Universitat Oberta de Catalunya

uoc.edu

Evaluación de análisis de clustering jerárquico en datos moleculares de alta dimensión.

Maria Isabel Lumbreras Herrera
Máster en Bioinformática y Bioestadística
Área 2. Subárea 13: Análisis de datos y técnicas de clustering

Carles Ventura Hoyo
Daniel Fernandez Martinez

24 de junio de 2020

FICHA DEL TRABAJO FINAL

Título del trabajo:	Evaluación de análisis de clustering jerárquico en datos moleculares de alta dimensión.
Nombre del autor:	*Maria Isabel Lumbreras Herrera.*
Nombre del consultor/a:	*Daniel Fernandez Martinez*
Nombre del PRA:	*Carles Ventura Royo*
Fecha de entrega (mm/aaaa):	24/06/2020
Titulación::	*Máster en Bioinformática y Bioestadística.*
Área del Trabajo Final:	*Área 2. Subárea 13: Análisis de datos y técnicas de clustering*
Idioma del trabajo:	*Castellano*
Palabras clave	*Clustering, K-means, distancia*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

EL objetivo de este estudio es realizar una clasificación de los pacientes de cáncer de mama en grupos molecularmente homogéneos**,** mediante la aplicación de clustering en función de los perfiles de expresión, y de establecer la correlación existente con la actual clasificación clínica y otros parámetros de posible interés para el tratamiento de los pacientes. Además, se presentará una alternativa al análisis jerárquico: el análisis de k-means; veremos las ventajas que este tiene sobre los modelos de clustering debido a que con este método disponemos de k-dimensiones en lugar de una sola. Por otro lado, se considerará si es necesario realizar un gráfico probabilístico con los resultados obtenidos.

Abstract (in English, 250 words or less):

The aim of this study is the classification of breast cancer patients into molecularly homogeneous groups through clustering based on their expression profiles, and the establishment of the existing correlation with the current clinical classification and other parameters of possible interest for the treatment of this disease. In addition, an alternative to hierarchical analysis will be presented: k-means analysis. We will see the advantages of this method over the clustering models, due to this method uses k-dimensions instead of just one dimension. On the other hand, it will be considered whether it is necessary to make a probabilistic graph with the obtained results.

Índice

1. Introducción 7
1.1 Contexto y justificación del Trabajo 7
1.2 Objetivos del Trabajo 8
1.3 Enfoque y método seguido 8
1.4 Planificación del Trabajo 9
1.5 Breve sumario de productos obtenidos 9
1.6 Breve descripción de los otros capítulos de la memoria 10
2. Métodos 12
2.1 Análisis de conglomerados 12
2.1.1 Clasificación de las técnicas de clasificación de análisis de conglomerados 13
2.1.3 Distancia euclídea 16
2.1.4 Método de Ward 17
2.2.5 K-means 18
2.2 Cappas de información 20
2.3 Modelos gráficos probabilísticos (MGP) 20

3. Resultados 21
3.1 Conjunto y origen de los datos 21
3.2 Composición de los datos, selección de variables y extracción de las características 21
3.3 Declaración de Ética 23
3.4 Filtrado de datos de expresión génica 23
3.5 Subtipos moleculares 25
3.6 Análisis jerárquico aglomerativo con la distancia euclídea y el método de Ward del conjunto de datos de expresión génica sobre el cáncer de mama 28
3.7 Análisis mediante el método de K-means del conjunto de datos de expresión génica sobre el cáncer de mama 35
3.8 Comparación de uHcl con K-means en el conjunto de datos de expresión génica de cáncer de mama 39
3.9 MGP sobre el conjunto de datos de expresión génica de cáncer de mama 39
3.10 Cappas de información sobre el conjunto de datos de expresión génica de cáncer de mama 40

4. Conclusiones 42
5. Glosario 44
6. Bibliografía 45

Lista de figuras

Ilustración 1. Diagrama de Gantt sobre la planificación de las tareas llevadas a cabo ……8

Ilustración 2. Tabla de clasificación de las distintas técnicas de análisis de conglomerados……..…....13

Ilustración 3. Enumeración de las variables escogidas para trabajar con el conjunto de datos clínico …….. 19

Ilustración 4. Resumen de las 6 primeras filas y las 7 primeras variables del conjunto de datos de expresión génica “dmseq …… 19

Ilustración 5. Resumen de las 6 primeras filas y las 7 primeras variables del conjunto de datos de expresión génica “mrnseq”…….. 22

Ilustración 6. Resumen de la base de datos de 1087 pacientes con cáncer de mama que contiene el subtipo molecular del tumor, proporcionada por TCGAquery_subtype(“brca”) de la librería TCGAbiolinks de Bioconductor……. 23

Ilustración 7. Proporción de subtipos moleculares en el conjunto de datos de expresión génica de cáncer de mama…… 23

Ilustración 8. Gráfico de barras con el número de genes distribuidos en cada subtipo molecular de cáncer de mama…… 24

Ilustración 9. Coeficientes de análisis aglomerativo con agnes() empleando los métodos "average", "single", "complete" y "Ward”…….……...……. 25

Ilustración 10. Comparación mediante cuatro dendogramas empleando clustering jerárquico aglomerativo con método Manhattan, distancia por máximos, distancia euclídea y método Camberra, usando en todos los casos la distancia de Ward en los datos escalados …….…… 26

Ilustración 11. Heatmap con distancia euclídea y método de Ward de los datos escalados de expresión génica de cáncer de mama. Pacientes en el eje x, y los genes en el eje y……. 27

Ilustración 12. Clustering jerárquico con los datos escalados, distancia euclídea y método de Ward……...……...…….. 27

Ilustración 13. Primeras 20 uniones clustering jerárquico con distancia euclídea y método de Ward.……...…..……. 28

Ilustración 14. Primeras 20 alturas clustering jerárquico con distancia euclídea y método de Ward.……..……. 28

Ilustración 15. Cantidad de genes agrupados en cada cluster empleando la distancia euclídea y el método de Ward…….. 28

Ilustración 16. Dendograma con distancia euclídea y método de Ward, que agrupa los pacientes separando por 5 colores en los datos escalados de expresión génica de cáncer de mama…….. 29

Ilustración 17. número pacientes con cada subtipo molecular ubicados en cada cluster…….. 33

Ilustración 18. Resumen de la agrupación en cluster de los 20 primeros genes mediante la función cluster_analysis() de la librería multiClust…….. 33

Ilustración 19. Dendograma obtenido mediante la función cluster_analysis() con distancia euclídea y método de Ward, separado por 5 colores en los datos escalados de expresión génica de cáncer de mama…….. 34

Ilustración 20. Algunos pacientes con el cluster al que se han asignado mediante el método k-mean con k=5…….. 36

Ilustración 11. Representación de la agrupación de pacientes por k-means con k=5 de los datos de expresión génica de cáncer de mama sobre las dos primeras componentes principales…….. 37

Ilustración 22. Veinte primeros genes con el número de cluster en el que se han ubicado mediante k-means con la función cluster_analysis() del paquete multiClust…….. 38

Ilustración 23. Expresión promedio en cada paciente en cada cluster, obtenida mediante la función avg_probe_exp() del paquete multiClust…….. 38

Ilustración 24. Comparación clusters de k-means y clustering jerárquico…….. 39

Ilustración 25. Red obtenida sobre la base de datos de expresión génica tras aplicar un modelo gráfico probabilístico…….. 40

Ilustración 26. Clasificación óptima de los primeros pacientes del conjunto de datos de expresión génica tras aplicar cappas de información…….. ...41

Ilustración 27. Clasificación con k=5 del conjunto de datos de expresión génica tras aplicar cappas de información…….. 41

Ilustración 28. Resultado de “consencusCluster” sobre el conjunto de datos de expresión génica….. 42

Ilustración 29. Cappas de información sobre el conjunto de datos de expresión génica…….. 42

1. Introducción

1.1 Contexto y justificación del Trabajo

El cáncer de mama es la neoplasia más común en mujeres a lo largo del mundo. Representa un 25% del total de los cánceres diagnosticados y supone la quinta causa de muerte asociada al cáncer. En 2012, se estimaron casi 1,7 millones de nuevos casos diagnosticados y 521.900 muertes asociadas.
Los tumores de cáncer de mama expresan ciertos marcadores que pueden ser analizados mediante inmunohistoquímica (IHC) proporcionando un sistema de clasificación y diagnóstico clínico. Los receptores de estrógenos (ER), el receptor de progesterona (PR) y el receptor del factor de crecimiento epidérmico humano 2 (HER2) son los marcadores más importantes. El 70% de los tumores esporádicos son ER+, de los que el 50% también son PR+, y la sobreexpresión de HER2 se observa en un 15% de los casos. Los tumores ER+ y PR+ responden mejor al tratamiento endocrino, mientras que los tumores con sobreexpresión de HER2 responden bien a terapias diana como el trastuzumab (Herceptin®, Roche). Aproximadamente, del 12 al 17% de todos los cánceres son negativos para ER, PR y HER2, y son conocidos como TNBC (Triple Negative Breast Cancer). Estos tumores son más agresivos que los de otros subtipos y, además, las terapias endocrinas y antiHER2 no son efectivas, dejando la quimioterapia como único tratamiento para estos pacientes.

Por otro lado, existe una clasificación por subtipos moleculares basada en expresión génica. De esta manera, el cáncer de mama se puede clasificar en Luminal A (que engloba a la mayoría de los receptores hormonales positivos poco proliferativos, Luminal B (que engloba a la mayoría de los receptores hormonales positivos muy proliferativos), Her2, Basal (en su mayoría triples negativos) y Normal (con un perfil de expresión más similar al tejido sano) (Perou et al. Nature 2000). [11]

La técnica de agrupación jerárquica (clustering jerárquico, HCL) es uno de los métodos de agrupamiento no supervisados más frecuentes; estos se basan exclusivamente en los datos de expresión para generar los resultados, sin hacer uso de ningún conocimiento previo relativo a la patología. Estos métodos se utilizan en los estudios de descubrimiento de clases en donde se intenta crear una nueva clasificación o descubrir nuevos subtipos moleculares con relevancia en el cáncer de mama, basándose solamente en las similitudes entre los perfiles de expresión.
Además, la técnica de k-means nos puede llevar a generar subtipos moleculares de cáncer de mama y encontrar información independiente a estos subtipos. Esto puede

ayudar a buscar predictores de respuesta y a acercar la clínica a una medicina personalizada ya que en función de las características moleculares de cada grupo puede ser más eficaz la toma de decisiones y la utilización idónea de los medicamentos.
Finalmente, si se precisa hacer un gráfico probabilístico, este nos permitirá evaluar los grupos obtenidos de forma diferente a otras técnicas convencionales, como el Significance Analisys of Microarrays (SAM).

1.2 Objetivos del Trabajo

1.1. Analizar el dataset por uHCL.
1.2. Evaluar distintas métricas.
1.3. Evaluar respecto a las clasificaciones clínicas y moleculares preexistentes.
2.1. Análizar el dataset por k-means
2.2. Evaluar la información adicional que se identifica respecto al uHCL.

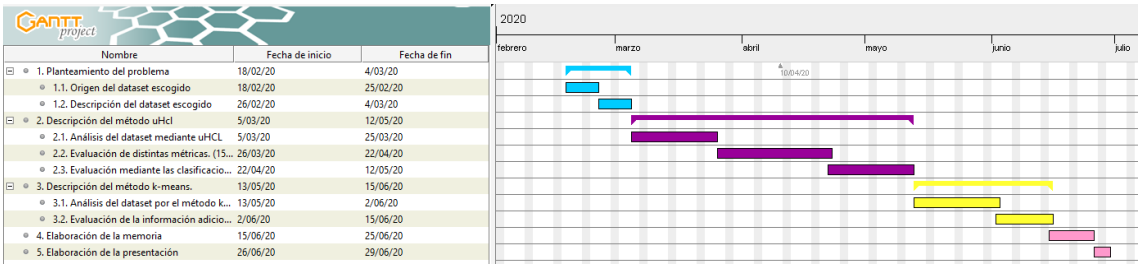
1.3 Enfoque y método seguido

El trabajo a desarrollar está enfocado en analizar la utilidad de diferentes herramientas matemáticas en un contexto con unas características especiales: el contexto clínico. Para ello, se van a seleccionar unas herramientas con mucho recorrido desde el punto de vista matemático y un problema que está bien caracterizado desde el lado clínico. Se podrían emplear otras herramientas más disruptivas, o acercase a una pregunta clínica más novedosa, pero el objetivo del trabajo es profundizar en esta herramienta, el uHCL, para evaluar los pros y contras de su uso para analizar datos moleculares masivos procedentes de muestras clínicas, y para ello se necesita trabajar con un set de datos que permita evaluar cómo está funcionando la herramienta. En cuanto a la selección del k-means como elemento de comparación, simplemente seleccionamos una herramienta sobre la que el equipo de investigación tiene experiencia previa, y un gran interés en seguir desarrollando y caracterizando aplicaciones, dado el gran potencial mostrado previamente. Sin duda hay multitud de herramientas que se podrían haber seleccionado para esta comparación (redes neuronales, arboles de decisión, etc.) pero no hay un criterio claro para decir que una es superior a la otra.

1.4 Planificación del Trabajo

La planificación del trabajo se ha llevado a cabo utilizando un diagrama de Grantt.

Anotar que en mitad del proceso tuvo que modificarse debido a un retraso provocado por la situación del COVID-19.



