EGFR).

A la hora de evaluar estos resultados hay que tener en cuenta que Metascape es una herramienta en constante cambio y evolución, por lo que el análisis obtenido es poco estable. Se pueden obtener resultados ligeramente diferentes en función de la actualización. Los resultados mostrados en este estudio se han obtenido durante el mes de Diciembre de 2021.

En el caso de la herramienta web STRING , también se ha hecho el análisis sobre las 34 proteínas correspondientes a los genes anteriores (tablas 2 y 3). La versión empleada es la 11.5. El resultado obtenido se muestra en la figura 12. En este gráfico, las líneas unen aquellas proteínas con asociación funcional, siendo el grosor de la línea proporcional a la fuerza de los resultados.

La función molecular de estas proteínas, según Gene Ontology, se muestran en la tabla asociada a la figura 12.

Molecular Function (Gene Ontology)				
GO-term	description	count in network	strength	false discovery rate
GO:0030898	Actin-dependent atpase activity	3 of 24	1.86	0.0354
GO:0000146	Microfilament motor activity	3 of 29	1.77	0.0398
GO:0003723	RNA binding	13 of 1649	0.66	0.0099
GO:1901363	Heterocyclic compound binding	22 of 5831	0.34	0.0398
GO:0097159	Organic cyclic compound binding	22 of 5916	0.33	0.0398
GO:0005488	Binding	32 of 12516	0.17	0.0442

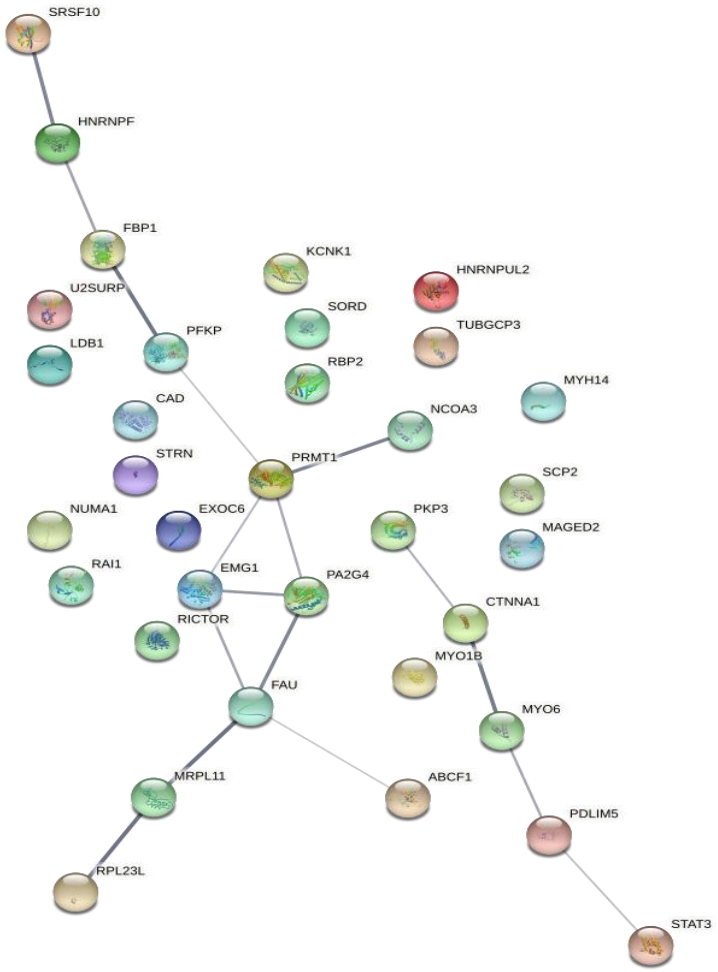


Figura 13: Análisis funcional de proteínas del análisis univariante obtenida en STRING

Si en vez de en la actividad que ejerce la proteínas, se hace un análisis de las rutas metabólicas donde participan, STRING sólo aporta un resultado estadísticamente significativo: Metabolismo de fructosa y manosa (Tabla 7):

KEGG Pathways				
pathway	description	count in network	strength	false discovery rate
hsa00051	Fructose and mannose metabolism	3 of 32	1.73	0.0136

Tabla 7: Rutas metabólicas significativas de las proteínas del análisis univariante obtenidas mediante STRING

Además, STRING permite obtener los tejidos donde estas proteínas se encuentran expresadas mayoritariamente (tabla 8)

Tissue expression (TISSUES)				
tissue	*description*	*count in network*	*strength*	*false discovery rate*
BTO:0000580	Blood cancer cell	10 of 1098	0.72	0.0090
BTO:0000763	Lung	10 of 1162	0.69	0.0117
BTO:0000203	Respiratory system	11 of 1436	0.64	0.0117
BTO:0000759	Liver	13 of 1882	0.6	0.0081
BTO:0000345	Digestive gland	16 of 2645	0.54	0.0058
BTO:0001491	Viscus	22 of 5020	0.4	0.0058

Tabla 8: Tejidos donde se encuentran expresadas las proteínas del análisis univariante. Obtenido por STRING.

