Elaboración de modelos predictivos con datos multimodales

Ricardo Gonzalo Sanz

Nombre del Consultor y Profesor responsable de la asignatura: Alex Sánchez Pla

Índice

1 FICHA DEL TRABAJO FINAL

Título del trabajo: Elaboración de modelos predictivos con datos multimodales

Nombre del autor: Ricardo Gonzalo Sanz

Nombre del consultor/a: Alex Sánchez Pla

Nombre del PRA: Alex Sánchez Pla

Fecha de entrega: 06/2019

Titulación: Máster universitario de Bioinformática y bioestadística (UOC-UB)

Área del trabajo Final: TFM-Bioinformática y Bioestadística

Idioma del trabajo: Castellano

Palabras Clave: Análisis de datos ómicos, integración datos ómicos, medicina personalizada

Resumen del Trabajo (máximo 250 palabras):

Tradicionalmente el análisis de los datos ómicos se ha realizado de forma individual sacando conclusiones solamente utilizando la ómica analizada. Últimamente se ha visto que la información que nos proporcionan las diferentes ómicas si se utilizan de forma conjunta, aporta mucha más información que analizarlas por separado. El método de integración de diferentes datos ómicos no está establecido. El principal objetivo de este trabajo es el definir un flujo de trabajo de integración de los datos ómicos. Se ha elegido un conjunto de datos que contiene cuatro tipos de datos: análisis de la expresión génica, análisis de la expresión de miRNA, análisis de las poblaciones celulares y una colección extensa de variables clínicas. Inicialmente se ha analizado cada conjunto de datos por separado y se han seleccionado aquellas variables en cada caso que han resultado más significativas para incluirlas en el posterior flujo de trabajo de la integración de todos los datos. Se ha utilizado el método DIABLO implementado en el paquete de R de Bioconductor llamado `mixOmics`. Entre los resultados obtenidos se observa que las variables clínicas seleccionadas son las que mejor separan las dos condiciones experimentales presentes en las muestras, seguidas de los miRNA. También se han identificado un grupo de variables procedentes de las diferentes ómicas estudiadas que están muy correlacionadas entre sí. Se han creado varias redes que relacionan estas variables entre ellas. El modelo creado clasifica bastante bien las muestras del conjunto de datos de test.

Abstract (in English, 250 words or less):

Traditionally, the analysis of the omic data has been carried out individually, drawing conclusions only using the single omics analyzed. Lately, it has been seen that the information provided by the different omics if used together, provides much more information than analyzing them separately. The method of integrating different omic data is not well established. The main objective of this work is to define a workflow for the integration of omic data. A data set that contains four types of data: analysis of gene expression, analysis of miRNA expression, analysis of cell populations and an extensive collection of clinical variables has been used in this work. In a first instance, each data set has been analyzed separately and the most significant variables

in each case have been selected to be included in the subsequent integration workflow. DIABLO method implemented in the Bioconductor R package called `mixOmics` has been used. Among the results obtained, it is observed that the clinical variables selected are the ones that best separate the two experimental conditions present in the samples, followed by the miRNAs. A group of variables has been also identified from the different omics studied that are highly correlated with each other. Several networks have been created that relate these variables to each other. The created model classifies the samples of the test data set quite well.

2 INTRODUCCIÓN

2.1 Contexto y justificación del trabajo

El continuo desarrollo de las tecnologías de alto rendimiento ha permitido que cada vez exista más información biológica disponible para su utilización por la comunidad científica. Esta información es conocida comúnmente como "datos ómicos" y puede comprender por ejemplo datos de transcriptómica, proteómica, metabolómica, etc. Esta información es cada vez más sencilla de conseguir y más precisa, por lo que su utilización genera mucho interés en el ámbito clínico para el diagnóstico preciso y único (medicina personalizada) de los pacientes en los que se han estudiado. Hasta hace unos años la tendencia era estudiar cada uno de estos datos ómicos de forma individual, muchas veces debido a que al tratarse de tecnologías muy caras en sus inicios, dificultaba mucho el poder estudiar más de una. En la actualidad, el precio de estas tecnologías es cada vez más asequible, por lo que cada vez es más habitual que los investigadores hayan utilizado varias aproximaciones en las mismas muestras. Este avance abre un nuevo campo de estudio al permitir la posibilidad de poder estudiarlas de manera conjunta permitiendo así una mayor comprensión, conocimiento y poder de predicción de las patologías estudiadas.

Actualmente el número de datos ómicos analizados en las mismas muestras puede ser muy elevado. Se puede dar el caso, que en unas mismas muestras se hayan analizado la transcriptómica, la expresión de los miRNAs, epigenómica y que también se hayan recogido datos clínicos. La integración de todos estos datos ómicos y su posible utilización en la creación de modelos predictivos para evaluar su posible utilización para la clasificación de pacientes, plantea una serie de dificultades debido a la propia naturaleza de las técnicas utilizadas. Algunos de los problemas que uno se puede encontrar es por ejemplo que los resultados tengan una escala diferente, que se traten de tipos de variables diferentes o que el número de variables a integrar o estudiar conjuntamente pueda ser muy diferente. Otro tipo de problema que se puede contabilizar es el hecho de que cuando se utilizan para crear modelos predictivos, en el modelo final unos datos ómicos tienden a estar sobre representados en detrimento de otros. Por todos estos problemas creemos que es muy interesante el estudio de diferentes métodos de integración de datos ómicos multimodales que permitan un obtener un mejor conocimiento y así poder diseñar una estrategia de análisis conjunta más eficiente.

2.2 Objetivos del trabajo

Originalmente se plantearon los siguientes objetivos generales a cumplir durante la realización de este trabajo: Los objetivos principales que espero alcanzar en este trabajo son los siguientes:

1) Estudio de las diferentes aproximaciones y metodologías existentes para la integración de datos ómicos multimodales
2) Creación de un modelo predictivo a partir de la integración de datos ómicos multimodales

Estos objetivos generales se dividieron en objetivos más específicos:

1) Estudio de las diferentes aproximaciones y metodologías existentes para la integración de datos ómicos multimodales

a) Identificar los diferentes métodos que existen para la integración de datos ómicos
b) Valorar diferentes métodos de pre-procesado de los datos
c) Estudiar las diferentes estrategias encontradas para la integración de datos ómicos y valorar como afectan los diferentes métodos de pre-procesado de los datos en los resultados

2) Creación de un modelo predictivo a partir de la integración de datos ómicos multimodales

a) Con la mejor aproximación analizada en el método anterior crear un modelo predictivo a partir de los datos multimodales
b) Evaluar la eficiencia del modelo predictivo multimodal creado
c) Probar el pipeline con otro conjunto de datos compuesto por diferentes tecnologías ómicas

2.3 Enfoque y método seguido

Para llevar a cabo este trabajo se ha empezado realizando una revisión bibliográfica lo más exhaustiva posible de los diferentes métodos de integración de datos ómicos multimodales que se han descrito en los últimos años. Se han seleccionado dos métodos de integración teniendo en cuenta su relevancia y su manera de abordar el problema, de manera que se han podido probar diferentes metodologías. De forma paralela se ha elegido un conjunto de datos en el que se han analizado tres tecnologías ómicas y un conjunto amplio de variables clínicas. Se ha probado el método de integración con el conjunto de datos y se han valorarado los resultados obtenidos. Finalmente se ha intentado probar este método de integración con un nuevo conjunto de datos.

2.4 Planificación del trabajo

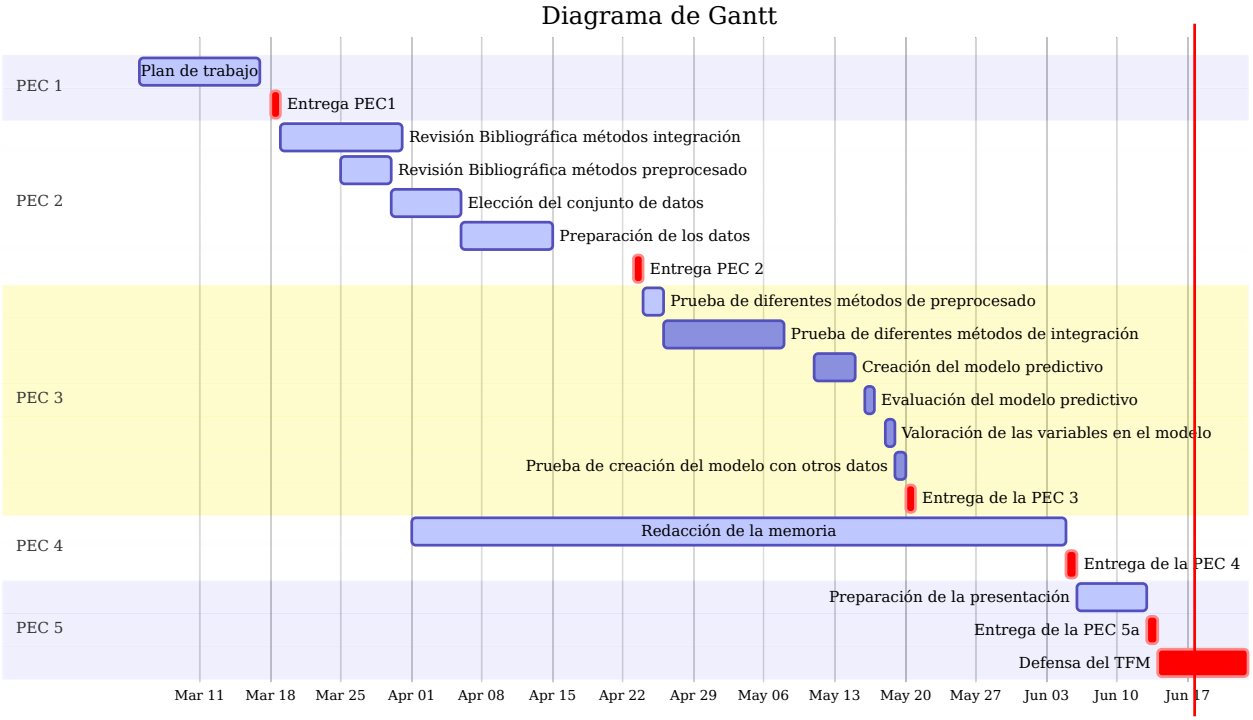
2.4.1 Tareas

En este apartado se van a describir las tareas que se han llevado a cabo para conseguir la realización de los objetivos descritos en el apartado anterior:

- Revisión bibliográfica de los métodos de integración de datos ómicos multimodales
- Elección del conjunto de datos a utilizar
- Preparación de los datos a utilizar
- Creación de un modelo predictivo con los datos resultantes del mejor modelo de integración
- Evaluación mediante parámetros gráficos o numéricos de los resultados del modelo predictivo
- Valoración de los diferentes tipos de variables incluidas en el modelo
- Probar la creación de otro modelo predictivo con otro conjunto de datos.

2.4.2 Calendario

En este apartado se van a temporizar cada una de las tareas que se han descrito en el apartado anterior utilizando un diagrama de Gantt, donde además se han incluido los hitos que se esperan encontrar en este trabajo:



2.4.3 Hitos

En la siguiente tabla se muestran los hitos que se esperan obtener durante la realización de este trabajo, así como la PEC a la que pertenecen y la fecha límite en la que deberían estar realizados:

Hito	**PEC**	**Fecha Límite**
Plan de trabajo	PEC 1	18/03/2019
Encontrar un conjunto de datos adecuado	PEC 2	24/04/2019
Preparación de los datos	PEC 2	24/04/2019
Prueba de los diferentes métodos de integración de datos	PEC 3	20/05/2019
Creación del modelo predictivo	PEC 3	20/05/2019
Entrega de la memoria	PEC 4	05/06/2019
Elaboración de la presentación	PEC 5a	13/06/2019
Defensa pública del trabajo	PEC 5b	26/06/2019

2.5 Breve sumario de los productos obtenidos

Una vez se haya finalizado el proyecto se espera haber obtenido los siguientes resultados que se definen a continuación:

- **Memoria**: El presente documento, donde se explica de una forma amplia todo el trabajo realizado en los meses que ha durado el proyecto. La memoria contiene una introducción y puesta en contexto del proyecto que se va ha llevado a cabo, se exponen los diferentes métodos y estrategias utilizados así como los resultados obtenidos.
- **Producto**: El producto o resultados finales de este trabajo serán varios: por un lado será el modelo predictivo creado con el conjunto de datos utilizado así como los resultados de su evaluación, y por otro lado, será la estrategia o pipeline definido para la integración de futuros datos ómicos multimodales, donde se espera que todos los tipos de datos ómicos estén representados de una manera suficientemente equitativa.
- **Presentación virtual**: Se realizará una presentación virtual donde se explicará todo el desarrollo del proyecto y los resultados obtenidos en un vídeo de una duración inferior a 20 minutos.
- **Autoevaluación del producto**: Se ha realizado una autoevaluación crítica sobre el desarrollo del mismo. Se ha evaluado la consecución de los diferentes objetivos planteados inicialmente, así como si se han cumplido en el tiempo previsto. Se ha evaluado también la calidad del producto obtenido.

2.6 Breve descripción de los otros capítulos de la memoria

En la memoria se encontrará un gran apartado llamado "desarrollo del trabajo", donde se podrá encontrar la introducción al tema que se trata, por un lado a los datos ómicos, a su análisis y finalmente a la integración de los datos. Posteriormente se comentará cómo se han tenido que preparar los diferentes datos ómicos para su posterior utilización en el análisis integrativo. Finalmente se explicará cómo se ha realizado el análisis de integración y se mostrarán los resultados obtenidos. Posteriormente hay un apartado de conclusiones sobre el trabajo realizado y se incluyen las referencias bibliográficas utilizas.

3 DESARROLLO DEL TRABAJO

3.1 Introducción

3.1.1 Introducción a las ciencias ómicas

Ómica es una palabra que se utiliza como sufijo para referirse al estudio de la totalidad o del conjunto de algo. Cuando se añade el sufijo ómica a una palabra siempre será sinónimo de estudio de una gran cantidad de datos.

La ómica es un amplio campo científico que comprende una gran variedad de disciplinas dirigidas al estudio de la abundancia y la caracterización estructural de diferentes familias de moléculas en una extensa colección de organismos [1]. Los estudios ómicos persiguen una variedad de objetivos que dependen en gran medida del campo de la investigación aplicada. Por ejemplo, en el campo clínico, los estudios ómicos pueden ser utilizados para la caracterización de la enfermedad o el seguimiento terapéutico de pacientes [2]. Alternativamente, en el campo ambiental, las ciencias ómicas evalúan las alteraciones que los organismos modelo pueden sufrir después de la exposición a un medio ambiente no habitual [3]. En todos estos casos, el análisis de las diferentes familias de compuestos involucrados es posible debido a la reciente revolución instrumental que ha permitido la medición analítica simultánea de miles de ADNs, ARNm, proteínas o metabolitos. La era de las ciencias ómicas se podría decir que empezó con el descubrimiento de la estructura del ADN por Watson y Crick en el año 1953 [4], una vez ya se sabía que en su estructura se codificaban los genes que eran los transmisores de la información genética. En años posteriores con el desarrollo del Proyecto Genoma Humano (1986), se consiguió secuenciar la totalidad de los 3.200 millones de nucleótidos que componen el material genético humano. No fue por eso hasta el año 2001 cuando se publicó el primer borrador de su secuencia, y hasta el 2004 cuando se publicó la secuencia completa [5]. Actualmente el sufijo ómica se añade inicialmente a muchas moléculas biológicas para así referirse a su estudio más amplio y más profundo (por ejemplo genómica, proteómica, lipidómica, metabolómica, . . .). Posteriormente, este sufijo ya se ha empezado a utilizar con muchos procesos biológicos tales como la fluxómica, la epigenómica, la transcriptómica, etc. A continuación se introducen algunas de las ciencias ómicas más utilizadas hasta el momento.

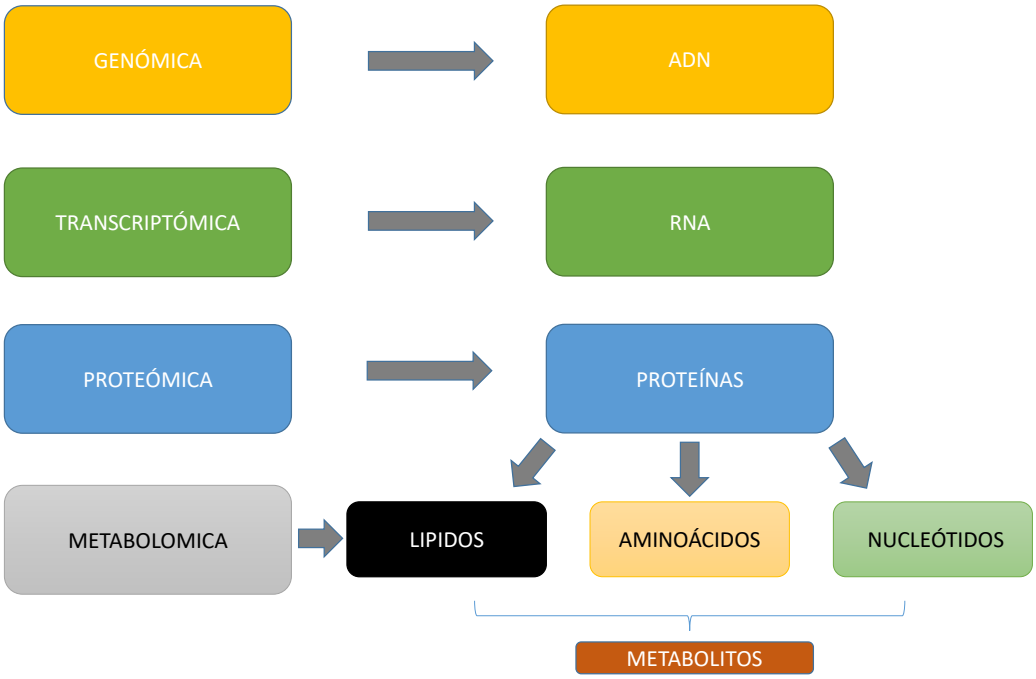


Figura 1: Relación de las moléculas biológicas con la ciencia ómica que las estudia.

Genómica

Tal como se ha comentado en la introducción de este apartado, el gran estudio realizado sobre la el ADN, provocó que la primera vez que se utilizó el sufijo ómica, fue sobre el estudio del genoma. Fue entonces cuando se acuñó el término de **Genómica**. La Genómica se encarga de estudiar la organización, función y la evolución de la información molecular del ADN contenida en el genoma completo. La genómica se puede dividir en dos grandes ramas: una de ellas que se encarga de la caracterización y localización de las secuencias que conforman el ADN, permitiendo así la creación de mapas genéticos de cada organismo. Esta sería la *genómica estructural*. Por otro lado existe la *genómica funcional*, que se encarga de estudiar y recopilar las funciones de los genes. En términos generales, la genómica trata de explicar el origen de un fenotipo determinado a partir de cambios que afecten a la estructura, función o flujos biológicos relacionados con los genes. Actualmente la genómica forma parte integral de las ciencias biomédicas, de tal manera que su uso es casi indispensable en la práctica médica con el establecimiento de mejores métodos de diagnóstico y pronóstico. En la figura 2 se muestran algunas de las aplicaciones que permite el uso de la genómica en la actualidad.

Transcriptómica

La Transcriptómica es el estudio del conjunto de ARN (ARNr, ARNt, ARNm y miARN) que existe en una célula, tejido u órgano. La expresión génica de cada célula, tejido u órgano es muy variable, ya que muestran qué genes se están expresando en un momento dado. Esta expresión génica es lo que se ha llamado transcriptoma. La transcriptómica, consiste en analizar miles de moléculas de ARN de todo tipo. Existen diferentes técnicas que se utilizan para estudiar el transcriptoma. La primera de ellas fueron los microarrays o micromatrices. Esta técnica consiste en fijar unos fragmentos de ADN de pequeña longitud en una superficie sólida (normalmente un cristal), que serán complementarios a diferentes zonas de los genes o miRNA a estudiar. El ARN de las muestras problema es convenientemente procesado (normalmente marcado y troceado en pequeños fragmentos), e hibridado contra la superficie de cristal mencionada anteriormente. En cada fragmento de ADN de la superficie solamente hibridará el ARN específico de la secuencia de ADN fijada. De esta manera a mayor hibridación, se producirá mayor señal que se traduce como mayor expresión del gen o miRNA determinado [6]. En la figura 3 se muestra un esquema de la distribución de los diferentes fragmentos de ADN fijados en la superficie sólida y un ejemplo de la secuencia que podrían tener.

Aunque todavía en uso, la técnica de microarrays está siendo desplazada por nuevas técnicas gracias al desarrollo de la tecnología de ultrasecuenciación. Actualmente, con la tecnología de RNAseq, es posible secuenciar cada molécula de ARN que hay en una determinada muestra, de manera que es posible contar el número de moléculas que hay de cada gen. A más número de moléculas de ARN más expresión del gen originario del ARN.

Proteómica

Proteómica es el estudio a gran escala de las proteínas, en particular de su estructura y función. Las proteínas

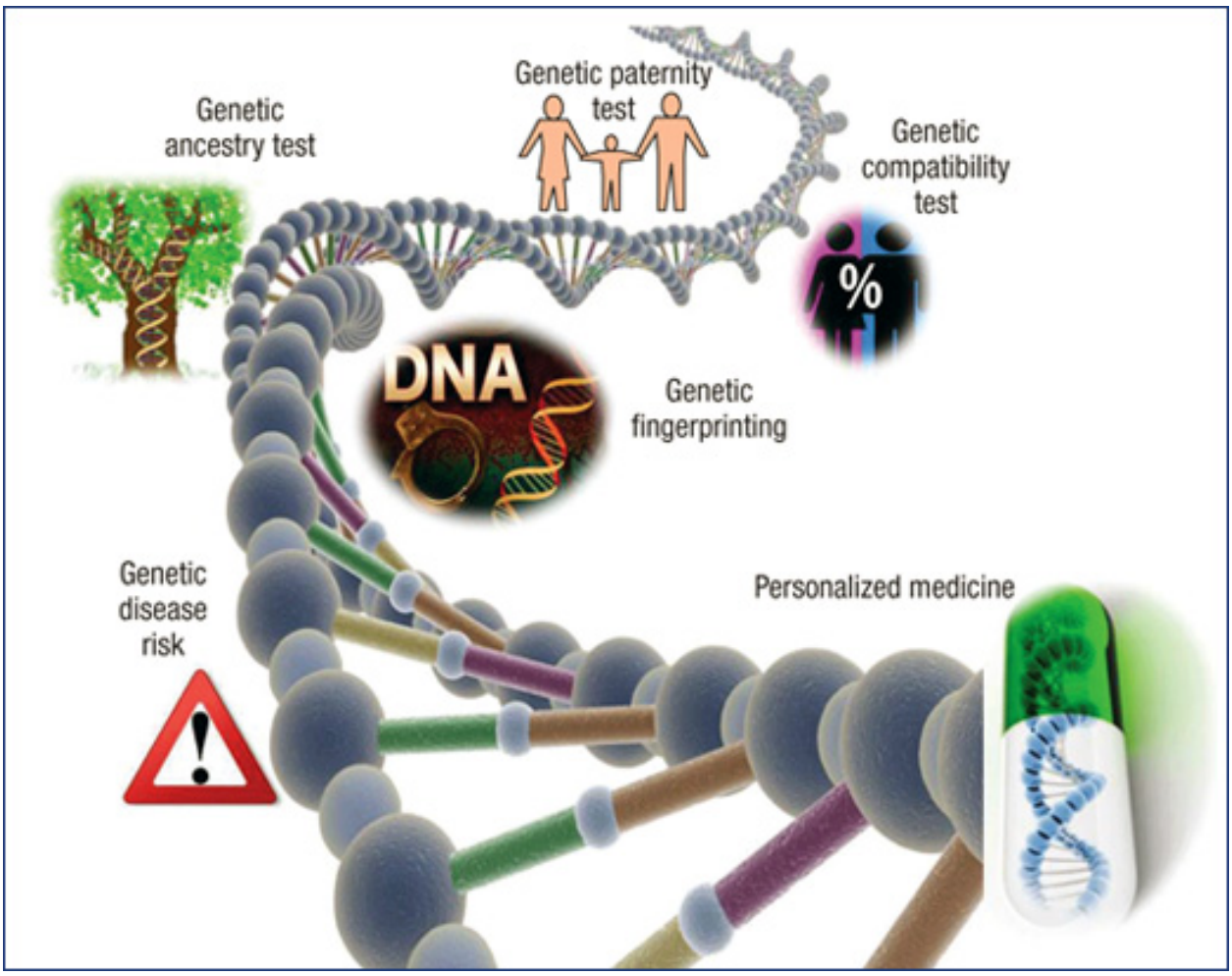


Figura 2: La genómica es una ciencia cada vez más utilizada en los laboratorios clínicos (http://www.esteticamedica.info/noticias/val/669-42/genomica-personalizada-conocer-los-genes-para-prevenir-enfermedades.html).