Nombre	Fecha de inicio	Fecha de fin
1. Planteamiento del problema	18/02/20	4/03/20
1.1. Origen del dataset escogido	18/02/20	25/02/20
1.2. Descripción del dataset escogido	26/02/20	4/03/20
2. Descripción del método uHcl	5/03/20	12/05/20
2.1. Análisis del dataset mediante uHCL	5/03/20	25/03/20
2.2. Evaluación de distintas métricas. (15...	26/03/20	22/04/20
2.3. Evaluación mediante las clasificacio...	22/04/20	12/05/20
3. Descripción del método k-means.	13/05/20	15/06/20
3.1. Análisis del dataset por el método k...	13/05/20	2/06/20
3.2. Evaluación de la información adicio...	2/06/20	15/06/20
4. Elaboración de la memoria	15/06/20	25/06/20
5. Elaboración de la presentación	26/06/20	29/06/20

Ilustración 1. Diagrama de Gantt sobre la planificación de las tareas llevadas a cabo

Hitos

1. Planteamiento del problema. (12 días)
1.1. Origen del *dataset* escogido. (6 días)
1.2. Descripción del *dataset* escogido. (6 días)
2. Descripción del método uHCL. (50 días)
2.1. Análisis del *dataset* mediante uHCL. (15 días)
2.2. Evaluación de distintas métricas. (20 días)
2.3. Evaluación mediante las clasificaciones clínicas y moleculares preexistentes. (15 días)
3. Descripción del método k-means. (25 días)
3.1. Análisis del *dataset* por el método k-means. (15 días)
3.2. Evaluación de la información adicional que se identifica respecto al método uHCL. (10 días)
4. Elaboración de la memoria (9 días)
5. Elaboración de la presentación (2 días)

1.6 Breve descripción de los otros capítulos de la memoria

Métodos:

Explicación sobre el análisis de conglomerados y clasificación de los distintos tipos, centrándonos en el método jerárquico aglomerativo con distancia euclídea y método de Ward, y en el método K-means.

También se incluye una breve descripción sobre los modelos gráficos probabilísticos y las cappas de información.

Resultados:

Explicación de los pasos seguidos a la hora de seleccionar las características clínicas más relevantes de la base de datos clínica.

Explicación del filtrado realizado a los datos de expresión génica.

Descripción de los pasos seguidos a la hora de aplicar los métodos de conglomerados y a la de elegir el mejor método y la distancia óptima para nuestro análisis.

Resultados obtenidos tras hacer MGP y cappas de información en la base de datos de expresión génica.

Conclusiones:

Planteamiento de las conclusiones finales basadas en los resultados obtenidos.

Glosario:

Sección dedicada a definir términos utilizados comúnmente durante el proyecto y que pueden no ser de conocimiento general.

Bibliografía

Incluye la bibliografía usada a lo largo del proyecto

2. Métodos

2.1 Análisis de conglomerados

El análisis de conglomerados (clusters), también llamados métodos de clasificación no supervisada, tienen como objetivo agrupar elementos de manera homogénea en función de sus similitudes. Lo que se pretende al aplicar este análisis es crear una partición entre unos datos que se sospecha que son heterogéneos en distintos grupos de manera que todos los elementos queden clasificados dentro de un grupo homogéneo con respecto al resto de elementos en el mismo, estructurando, a su vez, los elementos de un conjunto de forma jerárquica por su similitud (construyendo jerarquías) y dividiendo, así, las variables en grupos.
Es decir, partiendo de una muestra Ξ de m individuos, $X_1, \ldots, X_m$, cada uno de los cuales está representado por un vector n-dimensional, $X_j = (x_{j1}, x_{j2}, \ldots, x_{jn})'$, $j = 1, \ldots, m$ y debemos encontrar una partición de la muestra en regiones $\omega_1, \ldots, \omega_c$ de forma que

$$\bigcup_{i=1}^{c} w_i = \Xi$$

$\omega_i \cap \omega_j = \emptyset$; $i \neq j$

El número de formas en las que se pueden clasificar m observaciones en k grupos es un número de Stirling de segunda especie (Abramowitz y Stegun, 1968):

$$S_m^{(k)} = \frac{1}{k!}\sum_{i=0}^{k}(-1)^{k-i}\binom{k}{i}i^m$$

Al ser, normalmente, el número de grupos desconocido, el número de posibilidades es la suma de números de Stirling:

$\sum_{j=1}^{m} S_m^{(j)}$ para m observaciones.

Este es un número es excesivamente grande, por lo que habrá un gran número de clasificaciones. Para resolver este problema se usa el análisis de conglomerados.

2.1.1 Clasificación de las técnicas de clasificación de análisis de conglomerados.

Existen múltiples métodos de clasificaciones dentro del análisis de conglomerados. A grandes rasgos se pueden distinguir dos grupos: métodos jerárquicos y métodos no jerárquicos o de partición. La diferencia fundamental entre estos es que los algoritmos jerárquicos utilizan la matriz de distancias o similitudes entre elementos mientras que los métodos de partición utilizan la matriz de datos.

º El objetivo de los métodos jerárquicos es agrupar *clusters* para formar uno nuevo o bien separar alguno ya existente para dar origen a otros dos, de tal forma que se minimice alguna función de distancia o bien se maximice alguna medida de similitud. Los métodos jerárquicos se subdividen a su vez en aglomerativos y disociativos. Los aglomerativos comienzan el análisis con tantos grupos como individuos haya en el estudio. A partir de ahí se van formando grupos de forma ascendente, hasta que, al final del proceso, todos los casos están englobados en un mismo conglomerado. Los métodos disociativos o divisivos realizan el proceso inverso al anterior.

A su vez, existen diversos criterios para realizar el agrupamiento. Dentro de los aglomerativos tenemos:

1. Método del amalgamamiento simple.
2. Método del amalgamamiento completo.
3. Método del promedio entre grupos.
4. Método del centroide.
5. Método de la mediana.
6. Método de Ward.

Dentro de los métodos disociativos, destacan, además de los anteriores:

1. El análisis de asociación.
2. El detector automático de interacción.

º En cuanto a los métodos no jerárquicos, tienen como objetivo realizar una sola partición de los individuos en K grupos.

Según Pedret (1986) [1] se pueden agrupar los métodos no jerárquicos en cuatro familias:

1. Métodos de Reasignación

   Permiten que un individuo asignado a un grupo en un determinado paso del proceso sea reasignado a otro grupo en un paso posterior, si ello optimiza el criterio de selección. El proceso acaba cuando no quedan individuos cuya reasignación permita optimizar el resultado que se ha conseguido. Dentro de estos métodos están:

   a) El método *K*-Medias.
   b) El Quick-Cluster análysis.
   c) El método de Forgy.
   d)) El método de las nubes dinámicas.

2. Métodos de búsqueda de la densidad.

Dentro de estos métodos están los que proporcionan una aproximación tipológica y una aproximación probabilística.

En el primer tipo, los grupos se forman buscando las zonas en las cuales se da una mayor concentración de individuos. Entre ellos destacan:

a) El análisis modal de Wishart.
b) El análisis Taxmap.
c) El método de Fortin.

En el segundo tipo se parte del postulado de que las variables siguen una ley de probabilidad según la cual los parámetros varían

de un grupo a otro. Se trata de encontrar los individuos que pertenecen a la misma distribución. Entre los métodos de este tipo destaca el método de las combinaciones de Wolf.

3. Métodos directos.

   Permiten clasificar simultáneamente a los individuos y a las variables. El algoritmo más conocido dentro de este grupo es el Block-Clustering.

4. Métodos de reducción de dimensiones.

   Estos métodos consisten en la búsqueda de unos factores en el espacio de los individuos; cada factor corresponde a un grupo. Se les conoce como Análisis Factorial tipo Q.

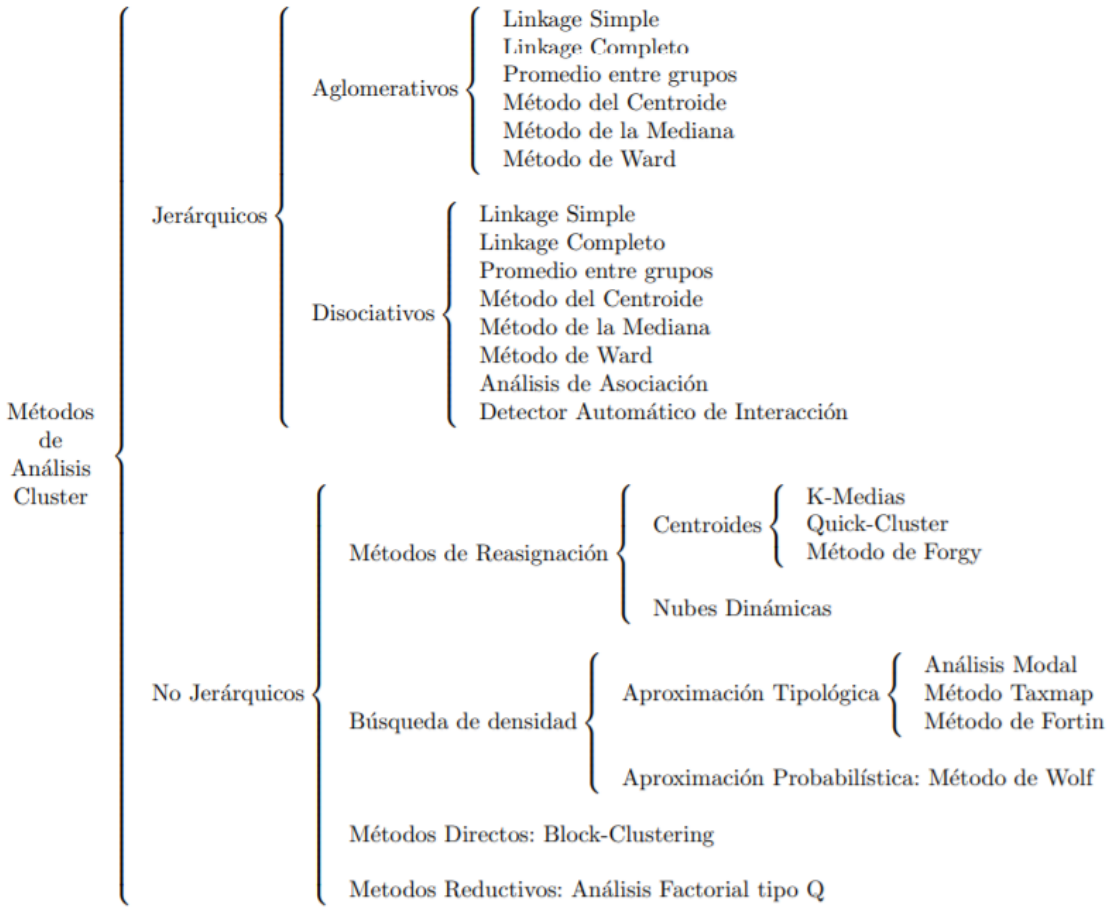


Ilustración 2. Tabla de clasificación de las distintas técnicas de análisis de conglomerados

Entre los diversos métodos existentes para el análisis de conglomerados, este estudio se centró en la aplicación del análisis jerárquico de conglomerados (uHcl) de forma aglomerativa y usando la distancia euclídea y el método de Ward, y el método no jerárquico de K-means.

2.1.2 Método de Ward

El método de Ward parte de los elementos directamente, en lugar de utilizar la matriz de distancias, y se define una medida global de la heterogeneidad de una agrupación de observaciones en grupos. Esta medida es la suma de las distancias euclÌdeas al cuadrado entre cada elemento y la media de su grupo:

$$\mathbf{W} = \sum_{g} \sum_{i \in g} (\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)'(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)$$

donde $\bar{x}_g$ es la media del grupo g. Con este criterio se supone que cada dato forma un grupo, g = n y por tanto W es cero. Después de esto se unen aquellos elementos que producen el mínimo incremento de W; o lo que es lo mismo, tomar los más próximos con la distancia euclídea. Así, nos quedamos con n-1 grupos, n − 2 de un elemento y uno de dos elementos. Repetimos el mismo proceso sucesivamente hasta tener un único grupo. Los valores de W van indicando el crecimiento del criterio al formar grupos y pueden utilizarse para decidir cuantos grupos naturales contienen nuestros datos. Puede demostrarse que, en cada etapa, los grupos que debe unirse para minimizar W son aquellos tales que:

$$\min \frac{n_a n_b}{n_a + n_b} (\overline{\mathbf{x}}_a - \overline{\mathbf{x}}_b)'(\overline{\mathbf{x}}_a - \overline{\mathbf{x}}_b)$$

2.1.3 Distancia euclídea

Los métodos jerárquicos parten de una matriz de distancias o similitudes entre los elementos de la muestra y construyen una jerarquía basada en una distancia. Si todas las variables son continuas, la distancia más utilizada es la distancia euclídea entre las variables estandarizadas. No es, en general, recomendable utilizar las distancias de Mahalanobis, ya que la única matriz de covarianzas disponible es la de toda la muestra, que puede mostrar unas correlaciones muy distintas de las que existen entre las variables dentro de los grupos.

Consideremos que tomamos dos individuos de la población (dos filas en la matriz de datos X):

$$x_i = (x_{i1}, \ldots, x_{in})'$$

$$x_j = (x_{j1}, \ldots, x_{jn})'$$

Recordemos la derivada de la norma L2 de un vector, que es la generalización a más de dos dimensiones de la distancia entre dos puntos en el plano

$$||x_i||_2 = \sqrt{x_i' x_i} = \sqrt{\sum_{l=1}^{n} x_{il}^2}$$

obteniéndose, a partir de ella, la distancia euclídea

$$d_2(x_i, x_j) = ||x_i - x_j||_2 = \sqrt{(x_i - x_j)'(x_i - x_j)} = \sqrt{\sum_{l=1}^{n} (x_{il} - x_{jl})^2}$$

Esta métrica tiene la propiedad, al igual que la norma L2, de que todos sus valores son invariantes respecto de las transformaciones ortogonales $\tilde{x}_i = \theta x_i$, donde θ es una matriz n × n que verifica

$$\theta'\theta = \theta\theta' = I$$

Así, nos queda la distancia euclídea de la siguiente manera:

$$d_2(\theta x_i, \theta x_j) = d_2(x_i, x_j)$$

verificándose que d_2 es invariante.