5 Discusión

El cáncer de mama Luminal A es el subtipo de cáncer de mama con mejor pronóstico y tasa de supervivencia intrínseca, aproximadamente 91.6% a los 5 años, por lo tanto la búsqueda de signaturas génicas que permitan mejorar este dato es un análisis ambicioso. Cuanto mayor es la supervivencia de base, más difícil se hace mejorarla.

Los análisis ejecutados en este trabajo muestran que sí existen algunos genes que permiten incrementar, en un estrecho margen, la tasa de supervivencia inicial, en función de su nivel de expresión.

La supervivencia se puede mejorar con la variación en la expresión de un solo gen, como muestra la lista de 34 genes significativos obtenidos en el análisis univariante (tablas 2 y 3). O con la combinación de dos genes, como es el caso de la combinación MRPL23+MTA3 (tabla 4). Cuando ambos genes se encuentran infraexpresados, se obtiene una supervivencia en el modelo superior al 99%.

Una vez estudiadas las posibles combinaciones de genes para mejorar el valor puntual de supervivencia a 5 años, se deciden explorar otras alternativas con las que buscar modelos que predigan una mejor respuesta al tratamiento.

Por ello, se ejecuta una regresión Lasso sobre el modelo. El resultado muestra una combinación de 13 genes que permiten obtener una curva de Kaplan Meier situada a mayor distancia del resto de modelos (tabla 5). Constituyen aquello genes significativos que ejercen influencia real sobre el modelo multivariante.

Dentro de estos 13 genes, hay 8 que coinciden en su influencia con los obtenidos en el análisis univariante:

- Mejoran la supervivencia cuando el nivel de expresión es alto: SCP2, MYO6, PDLIM5 y U2SURP
- Mejoran la supervivencia con niveles de expresión bajos: MRPL23, CAD, FBP1 y RBP2

Atendiendo a esta lista de genes, alguno de ellos sí ha mostrado en diversos estudios tener influencia o estar involucrados en procesos que modifican el pronóstico en cáncer de mama o en otros tipos de cáncer:

1. **FBP1** (*Fructosa-1,6-bifosfatasa 1*): El metabolismo de glúcidos generalmente se encuentra alterado en las células tumorales. Este gen se encuentra sobreexpresado en tumores luminales. La mayoría de artículos asocian un buen pronóstico cuando la expresión es baja[26], como ha ocurrido en este estudio, sin embargo, existen estudios que muestran evidencias contradictorias en este aspecto[27].
2. **CAD** (*Carbamoil-fosfato-sintetasa*): este gen se encuentra sobreexpresado en las células tumorales con elevadas concentraciones de 17-βEstradiol. Participa en la síntesis de pirimidinas, necesarias para la progresión de la célula tumoral[28], por tanto, se espera encontrar mejor pronóstico cuando el gen se

encuentra infraexpresado. A esto contribuye el tratamiento con tamoxifeno, que actúa bloqueando el receptor de estrógenos.

3. **RBP2** (*Retinoblastoma-binding-protein-2*): Actúa desmetilando histonas y está relacionado con la aparición de resistencia al tratamiento con tamoxifeno[29], por lo que un bajo nivel de expresión de este gen se asocia con un incremento de supervivencia.

4. **MRPL23** (*Mitochondrial ribosomal protein L23*): el perfil de proteínas ribosomales se modifica en los tejidos tumorales, y varía en función del tipo de tumor[30]. La proteína MRPL23 o RPL23 se ha visto que puede tener una coamplificación con el gen HER2. El cáncer luminal A es HER2 negativo. Los subtipos HER2 positivos tienen peor pronóstico. En este estudio se ha obtenido que un descenso en la expresión de MRPL23, y por tanto se asume que habría un descenso de HER2, incrementa la supervivencia.

   - **MTA3** (*Metastasis associated 1 family member*): El análisis bivariante ha indicado que la combinación de MRPL23 con MTA3, cuando ambos tienen bajo nivel de expresión, incrementa la tasa de supervivencia hasta el 99,3%. Este gen se regula positivamente gracias al receptor de estrógenos, y reprime selectivamente al factor de transcripción Snail y la transcripción de E-Cadherina, que ayudan al mantenimiento de la arquitectura epitelial[31]. La terapia endocrina, al actuar bloqueando el receptor de estrógenos, disminuiría la expresión de MTA3.

5. **U2SURP** *(U2 SnRNP Associated SURP Domain Containing)*: Este gen es uno de los componentes del spliceosoma, participando en el proceso de splicing. Se ha visto que actúa como predictor independiente del pronóstico del cáncer de mama[32], y que su inhibición frena el crecimiento celular. Sin embargo, en este estudio se han obtenido datos contradictorios, ya que un nivel de expresión alto ha mostrado un incremento de la supervivencia en la población estudiada.

