RNAS LARGOS NO CODIFICANTES PARA LA CLASIFICACIÓN DEL HEPATOCARCINOMA

Trabajo Fin de Máster

Máster en Métodos Computacionales en Ciencias

Facultad de Ciencias - Universidad de Navarra

Curso académico 2020-2021

Janire Erasun Martínez

Pamplona, 27 de Agosto de 2021

CERTIFICADO DE AUTORÍA

Los tutores abajo firmantes, **DECLARAN** que la memoria de Trabajo Fin de Máster (TFM) titulada: **RNAs largos no codificantes para la clasificación de hepatocarcinoma** forma parte del proyecto de investigación realizado por D. **Janire Erasun Martínez**, alumno del **Máster en Métodos Computacionales en Ciencias de la Universidad de Navarra**.

Así mismo, el alumno/a **CERTIFICA** que es autor/a de la memoria y que cumple con los requerimientos de originalidad y reconocimiento de fuentes recogidos en la normativa de la asignatura TFM.

Prof. D./Dña. Purificación Fortes Alonso
Tutor/a

Prof. D./Dña. Ibon Tamayo Uria
Tutor/a

D./Dña. Janire Erasun Martínez
Alumno/a

Pamplona, 27 de Agosto de 2021


Universidad
de Navarra

Tabla de contenido

ABSTRACT

El hepatocarcinoma es el tumor primario más común de los cánceres de hígado. La tasa de supervivencia a 5 años es del 18% debido al escaso número de tratamientos y a su pobre efectividad. Una mejor estratificación de los pacientes podría ayudar al diagnóstico, tratamiento y pronóstico de la enfermedad, e incluso revelar nuevas dianas accionables terapéuticamente. Los RNAs largos no codificantes (lncRNA) son moléculas con una función altamente reguladora a diferentes niveles. Además son mucho más específicas de tejido que los RNAs codificantes. Los lncRNAs están involucrados en diversas patologías y, entre ellas, el cáncer. Se ha descrito el transcriptoma no codificante del hepatocarcinoma pero nunca se ha usado para estratificar estos tumores. En este trabajo proponemos que dada su enorme especificidad de tejido y de tumor, los lncRNAs se podrían utilizar para estratificar a los pacientes en grupos de distinto pronóstico y descubrir firmas genéticas asociadas significativamente con la supervivencia. Todo ello permitiría el desarrollo de nuevas terapias y de un sistema que facilite el pronóstico y la selección de terapias más adecuadas para cada paciente.

1. INTRODUCCIÓN

1.1. El hepatocarcinoma

El hepatocarcinoma (HCC) es el tumor primario que causa el 75-90% de los cánceres de hígado. Implica un grave problema de salud ya que es la cuarta causa de muerte relacionada con cáncer. Actualmente, la incidencia es similar a la mortalidad, con una tasa de supervivencia a 5 años de sólo el 18% de los pacientes [1]. Esta alta mortalidad está relacionada con varios factores. Uno de ellos es que el hígado es un órgano esencial. Su función principal es procesar y detoxificar los nutrientes absorbidos por el tubo digestivo, almacenar algunos de ellos (por lo que tiene importantes funciones metabólicas) y secretar diversas sustancias (es la glándula más grande del organismo) [2]. Gran parte de estas funciones las realiza el hepatocito, una célula poliédrica que es el principal componente del hígado y que, cuando se maligniza, da lugar a un HCC [3]. La malignización ocurre, en parte, por la enorme capacidad de regeneración que tienen los hepatocitos [4]. En respuesta a un daño agudo (como el consumo esporádico de alcohol), algunos hepatocitos mueren y se genera una pequeña cicatriz que se resuelve generalmente con la regeneración de los hepatocitos circundantes que pasan a ocupar el hueco dejado en la zona dañada. Si el daño persiste en el tiempo (por ejemplo, el causado por el consumo habitual de alcohol o por infecciones por virus persistentes como el virus de la hepatitis B o C), se generan cicatrices que el hígado no tiene tiempo de reparar y que originan primero la fibrosis y después la cirrosis hepática, lo que modifica la estructura original del hígado e impide su funcionamiento óptimo [4]. El ambiente generado por la lesión continuada produce inflamación y los hepatocitos se dividen de forma desordenada intentando reparar el tejido dañado pero produciendo nódulos de regeneración. En ellos, el ambiente circundante y la rápida división celular, favorece la aparición de mutaciones y cambios epigenéticos con potencial de generar un hepatocarcinoma [5]. Algunos HCC se podrían generar a partir de células madre o progenitoras [6].

Por todo ello, las causas principales del HCC son las infecciones hepáticas crónicas, el consumo de alcohol, la obesidad o la ingesta habitual de drogas. Las infecciones con virus de la hepatitis B o C tienen tratamiento en la actualidad, por lo que está disminuyendo enormemente el número de pacientes que desarrollan HCC por estas causas. Sin embargo, la incidencia de esta enfermedad aumenta anualmente por la falta de políticas destinadas a disminuir la alcoholemia y la obesidad. Esta última causa hígado graso, esteatohepatitis no alcohólica y/o síndrome

metabólico y es la responsable del incremento de HCC en los últimos años [7]. También hay un porcentaje reducido de pacientes, alrededor de un 10%, que desarrollan HCC de forma espontánea [8]

También contribuye a la alta mortalidad que el HCC sea una enfermedad predominantemente asintomática. Cuando aparecen los síntomas, es muy común encontrar al paciente en un estadio muy avanzado. Las técnicas diagnósticas más utilizadas suelen ser no invasivas (TAC, ecografía y/o resonancia magnética) [9]. Una vez diagnosticado se miden en suero parámetros indicativos de la funcionalidad hepática (como la bilirrubina, la albúmina o las transaminasas) y los niveles de alfafetoproteína (AFP), una proteína de hígado embrionario que secretan las células de HCC y que es un marcador de mal pronóstico [10]. Otros marcadores de mal pronóstico se observan tras el análisis histopatológico de muestras obtenidas tras realizar una biopsia del tejido tumoral. Destaca, por ejemplo, la presencia de invasión vascular de células tumorales [11]. Sin embargo, ninguno de los marcadores pronóstico descritos hasta la fecha tiene un alto grado de sensibilidad y fiabilidad.

En aquellos casos en los que es posible diagnosticar al paciente en un estadío temprano (0 o A) se pueden aplicar tratamientos potencialmente curativos como la resección o el trasplante hepático, con lo que la supervivencia a 5 años es de un 60% [12]. Sin embargo, es muy común la recurrencia y la aparición de tumores de *novo* después del tratamiento [13, 14]. En estadíos intermedios (B) se suelen proponer tratamientos locales como la quimio o radioembolización, con lo que se obtiene una esperanza de vida de unos 16 meses [15]. En estadíos tardíos, donde se diagnostican la mayoría de los pacientes, sólo se pueden aplicar tratamientos sistémicos. Hasta hace pocos años el tratamiento en primera línea eran inhibidores de tirosina-quinasas, como el sorafenib, que conseguía prolongar la esperanza de vida en pocos meses. Recientemente el tratamiento de elección está basado en inhibidores del sistema inmune, con los que se ha conseguido una supervivencia de 16 meses [1]. Sin embargo, hay muchos pacientes que no responden a este tratamiento. Además, cuando el tumor progresa sólo se pueden aplicar tratamientos paliativos. Por todo ello, existe una gran necesidad de encontrar dianas que permitan el desarrollo de nuevas terapias. Tras décadas de estudio, se han identificado varios oncogenes o genes supresores tumorales que producen proteínas esenciales para la generación y progresión del HCC, pero que están implicadas en rutas que no se pueden bloquear o los fármacos dirigidos contra ellas han resultado ineficaces por otros motivos [1].

Comparativamente, el genoma no codificante se ha estudiado poco en el caso del HCC y podría ser una excelente fuente de marcadores de pronóstico y dianas terapéuticas.

1.2. El transcriptoma no codificante del HCC y los RNAs largos no codificantes (lncRNAs)

La mayor parte del genoma se transcribe para dar lugar a RNAs no codificantes. Entre ellos, destacan los lncRNAs por ser los más heterogéneos y, en general, menos conocidos. Los lncRNAs se caracterizan por ser parecidos a los RNAs mensajeros (mRNAs) puesto que se transcriben y se procesan de forma parecida, pero no son traducidos a proteína. Además, comparados con los mRNAs, los lncRNAs son más nucleares, menos abundantes, más específicos de tejido y mucho más numerosos, puesto que se estima que podría haber más de 100.000 genes que transcriben lncRNAs frente a sólo unos 30.000 que codifican proteínas. Comparados con otros RNAs no codificantes son más largos (más de 200 nucleótidos), y por ello tienen capacidad de unirse a otras moléculas de RNA, DNA o proteína y formar estructuras secundarias y terciarias que permiten su funcionalidad. La mayoría de los lncRNAs que se han estudiado en detalle sirven para regular la expresión génica tanto en *cis*, regulando la expresión de genes cercanos, como en *trans*, afectando la expresión o actividad de genes localizados lejos del que transcribe al lncRNA. Esta regulación ocurre a nivel de DNA (regulando la compactación del genoma o la transcripción), de RNA (afectando al procesamiento de intrones, la poliadenilación, la traducibilidad, la estabilidad o la localización) o de proteínas (modificando modificaciones post-transduccionales o la actividad enzimática). Estas actividades son esenciales para permitir la viabilidad de la célula, su capacidad de responder a estímulos y de volver a la homeostasis. Además, se han descrito muchos lncRNAs esenciales para la diferenciación y la división celular. Por ello, muchos de ellos se han relacionado con distintas enfermedades, incluido el cáncer.

Numerosos estudios han detectado lncRNA desregulados en HCC y otros tumores y se ha demostrado que algunos de ellos realizan funciones importantes para la proliferación celular, con lo que podrían ser excelentes dianas terapéuticas [16, 17]. Además, se ha descrito que los lncRNAs son mucho más específicos de tumor que los mRNAs. Por ello, se espera que también podrían utilizarse como biomarcadores, si se secretan a suero y para permitir estratificar a los tumores según su pronóstico de

manera mucho más fina que con las estrategias actuales basadas en genes codificantes [18, 19] [20]. Esto es especialmente relevante en el caso del HCC.

1.3. El transcriptoma codificante del HCC: estratificación y pronóstico

El enorme desarrollo de las técnicas de secuenciación y de bioinformática han permitido analizar los transcriptomas de varias cohortes de pacientes con HCC y conseguir firmas de genes codificantes que permiten estratificar a los pacientes [1]. De forma general, los tumores se dividen en dos clases: los proliferativos y los no proliferativos.

Los **tumores proliferativos** presentan una diferenciación muy pobre, inestabilidad cromosómica y mutaciones en *TP53*. Las rutas oncogénicas principales relacionadas con este grupo tumoral son la *TGF-beta, RAS/MAPK* y *PI3K/AKT*, y hay sobreexpresión en los genes involucrados en el ciclo celular y la supervivencia [1, 21]. Están asociados a un fenotipo más agresivo (clase de proliferación alta), presentan niveles de AFP (alfa-fetoproteína) más elevados y están asociados a un peor pronóstico. A nivel molecular, se encuentran en este grupo los subgrupos G1-G2-G3 [1,22], los subgrupos S1-S2 [23], el subgrupo *immune-high [24]* y los subrupos iClust1 y iClust3 [25].

Los **tumores no proliferativos** presentan un alto grado de diferenciación, estabilidad cromosómica y mutaciones en *CTNNB1*. Mantienen los marcadores de los hepatocitos, y las principales rutas oncogénicas implicadas son *JAK/STAT* y *Wnt/b-catenina [1]*. A nivel molecular, se encuentran los subgrupos G4-G5-G6 [1], el subgrupo S3 [23] y el subgrupo iClust2 [25].