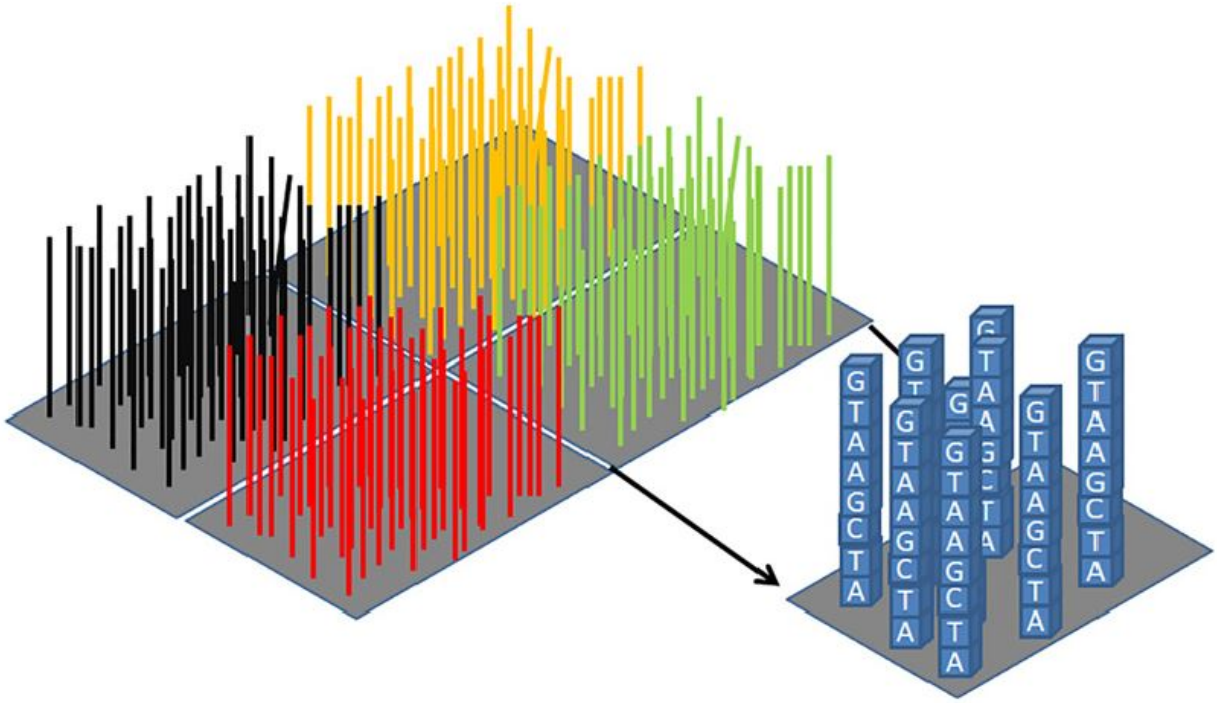


Figura 3: Vista esquemática de un microarray. En azul a la derecha se representan ocho fragmentos de ADN de 8 nucléotidos que son diana de una región específica de un gen. Diferentes partes de los genes son reconocidas por fragmentos similares al representado. En la figura están representados con líneas de colores negro, rojo, amarillo y verde

sufren modificaciones posteriores a su construcción llamadas modificaciones postraduccionales. Esto afecta tanto en la forma como la función de una proteína. Mientras el genoma es prácticamente invariable, el proteoma no sólo difiere de una célula en otra célula, sino que también cambia según las interacciones bioquímicas con el genoma y el ambiente. Este hecho ha provocado que la proteómica sea bastante más difícil de estudiar que la genómica y se tenga la sensación que va un paso por detrás de la genómica. A este hecho se le suma la característica, de que el número de proteínas es mucho mayor que el número de genes, ya que un solo gen puede dar lugar a varias proteínas debido a mecanismos como el splicing alternativo y a modificaciones post-traduccionales. La técnica principal que ha ayudado a la consolidación de la proteómica como ciencia de estudio, es la espectrometría de masas. Esta técnica se ha combinado a técnicas de separación o fraccionamiento tales como los geles 2D-PAGE (eletroforesis de poliacrilamida en dos dimensiones) o la cromatografía líquida de alta resolución (HPLC). En la figura 4 se muestra el flujo de trabajo que se sigue en un experimento de proteómica, desde que se recoge la muestra hasta que se procesa en un espectrómetro de masas.

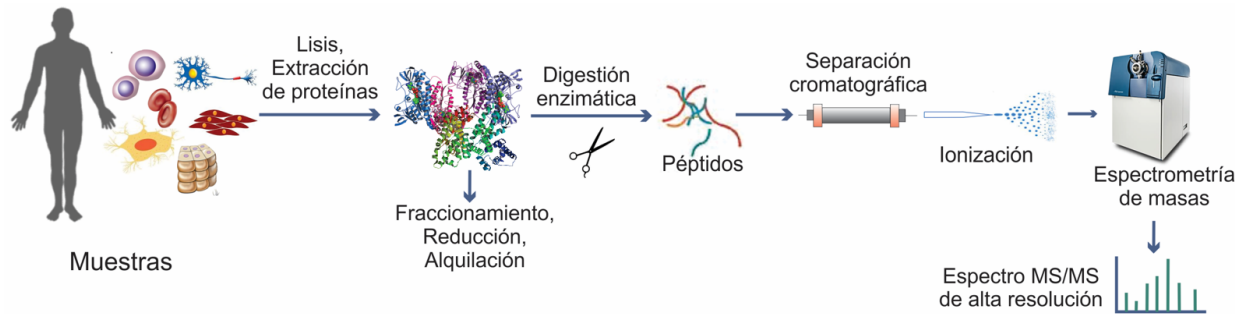


Figura 4: Flujo de trabajo que se sigue un experimento de proteómica (http://rai.unam.mx/pages/lmyp.html)

Metabolómica

La metabolómica se encarga de catalogar y cuantificar los metabolitos que se encuentran en los sistemas biológicos. Los metabolitos son subproductos del metabolismo que se pueden encontrar en diferentes muestras biológicas de fácil acceso, tales como orina, saliva o plasma sanguíneo. Procesos de actividad como la señalización celular, la transferencia de energía y comunicación entre las células, están controlados por los metabolitos. El metaboloma es una colección de todos los metabolitos en una celda en un momento determinado en el tiempo. Este hecho provoca que al igual que pasaba con la proteómica su estudio es muy complicado y muy variable, ya que los metabolitos pueden cambiar mientras se están intentando estudiar. Las dos técnicas que se utilizan principalmente en su estudio son la resonancia magnética nuclear (RMN) o la espectrometría de masas. En la figura 5 se muestra de manera esquemática el funcionamiento de un espectrómetro de masas. En la entrada del espectrometro de masas las muestras son fragmentadas e ionizadas. Una vez dentro del espectrofotómetro, las muestras ionizadas son aceleradas aprovechando su carga electrónica mediante unos imanes que son capaces de cambiar muy rápidamente su polaridad. Finalmente los fragmentos

de las muestras llegan a un detector que es capaz de determinar su masa molecular. A partir de esta masa molecular se deduce de qué compuesto se trata.

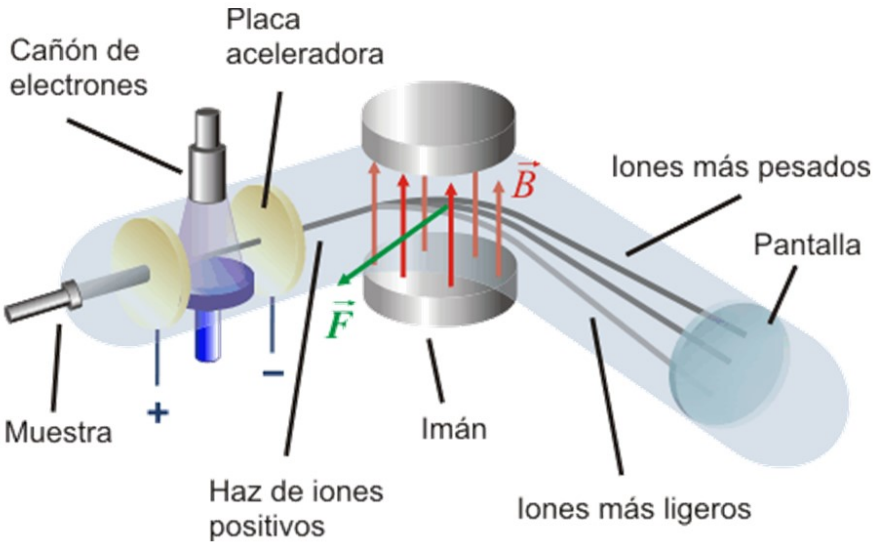


Figura 5: Esquema del funcionamiento de un espectrómetro de masas (http://www2.montes.upm.es/dptos/digfa/cfisica/magnet/espectrometro.html)

Muchas enfermedades genéticas pueden explicarse por el estudio de los cambios en el metaboloma, por lo que el análisis de los metabolitos puede ayudar a diagnosticar enfermedades o estudiar los efectos de una sustancia o de una intoxicación. En la figura 6 [7] se muestra el resultado de un posible análisis dirigido de un experimento de metabolómica.

3.1.2 Introducción al análisis de datos ómicos

En las ciencias ómicas existen dos estrategias principales en los experimentos que se pueden llevar a cabo con ellas, y que afectan a la manera en que se van a analizar los datos generados: los experimentos dirigidos y los no dirigidos [8]. Desde una perspectiva de análisis de datos, la diferencia entre estas dos estrategias es que los datos generados se manejan de una manera diferente. Los estudios dirigidos se centran en un conjunto reducido de componentes químicos conocidos y preseleccionados, mientras que los estudios no dirigidos intentan la investigación del perfil ómico completo con el objetivo de permitiendo una evaluación más completa del sistema biológico considerado. Sin embargo, ambas estrategias se pueden utilizar de forma complementaria. El objetivo del enfoque dirigido proporciona una detección más sensible y precisa de un preseleccionado número de compuestos químicos en un grupo de muestras, mientras que el enfoque no dirigido

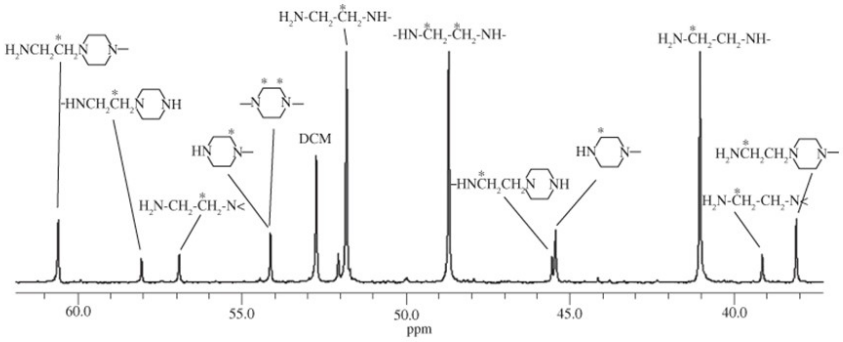


Figura 6: Ejemplo de resultado de un experimento de metabolómica. Cada pico del cromatograma se corresponde con un metabolito

aplicado en las mismas muestras todavía puede permitir la detección e identificación de un conjunto más amplio de genes o compuestos desconocidos. La considerable complejidad de los datos adquiridos en estudios no dirigidos requiere el uso de potentes estrategias de análisis de datos. En la figura 7 se muestra de manera general el flujo de trabajo que se debería seguir en cada experimento que involucre a las ciencias ómicas, tomando especial énfasis en el análisis de datos.

Independientemente de la técnica ómica que se pretenda analizar, se observará siempre una estructura común de los datos: se habrán analizado un gran de variables (genes, lípidos, proteínas, ...) sobre un pequeño número de muestras (en comparación con el gran número de variables que se estudiarán). Cada variable medirá la abundancia de las moléculas estudiadas en cada caso, utilizando distintos procedimientos. Se denominará n al número de muestras utilizadas y N al número de variables analizadas. Lo normal es que N sea mucho mayor que n: $n << N$. En un contexto estadístico clásico se da exactamente la situación contraria, donde N es mucho menor que n. Este hecho provocó el desarrollo y adaptación de métodos estadísticos clásicos para el análisis de los datos ómicos o también llamados de alto rendimiento. Las abundancias ya sean de genes, proteínas, metabolitos o de cualquiera de las ómicas estudiadas se recogerán en una matriz que se denotará de la siguiente manera:

$x = [x_{ij}]_{ij} = 1, \ldots n$

En este caso x_{ij} cuantifica la abundancia de la característica i en la muestra j. El valor de x_{ij} será un

Figura 7: Flujo de trabajo desde las preguntas o hipótesis iniciales hasta la interpretación biológica final

valor (usualmente) positivo. Sobre las columnas se distribuirán las muestras en estudio, y en las filas las características estudiadas. Si se corresponde por ejemplo a un valor procedente de un experimento de microarrays, entonces mide un nivel de fluorescencia y tomará valores positivos. Un valor mayor indicará una mayor expresión del gen en este ejemplo. Si se trabajará con datos de RNAseq, entonces el valor se referirá a conteos, esto es, el el número de lecturas cortas alineadas sobre un gen o sobre un exón o sobre un zona genómica de interés. Más lecturas indicará de nuevo más expresión del gen. Los valores observados en una misma fila (una misma característica sobre todas las muestras), se suele decir que son un perfil (de un modo genérico se le puede llamar perfil de expresión). En la matriz de expresión los valores observados para las distintas muestras son independientes entre ellos, aunque sus cambios reflejan una condición experimental común, por lo que de alguna manera también están relacionados, ya que los genes o metabolitos (o cualquiera que sea la característica que se esté estudiando) actúan de un modo coordinado. Como se verá más adelante, los datos de la matriz de expresión no son directamente comparables, ya que el nivel de ruido es elevado y por ello se han desarrollado técnicas para corregirlo. Son métodos de corrección de fondo y normalización. Cada tipo de análisis ómico puede tener un método de normalización particular, aunque muchas veces el fondo que hay detrás de la metodología estadística es muy similar entre todos ellos. De cada muestra se tiene además información extra que nos informa por ejemplo de si la muestra pertenece al grupo de muestras control o pertenece a un determinado tratamiento. A esta información o variables que describen las muestras se les llamará *metadatos*. A continuación (figura 8) [6] se muestra un esquema típico más detallado de un análisis de datos ómicos. Este flujo de trabajo es muy similar independientemente del tipo de análisis ómico llevado a

cabo.

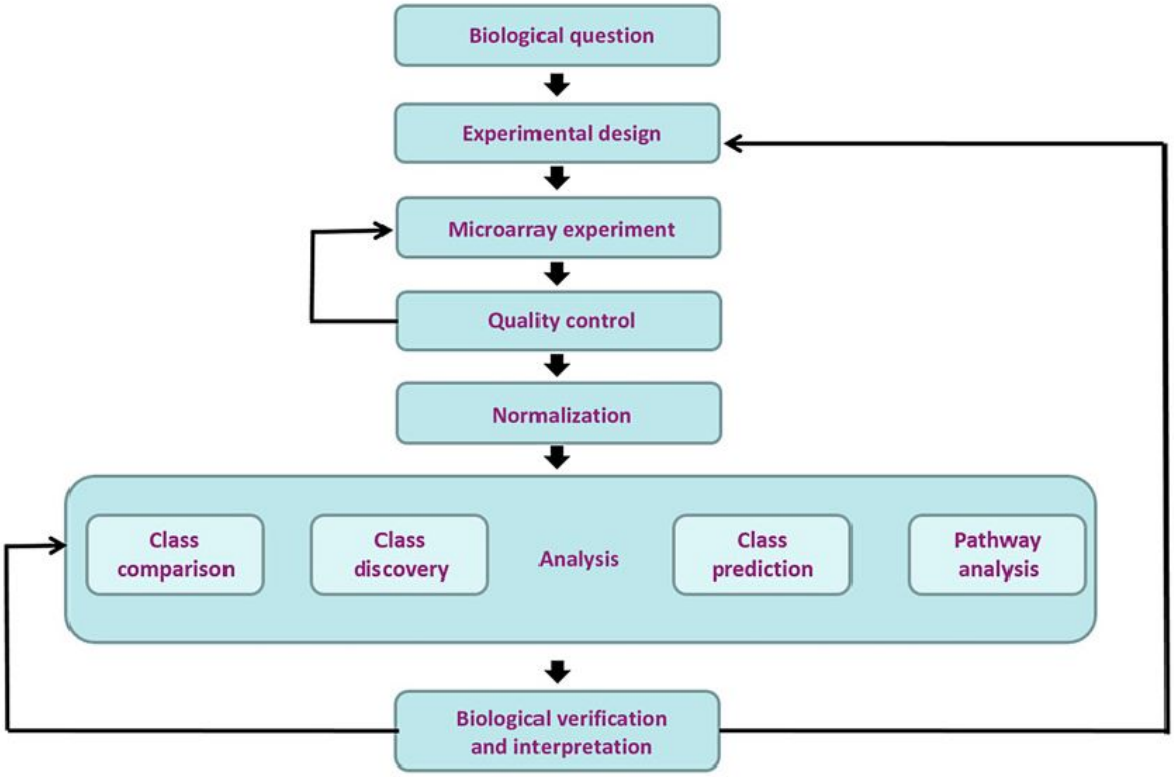


Figura 8: Pasos a seguir en un análisis de datos ómicos

Es muy importante remarcar que en este tipo de análisis toma especial importancia la fase de **diseño experimental**, ya que son experimentos por regla general muy complejos en cuanto a su realización y a su análisis, y no hay que olvidar que también suelen ser muy caros. Debido a esto es muy importante un buen diseño del experimento previo a su realización para que asegurarse que los resultados obtenidos podrán responder en sentido positivo o negativo a la hipótesis planteada en su inicio. Para realizar este diseño experimental es recomendable que esté supervisado por personal con experiencia en este tipo de análisis. Una vez realizado el experimento, la primera fase del análisis es el **control de calidad**. Esta etapa permitirá detectar si ha habido algún problema técnico con alguna muestra y eliminarla del análisis para que no introduzca ruido posteriormente. Este control de calidad sí suele ser característico de cada tipo de ómica estudiada, aunque puede haber técnicas estadísticas que se usan en muchos de ellos, como pueden ser las técnicas de reducción de la dimensión como el análisis de componentes principales. En la figura 9 [6] se muestra el resultado de un análisis de componentes principales sobre un análisis de datos de microarrays.

Mediante esta técnica se pueden representar las fuentes de variabilidad más importantes que hay en los datos. En la figura anterior se observa por ejemplo las muestras se separan a derecha y a izquierda del gráfico (teniendo en cuenta el eje de las X, donde se ha representado la primera componente que aglutina la primera

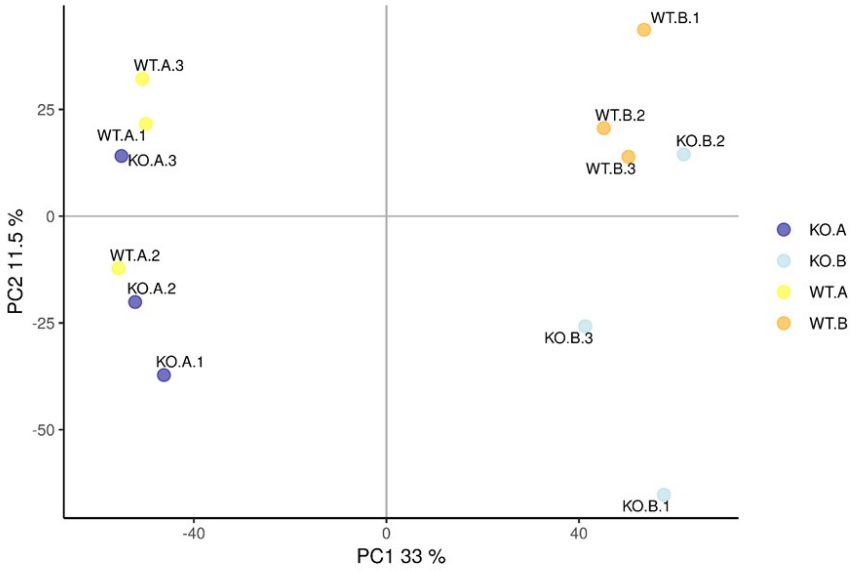


Figura 9: Análisis de componentes principales en un experimento de microarrays

fuente de variabilidad en los datos), según si pertenecen al grupo **A** o **B** (mirar leyenda de la figura). Esto indica que el tratamiento **A** o **B** es más importante que el tratamiento **WT** o **KO** por ejemplo.

Una vez se ha realizado el control de calidad, como se ha comentado antes, es importante eliminar el ruido presente en los datos de forma inherente mediante el proceso llamado **normalización**. Este ruido puede ser debido a pequeñas diferencias en el procesamiento de las muestras, pequeñas diferencias en la captación de la señal por los equipos, etc. Una vez se han normalizado los datos, la distribución de los mismos debería seguir un patrón muy similar entre todas las muestras, tal y como se muestra en la imagen 10 [6]. En el proceso de normalización a parte de eliminar el posible ruido y variabilidad no biológica que exista en las muestras, se escalan los datos de todas las muestras de manera que sean comparables entre ellas.

El siguiente paso es la comparación estadística entre las muestras según a la condición experimental a la que pertenezcan: el ejemplo más sencillo podría ser el comparar las muestras que pertenecen al grupo control con las muestras que han sido tratadas. Este proceso por regla general se suele llamar **análisis de la expresión diferencial**. En este proceso en líneas generales se compara la media de abundancia de cada característica analizada (gen, proteína, metabolito,. . .) de una condición versus la otra o otras condiciones. El resultado principal será una tabla llamada *toptable* donde en las filas se mostrará cada característica estudiada con el resultado del test estadístico (entre otros resultados), que indicará si esa característica se ve afectada por las condiciones experimentales estudiadas. Uno de los últimos pasos en el análisis de datos ómicos es la **interpretación biológica de los resultados**, que hace referencia a poner en el contexto biológico del

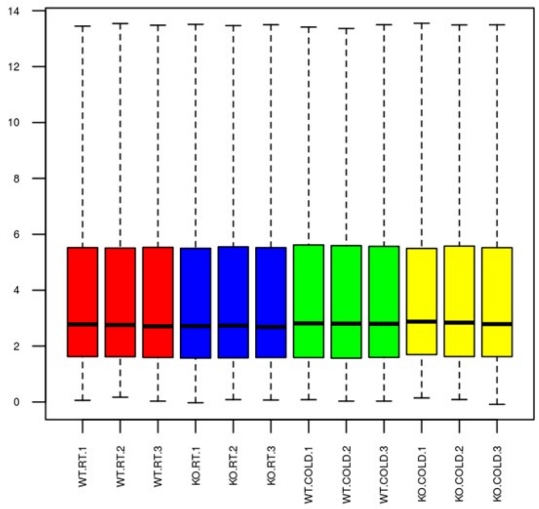


Figura 10: Distribución de los datos una vez normalizados. Se representa mediante un diagrama de cajas

organismo estudiado los resultados obtenidos del análisis. Este a su vez es un análisis complejo que se puede abordar desde diferentes aproximaciones que están fuera del objetivo de este trabajo. En este estudio se van a utilizar datos multimodales, que son aquellos datos que se van a utilizar en mismo análisis pero se han obtenido de manera diferente, muchas veces tienen estructuras diferentes, escalas diferentes o son de tipos diferentes.

3.1.3 Integración de datos ómicos

Los diferentes componentes de un sistema biológico no actúan de manera individual, sino más bien a través de un complejo sistema jerárquico, coordinado, dinámico de interacciones no lineales de un número muy elevado de moléculas biológicas (proteínas interaccionando con ADN, RNA, metabolitos u otras proteínas). Este hecho permite el funcionamiento del sistema biológico en si mismo. El desarrollo de las tecnologías de alto rendimiento ha permitido el disponer de gran cantidad de datos moleculares, que en un principio se estudiaban de forma individualizada obteniendo buenos resultados aunque pronto se vio que la mejor manera de entender el funcionamiento de los sistemas biológicos era intentar trabajar con todos esos datos de manera conjunta. Es lo que se ha llamado **integración de datos ómicos**. La biología de sistemas es el concepto y la aproximación conceptual necesaria para extraer e integrar la información desde esta grandísima cantidad de datos ómicos existente [9]. La aproximación de la biología de sistemas organiza de manera sistemática, integra y racionaliza los diferentes datos ómicos a través de análisis estadísticos, modelización y

visualización asistida por ordenador. Cada vez se está haciendo más evidente que se requiere la integración de los diferentes datos ómicos para un conocimiento más profundo de los complejos sistemas biológicos. El hecho de que cada vez haya más datos ómicos disponibles y que cada vez sean más precisos y más baratos, permite que sean usados como guía para la elección, diseño y seguimiento de aproximaciones terapéuticas, permitiendo que la biología de sistemas ayude a la práctica médica que intenta abordar la complejidad de las enfermedades multifactoriales para así permitir una mejor clasificación e identificación de estas enfermedades y el descubrimiento de nuevas dianas terapéuticas [10]. De esta manera ha nacido el término de **medicina personalizada** que hace referencia a la adaptación del tratamiento médico a las características individuales de cada paciente. Implica que las decisiones referentes al tratamiento o la prevención de enfermedades se tomarán en base a la integración de las características genómicas y moleculares del tumor, la información sobre la situación clínica y los hábitos del paciente. Los datos clínicos por separado tienden a ser muy descriptivos y su valor no deja de ser el resultado de una alteración molecular previa. En cambio lo datos ómicos proporcionan la alteración molecular sin informar de la expresión fenotípica de esa alteración. Por lo tanto parece comprensible que una integración de datos de datos clínicos y ómicos será de gran ayuda en la práctica médica diaria. Las ventajas de realizar un análisis multidimensional de los datos ómicos para obtener más información en vez de seguir una estrategia de análisis de datos ómicos de manera individual se puede resumir en los siguientes puntos [9]:

- La integración de múltiples datos ómicos es una estrategia para prevenir la pérdida de información debida al hecho de que la información procedente de una sola entidad (gen, proteína, tránscrito, etc) puede sugerir refinar el análisis de otros datos ómicos para llenar los vacíos de información o para corregir asociaciones de datos incorrectas.
- Si se obtienen información desde diferentes fuentes sobre el mismo gen o la misma ruta metabólica, es más difícil que se produzcan "falsos positivos".
- El examen de los sistemas de regulación biológicos desde un punto de vista integrativo, es una vía prometedora para deducir el funcionamiento y la regulación precisa del sistema que está siendo estudiado.

3.1.3.1 Tipos de integración

El término *análisis integrativo* se ha usado de forma muy extensa para principalmente describir estudios que integraban múltiples datos de muy diversos orígenes desde los siguientes puntos de vista [11]:

- integración del mismo tipo de datos ómicos procedentes de diferentes estudios
- integración de diferentes tipos de datos ómicos en la misma cohorte de muestras.