2.1.4 K-means

K-means es el proceso de asignar cada individuo al cluster (de los K prefijados) con el centroide más próximo (según MacQueen, en 1972).

El centroide es la posición definida por la media de cada una de las dimensiones (variables) de las observaciones que forman el *cluster.*

K-means es uno de los métodos que se emplean para el análisis de conglomerados no jerárquicos o de particiones. El centroide se calcula a partir de los miembros del cluster tras cada asignación en lugar de al final de cada ciclo, como ocurre con otros métodos de análisis de particiones.

El proceso que sigue este método es coger los K primeros casos como clusters unitarios, asignar cada uno de los m − K individuos restantes al cluster con el centroide más próximo y después de cada asignación recalcular el centroide del cluster obtenido. Tras asignar todos los individuos, se toman los centroides de los clusters existentes como puntos semilla fijos y se asigna cada dato al punto semilla más cercano. Finalmente, se alternan los pasos anteriores hasta que ningún individuo cambie de cluster; es decir, hasta que el proceso converja.

Si consideramos $C_1, \ldots, C_K$ como los sets formados por los índices de las observaciones de cada uno de los *clusters* (por ejemplo, el set C1C1 contiene los índices de las observaciones agrupadas en el *cluster* 1). Tenemos que:
º Toda observación pertenece al menos a uno de los *K clusters*

- $C_1 \cup C_2 \cup \ldots \cup C_K = \{1, \ldots, n\}$

º Ninguna observación pertenece a más de un *cluster* a la vez

- $C_k \cap C_{k'} = \emptyset$

Podemos definir de dos formas la varianza interna de un *cluster* ($\mathrm{W(C_K)}$)

1. Como la suma de las distancias euclídeas al cuadrado entre cada observación (x_i) y el centroide (μ):

$$W(C_k) = \sum_{x_i, \in C_k} (x_i - \mu_k)^2$$

2. Como la suma de las distancias euclídeas al cuadrado entre todos los pares de observaciones que forman el *cluster*, dividida entre el número de observaciones del *cluster:*

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

El objetivo es minizar la suma total de varianza interna $\sum_{k=1}^{k} \mathrm{W(C_K)}$. Obtener un resultado exacto de esto es muy complejo; sin embargo, el proceso para obtener una solución muy buena es el siguiente:

1. Se asigna aleatoriamente un número entre 1 y K a cada observación; una asignación inicial aleatoria de las observaciones a los *clusters*.
2. Iterar los siguientes pasos hasta que la asignación de las observaciones a los *clusters* no cambie o se alcance un número máximo de iteraciones establecido por el usuario.
    - 2.1 Para cada uno de los *clusters* calcular su centroide.
    - 2.2 Asignar cada observación al *cluster* cuyo centroide está más próximo.

2.2 Cappas de información

Nuestro grupo ha creado un método basado en aplicar en primer lugar un k-means para seleccionar aquellos genes más importantes para clasificar nuestro conjunto de datos y, posteriormente, un consensus cluster para escoger el número óptimo de grupos en los que dividir a nuestra muestra de pacientes. Una vez establecida la lista de genes de interés, estos genes se excluyen de la base de datos y se repite el análisis de manera recurrente. Este método ha demostrado su utilidad a la hora de determinar diferentes capas de información biológica en los datasets de tumores, por ejemplo, información sobre respuesta inmune del tumor e información molecular del propio tumor. [24]

2.3 Modelos gráficos probabilísticos (MPG)

Los modelos gráficos probabilísticos (MGP), fueron desarrollados en los campos de la Inteligecia Artificial, las Matemáticas y la Economía. Los MGP, compatibles con datos de alta dimensión, consisten en un modelo gráfico no dirigido basado en el criterio de información bayesiano (BIC). La obtención de un MGP se basa en la creación de un árbol de expansión con la probabilidad máxima y posteriormente una búsqueda hacia delante en la que se van añadiendo aristas hasta obtener un modelo óptimo que reduce el BIC lo máximo posible y conserva la capacidad de descomposición del grafo inicial. [25]
Este tipo de gráficos nos permiten, a partir de los datos de expresión y sin otra información a priori, construir una red de relaciones entre nuestras variables (genes o proteínas). Han demostrado su utilidad a la hora de caracterizar molecularmente diferentes grupos de tumores como cáncer de mama, cáncer de vejiga músculo-invasivo o melanoma. [24]

1. Resultados

3.1 Conjunto y origen de los datos

El conjunto de datos elegido trata sobre *Breast invasive carcinoma*, una de las bases de datos del **TCGA**, la cual se pueden descargar directamente desde la web **firebrowse** [http://firebrowse.org/] [6] donde hay datos de prácticamente todos los tipos de cáncer.

Por un lado, se escogió la base de datos clínica (archivo All_CDEs.txt de *Clinical_Pick_Tier1) que* contiene 131 parámetros sobre los datos clínicos de cáncer de mama.

Por otro lado, se descargó la base de datos de expresión mRNAseq (*BRCA.rnaseqv2__illuminahiseq_rnaseqv2__unc_edu__Level_3__RSEM_genes_normalized__data.data.txt*) con los genes normalizados y en el cual se encontraban los pacientes en columnas y las expresiones génicas en las filas.

3.2 Composición de los datos, selección de variables y extracción de las características

Tras cargar en R el conjunto de datos con las características clínicas (función read.delim() de la librería readr [13])se decidió trasponer y asignar los códigos de los pacientes como nombres de filas y las características clínicas como nombres de las columnas. Posteriormente se seleccionaron las características más relevantes, como fueron características de tamaño tumoral (t), afectación ganglionar (n), metástasis (m), estado (stage), estatus vital, días hasta la muerte, días hasta el último seguimiento, estatus de her2 y los niveles de expresión de los receptores hormonales, estrógenos (ER) y progesterona; además de la edad, la raza, el género y demás parámetros, entre otras características. Posteriormente se unieron las columnas “días hasta la muerte” (days_to_death) y “días hasta el último seguimiento” (days_to_last_followup) en una sola (“days”) ya que ambas eran complementarias.

```
 [1] "vital_status"
 [2] "age_at_initial_pathologic_diagnosis"
 [3] "breast_carcinoma_estrogen_receptor_status"
 [4] "breast_carcinoma_progesterone_receptor_status"
 [5] "breast_carcinoma_surgical_procedure_name"
 [6] "cytokeratin_immunohistochemistry_staining_method_micrometastasis_indicator"
 [7] "gender"
 [8] "her2_immunohistochemistry_level_result"
 [9] "history_of_neoadjuvant_treatment"
[10] "lab_proc_her2_neu_immunohistochemistry_receptor_status"
[11] "lab_procedure_her2_neu_in_situ_hybrid_outcome_type"
[12] "menopause_status"
[13] "metastatic_site_at_diagnosis-2"
[14] "metastatic_site_at_diagnosis-3"
[15] "metastatic_site_at_diagnosis-4"
[16] "metastatic_site_at_diagnosis"
[17] "metastatic_site_at_diagnosis_other"
[18] "number_of_lymphnodes_positive_by_he"
[19] "number_of_lymphnodes_positive_by_ihc"
[20] "pathologic_m"
[21] "pathologic_n"
[22] "pathologic_stage"
[23] "pathologic_t"
[24] "person_neoplasm_cancer_status"
[25] "race"
[26] "radiation_therapy"
[27] "system_version"
[28] "targeted_molecular_therapy"
[29] "year_of_initial_pathologic_diagnosis"
[30] "days"
```

Ilustración 3. Enumeración de las variables escogidas para trabajar con el conjunto de datos clínico (dataset compuesto por 1.097 filas y 30 columnas exportado como "datos_clinicos.txt").

Tras importar el *dataset* de expresión génica en R, se eliminaron las filas que no contenían el nombre del gen (contenidos en la primera columna). Posteriormente, gracias al paquete *tidyverse* [12], se modificó la primera columna con la identificación de los genes y nos quedamos sólo con lo que refería al *gene_symbol*, eliminando lo que procedía al símbolo | . Nota: más adelante y después de realizar el filtrado se asignará esta columna como nombre de filas en nuestros datos.
Además, volviendo a hacer uso de la librería *tidyverse* [12], se modificó la primera fila que contenía la identificación de los pacientes para que tuviera el mismo formato que en la base de datos clínica, tras hacer esto se asignó como nombre de columnas en nuestros datos de expresión génica.

	Hybridization <fctr>	tcga-3c-aaau <dbl>	tcga-3c-aali <dbl>	tcga-3c-aalj <dbl>	tcga-3c-aalk <dbl>	tcga-4h-aaak <dbl>	tcga-5l-aat0 <dbl>
31	A1BG	197.0897	237.3844	423.2366	191.0178	268.8809	203.7718
32	A1CF	0.0000	0.0000	0.9066	0.0000	0.4255	0.0000
33	A2BP1	0.0000	0.0000	0.0000	0.0000	3.8298	0.5866
34	A2LD1	102.9634	70.8646	161.2602	62.5072	154.3702	111.5354
35	A2ML1	1.3786	4.3502	0.0000	1.6549	3.4043	1.1732
36	A2M	5798.3746	7571.9793	8840.3989	10960.2193	9585.4426	12331.3213

Ilustración 4. Resumen de las 6 primeras filas y las 7 primeras variables del conjunto de datos de expresión génica "dmseq" (dataset compuesto por 20.502 filas y 1.213 columnas exportado como "datos_expresión_génica.txt").

Posteriormente, se volvió a recurrir a la función "merge" (librería dplyr) para unir la base de datos de expresión génica a la base de datos clínica buscando coincidencias entre las identificaciones de los pacientes presentes, y con el mismo formato tras realizar las transformaciones anteriores, en ambas bases de datos (*dataset* compuesto por 1.212 filas y 2.031 columnas exportado como *"datos_completos.txt")*

3.3 Declaración de Ética

Se accedió a los datos incluidos en este estudio desde TCGA de acuerdo con las pautas de acceso abierto y provenientes de artículos publicados previamente donde se obtuvieron los consentimientos informados por escrito de acuerdo con los Comités de Ética de las instituciones locales [27 , 29 - 31]. Hay más información disponible en: http://cancergenome.nih.gov/abouttcga/policies/informedconsent .[10]

3.4 Filtrado de datos de expresión génica

Estas bases de datos del TCGA usan el método RNAseq para ver la expresión de los genes, este método utiliza la secuenciación masiva para revelar la cantidad de ARN de una muestra; es decir, cuenta el número de copias directamente. Esto no quiere decir que un 0 es una copia de un gen, ya que existe un mínimo de detección bajo el cual la técnica automáticamente asigna un 0 a ese gen. Es un 0 técnico, que sí, puede ser que esa muestra tenga una expresión muy pequeña de ese gen, pero también puede ser que la técnica no haya detectado la expresión del mismo. Así, el primer paso de filtrado fue deshacerse de los ceros. Para ello, se calculó

el porcentaje de 0s que había en cada fila (gen) mediante la función “apply”:

*por_0<-apply(dmseq[,-1],1, function(x)(sum(x==0))*100/lenght(dmseq[,-1]))*

Posteriormente, se unió esta columna al *dataset*, pudiendo eliminar de esta manera las primeras filas que contenían más de un 25% de 0s.

Así nos quedamos con 12.653 genes en los cuales ninguna muestra tenía un 0 técnico. Pudimos perfectamente trabajar con esos 12.653 genes porque son una muestra lo suficientemente grande para hacer nuestros análisis. Una vez filtrada la base de datos, eliminamos la columna que contenía el porcentaje de 0s. Aún así, seguíamos teniendo 0s en la base de datos. Para deshacerse de estos, se imputaron valores aleatorios que se encontraban en la parte inferior de la distribución de los datos. Para ello se simuló en R el procedimiento que emplea el software “Perseus” (http://www.coxdocs.org/doku.php?id=perseus:user:activities:matrixprocessing:imputation:replacemissingfromgaussian). En resumen, lo que se hizo fue transformar la base de datos a logaritmos en base 2 (función log2() en R). Tras calcular el Log_2 de 0 se obtuvieron unos valores -Inf que, mediante la ayuda de la librería “MCMCglmm” [14] se imputaron a la base de la cola de una distribución normal con la función rtnorm(), basándose en los valores que empleaba el software “Perseus” se empleó *lower = m-1.8-0.3*s* como límite inferior y *upper = m-1.8+0.3*s* como límite superior de la distribución normal.

dmlog<-log2(dmseq[,-1])

library(MCMCglmm)

m<-mean(dmlog!=-Inf)

s<-sd(dmlog!=-Inf)

*nor <- rtnorm(sum(dmlog==-Inf),m,s, lower = m-1.8-0.3*s, upper = m-1.8+0.3*s)*

Finalmente, para reducir algo más la base de datos, se seleccionaron los 2000 genes con más variabilidad. Para ello se volvió a recurrir a la función “apply” calculando la varianza de todos los genes, se añadió como nueva columna al *dataset* y se ordenó en función de la misma (librería “dplyr” [15]), posteriormente se seleccionaron las 2000 primeras filas y eliminamos la columna de la varianza anteriormente añadida:

```
dv<-apply(dmlog,1, function(x)var(x))

library(dplyr)

#Ordenamos en función a la varianza de forma descendente
dmr<-arrange(dmr,desc(dv))

#Cogemos las 2000 primeras filas

mrnseq<-dmr[1:2000,]

# Eliminamos la columna de la varianza
mrnseq<-mrnseq[,-2]
```

Así, los posteriores análisis se realizaron con los 2000 genes más variables del conjunto de datos en logaritmo en base 2 y con los valores imputados de la cola de una distribución normal en sustitución de los valores perdidos.

