


Universitat Oberta de Catalunya – Universitat de Barcelona

Máster universitario en Bioinformática y bioestadística

Área 5 – TFM Estadística y Bioinformática – Aula 1

Análisis integrativo de datos ómicos y datos clínicos: Predicción de variables clínicas a partir de datos de expresión génica en pacientes con Enfermedad Pulmonar Obstructiva Crónica

Trabajo final de máster presentado por

GUILLERMO RAFAEL SUÁREZ CUARTÍN

Director: **Alexandre Sánchez Pla**

Barcelona, junio de 2019

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Análisis integrativo de datos ómicos y datos clínicos: Predicción de variables clínicas a partir de datos de expresión génica en pacientes con Enfermedad Pulmonar Obstructiva Crónica*
Nombre del autor:	*Guillermo Rafael Suárez Cuartín*
Nombre del consultor/a:	*Alexandre Sánchez Pla*
Nombre del PRA:	*Alexandre Sánchez Pla*
Fecha de entrega (mm/aaaa):	06/2019
Titulación:	*Máster Universitario de Bioinformática y Bioestadística*
Área del Trabajo Final:	*Trabajo final de Máster*
Idioma del trabajo:	*Castellano*
Palabras clave	*Ómica, genómica, proteómica, metabolómica, interactómica, datos clínicos, big data, machine learning, análisis multivariante, análisis integrativo*

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

Introducción: El análisis integrativo de datos clínicos y ómicos puede aumentar el poder de predicción que ambos tienen por separado, lo que ayudaría a una mejor comprensión de múltiples enfermedades. Este trabajo estudia este análisis como herramienta para la predicción de variables clínicas de gravedad de la enfermedad pulmonar obstructiva crónica (EPOC).

Métodos: Se escogió un conjunto de datos real de pacientes con EPOC (repositorio GEO: GSE42057), y se llevó a cabo revisión bibliográfica sobre los principales métodos para análisis integrativo. Se analizó la información clínica y ómica por separado con análisis de componentes principales (PCA), y se construyó un modelo mediante regresión de mínimos cuadrados parciales dispersos (sPLS) para predecir las variables clínicas. La evaluación del modelo se realizó mediante validación cruzada.

Resultados: El análisis de genes diferencialmente expresados entre pacientes EPOC grave y controles detectó la participación de vías biológicas como la diferenciación celular hematopoyética, la interacción citoquina-receptor y las inmunodeficiencias primarias. El PCA identificó que el primer componente principal explicaba el mayor porcentaje de la varianza

(79% datos clínicos; 57% datos ómicos). Seguidamente, se construyó el modelo con sPLS, obteniendo un coeficiente Q^2>0,0975 para el primer componente principal.

Conclusiones: El análisis integrativo de datos clínicos y ómicos tiene un amplio potencial para el estudio de la EPOC. Puede ser un abordaje útil para realizar predicciones de variables clínicas como la relación FEV1/FVC, el FEV1 y la distancia caminada en el test de la marcha de 6 minutos, a partir de un conjunto de datos de expresión génica.

Abstract (in English, 250 words or less):

Introduction: The integrative analysis of clinical and omic data can increase the prediction power that both have separately, which would help a better understanding of multiple diseases. This study aims to use this analysis as a tool for the prediction of clinical variables of severity of chronic obstructive pulmonary disease (COPD).

Methods: A real data set of patients with COPD was selected (GEO repository: GSE42057), and a literatura review of the main methods for integrative analysis was performed. Clinical and omic information was analyzed separately with principal component analysis (PCA), and a model was constructed by sparse partial least squares regression (sPLS) to predict clinical variables. The evaluation of the model was carried out through cross validation.

Results: The analysis of differentially expressed genes between patients with severe COPD and controls detected the participation of biological pathways such as hematopoietic cell differentiation, cytokine-receptor interaction and primary immunodeficiencies. The PCA identified that the first principal component explained the highest percentage of the variance (79% of clinical data, 57% of omic data). Afterwards, the model was constructed with sPLS, obtaining a coefficient Q^2> 0.0975 for the first principal component.

Conclusions: The integrative analysis of clinical and omic data has a wide potential for the study of COPD. It can be a useful approach to make predictions of clinical variables such as FEV1 / FVC ratio, FEV1 and walking distance in the 6-minute walk test, from a set of gene expression data.

ÍNDICE

LISTA DE TABLAS Y FIGURAS

CAPÍTULO 1: INTRODUCCIÓN

1.1 Contexto y justificación del trabajo

El conocimiento de las bases fisiopatológicas de las enfermedades es fundamental para el adecuado manejo de los pacientes que las padecen. Existe un gran número de noxas, humanas y animales, cuyo manejo ha mejorado significativamente gracias al desarrollo de técnicas para el análisis molecular. El progreso paralelo de las herramientas bioinformáticas para procesar y analizar esta cantidad creciente de datos ómicos, ha facilitado que esta información se traduzca en la creación de estrategias de prevención y tratamientos de un gran número de enfermedades.

Esto se debe a que a pesar de que los marcadores clínicos son hasta la fecha los mejores marcadores predictores para la evolución de una enfermedad, la gran cantidad de información que proporcionan los datos ómicos puede sin duda no solo aumentar el poder predictivo, sino que además también pueden ayudar a identificar potenciales nuevas dianas terapéuticas. El problema radica en la diferencia abismal entre ambos tipos de datos, clínicos y ómicos, y en cómo pueden combinarse desde el punto de vista estadístico para potenciar su poder de predicción. Por ello, se han desarrollado múltiples estrategias para realizar el análisis integrativo de esta información tan heterogénea, como, por ejemplo, técnicas de *machine learning*, modelos de predicción o exploración estadísticos, redes de información, entre muchas otras.

Otro problema que surge en el análisis integrativo de datos clínicos y ómicos es precisamente la elección del método a emplear. Esto se deberá basar tanto en las características de los datos como en la pregunta biológica o clínica que se desea responder. En esta línea, el presente Trabajo Final de Máster (TFM) se plantea realizar una revisión de los tipos de métodos más comúnmente usados en el análisis integrativo, así como escoger uno de ellos para ser utilizado en un conjunto de datos real.

Para realizar el análisis integrativo, se ha tomado como referencia a la EPOC, debido a su alta prevalencia, a la heterogeneidad de su presentación, y a la poca comprensión que se tiene actualmente de sus bases biológicas. Una estrategia que puede emplearse para estudiar una enfermedad de estas características, es evaluar el peso que tienen los perfiles de expresión génica en la evolución y gravedad de la patología. En la actualidad, esta estrategia ha sido poco estudiada en la EPOC. Por ello, en este TFM se postula que

los datos de expresión génica pueden ser de gran valor para la predicción de variables clínicas de gravedad de la enfermedad en pacientes con EPOC.

1.2 Objetivos del trabajo

Objetivos generales

1.2.1 Estudiar me.

1.2.2 Aplicar un método de análisis integrativo para el estudio de un conjunto de datos real de pacientes con EPOC y controles sanos.

1.2.3 Evaluar la utilidad del análisis integrativo para la predicción de variables clínicas en pacientes con EPOC.

Objetivos específicos

1.2.1 Realizar una revisión bibliográfica sobre los métodos de integración de datos ómicos y clínicos disponibles.

1.2.2 Estudiar el *machine learning* y su aplicabilidad científica en la integración de datos ómicos y clínicos.

1.2.3 Analizar la utilidad de los modelos estadísticos en el análisis integrativo.

1.2.4 Evaluar el uso de plataformas informáticas para la integración de datos clínicos y ómicos.

1.2.5 Seleccionar un conjunto de datos real para aplicar el análisis integrativo.

1.2.6 Identificar la estrategia de análisis integrativo más adecuada para el estudio del conjunto de datos propuesto.

1.2.7 Implementar el método escogido para analizar el conjunto de datos real.

1.2.8 Realizar la validación del modelo creado para el análisis integrativo.

1.2.9 Comparar los valores reales de las variables clínicas de los sujetos estudiados con los resultados predichos mediante el análisis integrativo.

1.2.10 Analizar los resultados obtenidos tras aplicar el método de análisis integrativo.

1.2.11 Realizar un análisis crítico de la utilidad de la metodología escogida en el manejo de los pacientes con EPOC.

1.2.12 Estimar otros posibles usos del análisis integrativo en la EPOC, atendiendo a los resultados obtenidos en el estudio realizado.

1.3 Enfoque y método seguido

El TFM planteado se constituye de dos partes principales. Una primera parte teórica, donde se lleva a cabo una revisión bibliográfica concienzuda, dirigida a la identificación de los principales métodos de análisis integrativo de datos ómicos y clínicos. Este apartado establece las bases para la segunda parte del TFM, práctica, donde se escoge una de las estrategias estudiadas para el análisis de un conjunto de datos real.

Para la segunda parte existen tres principales tipos de métodos a ser aplicados:

- Principio de *machine learning*.
- Modelos estadísticos.
- Plataformas de información en red.

La selección del método se basa en las características del conjunto de datos a estudiar y a la pregunta a responder, ya que algunas técnicas pueden implicar análisis más complejos, mientras que el uso de modelos estadísticos podría ser suficientemente potente para el estudio de un menor volumen de datos. Por ello, se decidió emplear una combinación de métodos estadísticos y de aprendizaje supervisado, dado el relativamente reducido tamaño de los datos.

La decisión de realizar el TFM en dos partes secuenciales permitió crear tareas específicas y establecer tiempos de cumplimiento bien delimitados para las mismas.

1.4 Planificación del trabajo

A continuación, se desglosarán las actividades realizadas durante el TFM:

1.4.1 Realizar una revisión bibliográfica en el repositorio de PubMed para identificar información relevante sobre el análisis integrativo de datos ómicos y clínicos (https://www.ncbi.nlm.nih.gov). Para ello se usarán las palabras clave: "*integrative analysis*"; "*omic data*"; "*clinical data*"; "*semantinc data integration*"; y "*machine learning*", principalmente. Objetivo específico: 1.2.1; Tiempo de dedicación: 9 horas.

1.4.2 Estudiar el *machine learning* como estrategia para la integración de datos clínicos y ómicos, identificando sus principios, sus puntos fuertes y debilidades. Objetivo específico: 1.2.2; Tiempo de dedicación: 12 horas.

1.4.3 Analizar los diferentes abordajes para la creación de modelos estadísticos dirigidos al análisis integrativo de datos ómicos y clínicos. Se elaborará un resumen sobre los principales métodos y sus ventajas y desventajas. Objetivo específico: 1.2.3; Tiempo de dedicación: 15 horas.

1.4.4 Identificar las principales plataformas de análisis integrativo mediante redes de información; analizar sus propiedades y posibles aplicaciones. Objetivo específico: 1.2.4; Tiempo de dedicación: 12 horas.

1.4.5 Seleccionar un conjunto de datos real y realizar un análisis inicial de las características de los mismos. Objetivo específico: 1.2.5; Tiempo de dedicación: 36 horas.

1.4.6 Identificar la estrategia de análisis integrativo más adecuada según la pregunta a responder y aplicar el método seleccionado al conjunto de datos reales. Objetivos específicos: 1.2.6 y 1.2.7; Tiempo de dedicación: 65 horas.

1.4.7 Efectuar un análisis crítico de los resultados obtenidos. Objetivos específicos: 1.2.8 al 1.2.10; Tiempo de dedicación: 22 horas.

1.4.8 Estimar la utilidad y aplicabilidad del análisis integrativo en la EPOC, de acuerdo a los resultados obtenidos en el estudio. Objetivos específicos: 1.2.11 y 1.2.12; Tiempo de dedicación: 5 horas.

La distribución de las tareas específicas del proyecto se puede observar en la figura 1.

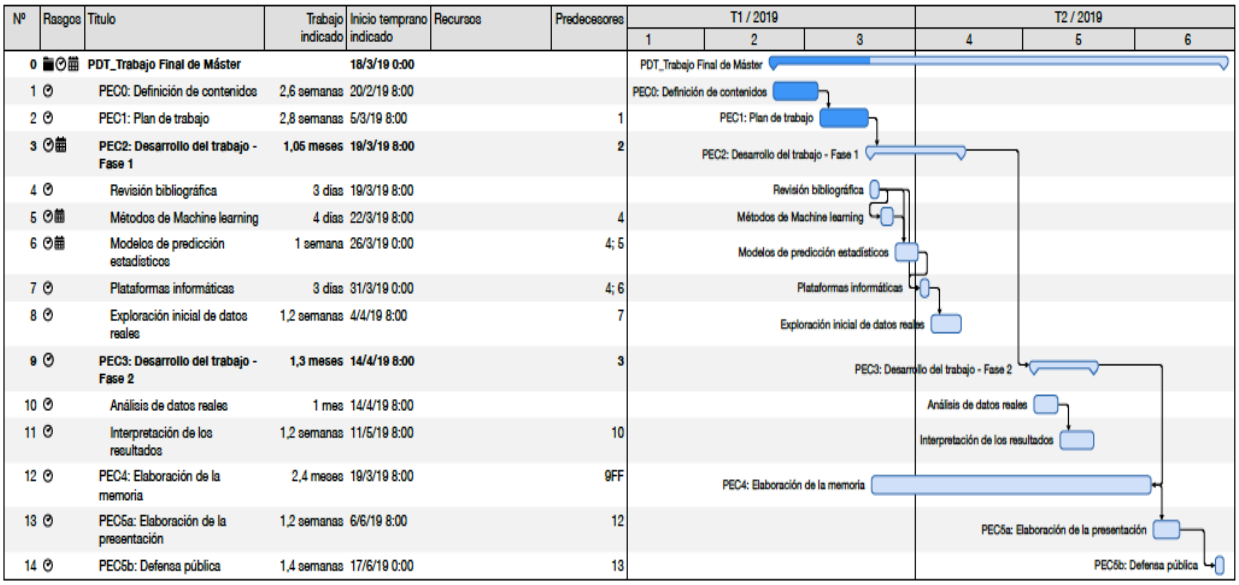


Figura 1. Diagrama de Gantt con las tareas a realizar.

En la tabla 1 se reflejan los principales hitos del proyecto de TFM con la fecha crítica correspondiente, en la cual dicha fase debe haber sido culminada.

Tabla 1. Hitos y fechas críticas del proyecto

Tareas	*Hito*	*Fecha crítica*
Definición de los objetivos	PEC0	04/03/2019
Plan de trabajo	PEC1	18/03/2019
Revisión bibliográfica y base teórica	PEC2	24/04/2019
Exploración inicial de los datos reales y selección del método de análisis integrativo	PEC2	24/04/2019
Análisis de los datos reales	PEC3	20/05/2019
Interpretación de los resultados	PEC3	20/05/2019
Cierre de la memoria	PEC4	05/06/2019
Elaboración de la presentación	PEC5a	13/06/2019
Defensa pública	PEC5b	26/06/2019

1.5 Breve sumario de productos obtenidos

El principal producto obtenido mediante el proyecto de TFM es la presente memoria, la cual recoge la metodología usada, los resultados obtenidos, así como su análisis crítico.

1.6 Breve descripción de otros capítulos de la memoria

Los siguientes capítulos constituyen el desarrollo del TFM, empezando por el marco teórico donde se describen los principales tipos de métodos de análisis integrativo, seguido por la selección del conjunto de datos real y la metodología a usar, la descripción de los resultados obtenidos y su análisis crítico. Por último, se explican las conclusiones del proyecto, la bibliografía consultada y los anexos, donde se refleja el código de R empleado para el análisis, entre otros elementos obtenidos.

CAPÍTULO 2: MARCO TEÓRICO

El creciente desarrollo de técnicas de análisis molecular y de herramientas bioinformáticas ha condicionado un aumento exponencial en la información disponible sobre múltiples enfermedades, los llamados datos "ómicos". Este tipo de datos se disponen en distintos niveles (genómica, proteómica, transcriptómica, metabolómica, entre otros), pero son tan amplios que se hace difícil su interpretación práctica [1].

Hasta la fecha, gran parte de las decisiones terapéuticas para el manejo de muchas enfermedades se basa en criterios pronósticos clínicos; pero estos son muy heterogéneos y no son necesariamente representativos de la complejidad de las enfermedades humanas [1, 2]. Por este motivo, en enfermedades cuyo origen es multifactorial debido a la interacción de mecanismos biológicos, genéticos y ambientales como el cáncer, se utiliza la información ómica además de la clínica para realizar un manejo más personalizado del paciente.

Por separado, los datos clínicos han demostrado una mejor capacidad de predicción que los ómicos en la mayoría de los estudios; sin embargo, su desarrollo ha aumentado significativamente el entendimiento actual de muchas enfermedades [1, 3, 4]. La integración de ambos tipos de información es complicada desde el punto de vista estadístico debido a las diferencias en las características de las mediciones. Por ello existen múltiples métodos disponibles para este tipo de análisis, incluyendo técnicas de *machine learning*, modelos estadístico y plataformas de redes de información, entre otras [2, 5, 6].

La elección del método más adecuado dependerá de las características de los datos a estudiar y de la pregunta a ser respondida. Dada la importancia de este tema para mejorar el manejo actual de diversas enfermedades, el TFM se centrará en el estudio de los métodos de análisis integrativos más importantes y posteriormente se implementará uno de ellos para el estudio de una base de datos real.

2.1 Machine learning

La alta complejidad y volumen de los datos biológicos han creado la necesidad del uso de métodos de análisis que se adapten continuamente y de forma eficaz. Los métodos de *machine learning* son ampliamente usados en bioinformática y

fundamentalmente consisten en el uso de algoritmos y modelos estadísticos por parte de equipos computarizados para realizar una determinada tarea basándose en el reconocimiento de patrones, sin necesidad de recibir comandos específicos [7]. Mediante estos métodos, se crean modelos computacionales a partir de un conjunto de datos de "entrenamiento" que le permiten al sistema hacer predicciones en otro conjunto de datos diferente. Actualmente se usan estos métodos en áreas muy diferentes, como lo son el diagnóstico médico, la bioinformática, informática química, análisis de redes sociales, educación, análisis bursátil, robótica, entre otras [7].

Es posible que exista más de un posible modelo computacional o algoritmo para resolver un mismo problema, y el rendimiento de un modelo específico depende de muchos factores, como las características de los datos de entrenamiento, los tipos de variables de entrada y salida, y las limitaciones del propio sistema como el tiempo y la memoria disponibles [7, 8].

Las técnicas de *machine learning* pueden ser clasificadas en dos grandes grupos, dependiendo de la presencia o no de los valores de salida en los datos de entrenamiento, llamadas aprendizaje supervisado y no supervisado, respectivamente [8]. La complejidad de los datos a analizar es clave en la selección del método más adecuado. Los denominados "datos de vistas múltiples" son aquellos conjuntos de datos que permiten la caracterización de un elemento, sistema o fenómeno a partir de distintos enfoques [6]. Existen diversos tipos de datos de vistas múltiples dependiendo del número de conjuntos de muestras y de características estudiadas. Los que miden las mismas características en distintos grupos de muestras (tipo 1), los que estudian los mismos rasgos en un solo conjunto de muestras, pero en diferentes condiciones (tipo 3), y aquellos que contienen varios grupos de características distintas para uno o múltiples conjuntos de muestras (tipos 2 y 4, respectivamente). Estos últimos son los más frecuentes y también son llamados datos "multi-ómicos" [6].