Algunos autores definen las dos aproximaciones anteriores como **integración horizontal** y **integración vertical** respectivamente. En la imagen 11 se puede observar un esquema de estos dos tipos de aproximaciones:

Desde el punto de vista de este trabajo, nos interesa más estudiar el tipo de integración vertical, donde se

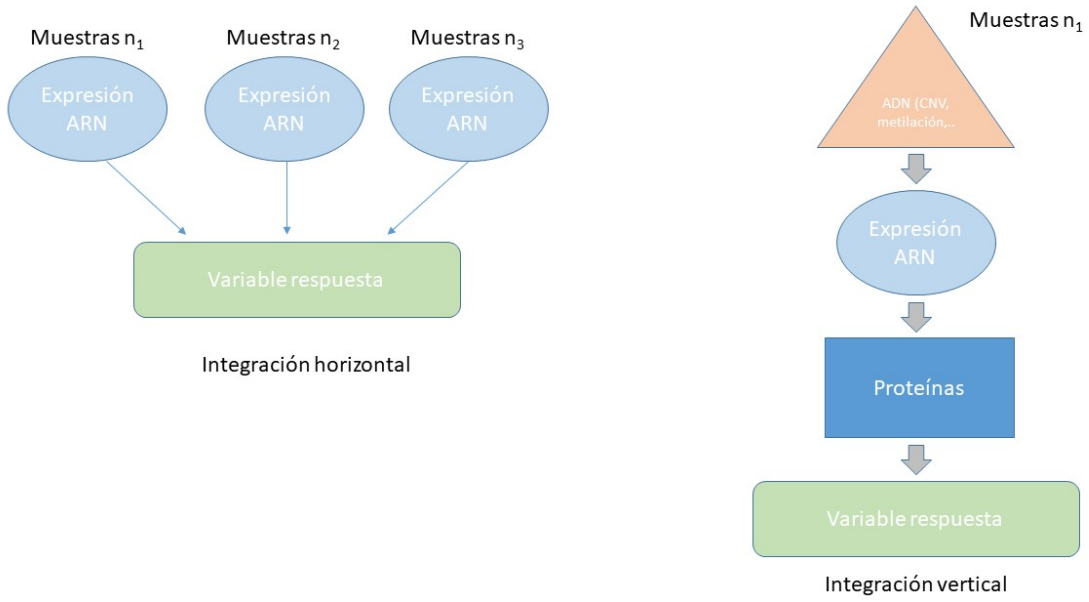


Figura 11: A la izquierda se muestra la integración horizontal y a la derecha la integración vertical

busca unir la información procedentes desde diferentes fuentes de datos ómicos. Existen muchos trabajos descritos sobre diferentes aproximaciones de cómo realizar este tipo de integración vertical, pero una limitación común en todos ellos es la falta de un marco preciso y unificador para resumir los métodos de integración, se ha realizado mucho trabajo al respecto pero todavía no hay un consenso sobre cuál es la mejor metodología a utilizar. Una de las aproximaciones más comunes es la de la **selección de variables**. La selección de variables ha jugado un papel central en el análisis de los datos ómicos a nivel individual, y teniendo en cuenta que los datos multiómicos es simplemente la agregación de diferentes capas de datos ómicos individuales, parece sencillo pensar que el análisis integrativo de datos multiómicos, puede requerir de técnicas de selección de variables. Por otro lado los datos ómicos son multidimensionales, pero solamente unas pocas variables de todas las que se han analizado deben estar relacionadas con la variable respuesta estudiada, por ello vuelve a tomar importancia la aproximación de la integración mediante la selección de variables. Muchos de los métodos estadísticos utilizados en la integración vertical de datos ómicos multidimensionales están basados en modelos y pueden ser clasificados como análisis de regresión (serían métodos supervisados) o análisis exploratorios (que serían métodos no supervisados), dependiendo de si la finalidad del estudio es la predicción, basada en unas características fenotípicas o no. Independientemente del tipo de análisis realizado, el paso de separar el ruido de las señales importantes, juega un papel crucial [12]. Una estrategia de integración aparentemente sencilla, que resulta ser extraordinariamente efectiva muchas veces, es tratar a los datos ómicos de las diferentes plataformas estudiadas de igual manera y realizar la integración de una manera paralela. En la figura 12 se muestra de manera esquemática esta aproximación:

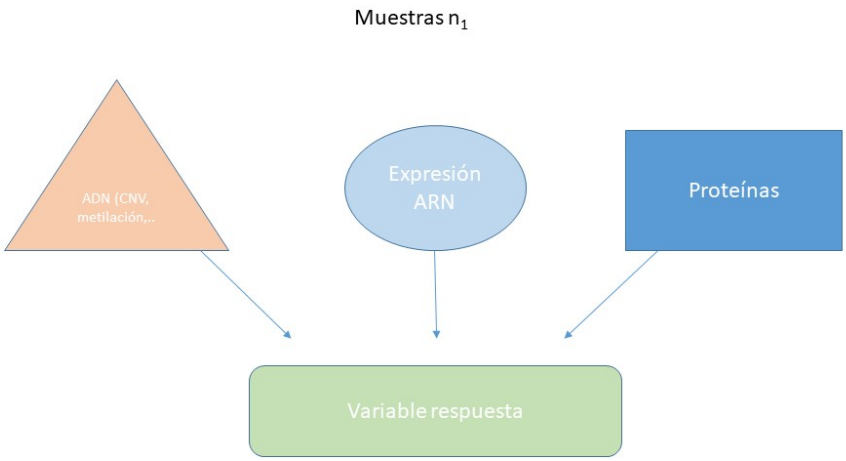


Figura 12: En la integración paralela todas las ómicas tienen la misma importancia

3.1.4 Métodos de integración. Mixomics

El paquete de `R` `mixOmics` (https://www.bioconductor.org/packages/release/bioc/html/mixOmics.html) propone una aproximación de análisis multivariante, que modelan las características como un conjunto, para la integración de datos ómicos. Esta enfoque puede por lo tanto proporcionar una imagen más completa de un sistema biológico, y complementa los resultados obtenidos por los métodos univariados. `mixOmics` ofrece una amplia gama de métodos multivariados para la exploración e integración de conjuntos de datos biológicos con un enfoque particular en la selección de variables [13]. `mixOmics` defiende la utilización de los métodos multivariantes ya que en primer lugar, son computacionalmente eficientes para manejar grandes conjuntos de datos, donde el número de las características biológicas (usualmente miles) son mucho más grandes que la cantidad de muestras (usualmente pocas decenas). En segundo lugar, realizan una reducción de dimensión al proyectar los datos en una subespacio mientras captura y resalta las mayores fuentes de variación de los datos, resultando en una poderosa visualización del sistema biológico bajo estudio. Por último, hacen pocos supuestos previos sobre la distribución de datos, cosa que los hacen altamente flexibles para responder preguntas de actualidad. En un inicio `mixOmics` se introdujo en el contexto de análisis supervisado, donde los objetivos son clasificar o discriminar grupos de muestras, para identificar el subconjunto más discriminante de características biológicas, y para predecir la clase de nuevas muestras. Posteriormente se amplió el método de *Partial Least Square*: Análisis discriminante (sPLS-DA) que se desarrolló originalmente para el análisis supervisado de un conjunto de datos. `mixOmics` ofrece una amplia gama de técnicas de reducción de

dimensiones multivariadas diseñadas a cada respuesta preguntas biológicas específicas, a través de análisis no supervisados o supervisados. Los métodos de análisis no supervisados incluyen el Análisis de Componentes Principales, basado en mínimos cuadrados parciales iterativos no lineales para valores perdidos, el análisis de componentes independientes, Regresión de mínimos cuadrados parciales (PLS), también conocido como Proyección a estructuras latentes, PLS multigrupo, y análisis de correlaciones canónicas regularizadas (rGCCA) basado en un algoritmo PLS. Los métodos de análisis supervisados incluyen PLS-Análisis discriminante PLS-DA, GCC-DA y PLS-DA multigrupo [13]. Además, `mixOmics` proporciona un nuevo método de variantes dispersas que permite la selección de características, la identificación de predictores clave (por ejemplo, genes, proteínas, metabolitos,...) que pueden constituir una firma molecular. La selección de las variables se realiza mediante regularización l_1 (LASSO [14]), el cual está implementado dentro de cada método estadístico utilizado. Una característica muy importante del paquete `mixOmics` es que ha invertido un gran esfuerzo en la fácil interpretación de los resultados creando una gran cantidad de formas de visualizar las asociaciones entre las variables, así como de crear protocolos y flujos de trabajo relativamente fáciles de seguir [15]. Dentro de los diferentes métodos descritos dentro del paquete `mixOmics`, en este trabajo se va a utilizar uno llamado **DIABLO** que permite la integración de varias ómicas hechas en las mismas muestras. En el siguiente apartado de la memoria se va a describir en que consiste este método.

3.1.4.1 DIABLO

El método central DIABLO se basa en el análisis de Correlación Canónica Generalizada [16], que a diferencia de lo que su nombre sugiere, generaliza PLS para múltiples conjuntos de datos realizados en las mismas muestras, y el método disperso de sGCCA [17]. Partiendo del paquete de `R` `RGCCA` de Tenenhaus et al, el equipo de `mixOmics` dice haber mejorado sustancialmente el método y el código para los diferentes tipos de análisis, incluida la N-integración sin supervisión (block.pls, block.spls) y los análisis supervisados (block.plsda, block.splsda) [18].

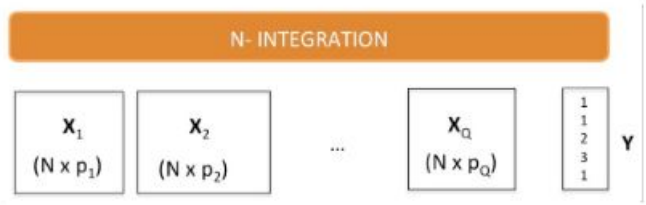


Figura 13: Esquema de la N-integración presentada por mixOmics

El objetivo de la integración en N con los métodos dispersos planteados por DIABLO es identificar variables correlacionadas (o expresadas conjuntamente) medidas en conjuntos de datos heterogéneos que también expliquen el resultado categórico de interés (análisis supervisado). La tarea de integración de datos múltiples

no es trivial, ya que el análisis puede verse fuertemente afectado por la variación entre los fabricantes o las plataformas tecnológicas ómicas a pesar de que se mide en las mismas muestras biológicas. Además, para evitar una selección errónea de variables, es mejor analizar cada información por separado primero para comprender bien de dónde provienen las principales fuentes de variación y también para guiar específicamente el proceso de integración [18]. Además, en el método implementado por DIABLO se extiende el sGCCA para el Análisis Discriminante, y se mejora sustancialmente el código `R` y desarrollando como viene siendo habitual en los proyectos llevados a cabo por `mixOmics`, resultados gráficos innovadores para interpretar los resultados.

3.2 Datos para el análisis

En esta sección se van a describir los datos que se han utilizado para realizar este trabajo, así como el procedimiento seguido en cada caso para prepararlos para poder ser utilizados con el método de integración seleccionado. Antes de ellos se va a definir el problema biológico sobre el cual se quiere trabajar: las muestras que se van a utilizar en el presente estudio provienen de pacientes a los cuales se les ha realizado un trasplante de pulmón en la unidad de pneumología del Hospital Vall d'Hebron de Barcelona. Algunos de estos pacientes responden bien (CONTROLES) al trasplante mientras que otros no responden bien (CASE). Los pacientes que no responden bien, lo pueden hacer a las pocas semanas de haber recibido el trasplante a al cabo de unos años. Si se conociese de alguna manera si el paciente va a responder bien o no al trasplante, mediante la utilización de algún biomarcador, sería posible prevenir el rechazo del nuevo órgano antes de que este suceda. El presente estudio estaba encaminado en la búsqueda de un biomarcador que fuese capaz de prevenir el rechazo o no del órgano.

3.2.1 Datos de expresión génica

En este apartado se van a describir los métodos que se han empleado para realizar el análisis de expresión diferencial a partir de los microarrays de expresión génica realizados (Human Clariom D arrays). Para realizar el análisis de la expresión diferencial se siguió el flujo de trabajo, con ligeras modificaciones debido al tipo de arrays utilizado, descrito en el capítulo tres del libro "Data Analysis for Omic Sciences: Methods and Applications", que trata sobre el análisis de microarrays [6]. A continuación se muestra el código utilizado y los resultados obtenidos.

Primero de todo se definen los directorios de trabajo de la siguiente manera:

```
mainDir <-getwd()
workingDir <- mainDir
celDir <- file.path(workingDir, "celfiles")
```

A continuación se cargan las librerías necesarias para realizar el análisis:

```r
library(xtable)
library(Biobase)
library(oligo)
library(ggplot2)
library(ggrepel)
source("https://raw.githubusercontent.com/uebvhir/UEB_PCA/master/UEB_plotPCA3.R")
library(genefilter)
library(clariomdhumantranscriptcluster.db)
library(limma)
```

En el repositorio de github que se menciona en el chunk de carga de librerías, existe una función para hacer un análisis de componentes principales. Posteriormente se lee el archivo targets:

A continuación se muestra el archivo targets:

	Group	ShortName	colores
P01.CEL	CASE	CASE.P01	green
P04.CEL	CASE	CASE.P04	green
P05.CEL	CONTROL	CONTROL.P05	yellow
P06.CEL	CASE	CASE.P06	green
P07.CEL	CONTROL	CONTROL.P07	yellow
P08.CEL	CASE	CASE.P08	green
P09.CEL	CONTROL	CONTROL.P09	yellow
P10.CEL	CASE	CASE.P10	green
P11.CEL	CONTROL	CONTROL.P11	yellow
P12.CEL	CASE	CASE.P12	green
P13.CEL	CONTROL	CONTROL.P13	yellow
P14.CEL	CASE	CASE.P14	green
P15.CEL	CASE	CASE.P15	green
P16.CEL	CONTROL	CONTROL.P16	yellow
P17.CEL	CASE	CASE.P17	green
P18.CEL	CONTROL	CONTROL.P18	yellow
P19.CEL	CASE	CASE.P19	green
P20.CEL	CONTROL	CONTROL.P20	yellow
P21.CEL	CASE	CASE.P21	green
P22.CEL	CONTROL	CONTROL.P22	yellow
P23.CEL	CASE	CASE.P23	green
P24.CEL	CONTROL	CONTROL.P24	yellow
P25.CEL	CASE	CASE.P25	green
P26.CEL	CONTROL	CONTROL.P26	yellow
P27.CEL	CASE	CASE.P27	green
P28.CEL	CONTROL	CONTROL.P28	yellow
P29.CEL	CASE	CASE.P29	green
P30.CEL	CONTROL	CONTROL.P30	yellow
P31.CEL	CASE	CASE.P31	green
P32.CEL	CASE	CASE.P32	green
P33.CEL	CASE	CASE.P33	green
P34.CEL	CONTROL	CONTROL.P34	yellow
P35.CEL	CONTROL	CONTROL.P35	yellow
P36.CEL	CASE	CASE.P36	green
P37.CEL	CONTROL	CONTROL.P37	yellow
P38.CEL	CASE	CASE.P38	green
P39.CEL	CONTROL	CONTROL.P39	yellow
P40.CEL	CASE	CASE.P40	green
P41.CEL	CONTROL	CONTROL.P41	yellow
P42.CEL	CONTROL	CONTROL.P42	yellow
P43.CEL	CASE	CASE.P43	green
P44.CEL	CONTROL	CONTROL.P44	yellow
P45.CEL	CASE	CASE.P45	green
P46.CEL	CASE	CASE.P46	green

P47.CEL	CASE	CASE.P47	green
P48.CEL	CONTROL	CONTROL.P48	yellow
P49.CEL	CONTROL	CONTROL.P49	yellow
P50.CEL	CASE	CASE.P50	green
P51.CEL	CONTROL	CONTROL.P51	yellow
P52.CEL	CASE	CASE.P52	green
P53.CEL	CONTROL	CONTROL.P53	yellow
P54.CEL	CONTROL	CONTROL.P54	yellow
P55.CEL	CASE	CASE.P55	green
P56.CEL	CASE	CASE.P56	green
P57.CEL	CONTROL	CONTROL.P57	yellow
P58.CEL	CONTROL	CONTROL.P58	yellow
P59.CEL	CASE	CASE.P59	green
P60.CEL	CONTROL	CONTROL.P60	yellow
P61.CEL	CONTROL	CONTROL.P61	yellow
P62.CEL	CONTROL	CONTROL.P62	yellow

Cuadro 2: Archivo targets que muestra la asociación entre las muestras y sus covariables

El siguiente paso es leer los archivos CEL que han resultado del experimento:

```
celFiles <- list.celfiles(celDir, full.names = TRUE)
rawData <- read.celfiles(celFiles, phenoData = pd)
colnames(rawData) <-rownames(pData(rawData)) <- pd@data$ShortName
```

Una vez se han leído los archivos CEL se puede realizar un primer control de calidad con los datos crudos. Se va a realizar un boxplot y un análisis de componentes principales:

```
colores <-  pd@data$colores
grupos <- as.factor(pd@data$Group)
colorPCA <- c("green", "yellow")

boxplot(rawData, which="all", cex.axis=0.6, col = colores,  las = 2,
          main="Boxplot for arrays intensity: Raw Data")

plotPCA3(exprs(rawData), labels = colnames(rawData), factor = grupos,
         title="Raw data", scale = FALSE, colores = colorPCA, size = 2.5)
```

A continuación se muestran las dos gráficas (14 y 15):

Se observa en la figura 14, que las intensidades de las muestras son un poco variables tal como es de esperar para los datos crudos. En la figura 15 del PCA, no se observa una agrupación clara para las muestras *CASE CONTROL*, pese a la gran cantidad de variabilidad que explica la primera componente. Una vez ha quedado de manifiesto que no se ha de descartar ninguna muestra por su mala calidad, se procede a normalizar utilizando el método de RMA [19].

```
eset_rma <- rma(rawData)
```

Una vez se han normalizado los datos se procede a realizar de nuevo un rápido control de calidad antes de

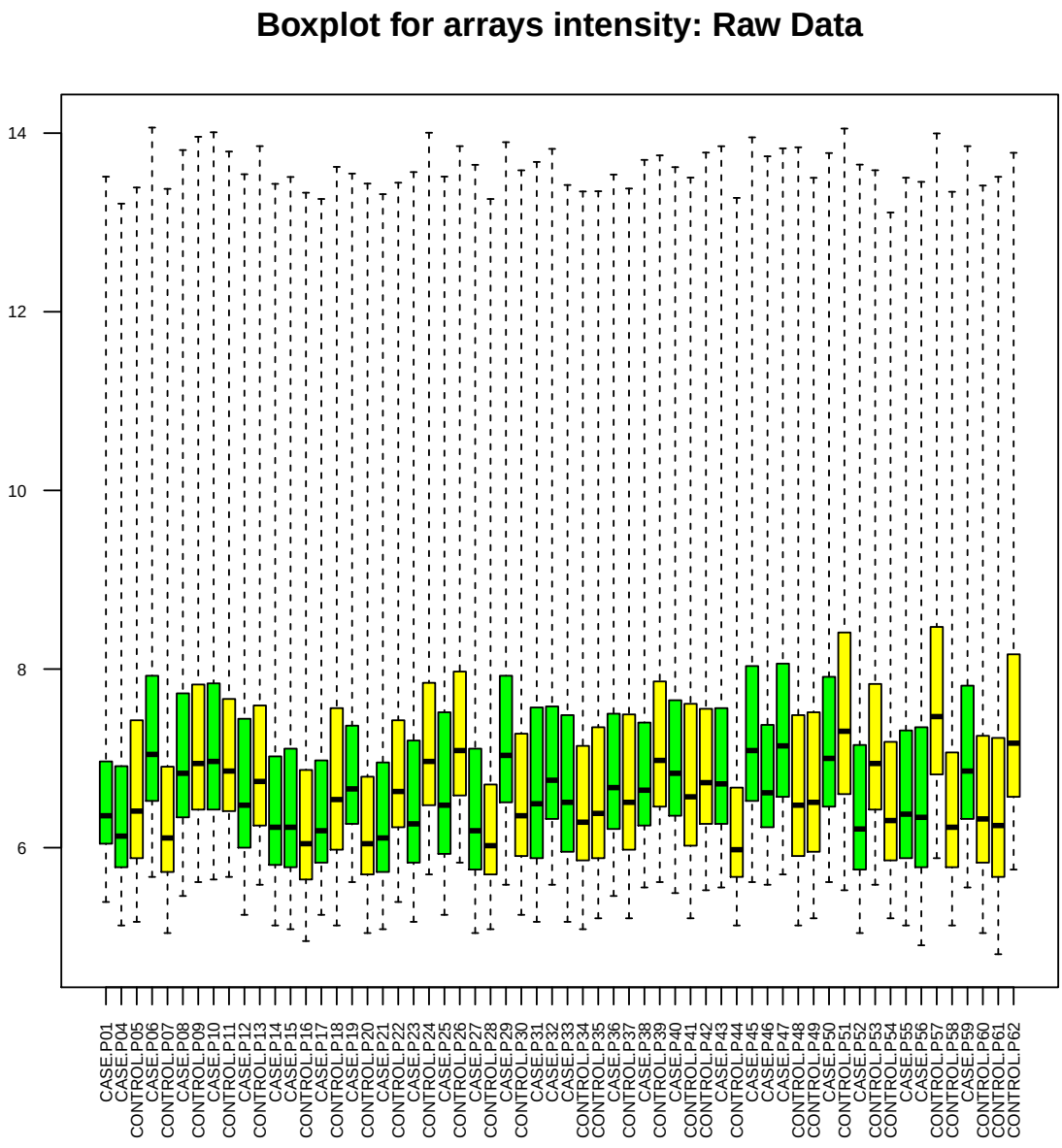


Figura 14: Boxplot de los datos crudos

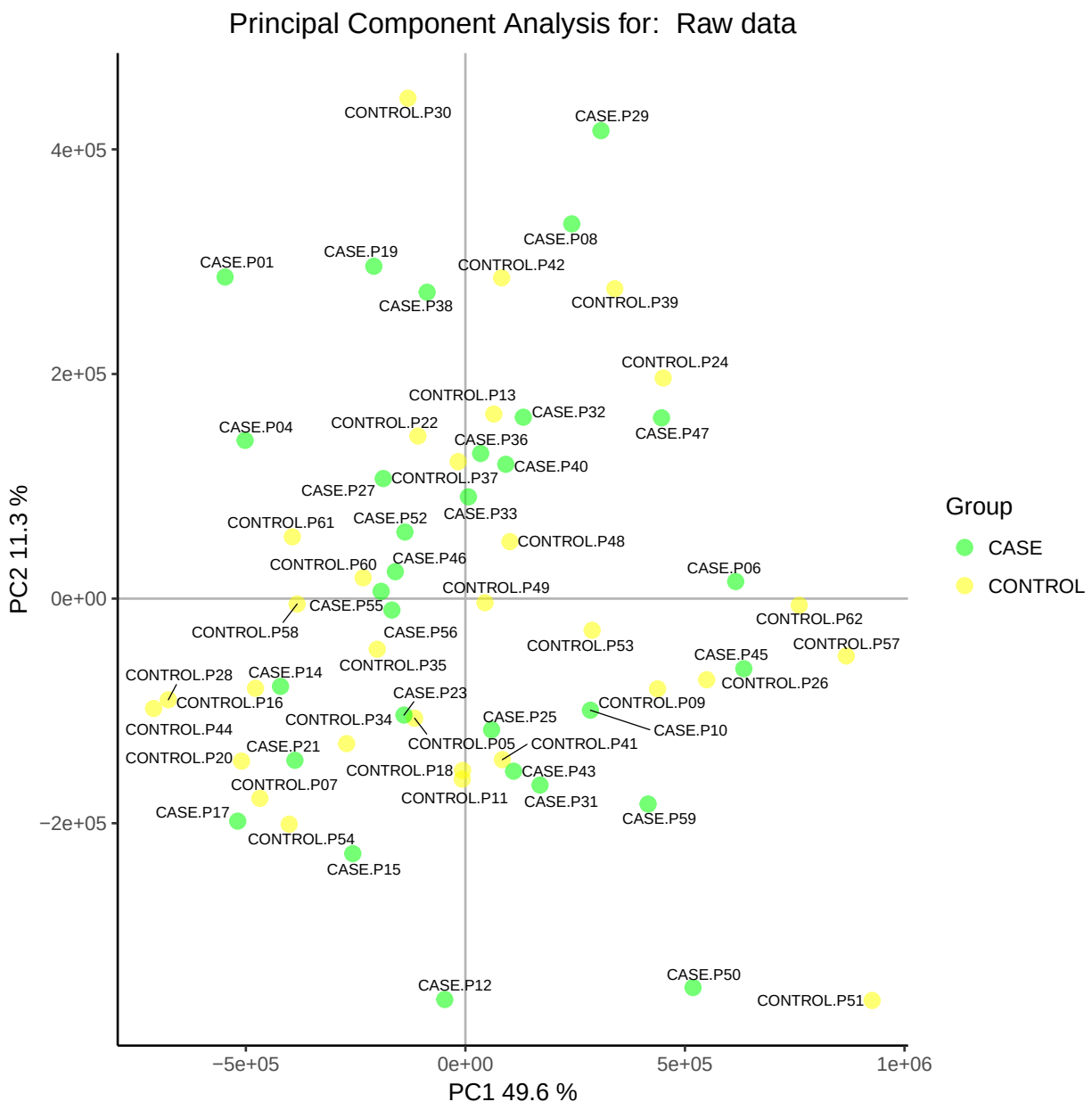


Figura 15: PCA de los datos crudos

empezar con el análisis de expresión diferencial:

```
boxplot(eset_rma,main="Boxplot of Normalized data", cex.axis=0.5, col=colores, las=2)
plotPCA3(exprs(eset_rma), labels = colnames(eset_rma), factor = grupos,
         title="Normalized data", scale = FALSE, colores = colorPCA, size = 2.5)
```

A continuación se muestran las dos gráficas (16 y 17):

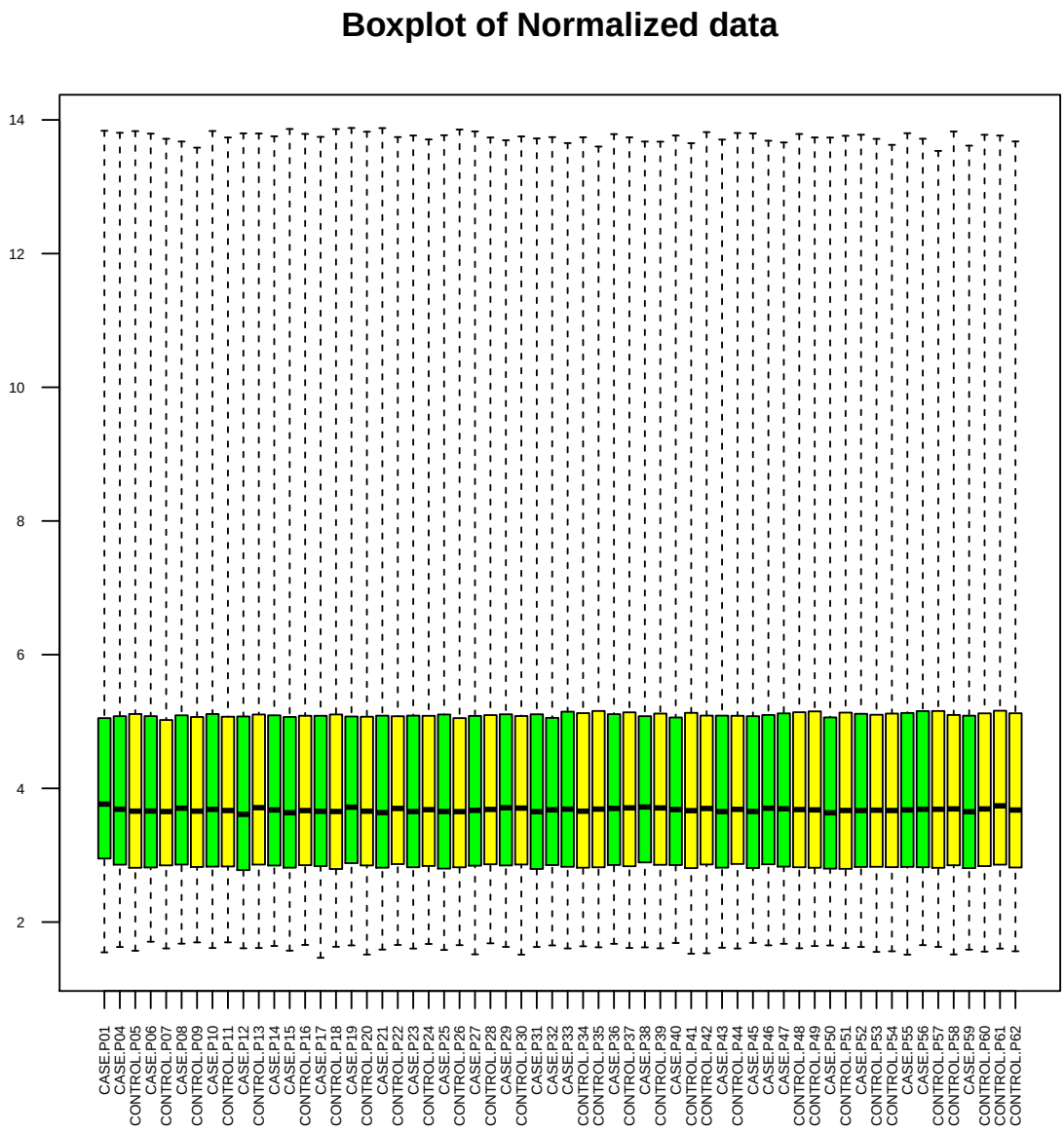


Figura 16: Boxplot de los datos normalizados

Se observa en la gráfica del boxplot (16) que ahora las intensidades de todas las muestras ya son prácticamente idénticas una vez se han normalizado los datos. En la imagen del PCA (17), se sigue sin observar una agrupación clara de las muestras de uno y otro grupo.