Hybridization <fctr>	tcga-3c-aaau <dbl>	tcga-3c-aali <dbl>	tcga-3c-aalj <dbl>	tcga-3c-aalk <dbl>	tcga-4h-aaak <dbl>
CPB1	16.161328	4.165470	11.1070878	6.916607	2.467775
SCGB2A2	15.969215	16.161796	-0.7826457	8.740803	14.983308
GSTM1	10.833656	6.513396	2.6659157	11.636676	11.365392
SCGB1D2	12.231115	10.533119	-0.1414619	6.534138	9.333393
TFF1	6.308730	7.538931	7.8185694	13.595509	12.354296
MUCL1	8.302445	7.296004	11.3658663	12.099103	7.263194

Ilustración 5. Resumen de las 6 primeras filas y las 7 primeras variables del conjunto de datos de expresión génica “mrnseq” (dataset compuesto por 20.000 filas y 1.213 columnas exportado como "datos_filtrados_expresión_génica.txt").

3.5 Subtipos moleculares

Molecularmente, el cáncer de mama se puede clasificar en: Luminal A (que engloba a la mayoría de los receptores hormonales positivos poco proliferativos), Luminal B (que engloba a la mayoría de los receptores hormonales positivos muy proliferativos), Her2+, Basal

(que son en su mayoría triple negativo) y Normal, que presentan un perfil molecular más cercano al del tejido mamario sano [28] . Además de esta clasificación, en los últimos años se han descrito 3 subtipos moleculares más: interferón, apocrino molecular y “Claudin-low”. [29]

La asignación de los 5 subtipos moleculares se obtuvo mediante el uso de un paquete de Bioconductor para el análisis integrado de datos, “TCGAbiolinks”[7]. La función TCGAquery_subtype() [8] proporcionada por esta librería, nos permite obtener los subtipos moleculares, entre otras variables, de un tumor dado, en este caso del “brca”.

```
'data.frame':   1087 obs. of  24 variables:
 $ patient                           : Factor w/ 1087 levels "tcga-3c-aaau",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ Tumor.Type                        : chr  "BRCA" "BRCA" "BRCA" "BRCA" ...
 $ Included_in_previous_marker_papers : chr  "NO" "NO" "NO" "NO" ...
 $ vital_status                      : chr  "Alive" "Alive" "Alive" "Alive" ...
 $ days_to_birth                     : chr  "-20211" "-18538" "-22848" "-19074" ...
 $ days_to_death                     : chr  "NA" "NA" "NA" "NA" ...
 $ days_to_last_followup             : chr  "4047" "4005" "1474" "1448" ...
 $ age_at_initial_pathologic_diagnosis: num  55 50 62 52 50 42 63 52 70 59 ...
 $ pathologic_stage                  : chr  "NA" "Stage_II" "Stage_II" "Stage_I" ...
 $ Tumor_Grade                       : chr  "NA" "NA" "NA" "NA" ...
 $ BRCA_Pathology                    : chr  "NA" "NA" "NA" "NA" ...
 $ BRCA_Subtype_PAM50                : chr  "LumA" "Her2" "LumB" "LumA" ...
 $ MSI_status                        : chr  "NA" "NA" "NA" "NA" ...
 $ HPV_Status                        : chr  "NA" "NA" "NA" "NA" ...
 $ tobacco_smoking_history           : chr  "NA" "NA" "NA" "NA" ...
 $ CNV Clusters                      : chr  "C6" "C6" "C6" "C1" ...
 $ Mutation Clusters                 : chr  "C7" "C9" "C4" "C5" ...
 $ DNA.Methylation Clusters          : chr  "C1" "C2" "C2" "C2" ...
 $ mRNA Clusters                     : chr  "C1" "C2" "C2" "C2" ...
 $ miRNA Clusters                    : chr  "C3" "C3" "C2" "C2" ...
 $ lncRNA Clusters                   : chr  "NA" "NA" "NA" "NA" ...
 $ Protein Clusters                  : chr  "NA" "C2" "NA" "C2" ...
 $ PARADIGM Clusters                 : chr  "C5" "C4" "C4" "C6" ...
 $ Pan-Gyn Clusters                  : chr  "NA" "C4" "NA" "C4" ...
```

Ilustración 6. Resumen de la base de datos de 1087 pacientes con càncer de mama que contiene el subtipo molecular del tumor, proporcionada por TCGAquery_subtype(“brca”) de la librería TCGAbiolinks de Bioconductor.

Tras obtener conjunto de datos anterior, haciendo uso de la función merge (), buscamos coincidencias entre los pacientes contenidos en este con los incluidos en nuestra base de datos de expresión génica. Obteniendo así un nuevo *dataset* (exportado como "datos_subt.txt") en el que se incluyen 1081 pacientes con 2023 variables, entre ellos los subtipos moleculares y las expresiones de cada gen.

```
brca_subtypes <- TCGAbiolinks::TCGAquery_subtype("brca")
brca_subtypes<-as.data.frame(brca_subtypes)

mrnseq1<-cbind(rownames(mrnseq),mrnseq)
```

```
library(tidyverse)
library(stringr)
# Sacamos la primera columna de "brca_subtypes" y cambiamos el tipo de letra a minúsculapatient<-as.vector(brca_subtypes[,1])

brca_subtypes<-brca_subtypes[,-1]
patient <- str_to_lower(patient, locale = "es")
brca_subtypes<-cbind(patient,brca_subtypes)
m5<-merge(mrnseq,brca_subtypes,by.x="rownames(mrnseq)", by.y="patient")
rownames(m5)<-m5[,1]
m5<-m5[,-1]
```

Y se vio la proporción que había de cada subtipo en el conjunto de datos:

```
prop.table(table(m5$BRCA_Subtype_PAM50))*100
```

```
     Basal      Her2      LumA      LumB    Normal
17.576318  7.585569 51.803885 19.333950  3.700278
```

Ilustración 7. Proporción de subtipos moleculares en el conjunto de datos de expresión génica de cáncer de mama.

```
ggplot(m4,aes(BRCA_Subtype_PAM50))+geom_bar()
```

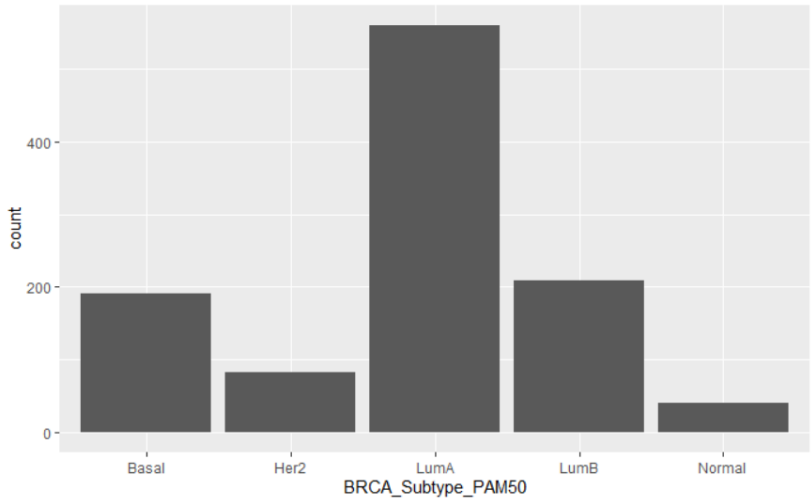


Ilustración 8. Gráfico de barras con el número de genes distribuidos en cada subtipo molecular de cáncer de mama

Observando los resultados de las tablas de proporciones junto con el gráfico de barras, se vio que más de la mitad de los genes eran del subtipo molecular Luminal A, los subtipos Luminal B y Basal estaban muy igualados y había una baja proporción de Her2 y Normal.

3.6 Análisis jerárquico aglomerativo con la distancia euclídea y el método de Ward del conjunto de datos de expresión génica sobre el cáncer de mama

Lo primero que se hizo para hacer uHcl en el conjunto de datos de expresión génica fue escalar los datos, mediante la función scale(). Escalar y centrar las variables de forma que todas ellas tengan media 0 y desviación estándar 1 antes de calcular la matriz de distancias asegura que todas las variables tengan el mismo peso cuando se realice el *clustering*.

```
datos.s<-scale(m5[,1:2000], center = TRUE, scale = TRUE) # Datos escalados
```

Tras realizar el paso anterior, se tuvo que decidir cuál sería el método óptimo a emplear. Para ello, se comparó el coeficiente resultante tras calcular el agrupamiento jerárquico aglomerativo, mediante la función agnes () (de la librería cluster [16]), empleando los métodos "average", "single", "complete" y "ward".

```
library(cluster)
library(knitr)
library(gtools)


# Comparamos los métodos


# vector con los métodos a comparar
m <- c( "average", "single", "complete", "ward")
names(m) <- c( "average", "single", "complete", "ward")


# función para calcular el coeficiente
ac <- function(x) {
 agnes(datos.s, method = x)$ac
                    }

map_dbl(m, ac)
```

```
        average    single  complete      ward
.. 0.4614845 0.3613321 0.5657465 0.9330632
```

Ilustración 9. Coeficientes de análisis aglomerativo con agnes() empleando los métodos "average", "single", "complete" y "Ward".

Tras obtener el coeficiente de 93.31% empleando el método de Ward, decidimos quedarnos con este para llevar a cabo el análisis.

Posteriormente, se compararon los resultados obtenidos empleando las distintas distancias sobre los datos escalados, empleando la función dist() del paquete cluster.

```
# Comparar distancias
distancias1<-dist(datos.s,method="manhattan")
cluster1<-hclust(distancias1,method = "ward.D2")
distancias2<-dist(datos.s,method="euclidean")
cluster2<-hclust(distancias2,method = "ward.D2")
distancias3<-dist(datos.s,method="maximum")
cluster3<-hclust(distancias3,method = "ward.D2")
distancias4<-dist(datos.s,method="canberra")
cluster4<-hclust(distancias4,method = "ward.D2")
distancias4<-dist(datos.s,method="canberra")
cluster4<-hclust(distancias4,method = "ward.D2")
#Realizamos la comparativa gráfica
op <- par(mfcol = c(2, 2)) #Nos permite presentar
par(las =1) #el gráfico en 4 partes
plot(cluster1,cex = 0.4, hang = -1,main="Método Manhatan")
plot(cluster2,cex = 0.4, hang = -1,main="Distancia euclídea")
plot(cluster3,cex = 0.4, hang = -1,main="Distancia por máximos")
plot(cluster4,cex = 0.4, hang = -1,main="Método Camberra")
```

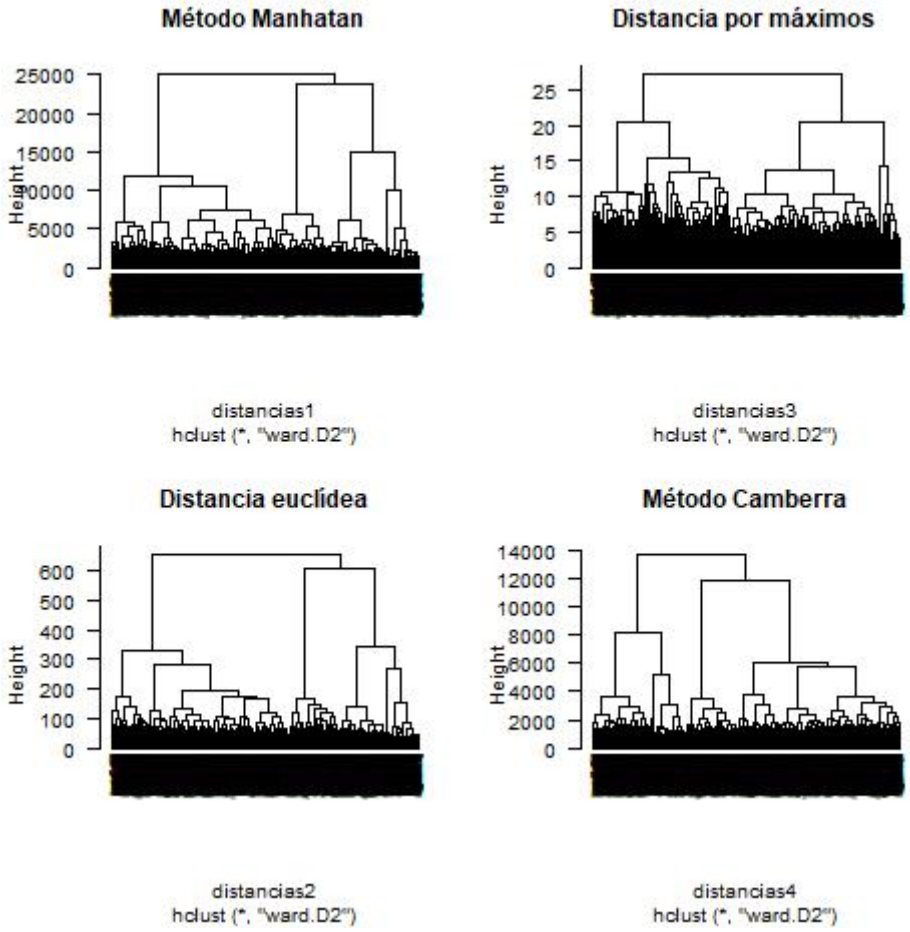


Ilustración 10. Comparación mediante cuatro dendogramas empleando clustering jerárquico aglomerativo con método Manhattan, distancia por máximos, distancia euclídea y método Camberra, usando en todos los casos la distancia de Ward en los datos escalados de expresión génica de cáncer de mama.