Estas clasificaciones se realizaron con metodologías y criterios diferentes. Hoshida reunió varios datasets, donde utilizó parte como datos de entrenamiento [23]. Los subgrupos de muestras los definió por el método SubclassMapping, basado en las subclases identificadas con un clustering jerárquico, k-means y la matriz de factorización no-negativa. Wheeler, en cambio, realizó clusterings de diferentes inputs, y acabó realizando un clustering intregativo de todos ellos con iClustering [25]. A su vez, numerosos estudios utilizan tanto las muestras tumorales como las de tejidos sanos. Por otro lado, *se* utilizaron los índices clínico-patológicos como predictores de la clasificación molecular, logrando un método económico aplicable en todo el mundo

[26]. Más tarde, *se* compararon los perfiles genómicos y transcriptómicos en cuatro modelos de ratón para evaluar la adecuación de los modelos, determinando que el knockout de TAK1 reflejaba mejor la firma mutacional humana y era transcripcionalmente similar a los tumores de bajo grado, mientas que el Animal Modelo Estélico (STAM) se asemejaba más a los tumores proliferativos transcriptómicamente, presentaba mutaciones frecuentes en CTNNB1 y rutas alteradas similares [27]. Además, *se corroboró* el uso de líneas celulares como modelos para evaluar la respuesta a fármaco de los subtipos moleculares específicos S1-S2 y S3 [28]. Por último, *se* encontraron drivers mutacionales adicionales en la hepatocarcinogénesis de pacientes de Mongolia, a la vez que se vieron subtipos moleculares comunes con el colangiocarcinoma, demostrando diferencias entre etnias [29, 30].

2. OBJETIVOS

Investigaciones anteriores han realizado estudios transcriptómicos de genes codificantes para conseguir estratificar a los pacientes con HCC en distintos grupos de relevancia clínica. Dado que los genes no codificantes son más específicos de tejido y de tumor que los genes codificantes, deberían permitir una estratificación de pacientes más detallada .

Por ello, usaremos distintas cohortes de pacientes para ver si el transcriptoma no codificante sirve para estratificar pacientes. Usaremos todo el genoma no codificante o distintos subgrupos de genes. Se evaluará la relación con la supervivencia de los grupos de pacientes resultantes. El objetivo final es encontrar una firma genética sólida que tenga asociación estadística con la supervivencia del paciente. Para ello compararemos distintas técnicas y realizaremos una evaluación crítica de cómo afecta cada una a los resultados.

3. MATERIALES Y MÉTODOS

Overview de la metodología

Se va a mostrar un pequeño esquema del proceso metodológico en R [31] enlazado a los principales scripts generados, disponibles en Github [32].

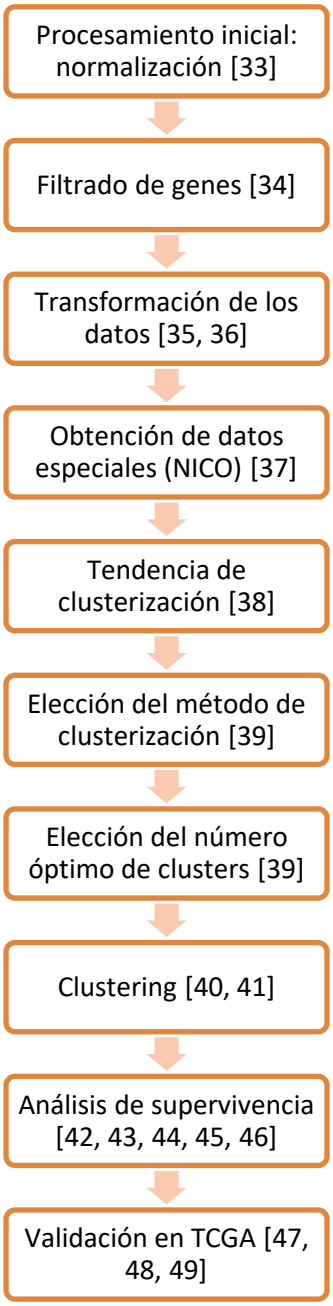


Ilustración 1. Overview de la metodología y los links para los scripts de cada apartado.

3.1. Muestras de pacientes

Disponemos del transcriptoma obtenido mediante RNASeq de 30 biopsias de pacientes del proyecto NASIR, un ensayo clínico nacional, financiado por la

farmacéutica BRISTOL-MYERS SQUIBB PHARMA EEIG y patrocinado por la Clínica Universidad de Navarra. Además, contamos con los datos de supervivencia de cada uno de los pacientes. Por otro lado, utilizamos los transcriptomas tumorales de libre acceso de 371 pacientes del proyecto LIHC del TCGA y sus respectivos datos de supervivencia para la validación de los resultados obtenidos (a través de *TCGABiolinks*) [50]

3.2. Análisis de los datos de secuenciación

3.2.1 Procesamiento inicial

Se hizo una secuenciación pair-end de las muestras de NASIR, obteniendo dos ficheros FASTQ para cada paciente. A continuación, comprobamos la calidad de las secuencias con FASTQC [51] y eliminamos las lecturas que tenían una calidad baja y adaptadores con *Trimmomatic* [52]. Después, alineamos las secuencias con Rsubread [53], utilizando el gtf v29 de Gencode [54] y obtuvimos los archivos en formato .BAM, para luego cuantificar las lecturas con la función FeatureCounts de la misma librería. Por último, usamos el método de mediana de ratios de DESeq2 [55] para normalizar las cuentas por diferencias en el tamaño y la composición de la librería, corrigiendo los sesgos técnicos que puedan contener. La Ilustración 2 resume el proceso:

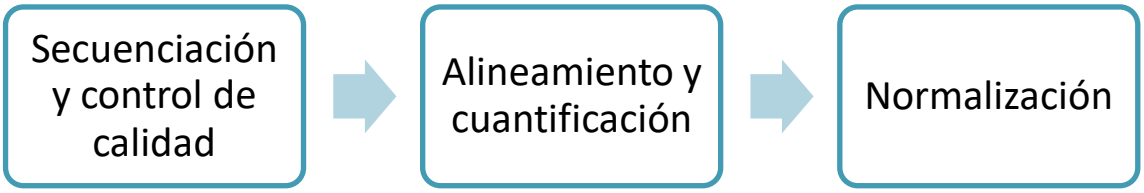


Ilustración 2. Pasos desde la obtención de la muestra hasta la normalización de las cuentas.

3.2.2 Filtrado de genes

Se realizó un filtrado de genes con el fin de eliminar aquellos con información irrelevante y así reducir la intensidad computacional de los análisis. Primero, se eliminaron todos los genes que codifican proteínas, enfrentando el dataset a la v38 de lncRNA de Gencode. A continuación, se mantuvieron aquellos genes con al menos 10 cuentas en un mínimo del 16% de los pacientes (5 de 30), un máximo $\geq$50 y una desviación estándar $\geq$ 10 . El esquema de los filtros se resume en la Ilustración 3.

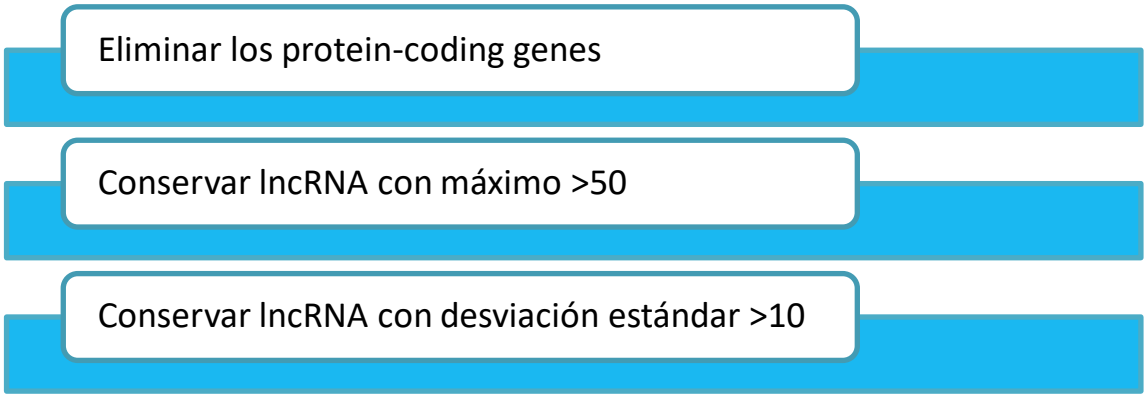


Ilustración 3. Listado de criterios de filtro.

3.2.3 Transformaciones de los datos

Aplicamos la transformación rlog [55] a la matriz original de cuentas y conservamos únicamente los genes que habían superado el filtrado. A continuación, transpusimos la matriz de cuentas para que los genes queden en las columnas, para así escalarlas por sus medias y desviaciones estándar y volvimos a transponer la matriz para volver a la forma del dataset original. Por último, aplicamos un reescalado al conjunto de datos [56] del 1 al 30 para que los distintos rangos no afecten al clustering. Una vez realizado el tratamiento de los datos, consideramos corregidas la asimetría y heterocedasticidad de los datos. La Ilustración 4 muestra el proceso del tratamiento de las cuentas.

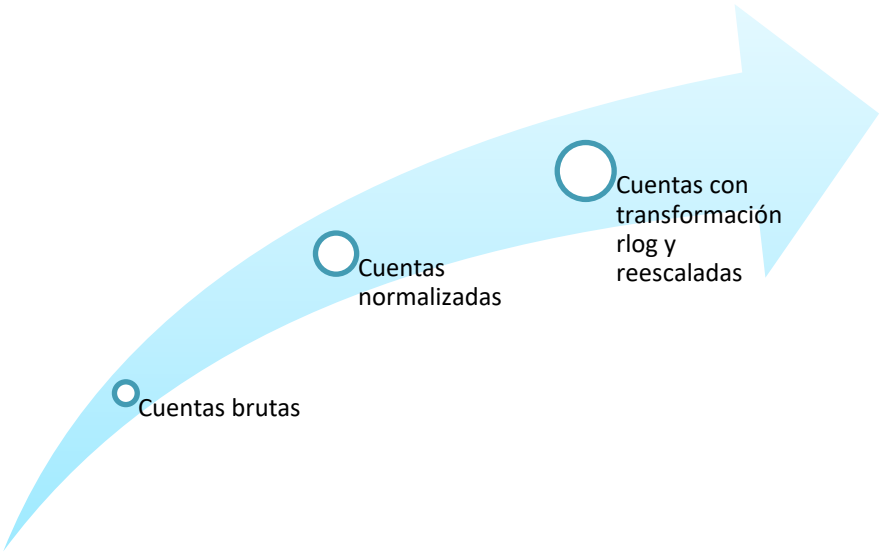


Ilustración 4. Proceso básico del tratamiento de las cuentas.

3.2.4 Datos especiales

El grupo de investigación de lncRNAs en HCC del CIMA ha identificado 100 lncRNAs con potencial oncogénico que ha denominado NICOs (de *"non-coding induced in cáncer with oncogenic potential"*). Estos genes se analizaron de modo independiente. Para ello, tras el reescalado, se creó un nuevo data-frame con la información de los NICOS.

3.2.5 Tendencia de clusterización del dataset

Antes de empezar a realizar clusterings, comprobamos que los datos tenían cierta tendencia a agruparse. Para ello, se calculó el estadístico de Hopkins (H) [57,58], que actúa como test de hipótesis, y es especialmente útil para ver si los datos son uniformes. La hipótesis nula indica que los datos están uniformemente distribuidos de una forma aleatoria siguiendo el proceso de Poisson (proceso estocástico que cuenta el número de sucesos raros). La hipótesis alternativa, en cambio, denota que el dataset no está distribuido uniformemente y puede contener clusters significativos.