Una vez se ha realizado el control de calidad se puede proceder a realizar el análisis de expresión diferencial. Lo primero de todo es realizar un filtrado inespecífico mediante el cual eliminamos todos aquellos genes que no han variado en ninguna de las muestras. Nos quedaremos con 35 % de los genes:

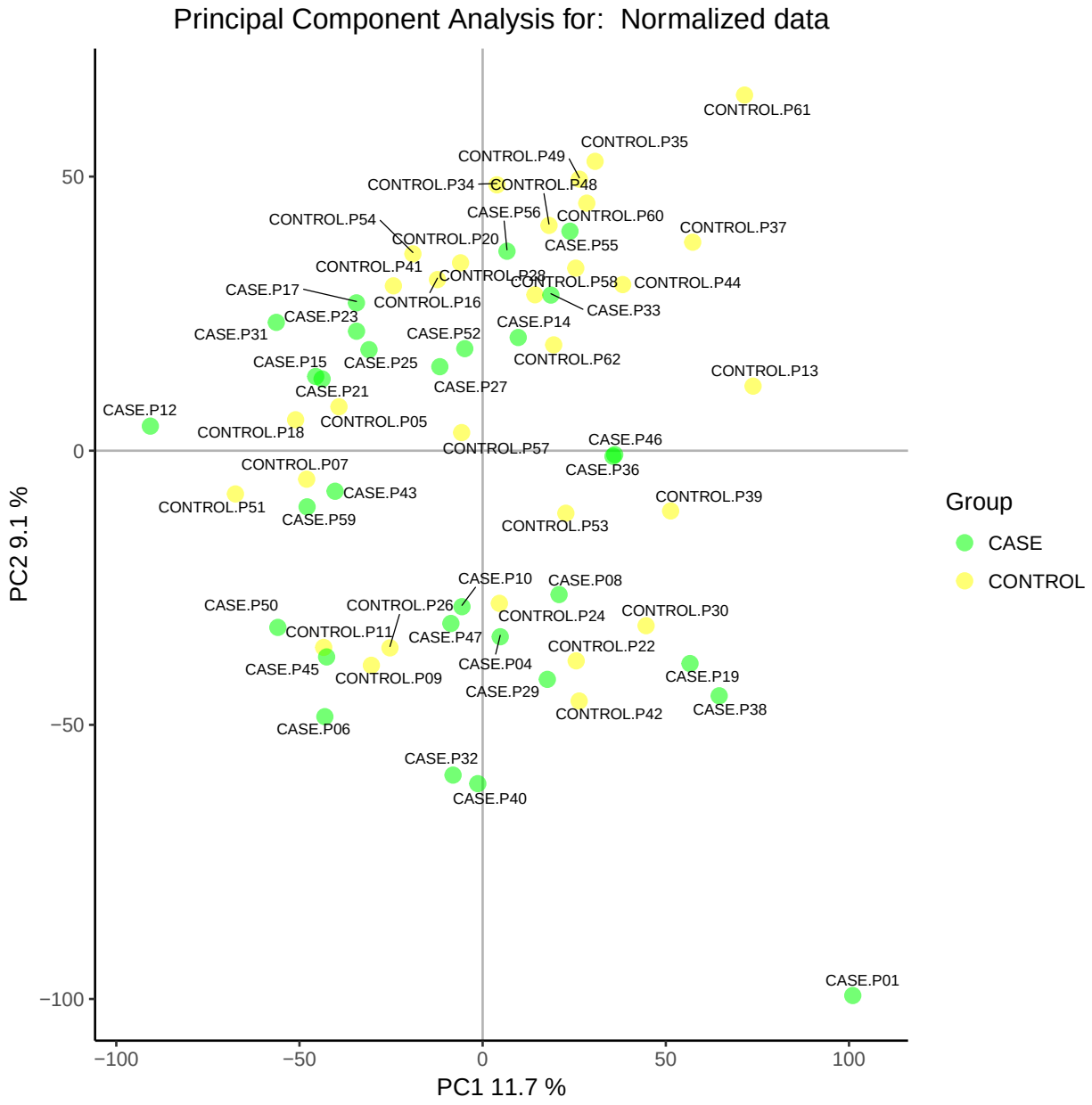


Figura 17: PCA de los datos normalizados

```r
annotation(eset_rma) <- "clariomdhumantranscriptcluster.db"
filtered <- nsFilter(eset_rma, require.entrez = TRUE,
         var.func=IQR, remove.dupEntrez = TRUE, require.GOBP = FALSE,
         require.GOCC = FALSE, require.GOMF = FALSE,
         var.filter = TRUE, var.cutoff = 0.66,
         filterByQuantile = TRUE, feature.exclude = "^AFFX")
dim(filtered$eset) #8239 60
eset_filtered <-filtered$eset
```

De esta manera nos quedaremos con **8239 genes** para seguir en el análisis. Ahora ya se puede hacer el diseño de la comparación, la matriz de contrastes y ajustar el modelo. Se utiliza el paquete `limma` [20], que utiliza los modelos lineales para realizar las comparaciones:

```r
grupo <- as.factor(targets$Group)
design2 <- model.matrix( ~ 0 + grupo)
colnames(design2)<-c("CASE","CONTROL")
rownames(design2)<-targets$ShortName
print(design2)
dim(design2) #60 2

contrastsMatrix2 <- makeContrasts(CASEvsCONTROL = CASE - CONTROL,
                                  levels = design2)
fit2<-lmFit(eset_filtered, design2)
fit.main2<-contrasts.fit(fit2, contrastsMatrix2)
fit.main2<-eBayes(fit.main2)
```

Una vez ajustado el modelo se puede generar la toptable:

```r
topTab_CASEvsCONTROL <- topTable (fit.main2, number = nrow(fit.main2),
                                  coef="CASEvsCONTROL", adjust="fdr")
```

A continuación se visualiza las 50 primeras filas de la toptable:

	Gene.Symbol	Entrez	logFC	AveExpr	t	P.Value	adj.P.Val	B
TC0400007933.hg.1	ANXA3	306	-0.96	5.66	-6.20	0.00	0.00	8.40
TC0100009656.hg.1	FCGR1B	2210	-1.04	7.56	-5.76	0.00	0.00	6.79
TC1900012043.hg.1	LILRA4	23547	0.63	4.20	5.33	0.00	0.00	5.25
TC0600011960.hg.1	TNFRSF21	27242	0.47	5.55	5.28	0.00	0.00	5.06
TC1700011600.hg.1	KCNJ2-AS1	400617	-0.61	4.74	-5.20	0.00	0.00	4.79
TC0400012621.hg.1	ACSL1	2180	-0.83	8.20	-5.19	0.00	0.00	4.75
TC0800011861.hg.1	LRRC6	23639	-0.67	4.50	-5.17	0.00	0.00	4.68
TC1000010273.hg.1	NRP1	8829	0.58	4.14	5.06	0.00	0.00	4.28
TC2100008510.hg.1	KCNJ15	3772	-0.82	9.78	-5.04	0.00	0.00	4.23
TC0100009697.hg.1	FCGR1CP	100132417	-0.82	7.57	-4.96	0.00	0.00	3.95
TC1700010221.hg.1	DHRS13	147015	-0.55	6.78	-4.95	0.00	0.00	3.91
TC0700009411.hg.1	MGAM2	93432	-0.94	5.51	-4.94	0.00	0.00	3.88

TC1100012126.hg.1	MMP8	4317	-1.89	7.55	-4.89	0.00	0.00	3.72
TC0300013471.hg.1	BCL6	604	-0.65	10.79	-4.83	0.00	0.00	3.50
TC0500008627.hg.1	SLC22A4	6583	-0.57	6.03	-4.77	0.00	0.01	3.32
TC0200008005.hg.1	DYSF	8291	-0.79	8.76	-4.60	0.00	0.01	2.74
TC1300009598.hg.1	GPR183	1880	0.79	8.34	4.59	0.00	0.01	2.71
TC0300011517.hg.1	PROK2	60675	-0.60	10.06	-4.52	0.00	0.01	2.47
TC1100006492.hg.1	PNPLA2	57104	-0.38	7.37	-4.49	0.00	0.01	2.37
TC1400008460.hg.1	PLD4	122618	0.44	5.99	4.39	0.00	0.02	2.06
TC0100017445.hg.1	TLR5	7100	-0.48	7.75	-4.36	0.00	0.02	1.95
TC0200011040.hg.1	ITM2C	81618	0.58	10.30	4.35	0.00	0.02	1.93
TC1100007513.hg.1	MIR3161	100423000	-0.61	5.57	-4.32	0.00	0.02	1.82
TC0500009562.hg.1	HRH2	3274	-0.50	7.34	-4.29	0.00	0.02	1.72
TC0100018308.hg.1	FCGR2A	2212	-0.42	10.20	-4.28	0.00	0.02	1.70
TC0300010949.hg.1	LTF	4057	-1.46	9.49	-4.27	0.00	0.02	1.67
TC0100010252.hg.1	FCER1A	2205	0.84	8.13	4.27	0.00	0.02	1.65
TC0400012622.hg.1	SLED1	643036	-0.65	4.75	-4.26	0.00	0.02	1.65
TC1700008769.hg.1	KCNJ2	3759	-0.70	7.21	-4.25	0.00	0.02	1.59
TC1300009638.hg.1	RPS26	6231	1.33	8.76	4.24	0.00	0.02	1.58
TC0600007825.hg.1	MAPK14	1432	-0.57	7.90	-4.24	0.00	0.02	1.57
TC2200007088.hg.1	LIMK2	3985	-0.55	8.90	-4.23	0.00	0.02	1.55
TC2000007102.hg.1	HCK	3055	-0.37	9.23	-4.23	0.00	0.02	1.54
TC1800009268.hg.1	DSC2	1824	-0.69	5.62	-4.23	0.00	0.02	1.53
TC1200009800.hg.1	SLC2A3	6515	-0.56	8.25	-4.22	0.00	0.02	1.51
TC1200009892.hg.1	OLR1	4973	-1.09	5.37	-4.20	0.00	0.02	1.45
TC0200011171.hg.1	SH3BP4	23677	0.32	6.70	4.18	0.00	0.02	1.37
TC2000009058.hg.1	TGM2	7052	-0.33	5.51	-4.17	0.00	0.02	1.34
TC0100018551.hg.1	CHIT1	1118	-0.83	5.29	-4.17	0.00	0.02	1.34
TC0100012421.hg.1	FAM41C	284593	-0.54	9.36	-4.16	0.00	0.02	1.31
TC1900009134.hg.1	SBNO2	22904	-0.47	7.31	-4.14	0.00	0.02	1.27
TC2000009969.hg.1	SIRPD	128646	-0.41	6.94	-4.14	0.00	0.02	1.26
TC0300007733.hg.1	FAM19A1	407738	0.52	5.56	4.13	0.00	0.02	1.24
TC0X00007586.hg.1	RPS26P11	441502	1.15	8.78	4.13	0.00	0.02	1.21
TC1000007471.hg.1	ALOX5	240	-0.46	8.94	-4.10	0.00	0.02	1.14
TC0100009866.hg.1	FCGR1A	2209	-0.86	6.98	-4.09	0.00	0.02	1.09
TC1000006652.hg.1	PFKFB3	5209	-0.48	7.40	-4.07	0.00	0.02	1.03
TC1900011657.hg.1	MCEMP1	199675	-0.73	7.41	-4.06	0.00	0.02	1.02
TC1900011742.hg.1	CEACAM6	4680	-0.80	6.72	-4.06	0.00	0.02	1.00
TC0X00006891.hg.1	GK	2710	-0.62	6.65	-4.06	0.00	0.02	0.99

Cuadro 3: 50 primeros genes más diferencialmente expresados

En la toptable 3 se puede observar que de los primeros 50 genes que se muestran todos ellos están diferencialmente expresados con un p valor ajustado inferior a 0.05. El cambio biológico, representado por la columna "logFC" no es muy elevado en ninguno de los genes que se muestran.

3.2.2 Datos de expresión de miRNA

En este apartado se van a mostrar los pasos que se han seguido para realizar el análisis de expresión diferencial en el caso de los miRNA. En este caso se utilizaron microarrays de Thermofisher del tipo miRNA 4.0. En lineas generales el procedimiento es muy similar al que se ha comentado anteriormente en el caso de los microarrays de expresión génica, por lo que algunos pasos se obviarán.

La lectura del targets y de los datos se hace de manera equivalente a como se ha hecho en el caso de los arrays Clariom D. A continuación se muestran las imágenes del control de calidad con los datos antes de normalizar 18 y 19

Se observa en la imagen de las intensidades de los arrays antes de normalizar (18) bastante variabilidad entre las muestras, especialmente en la muestra *CONTROL.P37* que se separa un poco del resto. En el caso de la

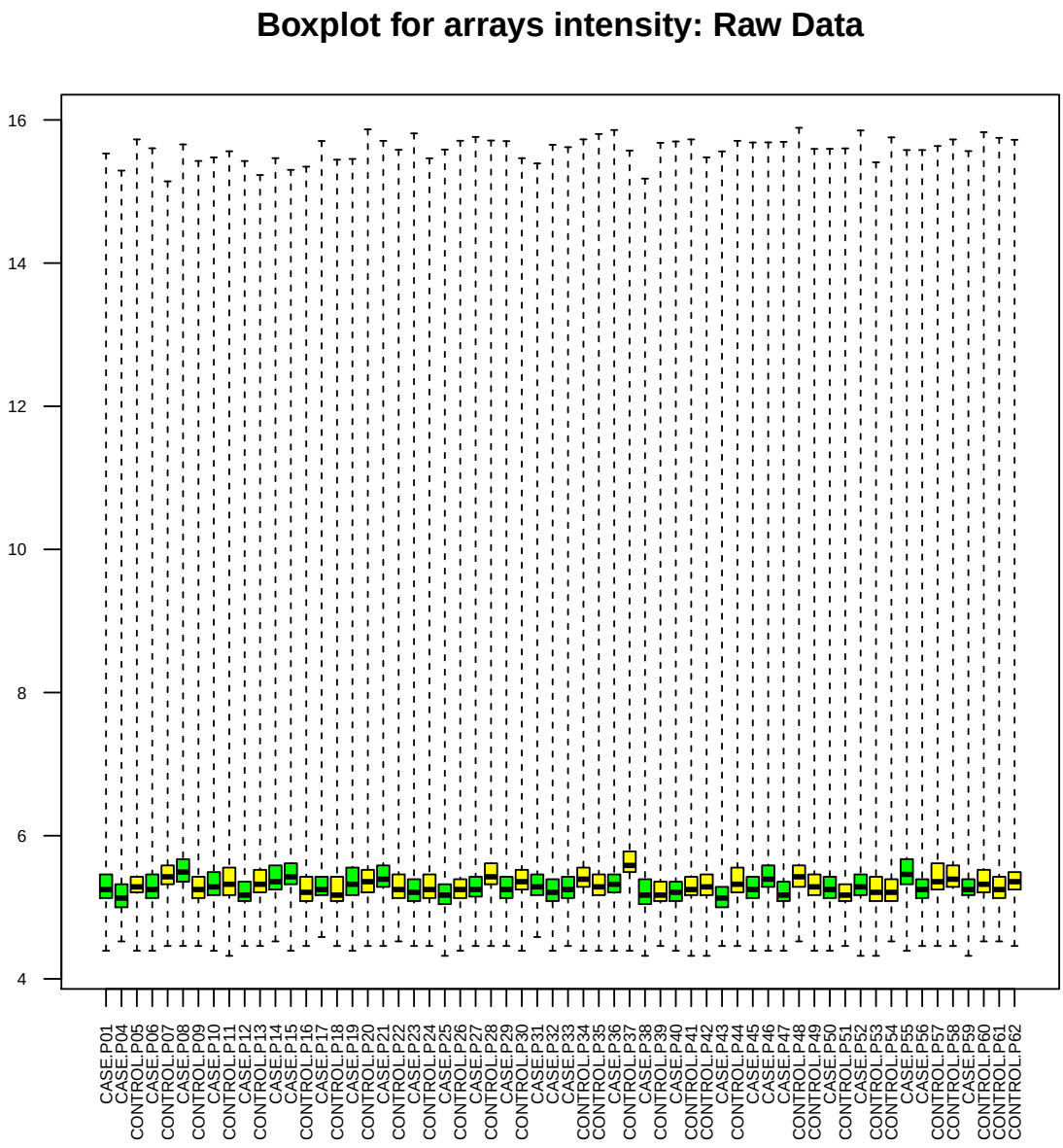


Figura 18: Boxplot de los datos antes de normalizar

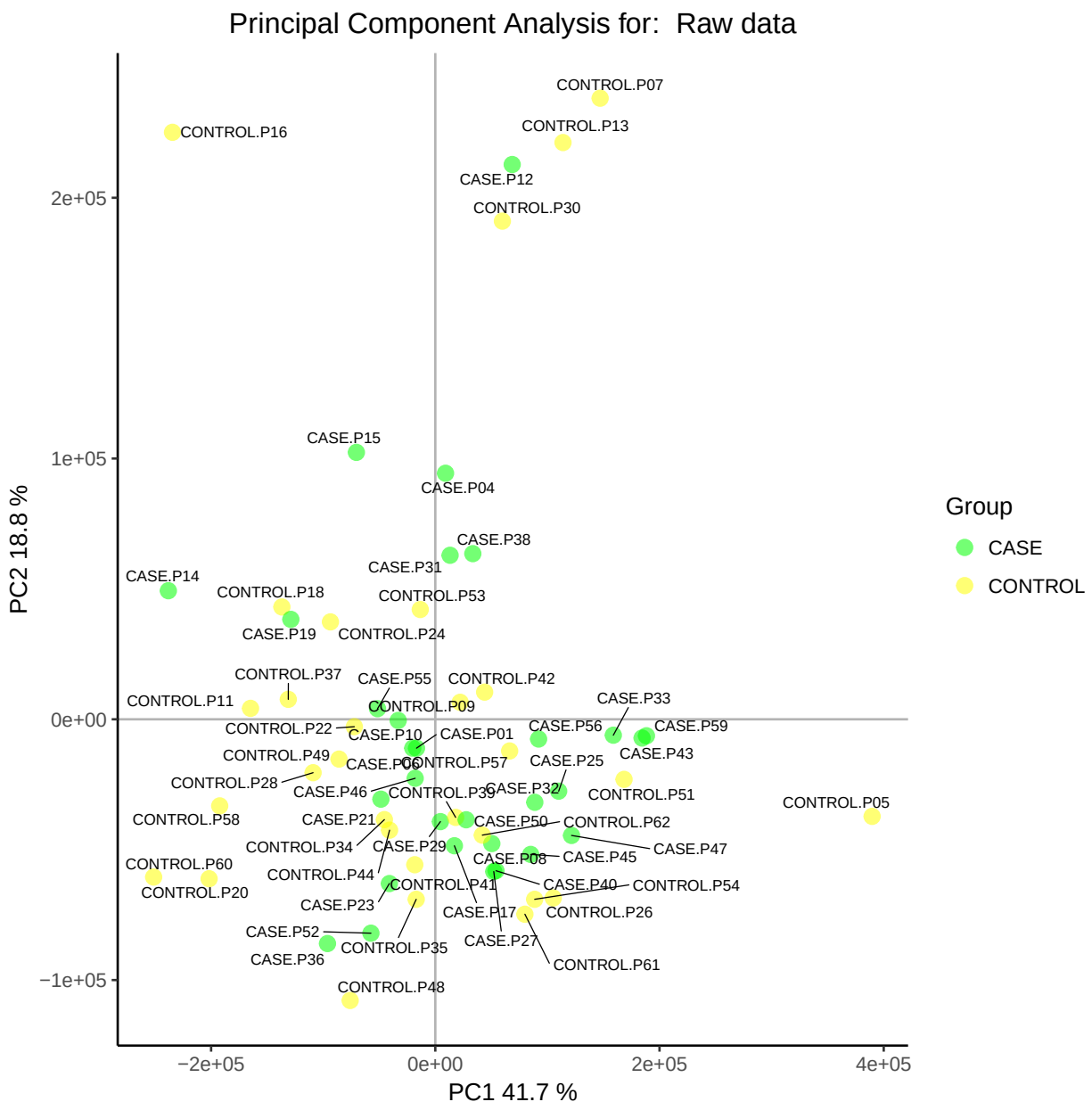


Figura 19: PCA de los datos antes de normalizar

imagen del PCA (19), se observan la mayoría de las muestras agrupadas en el centro y unas pocas muestras que se separan del resto en la parte superior e inferior derecha del gráfico. De todas las maneras no parece haber ninguna razón objetiva para eliminar ninguna de las muestras antes de proceder a la normalización de los dato, que se realiza de manera análoga a como se ha hecho en el caso de los arrays de expresión. Una vez normalizadas las muestras se procede a realizar de nuevo el control de calidad.