Basándose en los resultados obtenidos en la *Figura 2* , se decidió elegir la distancia euclídea para el análisis (*cluster2*).

Mediante la función heatmap.2() del paquete gplots [17], se obtuvo el mapa de calor con la distancia euclídea y el método de Ward.

```
heatmap.2(x = datos.s, scale = "none", col = bluered(256),
        distfun = function(x){dist(x, method = "euclidean")},
        hclustfun = function(x){hclust(x, method =
"ward.D2")},RowSideColors=heat.colors (nrow ( datos.s )),main="Heatmap con
distancia euclídea y método de Ward",
        density.info = "none",
        trace = "none", cexRow = 0.7)
```

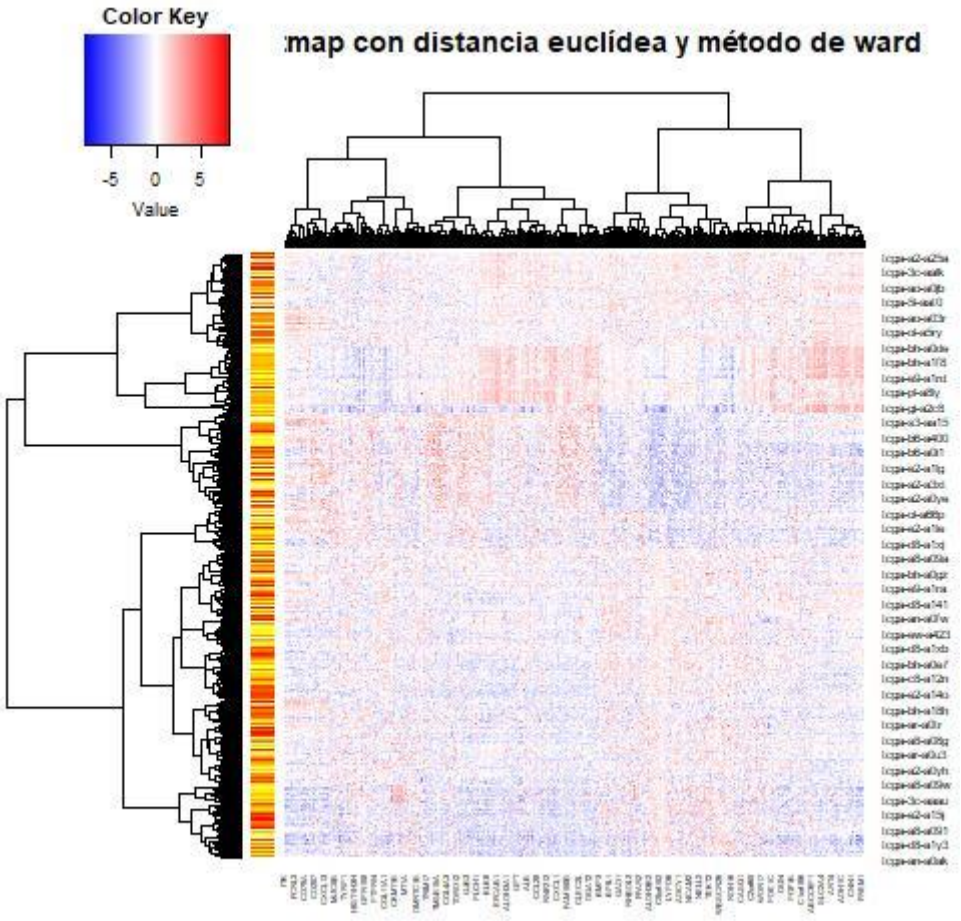


Ilustración 11. Heatmap con distancia euclídea y método de Ward de los datos escalados de expresión génica de cáncer de mama. Pacientes en el eje x, y los genes en el eje y.

Posteriormente se procedió a hacer un estudio algo más exhaustivo sobre estos resultados.

cluster2 # Clustering jerárquico con los datos escalados, distancia euclídea y método de Ward

```
Call:
hclust(d = distancias2, method = "ward.D2")

Cluster method   : ward.D2
Distance         : euclidean
Number of objects: 1081
```

Ilustración 12. Clustering jerárquico con los datos escalados, distancia euclídea y método de Ward.

head(cluster2$merge,20) # Primeras 20 uniones clustering jerárquico con distancia euclídea y método de Ward

```
      [,1] [,2]
 [1,] -644 -904
 [2,] -513 -523
 [3,] -886 -900
 [4,] -557 -558
 [5,] -609 -637
 [6,] -552 -646
 [7,] -543 -553
 [8,] -634 -979
 [9,] -896    3
[10,] -542 -605
[11,] -255 -607
[12,] -615 -624
[13,] -251 -628
[14,] -610 -618
[15,] -885 -902
[16,] -575    2
[17,] -625 -843
[18,] -538    7
[19,] -520    5
[20,] -548 -632
```

Ilustración 13. Primeras 20 uniones clustering jerárquico con distancia euclídea y método de Ward

head(cluster2$height,20) # Primeras 20 alturas clustering jerárquico con distancia euclídea y método de Ward

```
 [1] 22.38910 22.42533 22.44379 22.92471 23.10793 23.26919 23.43806 23.45660 23.56496 23.79992 23.80971
[12] 24.06736 24.19762 24.57614 24.58033 24.76766 25.09546 25.14266 25.27179 25.42661
```

Ilustración 14. Primeras 20 alturas clustering jerárquico con distancia euclídea y método de Ward

Según la tabla anterior, vemos, por ejemplo, que en el paso 1, se unen los cluster formados por una sola observación 644 y 904. Los signos negativos indican que cuando se unen están formados por una sola observación cada uno. En el paso 2, se unen los cluster 513 y 523. Y así, observamos agrupaciones individuales hasta llegar a la novena unión, donde vemos que se une el cluster 896 (formado por una sola observación) con el cluster que se formó en el paso 3 (es decir, al formado por 886 y 900).

Con la función cutree() se pudieron identificar grupos. En este caso, se cortó en cinco grupos para posteriormente compararlos con los 5 subtipos moleculares.

clust <- cutree(cluster2, k = 5)
table(clust)

```
clust
  1   2   3   4   5
125 495 167 177 117
```

Ilustración 15. Cantidad de genes agrupados en cada cluster empleando la distancia euclídea y el método de Ward.

plot(cluster2, cex = 0.6)
rect.hclust(cluster2, k = 5, border = 2:6)

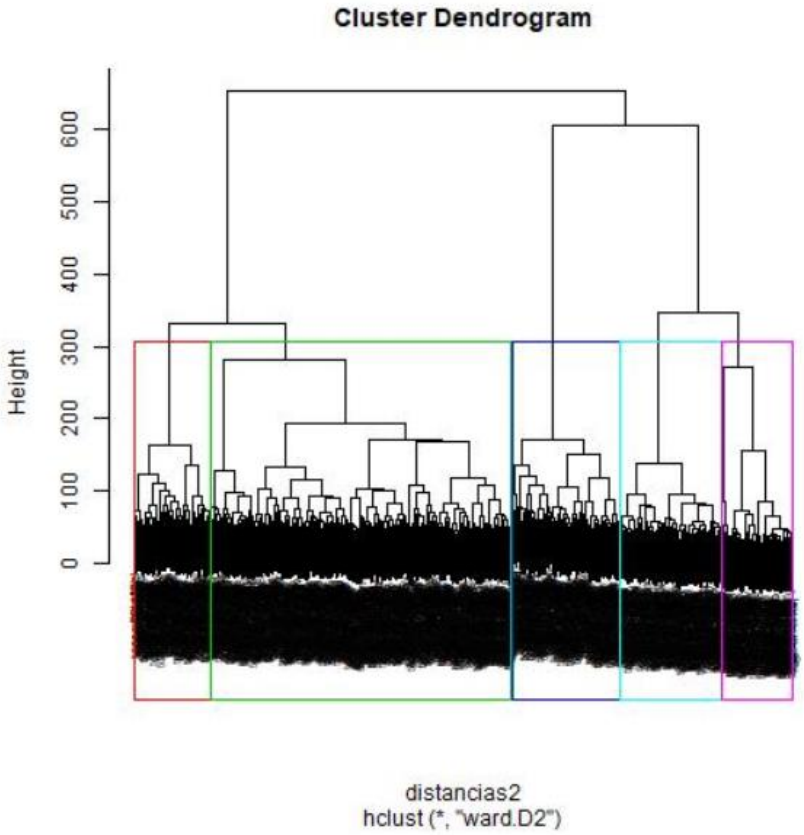


Ilustración 16. Dendograma con distancia euclídea y método de Ward, que agrupa los pacientes separando por 5 colores en los datos escalados de expresión génica de cáncer de mama.

Y estudiamos el número de subtipos moleculares en cada cluster

```
distancia_eu<-dist(datos.s,method="euclidean")
cluster_Ward<-hclust(distancia_eu,method = "ward.D2")
# Cut tree into 5 groups
sub_grp <- cutree(cluster_Ward, k = 5)

# Number of members in each cluster
table(sub_grp)

table(complete=sub_grp,region=m5$BRCA_Subtype_PAM50)
```

```
sub_grp
  1   2   3   4   5
125 495 167 177 117
        region
complete Basal Her2 LumA LumB Normal
       1     0    0   70   55      0
       2     5   72  284  129      5
       3     1    1  143    2     20
       4   168    0    0    2      7
       5    16    9   63   21      8
```

Ilustración 17. número pacientes con cada subtipo molecular ubicados en cada cluster

Vemos, por ejemplo, que en el cluster 1 se forma por Luminales A y Luminales B y el cluster 4 únicamente por Basales y Luminales B.

Para reforzar el análisis se hizo una agrupación por genes utilizando el paquete multiClust [9], empleando la función cluster_analysis se obtuvo:

- Un archivo CSV que contiene los nombres de muestra y sus respectivos clusteres.
- Archivo pdf del dendrograma de muestra, así como archivos atr, gtr y cdt para visualizar en Java TreeView. En el mapa de calor Java TreeView, las muestras se agrupan por el método indicado por el argumento "linkage_type", mientras que los genes se agrupan por el método indicado por el argumento "gene_distance".

```
library(multiClust)
hclust_analysis <- cluster_analysis(sel.exp=datos.s2,
   cluster_type="HClust",
   distance="euclidean", linkage_type="ward.D2",
   gene_distance="correlation",
   num_clusters=5, data_name="Breast Cancer",
   probe_rank="SD_Rank", probe_num_selection="Fixed_Probe_Num",
   cluster_num_selection="Fixed_Clust_Num")
head(hclust_analysis,20)
```

```
  A2ML1  ABCA10  ABCA12  ABCA13   ABCA4   ABCA6   ABCA8   ABCA9   ABCB5  ABCC11   ABCC2   ABCC6 ABCC6P1
      5       3       2       5       2       3       3       3       3       1       5       3       3
 ABCC8   ABCD2 ABHD12B  ABI3BP     ABO    ABP1   ACADL
     1       3       1       3       3       4       3
```

Ilustración 18. Resumen de la agrupación en cluster de los 20 primeros genes mediante la función cluster_analysis() de la librería multiClust.

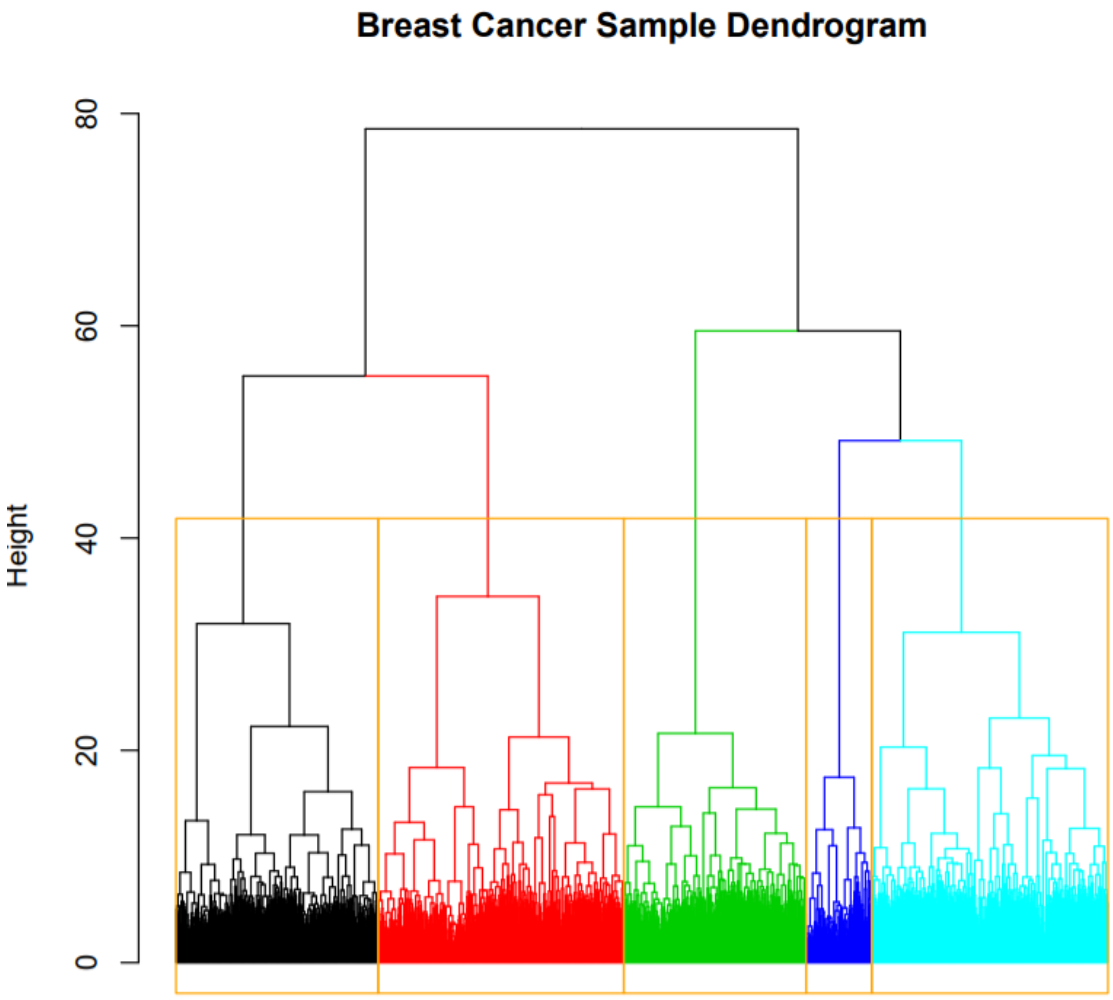


Ilustración 19. Dendograma obtenido mediante la función cluster_analysis() con distancia euclídea y método de Ward, separado por 5 colores en los datos escalados de expresión génica de cáncer de mama.