El cálculo se realiza mediante los siguientes pasos:

1. Muestra uniformemente n puntos del dataset D.
2. Para cada punto de D, encontrar su vecino más cercano. Calcular la distancia entre ellos.
3. Generar datos aleatorios con n puntos y la misma variación que el dataset D.
4. Para cada punto random, encontrar su vecino más cercano. Calcular la distancia entre ellos.
5. Calcular el estadístico de Hopkins (H) como la media de la distancia del vecino más cercano del dataset randomizado dividido por la suma de la media de las distancias del vecino más cercano más las del randomizado.

Cabe destacar que se ha usado la función "hopkins" de la librería *clustertend* [59] que, en vez de dar el coeficiente, realiza el siguiente cálculo:

$$1 - H$$

Así pues, los datos clusterizables son aquellos en los que se da un resultado menor a 0.5.

3.2.6 Elección del método de clusterización de los genes

Se consideraron varios algoritmos de aprendizaje no supervisado para la división de los genes en grupos en base a características que no son conocidas [59]. De entre todos los algoritmos evaluamos (breve resumen en la figura 5):

· **Agrupamiento por centroides**. Algoritmo K-means que realiza tres pasos para clusterizar: (i) cuando el usuario asigna el número de grupos o clusters que desea (k), el algoritmo establece k centroides de forma aleatoria, (ii) se asigna cada uno de los valores al centroide más cercano, (iii) se calcula el promedio de cada grupo y el resultado pasa a ser el nuevo centroide del grupo. Los dos últimos pasos se repiten hasta que los centroides no superen un umbral de movilidad. De hecho, lo que en realidad hace K-means es minimizar la función de la suma de las distancias cuadráticas de cada uno de los datos con el centroide. El resultado es un ajuste que maximiza la distancia entre grupos y minimiza la distancia intragrupal.

· **Modelos de distribución**: miden la probabilidad de que todos los datos de un clúster pertenezcan a la misma distribución. Estos modelos muchas veces sufren overfitting.

· **Modelos de conectividad**: se agrupan en base a la hipótesis de que los datos que se encuentran más cercanos en el espacio son los que muestran una mayor similitud. Dentro de estos, los más comunes son los clusterings jerárquicos (hierarchicalclustering), que son los escogidos para el trabajo. Estos modelos requieren la elección de los siguientes parámetros:

1. El tipo de algoritmo: Los aglomerativos (AGNES) comienzan con un cluster por cada dato y con las iteraciones se combinan los más similares entre sí, por lo que están indicados para identificar clusters pequeños. Los divisivos (DIANA) comienzan con un solo cluster que va dividiéndose.
2. La métrica de distancia: sirven para calcular la distancia entre dos puntos. Las más utilizadas son la distancia euclídea (distancia en línea recta entre dos puntos del espacio euclídeo) y la distancia de Manhattan (suma de las diferencias en valor absoluto de las coordenadas cartesianas), aunque se tienen en consideración otras como canberra o minkowski.
3. Medida de la disimilitud entre clusters: la mayoría calculan las disimilitudes entre dos clusters y establecen el valor más alto (complete linkage), el más bajo (single linkage) o el promedio (average linkage) como la distancia entre los clusters. El método de Ward minimiza la varianza total dentro del cluster.

Para determinar el método más adecuado, realizamos una validación interna con *clValid* [60]. A continuación, calculamos las correlaciones cophenéticas para determinar la métrica de distancia y el método de enlace idóneos, y se volvió a realizar la validación interna con estos parámetros escogidos.

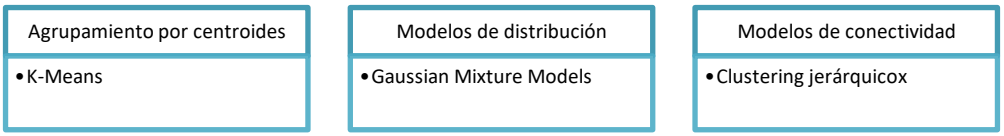


Ilustración 5. Esquema de los tipos de algoritmo considerados y un ejemplo de los más representativos de cada grupo.

3.2.7 Elección del número óptimo de clusters

La elección del número de clusters se puede determinar mediante varios métodos. El método de Elbow [58] busca minimizar la varianza del cluster, utilizando el algoritmo de partición para k grupos y calculando la suma del cuadrado intracluster (wss). El método de Silhouette [58] , por otro lado, busca maximizar el coeficiente del mismo nombre, que mide la calidad del clustering. Otro muy utilizado y válido para muchos métodos de agrupamiento es gap-statistic [58], que compara la variación de dentro del cluster para diferentes números de cluster k, con una distribución que no tiene grupos. Se busca maximizar el estadístico de gap. La ilustración 6 recoge un breve resumen del funcionamiento y obtención del número óptimo de clusters.

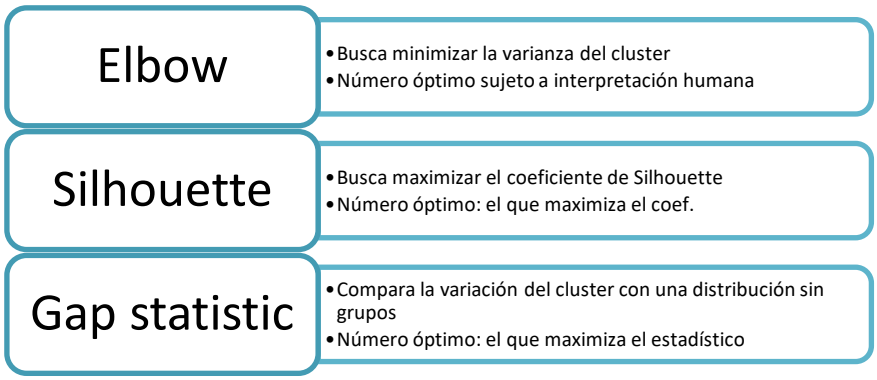
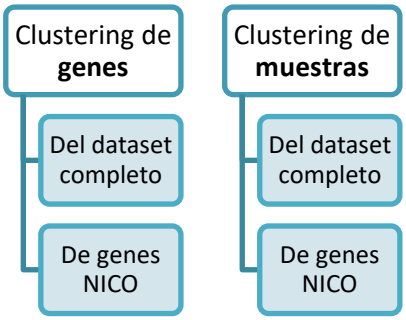


Ilustración 6. Esquema de los métodos más comunes, su funcionamiento y la elección del número óptimo de clusters para cada caso.

3.2.8 Clustering

Una vez que determinamos el número óptimo de clusters, procedimos a la división por grupos de los genes (ilustración 7). Primero, utilizamos el dataset completo para la división de los grupos mediante un clustering jerárquico con distancia euclídea y

average linkage. Se forzó al algoritmo para obtener un rango de entre 2 y 50 clusters. A continuación, realizamos subclusterings siguiendo el número óptimo de clusters iniciales y las mismas métricas y enlaces hasta dejar clusters de un máximo de 50 genes. Después, escogimos los genes NICO del dataset completo reescalado y realizamos los clusterings dejando un máximo de 15 genes por grupo. Por último, realizamos un clustering jerárquico de las muestras para cada grupo de genes para dividir a los pacientes en grupos de alto y bajo riesgo .



Ilustraciónn 7. Esquema de los clusterings realizados para cada dataset.

3.2.9 Análisis de supervivencia

Analizamos la supervivencia [61] entre los grupos de pacientes para cada cluster y aplicamos el término “overall survival” (OS) como indicador, que describe el tiempo que un paciente ha sobrevivido desde el diagnóstico. Utilizamos el método no paramétrico de Kaplan-Meier para estimar la probabilidad de supervivencia, asumiendo que el evento (fallecimiento) es independiente para cada paciente. Para ver si la diferencia entre grupos era estadísticamente significativa, se realizó un log-rank test, que asume como hipótesis nula que no hay diferencia entre los grupos de pacientes para la ocurrencia del evento. Se consideraron significativos p-valores $\leq$ 0.05 que tuvieran un número de pacientes balanceado (un mínimo de un 20% de los pacientes por grupo). Para aquellos clusters en los que sí existe diferencia estadísticamente significativa entre los grupos de pacientes, se realizó una combinación de pares de genes y se volvió a realizar un análisis de supervivencia de cada una de las combinaciones, para así establecer relaciones significativas y poder obtener una firma más robusta. En esta ocasión, se consideraron significativos los resultados que tuviesen un mínimo de un 20% de los pacientes en cada grupo y un p-valor <0.01 para conservar las firmas más robustas (Ilustración 8).

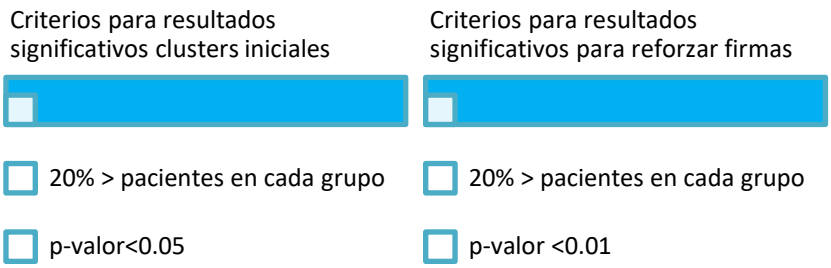


Ilustración 8. Criterios para considerar que un gen o cluster influyen en la supervivencia de los pacientes.

Por último, se evaluaron los genes NICO. Primero se realizaron los análisis de supervivencia para los clusters y después se analizaron los genes de forma individual. Además, se comparó la presencia estos genes en los genes de los sub-clusters significativos.

La ilustración 9 muestra un resumen de los datos a los que se les ha realizado el análisis de supervivencia.

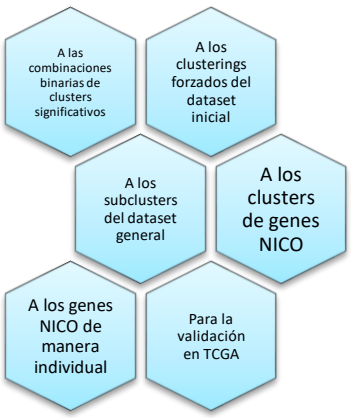


Ilustración 9. Figura que esquematiza los datos a los que se les realiza un análisis de supervivencia

3.2.10 Validación en el TCGA

Las muestras del TCGA de libre acceso ya habían sido alineadas y se habían cuantificado las cuentas con HTSeq, por lo que se partió de la matriz de cuentas. Se aplicaron los pasos de filtrado y transformación de datos de la misma manera que se ha descrito para las muestras de NASIR. A continuación, se vio cuántos genes tenía en común con el dataset de la cohorte independiente, y se especificó el número de genes significativos de NASIR que se encontraban en TCGA. Después, se realizaron las clusterizaciones sucesivas a los genes del TCGA siguiendo los métodos descritos

para la cohorte independiente y se realizaron los análisis de supervivencia para ver los genes o grupos en común.

4. RESULTADOS

4.1 Muestras

La cohorte NASIR tenía un total de 58721 genes al inicio del proceso, que se redujeron a 12308 cuando se eliminaron los genes codificantes. Cuando aplicamos el resto de criterios de inclusión, el dataset de trabajo se quedó con 3120 genes, como muestra la Ilustración 10.

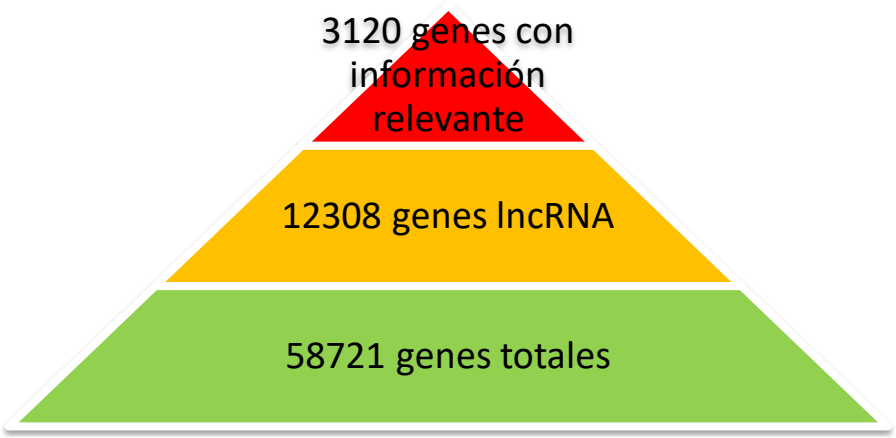


Ilustración 10. Pirámide que resume el número de genes que permanecen tras los filtrados.