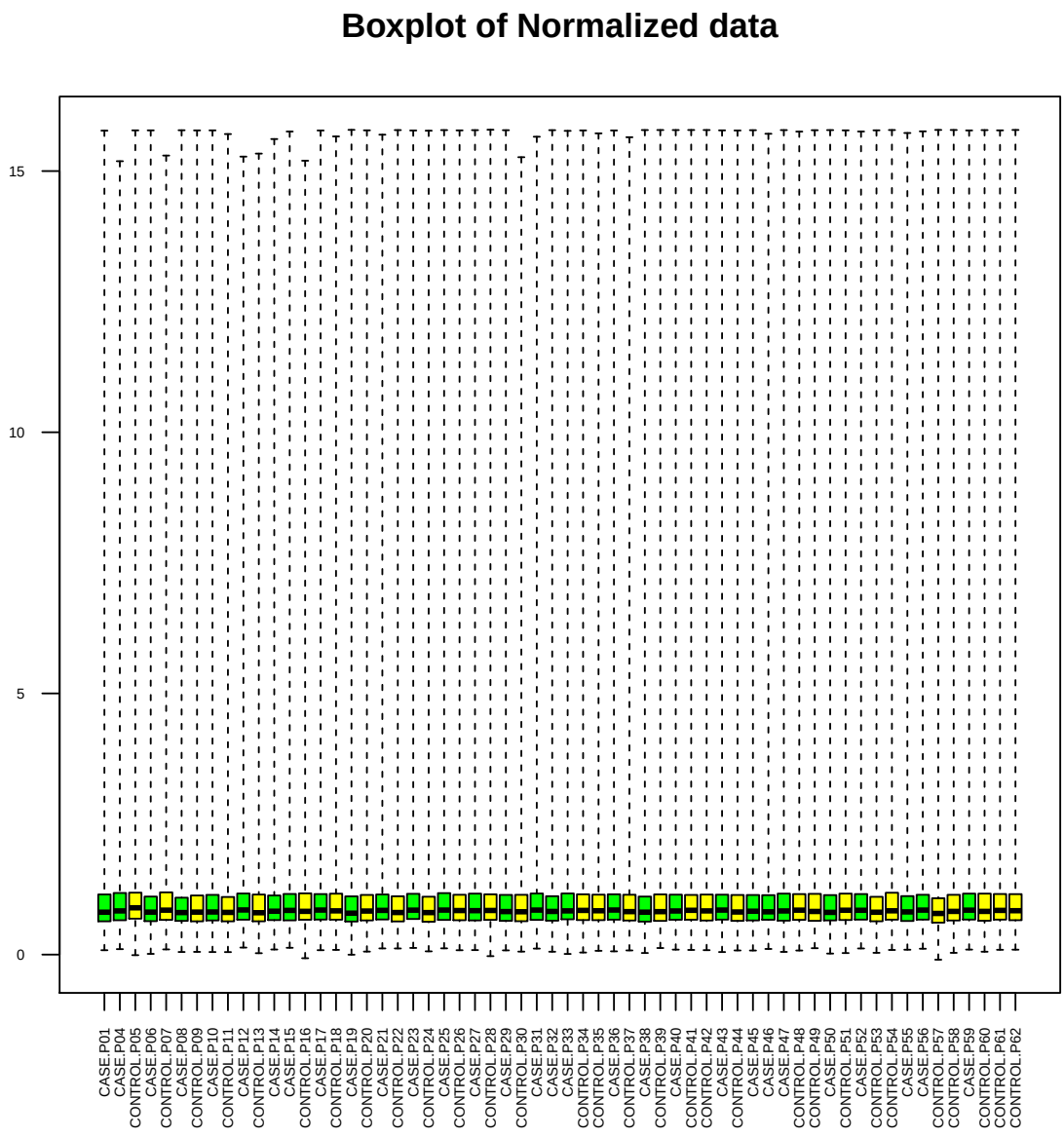


Figura 20: Boxplot de los datos una vez normalizados

En la imagen del boxplot de los datos normalizados (20) se observa que todas las muestras son muy similares. En la imagen del PCA (21), ya no se observa ninguna muestra que se aleje del resto. Se observa una mayor

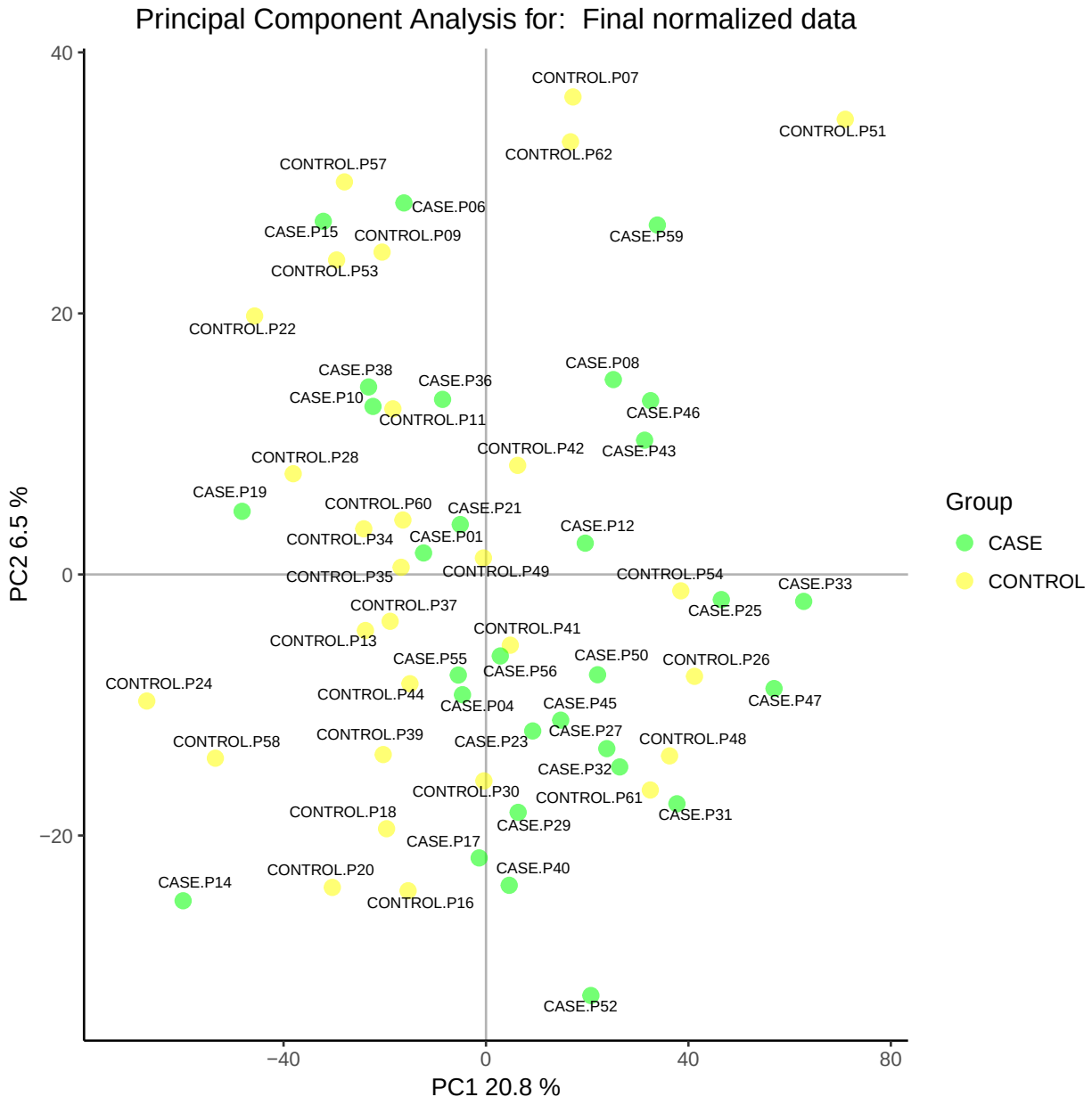


Figura 21: PCA de los datos una vez normalizados

agrupación de las muestra de la condición *CASE* a la derecha del gráfico y las de la condición *CONTROL* a la izquierda del gráfico.

En los arrays de miRNA 4.0 a parte de encontrarse las sondas para detectar los miRNAs para la especie humana se encuentran las sondas para los miRNAs de muchas otras especies, por lo que resulta conveniente el hacer un filtrado de las sondas para trabajar solamente con las de la especie humana.

```
#se lee archivo de anotaciones que se ha bajado de la web de affymetrix
anotacions <- read.csv(file.path(dataDir, "miRNA-4_0-st-v1.annotations.mod.csv"),
                       sep=",",header=TRUE)

#se redefinen el nombre de las columnas
library(dplyr)
anotacions2 <- add_rownames(anotacions,"Probe.Set.ID2")
colnames(anotacions2) <- c("Probe.Set.ID","Probe.Set.Name","Accession",
                           "Transcript.ID.Array.Design.","Sequence.Type",
                           "Species.Scientific.Name","Alignments","Sequence.Length",
                           "Sequence","Genome.Context","Clustered.miRNAs.within.10kb",
                           "Target.Genes","GeneChip.Array","Annotation.Date","Year",
                           "Sequence.Source")

#se seleccionan las de la especie humana
Hanotacions <- anotacions2[which(anotacions2$Species.Scientific.Name == "Homo sapiens"),]
```

Ahora ya si se puede hacer la matriz de contrastes y generar la toptable:

```
grupo <- as.factor(targets1$Group)
design <- model.matrix( ~ 0 + grupo)
colnames(design)<-c("CASE","CONTROL")
rownames(design)<-targets1$ShortName

contrastsMatrix <- makeContrasts(CASEvsCONTROL = CASE - CONTROL,
                                 levels = design)

fit <- lmFit(end.data1, design)
fit.main <- contrasts.fit(fit, contrastsMatrix)
fit.main <- eBayes(fit.main)

topTab_CASEvsCONTROL <- topTable (fit.main, number = nrow(fit.main),
```

```
                                    coef="CASEvsCONTROL", adjust="fdr")
```

A continuación se visualiza las 50 primeras filas de la toptable:

Symbol	logFC	AveExpr	t	P.Value	adj.P.Val	B
U78	-0.80	4.04	-4.65	0.00	0.12	1.82
ENSG00000212378	-0.82	4.70	-4.28	0.00	0.15	0.88
U78	-0.82	4.70	-4.28	0.00	0.15	0.88
U17b	-0.50	5.11	-3.90	0.00	0.28	-0.04
U75	-0.57	2.65	-3.89	0.00	0.28	-0.06
ENSG00000252199	-0.23	0.97	-3.81	0.00	0.28	-0.27
U75	-0.50	2.58	-3.77	0.00	0.28	-0.36
ACA7B	-0.42	2.42	-3.73	0.00	0.28	-0.45
ACA7	-0.42	2.42	-3.73	0.00	0.28	-0.45
ENSG00000206913	-0.42	2.42	-3.73	0.00	0.28	-0.45
U46	-0.25	1.44	-3.65	0.00	0.33	-0.64
hsa-mir-548o-2	0.19	0.89	3.56	0.00	0.40	-0.84
hsa-mir-3142	0.15	0.74	3.53	0.00	0.41	-0.91
hsa-mir-1973	-0.21	1.42	-3.41	0.00	0.50	-1.18
ACA48	-0.31	3.66	-3.40	0.00	0.50	-1.20
HBI-6	0.14	0.80	3.38	0.00	0.50	-1.24
U30	-0.50	5.05	-3.37	0.00	0.50	-1.26
hsa-miR-223-3p	-1.23	4.18	-3.36	0.00	0.50	-1.29
HBII-180C	-0.33	1.69	-3.29	0.00	0.53	-1.44
hsa-miR-27b-3p	-0.56	1.81	-3.28	0.00	0.53	-1.47
HBII-251	-0.38	1.46	-3.28	0.00	0.53	-1.47
hsa-miR-6734-3p	0.16	0.78	3.26	0.00	0.53	-1.51
ENSG00000264346	0.16	0.73	3.25	0.00	0.53	-1.53
ENSG00000200536	0.15	0.72	3.24	0.00	0.53	-1.56
ENSG00000252438	0.21	0.85	3.20	0.00	0.53	-1.63
ENSG00000253042	-0.15	0.75	-3.20	0.00	0.53	-1.63
ENSG00000221345	0.15	0.84	3.20	0.00	0.53	-1.64
hsa-miR-1290	-0.15	0.64	-3.17	0.00	0.53	-1.70
ENSG00000239098	0.15	0.71	3.16	0.00	0.53	-1.72
hsa-mir-4461	0.15	0.76	3.15	0.00	0.53	-1.74
ENSG00000238925	0.12	0.85	3.12	0.00	0.53	-1.80
U58C	-0.20	1.04	-3.12	0.00	0.53	-1.80
hsa-miR-151b	-0.45	7.40	-3.12	0.00	0.53	-1.80
hsa-mir-548h-3	0.15	0.77	3.12	0.00	0.53	-1.80
hsa-mir-8064	0.21	0.88	3.06	0.00	0.53	-1.92
hsa-miR-543	-0.35	1.13	-3.04	0.00	0.53	-1.97
hsa-mir-3937	0.27	1.25	3.04	0.00	0.53	-1.98
hsa-miR-1180-3p	0.32	9.87	3.03	0.00	0.53	-1.99
hsa-miR-4639-5p	0.19	0.71	3.03	0.00	0.53	-2.00
ACA44	-0.69	3.92	-3.02	0.00	0.53	-2.01
ENSG00000252840	-0.69	3.92	-3.02	0.00	0.53	-2.01
hsa-mir-1287	0.14	0.82	3.02	0.00	0.53	-2.01
hsa-miR-6808-3p	0.22	0.94	3.02	0.00	0.53	-2.02
ENSG00000212579	0.19	0.77	3.00	0.00	0.53	-2.04
hsa-miR-489-5p	-0.14	0.73	-3.00	0.00	0.53	-2.05
hsa-miR-4732-3p	0.34	8.64	2.98	0.00	0.53	-2.08
hsa-let-7a-5p	-0.26	12.94	-2.97	0.00	0.53	-2.11
ENSG00000252236	0.14	0.81	2.97	0.00	0.53	-2.11
hsa-miR-548j-5p	0.13	0.79	2.97	0.00	0.53	-2.12
U31	-0.29	1.38	-2.97	0.00	0.53	-2.12

Cuadro 4: 50 primeros miRNA más diferencialmente expresados

En la toptable 4 se puede observar que de los resultados obtenidos para los miRNA no han sido tan buenos como los obtenidos para la expresión génica. Ningún miRNA tiene un p valor ajustado inferior a 0.05.

3.2.3 Variables clínicas

En este apartado se va a mostrar como se ha llevado a cabo la selección de las variables clínicas que se han

utilizado posteriormente en el análisis de integración. Una vez se han depurado las variables de la base de datos se procede a realizar un análisis descriptivo de las mismas. Para las variables categóricas se han calculado las frecuencias (totales y en porcentaje entre paréntesis) y para las variables continuas se ha calculado la mediana y el rango intercuartílico.

```
dades <- read.table(file.path("dades.csv"), sep = ";", head = T,row.names = 1)
library(compareGroups)
#githubinstall::githubinstall("mmotaF")
library(mmotaF)
#githubinstall::githubinstall("anaStatsUEB")
library(anaStatsUEB)
```

```
res <- compareGroups(~. , data = dades[,-c(1,2,9:11,13,36,56,83,84)], max.xlev = 50,
                     method = 2)
restab <- createTable(res)
export2csv(restab, "descriptive.csv", which.table="descr", sep=";", nmax = TRUE)
descriptive <- read.table(file.path("descriptive.csv"), sep = ";", head = T)
```

	X	X.....ALL.....	N.
1		N=62	
2	Sexo:		62.00
3	Femenino	33 (53.2 %)	
4	Masculino	29 (46.8 %)	
5	Situacion_Familiar:		60.00
6	Viven en Pareja	40 (66.7 %)	
7	Viven Solos	20 (33.3 %)	
8	Profesion:		62.00
9	Blue collar	31 (50.0 %)	
10	Nunca ha trabajado	2 (3.23 %)	
11	White collar	29 (46.8 %)	
12	Nivel_de_estudios:		40.00
13	Primaria	6 (15.0 %)	
14	Secundaria	15 (37.5 %)	
15	Universitarios	19 (47.5 %)	
16	Cuidador_principal:		54.00
17	conyugue	33 (61.1 %)	
18	Hermano	1 (1.85 %)	
19	Hijos	6 (11.1 %)	
20	Padres	14 (25.9 %)	
21	Grado_de_implicacion_del_cuidador:		60.00
22	Alto	4 (6.67 %)	
23	Bajo	56 (93.3 %)	
24	Causa_de_exitus:		5.00
25	CLAD	3 (60.0 %)	
26	reTP por BOS	2 (40.0 %)	
27	Patologia_de_base:		62.00
28	BQ	4 (6.45 %)	
29	EPID	8 (12.9 %)	
30	EPOC	12 (19.4 %)	
31	FQ	27 (43.5 %)	
32	HAP	2 (3.23 %)	
33	Histiocitosis	2 (3.23 %)	
34	LAM	7 (11.3 %)	
35	Peso_kg_	58.0 [49.0;67.0]	61.00
36	Altura_cm_	164 [159;170]	61.00
37	FVC_pre_cc_	1630 [1185;2110]	62.00
38	FVC_pre_	42.0 [29.5;52.5]	62.00
39	FEV1_pre_cc_	780 [640;1038]	62.00

40	FEV1_pre_	25.5 [19.2;31.0]	62.00
41	TIFFENAU_pre	57.0 [42.2;69.2]	60.00
42	TLC_pre	111 [91.0;125]	49.00
43	RV_pre	258 [153;322]	50.00
44	DLCO_pre	15.5 [11.0;21.8]	34.00
45	KCO_pre	3.50 [1.00;17.8]	62.00
46	PO2_pre	60.0 [53.0;67.0]	60.00
47	PCO2_pre	42.5 [35.8;46.0]	60.00
48	O2_domiciliario:		61.00
49	No	11 (18.0 %)	
50	Si	50 (82.0 %)	

Cuadro 5: Análisis descriptivo de las primeras 50 variables

Además se ha realizado un análisis gráfico para cada una de las variables. Para las variables categóricas se ha realizado un diagrama de barras, donde cada barra indica una categoría y el eje vertical el porcentaje, el recuento absoluto se muestra en cada una de las categorías. Para las variables continuas se ha realizado un histograma. En la siguiente imagen (22) se muestran las primeras gráficas del archivo generado:

```
pdf("descriptivePlot.pdf")
desc_plot(dades[,-c(1,2,9:11,13,36,56,83,84)], rowcol = c(1,3),show.lg = TRUE,
          cex.lab = 0.01)
dev.off()
```

Una vez se ha realizado el análisis descriptivo de todas las variables se procede a realizar un test de comparación entre grupos para evaluar si existe relación entre las variables clínicas que se han estudiado hasta ahora y la variable respuesta en el estudio. Para las variables categóricas se ha realizado una prueba Chi-cuadrado, exceptuando en el caso de que las frecuencias absolutas en alguna de las celdas de la tabla sea inferior a 5, que se ha calculado el test exacto de Fisher. Para las variables numéricas se ha realizado la prueba U de Mann-Withney.

```
res <- compareGroups(type~. , data = dades[,-c(1,2,9:11,13,36,56,83,84)],
                     max.xlev = 10, method = 2 )
restab1 <- createTable(res)
export2csv(restab1, "comp.csv", sep=";", nmax = TRUE)
```

	X	X.....CASO.....	X....CONTROL.....	p.overall
1		N=31	N=31	
2	Sexo:			1.000
3	Femenino	17 (54.8 %)	16 (51.6 %)	
4	Masculino	14 (45.2 %)	15 (48.4 %)	
5	Situacion_Familiar:			1.000
6	Viven en Pareja	20 (66.7 %)	20 (66.7 %)	
7	Viven Solos	10 (33.3 %)	10 (33.3 %)	
8	Profesion:			0.545
9	Blue collar	13 (41.9 %)	18 (58.1 %)	
10	Nunca ha trabajado	1 (3.23 %)	1 (3.23 %)	
11	White collar	17 (54.8 %)	12 (38.7 %)	
12	Nivel_de_estudios:			0.737
13	Primaria	4 (14.8 %)	2 (15.4 %)	
14	Secundaria	9 (33.3 %)	6 (46.2 %)	
15	Universitarios	14 (51.9 %)	5 (38.5 %)	
16	Cuidador_principal:			0.429

17	conyugue	17 (63.0 %)	16 (59.3 %)	
18	Hermano	1 (3.70 %)	0 (0.00 %)	
19	Hijos	4 (14.8 %)	2 (7.41 %)	
20	Padres	5 (18.5 %)	9 (33.3 %)	
21	Grado_de_implicacion_del_cuidador:			1.000
22	Alto	2 (6.90 %)	2 (6.45 %)	
23	Bajo	27 (93.1 %)	29 (93.5 %)	
24	Patologia_de_base:			0.857
25	BQ	2 (6.45 %)	2 (6.45 %)	
26	EPID	4 (12.9 %)	4 (12.9 %)	
27	EPOC	4 (12.9 %)	8 (25.8 %)	
28	FQ	14 (45.2 %)	13 (41.9 %)	
29	HAP	1 (3.23 %)	1 (3.23 %)	
30	Histiocitosis	1 (3.23 %)	1 (3.23 %)	
31	LAM	5 (16.1 %)	2 (6.45 %)	
32	Peso_kg_	57.5 [49.0;64.8]	59.0 [51.0;71.0]	0.475
33	Altura_cm_	166 [158;171]	163 [160;170]	0.583
34	FVC_pre_cc_	1700 [1255;2152]	1610 [1055;1980]	0.481
35	FVC_pre_	46.0 [32.5;54.5]	41.0 [28.5;47.5]	0.202
36	FEV1_pre_cc_	760 [660;980]	800 [595;1070]	0.418
37	FEV1_pre_	26.0 [22.0;35.0]	25.0 [17.5;29.5]	0.307
38	TIFFENAU_pre	59.0 [47.0;73.0]	52.0 [39.5;68.0]	0.510
39	TLC_pre	111 [85.5;123]	112 [101;127]	0.541
40	RV_pre	230 [137;310]	278 [180;328]	0.336
41	DLCO_pre	16.5 [13.2;23.8]	13.0 [8.75;20.0]	0.262
42	KCO_pre	2.00 [1.00;18.5]	4.00 [1.00;16.5]	0.965
43	PO2_pre	57.0 [53.0;64.5]	61.0 [57.0;68.0]	0.173
44	PCO2_pre	41.0 [37.0;45.5]	43.0 [35.0;47.0]	0.705
45	O2_domiciliario:			0.952
46	No	5 (16.1 %)	6 (20.0 %)	
47	Si	26 (83.9 %)	24 (80.0 %)	
48	WT_pre_m_	312 [288;360]	288 [206;363]	0.165
49	Grupo_ABO:			0.036
50	0	13 (41.9 %)	17 (54.8 %)	

Cuadro 6: Análisis estadístico de las primeras 50 variables

Como se puede observar en la tabla anterior solamente unas pocas variables han resultado significativas al realizar la comparación entre los grupos.

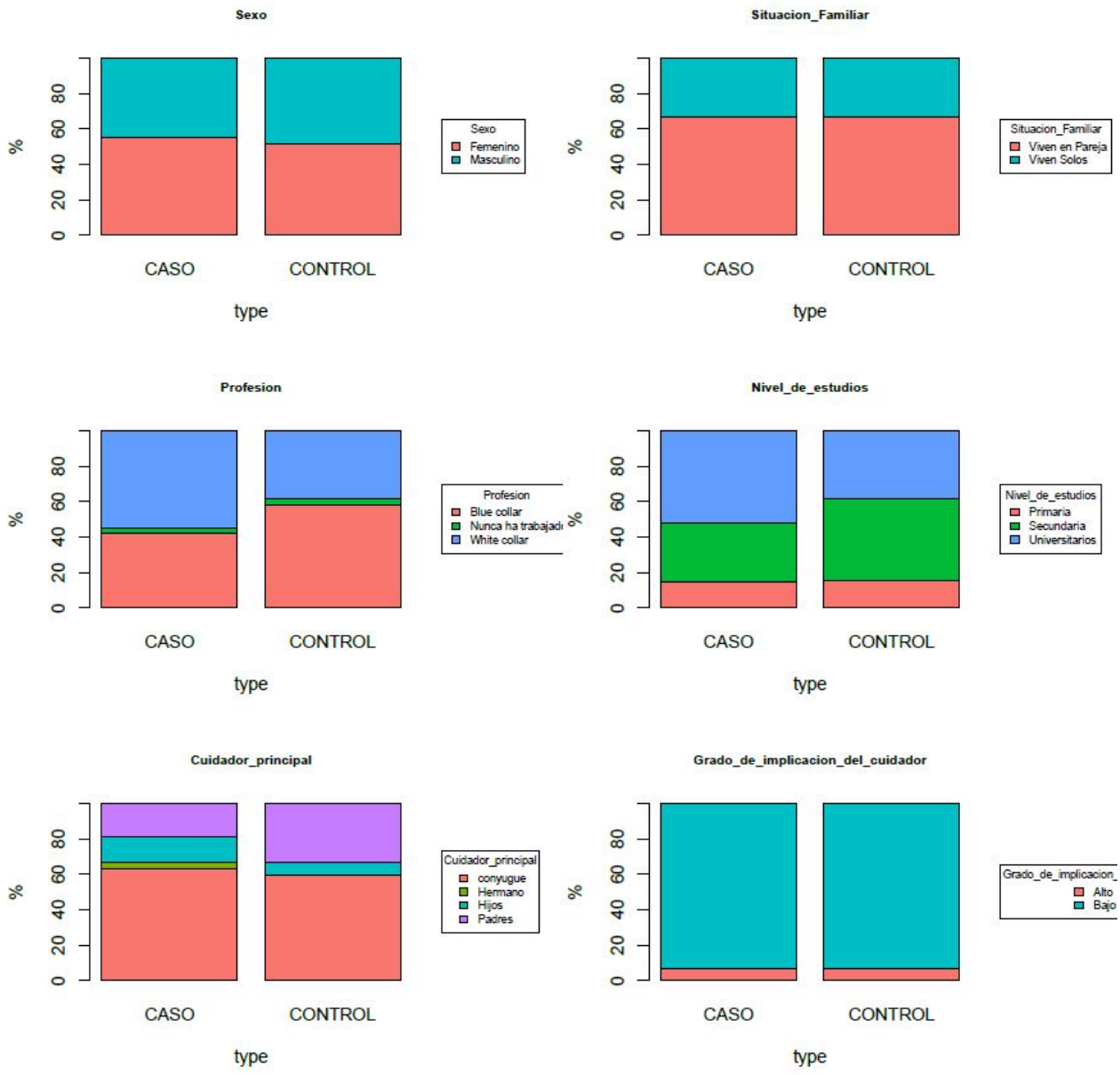


Figura 22: Descriptivo gráfico de las primeras variables analizadas

3.2.4 Datos sobre las poblaciones celulares

En este apartado se va a mostrar como se ha llevado a cabo la selección de las poblaciones celulares analizadas mediante citometría de flujo, para evaluar las que serán incluidas en el análisis de integración. De manera análoga a como se ha realizado en el apartado anterior se hecho un análisis descriptivo de numérico y gráfico de los datos:

	Var	X.ALL....N.62	X	N
1	Grupo:			62
2	CLAD	31 (50.0 %)		
3	LTS	31 (50.0 %)		
4	granulo_leuco	57.8 [50.5	68.0]	62
5	mono_linfo_leuco	42.2 [32.0	49.5]	62
6	mono_leuco	8.75 [6.97	10.9]	62
7	mono_CD14p_CD16p_mono	10.9 [7.64	13.4]	62
8	mono_CD14p_CD16n_mono	89.1 [86.6	92.4]	62
9	mono_CD14high_CD16n_mono	86.4 [82.3	90.6]	62
10	mono_CD14high_CD16p_mono	6.04 [4.68	8.04]	62
11	mono_CD14dim_CD16n_mono	1.95 [1.47	2.68]	60
12	mono_CD14dim_CD16p_mono	4.18 [3.03	5.44]	61
13	linfo_leuco	31.6 [23.3	36.7]	62
14	LB_linfo	4.53 [2.46	7.23]	62
15	NKs_linfo	8.80 [5.08	16.8]	62
16	NK_CD56_high	8.53 [4.71	14.1]	58
17	NK_CD56_dim	91.5 [85.9	95.5]	62
18	NK_CD56p_CD16p	93.8 [89.7	95.8]	61
19	NK_CD56p_CD16n	6.75 [4.56	10.8]	56
20	NK_CD16pCD56dim_NK	88.8 [83.4	93.2]	60
21	NKCD16pCD56high_NK	4.16 [2.34	8.23]	49
22	NKCD16n_CD56dim_NK	2.82 [1.73	4.16]	36
23	NKCD16n_CD56high_NK	4.02 [2.70	7.41]	52
24	LT_linfo	80.5 [73.9	88.3]	62
25	LT_CD4p_LT	55.8 [43.3	61.7]	62
26	LT_CD8p_LT	39.0 [32.7	49.6]	62
27	NKT_LT	8.62 [5.07	15.5]	62
28	sub_NKT_1_CD3p_CD56p_CD4n_CD8n_NKT	73.8 [60.3	79.0]	62
29	sub_NKT_2_CD3p_CD56p_CD4n_CD8p_NKT	14.9 [8.18	27.9]	55
30	sub_NKT_3_CD3p_CD56p_CD4p_CD8n_NKT	12.2 [7.00	21.1]	42
31	NKT_linfo	6.78 [4.00	11.6]	62
32	sub_NKT_1_CD3p_CD56p_CD4n_CD8n_linfo	4.80 [2.79	7.33]	62
33	sub_NKT_2_CD3p_CD56p_CD4n_CD8p_linfo	1.01 [0.42	1.94]	55
34	sub_NKT_3_CD3p_CD56p_CD4p_CD8n_linfo	0.73 [0.40	2.30]	42
35	Leuco	189473 [183779	192397]	62

Cuadro 7: Análisis descriptivo de las poblaciones celulares

En la figura (23) se muestra un fragmento del análisis descriptivo realizado.