3.7 Análisis mediante el método de K-means del conjunto de datos de expresión génica sobre el cáncer de mama

Para realizar el anàlisis por k-means en R recurrimos a la función kmeans() de la librería stat.

```
library(stat)
set.seed(123) # Semilla
k5 <- kmeans(datos.s, centers = 5, nstart = 25)

k5$cluster # pacientes con el cluster al que se han asignado
```

Vemos algunos de ellos:

tcga-3c-aaau	tcga-3c-aali	tcga-3c-aalj	tcga-3c-aalk	tcga-4h-aaak	tcga-5l-aat0	tcga-5l-aat1	tcga-5t-a9qa
4	3	3	2	2	2	2	4
tcga-a1-a0sb	tcga-a1-a0sd	tcga-a1-a0se	tcga-a1-a0sf	tcga-a1-a0sg	tcga-a1-a0sh	tcga-a1-a0si	tcga-a1-a0sj
5	2	2	2	2	2	3	2
tcga-a1-a0sk	tcga-a1-a0sn	tcga-a1-a0so	tcga-a1-a0sp	tcga-a1-a0sq	tcga-a2-a04n	tcga-a2-a04p	tcga-a2-a04q
5	3	5	5	4	2	5	5
tcga-a2-a04r	tcga-a2-a04t	tcga-a2-a04u	tcga-a2-a04v	tcga-a2-a04w	tcga-a2-a04x	tcga-a2-a04y	tcga-a2-a0ck
4	5	5	4	3	3	3	2
tcga-a2-a0cl	tcga-a2-a0cm	tcga-a2-a0co	tcga-a2-a0cp	tcga-a2-a0cq	tcga-a2-a0cr	tcga-a2-a0cs	tcga-a2-a0ct
3	5	2	2	4	2	4	4
tcga-a2-a0cu	tcga-a2-a0cv	tcga-a2-a0cw	tcga-a2-a0cx	tcga-a2-a0cy	tcga-a2-a0cz	tcga-a2-a0d0	tcga-a2-a0d1
2	2	3	3	3	2	5	3
tcga-a2-a0d2	tcga-a2-a0d3	tcga-a2-a0d4	tcga-a2-a0em	tcga-a2-a0en	tcga-a2-a0eo	tcga-a2-a0ep	tcga-a2-a0eq
5	4	4	4	2	2	2	3
tcga-a2-a0er	tcga-a2-a0es	tcga-a2-a0et	tcga-a2-a0eu	tcga-a2-a0ev	tcga-a2-a0ew	tcga-a2-a0ex	tcga-a2-a0ey
4	1	4	4	2	2	2	3
tcga-a2-a0st	tcga-a2-a0su	tcga-a2-a0sv	tcga-a2-a0sw	tcga-a2-a0sx	tcga-a2-a0sy	tcga-a2-a0t0	tcga-a2-a0t1
5	2	3	3	5	2	5	3
tcga-a2-a0t2	tcga-a2-a0t3	tcga-a2-a0t4	tcga-a2-a0t5	tcga-a2-a0t6	tcga-a2-a0t7	tcga-a2-a0yc	tcga-a2-a0yd
5	2	2	2	2	2	4	2
tcga-a2-a0ye	tcga-a2-a0yf	tcga-a2-a0yg	tcga-a2-a0yh	tcga-a2-a0yi	tcga-a2-a0yj	tcga-a2-a0yk	tcga-a2-a0yl
5	4	3	3	2	5	1	2
tcga-a2-a0ym	tcga-a2-a0yt	tcga-a2-a1fv	tcga-a2-a1fw	tcga-a2-a1fx	tcga-a2-a1fz	tcga-a2-a1g0	tcga-a2-a1g1
5	3	2	4	4	2	2	5
tcga-a2-a1g4	tcga-a2-a1g6	tcga-a2-a259	tcga-a2-a25a	tcga-a2-a25b	tcga-a2-a25c	tcga-a2-a25d	tcga-a2-a25e
4	1	2	2	4	2	3	4
tcga-a2-a25f	tcga-a2-a3kc	tcga-a2-a3kd	tcga-a2-a3xs	tcga-a2-a3xt	tcga-a2-a3xu	tcga-a2-a3xv	tcga-a2-a3xw
5	2	2	5	5	5	3	2
tcga-a2-a3xx	tcga-a2-a3xy	tcga-a2-a3xz	tcga-a2-a3y0	tcga-a2-a4rw	tcga-a2-a4rx	tcga-a2-a4ry	tcga-a2-a4s0
5	5	3	5	2	5	2	4
tcga-a2-a4s1	tcga-a2-a4s2	tcga-a2-a4s3	tcga-a7-a0cd	tcga-a7-a0ce	tcga-a7-a0cg	tcga-a7-a0ch	tcga-a7-a0cj
5	2	3	4	1	2	1	4
tcga-a7-a0d9	tcga-a7-a0da	tcga-a7-a0db	tcga-a7-a13d	tcga-a7-a13e	tcga-a7-a13f	tcga-a7-a13g	tcga-a7-a13h
1	5	1	5	1	1	1	2
tcga-a7-a26e	tcga-a7-a26f	tcga-a7-a26g	tcga-a7-a26h	tcga-a7-a26i	tcga-a7-a26j	tcga-a7-a2kd	tcga-a7-a3iy
2	3	5	2	5	4	3	2
tcga-a7-a3iz	tcga-a7-a3j0	tcga-a7-a3j1	tcga-a7-a3rf	tcga-a7-a425	tcga-a7-a426	tcga-a7-a4sa	tcga-a7-a4sb
4	4	2	4	2	2	2	4
tcga-a7-a4sc	tcga-a7-a4sd	tcga-a7-a4se	tcga-a7-a4sf	tcga-a7-a56d	tcga-a7-a5zv	tcga-a7-a5zw	tcga-a7-a5zx
2	5	5	3	3	5	2	2
tcga-a7-a6vv	tcga-a7-a6vw	tcga-a7-a6vx	tcga-a7-a6vy	tcga-a8-a06n	tcga-a8-a06o	tcga-a8-a06p	tcga-a8-a06q
5	5	3	5	4	4	2	4
tcga-a8-a06r	tcga-a8-a06t	tcga-a8-a06u	tcga-a8-a06x	tcga-a8-a06y	tcga-a8-a06z	tcga-a8-a075	tcga-a8-a076
3	4	3	4	4	4	3	3
tcga-a8-a079	tcga-a8-a07b	tcga-a8-a07c	tcga-a8-a07e	tcga-a8-a07f	tcga-a8-a07g	tcga-a8-a07i	tcga-a8-a07j
4	3	5	2	3	2	3	2
tcga-a8-a07l	tcga-a8-a07o	tcga-a8-a07p	tcga-a8-a07r	tcga-a8-a07s	tcga-a8-a07u	tcga-a8-a07w	tcga-a8-a07z
4	5	4	5	4	5	4	4
tcga-a8-a081	tcga-a8-a082	tcga-a8-a083	tcga-a8-a084	tcga-a8-a086	tcga-a8-a08a	tcga-a8-a08b	tcga-a8-a08c
3	4	4	3	4	4	3	4
tcga-a8-a08f	tcga-a8-a08g	tcga-a8-a08h	tcga-a8-a08i	tcga-a8-a08j	tcga-a8-a08l	tcga-a8-a08o	tcga-a8-a08p
4	4	2	3	3	3	2	3
tcga-a8-a08r	tcga-a8-a08s	tcga-a8-a08t	tcga-a8-a08x	tcga-a8-a08z	tcga-a8-a090	tcga-a8-a091	tcga-a8-a092
5	4	4	3	3	4	4	3
tcga-a8-a093	tcga-a8-a094	tcga-a8-a095	tcga-a8-a096	tcga-a8-a097	tcga-a8-a099	tcga-a8-a09a	tcga-a8-a09b
4	3	2	4	2	4	2	4
tcga-a8-a09c	tcga-a8-a09d	tcga-a8-a09e	tcga-a8-a09g	tcga-a8-a09i	tcga-a8-a09k	tcga-a8-a09m	tcga-a8-a09n
4	3	4	3	3	4	4	4
tcga-a8-a09q	tcga-a8-a09r	tcga-a8-a09t	tcga-a8-a09v	tcga-a8-a09w	tcga-a8-a09x	tcga-a8-a09z	tcga-a8-a0a1
3	4	4	4	3	3	4	2
tcga-a8-a0a2	tcga-a8-a0a4	tcga-a8-a0a6	tcga-a8-a0a7	tcga-a8-a0a9	tcga-a8-a0ab	tcga-a8-a0ad	tcga-ac-a23c
4	2	2	3	4	4	4	2
tcga-ac-a23e	tcga-ac-a23g	tcga-ac-a23h	tcga-ac-a2b8	tcga-ac-a2bk	tcga-ac-a2bm	tcga-ac-a2fb	tcga-ac-a2fe
4	2	1	2	5	4	1	2
tcga-ac-a2ff	tcga-ac-a2fg	tcga-ac-a2fk	tcga-ac-a2fm	tcga-ac-a2fo	tcga-ac-a2qh	tcga-ac-a2qi	tcga-ac-a2qj
1	2	1	1	2	5	2	5
tcga-ac-a3bb	tcga-ac-a3eh	tcga-ac-a3hn	tcga-ac-a3od	tcga-ac-a3qp	tcga-ac-a3qq	tcga-ac-a3tm	tcga-ac-a3tn
2	4	2	2	2	2	2	4

Ilustración 20. Algunos pacientes con el cluster al que se han asignado mediante el método k-mean con k=5.

k5$size #Tamaño de los 5 centros

```
[1] 324 273 176 123 185
```

Aquí podemos apreciar la diferencia con el tamaño de los conglomerados al hacer análisis jerárquico.

k5$withinss# Suma de cuadrados intra-cluster individual

```
[1] 380371.8 481192.5 322638.3 122276.6 302026.6
```

k5$tot.withinss# Suma de cuadrados intra-cluster total

```
[1] 1608506
```

Para representar el resultado tras aplicar el análisis de k-means se debe recurrir a la aplicación de un algoritmo para reducir la dimensionalidad, como, por ejemplo, el Análisis de Componentes Principales. Se usó la función PCA() de la librería FactoMiner [22] junto con la librería factoextra [20] para su representación.

```
library(FactoMineR)
pca_1<- prcomp(datos.s, scale = TRUE)
dim(pca_1$rotation) # Vemos el número de componentes

# Cálculo componentes principals con la función PCA()
pca <- PCA(X = datos.s, scale.unit = TRUE, ncp = 1081, graph = FALSE)

library(factoextra)
fviz_pca_ind(pca, geom.ind = "point",
          col.ind = as.factor(k5$cluster),
            axes = c(1, 2),
            pointsize = 1.5)
```

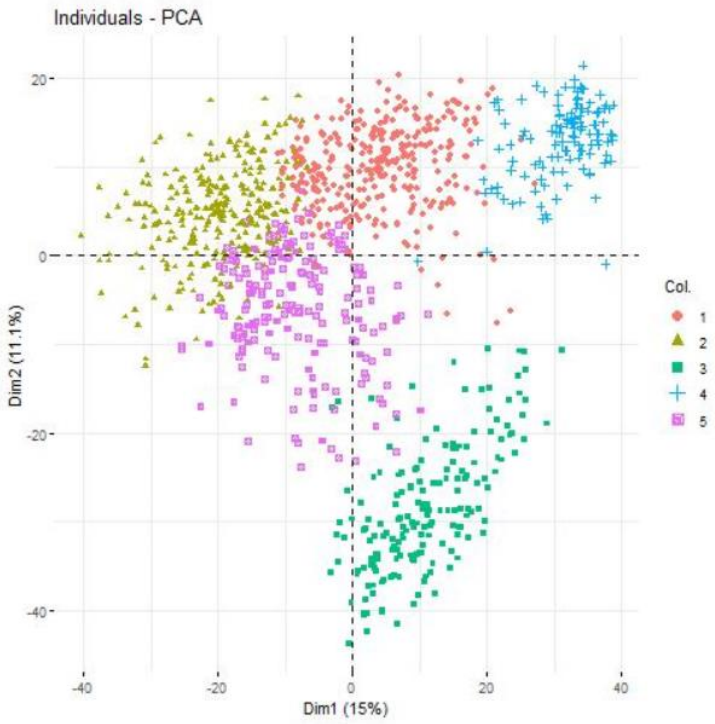


Ilustración 21. Representación de la agrupación de pacientes por k-means con k=5 de los datos de expresión génica de cáncer de mama sobre las dos primeras componentes principales.