4.2 Transformación de los datos

Con la transformación de datos consideramos corregida la extrema asimetría de los datos, acercándolos a una distribución normal (ejemplo de una muestra en la Ilustración 11).

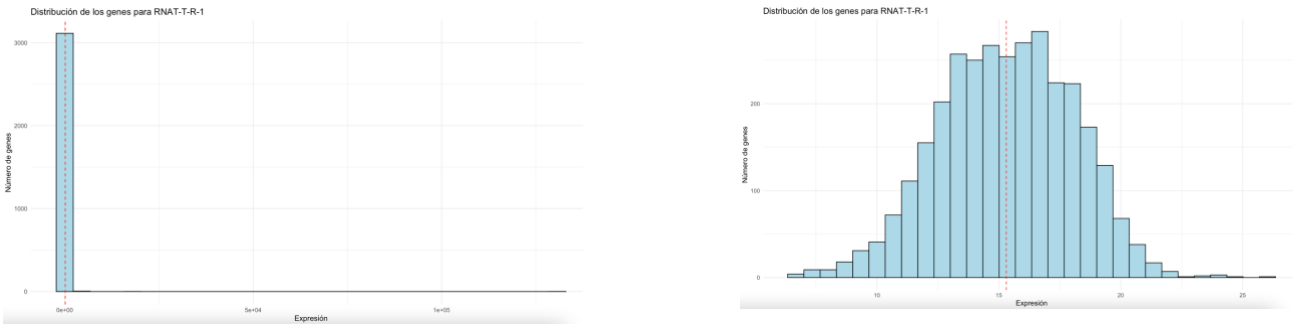


Ilustración 11. Comparación de la distribución antes (izquierda) y después (derecha) del tratamiento de datos.

4.3 Tendencia de clusterización

Antes de comenzar a realizar los agrupamientos, calculamos el coeficiente de Hopkins para ver la tendencia de agrupamiento del conjunto de datos. El coeficiente de los datos reales fue de 0.4, frente al 0.5 de los datos aleatorios, por lo que pudo confirmarse una cierta tendencia a clusterizar, tal y como se observa en la ilustración 12.

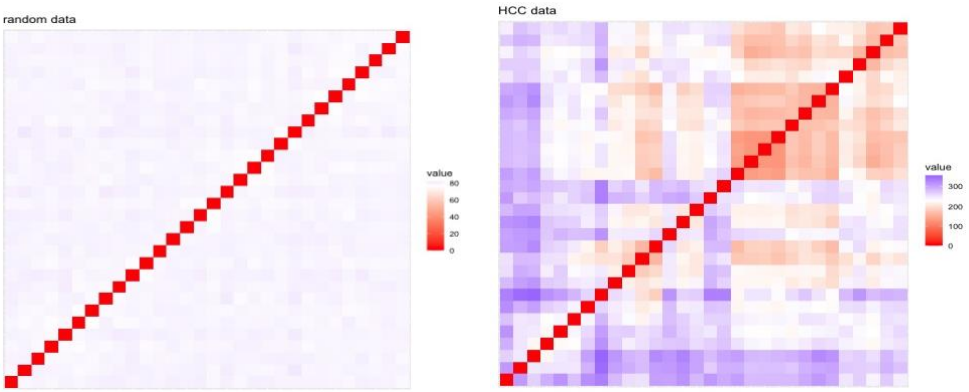


Ilustración 12. gráficos de las matrices de distancia ordenadas. Si existen agrupaciones, se forman patrones de bloque cuadrados. A la izquierda, la matriz de los datos aleatorios. A la derecha, losdatos de la cohorte NASIR.

4.4 Elección del método de clusterización y número de grupos

Una vez confirmada la tendencia a clusterizar, se calcularon los coeficientes cophenéticos (Tabla 1) para distintas combinaciones entre las métricas y métodos de enlace, estableciendo que los resultados óptimos se obtenían con una métrica euclídea y enlace average. Por otro lado, decidimos utilizar el clustering jerárquico como método de división de grupos debido a la enorme popularidad del algoritmo en el ámbito biológico. Igualmente, se realizó una validación interna de los datos para asegurar la adecuación del método tanto en el dataset original, como en las posteriores divisiones. Como se puede observar en la Tabla 2, el método jerárquico maximiza el coeficiente de Dunn, pero Diana obtiene mejores resultados en cuanto a conectividad y coeficiente de Silhouette. Sin embargo, se puede observar que las diferencias en estos dos coeficientes no son extremas entre los dos métodos. A su vez, en agrupamientos sucesivos, se ve una clara mejora del agrupamiento jerárquico sobre el resto, con unos coeficientes mejorados respecto al dataset inicial. Por último, estos métodos confirman que la división de grupos óptima es de 2.

Tabla1. Coeficientes cophenéticos de las combinaciones de métrica y método. A la izquierda, del dataset completo. A la derecha, de uno de los agrupamientos.

Coeficientes cophenéticos del dataset completo

	metrica_distancia	metodo_linkage	cophenetic
1	euclidean	average	0.529276952158219
2	minkowski	average	0.529276952158219
3	manhattan	average	0.5234569936502
4	canberra	average	0.499018867008647
5	canberra	ward.D2	0.422256035032567
6	euclidean	complete	0.403471581664611
7	minkowski	complete	0.403471581664611
8	euclidean	ward.D2	0.394230989387006
9	minkowski	ward.D2	0.394230989387006
10	manhattan	ward.D2	0.380647044366655
11	canberra	complete	0.376956761574247
12	manhattan	complete	0.374656322227144
13	euclidean	centroid	0.176196431334829
14	minkowski	centroid	0.176196431334829
15	manhattan	centroid	0.173630996353795
16	canberra	centroid	0.16464474104272
17	euclidean	single	0.159175141639053
18	minkowski	single	0.159175141639053
19	manhattan	single	0.141211655466567
20	canberra	single	0.139866480977186
21	euclidean	median	0.119287098815
22	minkowski	median	0.119287098815
23	manhattan	median	0.115693559601314
24	canberra	median	0.100882463166582

Coeficientes cophenéticos cluster 120

	metrica_distancia	metodo_linkage	cophenetic
1	euclidean	average	0.712663329067137
2	minkowski	average	0.712663329067137
3	manhattan	average	0.690608574376763
4	canberra	average	0.681088584393686
5	euclidean	ward.D2	0.67732216434648
6	minkowski	ward.D2	0.67732216434648
7	canberra	complete	0.650429601221253
8	euclidean	single	0.631998590883633
9	minkowski	single	0.631998590883633
10	manhattan	ward.D2	0.618598969387158
11	canberra	ward.D2	0.612400933869924
12	manhattan	single	0.605732669771159
13	manhattan	complete	0.583132358413087
14	canberra	single	0.575689126783464
15	euclidean	complete	0.571300224651478
16	minkowski	complete	0.571300224651478
17	euclidean	centroid	0.552542555352266
18	minkowski	centroid	0.552542555352266
19	manhattan	centroid	0.530074018770378
20	canberra	centroid	0.524192991071522
21	manhattan	median	0.434626555276622
22	canberra	median	0.397443113772026
23	euclidean	median	0.338032659541705
24	minkowski	median	0.338032659541705

Tabla2. Resultados de la validación interna. Se muestra como resumen qué algoritmo y número de clusters los optimizan. Arriba, de un subgrupo y, abajo, del dataset completo.

```
Clustering Methods:
 hierarchical pam kmeans clara diana

Cluster sizes:
 2 3 4 5 6

Validation Measures:
                                 2        3        4        5        6

hierarchical Connectivity   977.9179 1025.1183 1116.9238 1372.2313 1390.7417
             Dunn             0.3248   0.3248   0.3248   0.3248   0.3248
             Silhouette       0.0986   0.0678   0.0509   0.0492   0.0358
pam          Connectivity   984.1083 1742.1325 2109.7226 2381.5337 2730.0159
             Dunn             0.1788   0.1185   0.1581   0.1677   0.1154
             Silhouette       0.1231   0.0784   0.0733   0.0646   0.0591
kmeans       Connectivity   971.1667 1495.4425 1920.5234 2152.4738 2399.6631
             Dunn             0.0802   0.1190   0.1134   0.1198   0.1437
             Silhouette       0.1258   0.0969   0.0778   0.0750   0.0702
clara        Connectivity  1017.0298 1874.1214 2018.3901 2378.4948 2716.6377
             Dunn             0.0784   0.1082   0.1507   0.1267   0.1190
             Silhouette       0.1221   0.0749   0.0820   0.0731   0.0576
diana        Connectivity   947.1873 1418.5869 2140.4274 2472.5238 2666.7992
             Dunn             0.0861   0.0873   0.0878   0.0890   0.0893
             Silhouette       0.1266   0.0928   0.0654   0.0547   0.0524

Optimal Scores:

             Score    Method       Clusters
Connectivity 947.1873 diana        2
Dunn           0.3248 hierarchical 2
Silhouette     0.1266 diana        2

NULL
```

```
Clustering Methods:
 hierarchical kmeans pam diana clara

Cluster sizes:
 2 3 4 5 6

Validation Measures:
                                 2       3       4       5       6

hierarchical Connectivity   8.8032 12.6722 22.1278 25.1972 31.0575
             Dunn           0.6946  0.6946  0.6204  0.6439  0.7021
             Silhouette     0.1537  0.1391  0.1346  0.1187  0.1287
kmeans       Connectivity   8.8032 12.6722 22.1278 25.1972 31.0575
             Dunn           0.6946  0.6946  0.6204  0.6439  0.7021
             Silhouette     0.1537  0.1391  0.1346  0.1187  0.1287
pam          Connectivity  16.8929 18.7246 22.5937 29.5317 31.9345
             Dunn           0.5466  0.6204  0.6204  0.6204  0.6204
             Silhouette     0.1062  0.1187  0.1318  0.1142  0.0869
diana        Connectivity   8.9976 19.8893 24.5361 29.2813 32.5159
             Dunn           0.6188  0.5298  0.5368  0.5405  0.5767
             Silhouette     0.1512  0.1249  0.0877  0.0786  0.0974
clara        Connectivity  19.1651 21.9782 22.5937 29.5317 31.9345
             Dunn           0.4303  0.4303  0.6204  0.6204  0.6204
             Silhouette     0.0706  0.0935  0.1318  0.1142  0.0869

Optimal Scores:

             Score  Method       Clusters
Connectivity 8.8032 hierarchical 2
Dunn         0.7021 hierarchical 6
Silhouette   0.1537 hierarchical 2

NULL
```

4.5 Clustering de los genes

Se obtuvieron dos tipos de clusters. Por un lado, aquellos derivados del agrupamiento del dataset general y, por otro, aquellos con un máximo de 50 genes derivados de los sucesivos agrupamientos. De estos últimos, se obtuvieron un total de 159 clusters.

4.6 Análisis de supervivencia

Primero, realizamos un análisis de supervivencia de los clusters generales, pero ninguno denotó una diferencia significativa entre los dos grupos de pacientes. A continuación, realizamos un análisis de supervivencia de los agrupamientos sucesivos, donde seis clusters cumplieron los criterios que buscábamos (p-valor $\leq$ 0.05 y más de un 20% de los pacientes en cada grupo), confirmando que hay diferencia entre los grupos de pacientes de alto y bajo riesgo. Los genes de los clusters significativos se recogen en los anexos (anexo I).