A grandes rasgos, dependiendo del análisis que se desee realizar, pueden emplearse diferentes métodos integrativos de *machine learning*. Por ejemplo, la identificación de características clave que diferencien conjuntos de muestras, la optimización de la predicción o clasificación de un grupo de elementos, el estudio de la interacción entre características distintas de un mismo conjunto de muestras, o la inferencia de las relaciones entre elementos de distintos niveles dentro de un sistema complejo [6, 9–

11]. La tabla 2 resume los principales métodos de análisis integrativo que se pueden realizar con *machine learning* según el tipo de datos de vistas múltiples.

2.1.1 Concatenación de características

Se fundamenta en la unión de las características a estudio en un único vector, a través de la clasificación, regresión y selección de rasgos de datos de múltiples clases. Para ello emplea modelos discriminativos lineales dispersos, como máquinas de vectores de soporte (SVM), o métodos de contracción, como el método de LASSO [12]. La principal limitación de esta técnica es la pérdida de información, debido a que confluyen datos de diferentes tipos (continuos, discretos, categóricos, etc.), que a su vez poseen escalas distintas [6].

2.1.2 Modelos Bayesianos

Consisten en la incorporación de conocimiento disponible *a priori* dentro del modelo de predicción o exploración. Un claro ejemplo de estos modelos son los algoritmos que se basan en la información genómica de grandes bancos de información como el proyecto ENCODE (ENCyclopedia Of DNA Elements) [13] para la predicción de regiones determinantes de una determinada secuencia a estudio [14]. Sin embargo, su aplicación para datos de múltiples niveles está limitada por la capacidad de integración de los conocimientos previos.

En estos casos donde se requiere la conjunción de datos de niveles diferentes, la creación de redes Bayesianas [15]. Los nodos de estas redes representan las variables aleatorias que tienen asociadas una función de probabilidad condicional, y sus arcos son las relaciones de dependencia directas entre ellas. Actualmente se pueden usar como herramientas de evaluación de riesgo en grandes conjuntos de datos clínicos y pueden identificar cuantitativamente las variables que son más importantes para predecir diagnósticos específicos, así como su respectivo pronóstico [16, 17].

2.1.3 Modelos de aprendizaje conjunto

Se basan en el uso de árboles de clasificación, regresión y decisión para el aprendizaje. En ellos, se escoge la característica o el conjunto de ellas que identifica con mejor a las distintas clases de elementos, correspondiendo a un nodo. Para cada nodo

se especifican diferentes reglas que a su vez permiten sub-clasificar a las clases de elementos [18]. Unas de las principales ventajas de este método son su fácil manipulación e interpretación de los resultados y la posibilidad de emplear características de diferentes niveles sin necesidad de la normalización de los datos. En este sentido, Pittman y colaboradores introdujeron el concepto de meta-características (o meta-genes en su estudio), que son características que contienen la síntesis de múltiples medidas de datos de distintas clases [19]. Sin embargo, una desventaja del uso de este método es la posibilidad de observar una alta correlación dentro de un determinado conjunto de árboles de decisión [6].

2.1.4 Aprendizaje de kernel o de núcleos múltiples

Es una técnica de integración de datos que empieza por la formación de matrices de similitud a partir de cada conjunto de características por separado, y posteriormente combina estas matrices para generar una matriz kernel que es usada en el modelo de aprendizaje [20]. Las matrices de similitud individuales se pueden calcular mediante el aprendizaje métrico a través de funciones métricas entre los elementos de una misma clase y de clases diferentes [21].

Este método se está usando cada vez más en el análisis integrativo de datos ómicos. En un estudio realizado por Zhu y colaboradores, se llevó a cabo un análisis de datos genómicos, epigenómicos y transcriptómicos de distintos tipos de cáncer de forma individual e integrativa con información clínica, con la finalidad de evaluar su capacidad de pronóstico. No solo objetivaron que la integración con los datos clínicos era sustancialmente más eficiente que la evaluación de los datos ómicos por separado para establecer el pronóstico, sino que fue el método kernel de *machine learning* el que tenía mejores resultados de forma consistente [5].

2.1.5 Métodos basados en redes

Son técnicas para estudios de asociación como las relaciones gen-enfermedad. Para ello se emplea una red donde los nodos representan los objetos y los bordes (ponderados) indican la presencia de asociaciones. Por lo tanto, los problemas de asociación se pueden resolver mediante métodos de factorización de la matriz del núcleo (relacional) y/o métodos gráficos [6, 22]. Se pueden clasificar en problemas de

dos relaciones o de relaciones múltiples; sin embargo, el paso clave está en la integración de las matrices relacionales múltiples o adyacentes de dos conjuntos de datos de objetos diferentes, ya sea mediante métodos de caminata aleatoria o la fusión de datos por factorización matricial, entre otros [23–25].

2.1.6 Factorización matricial

Su fundamento es la extracción de nuevas características a partir de cada objeto a estudio y su posterior combinación en un nuevo conjunto de características. Seguidamente, un algoritmo de clasificación es aplicado sobre este nuevo conjunto de datos para obtener una decisión. Para la extracción de las características se emplean métodos de factorización matricial como lo son el análisis de componentes principales (PCA), el análisis factorial (FA), la factorización matricial no negativa (NMF), métodos de descomposición tensorial, entre otros [6, 22]. En general, permite el estudio y cuantificación de datos de distintas clases, independientemente de su naturaleza. Además, reduce notablemente las dimensiones de los datos y con ello la complejidad del análisis computacional. Por otra parte, el uso de estos métodos no permite incluir en el análisis las interacciones entre las características de distintos objetos [26, 27].

Esta técnica se ha usado ampliamente en la bioinformática para el estudio de agrupación de la expresión génica, así como para la identificación de patrones comunes en distintos tipos de cáncer, obteniendo muy buenos resultados [22, 28, 29].

2.1.7 Aprendizaje de modelos múltiples

Partiendo del principio de los métodos de análisis en redes, la idea básica es seleccionar una subred específica para cada objeto y luego integrarlas. El uso de las subredes permite la correcta elección de los modelos de aprendizaje más apropiados para cada objeto [6, 22]. El uso de subredes favorece el ajuste global del modelo, debido a que cada una ya ha sido entrenada con datos diferentes por separado. Además, se puede emplear incluso en conjuntos de datos con valores faltantes y de distinta naturaleza, maximizando el aprovechamiento de los datos. Por último, el aprendizaje multimodal es flexible, lo cual facilita el estudio de sistemas complejos.

Tabla 2. Métodos de análisis integrativo mediante machine learning según el tipo de datos a estudiar, y los principios en los que se basan. Modificado de Li et al. [6]

Método integrativo	Tipos de datos de vistas múltiples			
	Datos de clases múltiples (Tipo 1)	**Conjuntos de características múltiples (Tipo 2)**	**Datos tensoriales (Tipo 3)**	**Datos múlti-relacionales (Tipo 4)**
Concatenación de características		Clasificación Regresión Selección de características		
Modelos o redes bayesianas	Clasificación Selección de características	Clasificación Regresión Selección de características Análisis de vías		
Aprendizaje conjunto	Clasificación Selección de características	Clasificación Regresión Selección de características		
Aprendizaje de núcleos múltiples	Clasificación	Clasificación Regresión Agrupación		Estudio de asociación
Métodos basados en redes				Estudio de asociación
Factorización matricial	Clasificación Selección de características	Clasificación Selección de características Análisis de vías Agrupación	Clasificación Agrupación	Estudio de asociación
Aprendizaje de modelos múltiples		Clasificación Agrupación Estudio de asociación		

2.2 Herramientas para el análisis integrativo en red

Como fue mencionado anteriormente, el análisis de redes es uno de los métodos más usados actualmente para el análisis integrativo. Ya sea para el estudio de las interacciones entre datos ómicos y sus vías de regulación, así como para la interpretación integrativa de información clínica y expresión génica, los modelos en red constituyen una herramienta de gran utilidad [1].

La creación de bancos que contengan diferentes estratos de información ómica de una gran cantidad de individuos, puede ayudar a la comprensión de los procesos biológicos de las enfermedades, y también a optimizar el tratamiento y el pronóstico de las mismas [30]. Proyectos como la Iniciativa de Neuroimagen de la enfermedad de Alzheimer (ADNI) [31], el Atlas del Genoma del Cáncer (TCGA) [32], o el Consorcio Internacional del Genoma del Cáncer (ICGC) [33] se han esforzado en recopilar datos multi-ómicos para realizar análisis más completos de estas enfermedades complejas. En vista de este incremento exponencial en la información ómica disponible, se han desarrollado herramientas o interfaces que facilitan la integración de los datos ómicos, así como su interpretación.

Una de las primeras herramientas desarrolladas para el análisis integrativo es el Algoritmo de Reconocimiento de Ruta que utiliza la Integración de Datos en el Modelo Genómico (PARADIGM) [34]. Al combinar el número de copias y los datos de expresión génica, toma en cuenta diferentes tipos de relaciones dentro de las vías utilizando un modelo gráfico probabilístico y es capaz de proporcionar un valor para el estado de activación de cada ruta para cada muestra [34], siendo su principal uso el análisis de datos del TCGA. Lemon-Tree es un paquete de software de código abierto, extensible, que se ha ampliado recientemente para permitir la integración de datos multi-ómicos mediante redes de inferencia. La aplicación de esta herramienta identificó varios genes conductores novedosos en el glioblastoma [35]. La Herramienta de Análisis para Asociaciones de Redes Hereditarias y Medioambientales (ATHENA), es un software que combina varios métodos de filtrado de variables con técnicas de *machine learning* para generar modelos multivariables que predicen un resultado categórico o cuantitativo [36]. Este método es flexible y se puede expandir para incluir otros tipos de datos de alto rendimiento como datos de expresión génica y mediciones de biomarcadores [36].

En la tabla 3 se mencionan las herramientas más usadas para el análisis integrativo de datos ómicos.

Aunque se han hecho grandes avances en el desarrollo de herramientas para facilitar la integración e interpretación de datos multi-ómicos, aún están lejos de ser suficientes para abordar la extensa información que se necesita para comprender los mecanismos que subyacen en las enfermedades humanas.

2.3 Modelos estadísticos

Bajo este apartado se incluyen diversos abordajes para el análisis integrativo que se basan en la elaboración de modelos de predicción y validación mediante métodos estadísticos. Uno de los principales problemas al evaluar el valor pronóstico de los datos ómicos en conjunción con la información clínica es la sobreestimación del poder predictivo de los genes debido al proceso de selección y la subestimación de las variables clínicas debido a la omisión de genes relevantes [37]. Por lo tanto, construir una "puntuación ómica" usando únicamente los datos moleculares para evaluar su significancia mediante análisis multivariante ajustado por las variables clínicas, conlleva un alto riesgo de sobreajuste del modelo y sobreestimación de las variables ómicas. En estos casos se considera crucial la creación de un modelo de validación de los modelos de predicción [38]. Para la construcción del modelo de validación puede emplearse un conjunto de datos distinto al usado para el modelo predictivo, o puede extraerse aleatoriamente de un único conjunto de datos.

2.3.1 Modelos de predicción

Se definen como una función que permite asignar una clase o una estimación de la función de supervivencia, según corresponda, a cada nueva observación [2]. Para ello se obtiene una puntuación basada en un número de predictores que se asocian con el resultado de interés. Estos modelos pueden incluir solo predictores clínicos, solo predictores moleculares o una combinación de ambos [2]. Existen cinco estrategias principales para obtener los modelos de predicción, que se resumen en la tabla 4.

Tabla 3. Algunas herramientas para el análisis integrativo de datos ómicos.

Nombre	Características	Dirección web
Anduril [39]	Traduce datos a gran escala fragmentados, en predicciones comprobables, permitiendo una rápida integración de datos heterogéneos con conocimiento existente en bases de datos conocidas.	http://csbi.ltdk.helsinki.fi/site/
ATHENA [36]	Paquete de análisis multi-ómico con filtrado de datos y modelado de interacciones.	https://ritchielab.psu.edu/software/athena-downloads
Atlas [40]	Sistema basado en modelos de datos relacionales desarrollados para cada uno de los tipos de datos de origen. Los datos almacenados dentro de estos modelos relacionales se administran a través de llamadas de *Structured Query Language* (SQL) que se implementan en un conjunto de interfaces de programación de aplicaciones.	http://bioinformatics.ubc.ca/ atlas/
BIOZON [41]	Sistema que utiliza un único esquema de gráfico extenso y estrechamente conectado que incluye una ontología jerárquica de documentos y relaciones entre los distintos tipos de datos.	http://biozon.org.
BioXM [42]	Sistema desarrollado en torno al concepto de integración semántica orientada a objetos. Permite la configuración de aplicaciones web especializadas que se pueden implementar en grandes grupos de usuarios y comunidades a través de Intranet o Internet.	https://www.biomax.com/product/bioxm-knowledge-management-environment/
FSMKL [43]	Usa el principio de aprendizaje de múltiples núcleos, seleccionando características por puntuación estadística e introduce una medida de confianza para la asignación de clases.	https://github.com/jseoane/FSMKL
iCluster [44]	Incorpora en un solo marco un modelo flexible de las asociaciones entre diferentes tipos de datos y la estructura de varianza-covarianza dentro de los tipos de datos, reduciendo al mismo tiempo la dimensionalidad de los conjuntos de datos.	https://www.mskcc.org/departments/epidemiology-biostatistics/biostatistics/icluster
JIVE [45]	Cuantifica la cantidad de variación conjunta entre los tipos de datos, reduce su dimensionalidad y proporciona nuevas direcciones para la exploración visual de la estructura conjunta e individual.	https://genome.unc.edu/jive/

Joint Genomics [46]	Factoriza el espacio espacio de las variables en dos componentes, infiriendo la dimensionalidad de estos espacios mediante un proceso beta-Bernoulli.	https://sites.google.com/site/jointgenomics
Lemon-Tree [35]	Integración de datos multi-ómicos para la inferencia en red.	http://lemon-tree.googlecode.com
MetaBridge [47]	Emplea interacciones moleculares curadas de alta calidad para identificar enzimas con interacción directa, para su posterior integración con datos transcriptómicos y proteómicos.	https://metabridge.org.
MDI (Multiple Dataset Integration) [48]	Cada conjunto de datos se modela utilizando el modelo Dirichlet-multinomial, determinando las dependencias entre estos modelos mediante parámetros que describen el acuerdo entre los conjuntos de datos.	https://warwick.ac.uk/fac/sci/systemsbiology/research/software/
NetGestalt [49]	Permite la presentación simultánea de datos experimentales y de anotaciones a gran escala de muchas fuentes.	http://www.netgestalt.org/
Omics Fusion [50]	Plataforma que ofrece conexiones entre herramientas establecidas para el análisis, integración y visualización de datos de proteómica, metabolómica y transcriptómica.	https://fusion.cebitec.uni-bielefeld.de/
PARADIGM [34]	Integración de los datos de expresión génica para estimar el estado de activación de cada vía para cada muestra	http://sbenz.github.com/Paradigm
Similarity network fusion [51]	Construye redes de muestras (por ejemplo, pacientes) para cada tipo de datos disponibles y luego las fusiona en una red que representa el espectro completo de los datos.	http://compbio.cs.toronto.edu/SNF/SNF/Software.html
SNPLS [52]	Emplea el método de mínimos cuadrados parciales para identificar patrones modulares conjuntos utilizando datos de expresión génica apareados a gran escala y datos de respuesta a fármacos.	http://page.amss.ac.cn/shihua.zhang/
VisANT 4.0 [53]	Plataforma web para navegación exploratoria de la red de interacciones; inferencia y visualización de datos ómicos con conocimiento jerárquico integrado.	http://visant.bu.edu/

2.3.1.i Estrategia 1: "naive":

Construye el modelo de predicción combinado, tratando a las variables clínicas y moleculares por igual [54]. Con este enfoque, el valor predictivo de las variables clínicas puede subestimarse ante la presencia de un gran número de predictores moleculares [2].

2.3.1.ii Estrategia 2: "residuos":

Se basa en la valoración de la variación residual de los resultados al construir un modelo de predicción fijo para las variables clínicas mediante regresión logística o regresión de Cox, a lo que posteriormente se añade el valor de las variables moleculares mediante regresión de LASSO o *boosting* [55]. Al usar este método es necesario tener en cuenta que si se realiza previamente una selección de variables clínicas puede haber un sesgo en los valores de los coeficientes de dichas variables, que no se verán afectados por los predictores moleculares. Una forma de sortear este sesgo podría ser la construcción de subgrupos según las características clínicas [2].

2.3.1.iii Estrategia 3: "favorecimiento":

Consiste en construir un modelo predictivo para ambos tipos de variables, "favoreciendo" a las clínicas mediante diferentes enfoques como la regresión penalizada o la aproximación CoxBoost [56]. Se podría considerar un punto intermedio entre las dos primeras estrategias debido a que se aprovecha mejor el peso de las variables clínicas respecto a la estrategia 1, pero dicho peso puede verse influenciado por los predictores moleculares, a diferencia de la segunda estrategia.

2.3.1.iv Estrategia 4: "reducción de dimensionalidad":

Es otra forma de "favorecer" a las variables clínicas. Se realiza en dos pasos, empezando por una síntesis de los predictores moleculares en forma de nuevos componentes mediante una técnica de reducción de dimensionalidad, seguida de la construcción de un modelo predictivo que incluye estos nuevos componentes y las variables clínicas como covariables mediante el método de Cox o de *Random Forest*, entre otros [2, 56, 57].

2.3.1.v Estrategia 5: “reemplazo”:

Consiste en sustituir las variables clínicas con menor peso en el modelo o más afectadas por errores de medida, por los predictores moleculares [2].

2.3.2 Modelos de validación del valor predictivo añadido con dos conjuntos de datos

Construir estos modelos de validación permite determinar la precisión y validez externa de los modelos predictivos. Los distintos enfoques para realizar la validación del poder predictivo añadido se resumen en la tabla 5.

2.3.2.i Enfoque A: Comparación del modelo predictivo clínico con el modelo predictivo combinado:

Se construyen dos modelos predictivos con los datos de entrenamiento, uno clínico y uno combinado (clínico y molecular). Los resultados de ambos modelos se comparan posteriormente (sensibilidad, especificidad, tasa de error), usando un conjunto de datos de validación.

2.3.2.ii Enfoque B: Comparación de una puntuación clínica y una puntuación combinada:

En los casos donde no es posible comparar los resultados de los modelos de predicción en los datos de validación (por ejemplo, diferentes escalas en los conjuntos de datos), puede ser útil comparar la capacidad de discriminación de los modelos y su asociación con los resultados, en lugar de la predicción. En caso de querer comparar la puntuación para la predicción de clases, pueden emplearse las curvas de *Receiver Operating Characteristics* (ROC), pruebas de igualdad del área bajo la curva (AUC) o el índice-c; si en cambio se quiere comparar la puntuación del análisis de supervivencia, se puede usar la regresión de Cox, por ejemplo [2].

2.3.2.iii Enfoque C: Prueba de la puntuación molecular basada en datos de validación en un modelo multivariado que se ajusta por los predictores clínicos:

Durante la primera fase de entrenamiento se obtiene una única puntuación molecular, tomando en cuenta los predictores clínicos o no de los datos de entrenamiento (por ejemplo, usando el análisis de componentes principales), para

posteriormente comparar su valor predictivo añadido en los datos de validación, incluyendo los factores clínicos predictivos de este conjunto de datos [56]. Este enfoque es usado con frecuencia en el análisis de estudios con datos moleculares multi-dimensionales [58].