Volvemos a estudiar el agrupamiento por genes con la librería multiClust [9]:

```
# Función de análisis de cluster k-means
 kmeans_analysis <- cluster_analysis(sel.exp=datos.s2,
   cluster_type="Kmeans",
   distance=NULL, linkage_type=NULL,
   gene_distance=NULL, num_clusters=5,
   data_name="Breast Cancer", probe_rank="SD_Rank",
   probe_num_selection="Fixed_Probe_Num",
   cluster_num_selection="Fixed_Clust_Num")
head(kmeans_analysis,20)
```

```
[1] "A CSV file has been produced containing your sample and cluster assignment information"
    CPB1  SCGB2A2    GSTM1  SCGB1D2     TFF1    MUCL1      PIP   ADIPOQ    ADH1B   HMGCS2 SERPINA6    PRAME
       4        4        4        4        4        3        4        3        3        4        4        1
ANKRD30A    KCNJ3  UGT2B11      TAT     MUC6   CYP4Z1 CYP2B7P1   MYBPC1
       4        4        4        3        4        4        4        5
```

Ilustración 22. Veinte primeros genes con el número de cluster en el que se han ubicado mediante k-means con la función cluster_analysis() del paquete multiClust.

El resultado es únicamente un objeto "kmeans_analysis" que contiene el vector de nombres de muestra y número de cluster.

Con la función avg_probe_exp() del paquete multiClust[9] se pudo determinar la expresión promedio de cada gen para las muestras en un grupo particular

```
# Call the avg_probe_exp function
avg_matrix <- avg_probe_exp(sel.exp=datos.s2,
   samp_cluster=kmeans_analysis,
   data_name="Breast Cancer", cluster_type="Kmeans", distance=NULL,
   linkage_type=NULL, probe_rank="SD_Rank",
   probe_num_selection="Fixed_Probe_Num",
   cluster_num_selection="Fixed_Clust_Num")
    head(avg_matrix,20)
```

```
[1] "Your matrix containing the average gene probe expression in each cluster is finished"
                Cluster 1   Cluster 2    Cluster 3   Cluster 4   Cluster 5
tcga-3c-aaau  0.226191231 -0.56238895 -0.452313257  0.29987908 -0.53235127
tcga-3c-aali  0.576146916  0.67309839 -0.287036424 -0.09101038 -0.20447203
tcga-3c-aalj  0.248051746  0.49286115 -0.219453238  0.00791887 -0.63314846
tcga-3c-aalk  0.164630041  0.19748892  0.204654781  0.49132069  0.06828574
tcga-4h-aaak -0.049574090 -0.18909252  0.032593312  0.34966502  0.06819671
tcga-5l-aat0 -0.003377484  0.46226125  0.108346018  0.47216070  0.15205927
tcga-5l-aat1  0.177930944  0.91849545  0.007635993  0.38594006 -0.03640221
tcga-5t-a9qa  0.030321487 -1.32933463 -1.102600240 -0.15955874 -0.68724133
tcga-a1-a0sb -0.316931732 -1.37299323  0.235575539 -0.59060272  0.95335105
tcga-a1-a0sd  0.169921545  0.03122776  0.423402012  0.36807762  0.15202000
tcga-a1-a0se -0.204415568 -0.35346258  0.269576024  0.46426402  0.21038545
tcga-a1-a0sf -0.056158924  0.48359281  0.312003436  0.54135194  0.06820624
tcga-a1-a0sg  0.138206560  0.16500728 -0.050957370  0.39130438 -0.25042448
tcga-a1-a0sh -0.200015297 -0.65152530  0.073522880  0.07400203  0.01252169
tcga-a1-a0si  0.166575537  0.46378550  0.020255070  0.11510852 -0.01216315
tcga-a1-a0sj  0.096805163 -0.03932344  0.500565931  0.11521969 -0.06914351
tcga-a1-a0sk  0.619925358 -0.90380837 -0.222164610 -0.43833519  0.65112338
tcga-a1-a0sn  0.252942920  0.19724396 -0.243473373  0.17487724 -0.26431837
tcga-a1-a0so -0.154451726 -0.51049721 -0.668021296 -1.06739996  0.13296860
tcga-a1-a0sp  0.360090653  0.64200990 -0.048727344 -0.67617809  0.75092662
```

Ilustración 23. Expresión promedio en cada paciente en cada cluster, obtenida mediante la función avg_probe_exp() del paquete multiClust.

3.8 Comparación de uHcl con K-means en el conjunto de datos de expresión génica de cáncer de mama

 # Asignación de conglomerados de cada observación
clusteres.km <- k5$cluster

 # Comparación de k-means y clustering jerárquico
table(clusteres.km, clust)

	clust				
clusteres.km	1	2	3	4	5
1	0	0	8	0	115
2	3	160	158	1	2
3	5	178	1	1	0
4	117	155	0	1	0
5	0	2	0	174	0

Ilustración 24. Comparación clusters de k-means y clustering jerárquico.

3.9 MGP sobre el conjunto de datos de expresión génica de cáncer de mama

Para aplicar el modelo gráfico probabilístico en la base de datos de expresión génica de cáncer de mama, se hizo uso de la librería “gRapHD” [19].
Se minimizó el BIC y se hizo búsquedad del modelo óptimo.

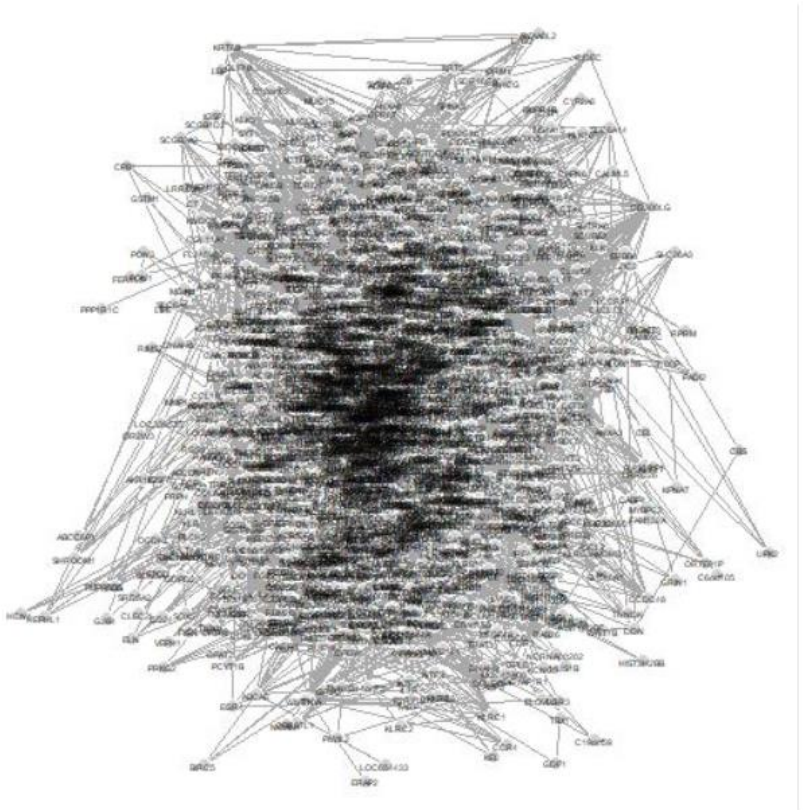


Ilustración 25. Red obtenida sobre la base de datos de expresión génica tras aplicar un modelo gráfico probabilístico.

3.10 Cappas de información sobre el conjunto de datos de expresión génica de cáncer de mama

Para la búsqueda de Cappas de información se hizo uso de la librería ConsensusClusterPlus de Bioconductor. Se cogió como Hilario >0.05 y nos quedamos con los pesos mayores a este valor.
Obteniendo finalmente una clasificación óptima de los pacientes.

	x
"tcga-3c-aaau"	1
"tcga-3c-aali"	1
"tcga-3c-aalj"	1
"tcga-3c-aalk"	1
"tcga-4h-aaak"	1
"tcga-5l-aat0"	1
"tcga-5l-aat1"	1
"tcga-5t-a9qa"	1
"tcga-a1-a0sb"	2
"tcga-a1-a0sd"	1
"tcga-a1-a0se"	1
"tcga-a1-a0sf"	1
"tcga-a1-a0sg"	1
"tcga-a1-a0sh"	1
"tcga-a1-a0si"	1
"tcga-a1-a0sj"	1
"tcga-a1-a0sk"	2
"tcga-a1-a0sn"	1
"tcga-a1-a0so"	2
"tcga-a1-a0sp"	2
"tcga-a1-a0sq"	1
"tcga-a2-a04n"	1
"tcga-a2-a04p"	2
"tcga-a2-a04q"	2
"tcga-a2-a04r"	1
"tcga-a2-a04t"	2
"tcga-a2-a04u"	2
"tcga-a2-a04v"	1
"tcga-a2-a04w"	1
"tcga-a2-a04x"	1
"tcga-a2-a04y"	1
"tcga-a2-a0ck"	1
"tcga-a2-a0cl"	1
"tcga-a2-a0cm"	2
"tcga-a2-a0co"	1
"tcga-a2-a0cp"	1
"tcga-a2-a0cq"	1
"tcga-a2-a0cr"	2
"tcga-a2-a0cs"	1
"tcga-a2-a0ct"	1
"tcga-a2-a0cu"	1
"tcga-a2-a0cv"	2
"tcga-a2-a0cw"	1

Ilustración 26. Clasificación óptima de los primeros pacientes del conjunto de datos de expresión génica tras aplicar cappas de información.

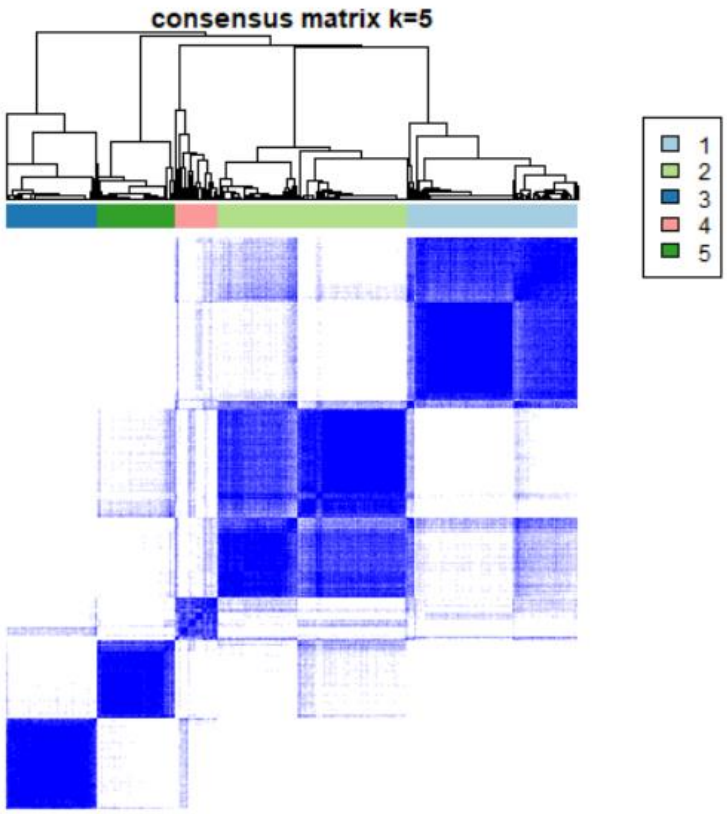


Ilustración 27. Clasificación con k=5 del conjunto de datos de expresión génica tras aplicar cappas de información.

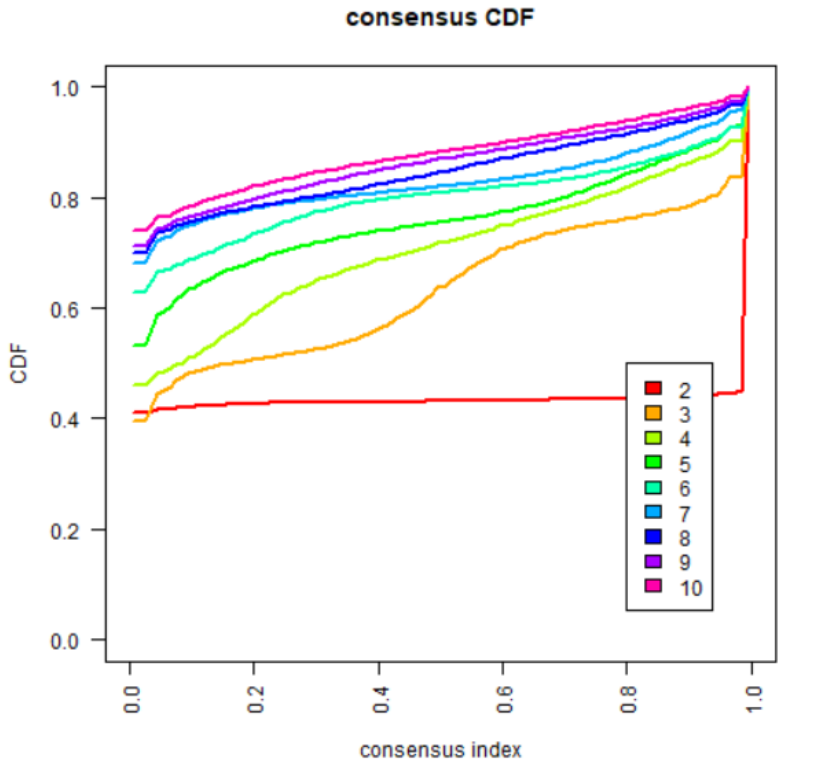


Ilustración 28. Resultado de "consencusCluster" sobre el conjunto de datos de expresión génica..