Después, generamos todas las posibles combinaciones binarias no repetitivas de los genes de cada uno de los clusters significativos y realizamos un análisis de supervivencia de cada una de las combinaciones para reforzar las firmas genéticas. Sólo consideramos significativos aquellos resultados con un p-valor<0.01 y con al menos un 20% de pacientes en cada uno de los grupos (anexo 2a). Uno de los clusters no obtuvo ninguna combinación con los criterios seleccionados, mientras que otro obtuvo combinaciones que no se relacionaban entre sí. Sin embargo, otros cuatro clusters establecieron firmas robustas de combinaciones relacionadas entre sí (anexo 2b, Ilustración 13). Después, se comprobó que ningún gen de forma individual determinaba la supervivencia de los pacientes.

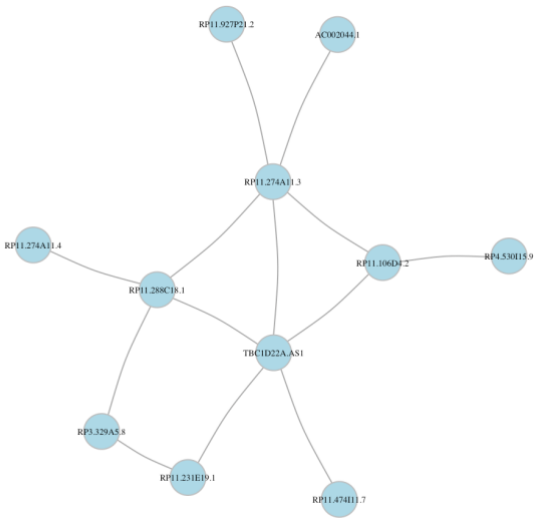


Ilustración 13.. Grafo de las combinaciones del cluster 114.

Por otro lado, se hicieron análisis de supervivencia para cada uno de los genes NICO de forma individual, donde se consiguió establecer que el gen MYLK-AS1 sí estaba relacionado con la supervivencia de los pacientes de nuestra cohorte (ilustración 14).

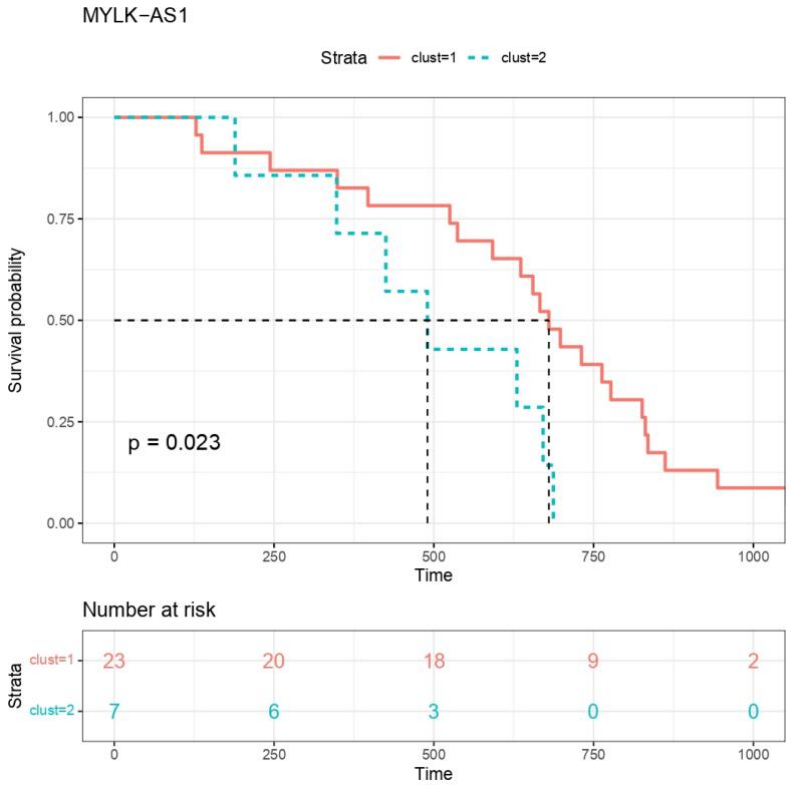


Ilustración 14. Curva de Kaplan Meier del gen MYLK-AS1.

4.7 Validación con el TCGA

A las muestras tumorales de pacientes del TCGA se les realizó el mismo tratamiento de datos que a los de nuestra cohorte (Anexo 3). En total, quedaron 2942 genes una vez se filtraron. Tras el filtrado, vimos que compartía 1427 genes con nuestra cohorte. Además, 45 genes significativos de NASIR se encontraban en TCGA (Ilustración 15). A continuación, se realizaron los clusterings sucesivos y el análisis de supervivencia, donde un cluster compuesto únicamente por el gen MAFTRR tuvo diferencias significativas entre los dos grupos de pacientes. Este gen no se encontraba en la lista de los de NASIR, por lo que no se consiguió validar ninguna firma. Por último, se hizo un análisis de supervivencia del gen MYLK-AS1, pero no pudo confirmarse tampoco su relación en la supervivencia de los pacientes para este dataset.

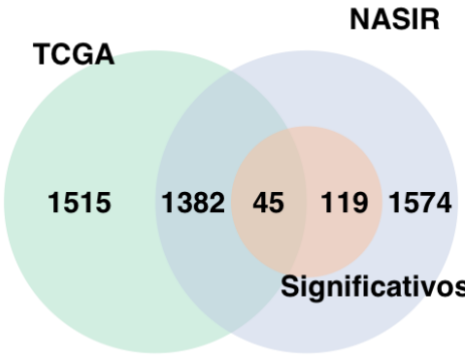


Ilustración 15. Diagrama de Venn del número de genes que comparten TCGA y NASIR, especificando los comunes dentro de los que son significativos en la cohorte independiente.

5. DISCUSIÓN

En este trabajo hemos identificado seis grupos de lncRNAs relacionados con la supervivencia de nuestra cohorte y hemos creado una firma genética válida para ésta mediante combinaciones binarias de los genes. Hemos de realizar experimentos adicionales para determinar firmas con el menor número de genes que tengan la mayor asociación con superviviencia. Después, deberíamos estudiar esas firmas para ver si sirven para dividir a pacientes con características adicionales, como proliferativos versus no proliferativos o asociados con mutaciones "driver" en HCC como TERT, TP53 o CTNNB1 o con determinadas etiologías (abuso de alcohol, infecciones con virus hepatotropos o obesidad, por ejemplo). Por último, sería interesante determinar los mecanismos que permiten la expresión de esos genes y su funcionalidad para determinar si alguno de ellos se puede considerar "driver" o "passenger" de HCC.

Antes de realizar todos estos estudios sería deseable confirmar que estas firmas son robustas y que también se asocian a supervivencia en una cohorte independiente. Esto es especialmente relevante en nuestro caso porque nuestra cohorte se compone de datos de solo 30 pacientes de un ensayo clínico en el que se deben cumplir requisitos especiales para el reclutamiento. Disponemos de un número muestral poco representativo y es probable que no haya heterogeneidad en los pacientes. En este trabajo mostramos que no pudieron validarse los descubrimientos usando los datos de TCGA. Sin embargo esto no nos desanima por varios motivos. Aunque es cierto que los datos de la TCGA pertenecen a 371 pacientes, y representan una colección de

datos moleculares de enorme relevancia, tienen una anotación clínica pobre. Las anotaciones de los pacientes, sobre todo la supervivencia, requieren de un seguimiento para nutrir las bases de datos que no ha tenido lugar de manera estricta. Además, en el caso de los pacientes de HCC, la TCGA ha utilizado un gran número de pacientes que desarrollan HCC sin inflamación previa. Esto se corresponde con solo el 10% de los pacientes de HCC y hace que la cohorte sea poco representativa. Por todo ello es urgente que se validen nuestros resultados en otras cohortes (RINKEN [62] o INSERM [63]). Aunque estamos trabajando en ello, en el caso de los datos clínicos del INSERM, por ejemplo, no son públicos y hemos de obtenerlos por colaboración.

Otra posible razón por la que no se ha conseguido la validación con los datos de la TCGA es que los archivos de libre acceso habían sido previamente alineados y cuantificados con HTSeq, mientras que nosotros utilizamos Rsubread. El uso de diferentes herramientas puede afectar drásticamente al resultado, por lo que lo óptimo hubiera sido seguir el mismo proceso desde el inicio. Esta diferencia podría acentuarse más con secuencias que, como los lncRNA, tienen pocas lecturas y diferencias abruptas de varianza. A pesar de que estos dos métodos son válidos para el estudio de los lncRNA, recientes estudios sugieren el uso de herramientas que utilizan pseudocounts como Kallisto o Salmon, ya que ofrecen una mejor actuación [64].

A lo largo del trabajo hemos comparado nuestra metodología con la de otros estudios [23,25]. Todas las clasificaciones utilizan datos de genes codificantes, que son los que más se expresan y, quizá por ello, las metodologías pueden ser dispares. En la clasificación de Wheeler analizaron de manera individual una serie de datos, como el mRNA o las islas CpG, para luego unificarlos en un cluster integrativo [25]. Hoshida, en cambio, recogió 8 cohortes independientes para obtener heterogeneidad de los pacientes y evitar el overfitting, las dividió en 9 grupos y utilizó tres de ellas como set de entrenamiento. Después, definió subgrupos con el método SubclassMapping y utilizó varios algoritmos de agrupación [23]. Otras investigaciones realizan análisis de expresión diferencial haciendo uso de los transcriptomas de tejidos peritumorales para encontrar lncRNA como potenciales biomarcadores [65]. Nosotros, en cambio, utilizamos exclusivamente la información de RNA-Seq de muestras tumorales y realizamos clusterings sucesivos hasta obtener grupos de genes del tamaño deseado. Este método no fuerza a hacer divisiones que no detecta en un solo paso, pero sí que puede forzar con las iteraciones a dividir grupos que no existen a

pesar de que los las validaciones mejoraron en los subclusters. Además, se ha utilizado la distancia euclídea, el enlace average y una división de dos grupos para todos los clusterings, por lo que es probable que haya agrupamientos en los que estos parámetros no sean los óptimos.

Respecto a los lncRNA significativos, no han sido descritos en la bibliografía, por lo que la información principal sobre éstos radica en características como la ubicación cromosómica o la posible conservación del locus en otros organismos modelo. Sin embargo, el gen NICO MYLK-AS1 sí ha sido descrito. Es un lncRNA antisentido ubicado en el brazo largo del cromosoma 3 que promueve el crecimiento y la invasión en hepatocarcinoma promoviendo la expresión de EGFR/HER2 (ruta señalizadoraEGFR/HER2-extracellular signal-regulatedkinase 1/2 (ERK1/2)). Se encuentra sobreexpresado en pacientes con HCC y está negativamente correlacionado con el pronóstico del paciente [66]. Como se ha dicho anteriormente, sería de enorme interés identificar los mecanismos moleculares que median la función del resto de los candidatos encontrados para identificar aquellos que también tengan un valor terapéutico.

6. CONCLUSIÓN

Hemos obtenido seis clusters relacionados con la supervivencia en nuestra cohorte y hemos conseguido reforzar las firmas genéticas mediante combinaciones binarias de genes. Además, hemos constatado la importancia del gen NICO MYLK-AS1 en la supervivencia de estos pacientes. Sin embargo, no hemos conseguido validar ninguno de los resultados en el TCGA. Con todo ello, hemos propuesto una línea de investigación a futuro que intentará conseguir una firma genética sólida usando diferentes datasets y pueda utilizarse para la estratificación de los pacientes.