2.3.2.iv Enfoque D: Comparación de modelos de predicción con y sin puntuación molecular mediante validación cruzada en los datos de validación:

Es similar al enfoque anterior, pero emplea la validación cruzada en lugar de las pruebas de significación para determinar la precisión de los modelos predictivos. Pueden construirse estos modelos mediante, por ejemplo, regresión logística o de Cox [2]. Este enfoque permite cuantificar el aumento en la precisión al incluir la puntuación molecular al modelo con predictores clínicos únicamente.

2.3.2.v Variante 1: Análisis de subgrupos: Tras construir el modelo predictivo con los datos de entrenamiento, se realiza la validación mediante uno de los cuatro enfoques descritos previamente, pero por separado según los subgrupos considerados dentro del conjunto de datos de validación [2].

2.3.2.vi Variante 2: Análisis de subgrupos con diferentes modelos ajustados para cada subgrupo: Se diferencia del enfoque anterior al considerar los subgrupos clínicos para construir los modelos de predicción con los datos de entrenamiento. Permite incluir posibles interacciones entre los predictores clínicos y moleculares; sin embargo, se requiere de un tamaño muestral grande para que los subgrupos tengan un tamaño suficiente para realizar el análisis [2].

2.3.2.vii Variante 3: Puntuación de predictores clínicos ajustada con los datos de entrenamiento: Similar a los enfoques C y D, pero con la diferencia de que los coeficientes de los predictores clínicos se ajustan en base a los datos de entrenamiento (y no con los datos de validación como en los enfoques previos). Así, tanto la puntuación clínica como molecular se ajustan a partir de los mismos datos [2].

2.3.2.viii Variante 4: La puntuación de los predictores clínicos se obtiene de la literatura: Dado que no se obtiene la puntuación clínica a partir de los datos de entrenamiento, no es necesario contar con la información clínica de dicho conjunto de datos [2].

2.3.3 Modelos de validación del valor predictivo añadido en datos de entrenamiento

En los casos en los que solo se cuente con un conjunto de datos que no permita partirlo de forma aleatoria para obtener los conjuntos de entrenamiento y validación, pueden emplearse otras pruebas para evaluar el poder predictivo.

2.3.3.i Pruebas globales con ajuste: Se basan en modelos lineales con predictores lineales con la intención de comprobar la hipótesis nula de que los predictores moleculares no aportan valor predictivo añadido. Una forma de hacerlo es ajustando los modelos de regresión regularizada usando *boosting* con la puntuación clínica como una compensación [55]. Otra forma sería con el método GlobalAncova, que se basa en análisis paralelos de varianza realizados para todos los predictores moleculares simultáneamente, con la clase como factor, permitiendo el ajuste por los predictores clínicos [59]. Estas pruebas no permiten realizar una comparación de los errores de predicción.

2.3.3.ii Enfoques de re-muestreo: Aplicable para los enfoques A y B, consiste en aplicar el principio de validación cruzada, donde en cada iteración, los datos excluidos funcionan como datos de validación mientras el resto lo hace como datos de entrenamiento. Estos enfoques son de gran utilidad cuando el conjunto de datos disponible no es lo suficientemente grande como para ser separado en dos. Una forma de aplicar este enfoque es la “pre-validación”, que consiste en una validación cruzada realizada dentro de un único conjunto de datos donde se obtiene un valor de puntuación para cada observación [60]. Seguidamente, se ajusta un único modelo de regresión multivariado con esta puntuación pre-validada, con las variables clínicas como predictores. El valor predictivo añadido se obtiene al evaluar la significación del coeficiente de regresión de la puntuación [60]. Este método no permite evaluar la ganancia en precisión del modelo, sino solo su significancia.

Tabla 4. Resumen de las distintas estrategias para la creación de modelos predictivos. Modificado de [2].

	Naive	**Residuos**	**Favorecimiento**	**Reducción de dimensionalidad**	**Reemplazo**
Enfoque de un solo paso	+	-	+	-	-
Predictores clínicos y moleculares tienen el mismo peso	+	-	-	-	-
Depende de un parámetro crucial	-	-	+	+	-
La contribución de las variables clínicas se puede afectar por los predictores moleculares	+	-	+	+	+/-
Se ajusta a un modelo clínico único	-	+	-	-	+
Ajusta un modelo molecular a los residuos del modelo clínico del primer paso	-	+	-	-	-
Reemplaza una variable clínica "débil" con predictores moleculares	-	-	-	-	+
Adecuado para evaluar el valor predictivo añadido	-	+	+/-	+/-	+/-

Tabla 5. Resumen de los enfoques generales para evaluar la validez del valor predictivo agregado. Modificado de [2].

	A	B	C	D
Usa modelos/puntuaciones combinados	+	+	-	-
Basado solo en puntuaciones	-	+	+	+
Se basa en la ganancia de precisión estimada directamente en los datos de validación	+	+	-	-
Se basa en la ganancia de precisión estimada a través del re-muestreo en los datos de validación	-	-	-	+
Se basa en pruebas de significación en modelos multivariables ajustados a los datos de validación	-	-	+	-
Considera la puntuación molecular como un "nuevo predictor"	-	-	+	+
Se ajusta el/los modelo(s) a los datos de validación	-	-	+	+
Se ajusta un modelo a los datos clínicos del conjunto de entrenamiento.	+	+	-	-

2.4 Generalidades de la Enfermedad Pulmonar Obstructiva Crónica

La enfermedad pulmonar obstructiva crónica (EPOC) es la enfermedad respiratoria crónica más frecuente y se estima que para el año 2020 sea la tercera causa de muerte en el mundo [61]. Es el resultado de una combinación de exposiciones ambientales, principalmente al humo del tabaco, con la presencia de susceptibilidad genética. Se caracteriza clínicamente por presentar disnea, tos crónica e intolerancia al ejercicio, y funcionalmente como una obstrucción irreversible del flujo aéreo en la espirometría, con una relación entre el volumen espiratorio forzado al primer segundo (FEV1) y la capacidad vital forzada (FVC) menor de 0,7 además de la disminución o no del FEV1 por debajo del 80% del valor predicho según las características antropométricas del sujeto [62].

La clasificación clínica de la EPOC se basa en cuatro grupos (A, B, C y D), dependiendo de los síntomas y del número de exacerbaciones e ingresos hospitalarios en el último año. Funcionalmente se clasifica en cuatro estadios de gravedad según la obstrucción al flujo aéreo en la espirometría: 1 o leve (FEV1 >80%), 2 o moderada (FEV1 $\geq$ 50% y < 80%), 3 o grave (FEV1 $\geq$ 30% y < 50%), y 4 o muy grave (FEV1 < 30%) [62].

La EPOC representa un verdadero reto terapéutico debido a la heterogeneidad de su presentación, siendo la inflamación local y sistémica un punto clave en la fisiopatología de la enfermedad [63]. Sin embargo, son múltiples las vías biológicas que participan en esta enfermedad. Diversos estudios recientes se han enfocado en el análisis de datos ómicos para ampliar el conocimiento sobre la EPOC, ya sea para el estudio de las vías implicadas en el desarrollo de la EPOC [64, 65], o para definir subtipos de la enfermedad [66]. A pesar de ello, la utilidad de la información ómica para la prevención, pronóstico y tratamiento de los pacientes con EPOC es muy limitada, por lo que aún no se ha integrado en el manejo clínico habitual.

CAPÍTULO 3: MATERIALES Y MÉTODOS

3.1 Selección del conjunto de datos reales

Para la realización del TFM se escogieron los datos procedentes del estudio realizado por Bahr y colaboradores [64], en el cual se analizaron datos de expresión génica de sujetos entre 45 y 80 años, fumadores activos o exfumadores (sanos y con EPOC) con una dosis acumulada de 10 o más paquetes-año, e identificados como de raza blanca no hispánica o africana-americana. Todos los sujetos formaban parte de la cohorte del estudio COPDGene y se encontraban en fase de estabilidad clínica, determinada por la ausencia de exacerbaciones respiratorias en los 30 días previos a la inclusión en el estudio [65].

3.2 Análisis de datos clínicos

Se recogieron datos clínicos de todos los sujetos, como el género, la edad, el hábito tabáquico y la dosis acumulada de tabaco, el índice de masa corporal, porcentaje del valor predicho del volumen espiratorio forzado en el primer segundo (FEV1), la relación entre el FEV1 y la capacidad vital forzada (FEV1/FEVC), el porcentaje de enfisema y de atrapamiento aéreo en la tomografía computarizada de tórax y la distancia (en pies) caminada con el test de la marcha de 6 minutos, así como la clasificación de la Global Initiative for Chronic Obstructive Lung Disease (GOLD). Al realizar el análisis clínico, se extrajeron los datos de relevancia clínica mencionados previamente, se recalibraron los niveles de las variables categóricas y se crearon otras variables como estado de la enfermedad: control/enfermo de EPOC, y gravedad de la enfermedad: Control/EPOC leve-moderado/EPOC grave (según el porcentaje del valor predicho del FEV1) para facilitar la comparación de los datos. Los pacientes con EPOC se clasificaron como habitualmente en la clínica según el FEV1, siendo leves/moderados aquellos con FEV1 mayor al 50% del predicho (GOLD 1 y 2), y graves aquellos con un FEV1 menor al 50% del predicho (clasificación GOLD 3 y 4).

Para el análisis predictivo se emplearon únicamente las variables continuas (de las disponibles en el conjunto de datos) que tienen interés de predicción desde el punto de vista clínico y biológico, como el FEV1, la relación FEV1/FVC, la distancia caminada en el

test de la marcha (medida en pies) y el porcentaje de enfisema y atrapamiento aéreo en la tomografía computarizada de tórax. Para la comparación de variables categóricas entre los tres grupos de gravedad de la enfermedad se empleó el test de chi-cuadrado, y para las variables continuas, los tests de ANOVA o Kruskal-Wallis según la distribución normal o no de las mismas.

3.3 Análisis de datos ómicos

El estudio de los datos de expresión génica se realizó a partir de ARN aislado de células mononucleares en sangre periférica. Para el estudio de microarrays se midieron 54675 transcritos mediante la matriz de genes Affymetrix Human Genome U133 plus 2.0 (número de acceso del repositorio GEO: GSE 42057). El análisis de los datos se realizó con el software RStudio versión 1.2.1335. El código empleado se recoge en el anexo 1.

Tras realizar un análisis exploratorio de los datos de expresión génica y comprobar que no había alteraciones importantes en las señales, se realizó un control de calidad a nivel de sonda (PLM) y se analizaron gráficamente la expresión logarítmica relativa y los errores estándar normalizados sin escalar. Seguidamente se llevó a cabo la normalización de los datos con el método de la medida de expresión media robusta de múltiples matrices (RMA). Finalmente, se ejecutó un filtraje no específico para eliminar aquellos genes con poca variación entre condiciones y los que no disponían de anotaciones.

El análisis de genes diferencialmente expresados se realizó con el estadístico empírico de Bayes del paquete “limma” empleando las tres condiciones de la variable “severity” creada previamente, comparando controles vs pacientes con EPOC leve/moderado, pacientes con EPOC leve/moderado vs grave y controles vs EPOC grave, siendo esta última comparación la que objetivó una mayor diferencia entre los genes diferencialmente expresados.

Para la comparación de genes diferencialmente expresados se usó el método de Benjamini y Hochberg de la tasa de falsos descubrimientos (FDR). Finalmente, se empleó el paquete “annotate” para conseguir las anotaciones de los genes diferenciados en las

bases de datos “Gene Ontology” (GO) y “Kyoto Encyclopedia of Genes and Genomes” (KEGG).

3.4 Análisis integrativo

Para el análisis integrativo de este conjunto de datos, se decidió realizar una aproximación inicial con métodos estadísticos de reducción de la dimensionalidad, específicamente el análisis de componentes principales; esto debido al gran número de datos ómicos que se manejan y al tipo de estudio que se desea realizar. Para el análisis se utilizan las funciones contenidas en el paquete “MixOmics” [67]. Brevemente, se realiza una exploración inicial mediante análisis de componentes principales por separado para los datos clínicos y ómicos, identificando a grandes rasgos el número de componentes que explican el mayor porcentaje de la varianza para cada caso. Los datos faltantes (NAs) presentes en algunas de las variables clínicas (“walkdist”, porcentaje de enfisema y de atrapamiento aéreo) fueron imputados mediante el análisis iterativo de mínimos cuadrados parciales no lineales (NIPALS).

Seguidamente, a partir de los resultados observados con el análisis de reducción de dimensiones, se emplea una aproximación supervisada, en concreto, el análisis de mínimos cuadrados parciales (PLS). Esta permite hallar las relaciones entre los dos tipos de datos (clínicos y ómicos), así como la regresión de mínimos cuadrados parciales dispersos (sPLS), mediante la producción de combinaciones lineales dispersas de los predictores originales [68]. Esta última fue la técnica que finalmente fue escogida para construir el modelo de predicción.

Debido a que contamos con un tamaño muestral limitado, se realizó la validación del modelo creado mediante sPLS con el método de validación cruzada “Mfolds” de la función “perf” (MixOmics), fijando el número de evaluaciones en 50 veces. Por último, se evaluaron las predicciones iniciales del modelo con la raíz del error cuadrático medio (RMSE) y comparando directamente los resultados obtenidos.

CAPÍTULO 4: RESULTADOS

4.1 Análisis descriptivo de los datos clínicos

En el estudio original de Bahr et al. [64], se incluyeron 136 sujetos fumadores, de los cuales 42 eran controles sanos y 94 diagnosticados de EPOC con distintos estadios de gravedad. Las características clínicas de la población según su condición y gravedad se resumen en la tabla 6.

Tabla 6. Características clínicas de la población del estudio [64].

	Control N=42	**EPOC Leve o Moderado N=52**	**EPOC grave N= 42**	**p-valor**
Género masculino, n (%)	22 (52,38%)	25 (48,08%)	27 (64,29%)	0,278
Edad, media (DE)	60,46 (9,07)	62,16 (8,99)	66,90 (6,35)	0,001
Índice de masa corporal	26,84 (7,93)	27,98 (6,44)	25,79 (7,57)	0,034
EPOC familiar, n (%)	6 (14,29%)	7 (13,46%)	6 (14,29%)	0,999
Tabaquismo activo, n (%)	13 (30,95%)	17 (32,69%)	5 (11,9%)	0,044
Dosis acumulada de tabaco (paquetes-año)	38,75 (31,13)	39,35 (28,95)	48 (29,68)	0,035
Distancia caminada en el TM6M, media (DE)	1697 (331,24)	1433,40 (414,46)	1089,76 (310,62)	<0,001
Relación FEV1/FVC	0,77 (0,06)	0,65 (0,09)	0,38 (0,12)	<0,001
FEV1 porcentaje predicho, media (DE)	97,93 (13,85)	69,04 (11,61)	32,90 (10,11)	<0,001
% Enfisema	0,76 (1)	2,58 (4,15)	19,8 (10,99)	<0,001
% Atrapamiento aéreo	8,27 (6,21)	15,23 (19,83)	57,84 (16,24)	<0,001

Los datos se muestran como mediana (rango intercuartiles), excepto cuando se manifieste lo contrario. EPOC: Enfermedad pulmonar obstructiva crónica; DE: Desviación estándar; FEV1: Volumen espiratorio forzado en 1 segundo; TM6M: Test de la marcha de 6 minutos; FVC: Capacidad vital forzada.

4.2 Análisis descriptivo de los datos ómicos

Al realizar las comparaciones entre controles con pacientes EPOC leve o moderado y entre estos y los pacientes con EPOC grave no se encontraron diferencias significativas entre los genes diferencialmente expresados; sin embargo, entre los controles y pacientes con EPOC grave se observaron 128 genes con diferencias significativas; 84 de los cuales se encontraban infra-expresados y 44 sobre-expresados. Se observaron algunos genes con un cambio de expresión > 1, pero no parecen corresponder con aquellos con p-valor más bajo. La figura 2 muestra un gráfico de volcán con los genes más diferencialmente expresados (en rojo, los genes con p-valor <0,05 y en naranja aquellos con un cambio log2 >1). La tabla 7 resume los 20 genes más diferencialmente expresados entre los controles y los pacientes con EPOC grave, que se usarán luego para el análisis integrativo.

Tras realizar una revisión de las vías biológicas en las que participan estos genes mediante las bibliotecas GEO (anexo 2) y KEGG (anexos 3-6), se identificaron principalmente las vías de diferenciación de células hematopoyéticas, así como diferentes vías de activación del sistema inmune (principalmente en la interacción entre citoquinas y sus receptores), así como en el desarrollo de inmunodeficiencias primarias.

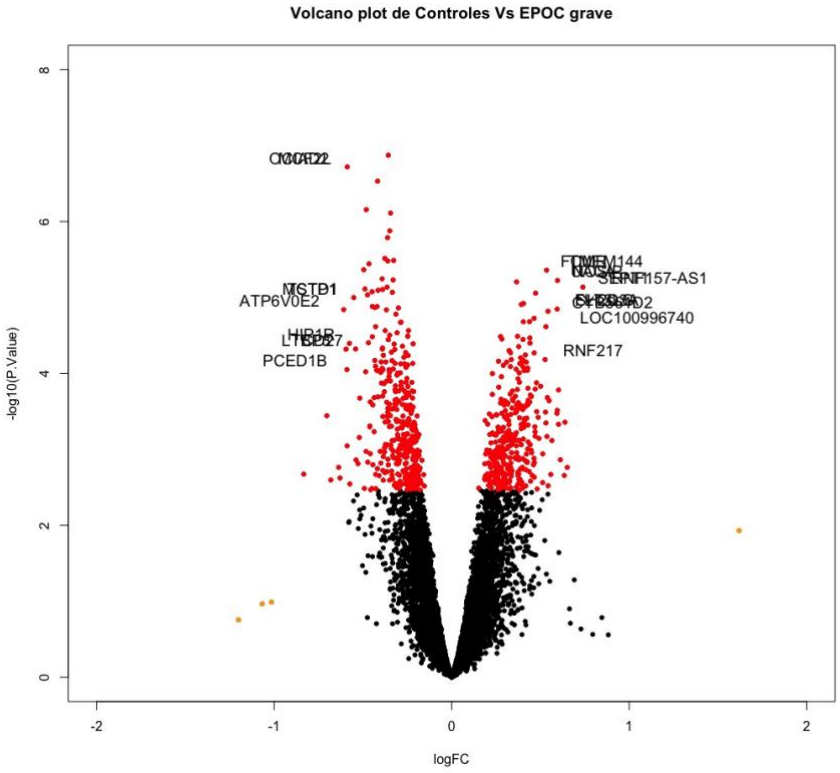


Figura 2. Gráfico de volcán de los genes más diferencialmente expresados entre controles y EPOC graves.