6. **SCP2** (*Sterol carrier protein 2*): Este gen participa en el metabolismo y transporte de lípidos intracelulares, fundamentales para el desarrollo y continuación del ciclo celular. Regula la angiogénesis y migración tumoral. Hay estudios que muestran que un descenso en la expresión de este gen disminuye la viabilidad de las células tumorales[33], sin embargo, en este trabajo se ha encontrado que un aumento de su expresión predice mejor respuesta al tratamiento.

Muchos de estos genes muestran resultados contradictorios con la bibliografía publicada en relación con la progresión y pronóstico del cáncer, por lo que se deben realizar más estudios para confirmar la implicación de estos genes en la supervivencia. Los resultados obtenidos son exploratorios.

Además de estudiar los genes de forma individual, el análisis funcional de los mismos, realizado con las herramientas web Metascape y STRING, ha

mostrado diferentes rutas metabólicas significativas donde intervienen estos genes. Para ampliar el rango de búsqueda, en este análisis se han incluido los 34 genes obtenidos en el análisis univariante (tablas 2 y 3).

La ruta metabólica mostrada por STRING, y que también se encuentra en el análisis con Metascape, es el metabolismo de fructosa y manosa. Como se ha comentado anteriormente, el metabolismo de glucosa en las células tumorales se encuentra alterado. Estas células consumen mucha más glucosa que una célula normal, incrementándose incluso la glucolisis anaerobia en el citosol, lo que se conoce como efecto Warburg[34]. Ante esta alteración del metabolismo, muchos estudios asumen el incremento del metabolismo de fructosa como una vía que emplea la célula tumoral para la obtención de energía, y por tanto, se asocia con la progresión tumoral[35, 36].

Con este trabajo se han identificado distintas signaturas génicas (genes individuales o combinaciones de genes) que mejoran la respuesta al tratamiento en pacientes con cáncer de mama luminal A en tratamiento endocrino. Además, alguno de estos genes participa en una ruta metabólica involucrada en la progresión del cáncer.

Conocer estos resultados puede orientar, por ejemplo, a futuros estudios en la búsqueda de nuevas dianas terapéuticas que, empleadas de forma concomitante al tratamiento endocrino, mejoren la supervivencia de estos pacientes. Así como mostrar la participación e influencia de vías metabólicas en el desarrollo y respuesta del cáncer de mama.

6 Conclusiones

6.1 Conclusiones

- Existen diferentes estrategias para valorar la respuesta al tratamiento. Se puede intentar obtener la mejor de las supervivencias a un tiempo determinado (análisis univariante y multivariante), o buscar aquel modelo que genere la mayor diferencia entre curvas de supervivencia (Regresión Lasso).
- El cáncer de mama luminal A tiene una elevada tasa de supervivencia basal, por lo que incrementar la supervivencia del mismo es difícil.
- Existen 34 genes que, en función de su nivel de expresión, predicen una mejor respuesta al tratamiento /tablas 2 y 3).
- La combinación de genes MRPL23 y MTA3 con baja expresión, genera una supervivencia del 99.3%. Sin embargo, para conocer la relación de estos genes con el cáncer de mama se necesitan más estudios.
- De entre los 808 genes estudiados, solo 13 se consideran significativos en el modelo para la mejora de la respuesta al tratamiento (tabla 5). Sin embargo, sus coeficientes son muy próximos a 0, por lo que el modelo debe interpretarse con prudencia.
- Tanto en el análisis univariante como por regresión Lasso se obtiene una mejora de la supervivencia cuando:
    - SCP2, MYO6, PDLIM5 y U2SURP están sobreexpresados.
    - MRPL23, CAD, FBP1 y RBP2 están infraexpresados
- La ruta del metabolismo de fructosa y manosa tiene relación con la supervivencia en el cáncer de mama.

6.2 Líneas de futuro

Este trabajo es un análisis exploratorio. Se necesitan más estudios para validar los resultados obtenidos, así como realizar la corrección por *multiple testing* de los genes obtenidos como significativos en todos los análisis.

Además, se puede repetir el estudio con la supervivencia libre de recaída (tiempo desde la remisión del tumor hasta que vuelve a aparecer) y analizar si existen diferencias en los resultados obtenidos al variar este dato.

Los resultados obtenidos se basan en pacientes del subtipo luminal A. Sería interesante conocer la diferencia de resultados con otros tipos de cáncer de mama donde el tratamiento endocrino sea una posible terapia, y así obtener aquellos genes, tanto comunes como diferentes, que sirvan como predictores de la respuesta.

6.3 Seguimiento de la planificación

La planificación del trabajo en base al diagrama de Gantt propuesto en la planificación del trabajo creo que ha sido la correcta. Sin embargo, siempre existen factores inesperados que pueden alterar la idea original.

En este caso, principalmente por motivos laborales inesperados y cuya fecha incial iba a ser posterior a la entrega de este trabajo, la fase 2 (entrega de la PEC3 y desarrollo de la memoria) ha estado algo afectada, retrasándose más de lo previsto.