7. ANEXOS

Anexo 1: Análisis de supervivencia de NASIR

Anexo 1a : Listado de genes de clusters significativos

	Cluster 77	Cluster 95	Cluster 114	Cluster 154	Cluster 155	Cluster 156
1	XXbac-B461K10.4	LINC01587	RP5-1125A11.4	WARS2-IT1	AC000111.6	RBMS3-AS2
2	C9orf106	IL10RB-DT	RP1-20B11.2	LINC02157	SLC24A3-AS1	CACNA2D1-AS1

3	CTD-2600O9.1	KLF7-IT1	RP4-530I15.9	RP11-384P7.7	LINC02593	AC025165.8
4	SPATA13-AS1	C6orf3	AC092620.2	RP11-759A24.2	LINP1	AC016735.1
5	RP5-837I24.2	LINC02398	RP11-544A12.8		GNA14-AS1	KIRREL1-IT1
6	AC006547.13	RP11-598F7.3	RP5-991G20.2		CFTR-AS1	AC003991.3
7	RP5-837I24.4	RP5-1142A6.9	CTA-339C12.1		RP11-388P9.2	HCG11
8	MAP3K20-AS1	RP11-484L8.1	RP11-231E19.1		RP11-334A14.8	HOXB-AS1
9	RP5-894A10.2	RP11-449P15.2	CTA-305I2.1		LINC00707	AOAH-IT1
10	RP11-867G23.3	RP11-622C24.2	RP11-166B2.3		RP11-806K15.1	DNM3OS
11	RP11-331K21.1	RP11-426J5.3	RP11-474I11.7		RP11-510M2.5	RP11-40C11.2
12	PITPNM2-AS1	RP11-46D1.2	RP11-464D20.6		PWRN2	RP1-225E12.2
13	CTC-458I2.2		AC010761.13		RP3-523K23.2	OSBPL10-AS1
14	RP11-136O12.2		RP11-927P21.2		RP4-555D20.2	HOTAIRM1
15	CTD-2373N4.5		RP11-214K3.19		RP11-2N1.1	RP1-187B23.1
16	AC012613.2		RP11-274B21.10		CH17-360D5.3	RP11-539E19.2
17	RP11-359M6.1		CTB-181H17.1		RP11-236L14.2	AP001189.4
18	CTD-3076O17.2		RP3-329A5.8		RP11-262A16.1	ADAMTS9-AS2
19	RP11-90E5.1		RP11-416N2.4		RP11-454H19.2	LINC00920
20	CTD-2647L4.4		RP11-240G22.5		RP11-71L14.3	MEF2C-AS1
21	RP11-702L15.4		RP11-106D4.2		RP11-327J17.7	RP11-526F3.1
22	RP11-16E23.3		RP11-391L3.4		LINC00842	CTC-459M5.2
23	RP11-737O24.1		RP11-274A11.3		RP11-202D18.3	SEMA5A-AS1
24	RP11-53B2.3		RP11-391L3.3			RP11-642D21.2
25	RP11-53B2.5		RP11-288C18.1			RP11-642D21.1
26	RP11-53B2.4		AC002044.1			RP11-728F11.4
27	RP11-53B2.1		RP11-274A11.4			SENCR
28	LDLRAD4-AS1		TBC1D22A-AS1			AF131215.3
29	RP11-691H4.4		UCKL1-AS1			AP002954.3
30	CTC-444N24.8		SALRNA2			RP11-357H14.16
31	ERVK9-11		RP5-1022I14.2			DIO3OS
32	RP5-894A10.6		RP11-141M1.5			RP11-588K22.2
33	RP4-564F22.6					CTD-3157E16.1
34	RP11-131L12.3					CTD-2373H9.5
35	AC008985.1					RP11-59C5.3
36	LL0XNC01-36H8.1					RP5-1057I20.6
37	RP13-977J11.6					RP11-327J17.9
38	CTC-457L16.1					AC020951.1
39	CTC-463A16.1					EPB41L4A-DT
40	RP13-554M15.7					RP11-329N15.3
41	RP11-197N18.8					CTD-2373H9.3

42	RP11-640L9.2					CTD-3035K23.3
43	RP11-87P3.1					PLAC4
44						CTD-2536I1.2
45						RP11-426C22.12
46						COL1A2-AS1
47						RP11-496B10.6
48						RP11-141H1.1
49						RP11-254F19.5
50						RP11-567P19.2

Anexo 1b : p-valor y número de pacientes por cada cluster significativo

	cluster 77	cluster 95	cluster 114	cluster 154	cluster 155	cluster 156
p-valor	0.022	0.002	0.041	0.047	0.01	0.042
número pacientes grupo 1	18	21	13	14	18	13
número pacientes grupo 2	12	9	17	16	12	17

Anexo 2: Resultados de las combinaciones

Anexo 2a : Listado de combinaciones significativas

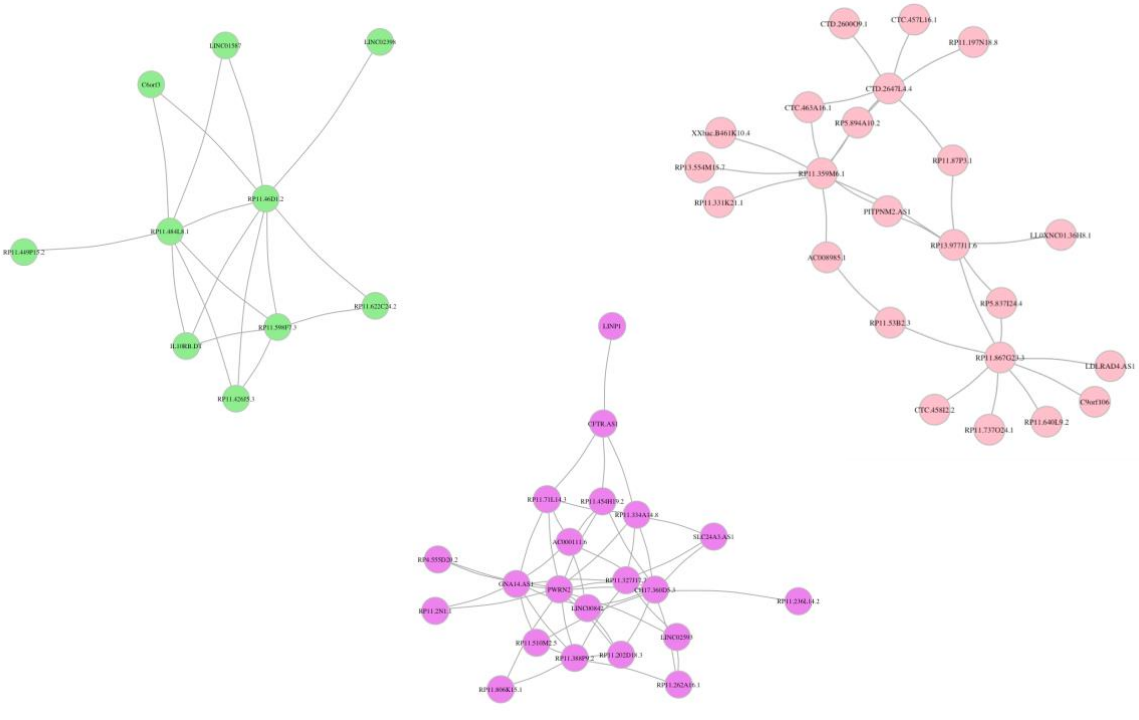
	cluster	gen1	gen2	p-valor	num_pacientes1	num_pacientes2
1	cluster114	RP11.231E19.1	RP3.329A5.8	0.00909202714803915	11	19
2	cluster114	RP11.231E19.1	TBC1D22A.AS1	0.00915014911702459	14	16
3	cluster114	RP11.474I11.7	TBC1D22A.AS1	0.00322535926642199	13	17
4	cluster114	RP11.927P21.2	RP11.274A11.3	0.00933794319299103	8	22
5	cluster114	RP3.329A5.8	RP11.288C18.1	0.000671242795413745	14	16
6	cluster114	RP11.106D4.2	RP11.274A11.3	0.00342870098319957	12	18
7	cluster114	RP11.106D4.2	TBC1D22A.AS1	0.00357780515291293	14	16
8	cluster114	RP11.274A11.3	RP11.288C18.1	0.00610936151383243	12	18
9	cluster114	RP11.274A11.3	AC002044.1	0.00500352559741994	11	19
10	cluster114	RP11.274A11.3	TBC1D22A.AS1	0.00286558386227528	11	19

11	cluster114	RP11.288C18.1	RP11.274A11.4	0.00746834340260909	12	18
12	cluster114	RP11.288C18.1	TBC1D22A.AS1	0.00610936151383243	12	18
13	cluster114	RP4.530I15.9	RP11.106D4.2	0.0053559725526502	14	16
14	cluster155	CFTR.AS1	RP11.334A14.8	0.00509933805809942	13	17
15	cluster155	CFTR.AS1	RP11.454H19.2	0.000314939307785413	8	22
16	cluster155	CFTR.AS1	RP11.71L14.3	0.00938558024932851	13	17
17	cluster155	RP11.388P9.2	RP11.806K15.1	0.00759619727103148	13	17
18	cluster155	RP11.388P9.2	RP11.510M2.5	0.00494154667071006	12	18
19	cluster155	RP11.388P9.2	PWRN2	0.00694314122034853	11	19
20	cluster155	RP11.388P9.2	RP11.262A16.1	0.00759619727103148	13	17
21	cluster155	RP11.388P9.2	RP11.327J17.7	0.00261550000936239	12	18
22	cluster155	RP11.388P9.2	RP11.202D18.3	0.00369656714766526	11	19
23	cluster155	RP11.334A14.8	PWRN2	0.00694314122034853	11	19
24	cluster155	RP11.334A14.8	CH17.360D5.3	0.00189531946715688	13	17
25	cluster155	RP11.334A14.8	RP11.71L14.3	0.00550817604132328	11	19
26	cluster155	RP11.334A14.8	RP11.327J17.7	0.00762240154684613	13	17
27	cluster155	AC000111.6	CH17.360D5.3	0.00298706551065612	14	16
28	cluster155	RP11.806K15.1	PWRN2	0.00694314122034853	11	19
29	cluster155	AC000111.6	RP11.454H19.2	0.00136947645867312	11	19
30	cluster155	RP11.510M2.5	CH17.360D5.3	0.00929587173117899	15	15
31	cluster155	PWRN2	RP4.555D20.2	0.00694314122034853	11	19
32	cluster155	AC000111.6	RP11.71L14.3	0.00592273000105233	10	20
33	cluster155	PWRN2	RP11.2N1.1	0.00694314122034853	11	19
34	cluster155	PWRN2	CH17.360D5.3	0.00694314122034853	11	19
35	cluster155	PWRN2	RP11.454H19.2	0.00694314122034853	11	19
36	cluster155	PWRN2	RP11.71L14.3	0.00694314122034853	11	19
37	cluster155	PWRN2	RP11.327J17.7	0.00762240154684613	13	17
38	cluster155	PWRN2	LINC00842	0.00694314122034853	11	19
39	cluster155	PWRN2	RP11.202D18.3	0.00694314122034853	11	19
40	cluster155	AC000111.6	LINC00842	0.00298706551065612	14	16
41	cluster155	CH17.360D5.3	RP11.236L14.2	0.00929587173117899	15	15
42	cluster155	CH17.360D5.3	RP11.262A16.1	0.00670417509427366	14	16
43	cluster155	CH17.360D5.3	RP11.454H19.2	0.00929587173117899	15	15
44	cluster155	CH17.360D5.3	RP11.327J17.7	0.0055061404859494	14	16
45	cluster155	CH17.360D5.3	LINC00842	0.00670417509427366	14	16
46	cluster155	CH17.360D5.3	RP11.202D18.3	0.00743490228539834	14	16
47	cluster155	LINC00842	RP11.202D18.3	0.00510023704520699	13	17
48	cluster155	SLC24A3.AS1	RP11.334A14.8	0.00694314122034853	11	19
49	cluster155	SLC24A3.AS1	CH17.360D5.3	0.00670417509427366	14	16
50	cluster155	AC000111.6	GNA14.AS1	0.00298706551065612	14	16