Tabla 7. Lista de los 20 genes más diferencialmente expresados entre controles y pacientes con EPOC grave

Genes	Cambio log2 estimado	Expresión log2 media	Estadístico T	P-valor	P-valor ajustado
OCIAD2	-0,3573	8,8249	-5,5608	<0,001	0,0010
FCMR	-0,5879	8,7496	-5,4851	<0,001	0,0010
NOSIP	-0,4173	9,6221	-5,3913	<0,001	0,0010
SEPT1	-0,4812	8,0118	-5,2017	<0,001	0,0016
TSTD1	-0,3440	9,1112	-5,1783	<0,001	0,0016
SH2D3A	-0,3485	6,2010	-5,0572	<0,001	0,0022
ATP6V0E2	-0,3612	8,2014	-5,0107	<0,001	0,0024
CYB561D2	-0,3768	6,0376	-4,8665	<0,001	0,0033
LOC100996740	-0,3268	10,5262	-4,8514	<0,001	0,0033
HIP1R	-0,3600	6,4418	-4,8488	<0,001	0,0033
CD27	-0,4653	8,9539	-4,8281	<0,001	0,0033
LTBP3	-0,4953	7,1622	-4,7860	<0,001	0,0034
RNF217	0,5349	4,5943	4,7830	<0,001	0,0034
PCED1B	-0,3919	8,7975	-4,7211	<0,001	0,0035
MCF2L	-0,3301	5,5413	-4,7131	<0,001	0,0035
TMEM144	0,5963	5,1315	4,7096	<0,001	0,0035
UACA	0,3660	4,2135	4,6994	<0,001	0,0035
RNF157-AS1	-0,3637	3,4678	-4,6617	<0,001	0,0035
MCTP1	0,7393	7,1006	4,6617	<0,001	0,0035
FLT3LG	-0,4885	7,3207	-4,6480	<0,001	0,0035

4.3 Análisis integrativo

Se inició el estudio con un análisis de componentes principales de los datos clínicos y ómicos por separado. En ambos casos, el primer componente principal explicaba la mayor parte de la varianza (79% para los clínicos, y 57% para los ómicos). Al comparar gráficamente según la condición y gravedad, en el conjunto de datos clínicos parecen existir grupos claramente separados, principalmente entre los controles y los EPOC graves. Menos evidente parece en el caso de los datos ómicos. Seguidamente, se realizó el análisis mediante sPLS, con el cual se creó una red de relevancia con los dos primeros componentes principales en función de la expresión génica y la información clínica (Figura 3).

Se realizó la evaluación del modelo creado con sPLS mediante el método de validación cruzada, observando que el primer componente principal es el que tiene el mayor poder predictivo, con un coeficiente $Q^2 > 0{,}0975$ (Figura 4).

Las predicciones de variables clínicas realizadas fueron evaluadas con la raíz del error cuadrático medio (RMSE), sin observar un error particularmente grande para algunas de las variables. La tabla 8 refleja los valores de RMSE para las predicciones de los primeros dos componentes principales obtenidos mediante sPLS.

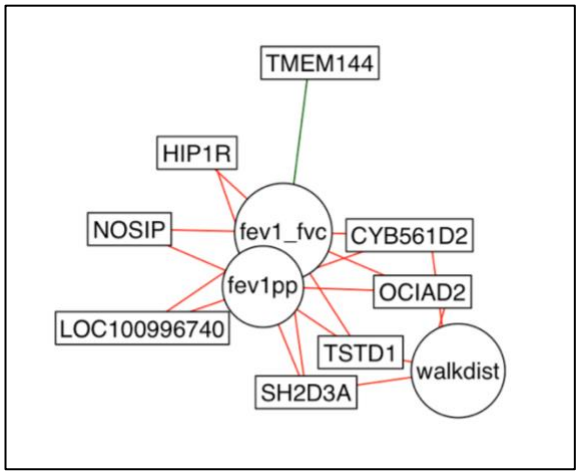


Figura 3. Red de relevancia con los dos primeros componentes principales creado a partir de sPLS. Las líneas verdes representan las correlaciones positivas, y las rojas, negativas.

Figura 4. Evaluación del modelo de predicción obtenido con la regresión de mínimos cuadrados parciales dispersos (sPLS) mediante validación cruzada (número de repeticiones: 50).

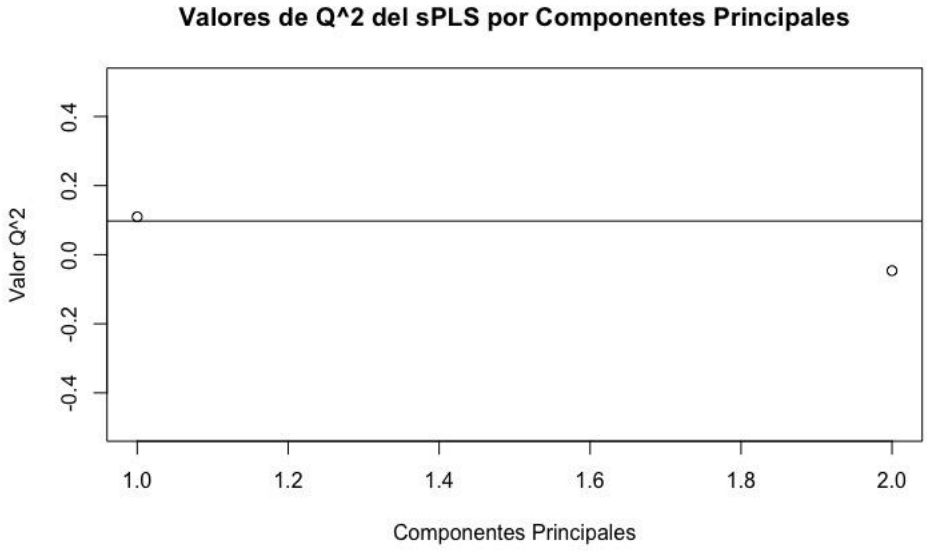


Tabla 8. Valores de la raíz del error cuadrático medio para las predicciones de los dos primeros componentes principales del análisis con sPLS.

RMSE	Distancia caminada	Relación FEV1/FVC	FEV1	Enfisema	Atrapamiento aéreo
sPLS Componente 1	383.4563	0.1717404	24.63936	236.448	98.54509
sPLS Componente 2	381.7444	0.1717072	24.60861	226.546	95.71014

RMSE: Raíz del error cuadrático medio; sPLS: Regresión de mínimos cuadrados parciales dispersos; FEV1: Volumen espiratorio forzado en el primer segundo; FVC: Capacidad vital forzada.

Finalmente, se realizó una comparación entre las predicciones realizadas con el primer componente principal del sPLS (el de mayor valor predictivo objetivado mediante validación cruzada), con los valores observados reales (y penalizados con el método NIPALS). Como ejemplo, la tabla 9 muestra la comparación para los primeros 6 sujetos.

Tabla 9. Comparación de los valores predichos con sPLS y los observados/imputados para los 6 primeros sujetos.

VALORES PREDICHOS					
Sujeto	**Distancia caminada**	**Relación FEV1/FVC**	**FEV1**	**Enfisema**	**Atrapamiento aéreo**
GSM1031549	1390,01	0,58	66,15	-10,16	33,84
GSM1031550	1073,91	0,42	44,88	-27,91	54,87
GSM1031551	1407,96	0,59	67,35	-9,15	32,64
GSM1031552	990,41	0,37	39,27	-32,61	60,43
GSM1031553	1596,75	0,68	80,05	1,46	20,07
GSM1031554	1662,56	0,72	84,48	5,15	15,70
VALORES OBSERVADOS E IMPUTADOS					
Sujeto	**Distancia caminada**	**Relación FEV1/FVC**	**FEV1**	**Enfisema**	**Atrapamiento aéreo**
GSM1031549	1600,00	0,46	36,96	168,52	33,47
GSM1031550	1805,01	0,58	53,97	138,47	34,52
GSM1031551	1620,00	0,54	52,98	105,38	29,70
GSM1031552	1010.01	0,33	31,99	69,90	18,87
GSM1031553	1641,00	0,78	93,99	1.58	19,11
GSM1031554	1460,00	0,73	96,03	-107,94	13,90

sPLS: Regresión de mínimos cuadrados parciales dispersos; FEV1: Volumen espiratorio forzado en el primer segundo; FVC: Capacidad vital forzada.

CAPÍTULO 5: ANÁLISIS Y DISCUSIÓN DE LOS RESULTADOS

El presente TFM plantea el uso del análisis integrativo de datos clínicos y ómicos como una herramienta para la predicción de diferentes variables de gravedad de enfermos con EPOC a partir de su perfil de expresión genética con una precisión aceptable para las principales características estudiadas, como lo son los valores espirométricos (FEV1 y la relación FEV1/FVC), así como la distancia caminada en el test de la marcha de 6 minutos.

Son muchas las vías que han sido implicadas en la patogénesis de la EPOC [65]. Sin embargo, la presentación de la enfermedad es sumamente heterogénea, como lo son el vasto número de genes implicados en las diferentes presentaciones de dicha entidad. Así pues, la predicción de la situación funcional de un determinado individuo a partir de su patrón de diferenciación génica puede ser una herramienta de utilidad para estimar el pronóstico del mismo, y anticipar medidas que permitan optimizar su manejo. Hasta la fecha, algunos estudios han empleado métodos de análisis integrativo para el estudio de la EPOC, logrando identificar subtipos de la enfermedad en base a diversos tipos de datos ómicos [66, 69]. A pesar de ello, a conocimiento del autor, no existen estudios que utilicen el análisis con sPLS para predecir variables funcionales clínicas a partir de los datos de expresión genética.

En un primer paso se realizó el análisis exploratorio del conjunto de datos, categorizando a los participantes según su condición de enfermo o no, y según la gravedad de su EPOC de acuerdo a la clasificación GOLD. Al describir la población, era de esperar, en base a la experiencia clínica, que los pacientes más graves tuvieran peores valores en las pruebas funcionales, mayor porcentaje de enfisema y atrapamiento aéreo en las pruebas de imagen [62]. Asimismo, en vista de que los pacientes más graves pueden tener un peor estado nutricional y que la EPOC puede agravarse con la edad, era esperable también que los grupos tuvieran diferencias significativas en estos aspectos.

El estudio de los genes diferencialmente expresados solo encontró diferencias significativas al comparar sujetos fumadores sanos con los pacientes con EPOC grave. Este análisis arrojó información interesante respecto a las vías en las que están

implicados dichos genes, destacando vías de activación de respuesta inmunológica (citoquinas-receptores de citoquinas), de la diferenciación de células inmunes y de algunas inmunodeficiencias primarias. Estos resultados siguen la línea de los hallazgos de estudios previos dirigidos al análisis de la inmunopatología de la enfermedad [70]. Dado que la EPOC es una enfermedad en la que la que la indemnidad del sistema inmune es fundamental para evitar la progresión rápida o las exacerbaciones, es de gran relevancia estudiar estas vías para identificar posibles estrategias de prevención y/o potenciales dianas terapéuticas.

El uso de las estrategias de reducción de la dimensionalidad y de los métodos de aprendizaje supervisado han demostrado ser herramientas muy versátiles que pueden ser de gran ayuda para el análisis integrativo de datos ómicos y clínicos, así como para elaborar modelos de predicción y clasificación [71, 72]. En el presente estudio se ha podido observar como la combinación de estrategias para realizar el análisis integrativo puede ser una aproximación válida en conjuntos de datos heterogéneos como este. Mediante el PCA se pudo identificar el número de componentes principales que explicaban la mayor parte de la varianza en los datos ómicos y clínicos, los cuales fueron utilizados seguidamente con una estrategia de aprendizaje supervisado como el sPLS.

Mediante este último método se obtuvieron predicciones con una buena precisión y con un RMSE relativamente bajo, principalmente para las variables de mayor uso y peso en la evaluación del paciente como lo son los valores de la espirometría y la distancia caminada en el test de la marcha. Sin embargo, en algunas predicciones, especialmente las de las variables con más valores faltantes como el porcentaje de enfisema, no se ha objetivado una gran precisión.

El presente estudio tiene ciertas limitaciones. Por una parte, el estudio de Bahr y colaboradores reporta posibles limitaciones concernientes a la obtención y procesado de los datos ómicos [64]. Sin embargo, estas no afectaron significativamente a los resultados del estudio. Además, dada la prevalencia y heterogeneidad de la EPOC, el tamaño muestral y el limitado número de características disponibles en el conjunto de datos puede no ser representativo de todos los pacientes con esta enfermedad. A pesar

de ello, los resultados son prometedores y pueden sentar las bases para futuros estudios con un mayor número de participantes y de variables clínicas.

Por otra parte, los resultados obtenidos mediante este trabajo son una muestra del potencial que tiene el análisis integrativo de datos ómicos y clínicos para el estudio de la EPOC. Esto se debe a que, no solo es posible realizar predicciones, sino también ayuda a conocer mejor la enfermedad mediante la identificación de subgrupos de pacientes según sus datos de expresión génica y a través de la identificación de vías biológicas involucradas en formas más graves de la enfermedad. A su vez, puede potencialmente ayudar a identificar nuevas dianas terapéuticas y por lo tanto realizar un abordaje más personalizado de la enfermedad.

Por lo tanto, es posible concluir que a través del uso de métodos de reducción de la dimensionalidad y de aprendizaje supervisado es posible predecir con cierta precisión, algunas variables clínicas de gravedad de la EPOC como la distancia caminada en el test de la marcha de 6 minutos y los valores espirométricos, para este conjunto de datos.

CAPÍTULO 6: CONCLUSIONES

Los resultados obtenidos del estudio realizado en el marco del presente trabajo final de máster han permitido llegar a las siguientes conclusiones:

- El análisis integrativo de datos ómicos y clínicos permite ampliar los conocimientos sobre una determinada enfermedad, lo cual ayuda a optimizar su manejo. Asimismo, los tipos de metodología para realizar el análisis integrativo son muy variados y deben ser escogidos de acuerdo a las características de los datos a analizar, así como a la pregunta que se desea responder.

- La combinación de estrategias de reducción de la dimensionalidad como el análisis de componentes principales, y de aprendizaje supervisado como la regresión de mínimos cuadrados parciales dispersos, es un abordaje útil para realizar predicciones de variables clínicas, como la relación FEV1/FVC, el FEV1 y la distancia caminada en el test de la marcha de 6 minutos, a partir de un conjunto de datos de expresión génica en pacientes con enfermedad pulmonar obstructiva crónica.

- El análisis integrativo de información clínica y ómica es una estrategia con un amplio potencial para el conocimiento de la enfermedad pulmonar obstructiva crónica, así como para predecir la gravedad de su presentación entre pacientes, y potencialmente identificar posibles vías de acción para el manejo de la enfermedad.

CAPÍTULO 7: GLOSARIO

EPOC	Enfermedad pulmonar obstructiva crónica
Espirometría	Prueba de exploración de la función pulmonar
FEV1	Volumen espiratorio forzado en el primer segundo
FEV1/FVC	Cociente que determina la obstrucción de la vía aérea
FVC	Capacidad vital forzada
GO	*Gene ontology*
KEGG	*Kyoto Encyclopedia of Genes and Genomes*
Machine learning	Método de análisis de datos que automatiza la construcción de modelos analíticos
NIPALS	Estimación no lineal por mínimos cuadrados iterativos
PCA	Análisis de componentes principales
PLS	Mínimos cuadrados parciales
R	Software y lenguaje de programación para análisis estadístico
RMSE	Raíz del error cuadrático medio
sPLS	Mínimos cuadrados parciales dispersos
TFM	Trabajo final de Máster
TM6M	Test de la marcha de 6 minutos. Prueba funcional cardio-pulmonar para determinar la distancia que puede caminar un individuo en 6 minutos

CAPÍTULO 8: BIBLIOGRAFÍA

1. Karczewski KJ, Snyder MP. Integrative omics for health and disease. Nat. Rev. Genet. NIH Public Access; 2018. p. 299–310.

2. Boulesteix AL, Sauerbrei W. Added predictive value of high-throughput molecular data to clinical data and its validation. *Brief. Bioinform.* Oxford University Press; 2011; 12: 215–229.

3. Losko S, Heumann K. Semantic data integration and knowledge management to represent biological network associations. *Methods Mol. Biol.* 2017. p. 403–423.

4. De Sanctis G, Colombo R, Damiani C, Sacco E, Vanoni M. -Omics and Clinical Data Integration. In: Vlahou A, Mischak H, Zoidakis J, Magnia F, editors. *Integr. Omi. Approaches Syst. Biol. Clin. Appl.* First edit. John Wiley & Sons, Inc.; 2017. p. 248–273.

5. Zhu B, Song N, Shen R, Arora A, Machiela MJ, Song L, Landi MT, Ghosh D, Chatterjee N, Baladandayuthapani V, Zhao H. Integrating Clinical and Multiple Omics Data for Prognostic Assessment across Human Cancers. *Sci. Rep.* Nature Publishing Group; 2017; 7: 16954.

6. Li Y, Wu FX, Ngom A. A review on machine learning principles for multi-view biological data integration. *Brief. Bioinform.* Oxford University Press; 2018; 19: 325–340.

7. Baştanlar Y, Özuysal M. Introduction to Machine Learning. Humana Press, Totowa, NJ; 2014. p. 105–128.

8. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. *Science* American Association for the Advancement of Science; 2015; 349: 255–260.

9. Li Y, Chen C-Y, Kaye AM, Wasserman WW. The identification of cis-regulatory elements: A review from a machine learning perspective. *Biosystems.* 2015; 138: 6–17.

10. Kristensen VN, Lingjærde OC, Russnes HG, Vollan HKM, Frigessi A, Børresen-Dale A-L. Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer* 2014; 14: 299–313.

11. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.* 2015; 16: 85–97.

12. Gligorijević V, Pržulj N. Methods for biological data integration: perspectives and challenges. *J. R. Soc. Interface* The Royal Society; 2015; 12.

13. ENCODE Project Consortium TEP. An integrated encyclopedia of DNA elements in the human genome. *Nature* NIH Public Access; 2012; 489: 57–74.

14. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* Cold Spring Harbor Laboratory Press; 2011; 21: 447.

15. Tian T. Bayesian Computation Methods for Inferring Regulatory Network Models Using Biomedical Data. *Transl. Biomed. Informatics* Springer, Singapore; 2016. p. 289–307.

16. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artif. Intell. Med.* 2001; 23: 89–109.

17. Onisko A, Druzdzel MJ, Austin RM. Application of Bayesian network modeling to pathology informatics. *Diagn. Cytopathol.* 2019; 47: 41–47.

18. Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W. Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression. *Ann. Behav. Med.* 2003; 26: 172–181.

19. Pittman J, Huang E, Dressman H, Horng C-F, Cheng SH, Tsou M-H, Chen C-M, Bild A, Iversen ES, Huang AT, Nevins JR, West M. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proc. Natl. Acad. Sci. U. S. A.* National Academy of Sciences; 2004; 101: 8431–8436.

20. Mariette J, Villa-Vialaneix N. Unsupervised multiple kernel learning for heterogeneous data integration. Wren J, editor. *Bioinformatics* 2018; 34: 1009–1015.

21. Li D, Tian Y. Survey and experimental study on metric learning methods. *Neural Networks* 2018; 105: 447–462.

22. Yu XT, Zeng T. Integrative analysis of omics big data. *Methods Mol. Biol.* 2018. p. 109–135.

23. Lan W, Wang J, Li M, Liu J, Wu F-X, Pan Y. Predicting MicroRNA-Disease Associations Based on Improved MicroRNA and Disease Similarities. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 2018; 15: 1774–1782.

24. Xuan P, Han K, Guo Y, Li J, Li X, Zhong Y, Zhang Z, Ding J. Prediction of potential disease-associated microRNAs based on random walk. *Bioinformatics* 2015; 31: 1805–1815.

25. Zitnik M, Zupan B. Jumping across biomedical contexts using compressive data fusion. *Bioinformatics* Oxford University Press; 2016; 32: i90–i100.

26. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature* 1999; 401: 788–791.

27. Li Y, Ngom A. The non-negative matrix factorization toolbox for biological data mining. *Source Code Biol. Med.* BioMed Central; 2013; 8: 10.

28. Kim H, Park H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics* Narnia; 2007; 23: 1495–1502.

29. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* Elsevier; 2013; 3: 246–259.