Como propuesta de mejora, se podría haber incluido un periodo en el diagrama inicial libre, para posibles contratiempos, y haber ido entregando las distintas fases antes del tiempo estipulado.

7 Bibliografía

1. Observatorio AECC. Disponible en: https://observatorio.aecc.es/explora/ dimensiones-del-cancer
2. Tan P, Ellis I, Allison K, Brogi E, Fox S, Lakhani S et al. The 2019 World Health Organization classification of tumours of the breast. 2020;77(2):181-185. doi: 10.1111/his.14091.
3. Alcaide Lucena M, Rodríguez González CJ, de Reyes Lartategui S et al. Clasificación actual del cáncer de mama. Implicación en el tratamiento y pronóstico de la enfermedad. Cir Andal. 2021;32(2):155-59. DOI: 10.37351/2021322.9
4. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747–752. DOI: 10.1038/35021093
5. Ayala de la Peña, F., Andrés, R., Garcia-Sáenz, J.A. et al. SEOM clinical guidelines in early stage breast cancer (2018). Clin Transl Oncol. 2019; 21:18-30. DOI: 10.1007/s12094-018-1973-6
6. Cardoso F, Kyriakides S, Ohno S, et al. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Annals of Oncology 2019; 30: 1194–1220. DOI: 10.1093/annonc/mdz17
7. Kimberly H. Allison, M. Elizabeth H. Hammond, et al. Estrogen and Progesterone Receptor Testing in Breast Cancer: ASCO/CAP Guideline Update. Journal of Clinical Oncology 2020 38:12, 1346-1366. DOI: 10.1200/JCO.19.02309
8. Lyndsay N., Nofisat L, McShane F, et al. Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy for Women With Early-Stage Invasive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline. J Clin Oncol 2016; 34:10:1134-1150. DOI: 10.1200/JCO.2015.65.2289
9. Rueda, O.M., Sammut, SJ., Seoane, J.A. *et al.* Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. *Nature* **567,** 399–404 (2019). https://doi.org/10.1038/s41586-019-1007-8.
10. Liu J, Lichtenberg T, Hoadley K, et al. An integrated TCGA Pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell 2018; 173(2): 400-416. doi:10.1016/j.cell.2018.02.052
11. Rueda, O.M., Sammut, SJ., Seoane, J.A. *et al.* Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. *Nature* **567,** 399–404 (2019). https://doi.org/10.1038/s41586-019-1007-8.

12. Brca_metabric, disponible en: http://www.cbioportal.org/study/summary?id=brca_metabric

13. Ohnstad H, Borgen E, Falk R, et al., Prognostic value of PAM50 and risk of recurrence score in patients with early-stage breast cancer with long-term follow-up. Breast Cancer Research (2017) 19:120. DOI 10.1186/s13058-017-0911-9

14. The Cancer Genome Atlas Program (TCGA): GDC Data Portal. Disponible en: https://portal.gdc.cancer.gov/exploration.

15. Study METABRIC. Disponible en: https://ega-archive.org/studies/EGAS00000000083.

16. Kleinbaum G, Klein M., Survival analysis, a self-learning text. Third Ed. London: Springer; 2020.

17. San José B., Pérez E., Madero R., Métodos estadísticos en estudios de supervivencia. APC (2009) Vol 7, 55-59. DOI: 10.1016/S1696-2818(09)70453-6.

18. Kumar Goel M., Khanna P., Kishore J., Understanding survival analysis: Kaplan-Meier estimate. International Journal of Ayurveda Research (2010), Vol1-4. DOI: 10.4103/0974-7788.76794.

19. CRAN - Package Survival. Disponible en: https://cran.r-project.org/web/packages/survival/index.html

20. CRAN - Package rms. Disponible en: https://cran.r-project.org/web/packages/rms/index.html

21. CRAN - Package survmines. Disponible en: https://cran.r-project.org/web/packages/survminer/index.html

22. Faraway J, Linear Models with R. Second Ed. ISBN: 13: 978-1-4398-8734-9. CRC Press.

23. CRAN - Package survmines. Disponible en: https://cran.r-project.org/web/packages/glmnet/index.html

24. Zhou et al., Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nature Communications (2019) 10(1):1523.

25. Szklarczyk et al. Nucleic acids research 47.D1 (2018): D607-D613.2

26. Dong C, et al. Loss of FBP1 by Snail-Mediated repression provides metabolic advantages in basal-like breast cancer, Cancer cell 23 (2013) 316-331

27. Li K. et al., Fructose-1,6-bipfosphatase is a novel regulator of Wnt/beta-Catenin pathway in breast cancer, Biomedicine & Pharmacotherapy 84 (2016)1144-1149

28. Khan S. et al., Estrogen receptor/Sp1 complexes are required for induction of cad gene expression by 17β-estradiol in breast cancer

cells. Endocrinology 144(6):2325-2335.