51	cluster155	SLC24A3.AS1	RP11.327J17.7	0.00762240154684613	13	17
52	cluster155	LINC02593	PWRN2	0.00694314122034853	11	19
53	cluster155	LINC02593	RP11.262A16.1	0.00694314122034853	11	19
54	cluster155	LINC02593	RP11.327J17.7	0.00762240154684613	13	17
55	cluster155	LINP1	CFTR.AS1	0.00938558024932851	13	17
56	cluster155	GNA14.AS1	RP11.388P9.2	0.00759619727103148	13	17
57	cluster155	GNA14.AS1	RP11.510M2.5	0.0025612468108674	11	19
58	cluster155	GNA14.AS1	PWRN2	0.00694314122034853	11	19
59	cluster155	GNA14.AS1	RP4.555D20.2	0.0025612468108674	11	19
60	cluster155	GNA14.AS1	RP11.2N1.1	0.00395499850330191	10	20
61	cluster155	GNA14.AS1	RP11.71L14.3	0.00743991588924254	12	18
62	cluster155	GNA14.AS1	RP11.327J17.7	0.00483790832026206	14	16
63	cluster155	GNA14.AS1	LINC00842	0.00163380845367916	10	20
64	cluster156	CTD.2373H9.5	AC020951.1	0.00279058918365579	14	16
65	cluster156	ADAMTS9.AS2	RP11.59C5.3	0.000902569645915979	7	23
66	cluster156	CTC.459M5.2	RP5.1057I20.6	0.00309667269503129	13	17
67	cluster156	RP11.642D21.2	CTD.3035K23.3	0.00675815365813651	15	15
68	cluster156	RP11.642D21.1	COL1A2.AS1	0.00733638197454484	13	17
69	cluster156	CACNA2D1.AS1	RP11.567P19.2	0.00483792324549543	13	17
70	cluster156	SENCR	RP11.141H1.1	0.000902569645915979	7	23
71	cluster77	CTD.2600O9.1	CTD.2647L4.4	0.00441327447133807	11	19
72	cluster77	XXbac.B461K10.4	RP11.359M6.1	0.00375715546057332	11	19
73	cluster77	RP5.837I24.4	RP11.867G23.3	0.00318140010673832	14	16
74	cluster77	RP5.837I24.4	RP13.977J11.6	0.00789394323674983	10	20
75	cluster77	RP5.894A10.2	RP11.359M6.1	0.00458655987344217	12	18
76	cluster77	RP5.894A10.2	CTD.2647L4.4	0.00559229010334777	10	20
77	cluster77	RP11.867G23.3	CTC.458I2.2	0.00678174736001643	9	21
78	cluster77	RP11.867G23.3	RP11.737O24.1	0.00261819708703676	11	19
79	cluster77	RP11.867G23.3	RP11.53B2.3	0.00432431092253542	8	22
80	cluster77	RP11.867G23.3	LDLRAD4.AS1	0.00678174736001643	9	21
81	cluster77	RP11.867G23.3	RP13.977J11.6	0.00432431092253542	8	22
82	cluster77	RP11.867G23.3	RP11.640L9.2	0.00432431092253542	8	22
83	cluster77	RP11.331K21.1	RP11.359M6.1	0.00866945217705155	14	16
84	cluster77	PITPNM2.AS1	RP11.359M6.1	0.00458655987344217	12	18
85	cluster77	PITPNM2.AS1	RP13.977J11.6	0.00212579035794708	13	17
86	cluster77	C9orf106	RP11.867G23.3	0.00573725527845543	10	20
87	cluster77	RP11.359M6.1	CTD.2647L4.4	0.00375715546057332	11	19
88	cluster77	RP11.359M6.1	AC008985.1	0.000743368290792738	11	19
89	cluster77	RP11.359M6.1	RP13.977J11.6	0.00290992797948062	12	18
90	cluster77	RP11.359M6.1	CTC.463A16.1	0.00260136808848123	12	18

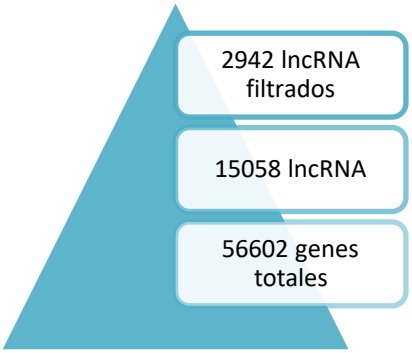
91	cluster77	RP11.359M6.1	RP13.554M15.7	0.00178358377947674	11	19
92	cluster77	CTD.2647L4.4	CTC.457L16.1	0.00533430974200045	6	24
93	cluster77	CTD.2647L4.4	CTC.463A16.1	0.00441327447133807	11	19
94	cluster77	CTD.2647L4.4	RP11.197N18.8	0.00559229010334777	10	20
95	cluster77	CTD.2647L4.4	RP11.87P3.1	0.00610873327411297	10	20
96	cluster77	RP11.53B2.3	AC008985.1	0.00537011030422269	10	20
97	cluster77	LL0XNC01.36H8.1	RP13.977J11.6	0.00513353052134208	14	16
98	cluster77	RP13.977J11.6	RP11.87P3.1	0.00646362369810425	14	16
99	cluster95	LINC01587	RP11.46D1.2	0.00113971129218254	11	19
100	cluster95	IL10RB.DT	RP11.598F7.3	0.00694314122034853	11	19
101	cluster95	IL10RB.DT	RP11.484L8.1	0.000843297363478081	8	22
102	cluster95	IL10RB.DT	RP11.46D1.2	0.00110537693174475	7	23
103	cluster95	C6orf3	RP11.484L8.1	0.00970195792949017	8	22
104	cluster95	C6orf3	RP11.46D1.2	0.00970195792949017	8	22
105	cluster95	LINC02398	RP11.46D1.2	0.00533430974200045	6	24
106	cluster95	RP11.598F7.3	RP11.484L8.1	0.0046924681561052	7	23
107	cluster95	RP11.598F7.3	RP11.622C24.2	0.00235669164675533	10	20
108	cluster95	RP11.598F7.3	RP11.426J5.3	0.00894447339078716	6	24
109	cluster95	RP11.598F7.3	RP11.46D1.2	0.00694314122034853	11	19
110	cluster95	RP11.484L8.1	RP11.449P15.2	0.0046924681561052	7	23
111	cluster95	RP11.484L8.1	RP11.426J5.3	0.00533430974200045	6	24
112	cluster95	RP11.484L8.1	RP11.46D1.2	0.0046924681561052	7	23
113	cluster95	RP11.622C24.2	RP11.46D1.2	0.00533430974200045	6	24
114	cluster95	RP11.426J5.3	RP11.46D1.2	0.00533430974200045	6	24
115	cluster95	LINC01587	RP11.484L8.1	0.00113971129218254	11	19

*An**exo 2b: Grafos de los genes con las combinaciones más robustas.***

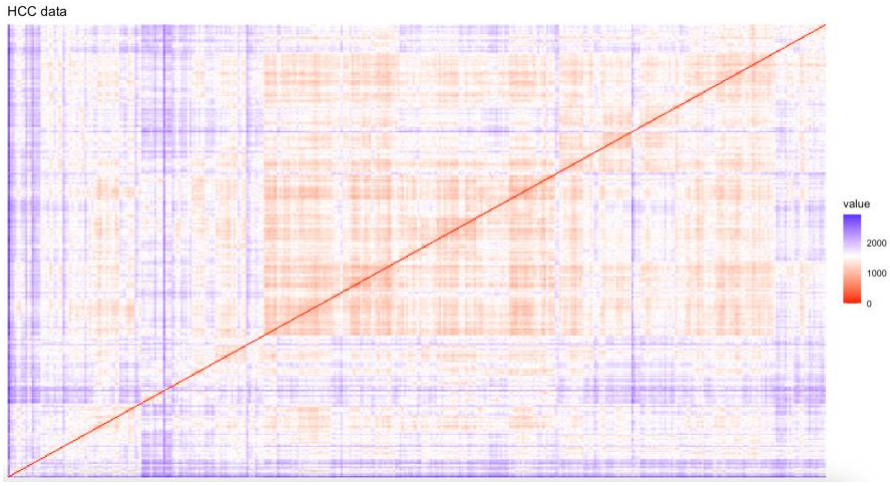


Anexo 3: TCGA

Anexo 3a: Pirámide de filtrado de los genes



Anexo 3b: Matriz de disimilitud de las muestras. El coeficiente de Hopkins es de 0.31.



Anexo 3c: Tabla con los coeficientes cophenéticos

	metrica_distancia	metodo_linkage	cophenetic
3	euclidean	average	0.502086315862586

27	minkowski	average	0.502086315862586
15	canberra	average	0.483991171108076
9	manhattan	average	0.479084952547383
2	euclidean	complete	0.37429966783388
26	minkowski	complete	0.37429966783388
4	euclidean	ward.D2	0.367543927141351
28	minkowski	ward.D2	0.367543927141351
14	canberra	complete	0.355840390746499
10	manhattan	ward.D2	0.341515445702309
16	canberra	ward.D2	0.337586649333173
17	canberra	centroid	0.325536300940116
5	euclidean	centroid	0.305333358639743
29	minkowski	centroid	0.305333358639743
8	manhattan	complete	0.297776494081479
11	manhattan	centroid	0.283866826950673
13	canberra	single	0.196778552883814
18	canberra	median	0.153378680065094
7	manhattan	single	0.141629890926716
1	euclidean	single	0.136620471665099
25	minkowski	single	0.136620471665099
12	manhattan	median	0.104394291706721
6	euclidean	median	0.0974137813125466
30	minkowski	median	0.0974137813125466

Anexo 3d: sumario de validación

```
Clustering Methods:
 hierarchical pam kmeans clara diana

Cluster sizes:
 2 3 4 5 6

Validation Measures:
                                  2         3         4         5         6

hierarchical Connectivity   675.7833  684.4440  726.7155 1428.8730 1493.8179
             Dunn             0.4569    0.4569    0.4569    0.4398    0.4398
             Silhouette       0.0422    0.0274    0.0184    0.0242    0.0198
pam          Connectivity  1578.2401 1873.9111 2204.7044 2190.8456 2409.9651
             Dunn             0.1801    0.2234    0.2414    0.2414    0.2234
             Silhouette       0.0496    0.0491    0.0392    0.0463    0.0420
kmeans       Connectivity  1164.5766 1835.8794 1987.8651 2011.0516 2262.2270
             Dunn             0.2317    0.1966    0.1965    0.1967    0.1966
             Silhouette       0.0582    0.0481    0.0444    0.0504    0.0450
clara        Connectivity  1295.7940 1899.1393 2569.6079 2651.6373 2770.5329
             Dunn             0.2214    0.1729    0.2234    0.2715    0.1762
             Silhouette       0.0524    0.0425    0.0339    0.0356    0.0333
diana        Connectivity  1068.1095 1477.1687 2144.0944 2288.2087 2555.7464
             Dunn             0.2769    0.2124    0.2144    0.2218    0.2239
             Silhouette       0.0596    0.0472    0.0382    0.0350    0.0307

Optimal Scores:

             Score    Method       Clusters
Connectivity 675.7833 hierarchical 2
Dunn           0.4569 hierarchical 2
Silhouette     0.0596 diana        2

NULL
```