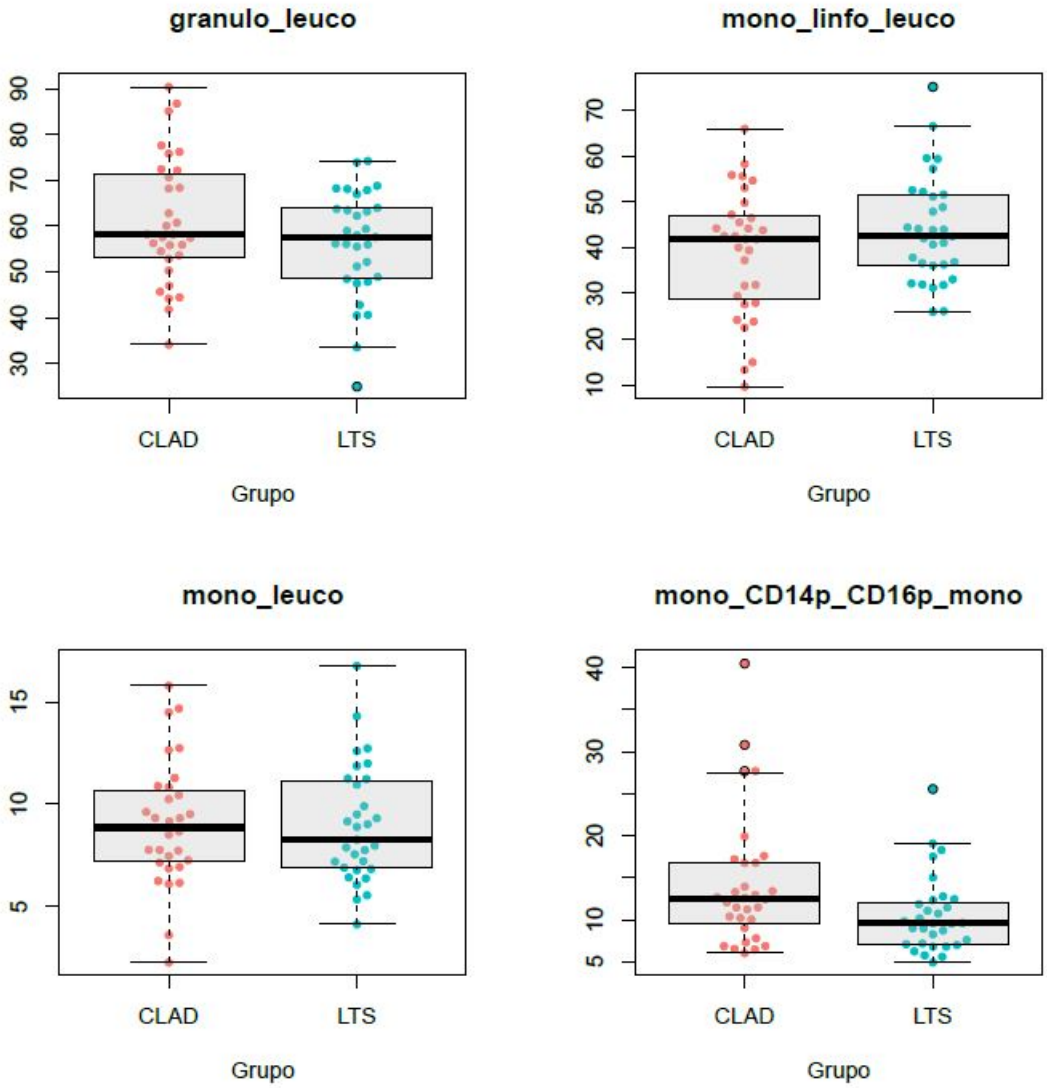


Figura 23: Descriptivo gráfico de las primeras variables analizadas

Se realiza un test de comparación entre grupos para evaluar si existen diferencias significativas entre los valores de cada variable (las proporciones celulares) respeto al grupo de respuesta.En los casos en que las variables son normales se ha empleado un test t de Student, mientras que en el resto se ha empleado la prueba U de Mann-Withney, que no asume normalidad. A continuación se muestran los resultados obtenidos:

	variable	N	p.value	adj.p.value
1	granulo_leuco	62	0.13	0.33
2	mono_linfo_leuco	62	0.13	0.33
3	mono_leuco	62	0.91	0.97
4	mono_CD14p_CD16p_mono	62	0.02	0.16
5	mono_CD14p_CD16n_mono	62	0.02	0.16
6	mono_CD14high_CD16n_mono	62	0.05	0.21
7	mono_CD14high_CD16p_mono	62	0.01	0.14
8	mono_CD14dim_CD16n_mono	60	0.95	0.98
9	mono_CD14dim_CD16p_mono	61	0.98	0.98
10	linfo_leuco	62	0.09	0.32
11	LB_linfo	62	0.68	0.80

12	NKs_linfo	62	0.53	0.74
13	NK_CD56_high	58	0.61	0.78
14	NK_CD56_dim	62	0.23	0.43
15	NK_CD56p_CD16p	61	0.00	0.12
16	NK_CD56p_CD16n	56	0.02	0.16
17	NK_CD16pCD56dim_NK	60	0.19	0.37
18	NKCD16pCD56high_NK	49	0.30	0.48
19	NKCD16n_CD56dim_NK	36	0.72	0.82
20	NKCD16n_CD56high_NK	52	0.05	0.21
21	LT_linfo	62	0.38	0.58
22	LT_CD4p_LT	62	0.42	0.61
23	LT_CD8p_LT	62	0.13	0.33
24	NKT_LT	62	0.10	0.32
25	sub_NKT_1_CD3p_CD56p_CD4n_CD8n_NKT	62	0.64	0.78
26	sub_NKT_2_CD3p_CD56p_CD4n_CD8p_NKT	55	0.85	0.94
27	sub_NKT_3_CD3p_CD56p_CD4p_CD8n_NKT	42	0.63	0.78
28	NKT_linfo	62	0.05	0.21
29	sub_NKT_1_CD3p_CD56p_CD4n_CD8n_linfo	62	0.16	0.37
30	sub_NKT_2_CD3p_CD56p_CD4n_CD8p_linfo	55	0.30	0.48
31	sub_NKT_3_CD3p_CD56p_CD4p_CD8n_linfo	42	0.24	0.43
32	Leuco	62	0.17	0.37

Cuadro 8: Análisis estadístico de las poblaciones celulares

3.4 Resultados del análisis de integración

Una vez ya se han preparado los datos de cada ómica por separado, el siguiente paso es integrarlos todos para ver si con la información conjunta que proporcionan todos ellos, es más fácil clasificar a los pacientes entre los dos grupos. A continuación se va ir describiendo paso a paso el análisis realizado, a la vez que se va mostrando el código utilizado.

El primer paso es cargar las librerías necesarias para realizar el análisis.

```
library(mixOmics)
```

A continuación se cargan los datos de cada ómica que se va a utilizar. Primero se cargan los **genes** diferencialmente expresados y se muestran las primeras filas y columnas:

```
genes <- read.csv2(file.path("ClariomD_TopTab.csv"), sep = ";", header = TRUE, dec = ".")
```

Affymetrix.ID	Gene.Symbol	Entrez	logFC	AveExpr	t	P.Value	adj.P.Val
TC0400007933.hg.1	ANXA3	306	-0.96	5.66	-6.20	0.00	0.00
TC0100009656.hg.1	FCGR1B	2210	-1.04	7.56	-5.76	0.00	0.00
TC1900012043.hg.1	LILRA4	23547	0.63	4.20	5.33	0.00	0.00
TC0600011960.hg.1	TNFRSF21	27242	0.47	5.55	5.28	0.00	0.00
TC1700011600.hg.1	KCNJ2-AS1	400617	-0.61	4.74	-5.20	0.00	0.00
TC0400012621.hg.1	ACSL1	2180	-0.83	8.20	-5.19	0.00	0.00
TC0800011861.hg.1	LRRC6	23639	-0.67	4.50	-5.17	0.00	0.00
TC1000010273.hg.1	NRP1	8829	0.58	4.14	5.06	0.00	0.00
TC2100008510.hg.1	KCNJ15	3772	-0.82	9.78	-5.04	0.00	0.00
TC0100009697.hg.1	FCGR1CP	100132417	-0.82	7.57	-4.96	0.00	0.00
TC1700010221.hg.1	DHRS13	147015	-0.55	6.78	-4.95	0.00	0.00
TC0700009411.hg.1	MGAM2	93432	-0.94	5.51	-4.94	0.00	0.00
TC1100012126.hg.1	MMP8	4317	-1.89	7.55	-4.89	0.00	0.00
TC0300013471.hg.1	BCL6	604	-0.65	10.79	-4.83	0.00	0.00
TC0500008627.hg.1	SLC22A4	6583	-0.57	6.03	-4.77	0.00	0.01
TC0200008005.hg.1	DYSF	8291	-0.79	8.76	-4.60	0.00	0.01
TC1300009598.hg.1	GPR183	1880	0.79	8.34	4.59	0.00	0.01
TC0300011517.hg.1	PROK2	60675	-0.60	10.06	-4.52	0.00	0.01
TC1100006492.hg.1	PNPLA2	57104	-0.38	7.37	-4.49	0.00	0.01
TC1400008460.hg.1	PLD4	122618	0.44	5.99	4.39	0.00	0.02

Cuadro 9: toptable del análisis de genes

A continuación se cargan los **miRNA** y de la misma que se ha hecho antes se muestran las primeras filas y columnas:

```
mirna <- read.csv2(file.path("mirna_expr.csv"), sep = ";", dec = ",", header = TRUE)
```

X	Probe.Set.Name	Symbol	logFC	AveExpr	t	P.Value	adj.P.Val
1	14q0_st	14q0	0.05	0.67	1.27	0.21	0.88
2	14qI-1_st	14qI-1	0.06	1.76	0.59	0.56	0.94
3	14qI-1_x_st	14qI-1	0.09	1.18	0.96	0.34	0.92
4	14qI-2_st	14qI-2	0.02	0.73	0.27	0.79	0.98
5	14qI-3_x_st	14qI-3	0.04	0.68	0.83	0.41	0.92
6	14qI-4_st	14qI-4	-0.10	0.81	-1.80	0.08	0.75
7	14qI-4_x_st	14qI-4	-0.07	0.82	-1.37	0.18	0.85
8	14qI-5_st	14qI-5	0.10	0.72	2.27	0.03	0.65
9	14qI-6_st	14qI-6	0.01	0.72	0.17	0.86	0.99
10	14qI-7_st	14qI-7	-0.09	1.49	-1.38	0.17	0.85
11	14qI-8_st	14qI-8	0.06	0.83	1.25	0.22	0.89
12	14qI-8_x_st	14qI-8	0.06	0.81	1.23	0.22	0.89
13	14qI-9_x_st	14qI-9	0.04	0.87	0.78	0.44	0.92
14	14qII-10_st	14qII-10	0.00	0.79	0.03	0.97	1.00
15	14qII-11_st	14qII-11	0.12	0.75	2.40	0.02	0.62
16	14qII-12_st	14qII-12	-0.11	0.96	-1.59	0.12	0.80
17	14qII-12_x_st	14qII-12	-0.13	0.80	-2.49	0.02	0.62
18	14qII-13_st	14qII-13	0.10	0.83	1.90	0.06	0.74
19	14qII-14_st	14qII-14	-0.00	0.78	-0.07	0.94	0.99
20	14qII-14_x_st	14qII-14	0.01	1.01	0.20	0.84	0.98

Cuadro 10: toptable del análisis de miRNA

Ahora se cargan los datos de las **variables clínicas**:

```
clinical <- read.csv2(file.path("ClinicalData_1_clean.csv"), header = TRUE,
                      dec = ",", sep = ";")

clinical <- clinical[order(clinical$Row.names),]
rownames(clinical) <- clinical[, 1]
clinical <- clinical[, -c(1, 11)]

#se eliminan las filas que no están en las ómicas
clinical.sel <- clinical[-c(2, 3, 5),]
#se eliminan aquellas columnas no numéricas
clinical.sel <- clinical.sel[, -c(1, 5:9)]
```

	Edad_donante	Creat_1A	Col_FIN	Leucos_FIN	N.Infec.1095	FVCcc	FVC.	FEV1cc	FEV1.	X25.75.
P01	26	1.10	197	6100.00	14	5180	111	3490	96	46
P04	17	1.03	192	5500.00	0	3250	88	2800	103	103
P06	32	0.95	265	7000.00	4	2380	89	1680	88	42
P07	26	0.60			5	2190	49	1570	43	22
P08	24	0.97	192	5700.00	0	4120	111	2900	112	77
P09	20	0.59			9	2370	66	1520	53	23
P10	22	0.85	197	9600.00	8	3110	91	2360	91	60
P11	28	0.60			1	3600	81	2600	72	36
P12	7	1.20	176	6400.00	1	4280	79	3730	87	96
P13	11	1.10	180	10500.00	3	2440	64	1510	51	25
P14	14	2.20	218	6000.00	9	2320	76	1940	84	64
P15	23	1.00	176	5300.00	1	5720	142	4680	155	136
P16	34	1.10	220	9500.00	7	2910	64	1560	48	21
P17	34	0.90	203	5390.00	6	2930	100	2810	130	143
P18	12	0.90	159	12100.00	4	2280	45	930	23	8

Cuadro 11: Variables clínicas seleccionadas

A continuación se muestran todas las variables clínicas que se van a utilizar en el análisis:

```
##  [1] "Edad_donante" "Creat_1A"     "Col_FIN"      "Leucos_FIN"
##  [5] "N.Infec.1095" "FVCcc"        "FVC."         "FEV1cc"
##  [9] "FEV1."        "X25.75."      "TIFFENAU"     "DLCO.."
## [13] "KCO.."        "RV"
```

A continuación se cargan los datos de las **poblaciones celulares**:

	mono_CD14p_CD16p_mono	mono_CD14p_CD16n_mono	mono_CD14high_CD16n_mono	mono_CD14high_CD16p_mono
P01	8.94	91.06	89.62	4.02
P04	11.06	88.94	85.89	4.97
P06	5.76	94.24	92.71	4.14
P07	16.74	83.26	81.90	6.17
P08	12.78	87.22	80.41	5.16
P09	9.02	90.98	86.85	5.13
P10	8.72	91.28	88.43	3.68
P11	10.03	89.97	85.17	5.01
P12	18.30	81.70	79.70	10.37
P13	7.31	92.69	90.97	5.33
P14	9.53	90.47	89.53	6.88
P15	5.59	94.41	91.81	3.51
P16	16.76	83.24	82.13	13.29
P17	9.82	90.18	89.13	6.31
P18	30.80	69.20	65.97	26.31

Cuadro 12: Poblaciones celulares seleccionadas

Finalmente se carga en un archivo la clasificación de cada paciente a que condición experimental pertenece:

```r
grupos <- read.csv2(file.path("grupos.csv"), sep = ";", header = TRUE)


Y <- grupos$Group
summary(Y)
```

```
## CASE  CTL
##   30   29
```

Vemos que contamos con 30 individuos para la condición **CASE** y de 29 para la condición **CTL**. Los nombres de las muestras y de los pacientes han de ser exactamente los mismos en todas las ómicas que se pretendan integrar.

	X	Group
1	P01	CASE
2	P04	CASE
3	P06	CASE
4	P07	CTL
5	P08	CASE
6	P09	CTL
7	P10	CASE
8	P11	CTL
9	P12	CASE
10	P13	CTL
11	P14	CASE
12	P15	CASE
13	P16	CTL
14	P17	CASE
15	P18	CTL
16	P19	CASE
17	P20	CTL
18	P21	CASE
19	P22	CTL
20	P23	CASE
21	P24	CTL
22	P25	CASE
23	P26	CTL
24	P27	CASE
25	P28	CTL
26	P29	CASE
27	P30	CTL
28	P31	CASE
29	P32	CASE
30	P33	CASE
31	P34	CTL
32	P35	CTL
33	P36	CASE
34	P37	CTL
35	P38	CASE
36	P39	CTL
37	P40	CASE
38	P41	CTL
39	P42	CTL
40	P43	CASE
41	P44	CTL
42	P45	CASE
43	P46	CASE
44	P47	CASE
45	P48	CTL
46	P49	CTL
47	P50	CASE
48	P51	CTL
49	P52	CASE
50	P53	CTL
51	P54	CTL
52	P55	CASE
53	P56	CASE
54	P57	CTL
55	P58	CTL
56	P59	CASE
57	P60	CTL
58	P61	CTL
59	P62	CTL

Cuadro 13: Correspondencia entre los pacientes y la condición experimental a la que pertenecen

A continuación se procede a seleccionar los genes y miRNA que se van a utilizar en el análisis de integración.

No se pueden utilizar todos ya que sino se estaría introduciendo mucho ruido en el análisis, además de ralentizar de forma considerable todos los cálculos. Tanto en el caso de los mRNA como de los miRNA se ha elegido el mismo punto de corte: se utilizarán todos aquellos genes que tengan un **pvalor < 0.01**. Se seleccionan primero los genes:

```
gen.sel <- subset(genes, P.Value < 0.01)
#se quita la muestra P05 que en los miRNA no está
gen.sel <- gen.sel[, c(2, 11, 12, 14:70)]
rownames(gen.sel) <- gen.sel$Gene.Symbol
gen.sel <- gen.sel[, -1]
```

A continuación se seleccionan los miRNA:

```
mirna.sel <- subset(mirna, P.Value < 0.01)
mirna.sel <- mirna.sel[, c(2, 11:69)]
rownames(mirna.sel) <- mirna.sel$Probe.Set.Name
mirna.sel <- mirna.sel[, -1]
```

Como ya se ha comentado es indispensable que las muestras en todos los conjuntos de datos que se vayan a utilizar se llamen de la misma manera. Se comprueba:

```
colnames(gen.sel) %in% colnames(mirna.sel)
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [29] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [43] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [57] TRUE TRUE TRUE
```

```
colnames(gen.sel) %in% rownames(clinical.sel)
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [29] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [43] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [57] TRUE TRUE TRUE
```

```
rownames(clinical.sel) %in% rownames(cell.sel)
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [29] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```
## [43] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [57] TRUE TRUE TRUE
```

El método de DIABLO necesita que al contrario de lo que sucede con los análisis ómicos habituales, donde las muestras están en las columnas y las variables en las filas, las muestras estén en las filas y las variables en las columnas. Por ello hay que transponer todas las matrices de datos.

```
gen.selt <- t(gen.sel)
mirna.selt <- t(mirna.sel)
```

Debido a que no ha sido posible encontrar otro conjunto de datos donde probar el pipeline creado, se ha decidido dividir el conjunto de datos que tenemos en dos partes: una se utilizará para entrenar al algoritmo de clasificación (un 67 % de las muestras), y la otra se utilizará para comprobar como ha funcionado (el 33 % restante). El conjunto de datos que se utilizará como entrenamiento se llamará utilizando el prefijo *train*, y el conjunto de datos que se utilice para comprobar el funcionamiento se llamará utilizando el prefijo *test*.

```
set.seed(123)
ind <- sample(nrow(grupos), 0.67*dim(grupos)[1])

train.gen.selt <- gen.selt[ind, ]
test.gen.selt <- gen.selt[-ind, ]
train.mirna.selt <- mirna.selt[ind, ]
test.mirna.selt <- mirna.selt[-ind, ]
train.clinical.sel <- clinical.sel[ind, ]
test.clinical.sel <- clinical.sel[-ind, ]
train.cell.sel <- cell.sel[ind, ]
test.cell.sel <- cell.sel[-ind, ]
train.grupos <- grupos[ind, ]
test.grupos <- grupos[-ind, ]
```

Se comprueba que hayamos escogido las mismas muestras en todos los conjuntos de datos:

```
rownames(train.gen.selt) %in% rownames(train.mirna.selt)
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [29] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```
rownames(train.cell.sel) %in% rownames(train.mirna.selt)
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```
## [29] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```
rownames(train.gen.selt) %in% rownames(train.clinical.sel)
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [29] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```
train.grupos$X %in% rownames(train.cell.sel)
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [29] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

Una vez ya están todos los datos preparados, se puede empezar con el protocolo de integración de datos propiamente dicho. Cargamos todos los conjunto de datos en un solo objeto, creando una lista. En otra variable guardamos a qué grupo pertenece cada paciente:

```
X <- list(mRNA = train.gen.selt,
          miRNA = train.mirna.selt,
          cell = train.cell.sel,
          clinical = train.clinical.sel)

Y <- train.grupos$Group
```

A continuación se elige, de manera arbitraria el número de variables que queremos mantener de cada tipo y en cada componente. En este caso se han elegido según la cantidad de variables en el conjunto de datos de partida.

```
list.keepX <- list(mRNA = c(16, 17), miRNA = c(18,5), cell = c(5, 5), clinical = c(6 ,6))
```

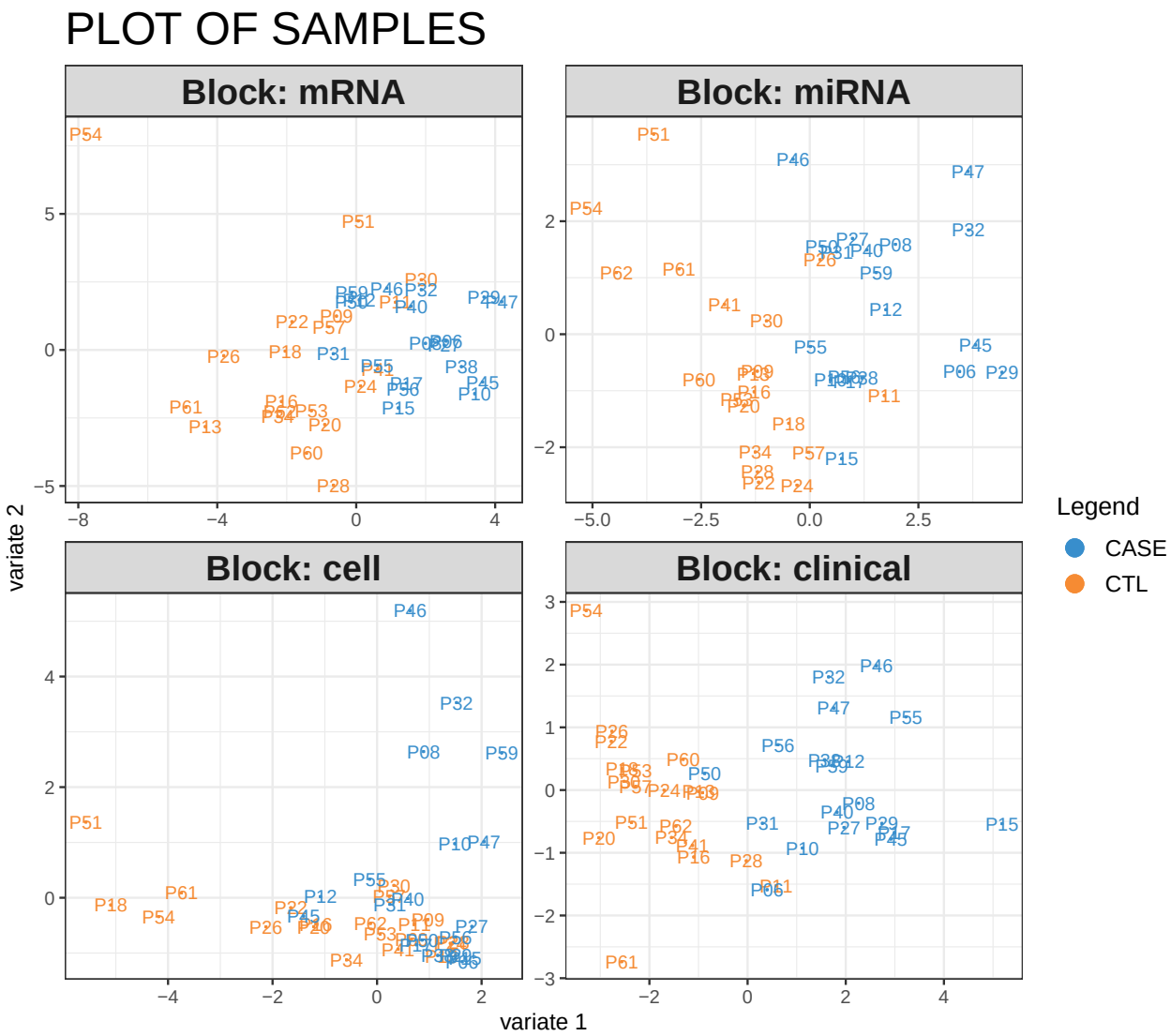
Ahora ya se puede aplicar el método de sPLS-DA implementado en DIABLO. Se elige mantener dos componentes de cada ómica analizada, escalar los datos y se elige el modo de *regresión*:

```
MyResult.diablo <- block.splsda(X, Y, keepX = list.keepX, ncomp = 2,
                                scale = TRUE, mode = "regression")
```

A continuación ya se pueden visualizar los primeros resultados. En este caso las gráficas para ver como se agrupan las muestras:

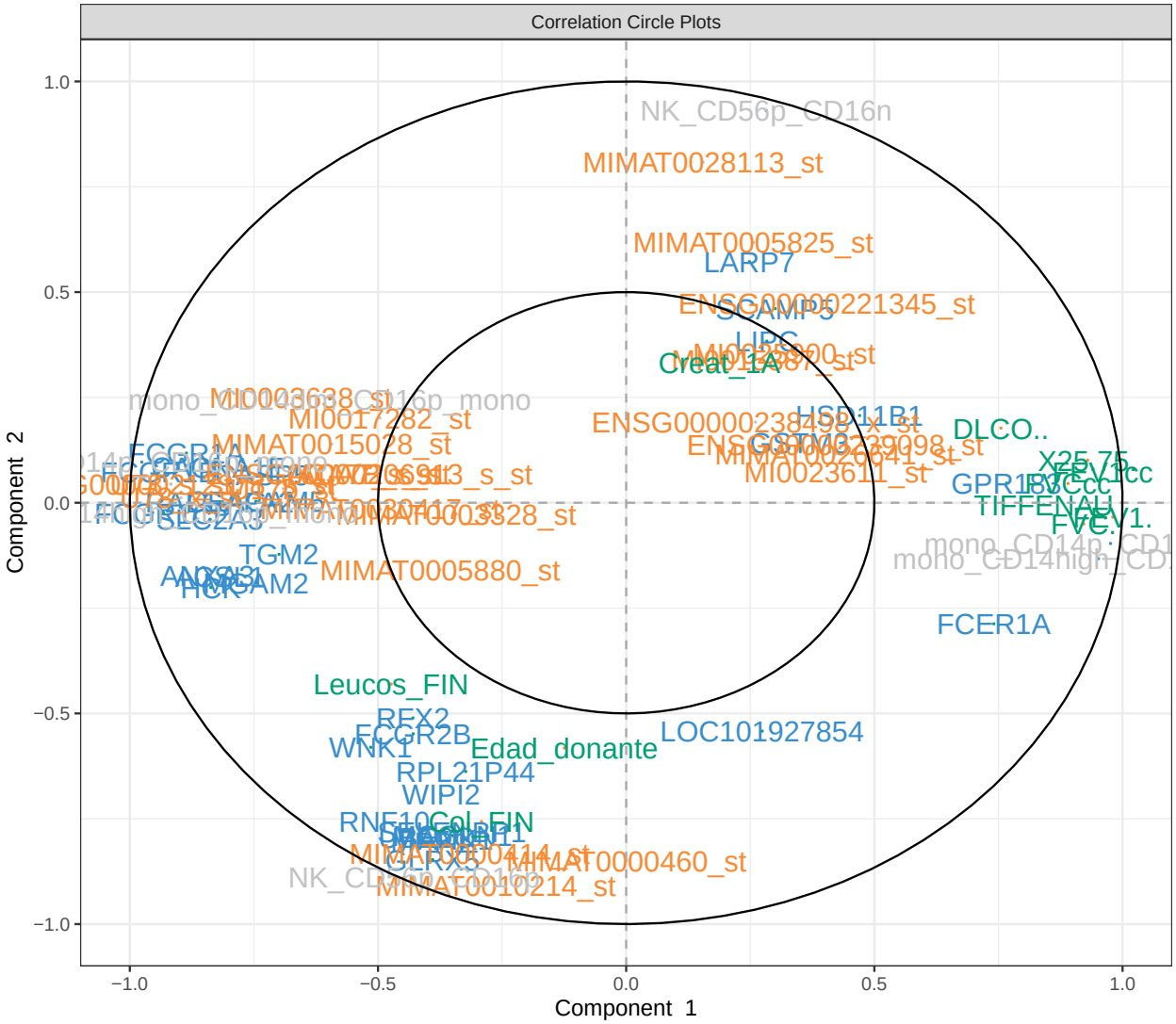
```
plotIndiv(MyResult.diablo,
            ind.names = TRUE,
            legend = TRUE, cex=c(3, 3),
```

```
            title = 'PLOT OF SAMPLES')
```



Se observa en la gráfica anterior como se agrupan las muestras en cada ómica estudiada después de haber realizado el primer análisis. En los cuatro tipos de datos estudiados se observa una relativa buena separación de las muestras según si pertenecen al grupo *CASE* o *CTL*. Dónde mejor se observa esta separación es con los datos de origen clínico, donde casi ni se observa ninguna muestra que se clasifica con el otro grupo. El siguiente grupo donde mejor se clasifican es en los miRNA, y donde parece que se clasifican peor es en los datos de las poblaciones celulares. A continuación se muestran los datos para ver como se agrupan las variables:

```
plotVar(MyResult.diablo,
          var.names = c(TRUE, TRUE, TRUE, TRUE),
          pch = c(5, 5, 5, 5))
```



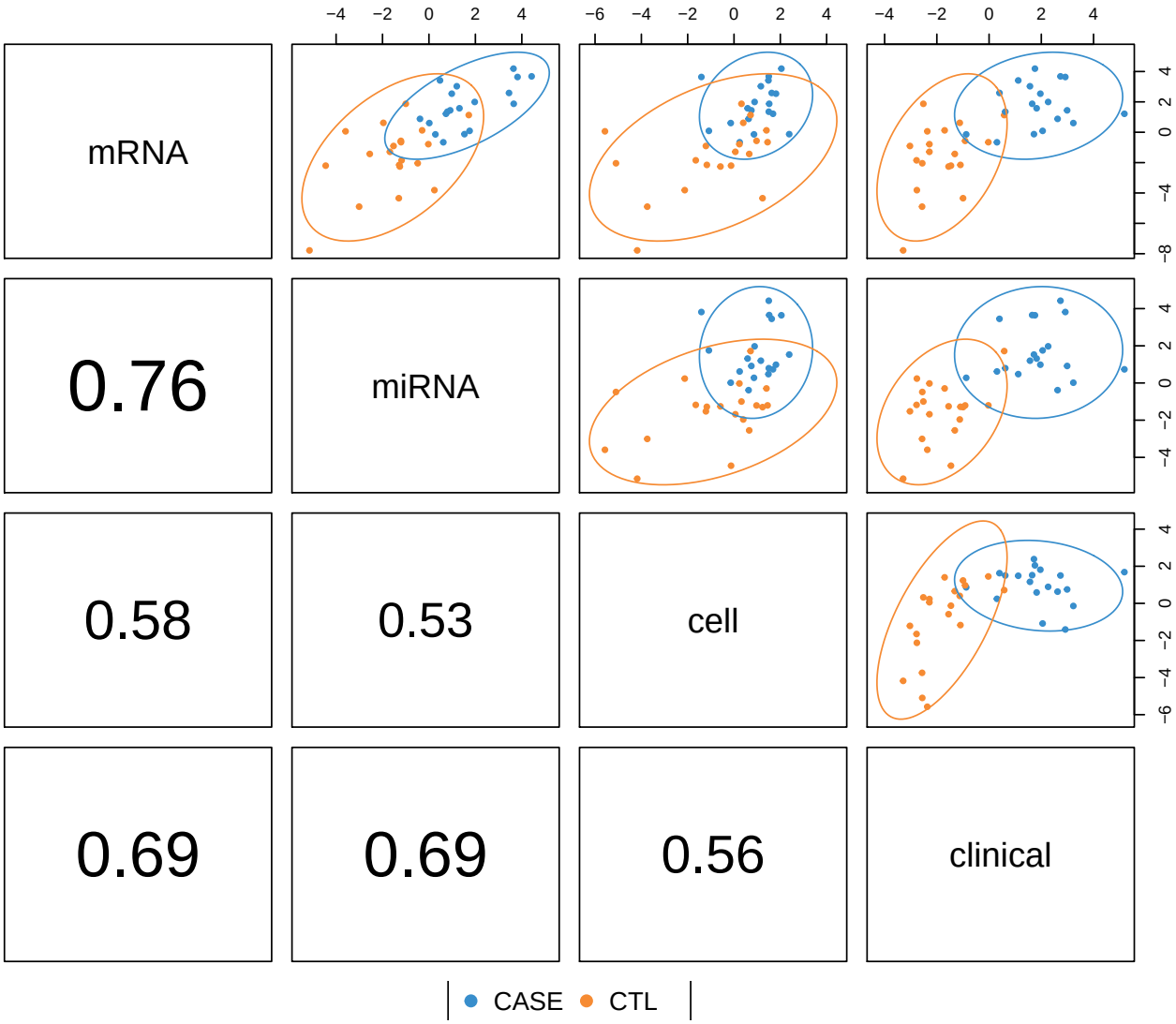
En este caso se ha generado un círculo de correlación donde aquellas variables que más alejadas estén del centro y más cerca del exterior, serán aquellas que más correlacionadas estarán. Las que estén más cerca del centro del círculo serán aquellas que estarán menos correlacionadas. En naranja se muestran los miRNA, en gris se muestran las poblaciones celulares, en azul se muestran los genes y en verde se muestran las variables clínicas. Se observan diferentes grupos de variables bastante correlacionadas en diferentes puntos del gráfico:

- en la parte derecha, hay un grupo de variables de origen principalmente clínico, celular y génico que correlacionan bastante bien entre ellas.
- en la parte inferior izquierda, hay otro grupo de variables que esta vez cuenta con componentes de cada tipo de dato analizado, aunque principalmente contiene miRNAs.
- en la parte izquierda del gráfico hay otro grupo de variables con alta correlación entre ellas, formado

principalmente por genes y alguna población celular.

A continuación se ha realizado un gráfico que muestra la correlación por parejas entre cada tipo de datos analizado:

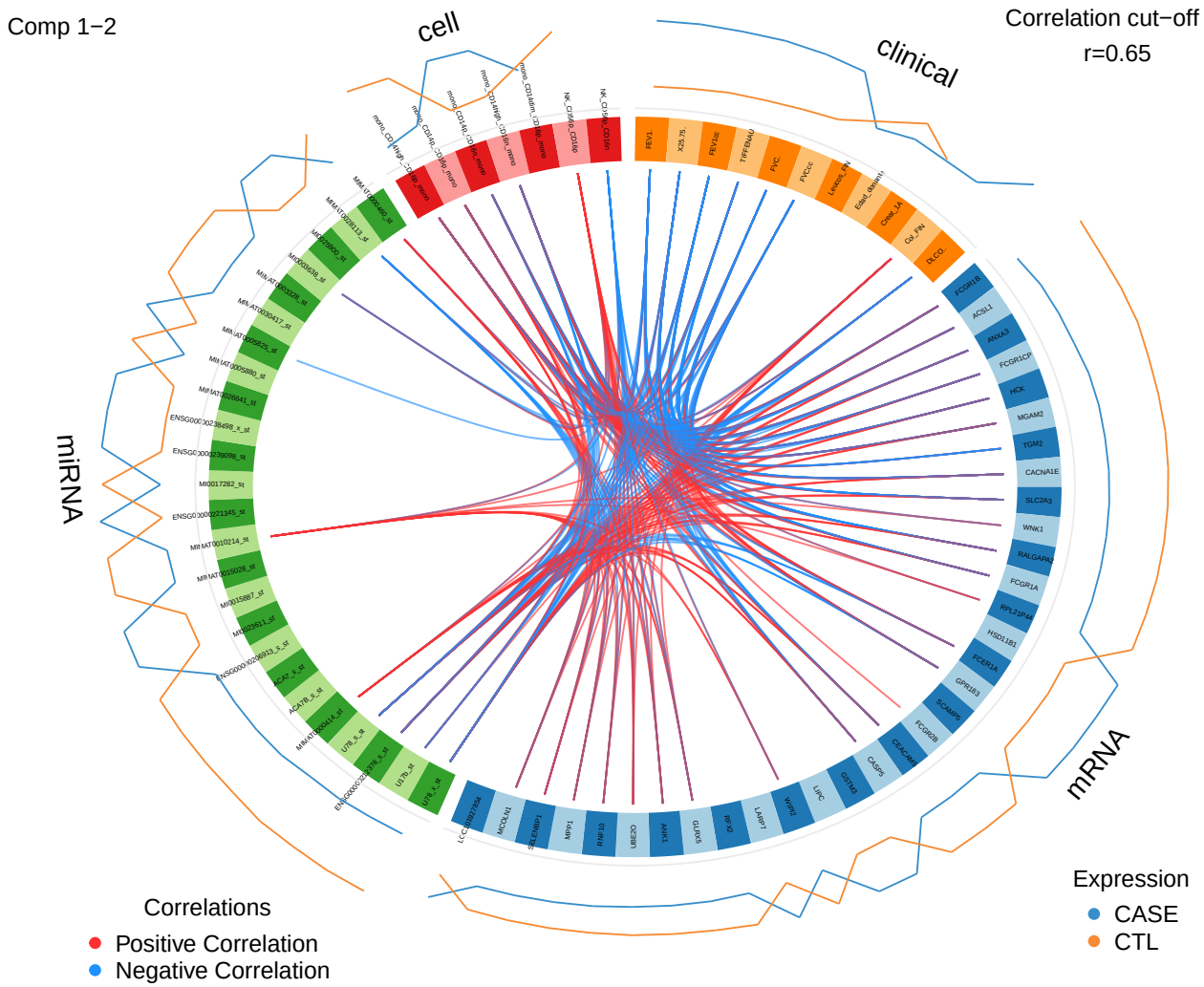
```
plotDiablo(MyResult.diablo, ncomp = 1)
```



Como se puede observar en el gráfico anterior, la mayor correlación se observa entre los mRNA y los miRNA con un valor de 0.76, seguida por la correlación entre el mRNA o el miRNA con las variables clínicas. En ambos casos, la correlación es de 0.69.

Otra manera más gráfica si cabe de mirar la correlación y la relación entre los diferentes datos analizados es utilizar los conocidos "circusplot". En este caso se ha elegido un nivel de corte de 0.65, para que solamente nos enseñe aquellas correlaciones que superan ese valor:

```
circosPlot(MyResult.diablo, cutoff = 0.65)
```

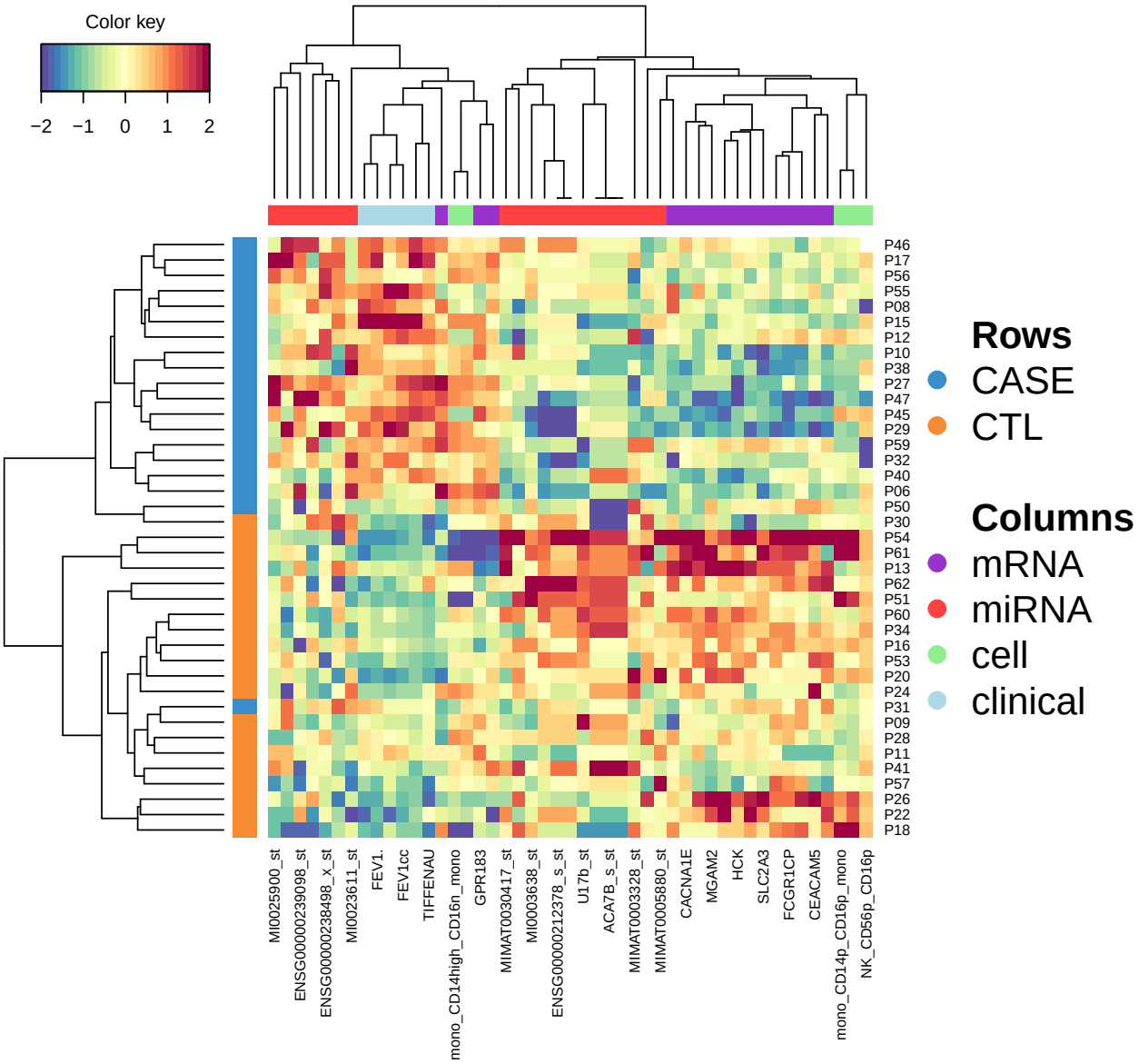


En rojo se nos muestran las correlaciones positivas y en azul las correlaciones negativas. Este gráfico nos relaciona una a una cada variables analizada con las de los otros conjuntos de datos (hay que tener en cuenta que solamente nos muestra datos para aquellas correlaciones que superen en nivel de corte establecido) y nos indica si esas correlaciones son positivas o negativas.

Otro gráfico que se puede generar es el "Clustered Image Maps" (CIM). Este es un gráfico muy similar a los heatmaps habituales, donde se muestra el valor de cada variable escalado entre +2 y -2 en el centro del gráfico, y tanto en la parte superior (columnas), como en la parte izquierda (filas) se ha hecho un clúster jerárquico.

```
cimDiablo(MyResult.diablo,
            color.blocks = c('darkorchid', 'brown1', 'lightgreen', "lightblue"),
            comp = 1,
```

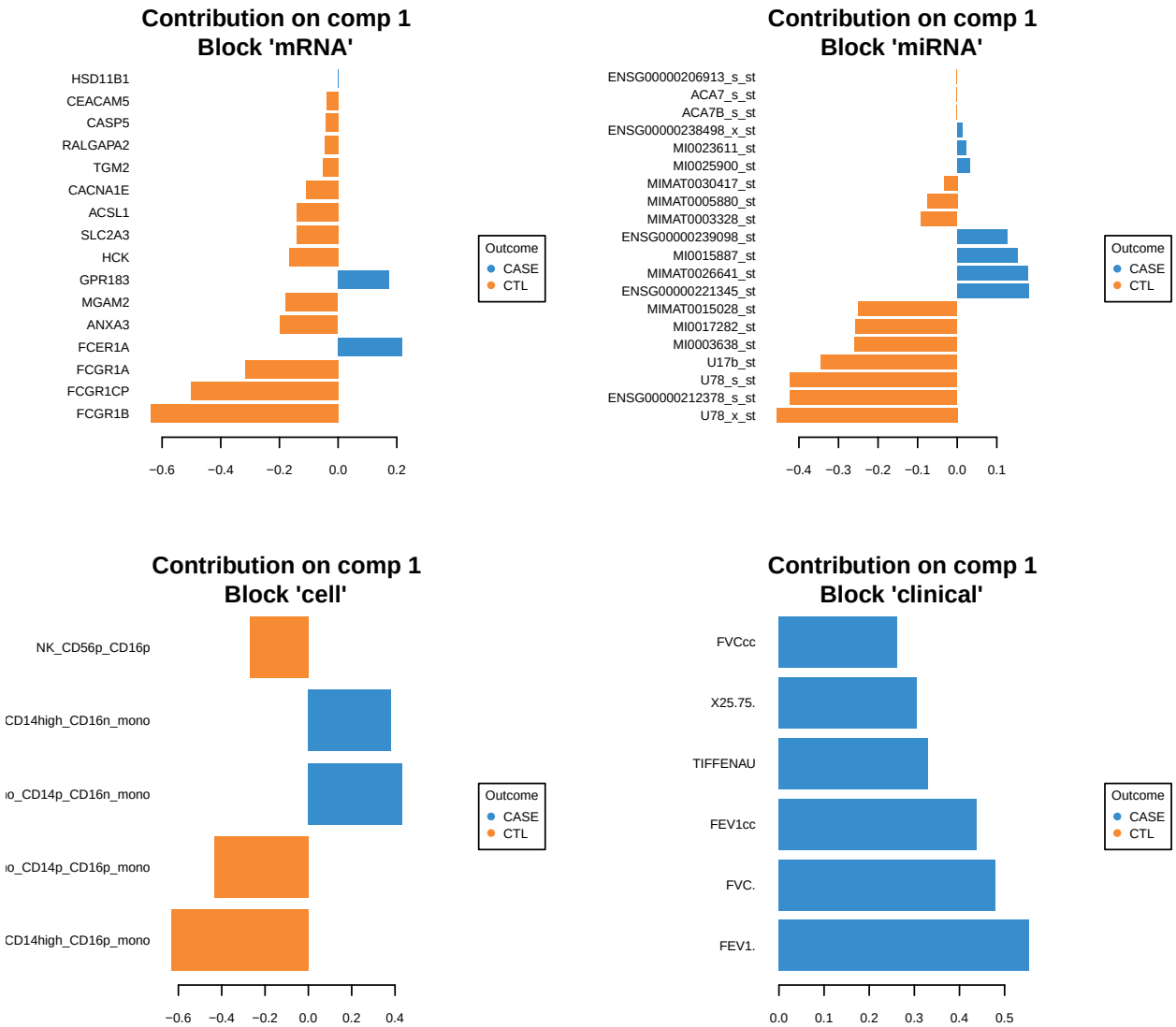
```
            margin = c(10,15),
            legend.position = "right")
```



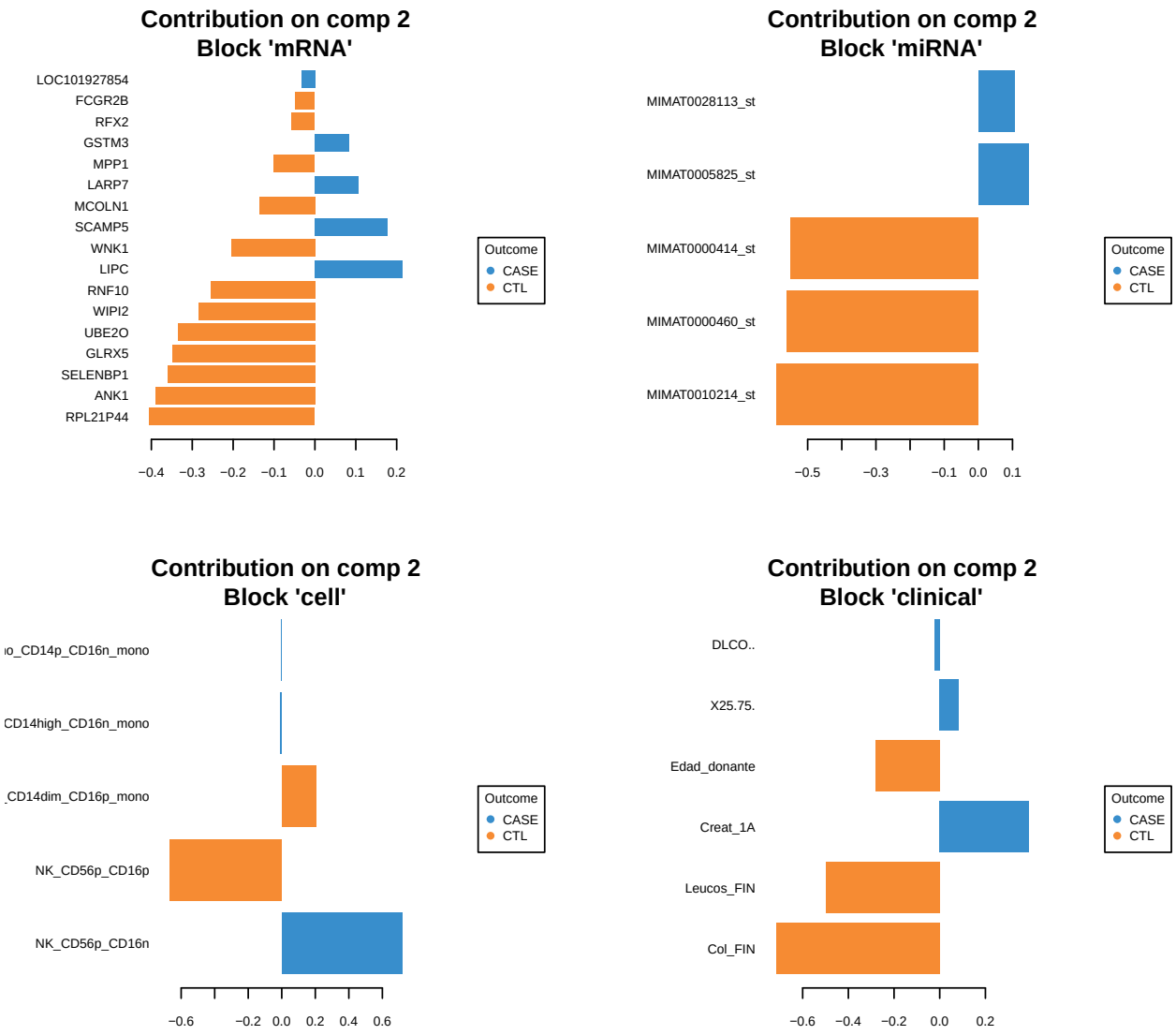
Se observa como el clúster jerárquico de las filas ha clasificado casi perfectamente a las muestras según si pertenecen al grupo *CASE* o *CTL*, a excepción del paciente *P31* que se ha clasificado con el grupo contrario. En las columnas se observa una agrupación de las variables relativamente buena, aunque hay varias que se mezclan con datos de otro tipo.

Otro gráfico que puede tener interés es el ver qué variables en cada conjunto de datos utilizado, son las que tienen más importancia en los dos componentes analizados:

```r
plotLoadings(MyResult.diablo, comp = 1, contrib = "max")
```




```r
plotLoadings(MyResult.diablo, comp = 2, contrib = "max")
```



Otro gráfico interesante es una red que se ha formado a partir de la matriz de similaridades entre los cuatro conjunto de datos utilizados. Se han utilizado diferentes niveles de corte para ver como varía esta red que se forma:

```
network(MyResult.diablo, blocks = c(1, 2, 3, 4),
        color.node = c('darkorchid', 'brown1', 'lightgreen', "lightblue"),
        cutoff = 0.95,
        save = 'pdf',
        name.save = file.path('CorNetwork95'))
```

En la figura 24 se muestra la red construida con un nivel de corte de 0.75. Se observan dos redes separadas.La de la derecha está formada por componentes de los cuatro tipos de datos, pero de las variables clínicas, solamente se observa una. Los otros tres conjuntos de datos están mucho más representados. En la red de la izquierda se observan los cuatro tipos de datos, pero parece que hay menos componentes de los miRNA. De

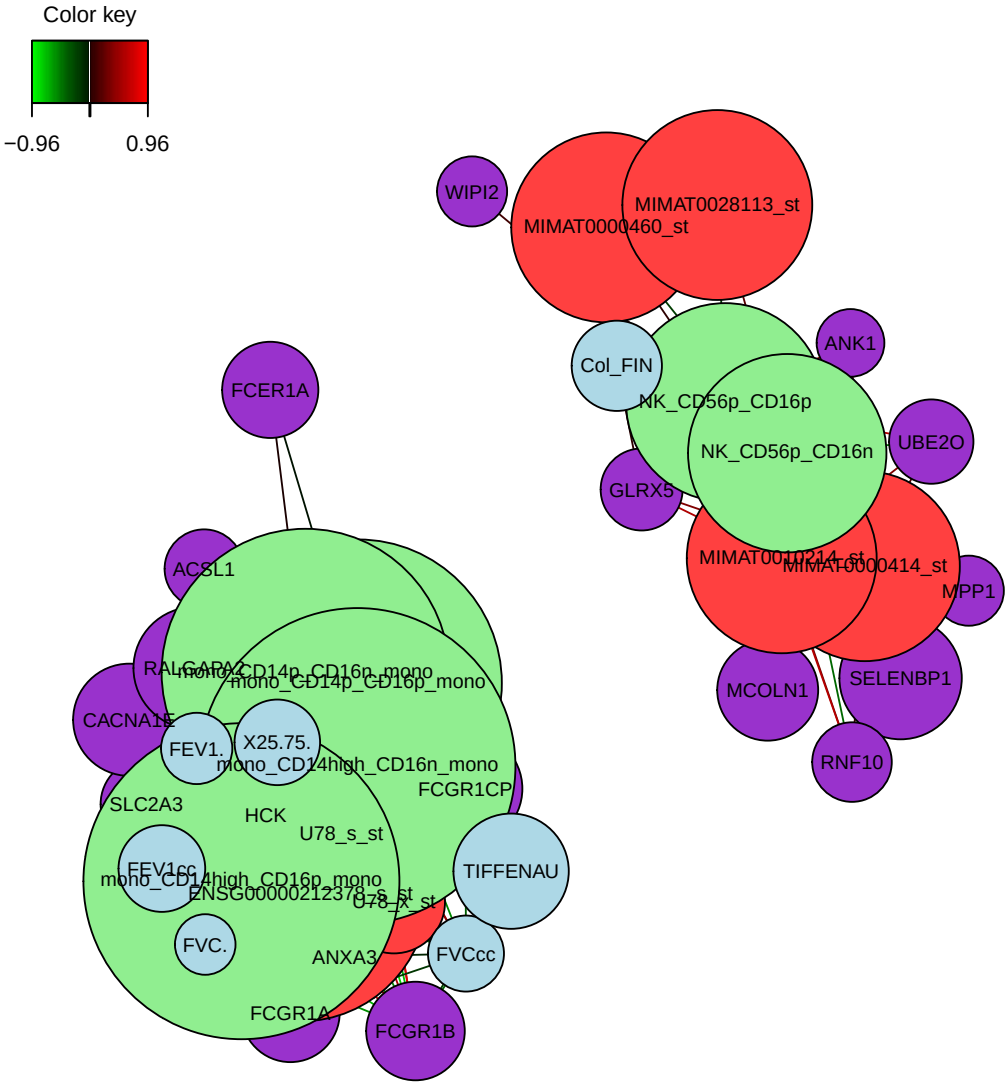


Figura 24: Red que muestra las relaciones entre los diferentes conjuntos de datos. Corte = 0.75

todas formas no se pueden sacar muchas conclusiones, ya que el nivel de corte es el más relajado y aparecen demasiados componentes para hacer las redes.

En la figura 25 se muestra la red construida con un nivel de corte de 0.85. Se siguen observando dos redes independientes entre sí. La de la derecha ya no tiene ningún componente del tipo *clínico*, mientras que en la de la izquierda sí que se observan componentes de los cuatro tipos de datos.

Finalmente, en la figura 26 se muestra la red construida con un nivel de corte de 0.95, el más restrictivo de todos. Solamente han pasado este punto de corte tres componentes de la red que se había formado a la izquierda, una variable clínica y dos variables de las poblaciones celulares.

Una vez ya se han visto algunos de los resultados posibles que se pueden obtener con el método DIABLO, es momento de mirar como clasifica el conjunto de datos que se ha preparado al inicio para evaluarlo:

```
X.test <- list(mRNA = test.gen.selt,
               miRNA = test.mirna.selt,
               cell = test.cell.sel,
               clinical = test.clinical.sel)
Mypredict.diablo <- predict(MyResult.diablo, newdata = X.test)
```

Se puede crear una matriz de confusión para analizar mejor los resultados:

```
confusion.mat <- get.confusion_matrix(
    truth = test.grupos$Group,
  predicted = Mypredict.diablo$MajorityVote$centroids.dist[,2])

confusion.mat
```

```
##      predicted.as.CASE predicted.as.CTL predicted.as.NA
## CASE                 9                1               1
## CTL                  1                5               3
```

Tal como se observa en la matriz de confusión anterior, de los 11 casos que eran de la condición *CASE*, **9** los ha clasificado bien, **1** lo ha clasificado mal como *CTL*, y **1** no lo ha sabido clasificar. En el caso de los *CTL*, de los **9** que había, **5** los ha clasificado bien, **1** lo ha clasificado mal en el grupo de los *CASE*, y a **3** no los ha sabido clasificar.

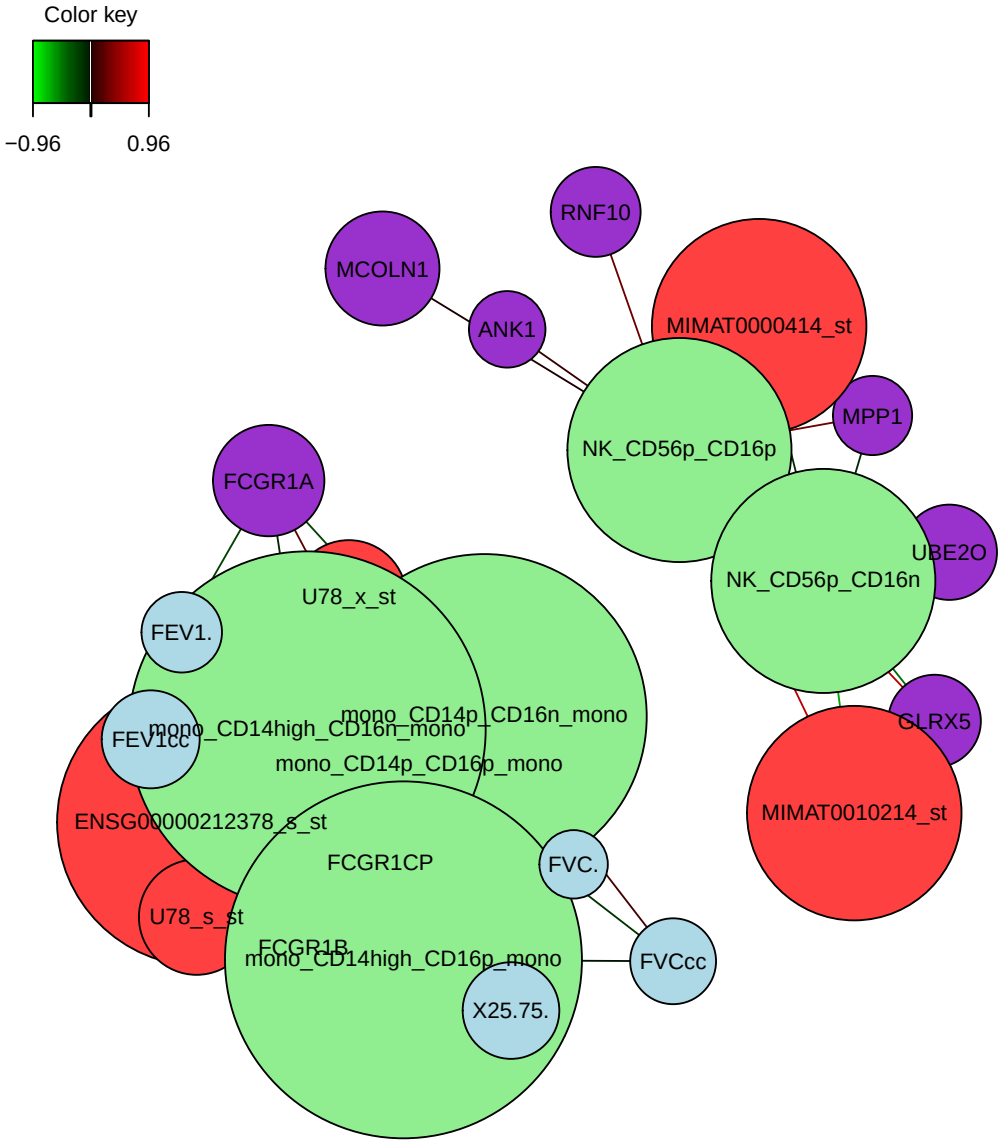


Figura 25: Red que muestra las relaciones entre los diferentes conjuntos de datos. Corte = 0.85

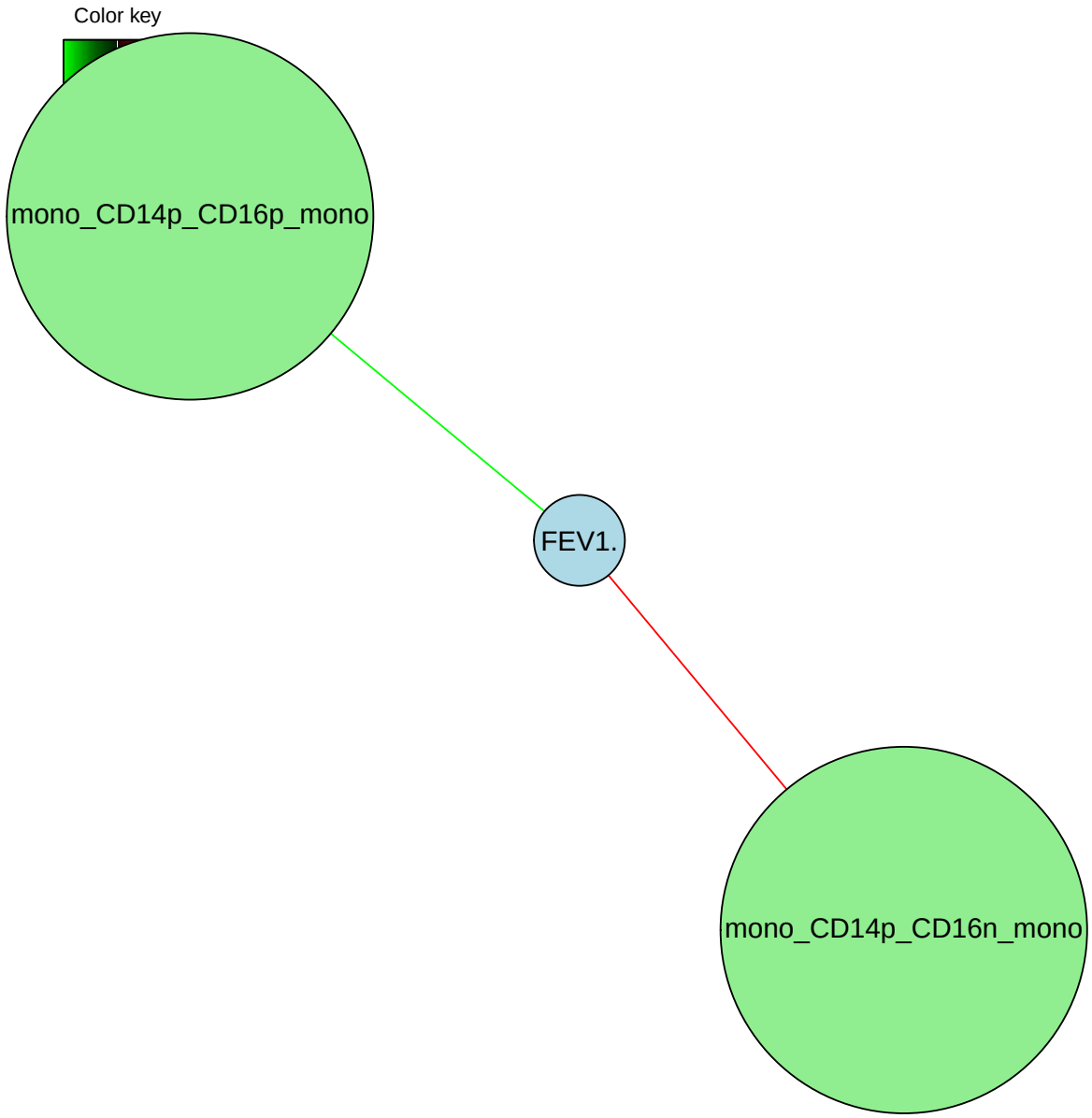


Figura 26: Red que muestra las relaciones entre los diferentes conjuntos de datos. Corte = 0.95

3.5 Discusión

En este trabajo se ha realizado una revisión de los métodos de integración actualmente descritos y se ha visto que el problema de la integración de datos, aunque se tiene muy claro que es muy importante y que es el paso necesario que se tiene que dar en el campo del análisis de datos ómicos, no existe ni mucho menos un consenso claro al respecto. Todavía existen muchos métodos estadísticos y aproximaciones que se pueden utilizar para integrar los datos. Esto hace que ante el problema de integrar los datos no se sepa cuál es la mejor aproximación. Un factor importante que alimenta este problema es la gran diferencia que hay entre los tipos de datos ómicos que se quieren integrar: resulta mucho más sencillo integrar transcriptómica con datos de miRNA, la relación se establece mucho más directamente, en cambio, integrar transcriptómica con metabolómica no es tan sencillo ya que el tipo de datos son diferentes y la relación no es tan directa. Más problema introduce el hecho de mezclar variables clínicas con los datos ómicos. Las variables clínicas tienden a explicar mucha más la relación entre las muestras y su condición experimental, ya que expresan características fenotípicas finales, en cambio la transcriptómica o la metabolómica, explican pasos previos al fenotipo que no lo acaban de explicar tan claramente. Al intentar integrar los datos clínicos frente a datos ómicos, los primeros tienden a estar más representados que los segundos en el modelo final. Se ha visto que el análisis individual de cada ómica, al menos las que se han analizado aquí, aunque conlleve un tiempo de trabajo considerable en preparación y el propio análisis de los datos, ya está muy establecido y fácilmente se puede llegar a las conclusiones finales de saber qué variables están más relacionadas con un fenotipo o con otro. Sobre el método de integración utilizado se ha constatado que es muy vistoso, ya que los autores han realizado un esfuerzo considerable en las representaciones gráficas que se pueden extraer del mismo, pero quizás le falta un poco más de esfuerzo en la parte de resultados numéricos.

4 CONCLUSIONES

4.1 Conclusiones del trabajo

En este trabajo se han analizado varias aproximaciones ómicas de manera individual desde los datos crudos hasta el análisis de expresión diferencial, pasando por el análisis descriptivo o el control de calidad según sea el origen de los datos analizados. Con los resultados que se han obtenido con cada análisis ómico, se han seleccionado aquellas variables más significativas y se ha probado el método DIABLO implementado dentro del paquete `mixOmics` para intentar obtener un modelo que mejor las relacione y sea capaz de separar mejor las muestras según la condición experimental a la que pertenecen. Se han obtenido diferentes modelos de integración/relación entre las variables utilizadas. Dependiendo del nivel de corte utilizado según la mejor correlación entre las variables, se obtienen diferentes redes que relacionan las variables. Cuando se ha evaluado el funcionamiento del modelo creado con otro conjunto de datos (subconjunto de datos llamado “test”), se ha visto que el modelo que se ha creado es capaz de clasificar bastante bien los nuevos individuos. No hay que

olvidar que lo mejor hubiese sido evaluar el modelo con nuevos pacientes independientes, que no se hayan analizado conjuntamente con los datos utilizados para crear el conjunto de datos de entrenamiento. Es lo que se llamaría una validación externa. Mediante estos resultados se ha conseguido conocer un poco más que factores pueden tener un papel importante en el rechazo del órgano en los pacientes trasplantados.

4.2 Reflexión sobre los objetivos planteados

Inicialmente se plantearon dos objetivos principales, el primero de ellos era el estudio de diferentes aproximaciones y metodologías para la integración de los datos ómicos multimodales, y el segundo era la creación de un modelo predictivo a partir del modelo seleccionado. El primer objetivo se ha cumplido prácticamente en su totalidad ya que se ha realizado una revisión profunda de los métodos de integración que existen en la actualidad, a excepción de valorar los diferentes métodos de pre-procesado de los datos, ya que se ha considerado que los métodos que existen y que están considerados como los aceptados son los mejores que hay en la actualidad. El segundo objetivo se ha cumplido prácticamente en su totalidad también, a excepción del último subobjetivo que era la de probar el pipeline creado con un conjunto de datos diferente. Como no se ha podido encontrar otro conjunto de datos, lo que se ha hecho en su lugar es dividir el conjunto de datos en dos, uno que se ha utilizado para entrenar el algoritmo de predicción y otro que se ha utilizado para evaluar la eficiencia del mismo.

4.3 Análisis crítico de la planificación

Sobre la planificación que se planteó en un principio se ha visto una vez que se ha acabado el trabajo que quizás algunos puntos se podrían haber planteado de otra manera. Concretamente, la revisión bibliográfica de los métodos de integración existentes, ya que se ha visto que al no haber un método establecido, la información es muy abundante y dispersa. Existen muchos métodos y alternativas de hacer la integración de los datos de muchas maneras diferentes y con aproximaciones diferentes. Por otro lado, y en contra de lo que acabo de decir, también hay que pensar que cuanto más se lee sobre un mismo tema, más “perdido” está uno. Otro punto dónde se ha invertido más tiempo del inicialmente planteado ha sido en la preparación de los datos. En un principio se pensó que sería una cosa rápida y sencilla, pero despueś se ha visto que analizar por separado cada ómica utilizada ha sido hacer “un análisis en cada caso”, que ya de por sí requiere un tiempo considerable de preparación de datos, control de calidad (o análisis descriptivo en los demás casos) y el propio análisis de comparación entre grupos. Finalmente, comentar que la redacción de la memoria que en un primer momento se había planteado empezar en abril, no se empezó en aquel momento y se dejó para hacerla al final cuando se tuvieran más resultados. Un error, ya que la redacción de la memoria conlleva mucho más tiempo del que se piensa en un inicio.

4.4 Posibles líneas de futuro

Sería muy interesante el poder comparar este método de integración de datos ómicos con algún otro método de los muchos que existen para comprobar si los resultados que se obtienen son similares, o por el contrario sucede lo que suele pasar muchas veces en el análisis de datos ómicos, que la comparación de dos métodos comunes suele dar resultados muy diferentes. Por otro lado también sería interesante comprobar el funcionamiento del flujo de trabajo que se han creado con otros datos que incluyan otro tipo de datos ómicos. Valorar si el el método funciona igual de bien y ver como influye el hecho de introducir otras aproximaciones ómicas.

5 REFERENCIAS

1. Manzoni C, Kia DA, Vandrovcova J, et al (2016) Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Briefings in bioinformatics 19:286-302. doi: 10.1093/bib/bbw114

2. Butcher EC, Berg EL, Kunkel EJ (2004) Systems biology in drug discovery. Nature Biotechnology 22:1253-1259. doi: 10.1038/nbt1017

3. Aardema MJ, MacGregor JT (2002) Toxicology and genetic toxicology in the new era of «toxicogenomics»: impact of «-omics» technologies. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 499:13-25. doi: https://doi.org/10.1016/S0027-5107(01)00292-5

4. R. H (2007) The Watson-Crick model of the DNA doublehelix. The history of the discovery and the role of the protein paradigm. Acta Hist Leopoldina 48:113-158

5. Venter JC, Adams MD, Myers EW, et al (2001) The Sequence of the Human Genome. Science 291:1304-1351. doi: 10.1126/science.1058040

6. Gonzalo R, Sánchez A (2018) Chapter Three - Introduction to Microarrays Technology and Data Analysis. En: Jaumot J, Bedia C, Tauler R (eds) Data Analysis for Omic Sciences: Methods and Applications. Elsevier, pp 37-69

7. GonzÃ¡lez Garcia EAT Filiberto AND Miguez (2008) CaracterizaciÃcualitativa de poliaminas libres en endurecedores de resinas epoxÃdicas del tipo etilenaminas por espectroscopia de resonancia magnÃnuclear. PolÃmeros 18:45-51

8. Cajka T, Fiehn O (2016) Toward Merging Untargeted and Targeted Methods in Mass Spectrometry-Based Metabolomics and Lipidomics. Analytical Chemistry 88:524-545. doi: 10.1021/acs.analchem.5b04491

9. De Sanctis G, Colombo R, Damiani C, et al (2018) -Omics and Clinical Data Integration. En: Integration of Omics Approaches and Systems Biology for Clinical Applications. John Wiley & Sons, Ltd, pp 248-273

10. Hood L, Tian Q (2012) Systems approaches to biology and disease enable translational systems medicine. Genomics, proteomics & bioinformatics 10:181-185. doi: 10.1016/j.gpb.2012.08.004

11. Wu C, Zhou F, Ren J, et al (2019) A Selective Review of Multi-Level Omics Data Integration Using Variable Selection. High-Throughput 8:4. doi: 10.3390/ht8010004

12. Richardson S, Tseng GC, Sun W (2016) Statistical Methods in Integrative Genomics. Annual review of statistics and its application 3:181-209. doi: 10.1146/annurev-statistics-041715-033506

13. Rohart BAS Florian AND Gautier (2017) mixOmics: An R package for 'omics feature selection and multiple data integration. PLOS Computational Biology 13:1-19. doi: 10.1371/journal.pcbi.1005752

14. Tibshirani R (1994) Regression Shrinkage and Selection Via the Lasso. JOURNAL OF THE ROYAL

STATISTICAL SOCIETY, SERIES B 58:267-288

15. González I, Cao K-AL, Davis MJ, Déjean S (2012) Visualising associations between paired 'omics' data sets. BioData Min 5:19-19. doi: 10.1186/1756-0381-5-19

16. Tenenhaus A, Tenenhaus M (2011) Regularized Generalized Canonical Correlation Analysis. Psychometrika 76:257. doi: 10.1007/s11336-011-9206-8

17. Tenenhaus A, Philippe C, Guillemot V, et al (2014) Variable selection for generalized canonical correlation analysis. Biostatistics 15:569-583. doi: 10.1093/biostatistics/kxu001

18. Singh A, Shannon CP, Tebbutt SJ, et al (2019) DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. doi: 10.1093/bioinformatics/bty1054

19. Bolstad B, Irizarry R, Åstrand M, Speed T (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19:185-193. doi: 10.1093/bioinformatics/19.2.185

20. Ritchie ME, Phipson B, Wu D, et al (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43:e47. doi: 10.1093/nar/gkv007

21. Hernández-de-Diego R, Tarazona S, Martínez-Mira C, et al (2018) PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. Nucleic Acids Res 46:W503-W509. doi: 10.1093/nar/gky466

22. Mirza B, Wang W, Wang J, et al (2019) Machine Learning and Integrative Analysis of Biomedical Big Data. Genes 10:87. doi: 10.3390/genes10020087

23. Kim M, Tagkopoulos I (2018) Data integration and predictive modeling methods for multi-omics datasets. Mol Omics 14:8-25. doi: 10.1039/C7MO00051K

24. Zhu B, Song N, Shen R, et al (2017) Integrating Clinical and Multiple Omics Data for Prognostic Assessment across Human Cancers. Sci Rep 7:16954-16954. doi: 10.1038/s41598-017-17031-8

25. Rappoport N, Shamir R (2018) Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res 46:10546-10562. doi: 10.1093/nar/gky889

26. Jain S, Kotsampasakou E, Ecker GF (2018) Comparing the performance of meta-classifiers—a case study on selected imbalanced data sets relevant for prediction of liver toxicity. Journal of Computer-Aided Molecular Design 32:583-590. doi: 10.1007/s10822-018-0116-z

27. Boulesteix A-L, Sauerbrei W (2011) Added predictive value of high-throughput molecular data to clinical data and its validation. Briefings in Bioinformatics 12:215-229. doi: 10.1093/bib/bbq085

28. Sperisen P, Cominetti O, Martin F-PJ (2015) Longitudinal omics modeling and integration in clinical

metabonomics research: challenges in childhood metabolic health research. Front Mol Biosci 2:44-44. doi: 10.3389/fmolb.2015.00044