29. Choi H, et al., Role of RBP2-induced ER an IGF1R-ErbB signaling in tamoxifen resistance in breast cancer. JNCIJ Natl Cancer Inst (2018) 110(4):djx207. DOI: 10.1093/jnci/djx207.

30. Huang Z, Duan H, Li H, Identification of Gene Expression Pattern Related to Breast Cancer Survival Using Integrated TCGA Datasets and Genomic Tools, Bioinformatic applications in life sciences and technology (2015). Doi: 10.1155/2015/878546

31. Fujita N, Jaye D, Kajita M, et al., MTA3, a Mi-2/NuRD complex subunit, regulates an invasive growth pathway in breast cancer, Cell, 2003 Apr;13(2):207-219. Doi: 10.1016/S0092-8674(03)00234-4

32. An J, Luo Z, An W, et al., Identification of spliceosome components pivotal to breast cancer survival. RNA Biology, 18:6, 833-842, DOI: 10.1080/15476286.2020.1822636

33. Zou J, Pei X, Xing D, Wu X, Chen S. LINC00261 elevation inhibits angiogenesis and cell cycle progression of pancreatic cancer cells by upregulating SCP2 via targeting FOXP3. J Cell Mol Med. 2021 Oct;25(20):9826-9836. doi: 10.1111/jcmm.16930.

34. Liberti M, Locasale JW. The Warburg Effect: How Does it Benefit Cancer Cells? Trends Biochem Sci. 2016 Mar;41(3):211-218. doi: 10.1016/j.tibs.2015.12.001.

35. Fan X, Liu H, Liu M, et al., Increased utilization of fructose has a positive effect on the development of breast cancer. PeerJ. 2017 Sep 27;5:e3804. doi: 10.7717/peerj.3804.

36. Krause N, Wegner A. Fructose Metabolism in Cancer. Cells. 2020 Dec 8;9(12):2635. doi: 10.3390/cells9122635.

8 Anexos

Anexo 1: Script para la obtención de los dataset de trabajo

```
clin_data<-read_delim("Datos/41586_2019_1007_MOESM8_ESM.txt", "\t",
escape_double = FALSE, trim_ws = TRUE) #Importar datos clínicos

LumA<-subset(clin_data, Pam50Subtype=="LumA") #Pacientes con cáncer de
mama luminal A

LumA_HT<-subset(LumA,((HT!="NO/NA")&(CT=="NO/NA")))#Tratamiento con
hormonoterapia y no quimioterapia

expr_data<-read_delim("Datos/brca_metabric/data_expression_median.txt"
, "\t", escape_double = FALSE, trim_ws = TRUE) #Importar datos de
expresión de METABRIC

m<-as.matrix(expr_data[,-c(1,2)])

rownames(m)<-expr_data$Hugo_Symbol

LumA_HT2<-LumA_HT[which(LumA_HT$METABRIC.ID %in% colnames(m)),] #Datos
clínicos de los pacientes con datos de expresión

m <- m[,which(colnames(m) %in% LumA_HT2$METABRIC.ID)] # Datos de expre
sión de los pacientes de la muestra
```

Anexo 2: Listado de genes análisis univariado

```
gl <- read.table("Datos/gene_list_forElena.txt") #Importar la lista de
genes

gene_list <- gl$V2
```