30. Dihazi H, Asif AR, Beißbarth T, Bohrer R, Feussner K, Feussner I, Jahn O, Lenz C, Majcherczyk A, Schmidt B, Schmitt K, Urlaub H, Valerius O. Integrative omics - from data to biology. *Expert Rev. Proteomics* 2018. p. 463–466.

31. Weiner MW, Aisen PS, Jack CR, Jagust WJ, Trojanowski JQ, Shaw L, Saykin AJ, Morris JC, Cairns N, Beckett LA, Toga A, Green R, Walter S, Soares H, Snyder P, Siemers E, Potter W, Cole PE, Schmidt M, Alzheimer's Disease Neuroimaging Initiative. The Alzheimer's Disease Neuroimaging Initiative: Progress report and future plans. *Alzheimer's Dement.* 2010; 6: 202–211.e7.

32. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* 2013; 45: 1113–1120.

33. Hudson (Chairperson) TJ, Anderson W, Aretz A, Barker AD, Bell C, Bernabé RR, Bhan MK, Calvo F, Eerola I, Gerhard DS, Guttmacher A, Guyer M, Hemsley FM, Jennings JL, Kerr D, Klatt P, Kolar P, Kusuda J, Lane DP, Laplace F, Lu Y, Nettekoven G, Ozenberger B, Peterson J, Rao TS, Remacle J, Schafer AJ, Shibata T, Stratton MR, Vockley JG, et al. International network of cancer genome projects. *Nature* 2010; 464: 993–998.

34. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* Oxford University Press; 2010; 26: i237.

35. Bonnet E, Calzone L, Michoel T. Integrative Multi-omics Module Network Inference with Lemon-Tree. *PLoS Comput. Biol.* Public Library of Science; 2015; 11: e1003983.

36. Holzinger ER, Dudek SM, Frase AT, Pendergrass SA, Ritchie MD. ATHENA: The analysis tool for heritable and environmental network associations. *Bioinformatics* Oxford University Press; 2014; 30: 698–705.

37. Truntzer C, Maucort-Boulch D, Roy P. Comparative optimism in models involving both classical clinical and gene expression information. *BMC Bioinformatics* BioMed Central; 2008; 9: 434.

38. George SL. Statistical Issues in Translational Cancer Research. *Clin. Cancer Res.* American Association for Cancer Research; 2008; 14: 5954–5958.

39. Ovaska K, Laakso M, Haapa-Paananen S, Louhimo R, Chen P, Aittomäki V, Valo E, Núñez-Fontarnau J, Rantanen V, Karinen S, Nousiainen K, Lahesmaa-Korpinen A-M, Miettinen M, Saarinen L, Kohonen P, Wu J, Westermarck J, Hautaniemi S. Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Med.* BioMed Central; 2010; 2: 65.

40. Shah SP, Huang Y, Xu T, Yuen MMS, Ling J, Ouellette BFF. Atlas - a data warehouse for integrative bioinformatics. *BMC Bioinformatics* BioMed Central; 2005; 6: 34.

41. Birkland A, Yona G. BIOZON: a system for unification, management and analysis of heterogeneous biological data. *BMC Bioinformatics* BioMed Central; 2006; 7: 70.

42. Maier D, Kalus W, Wolff M, Kalko SG, Roca J, Marin de Mas I, Turan N, Cascante M, Falciani F, Hernandez M, Villà-Freixa J, Losko S. Knowledge management for systems biology a general and visually driven framework applied to translational medicine. *BMC Syst. Biol.* BioMed Central; 2011; 5: 38.

43. Seoane JA, Day INM, Gaunt TR, Campbell C. A pathway-based data integration framework for prediction of disease progression. *Bioinformatics* Oxford University Press; 2014; 30: 838–845.

44. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* Oxford University Press; 2009; 25: 2906–2912.

45. Lock EF, Hoadley KA, Marron JS, Nobel AB. JOINT AND INDIVIDUAL VARIATION EXPLAINED (JIVE) FOR INTEGRATED ANALYSIS OF MULTIPLE DATA TYPES. *Ann. Appl.*

Stat. NIH Public Access; 2013; 7: 523–542.

46. Ray P, Zheng L, Lucas J, Carin L. Bayesian joint analysis of heterogeneous genomics data. *Bioinformatics* 2014; 30: 1370–1376.

47. Hinshaw SJ, H Y Lee A, Gill EE, E W Hancock R. MetaBridge: enabling network-based integrative analysis via direct protein interactors of metabolites. Berger B, editor. *Bioinformatics* Narnia; 2018; 34: 3225–3227.

48. Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* Oxford University Press; 2012; 28: 3290–3297.

49. Shi Z, Wang J, Zhang B. NetGestalt: integrating multidimensional omics data over biological networks. *Nat. Methods* NIH Public Access; 2013; 10: 597–598.

50. Brink BG, Seidel A, Kleinbölting N, Kleinbölting K, Nattkemper TW, Albaum SP. Omics Fusion-A Platform for Integrative Analysis of Omics Data. *J. Integr. Bioinform.* 2016; 13: 296.

51. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A. Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 2014; 11: 333–337.

52. Chen J, Zhang S. Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data. *Bioinformatics* 2016; 32: 1724–1732.

53. Hu Z, Chang Y-C, Wang Y, Huang C-L, Liu Y, Tian F, Granger B, Delisi C. VisANT 4.0: Integrative network platform to connect genes, drugs, diseases and therapies. *Nucleic Acids Res.* Oxford University Press; 2013; 41: W225-31.

54. Hothorn T, Bühlmann P, Dudoit S, Molinaro A, van der Laan MJ. Survival ensembles. *Biostatistics* 2005; 7: 355–373.

55. Boulesteix A-L, Hothorn T. Testing the additional predictive value of high-dimensional molecular data. *BMC Bioinformatics* BioMed Central; 2010; 11: 78.

56. Bøvelstad HM, Nygård S, Borgan Ø. Survival prediction from clinico-genomic models - a comparative study. *BMC Bioinformatics* BioMed Central; 2009; 10: 413.

57. Boulesteix A-L, Strobl C, Augustin T, Daumer M. Evaluating microarray-based classifiers: an overview. *Cancer Inform.* SAGE Publications; 2008; 6: 77–97.

58. Gu W, Pepe M. Measures to summarize and compare the predictive capacity of markers. *Int. J. Biostat.* Berkeley Electronic Press; 2009; 5: Article 27.

59. Hummel M, Meister R, Mansmann U. GlobalANCOVA: exploration and assessment of gene group effects. *Bioinformatics* 2008; 24: 78–85.

60. Tibshirani RJ, Efron B. Pre-validation and inference in microarrays. *Stat. Appl. Genet. Mol. Biol.* 2002; 1: Article1.

61. López-Campos JL, Tan W, Soriano JB. Global burden of COPD. Respirology John Wiley & Sons, Ltd (10.1111); 2016. p. 14–23.

62. Singh D, Agusti A, Anzueto A, Barnes PJ, Bourbeau J, Celli BR, Criner GJ, Frith P, Halpin

DMG, Han M, López Varela MV, Martinez F, Montes de Oca M, Papi A, Pavord ID, Roche N, Sin DD, Stockley R, Vestbo J, Wedzicha JA, Vogelmeier C. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease: The GOLD Science Committee Report 2019. *Eur. Respir. J.* European Respiratory Society; 2019; : 1900164.

63. Amsellem V, Gary-Bobo G, Marcos E, Maitre B, Chaar V, Validire P, Stern JB, Noureddine H, Sapin E, Rideau D, Hue S, Le Corvoisier P, Le Gouvello S, Dubois-Randé JL, Boczkowski J, Adnot S. Telomere dysfunction causes sustained inflammation in chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.* 2011; 184: 1358–1366.

64. Bahr TM, Hughes GJ, Armstrong M, Reisdorph R, Coldren CD, Edwards MG, Schnell C, Kedl R, LaFlamme DJ, Reisdorph N, Kechris KJ, Bowler RP. Peripheral Blood Mononuclear Cell Gene Expression in Chronic Obstructive Pulmonary Disease. *Am. J. Respir. Cell Mol. Biol.* American Thoracic Society; 2013; 49: 316–323.

65. Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, Beaty TH, Curran-Everett D, Silverman EK, Crapo JD. Genetic epidemiology of COPD (COPDGene) study design. *COPD* NIH Public Access; 2010; 7: 32–43.

66. Kim S, Herazo-Maya JD, Kang DD, Juan-Guardela BM, Tedrow J, Martinez FJ, Sciurba FC, Tseng GC, Kaminski N. Integrative phenotyping framework (iPF): integrative clustering of multiple omics data identifies novel lung disease subphenotypes. *BMC Genomics* BioMed Central; 2015; 16: 924.

67. Rohart F, Gautier B, Singh A, Lê Cao K-A. mixOmics: An R package for 'omics feature selection and multiple data integration. Schneidman D, editor. *PLOS Comput. Biol.* Public Library of Science; 2017; 13: e1005752.

68. Chun H, Keleş S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Series B. Stat. Methodol.* Wiley-Blackwell; 2010; 72: 3–25.

69. Chang Y, Glass K, Liu Y-Y, Silverman EK, Crapo JD, Tal-Singer R, Bowler R, Dy J, Cho M, Castaldi P. COPD subtypes identified by network-based clustering of blood gene expression. *Genomics* Academic Press; 2016; 107: 51–58.

70. Caramori G, Casolari P, Barczyk A, Durham AL, Stefano A Di, Adcock I. COPD immunopathology. *Semin. Immunopathol.* Springer; 2016; 38: 497.

71. Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief. Bioinform.* Narnia; 2016; 17: 628–641.

72. Pascoe SJ, Wu W, Collison KA, Nelsen LM, Wurst KE, Lee LA. Use of clinical characteristics to predict spirometric classification of obstructive lung disease. *Int. J. Chron. Obstruct. Pulmon. Dis.* Dove Press; 2018; 13: 889–902.

CAPÍTULO 9: ANEXOS

Anexo 1: Código R usado para llevar a cabo el análisis integrativo

Preparación de los datos

```
setwd("/Users/gsuarezcuartin/Documents/Máster/TFM")
workingDir <-getwd()
if(!dir.exists("data")) dir.create("data")
if(!dir.exists("results")) dir.create("results")
if(!dir.exists("celfiles")) dir.create("celfiles")
dataDir <-file.path(workingDir, "data")
resultsDir <- file.path(workingDir,"results")
celfilesDir <- file.path(workingDir,"celfiles")
setwd(workingDir)
```

Empezamos cargando los paquetes y los datos:

```
library(affy)

library(Biobase)

library(GEOquery)

library(hgu133plus2cdf)

gset <- getGEO("GSE42057", GSEMatrix =TRUE, getGPL=FALSE) #Obtenemos l
os datos

## Found 1 file(s)

## GSE42057_series_matrix.txt.gz

## Parsed with column specification:
## cols(
##   .default = col_double(),
##   ID_REF = col_character()
## )

#Nos aseguramos de usar solo una plataforma
if (length(gset) > 1) idx <- grep("GPL570", attr(gset, "names")) else
idx <- 1
gsetx <- gset[[idx]]
gset[[1]]

## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 12531 features, 136 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: GSM1031549 GSM1031550 ... GSM1031684 (136 total)
##   varLabels: title geo_accession ... tissue:ch1 (56 total)
##   varMetadata: labelDescription
## featureData: none
## experimentData: use 'experimentData(object)'
```

```
##   pubMedIds: 23590301
## 25494452
## 30459441
## Annotation: GPL570
```

Parte 1: Datos Clínicos

##1.1 Exploración de los datos clínicos

Empezamos por obtener dichos datos:

```
#Obtenemos la tabla con los datos de los pacientes
dbgset<- pData(phenoData(gset[[1]]))[,c(1,2,44:55)]
head(dbgset)
```

```
##                               title geo_accession age_enroll:ch1
## GSM1031549    COPD subject (10062C)    GSM1031549           64.7
## GSM1031550    COPD subject (10071D)    GSM1031550           66.2
## GSM1031551    COPD subject (10087S)    GSM1031551           65.5
## GSM1031552    COPD subject (10097V)    GSM1031552           75.8
## GSM1031553 Control subject (10102O)    GSM1031553           61.2
## GSM1031554 Control subject (10104S)    GSM1031554           50.6
##            ats_packyears:ch1 bmi:ch1 distwalked:ch1 fev1_fvc_utah:c
h1
## GSM1031549                38   27.71           1600              0.
46
## GSM1031550                18   32.64           1805              0.
65
## GSM1031551                63   26.58           1620              0.
49
## GSM1031552                41   23.62           1010              0.
43
## GSM1031553                39   30.82           1641              0.
78
## GSM1031554              34.6   26.87           1460              0.
76
##            fev1pp_utah:ch1 finalgold:ch1 gender:ch1 parentalcopd:ch
1
## GSM1031549              37             3          1
1
## GSM1031550              54             2          1
2
## GSM1031551              53             2          0
0
## GSM1031552              32             3          0
0
## GSM1031553              94             0          1
1
## GSM1031554              96             0          0
2
##            pctemph_slicer:ch1 pctgastrap_slicer:ch1 smokcignow:ch1
## GSM1031549                 NA                    NA              0
## GSM1031550                 NA                    NA              0
## GSM1031551                 NA                    NA              0
```

```
## GSM1031552                NA                  NA              0
## GSM1031553           1.58281             19.1088              0
## GSM1031554                NA                  NA              1
```

Nos hemos quedado con las siguientes variables clínicas: edad, dosis acumulada de tabaco (paquetes-año), índice de masa corporal, distancia caminada en el test de marcha, relación espirométrica entre el FEV1/FVC, el porcentaje predicho de FEV1, escala GOLD de gravedad de la EPOC (-1=enfermos sin obstrucción (FEV1/FVC >0.7), 0=controles, 1=leve, 2=moderada, 3=grave, 4=muy grave), género (0=mujer, 1=hombre), historia de padres con EPOC (0=no, 1=desconocido, 2=Si), porcentaje de enfisema y de atrapamiento aéreo en estudios de imagen (tomografía computarizada) y tabaquismo activo (0=no, 1=si).

```
summary(dbgset)
```

```
##                       title     geo_accession      age_enroll:ch1
##  Control subject (10102O):  1   Length:136         Length:136
##  Control subject (10104S):  1   Class :character   Class :character
##  Control subject (10136F):  1   Mode  :character   Mode  :character
##  Control subject (10391V):  1
##  Control subject (10465Y):  1
##  Control subject (10687Q):  1
##  (Other)                  :130
##  ats_packyears:ch1    bmi:ch1          distwalked:ch1
##  Length:136         Length:136         Length:136
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  fev1_fvc_utah:ch1  fev1pp_utah:ch1    finalgold:ch1
##  Length:136         Length:136         Length:136
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   gender:ch1        parentalcopd:ch1   pctemph_slicer:ch1
##  Length:136         Length:136         Length:136
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  pctgastrap_slicer:ch1 smokcignow:ch1
##  Length:136            Length:136
##  Class :character      Class :character
##  Mode  :character      Mode  :character
##
```

Vemos que todos los datos son interpretados como variables tipo “character”, por lo que primero vamos a especificar cuáles de ellas son realmente numéricas y categóricas.

```
#Creo una variable para diferenciar enfermos de controles:
dbgset$condition <- paste0(substr(dbgset$title,1,4))
#Especificamos los valores faltantes
dbgset$`pctemph_slicer:ch1`[dbgset$`pctemph_slicer:ch1`==0]<- NA
dbgset$`pctgastrap_slicer:ch1`[dbgset$`pctgastrap_slicer:ch1`==0]<- NA

#Cambiamos los tipos de variables:
dbgset[,c(3:8,12,13)]<- sapply(dbgset[, c(3:8,12,13)], as.numeric)

dbgset[,c(9:11,14,15)]<- lapply(dbgset[, c(9:11,14,15)], as.factor)
```

Para facilitar la interpretación y manejo de los datos, voy a renombrar las categorías, crearé una variable para agrupar los grados de severidad de la enfermedad y cambiaré los valores = 0 en el test de la marcha por “NA”, ya que el resultado cero implica que no se realizó el test por algún motivo.

```
#Renombramos categorías
levels(dbgset$`gender:ch1`)<- list('female'='0', 'male'='1')
levels(dbgset$`parentalcopd:ch1`)<-list('No'='0', 'Unk'='1','Yes'='2')
levels(dbgset$`smokcignow:ch1`)<- list('No'='0', 'Yes'='1')

#Agrupamos variables de gravedad
dbgset$severity[dbgset$`finalgold:ch1`=='0']<- 'Control'
dbgset$severity[dbgset$`finalgold:ch1`== '-1']<- 'MildMod'
dbgset$severity[dbgset$`finalgold:ch1`== '1']<- 'MildMod'
dbgset$severity[dbgset$`finalgold:ch1`== '2']<- 'MildMod'
dbgset$severity[dbgset$`finalgold:ch1`=='3']<- 'Severe'
dbgset$severity[dbgset$`finalgold:ch1`=='4']<- 'Severe'
dbgset$severity<- as.factor(dbgset$severity)

#Asignamos NA's a los valores 0 del test de marcha
is.na(dbgset$`distwalked:ch1`) <- !dbgset$`distwalked:ch1`

summary(dbgset)
```

```
##                        title     geo_accession      age_enroll:ch1
##  Control subject (10102O):  1   Length:136         Min.   :45.00
##  Control subject (10104S):  1   Class :character   1st Qu.:57.08
##  Control subject (10136F):  1   Mode  :character   Median :62.75
##  Control subject (10391V):  1                      Mean   :63.10
##  Control subject (10465Y):  1                      3rd Qu.:69.75
##  Control subject (10687Q):  1                      Max.   :79.60
##  (Other)                 :130
##  ats_packyears:ch1    bmi:ch1      distwalked:ch1 fev1_fvc_utah:ch1
##  Min.   : 11.00    Min.   :15.57   Min.   : 200   Min.   :0.2300
##  1st Qu.: 28.93    1st Qu.:24.25   1st Qu.:1130   1st Qu.:0.4400
##  Median : 41.70    Median :26.95   Median :1460   Median :0.6500
##  Mean   : 47.79    Mean   :27.91   Mean   :1416   Mean   :0.5958
##  3rd Qu.: 62.25    3rd Qu.:31.19   3rd Qu.:1659   3rd Qu.:0.7525
```

```
##  Max.   :145.00    Max.   :47.76   Max.   :2485   Max.   :0.8500
##                                    NA's   :5
##  fev1pp_utah:ch1  finalgold:ch1  gender:ch1 parentalcopd:ch1
##  Min.   :  9.00   -1:10         female:62   No :106
##  1st Qu.: 43.75   0 :42         male  :74   Unk: 11
##  Median : 68.00   1 : 8                     Yes: 19
##  Mean   : 66.80   2 :34
##  3rd Qu.: 85.25   3 :25
##  Max.   :135.00   4 :17
##
##  pctemph_slicer:ch1 pctgastrap_slicer:ch1 smokcignow:ch1 condition
##  Min.   : 0.03225   Min.   : 0.3955       No  :100       Cont:42
##  1st Qu.: 0.72632   1st Qu.: 7.9749       Yes : 35       COPD:94
##  Median : 2.50719   Median :16.0263       NA's:  1
##  Mean   : 8.32981   Mean   :25.9467
##  3rd Qu.:15.36130   3rd Qu.:47.6611
##  Max.   :45.06300   Max.   :77.5963
##  NA's   :19         NA's   :21
##     severity
##  Control:42
##  MildMod:52
##  Severe :42
##
```