8. BIBLIOGRAFÍA

1. Llovet, J., et al., 2021. Hepatocellular carcinoma. *Nature Reviews Disease Primers*, 7(1).
2. Junqueira, L. and José, C., 2015. *Histología Básica*. 12th ed. Madrid: Editorial Médica Panamericana, pp.318-330.
3. Costanzo and Linda S., 2014. *Fisiología*. Barcelona: Elsevier.
4. Cordero-Espinoza, L. and Huch, M., 2018. The balancing act of the liver: tissue regeneration versus fibrosis. *Journal of Clinical Investigation*, 128(1), pp.85-96.
5. Sia, D., et al., 2017. Liver Cancer Cell of Origin, Molecular Class, and Effects on Patient Prognosis. *Gastroenterology*, 152(4), pp.745-761.
6. Marquardt, J., Andersen, J. and Thorgeirsson, S., 2015. Functional and genetic deconstruction of the cellular origin in liver cancer.*Nature Reviews Cancer*, 15(11), pp.653-667.
7. Estes, C., et al., 2018. Modeling the epidemic of nonalcoholic fatty liver disease demonstrates an exponential increase in burden of disease. *Hepatology*, 67(1), pp.123-133.
8. Yen, Y., et al., 2021. Characteristics and etiologies of hepatocellular carcinoma in patients without cirrhosis: When East meets West. *PLOS ONE*, 16(1).
9. Cun.es. 2021. *Hepatocarcinoma: qué es, causas, síntomas y tratamiento. Clínica Universidad de Navarra*. [online] Available at: <https://www.cun.es/enfermedades-tratamientos/enfermedades/hepatocarcinoma>.
10. Medlineplus.gov. 2021. *Prueba de marcador tumoral AFP (alfafetoproteína): Prueba de laboratorio de MedlinePlus*. [online] Available at: <https://medlineplus.gov/spanish/pruebas-de-laboratorio/prueba-de-marcador-tumoral-afp-alfafetoproteina/>.
11. Lee YH, et al., Vascular invasion in hepatocellular carcinoma: prevalence, determinants and prognostic impact. J ClinGastroenterol. 2014 Sep;48(8):734-41. doi: 10.1097/MCG.0b013e3182a8a254. PMID: 24100755.
12. Pérez de Luque, et al.,2006. Survival of patientsreceiving a livertransplantforhepatocellular carcinoma, and risk of tumor recurrence. *REVISTA ESPAÑOLA DE ENFERMEDADES DIGESTIVAS Copyright © 2006 ARÁN EDICIONES, S. L.*, Vol. 98., pp.pp. 899-906.
13. Saab S, et al., *De novo* Hepatocellular Carcinoma after Liver Transplantation. *JClinTranslHepatol*. 0;3(4):284-287. doi: 10.14218/JCTH.2015.00033.
14. Zhang JA, Kwee SA, Wong LL. Late recurrence of hepatocellular carcinoma after liver transplantation. *Hepatoma Res* 2017;3:58-66. http://dx.doi.org/10.20517/2394-5079.2017.05
15. Sangro B, Iñarrairaegui M, Bilbao JI. Radioembolization for hepatocellular carcinoma. JHepatol. 2012 Feb;56(2):464-73. doi: 10.1016/j.jhep.2011.07.012. Epub 2011 Aug 2. PMID: 21816126.
16. Qiu, L., et al., 2017. Long Non-Coding RNAs in Hepatitis B Virus-Related Hepatocellular Carcinoma: Regulation, Functions, and Underlying Mechanisms. *International Journal of Molecular Sciences*, 18(12), p.2505.
17. Unfried, J. and Fortes, P., 2020. LncRNAs in HCV Infection and HCV-Related Liver Disease. *International Journal of Molecular Sciences*, 21(6), p.2255.

18. Vučićević, D., et al., 2015. Long ncRNA expression associates with tissue-specific enhancers. *Cell Cycle*, 14(2), pp.253-260.
19. Sukowati, C., et al., 2021. Circulating Long and Circular Noncoding RNA as Non-Invasive Diagnostic Tools of Hepatocellular Carcinoma. *Biomedicines*, 9(1), p.90.
20. Liebman, H., et al., 1984. Des-γ-Carboxy (Abnormal) Prothrombin as a Serum Marker of Primary Hepatocellular Carcinoma. *New England Journal of Medicine*, 310(22), pp.1427-1431.
21. Calderaro, et al., 2019. Molecular and histological correlations in liver cancer. *Journal of Hepatology*, 71(3), pp.616-630.
22. Calderaro, J., et al., 2017. Histological subtypes of hepatocellular carcinoma are related to gene mutations and molecular tumour classification. *Journal of Hepatology*, 67(4), pp.727-738.
23. Hoshida, Y., et al., 2009. Integrative Transcriptome Analysis Reveals Common Molecular Subclasses of Human Hepatocellular Carcinoma. *Cancer Research*, 69(18), pp.7385-7392.
24. Kurebayashi, Y., et al., 2018. Landscape of immune microenvironment in hepatocellular carcinoma and its additional impact on histological and molecular classification. *Hepatology*, 68(3), pp.1025-1041.
25. Ally, A., et al., 2017. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell*, 169(7), pp.1327-1341.e23.
26. Tan, P., et al., 2015. Clinicopathological indices to predict hepatocellular carcinoma molecular classification. *Liver International*, 36(1), pp.108-118.
27. Dow, M., et al., 2018. Integrative genomic analysis of mouse and human hepatocellular carcinoma. *Proceedings of the National Academy of Sciences*, 115(42), pp.E9879-E9888.
28. Hirschfield, H., et al., 2018. In vitro modeling of hepatocellular carcinoma molecular subtypes for anti-cancer drug assessment. *Experimental& Molecular Medicine*, 50(1), pp.e419-e419.
29. Chaisaingmongkol, J., et al., 2017. Common Molecular Subtypes Among Asian Hepatocellular Carcinoma and Cholangiocarcinoma. *Cancer Cell*, 32(1), pp.57-70.e3.
30. Candia, J., et al., 2020. The genomic landscape of Mongolian hepatocellular carcinoma. *Nature Communications*, 11(1).
31. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/
32. GitHub. 2021. *GitHub - janire98/lncrna: Master's degree (Computational methods in Science)*. [online] Available at: <https://github.com/janire98/lncrna>.
33. https://github.com/janire98/lncrna/blob/main/scripts/001_deseq_subread.R
34. GitHub. 2021. *lncrna/002_filtro_genes.R at main · janire98/lncrna*. [online] Available at: <https://github.com/janire98/lncrna/blob/main/scripts/002_filtro_genes.R>.
35. GitHub. 2021. *lncrna/003_rescalado_variables.R at main · janire98/lncrna*. [online] Available at: <https://github.com/janire98/lncrna/blob/main/scripts/003_rescalado_variables.R>.
36. GitHub. 2021. *lncrna/003_2_previsualizacion_tests.R at main · janire98/lncrna*. [online] Available at: <https://github.com/janire98/lncrna/blob/main/scripts/003_2_previsualizacion_tests.R>.

37. GitHub. 2021. *lncrna/004_obtencion_counts_nico.R at main · janire98/lncrna*. [online] Available at: <https://github.com/janire98/lncrna/blob/main/scripts/004_obtencion_counts_nico.R>.
38. GitHub. 2021. *lncrna/005_tendencia_cluster.R at main · janire98/lncrna*. [online] Available at: <https://github.com/janire98/lncrna/blob/main/scripts/005_tendencia_cluster.R>.
39. GitHub. 2021. *lncrna/006y7_metodo_cluster.R at main · janire98/lncrna*. [online] Available at: <https://github.com/janire98/lncrna/blob/main/scripts/006y7_metodo_cluster.R>.
40. GitHub. 2021. *lncrna/008_1_clusterings_automatizado.R at main · janire98/lncrna*. [online] Available at: <https://github.com/janire98/lncrna/blob/main/scripts/008_1_clusterings_automatizado.R>.
41. GitHub. 2021. *lncrna/008_2_clustering_general.R at main · janire98/lncrna*. [online] Available at: <https://github.com/janire98/lncrna/blob/main/scripts/008_2_clustering_general.R>.
42. GitHub. 2021. *lncrna/009_1_analisis_supervivencia.R at main · janire98/lncrna*. [online] Available at: <https://github.com/janire98/lncrna/blob/main/scripts/009_1_analisis_supervivencia.R>.
43. GitHub. 2021. *lncrna/009_2_supervivencia_gen_por_gen.R at main · janire98/lncrna*. [online] Available at: <https://github.com/janire98/lncrna/blob/main/scripts/009_2_supervivencia_gen_por_gen.R>.
44. GitHub. 2021. *lncrna/009_3_combinacion_genes.R at main · janire98/lncrna*. [online] Available at: <https://github.com/janire98/lncrna/blob/main/scripts/009_3_combinacion_genes.R>.
45. GitHub. 2021. *lncrna/009_4_supervivencia_combinaciones.R at main · janire98/lncrna*. [online] Available at: <https://github.com/janire98/lncrna/blob/main/scripts/009_4_supervivencia_combinaciones.R>.
46. GitHub. 2021. *lncrna/009_5_grafos_combinacion_genes.R at main · janire98/lncrna*. [online] Available at: <https://github.com/janire98/lncrna/blob/main/scripts/009_5_grafos_combinacion_genes.R>.
47. GitHub. 2021. *lncrna/010_1_desarrollo_tcga.R at main · janire98/lncrna*. [online] Available at: <https://github.com/janire98/lncrna/blob/main/scripts/010_1_desarrollo_tcga.R>.
48. GitHub. 2021. *lncrna/010_2_genes_comunes_tcga.R at main · janire98/lncrna*. [online] Available at: <https://github.com/janire98/lncrna/blob/main/scripts/010_2_genes_comunes_tcga.R>.
49. GitHub. 2021. *lncrna/010_3_diagrama_venn.R at main · janire98/lncrna*. [online] Available at: <https://github.com/janire98/lncrna/blob/main/scripts/010_3_diagrama_venn.R>.
50. Colaprico, A., et al., TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Research (05 May 2016) 44 (8): e71. (doi:10.1093/nar/gkv1507)

51. Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
52. Bolger, A. M., Lohse, M., &Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.
53. Liao Y, Smyth GK and Shi W (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Research 47(8), e47.
54. Gencodegenes.org. 2021. *GENCODE - Human Release 38*. [online] Available at: <https://www.gencodegenes.org/human/>.
55. Love, M.I., Huber, W., Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 Genome Biology 15(12):550 (2014)
56. Hadley Wickham and Dana Seidel (2020). scales: Scale Functions for Visualization. R package version 1.1.1. https://CRAN.R-project.org/package=scales
57. Datanovia.2021. *Assessing Clustering Tendency - Datanovia*. [online] Available at: <https://www.datanovia.com/en/lessons/assessing-clustering-tendency/>.
58. Xsliulab.github.io. 2021. [online] Available at: <https://xsliulab.github.io/Workshop/week10/r-cluster-book.pdf>.
59. Kevin Wright, Luo YiLan and Zeng RuTong (2021). clustertend: Check the Clustering Tendency. R package version 1.5. https://CRAN.R-project.org/package=clustertend
60. Brock, G., et al., (2008). clValid: An R Package for Cluster Validation. Journal of Statistical Software, 25(4), 1-22. URL https://www.jstatsoft.org/v25/i04/
61. Kleinbaum, D. and Klein, M., n.d.*Survival Analysis*. 3rd ed.
62. Hirata, M., et al., 2017. Cross-sectional analysis of BioBank Japan clinical data: A large cohort of 200,000 patients with 47 common diseases. *Journal of Epidemiology*, 27(3), pp.S9-S21.
63. Inserm. 2021. *Public Health Research · Inserm*. [online] Available at: <https://www.inserm.fr/en/our-research/public-health-research/>.
64. Zheng, H., et al., 2019. Benchmark of long non-coding RNA quantification for RNA sequencing of cancer samples. *GigaScience*, 8(12).
65. Li, G., et al., 2019. Identification of diagnostic long non-coding RNA biomarkers in patients with hepatocellular carcinoma. *Molecular Medicine Reports*,.
66. Liu, J., et al., 2020. Long noncoding RNA MYLK-AS1 promotes growth and invasion of hepatocellular carcinoma through the EGFR/HER2-ERK1/2 signaling pathway. *International Journal of Biological Sciences*, 16(11), pp.1989-2000.