ACACA, LRPPRC, MYH14, HSPA9, SNRNP200, GTF2I, PCK2, PNLIP, DDX5, TAGLN2, COPA, HNRNPF, FUBP3, SND1, TRAP1, IQGAP1, TOP2B, NUP205, RBM14, HNRNPR, CTTN, SNRNP70, SF3B2, NCOA5, MYO1C, DDX6, MTOR, U2AF2, HP1BP3, IARS1, CKAP4, CAD, RRBP1, TRRAP, TENM1, HK1, NUP155, PDS5B, HNRNPH2, HDLBP, MDH2, CNOT1, KIAA0368, EPRS1, ACTC1, PDLIM1, CAPN1, MYH10, EEF1A2, CTNNA1, COPG1, TKT, HSPA5, ACTN1, RPS17, MYO6, DNM1L, ITPR3, U2AF1, ALDOC, CCT4, EZR, PAICS, AP1B1, PA2G4, ESRP1, NUP210, TRIP12, RPS13, SNRPD1, ARPC2, PHGDH, PRPF6, SMARCA5, CCT6A, DNAJC13, HNRNPA0, USP5, PRDX2, SMC1A, AP2A1, CYFIP1, BAZ1B, RPS21, VCP, TPR, GLRX3, IPO4, RPS27, ADAR, ASH2L, NUP98, SLIRP, NUP93, DARS1, XPO5, DNM2, ARCN1, DEK, RPL35A, BAG3, CSDE1, IMMT, PDCD6IP, EXOC4, IQGAP3, SERBP1, PRDX5, MRPS31, PYCR1, ANXA6, GSPT1, H2BC5, INF2, SART1, PRPF31, SMCHD1, TUBGCP3, NUP188, NCKAP1, HMGB2, NQO1, STUB1, VAV2, ALDH16A1, SCRIB, GALK1, DDX42, MYOF, INPP4B, UBAP2L, AP3D1, GLYR1, VWA8, MARS1, RARS1, EIF2S1, CAPZA1, TRIP6, AP2M1, SART3, KIF5B, MCM2, PDS5A, SMC3, PHPT1, RPN1, RAB1A, CACYBP, TMPO, GSTK1, SRP68, PSMC3, INTS7, NUP160, WNT1, DYNC1I2, PI4KA, VASP, FTSJ1, DBNL, SSRP1, KTN1, EPS8L2, NAPA, GPI, CCT7, AP1G1, CUL3, WNT3, MYO1D, TUBG1, SORD, RPLP0, AGR2, ARPC1B, TANK, FAM83H, PFKP, HDAC1, MTA2, PRPF4B, STAU1, CTNND1, ATAD3A, UBTF, ABCF1, EIF3L, POLR2A, AKT2, FOLH1, SNRPD2, MISP, RSPO1, NT5C2, ZNF185, COPE, PSMD3, MCM4, TEX10, CLINT1, SCP2, SCYL2, PEBP1, S100A16, EIF1AX, TRA2A, CELF1, PPA1, SH3GLB2, EPRS1, SEC23B, FBP1, PSMD2, AP2B1, PRKD2, NAMPT, LARS1, KCNK1, HUWE1, NUP107, CMAS, GAK, ATAD2, RPA1, WASHC4, COPZ1, GDI2, CLIC3, YWHAB, SNW1, AICDA, IPO5, ZC3H18, CCDC47, PDLIM5, ANXA4, SRP72, XPOT, SNRNP40, PSMC5, AIMP1, H2AW, BIRC6, NUP35, SGPL1, ACOX1, MECP2, USP9X, PSME4, MSH2, DNPEP, FAM50A, EMD, SET, RECQL, SEC24C, SNX27, DNAH17, CLASP1, EIF3G, RNF213, HTT, CLPB, STAT3, RPL31, MRPL49, RBX1, RHOC, BCLAF1, EIF3H, TFAP2C, FAM120A, FAM83B, RPS6KB1, PRKAA1, MTDH, TBCD, USP39, RICTOR, NUP153, COPG2, AHCTF1, C7orf50, TOR1AIP1, CD2BP2, PRPF3, DYNC1LI1, AP3B1, PDAP1, ZC3H4, CBX5, VDAC2, PAWR, CORO1C, PDXK, PHF2, FAM98A, GIPC1, MLPH, PIK3C2A, PYM1, NIBAN2, CMYA5, STAT1, MARK2, DYNC1LI2, NBEAL2, EIF2A, PHIP, EXOC6, DNMT3A, NUP62, ACAD11, MCM6, EXOC8, NEDD1, INTS5, PSMD14, VPS11, EPB41L5, GTF3C3, RBBP4, FXR1, RBM8A, PSMC6, HMGN1, FAM98B, PPP1R12A, SEC22B, VPS35, SYNC, SKP1, CAPNS1, USP14, RPL35A, PMM2, DDB1, DNAJC9, COG5, VPS16, SMU1, POLR2B, ARFGEF3, SEH1L, NUP54, EXOC2, ST13, TMEM209, ARFGAP2, CCDC22, MED23, NPLOC4, NUP133, SMPD4, CNOT10, PRRC2B, UBE3C, ACTR1A, PSMA7, SMC4, EPS15L1, EXOC7, PICALM, PAPOLA, TUBGCP6, MMS19, PCNA, G3BP2, COG4, MKI67, FGF3, DRG1, LAD1, XPO7, DHX36, CDV3, NCAPG, GEMIN5, ACOX3, PAK4, AP2A2, EIF3F, LARP4, RAE1, PGLYRP2, AHNAK2, MSMO1, ENO1, HSD17B4, GPD2, ALDOA, ANXA2, HSP90AA1, XRCC6, GLUD1, MACROH2A1, PREX1, KRT5, G6PD, PGK1, PSMA3, MTHFD1, PCBP1,