Ilustración 29. Cappas de información sobre el conjunto de datos de expresión génica..

1.5 Breve sumario de productos obtenidos

2. Datos filtrados y sin valores faltantes
3. Conjunto de datos formados por la expresión génica y, a su vez, por las características clínicas más relevantes.
4. Subtipos moleculares de cáncer de mama en cada paciente
5. Obtención de 5 clusters mediante la aplicación de clustering jerárquico con la distancia euclídea y el método de Ward en la base de datos de expresión génica.
6. Resultados obtenidos tras aplicar el método de K-means en la base de datos de expresión génica.
7. Red obtenida tras aplicar un algoritmo para implementar un modelo gráfico probabilístico sobre la base de datos de expresión génica.
8. Pesos obtenidos tras aplicar un algoritmo sobre las cappas de información de la base de datos de expresión génica.

3. Conclusiones

La principal conclusión del trabajo es que distintas metodologías de agrupamiento ofrecen resultados distintos. Esto es de especial relevancia en estudios enfocados a la respuesta de preguntas clínicas, ya que la metodología de análisis de los datos empleados será crucial a la hora de obtener unos resultados con mayor o menor impacto clínico. En nuestro caso concreto, hemos empleado una clasificación molecular obtenida asignando cada muestra con un algoritmo predefinido en un estudio previo. Al emplear dos metodologías de agrupamiento distintas, hemos obtenido resultados dispares entre ambas, y además no hemos sido capaces de reproducir fielmente la clasificación de los subgrupos.

En cuanto al cumplimiento de los objetivos planteados, se ha logrado una consecución de objetivos bastante grande. Sin embargo, hay dos puntos (la evaluación de las clasificaciones respecto a la información clínica y evaluar si el k-means puede mejorar el clustering jerárquico) no han podido realizarse de forma completa. Hay dos motivos para ello. La información clínica disponible en el TCGA es escueta, no ofrece una imagen completa del estatus clínico de los pacientes, lo que dificulta un análisis correcto del impacto de las clasificaciones respecto a la evolución clínica. Por otro lado, el método de k-means recursivo es una metodología diseñada por el grupo investigador, y la situación de confinamiento durante el estado de alarma ha limitado nuestra capacidad de reunirnos y compartir información, así como el acceso a los equipos informáticos de alto rendimiento de los que dispone el grupo. Esto ha supuesto que estos análisis no se hayan podido realizar de forma adecuada en plazo.

En cuanto a la planificación del trabajo, considero que ha sido adecuada. Se han contemplado multitud de metodologías, seleccionado las más habituales en la bibliografía y he aprendido a emplearlas en un tipo de datos (datos moleculares de alta dimensión) de difícil manejo. También he aprendido nociones básicas del preprocesamiento de este tipo de datos y he tenido contacto con la información clínica que va asociada a las muestras de pacientes. Desde el punto de vista formativo, el trabajo ha sido enriquecedor, y considero que he aprendido mucho. En aspectos a mejorar, habría sido deseable un mayor número de reuniones con mi tutor, que han sido bastante limitadas por la situación que hemos vivido estos meses. Esto ha hecho que no hayamos sido muy eficientes a la hora de planificar y evaluar algunos análisis.

Respecto a las líneas de trabajo futuro, en este trabajo se ha realizado un gran esfuerzo en recoger distintas herramientas que se emplean por el grupo de investigación, basadas en distintos softwares y lenguajes de programación, y establecer un flujo de trabajo completo en R. Esto nos va a permitir aumentar la productividad en proyectos a futuro. Por otro lado, el escenario clínico evaluado (subtipos moleculares de cáncer de mama) es un escenario ya estudiado y establecido. El siguiente reto a plantearse sería evaluar un escenario clínico no resuelto y ver si estas herramientas permiten identificar grupos o entidades moleculares con valor clínico.

4. Glosario

- Aprendizaje no supervisado: Rama del aprendizaje automático que se basa en permitir a los algoritmos trabajar sin incorporar condiciones por parte del humano.

- Cluster: cada una de las divisiones que lleva a cabo un algoritmo de agrupamiento sobre un set de datos.

- Algoritmo: Conjunto de procesos que derivan en el tratamiento de un objeto para modificarlo u obtener información de él.

- Matriz de expresión genética: Matriz de valores, donde cada valor indica el nivel de expresión de un probe en una muestra.

5. Bibliografía

1. Pedret R; Sagnier L; Camp F. 2003. Herramientas para segmentar mercados y posicionar productos. Barcelona, España, Planeta. 329p. Fecha consulta: 05/03/2020

2. Gallardo, J. s.f. Introducción al Análisis Cluster. Universidad de Granada, Granada, España. Disponible en http://www.ugr.es/~gallardo/pdf/cluster-g.pdf. Fecha consulta: 05/03/2020

3. Hu, C. W., Kornblau, S. M., Slater, J. H., & Qutub, A. A. (2015).*Progeny Clustering*: A Method to Identify Biological Phenotypes.*Scientific Reports*,*5*, 12894. http://doi.org/10.1038/srep12894. Fecha consulta: 15/03/2020

4. Autor: Nathan Lawlor (19/01/2018). Título: A Guide to multiClust . Nombre de la página: Bioconductor. Disponible en: http://www.bioconductor.org/packages/3.7/bioc/vignettes/multiClust/inst/d oc/multiClust.html . Fecha consulta: 15/03/2020

5. Procedencia: Rstudio-pubs Título: Heatmaps y aprendizaje no supervisado Disponible en: https://rstudio-pubs-static.s3.amazonaws.com/310338_fc5c392188a14507b6325570c6a5e82 1.html. Fecha de consulta: 20/03/2020.

6. Firebrowse. Repositorio de datos sobre distintos tipos de cáncer de TCGA. Disponible en: http://firebrowse.org/. Fecha de consulta: 18/02/2020.

7. TCGAbiolinks: TCGAbiolinks: un paquete R/ Bioconductor para el análisis integrado con datos GDC. Información disponible en: https://rdrr.io/bioc/TCGAbiolinks/man/TCGAquery_subtype.html. Fecha de consulta: 22/4/2020.

8. TCGAquery_subtype function: recupera subtipos moleculares para un tumor dado. Función del paquete Bioconductor TCGAbiolinks. Información disponible en :

https://rdrr.io/bioc/TCGAbiolinks/man/TCGAquery_subtype.html. Fecha de consulta: 22/4/2020.

9. Autor: Nathan Lawlor (19/01/2018). Título: A Guide to multiClust . Nombre de la página: Bioconductor. Disponible en: Fecha de consulta: 22/4/2020.http://www.bioconductor.org/packages/3.7/bioc/vignettes/multiClust/inst/d oc/multiClust.html . Fecha consulta: 03/03/2018

10. TCGA: The Cancer Genome Atlas Program. National Cancer Institute. Disponible en: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga. Fecha de consulta: 18/2/2020.

11. Nature article of Molecular portraits of human breast tumours. August 2000. Disponible en : https://www.nature.com/articles/35021093/. Fecha de consulta: 20/4/2020.

12. Package tidyverse . The 'tidyverse' is a set of packages that work in harmony because they share common data representations and 'API' design. This package is designed to make it easy to install and load multiple 'tidyverse' packages in a single step. Información disponible en: https://cran.r-project.org/web/packages/tidyverse/index.html. Fecha de consulta: 18/2/2020.

13. Package readr. The goal of 'readr' is to provide a fast and friendly way to read rectangular data (like 'csv', 'tsv', and 'fwf'). It is designed to flexibly parse many types of data found in the wild, while still cleanly failing when data unexpectedly changes. Información disponible en: https://cran.r-project.org/web/packages/readr/index.html. Fecha de consulta: 20/2/2020.

14. Package MCMCGlmm: MCMC Generalised Linear Mixed Models. Información disponible en: https://cran.r-

project.org/web/packages/MCMCglmm/index.html. Fecha de consulta: 20/2/2020.

15. Package dplyr: A grammar of data manipulation. A fast, consistent tool for working with data frame like objects, both in memory and out of memory. Información disponible en: https://cran.r-project.org/web/packages/dplyr/index.html. Fecha de consulta: 21/2/2020.

16. Package cluster: “Finding Groups in Data”: Cluster Analysis Extended Rousseeuw et al. Methods for Cluster analysis. Much extended the original from Peter Rousseeuw, Anja Struyf and Mia Hubert, based on Kaufman and Rousseeuw (1990) "Finding Groups in Data". Información disponible en https://cran.r-project.org/web/packages/cluster/index.html. Fecha de consulta: 21/2/2020.

17. Package gplots: Various R programming tools for plotting data, including: - calculating and plotting locally smoothed summary function as ('bandplot', 'wapply'), - enhanced versions of standard plots ('barplot2', 'boxplot2', 'heatmap.2', 'smartlegend'), - manipulating colors ('col2hex', 'colorpanel', 'redgreen', 'greenred', 'bluered', 'redblue', 'rich.colors'), - calculating and plotting two-dimensional data summaries ('ci2d', 'hist2d'), - enhanced regression diagnostic plots ('lmplot2', 'residplot'), - formula-enabled interface to 'stats::lowess' function ('lowess'), - displaying textual data in plots ('textplot', 'sinkplot'), - plotting a matrix where each cell contains a dot whose size reflects the relative magnitude of the elements ('balloonplot'), - plotting "Venn" diagrams ('venn'), - displaying Open-Office style plots ('ooplot'), - plotting multiple data on same region, with separate axes ('overplot'), - plotting means and confidence intervals ('plotCI', 'plotmeans'), - spacing points in an x-y plot so they don't overlap ('space'). Información disponible en https://cran.r-project.org/web/packages/gplots/index.html. Fecha de consulta: 22/2/2020.

18. Package knitr: A General-Purpose Package for Dynamic Report Generation in R. Información disponible en https://cran.r-project.org/web/packages/knitr/knitr.pdf. Fecha de consulta: 18/2/2020.

19. Package grapHD: The gRapHD package is designed for efficient selection of high-dimensional undirected graphical models. The package provides tools for selecting trees, forests and decomposable models minimizing information criteria such as AIC or BIC, and for displaying the independence graphs of the models. It has also some useful tools for analysing graphical structures. It supports the use of discrete, continuous, or both types of variables. Información disponible en https://www.rdocumentation.org/packages/gRapHD/versions/0.2.5. Fecha de consulta: 18/2/2020.

20. Package factoextra. Provides some easy-to-use functions to extract and visualize the output of multivariate data analyses, including 'PCA' (Principal Component Analysis), 'CA' (Correspondence Analysis), 'MCA' (Multiple Correspondence Analysis), 'FAMD' (Factor Analysis of Mixed Data), 'MFA' (Multiple Factor Analysis) and 'HMFA' (Hierarchical Multiple Factor Analysis) functions from different R packages. It contains also functions for simplifying some clustering analysis steps and provides 'ggplot2' - based elegant data visualization. Información disponible en https://cran.r-project.org/web/packages/factoextra/index.html. Fecha de consulta: 18/2/2020.

21. Package sparcl. Implements the sparse clustering methods of Witten and Tibshirani (2010): "A framework for feature selection in clustering"; published in Journal of the American Statistical Association 105(490): 713-726. Información disponible en https://cran.r-project.org/web/packages/sparcl/index.html. Fecha de consulta: 20/2/2020.

22. Package factoMineR. Exploratory data analysis methods to summarize, visualize and describe datasets. The main principal component methods are available, those with the largest potential in terms of applications: principal component analysis

(PCA) when variables are quantitative, correspondence analysis (CA) and multiple correspondence analysis (MCA) when variables are categorical, Multiple Factor Analysis when variables are structured in groups, etc. and hierarchical cluster analysis. F. Husson, S. Le and J. Pages (2017). Información disponible en https://cran.r-project.org/web/packages/FactoMineR/index.html. Fecha de consulta: 25/2/2020.

23. Package ConsensusClusterPlus of Bioconductor. Algorithm for determining cluster count and membership by stability evidence in unsupervised analysis. Información disponible en https://bioconductor.org/packages/release/bioc/html/ConsensusClusterPlus.html. Fecha de consulta: 16/3/2020.

24. Gámez-Pozo et al. Cancer Research 2015, Trilla-Fuertes et al. Scientific Reports 2019, Trilla-Fuertes et al. BMC Cancer 2019. Disponible en https://scholar.google.es/citations?user=FYbzEeAAAAAJ&hl=es. Fecha de consulta: 20/5/2020.

25. Schwarz Estimating the dimensión of a model. Ann Stat 1978; Lauritzen, Graphical Models, Oxford UK,1996. Disponible en https://scholar.google.es/citations. Fecha de consulta: 20/5/2020.

26. 16-.Kakushadze, Z., & Yu, W. (2017). *K-means and cluster models for cancer signatures. *Biomolecular Detection and Quantification*, *13*, 7–31. http://doi.org/10.1016/j.bdq.2017.07.001. Fecha de consulta: 30/5/2020.

27. Autor: Andrea Trevino. (12/06/2016) Título: Introduction to k-means clustering. Nombre de la página: datascience. Disponible en: https://www.datascience.com/blog/k-means-clustering. Fecha de consulta: 06/06/2019

28. Deconstructing the molecular portraits of breast cancer. Perou et al. 2000; Parker et al. 2009. Disponible en:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5528267/. Fecha de consulta: 30/5/2020.

29. The contribution of gene expresión profiling to breast cancer classification, prognostication and prediction: a restrospective of te last decade. Weigelt et al. 2010. Disponible en: https://onlinelibrary.wiley.com/doi/full/10.1002/path.2648. Fecha de consulta: 30/5/2020.