Comparamos las variables entre los grupos. Empezamos por las categóricas usando el test de chi cuadrado:

Género:

```
table(dbgset$`gender:ch1`, dbgset$severity)
```

```
##
##          Control MildMod Severe
##   female      20      27     15
##   male        22      25     27
```

```
chisq.test(dbgset$`gender:ch1`, dbgset$severity, correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  dbgset$`gender:ch1` and dbgset$severity
## X-squared = 2.5619, df = 2, p-value = 0.2778
```

EPOC familiar:

```
table(dbgset$`parentalcopd:ch1`, dbgset$severity)
```

```
##
##       Control MildMod Severe
##   No       33      41     32
##   Unk       3       4      4
##   Yes       6       7      6
```

```
#En este caso usaremos el test de Fisher porque hay celdas con recuent
o <5
fisher.test(dbgset$`parentalcopd:ch1`, dbgset$severity)

##
##  Fisher's Exact Test for Count Data
##
## data:  dbgset$`parentalcopd:ch1` and dbgset$severity
## p-value = 1
## alternative hypothesis: two.sided
```

Tabaquismo actual:

```
table(dbgset$`smokcignow:ch1`, dbgset$severity)

##
##       Control MildMod Severe
##   No       28      35     37
##   Yes      13      17      5

chisq.test(dbgset$`smokcignow:ch1`, dbgset$severity, correct=FALSE)

##
##  Pearson's Chi-squared test
##
## data:  dbgset$`smokcignow:ch1` and dbgset$severity
## X-squared = 6.2528, df = 2, p-value = 0.04388
```

Para las variables continuas, primero haremos un estudio descriptivo por grupo de medias y medianas:

```
library(psych)
describeBy(dbgset[,c(3:8,12,13)], dbgset$severity)
##  Descriptive statistics by group
## group: Control
##                        vars  n    mean     sd  median trimmed    mad
min
## age_enroll:ch1            1 42   60.46   9.07   60.30   60.04  10.16
45.80
## ats_packyears:ch1         2 42   44.49  28.48   38.75   40.85  24.46
11.00
## bmi:ch1                   3 42   27.98   5.00   26.84   27.67   4.94
19.69
## distwalked:ch1            4 41 1697.00 331.24 1657.00 1685.27 292.07
970.00
## fev1_fvc_utah:ch1         5 42    0.77   0.04    0.77    0.77   0.04
0.70
## fev1pp_utah:ch1           6 42   97.93  13.85   96.50   96.65  14.08
80.00
## pctemph_slicer:ch1        7 39    1.37   1.56    0.76    1.08   0.53
0.09
## pctgastrap_slicer:ch1     8 39    8.40   5.49    8.27    7.89   4.87
0.40
##                            max   range  skew kurtosis    se
```

```
## age_enroll:ch1          79.60   33.80  0.28    -0.85  1.40
## ats_packyears:ch1      145.00  134.00  1.32     1.90  4.40
## bmi:ch1                 40.76   21.07  0.53    -0.55  0.77
## distwalked:ch1        2485.00 1515.00  0.34    -0.27 51.73
## fev1_fvc_utah:ch1        0.85    0.15 -0.16    -1.02  0.01
## fev1pp_utah:ch1        135.00   55.00  0.72    -0.28  2.14
## pctemph_slicer:ch1       7.85    7.76  2.40     6.02  0.25
## pctgastrap_slicer:ch1   24.37   23.98  0.95     0.72  0.88
## ------------------------------------------------------------
## group: MildMod
##                       vars  n    mean     sd  median trimmed    mad
min
## age_enroll:ch1           1 52   62.16   8.99   62.20   62.03  10.75
45.00
## ats_packyears:ch1        2 52   44.43  24.90   39.35   41.72  22.54
12.00
## bmi:ch1                  3 52   29.20   5.90   27.98   28.79   4.06
17.37
## distwalked:ch1           4 52 1433.40 414.46 1495.00 1459.17 294.30
200.00
## fev1_fvc_utah:ch1        5 52    0.63   0.10    0.65    0.64   0.07
0.39
## fev1pp_utah:ch1          6 52   69.04  11.61   68.00   68.50  10.38
50.00
## pctemph_slicer:ch1       7 46    4.86   6.32    2.58    3.59   3.45
0.03
## pctgastrap_slicer:ch1    8 44   20.00  16.22   15.22   17.65  11.34
0.68
##                          max   range  skew kurtosis    se
## age_enroll:ch1         79.50   34.50  0.12    -1.01  1.25
## ats_packyears:ch1      135.00  123.00  1.24     1.93  3.45
## bmi:ch1                 47.76   30.39  0.84     1.22  0.82
## distwalked:ch1        2109.00 1909.00 -0.74     0.65 57.47
## fev1_fvc_utah:ch1        0.79    0.40 -0.69     0.01  0.01
## fev1pp_utah:ch1        115.00   65.00  1.02     2.67  1.61
## pctemph_slicer:ch1      24.63   24.59  1.86     2.57  0.93
## pctgastrap_slicer:ch1   67.52   66.84  1.26     1.04  2.45
## ------------------------------------------------------------
## group: Severe
##                       vars  n    mean     sd  median trimmed    mad
min
## age_enroll:ch1           1 42   66.90   6.35   66.90   66.91   7.26
54.10
## ats_packyears:ch1        2 42   55.27  25.66   48.00   52.29  20.39
12.00
## bmi:ch1                  3 42   26.25   6.03   25.79   25.82   5.77
15.57
## distwalked:ch1           4 38 1089.76 310.62 1095.50 1102.47 260.20
212.00
## fev1_fvc_utah:ch1        5 42    0.38   0.09    0.38    0.38   0.09
0.23
## fev1pp_utah:ch1          6 42   32.90  10.11   33.00   33.65  11.86
9.00
## pctemph_slicer:ch1       7 32   21.80  10.00   19.80   21.86  10.64
```

```
0.97
## pctgastrap_slicer:ch1    8 32   55.51  13.27   57.84   57.05  11.59
20.29
##                            max   range  skew kurtosis    se
## age_enroll:ch1           79.40   25.30  0.01    -0.94  0.98
## ats_packyears:ch1       124.00  112.00  1.00     0.50  3.96
## bmi:ch1                  43.32   27.75  0.64     0.18  0.93
## distwalked:ch1         1670.00 1458.00 -0.47     0.64 50.39
## fev1_fvc_utah:ch1         0.61    0.38  0.24    -0.31  0.01
## fev1pp_utah:ch1          46.00   37.00 -0.43    -0.86  1.56
## pctemph_slicer:ch1       45.06   44.09  0.06    -0.32  1.77
## pctgastrap_slicer:ch1    77.60   57.31 -0.92     0.45  2.35

summary(subset(dbgset[,c(3:8,12,13)], dbgset$severity=='Control'))

##  age_enroll:ch1  ats_packyears:ch1    bmi:ch1      distwalked:ch1
##  Min.   :45.80   Min.   : 11.00    Min.   :19.69   Min.   : 970
##  1st Qu.:53.23   1st Qu.: 22.12    1st Qu.:24.16   1st Qu.:1460
##  Median :60.30   Median : 38.75    Median :26.84   Median :1657
##  Mean   :60.46   Mean   : 44.49    Mean   :27.98   Mean   :1697
##  3rd Qu.:66.35   3rd Qu.: 53.25    3rd Qu.:32.09   3rd Qu.:1810
##  Max.   :79.60   Max.   :145.00    Max.   :40.76   Max.   :2485
##                                                    NA's   :1
##  fev1_fvc_utah:ch1 fev1pp_utah:ch1  pctemph_slicer:ch1
##  Min.   :0.7000    Min.   : 80.00   Min.   :0.08864
##  1st Qu.:0.7425    1st Qu.: 87.25   1st Qu.:0.46147
##  Median :0.7700    Median : 96.50   Median :0.76225
##  Mean   :0.7698    Mean   : 97.93   Mean   :1.36673
##  3rd Qu.:0.8000    3rd Qu.:105.75   3rd Qu.:1.46471
##  Max.   :0.8500    Max.   :135.00   Max.   :7.85244
##                                     NA's   :3
##  pctgastrap_slicer:ch1
##  Min.   : 0.3955
##  1st Qu.: 4.5027
##  Median : 8.2715
##  Mean   : 8.4000
##  3rd Qu.:10.7149
##  Max.   :24.3711
##  NA's   :3

summary(subset(dbgset[,c(3:8,12,13)], dbgset$severity=='MildMod'))

##  age_enroll:ch1  ats_packyears:ch1    bmi:ch1      distwalked:ch1
##  Min.   :45.00   Min.   : 12.00    Min.   :17.37   Min.   : 200
##  1st Qu.:55.48   1st Qu.: 26.38    1st Qu.:25.75   1st Qu.:1262
##  Median :62.20   Median : 39.35    Median :27.98   Median :1495
##  Mean   :62.16   Mean   : 44.43    Mean   :29.20   Mean   :1433
##  3rd Qu.:69.33   3rd Qu.: 55.33    3rd Qu.:32.19   3rd Qu.:1654
##  Max.   :79.50   Max.   :135.00    Max.   :47.76   Max.   :2109
##
##  fev1_fvc_utah:ch1 fev1pp_utah:ch1  pctemph_slicer:ch1
##  Min.   :0.3900    Min.   : 50.00   Min.   : 0.03225
##  1st Qu.:0.5900    1st Qu.: 60.75   1st Qu.: 0.90163
##  Median :0.6500    Median : 68.00   Median : 2.58373
##  Mean   :0.6298    Mean   : 69.04   Mean   : 4.86495
```

```
##  3rd Qu.:0.6800    3rd Qu.: 75.00   3rd Qu.: 5.05148
##  Max.   :0.7900    Max.   :115.00   Max.   :24.62650
##                                     NA's   :6
##  pctgastrap_slicer:ch1
##  Min.   : 0.6805
##  1st Qu.: 8.1660
##  Median :15.2248
##  Mean   :20.0013
##  3rd Qu.:27.9914
##  Max.   :67.5233
##  NA's   :8
```

```
summary(subset(dbgset[,c(3:8,12,13)], dbgset$severity=='Severe'))
```

```
##  age_enroll:ch1  ats_packyears:ch1    bmi:ch1      distwalked:ch1
##  Min.   :54.10   Min.   : 12.00    Min.   :15.57   Min.   : 212
##  1st Qu.:62.02   1st Qu.: 38.20    1st Qu.:21.96   1st Qu.: 944
##  Median :66.90   Median : 48.00    Median :25.79   Median :1096
##  Mean   :66.90   Mean   : 55.27    Mean   :26.25   Mean   :1090
##  3rd Qu.:71.78   3rd Qu.: 67.88    3rd Qu.:29.53   3rd Qu.:1279
##  Max.   :79.40   Max.   :124.00    Max.   :43.32   Max.   :1670
##                                                    NA's   :4
##  fev1_fvc_utah:ch1 fev1pp_utah:ch1 pctemph_slicer:ch1
##  Min.   :0.2300    Min.   : 9.00   Min.   : 0.9713
##  1st Qu.:0.3225    1st Qu.:25.25   1st Qu.:16.5190
##  Median :0.3800    Median :33.00   Median :19.7965
##  Mean   :0.3798    Mean   :32.90   Mean   :21.7968
##  3rd Qu.:0.4400    3rd Qu.:41.75   3rd Qu.:27.5161
##  Max.   :0.6100    Max.   :46.00   Max.   :45.0630
##                                    NA's   :10
##  pctgastrap_slicer:ch1
##  Min.   :20.29
##  1st Qu.:49.14
##  Median :57.84
##  Mean   :55.51
##  3rd Qu.:65.38
##  Max.   :77.60
##  NA's   :10
```

Revisamos la distribución normal:

```
library("car")

par(mfrow=c(2,4))
qqPlot(dbgset$`age_enroll:ch1`)

## [1]  22 120

qqPlot(dbgset$`ats_packyears:ch1`)

## [1]  36 108

qqPlot(dbgset$`bmi:ch1`)
```

```
## [1] 29 45
```

```
qqPlot(dbgset$`distwalked:ch1`)
```

```
## [1] 58 91
```

```
qqPlot(dbgset$`fev1_fvc_utah:ch1`)
```

```
## [1] 37 62
```

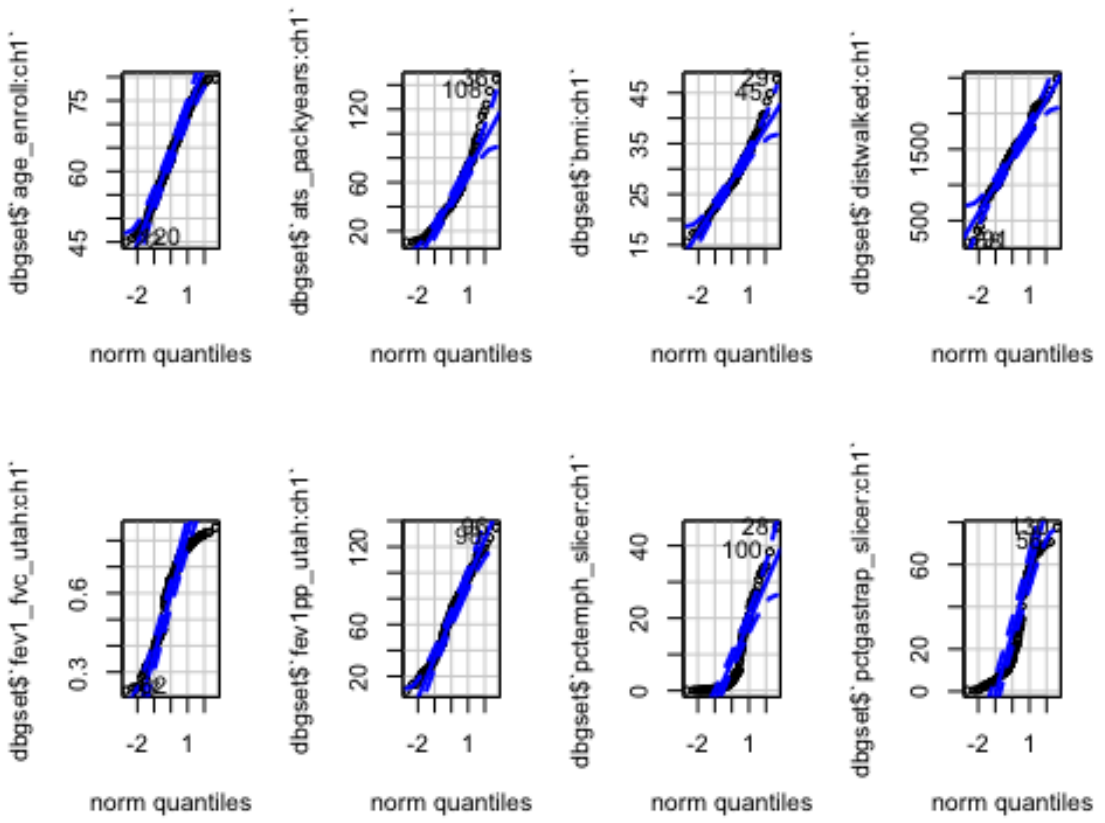
```
qqPlot(dbgset$`fev1pp_utah:ch1`)
```

```
## [1] 96 98
```

```
qqPlot(dbgset$`pctemph_slicer:ch1`)
```

```
## [1]  28 100
```

```
qqPlot(dbgset$`pctgastrap_slicer:ch1`)
```




```
## [1] 130  56
```

Vemos en las curvas QQ que las variables edad, porcentaje predicho del FEV1 y la distancia caminada en el test de la marcha siguen una clara distribución normal, mientras que el resto no. Lo comprobamos con el test de Shapiro-Wilks:

```
#Usamos el test de Shapiro-Wilks para comprobar la normalidad:
shapiro.test(dbgset$`age_enroll:ch1`)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dbgset$`age_enroll:ch1`
## W = 0.98205, p-value = 0.07045

shapiro.test(dbgset$`ats_packyears:ch1`)

##
##  Shapiro-Wilk normality test
##
## data:  dbgset$`ats_packyears:ch1`
## W = 0.91692, p-value = 4.151e-07

shapiro.test(dbgset$`bmi:ch1`)

##
##  Shapiro-Wilk normality test
##
## data:  dbgset$`bmi:ch1`
## W = 0.97337, p-value = 0.00909

shapiro.test(dbgset$`distwalked:ch1`)

##
##  Shapiro-Wilk normality test
##
## data:  dbgset$`distwalked:ch1`
## W = 0.98624, p-value = 0.212

shapiro.test(dbgset$`fev1_fvc_utah:ch1`)

##
##  Shapiro-Wilk normality test
##
## data:  dbgset$`fev1_fvc_utah:ch1`
## W = 0.91862, p-value = 5.267e-07

shapiro.test(dbgset$`fev1pp_utah:ch1`)

##
##  Shapiro-Wilk normality test
##
## data:  dbgset$`fev1pp_utah:ch1`
## W = 0.98361, p-value = 0.1025

shapiro.test(dbgset$`pctemph_slicer:ch1`)

##
##  Shapiro-Wilk normality test
##
## data:  dbgset$`pctemph_slicer:ch1`
## W = 0.76323, p-value = 1.875e-12

shapiro.test(dbgset$`pctgastrap_slicer:ch1`)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dbgset$`pctgastrap_slicer:ch1`
## W = 0.85092, p-value = 2.136e-09
```

Comprobamos lo observado en las curvas. Procedemos a hacer las comparaciones de medias, empezando por las variables con distribución normal:

Edad:

```
ageaov<- aov(dbgset$'age_enroll:ch1' ~ severity, data = dbgset)
summary.aov(ageaov)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## severity      2    945   472.6   6.871 0.00145 **
## Residuals   133   9148    68.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Distancia caminada en pies:

```
wtaov<- aov(dbgset$`distwalked:ch1` ~ severity, data = dbgset)
summary.aov(wtaov)
```

```
##              Df   Sum Sq Mean Sq F value   Pr(>F)
## severity      2  7297503 3648752   27.93 8.57e-11 ***
## Residuals   128 16719275  130619
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 5 observations deleted due to missingness
```

FEV1:

```
fevaov<- aov(dbgset$`fev1pp_utah:ch1` ~ severity, data = dbgset)
summary.aov(fevaov)
```

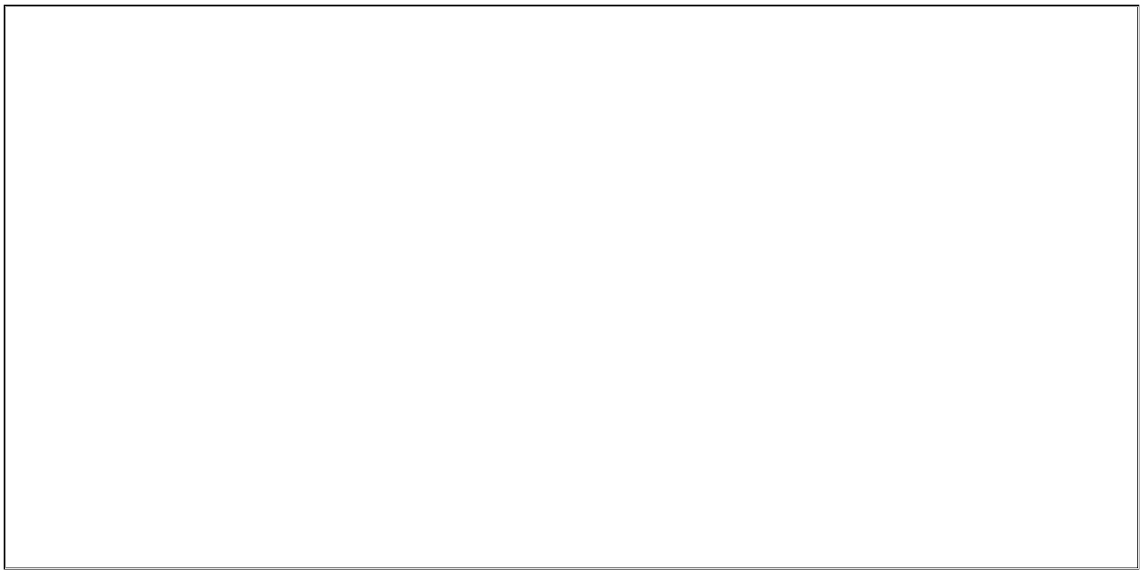
```
##              Df Sum Sq Mean Sq F value Pr(>F)
## severity      2  89211   44606   313.4 <2e-16 ***
## Residuals   133  18930     142
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ahora con las variables no-normales:

```
kruskal.test(dbgset$`ats_packyears:ch1` ~ severity, data = dbgset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  dbgset$`ats_packyears:ch1` by severity
## Kruskal-Wallis chi-squared = 6.7276, df = 2, p-value = 0.0346
```

```
kruskal.test(dbgset$`bmi:ch1` ~ severity, data = dbgset)

##
##  Kruskal-Wallis rank sum test
##
## data:  dbgset$`bmi:ch1` by severity
## Kruskal-Wallis chi-squared = 6.7435, df = 2, p-value = 0.03433

kruskal.test(dbgset$`fev1_fvc_utah:ch1` ~ severity, data = dbgset)

##
##  Kruskal-Wallis rank sum test
##
## data:  dbgset$`fev1_fvc_utah:ch1` by severity
## Kruskal-Wallis chi-squared = 105.11, df = 2, p-value < 2.2e-16

kruskal.test(dbgset$`pctemph_slicer:ch1` ~ severity, data = dbgset)

##
##  Kruskal-Wallis rank sum test
##
## data:  dbgset$`pctemph_slicer:ch1` by severity
## Kruskal-Wallis chi-squared = 60.807, df = 2, p-value = 6.252e-14

kruskal.test(dbgset$`pctgastrap_slicer:ch1` ~ severity, data = dbgset)

##
##  Kruskal-Wallis rank sum test
##
## data:  dbgset$`pctgastrap_slicer:ch1` by severity
## Kruskal-Wallis chi-squared = 67.514, df = 2, p-value = 2.186e-15
```



Parte 2: Datos Ómicos

Vamos con los datos de expresión:

```
#Creo un elemento que luego usaré como nombres de columnas en la matri
z de expresión y con los datos brutos:
code <- paste0(substr(dbgset$title,1,4), substr(dbgset$geo_accession,8
,10))

#Cargo los datos brutos de los archivos .cel que he descargado del rep
ositorio
rawdat<- ReadAffy(celfile.path = "/Users/gsuarezcuartin/Documents/Mást
er/TFM/data")

#Cambio los nombres de las muestras para facilitar el estudio
sampleNames(rawdat)<- code
sampleNames(rawdat)[1:10]

##  [1] "COPD549" "COPD550" "COPD551" "COPD552" "Cont553" "Cont554" "C
ont555"
##  [8] "COPD556" "COPD557" "COPD558"

#Cargo de nuevo los datos en otro objeto al que no modificaré los nomb
res
#para usar luego en el análisis integrativo
rawdat0<- ReadAffy(celfile.path = "/Users/gsuarezcuartin/Documents/Más
ter/TFM/data")

#Extraemos los datos para crear la matriz de expresión
exs<-exprs(gsetx)

#Modifico los nombres de columnas para facilitar su comprensión
colnames(exs)<- code
str(exs)

##  num [1:12531, 1:136] 7.97 8.72 11.95 6.46 7.34 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:12531] "117_at" "1294_at" "1405_i_at" "1552256_a_at"
...
##   ..$ : chr [1:136] "COPD549" "COPD550" "COPD551" "COPD552" ...
```

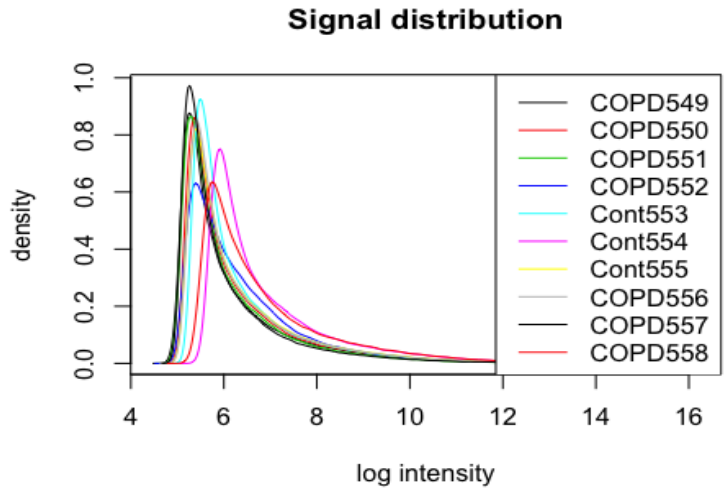
Tenemos entonces 12531 filas que representan las sondas, y 136 columnas que representan los sujetos incluidos en el estudio.

2.1 Exploración de los datos

Procedemos a realizar la exploración de los datos brutos y control de calidad de los mismos. Dado que son un gran número de arrays, usaremos a manera de representación las 10 primeras muestras.

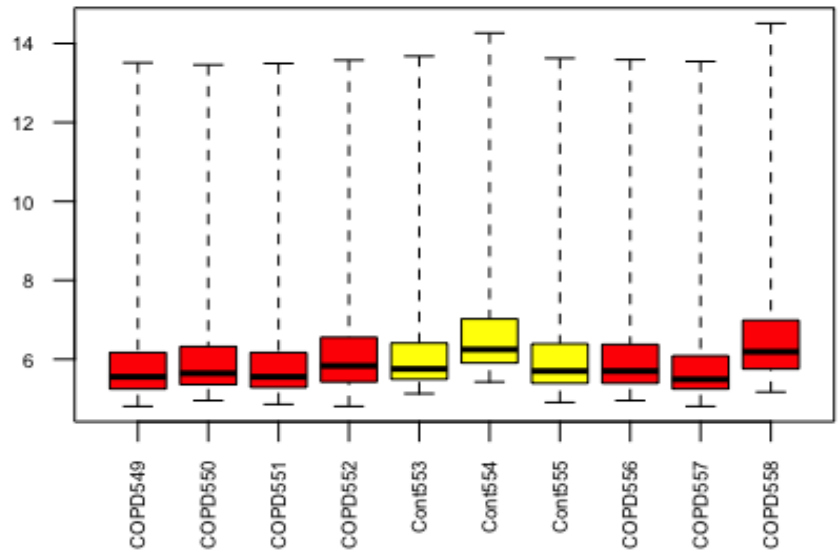
```
#Empezamos valorando la distribución de la señal de las primeras 10 mu
estras
hist(rawdat[,1:10], main = "Signal distribution", col = 1:ncol(rawdat)
, lty = 1)
```

```
legend(x = "topright", legend = colnames(rawdat[,1:10]), col = 1:ncol(
rawdat), lty = 1)
```



Vemos que, en los gráficos de densidad de los arrays, no parecen existir problemas significativos, aunque algunos parecen tener menor densidad e intensidad, por lo que podría ser necesario la normalización de los datos.

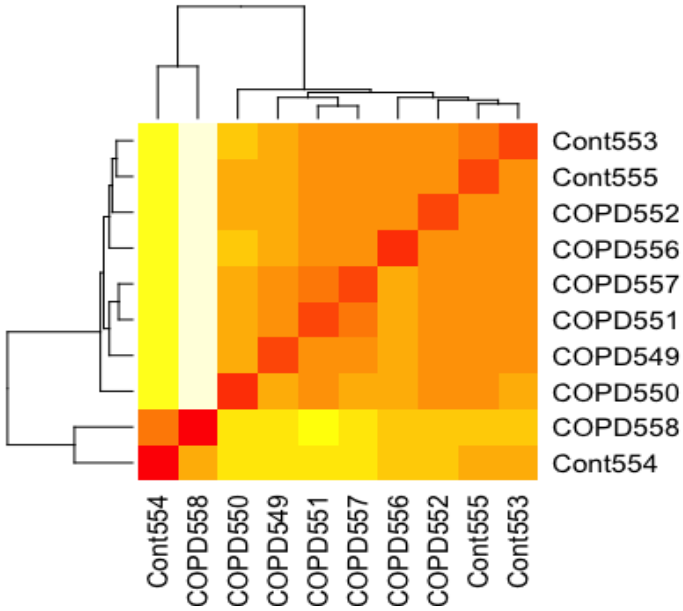
```
info<- data.frame(grupo=substr(sampleNames(rawdat), 1, 4))
boxplot(rawdat[,1:10], cex.axis = 0.6, col = c('yellow','red')[info$gr
upo],
        las = 2, names = sampleNames(rawdat[,1:10]))
```



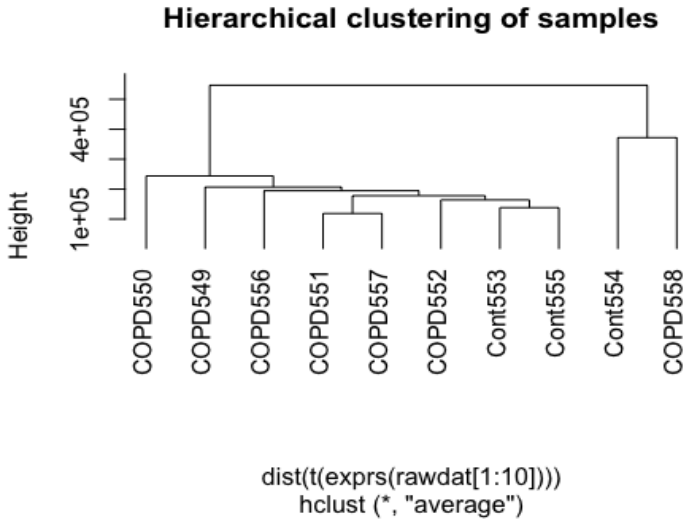
En el boxplot de los 10 primeros arrays de nuevo muestra una discreta heterogeneidad, pero no se observan cambios que sugieran grandes problemas en los mismos.

Seguidamente realizaremos un mapa de colores para visualizar la matriz de distancia entre muestras; en vista del gran número de muestras, por lo que usaremos nuevamente las 10 primeros arrays para la exploración:

```
#Mapa de colores
matdist2 <- dist(t(exprs(rawdat[1:10])))

heatmap(as.matrix(matdist2), col=heat.colors(16))
```




```
#Dendograma
clust.euclid.average <- hclust(dist(t(exprs(rawdat[1:10]))), method =
"average")

plot(clust.euclid.average, labels = sampleNames(rawdat[1:10]),
     main ="Hierarchical clustering of samples", hang = -1)
```

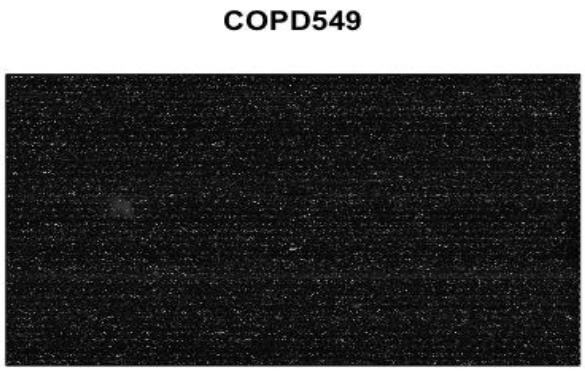


A grandes rasgos, en ambos gráficos parece que, salvo el último grupo, las muestras se asocian por condición de enfermos o controles. Sin embargo, dado que incluso dentro de los enfermos existe una gran heterogeneidad, es esperable que los grupos no sean del todo definidos.

2.2 Control de calidad

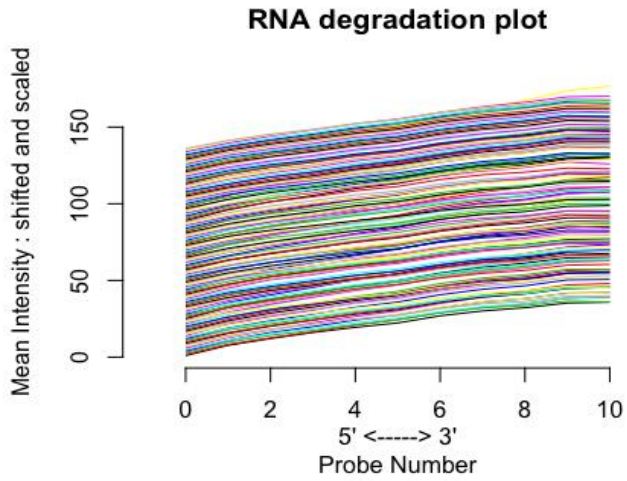
Buscamos una imagen del escaneado del primer array para descartar algún problema con el mismo:

```
image(rawdat[, 1])
```



No observamos algún problema significativo en la imagen. Seguidamente realizamos un gráfico de degradación del RNA de todos los arrays.

```
deg <- AffyRNAdeg(rawdat, log.it = T)
plotAffyRNAdeg(deg, lwd=2, col=1:ncol(rawdat))
```




```
#No he añadido leyenda debido al gran número de arrays y la ausencia d
e anomalías.
```

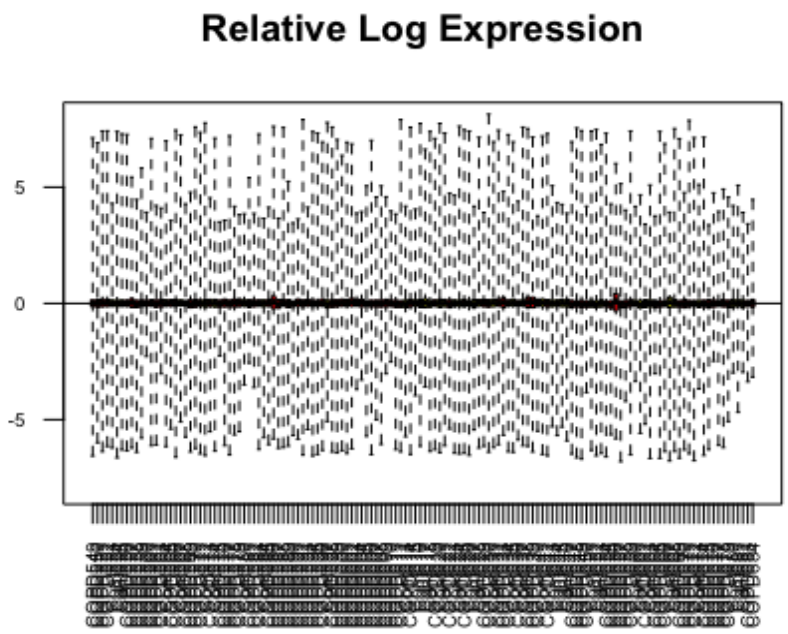
Las líneas de degradación son bastante paralelas, lo que sugiere una degradación de RNA similar entre los chips, traduciendo una calidad similar.

Realizaremos ahora un control de calidad a nivel de sonda (PLM):

```
library(affyPLM)

computePLM <- T
if (computePLM) {
  Pset <- fitPLM(rawdat)
  save(Pset, file = file.path(dataDir, "PLM.Rda"))
  } else {
    load(file = file.path(dataDir, "PLM.Rda"))
    }
```

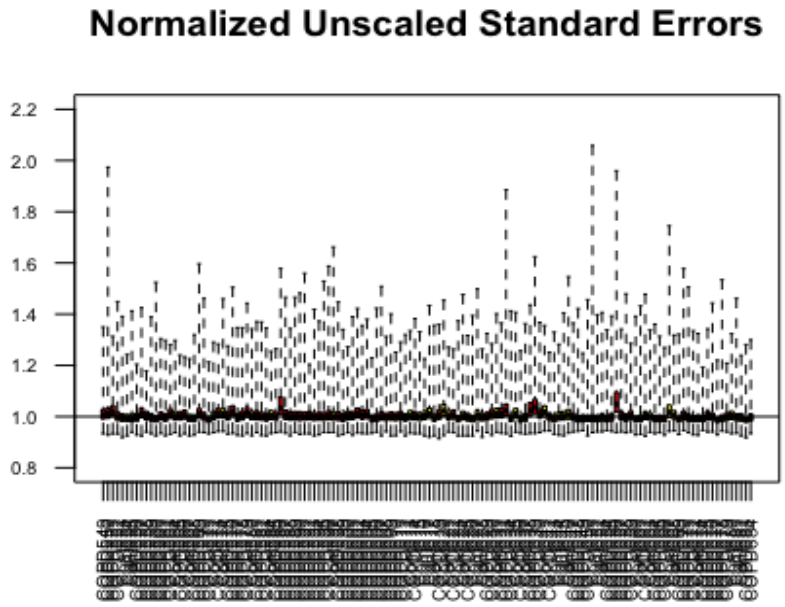
Posterior a este ajuste PLM, obtendremos los gráficos de expresiones relativas:

```
RLE(Pset, main = "Relative Log Expression",names =sampleNames(rawdat),
    las= 2, col= c('yellow','red')[info$grupo], cex.axis = 0.6, ylim =
c(-8,8))
```



..Y con los errores normalizados:

```
NUSE(Pset,main="Normalized Unscaled Standard Errors", las=2,
 names= sampleNames(rawdat), las=2, col=c('yellow','red')[info$grupo],
 cex.axis = 0.6, ylim = c(0.8,2.2))
```



Observamos que en ambos diagramas los datos están centrados y son similares, y en el caso del RLE existe una clara simetría de los datos. Solo destacan dos cajas ligeramente por encima del resto en el gráfico de NUSE. Deducimos que no existen problemas significativos en la calidad de los datos.

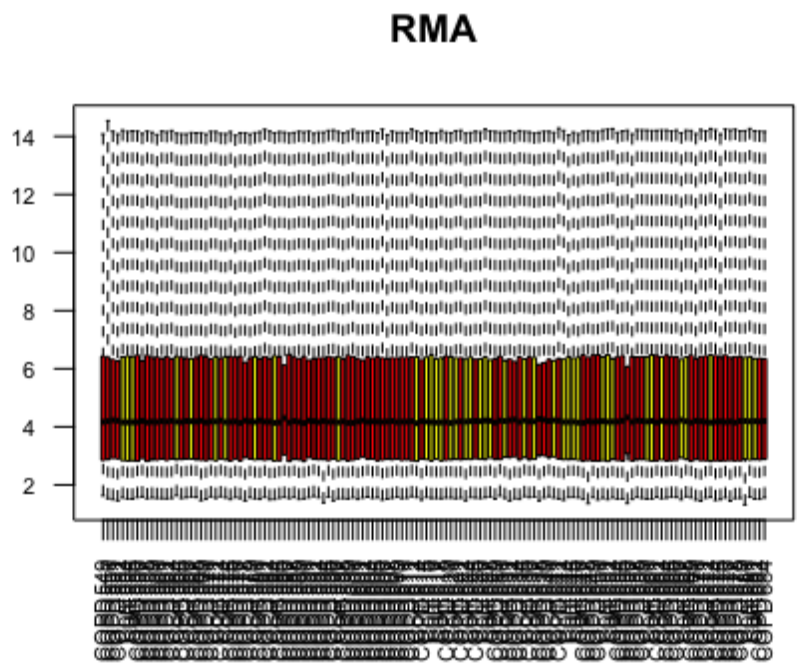
2.3 Normalización de los datos

Procedemos a realizar la normalización de los datos, usando el método RMA, que normaliza y corrige el ruido de fondo. Usaremos los datos con los nombres originales, ya que son los que nos servirán luego para el análisis integrativo:

```
gset_rma <- rma(rawdat0)

## Background correcting
## Normalizing
## Calculating Expression
```

Realizamos luego un boxplot para confirmar que los datos están normalizados a valores comparables.

```
boxplot(gset_rma,main="RMA", names=sampleNames(rawdat), cex.axis=0.7,
        col=c('yellow','red')[info$grupo] ,las=2)
```



Confirmamos que los valores son ahora comparables.