H1-1, PGD, VANGL1, TMPO, ESR1, NPM1, YBX1, JUP, STRN, RPL9, RAET1E, TUBB4B, RPS4X, RPS15A, RPS16, RAN, HNRNPH3, PFKL, RPL7A, RPS3A, PICALM, CLPX, HRAS, PFN1, RPL10A, RPL10A, CASP8, FAU, LMNB2, RXFP3, ATP1A3, ATP1B1, ANPEP, SRSF2, MDC1, RPL23, RPL13, EWSR1, HRAS, EEF1G, RPS14, RPSA, EIF5A, ESD, CBX3, H4C1, RPL23, NACA, MYEF2, PRMT1, RPL24, RPS10P5, TRIM33, DAZAP1, RSL1D1, EEF1D, YWHAE, DLG1, STRN4, MRPL11, LUC7L2, EIF2S3, EML4, GATA3, KPNB1, CNN2, S100P, CPSF1, CSTF2, MOB4, ESRP2, FBXL20, DDX21, CSTF3, NCOA3, RBM12B, PABPN1, SUMO1, DNAJA1, TRIM24, MYNN, TAF15, CSTF1, GART, SUN1, LUC7L3, PPP1CA, OLA1, PTBP3, NOP56, TIAL1, RBM12, CUL4B, TMEM160, SF3B4, SON, OGT, NXF1, TXNDC17, CTBP1, CLCC1, PARN, STRIP1, TRPS1, ARMT1, ZFR, NAP1L1, WTAP, RXRA, NOP2, DSTN, CDK1, NUDT16L1, AKAP8, UBR2, WDR3, WDR36, PFN2, UBXN1, ATP1A1, ATIC, MBNL1, ANXA7, PDCD6, NFIC, NOC4L, TLE1, SMARCD2, SUGP2, BSCL2, FBRS, CEP131, ABCB9, WDR33, CLASP1, SRI, CTCF, SNRPB, CPSF2, IFNG, ITGB4, DSG2, PHB2, AKAP8L, RAVER1, RAI1, LDB1, RRP12, RBM22, NFYC, HMCES, ZNF207, DCAF7, BCL9L, CUL4A, TLE3, NOL6, DHX38, RBM33, HNRNPLL, PPFIBP1, NKRF, RTF2, DPF2, CASZ1, AP1S1, PTH, MRPL38, FADD, EML1, GATAD2A, PCAT1, MRPL23, RAB11A, TCF25, NUP37, EMG1, SIRT5, ARF6, ROCK2, AES, DHRS2, STK25, TIMM44, PSME1, DIP2B, EIF2AK2, CHD7, PTPA, INTS8, RAPH1, UQCRC2, NOA1, RAB21, SREBF2, SARS2, ITPR2, POLQ, ZC3HAV1, SEC23A, NAGA, MRPS30, EXOC3, DNAJA2, TNKS1BP1, CHTF18, PLCB3, MRPL3, PCM1, ACAD10, HSPH1, PHF14, VPS33A, BCKDHA, PES1, HSDL2, NELFE, CNOT9, LSM3, PPP1R3A, MRRF, ACTL6A, FZD9, NDC1, LRWD1, MRPL40, DOP1B, MAP7, PRPF4B, MTHFD2, CMC2, MARK3, ARL6IP4, BYSL, SULT2B1, DCTN2, MTA3, MAGED2, ZNF512, EIF3M, GPD2, ZNF24, GNAI2, KIDINS220, LARP1, SNTB2, VAPA, RNPEP, YY1, ZNF385A, PRKDC, FLNB, RPS3, HM13, UGDH, RPL15, G3BP1, PCLO, HMGB1, SRSF7, RYR3, RPS25, RBM25, ERH, HIP1R, AIFM1, ADNP, AHCY, ANXA11, BTF3, BUB3, CAND1, CAPZB, CCAR2, CHD4, CIRBP, CLIC1, COPB2, CPNE3, CPSF6, CPSF7, CRIP2, CSE1L, CTBP2, DDX1, DDX23, DDX46, DHRS4, DHX15, EIF3A, EIF3B, EIF3C, EIF3D, EIF3E, EIF3I, ELAVL1, FEN1, FIP1L1, FKBP4, GCN1, GSTM3, GTPBP4, HCFC1, HMGB3, JPT2, KRT86, LDHA, MCM3, MCM5, MCM7, MYO1B, NAT10, NMD3, NME2, HNRNPUL2, ERBB4, HNRNPUL1, ZNF638, RANGAP1, HNRNPAB, EIF4A3, CCT3, VARS1, SF1, SUPT16H, AARS1, UNC45A, NAP1L4, FBL, ZNF217, KHDRBS1, KRT80, TRIM25, HNRNPDL, RPL13A, RPS18, PAXIP1, LGALS3, H2AC1, S100A11, TUBB3, LLGL2, SEC16A, RPL19, SRSF10, EIF4B, RPS15, THRAP3, USP7, SNRPA1, GMPS, ZNF326, RPS20, L1RE1, EIF4H, YWHAQ, NUDT21, NUMA1, OTUB1, PARP1, PKP3, PPP1CB, PRPF19, PRPF8, PSMD5, PSPC1, PUF60, RALY, RBM10, RBM15, RBM3, RBM39, RBM4, RBM47, RBP2, RCC2, RNPS1, RTCB, SAFB, SAFB2, SF3B1, SF3B3, SLTM, SNRPA, SRRM2, SRRT, SRSF1, SRSF9, STIP1, TLN1, TNPO3, TPM3, TRA2B, TUFM, U2SURP, UBR4, UPF1, WDR1, WISP3, XPO1, XRCC5, XRN2, YLPM1, ZYX

Anexo 3: Script para el análisis univariante

```
time <- LumA_HT2$T/30 #Transformar el tiempo en meses
status <- LumA_HT2$DeathBreast
# Reiniciar las variables
pval<- c()
gene <- c()
survhigh5 <- c()
survlow5 <- c()
diff5 <- c()
surv5 <- c()
# Iteración en la lista de genes
for (i in c(1:length(gene_list)))
{
  g <- as.character(gene_list[i])
  if( g %in% rownames(m))
  {
    gene <- c(gene, g)
    e <- m[match(g,rownames(m)),]



#Divide el vector e en 3 terciles de expresión
```

```
    clus <- ifelse(e<quantile(e,probs=1/3),"low",ifelse(e>quantile(e,p
robs=2/3),"high",NA)) # change as appropriate

    dfsurv <- data.frame(time, status, expr=clus)

    fit <- survfit(Surv(time, status) ~ expr, data = dfsurv)

    p1 <-surv_pvalue(fit)


    #Creación de las curvas de supervivencia

    png(file=paste("Univariate/", g,".png", sep=""))

    plot(fit, col=c("blue", "red"), show.nrisk=F, hr.pos=NA, legend.po
s = "topright", stitle=NA, cex=0.5)

    legend("top","LumA hormone therapy only",bty="n",cex=0.9)

    legend(x=0, y=0.2, legend=as.character(p1$pval.txt), cex=0.8, box.
lty = 0)

    legend(x=120, y=0.3,paste(paste("Patients divided by tertiles of "
, g, " expression.", sep=""), "Upper tertile compared to lower tertile
", sep="\n"), cex=0.7, box.lty=0)

    dev.off()


    #Supervivencia a 5 años

    pos5 <- which(fit$time>60)

    pos51 <- pos5[1]

    pos52 <- pos5[which.max(diff(pos5))+1]


    #Supervivencia alta expresión

    survhigh5 <- fit$surv[pos51]

    #Supervivencia baja expresión

    survlow5 <- fit$surv[pos52]

    diff5 <- c(diff5, survhigh5-survlow5)

    surv5 <- c(surv5, max(survhigh5, survlow5))


    pval<- c(pval, p1$pval.txt)

    cat(i)

  }

}

res <- data.frame(gene, pval, survhigh5, survlow5, diff5) # Dataset de
resultado
```

Anexo 4: Script para el análisis bivariante

```
# Genes significativos del análisis univariante
```

```
gl_sig<-read_excel("Datos/gl_sig.xlsx")

X <- m[which(rownames(m) %in% gene_list),] #Datos de expresión de
todos los genes

X <- scale(t(X))

X <- as.data.frame(X)


# Combinación de genes


time <- LumA_HT2$T/30

status <- LumA_HT2$DeathBreast


gene1 <- gl_sig$Gen[1] #Numero correspondiente a gene1, cambiar para
realizar todas las combinaciones

gene_list1 <- gene_list[-which(gene_list %in% gene1)] # Combino con
todos los genes menos consigo mismo


# Reinicio de las variables

pval1<- c()

diff5 <- c()

surv5 <- c()

gene2<-c()

survhigh5<-c()

survlow5<-c()


  for (i in c(1:length(gene_list1)))

    {

    g2 <- as.character(gene_list1[i])

    e_h <- X[,match(gene1, names(X))]


    if( g2 %in% names(X))

      {

      gene2 <- c(gene2, g2)

      e_l <- X[,match(g2, names(X))]

      I <- e_h+e_l # Cálculo de la expresión sumando

      # I <- e_h-e_l Cálculo de la expresión restando

      clus1 <- ifelse(I<quantile(I,probs=1/3),"low",ifelse(I>quantile(
I,probs=2/3),"high",NA))

      dfsurv1 <- data.frame(time, status, expr=clus1)

      fit1 <- survfit(Surv(time, status) ~ expr, data = dfsurv1)
```

```
      p1 <-surv_pvalue(fit1)



      pos5 <- which(fit1$time>60)
      pos51 <- pos5[1]
      pos52 <- pos5[which.max(diff(pos5))+1]

      survhigh5 <- c(survhigh5, fit1$surv[pos51])
      survlow5 <- c(survlow5, fit1$surv[pos52])
      diff5 <- c(diff5, survhigh5-survlow5)
      surv5 <- c(surv5, max(survhigh5, survlow5))



      pval1<- c(pval1, p1$pval)
      cat(i)
    }
  }

res1<-data.frame(gene2, pval1, survhigh5, survlow5)
```

Anexo 5: Script Regresión Lasso

```
# Regresión Lasso
library(glmnet)
fit <- Surv(time,status)

cv.lassoFit <- cv.glmnet(X, fit, family="cox", alpha=1) # alpha=1 para
hacer regresión de Lasso
idealLambda <- cv.lassoFit$lambda.min
idealLambda1se <- cv.lassoFit$lambda.1se

co <- coef(cv.lassoFit, s=idealLambda)
rownames(co)[which(co!=0)] # Lista de genes predictores
co[which(co!=0)] # Coeficientes de los predictors
```