2.4 Filtraje

Realizaremos un filtraje no específico para eliminar aquellos genes con poca variación entre condiciones y los que no dispongan de anotaciones.

```
library(genefilter)

library(hgu133plus2.db)

filt <- nsFilter(gset_rma, require.entrez=TRUE,
                 remove.dupEntrez=TRUE, var.func=IQR,
                 var.cutoff=0.5, var.filter=TRUE,
                 filterByQuantile=TRUE, feature.exclude="^AFFX")

#Miramos la distribución de elementos filtrados
print(filt$filter.log)

## $numDupsRemoved
## [1] 21739
##
## $numLowVar
## [1] 10093
##
## $numRemoved.ENTREZID
## [1] 12740
##
## $feature.exclude
## [1] 10
```

```
#Y guardamos el objeto resultante
gset_filt <-filt$eset
```

Ahora guardaremos los datos resultantes del normalizado y filtraje en un elemento tipo ExpressionSet.

```
save(gset_rma, gset_filt, file=file.path(resultsDir, "normdat.Rda"))

#Obtenemos también los datos en formato de archivo de texto:
write.csv2(exprs(gset_rma), file.path(resultsDir, "normdat.csv2"))
```

2.5 Selección de genes diferencialmente expresados

Cargamos el paquete limma:

```
library(limma)
```

Hacemos el análisis: Nos interesa comparar los genes diferencialmente expresados entre los pacientes sanos (controles) y los enfermos (COPD) leves-moderados o graves.

```
#Diseñamos la matriz
sevfact<- factor(dbgset$severity)
design<- model.matrix(~ sevfact + 0)
colnames(design)<- c("Control", "MildMod", "Severe")
rownames(design) <- sampleNames(rawdat)
head(design)
```

```
##         Control MildMod Severe
## COPD549       0       0      1
## COPD550       0       1      0
## COPD551       0       1      0
## COPD552       0       0      1
## Cont553       1       0      0
## Cont554       1       0      0
```

Procedemos a crear la matriz de contrastes con las comparaciones que queremos hacer:

```
#Creamos la matriz de contrastes
cont.matrix <- makeContrasts (ControlvsMilMod = MildMod-Control,
                              ControlvsSevere = Severe-Control,
                              MildModvsSevere = Severe-MildMod, levels
=design)
print(cont.matrix)
```

```
##          Contrasts
## Levels    ControlvsMilMod ControlvsSevere MildModvsSevere
##   Control              -1              -1               0
##   MildMod               1               0              -1
##   Severe                0               1               1
```

Finalmente creamos el modelo con el conjunto de genes normalizados y filtrados:

```
#Creamos el modelo a comparar
fitlm <- lmFit(gset_filt, design)
fit <-contrasts.fit(fitlm, cont.matrix)

#Procesamos los datos con el estadístico empírico de Bayes
fit <- eBayes(fit)
```

Ahora realizaremos una tabla para las tres comparaciones: Controles vs EPOC leve-moderado, Controles Vs EPOC grave y EPOC leve-moderado vs grave.

```
#Construimos la tabla de los genes más diferenciados Controles Vs EPOC
Leve-Moderado:
topTCvsMM <- topTable(fit, number=nrow(fit), coef="ControlvsMilMod", a
djust="fdr")
head(topTCvsMM)
```

```
##                   logFC  AveExpr         t     P.Value adj.P.Val
B
## 218897_at    -0.2850410 4.189976 -3.384139 0.000929524 0.9843915 -1.
637307
## 228806_at    -0.3361878 4.877183 -3.128500 0.002143512 0.9843915 -2.
117245
## 205514_at    -0.2490725 2.754582 -3.078300 0.002511546 0.9843915 -2.
207895
## 234072_at     0.2089254 4.487781  3.070605 0.002572870 0.9843915 -2.
221684
## 239435_x_at  0.3088578 6.790405  3.067139 0.002600942 0.9843915 -2.
227885
## 207738_s_at -0.2102501 4.311065 -3.019389 0.003017601 0.9843915 -2.
312733
```

```
#Construimos la tabla de los genes más diferenciados Controles Vs EPOC
Grave:
topTCvsS <- topTable(fit, number=nrow(fit), coef="ControlvsSevere", ad
just="fdr")
head(topTCvsS)
```

```
##                   logFC  AveExpr         t      P.Value    adj.P.Val
## 225314_at    -0.3573427 8.824944 -5.560754 1.341701e-07 0.0009636436
## 221602_s_at -0.5879187 8.749571 -5.485056 1.909528e-07 0.0009636436
## 217950_at    -0.4172597 9.622054 -5.391312 2.944568e-07 0.0009906509
## 227552_at    -0.4812024 8.011767 -5.201715 6.974280e-07 0.0015642055
## 226482_s_at -0.3440423 9.111201 -5.178269 7.748962e-07 0.0015642055
## 219513_s_at -0.3485224 6.201044 -5.057196 1.328707e-06 0.0022351071
##                     B
## 225314_at    7.151644
## 221602_s_at 6.829900
## 217950_at    6.435254
## 227552_at    5.650310
```

```
## 226482_s_at 5.554502
## 219513_s_at 5.064307

#Construimos la tabla de los genes más diferenciados EPOC leve-Mod Vs
EPOC Grave:
topTMMvsS <- topTable(fit, number=nrow(fit), coef="MildModvsSevere", a
djust="fdr")
head(topTMMvsS)

##                 logFC  AveExpr         t      P.Value  adj.P.Val
B
## 225314_at  -0.2698951 8.824944 -4.417703 1.999386e-05 0.09328927 2.
480490
## 58780_s_at  0.4092180 7.270406  4.321262 2.946481e-05 0.09328927 2.
148623
## 203066_at   0.3902173 9.663447  4.296365 3.253654e-05 0.09328927 2.
063821
## 213222_at   0.4483731 7.560763  4.264136 3.697187e-05 0.09328927 1.
954584
## 219543_at   0.2653044 4.606935  4.093360 7.197260e-05 0.09852547 1.
386054
## 210166_at   0.3976115 7.654864  4.082791 7.495531e-05 0.09852547 1.
351448
```

Veremos ahora el resumen de las diferencias en las comparaciones con un p-valor<0.01 ajustado por el método de Benjamini y Hochberg de la tasa de falsos descubrimientos (fdr):

```
#Creamos una tabla para resumir los hallazgos
COPDsym <- mget(rownames(fit), hgu133plus2SYMBOL)
COPDtest <- decideTests(fit, method = "separate", adjust.method = "fdr
",
                        p.value = 0.01)
sum.COPDtest.rows<-apply(abs(COPDtest),1,sum)
print(summary(COPDtest))

##        ControlvsMilMod ControlvsSevere MildModvsSevere
## Down                 0              84               0
## NotSig           10093            9965           10093
## Up                   0              44               0
```

Vemos que, de las 3 comparaciones, donde existen realmente diferencias significativas es entre los controles y enfermos con EPOC grave, por lo que solo usaremos esta información de aquí en adelante.

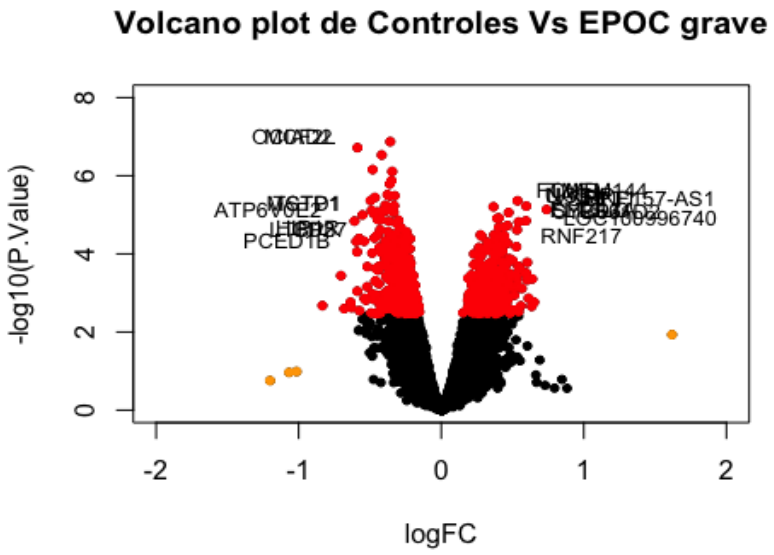
```
#Buscamos los nombres de los 20 genes más diferencialmente expresados
en la tabla a usar
TopGen <- unlist(mget(rownames(topTCvsS[1:20,]), hgu133plus2SYMBOL))
TT<- topTCvsS[1:20,]
```

```
row.names(TT)<- TopGen
round(TT[,-6],4)
```

```
##                 logFC AveExpr       t P.Value adj.P.Val
## OCIAD2       -0.3573  8.8249 -5.5608       0    0.0010
## FCMR         -0.5879  8.7496 -5.4851       0    0.0010
## NOSIP        -0.4173  9.6221 -5.3913       0    0.0010
## SEPT1        -0.4812  8.0118 -5.2017       0    0.0016
## TSTD1        -0.3440  9.1112 -5.1783       0    0.0016
## SH2D3A       -0.3485  6.2010 -5.0572       0    0.0022
## ATP6V0E2     -0.3612  8.2014 -5.0107       0    0.0024
## CYB561D2     -0.3768  6.0376 -4.8665       0    0.0033
## LOC100996740 -0.3268 10.5262 -4.8514       0    0.0033
## HIP1R        -0.3600  6.4418 -4.8488       0    0.0033
## CD27         -0.4653  8.9539 -4.8281       0    0.0033
## LTBP3        -0.4953  7.1622 -4.7860       0    0.0034
## RNF217        0.5349  4.5943  4.7830       0    0.0034
## PCED1B       -0.3919  8.7975 -4.7211       0    0.0035
## MCF2L        -0.3301  5.5413 -4.7131       0    0.0035
## TMEM144       0.5963  5.1315  4.7096       0    0.0035
## UACA          0.3660  4.2135  4.6994       0    0.0035
## RNF157-AS1   -0.3637  3.4678 -4.6617       0    0.0035
## MCTP1         0.7393  7.1006  4.6617       0    0.0035
## FLT3LG       -0.4885  7.3207 -4.6480       0    0.0035
```

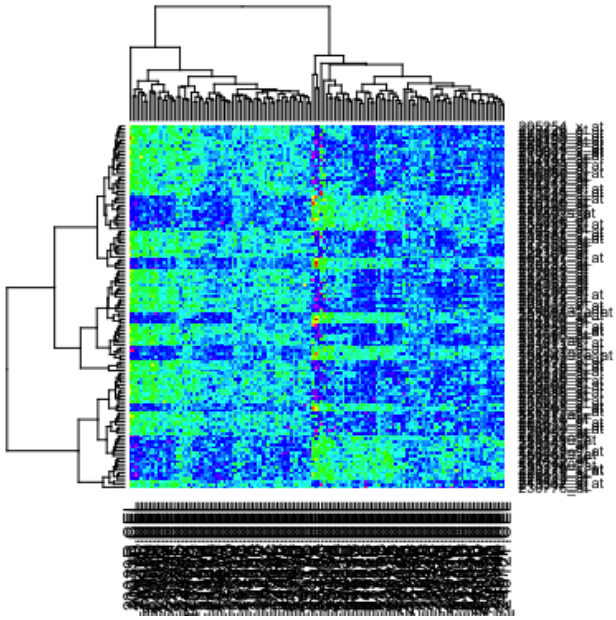
Procedemos a obtener los gráficos para la comparación de controles Vs EPOC grave

```
library(calibrate) #Cargamos el paquete para añadir texto

with(topTCvsS, plot(logFC, -log10(P.Value), pch=20,
                    main="Volcano plot de Controles Vs EPOC grave",
                    xlim=c(-2,2), ylim=c(0,8)))
#Especificamos los colores: rojo para p-valor < 0.05
with(subset(topTCvsS, adj.P.Val<.05 ), points(logFC, -log10(P.Value),
                                              pch=20, col="red"))
#naranja para logFC > 1
with(subset(topTCvsS, abs(logFC)>1), points(logFC, -log10(P.Value),
                                            pch=20, col="orange"))
#y verde para p-valor < 0.05 y logFC > 1
with(subset(topTCvsS, adj.P.Val<.05 & abs(logFC)>1), points(logFC, -lo
g10(P.Value),
                                                            pch=20, col="g
reen"))
#Añadimos las etiquetas a los genes con p-valor < 0.01 y logFC > 0.5
#Por motivos de espacio, marcaremos los 20 genes más diferencialmente
expresados.
with(subset(topTCvsS, adj.P.Val<.01 & abs(logFC)>0.5),
     textxy(logFC, -log10(P.Value),
            labs=unlist(mget(rownames(topTCvsS[1:20,]), hgu133plus2SYM
BOL)),
            cex=0.8, offset = 0.8)
     )
```



Observamos que no hay puntos verdes, por lo que no hay genes que cumplan ambas condiciones con este conjunto de datos; es decir, que hay algunos genes con un cambio de expresión > 1, pero no parecen corresponder con aquellos con p-valor más bajo. Creamos un heatmap para obtener una idea general de los resultados:

```
probeNames<-rownames(COPDtest)
probeNames.sel<-probeNames[sum.COPDtest.rows!=0]
expclust <-exprs(gset_filt)[probeNames.sel,]
heatmap(expclust, col=rainbow(100), cexCol=0.9)
```



A grandes rasgos, vemos que no son muchos los genes que estan sobreexpresados en este conjunto de datos.

##2.6 Anotación de resultados

Obtendremos el análisis de los genes encontrados en el paso previo.

```
library(annotate)

## Loading required package: XML

#Buscaremos los 5 genes más diferencialmente expresados y obtendremos
las rutas biológicas en las que participan, usando el conjunto de dato
s de Homo sapiens:
topCOPD1 <-rownames(topTCvsS[c(1:5),])
COPDsym2 <- getSYMBOL(topCOPD1, "hgu133plus2.db")
GOcopd1 <- mget(topCOPD1, hgu133plus2GO)
for (i in 1:length(GOcopd1)){
  for (j in 1:length(GOcopd1[[i]])){
    GOcopd <- GOcopd1[[i]][[j]][[1]]
    cat(topCOPD1[i],COPDsym2[i],GOcopd,substr(Term(GOcopd),1,100), "\n
")
    }}

## 225314_at OCIAD2 GO:0009617 response to bacterium
## 225314_at OCIAD2 GO:0005768 endosome
## 221602_s_at FCMR GO:0002376 immune system process
## 221602_s_at FCMR GO:0006968 cellular defense response
## 221602_s_at FCMR GO:0043066 negative regulation of apoptotic proces
s
## 221602_s_at FCMR GO:0005576 extracellular region
## 221602_s_at FCMR GO:0016021 integral component of membrane
## 217950_at NOSIP GO:0007275 multicellular organism development
## 217950_at NOSIP GO:0016567 protein ubiquitination
## 217950_at NOSIP GO:0043086 negative regulation of catalytic activit
y
## 217950_at NOSIP GO:0050999 regulation of nitric-oxide synthase acti
vity
## 217950_at NOSIP GO:0051001 negative regulation of nitric-oxide synt
hase activity
## 217950_at NOSIP GO:0000139 Golgi membrane
## 217950_at NOSIP GO:0005634 nucleus
## 217950_at NOSIP GO:0005634 nucleus
## 217950_at NOSIP GO:0005737 cytoplasm
## 217950_at NOSIP GO:0005829 cytosol
## 217950_at NOSIP GO:0003723 RNA binding
## 217950_at NOSIP GO:0005515 protein binding
## 217950_at NOSIP GO:0061630 ubiquitin protein ligase activity
## 227552_at SEPT1 GO:0000281 mitotic cytokinesis
## 227552_at SEPT1 GO:0000921 septin ring assembly
## 227552_at SEPT1 GO:0017157 regulation of exocytosis
## 227552_at SEPT1 GO:0060271 cilium assembly
## 227552_at SEPT1 GO:0005815 microtubule organizing center
## 227552_at SEPT1 GO:0005940 septin ring
## 227552_at SEPT1 GO:0008021 synaptic vesicle
## 227552_at SEPT1 GO:0015630 microtubule cytoskeleton
## 227552_at SEPT1 GO:0030496 midbody
## 227552_at SEPT1 GO:0031105 septin complex
```

```
## 227552_at SEPT1 GO:0032160 septin filament array
## 227552_at SEPT1 GO:0003924 GTPase activity
## 227552_at SEPT1 GO:0005515 protein binding
## 227552_at SEPT1 GO:0005525 GTP binding
## 227552_at SEPT1 GO:0032947 protein-containing complex scaffold acti
vity
## 227552_at SEPT1 GO:0042802 identical protein binding
## 226482_s_at TSTD1 GO:0070221 sulfide oxidation, using sulfide:quino
ne oxidoreductase
## 226482_s_at TSTD1 GO:0005737 cytoplasm
## 226482_s_at TSTD1 GO:0005737 cytoplasm
## 226482_s_at TSTD1 GO:0005829 cytosol
## 226482_s_at TSTD1 GO:0005829 cytosol
## 226482_s_at TSTD1 GO:0036464 cytoplasmic ribonucleoprotein granule
## 226482_s_at TSTD1 GO:0048471 perinuclear region of cytoplasm
## 226482_s_at TSTD1 GO:0050337 thiosulfate-thiol sulfurtransferase ac
tivity
```

Vemos que muchos participan en la estructura celular y en regulación de la respuesta inmunológica-inflamatoria y de respuesta bacteriana.

```
#Buscamos la información contenida en Gene Ontology
library(GOstats)

library(org.Hs.eg.db)

#Especificamos la comparación que usaremos y los nombres de los genes
topTab <- topTCvsS
topTab$ID = rownames(topTab)
ezUni<- getEG(as.character(topTab$ID), "hgu133plus2.db")

# Reduzco los datos a aquellos con p-valor sin ajustar <0.01:
hwhich<-topTab["P.Value"]<0.01
hIds <- getEG(as.character(topTab$ID[hwhich]), "hgu133plus2.db")

# Creamos los "hiperparámetros" para realizar el análisis
hparams <- new("GOHyperGParams", geneIds=hIds, universeGeneIds=ezUni,
               annotation="org.Hs.eg.db", ontology="BP",
               pvalueCutoff=0.01, conditional=FALSE,
               testDirection="over")

#Buscamos la información contenida en la Kyoto Encyclopedia of Genes a
nd Genomes
hKEGG <- new("KEGGHyperGParams",
             geneIds=hIds, universeGeneIds=ezUni,
             annotation="org.Hs.eg.db",pvalueCutoff=0.01,
             testDirection="over")
```

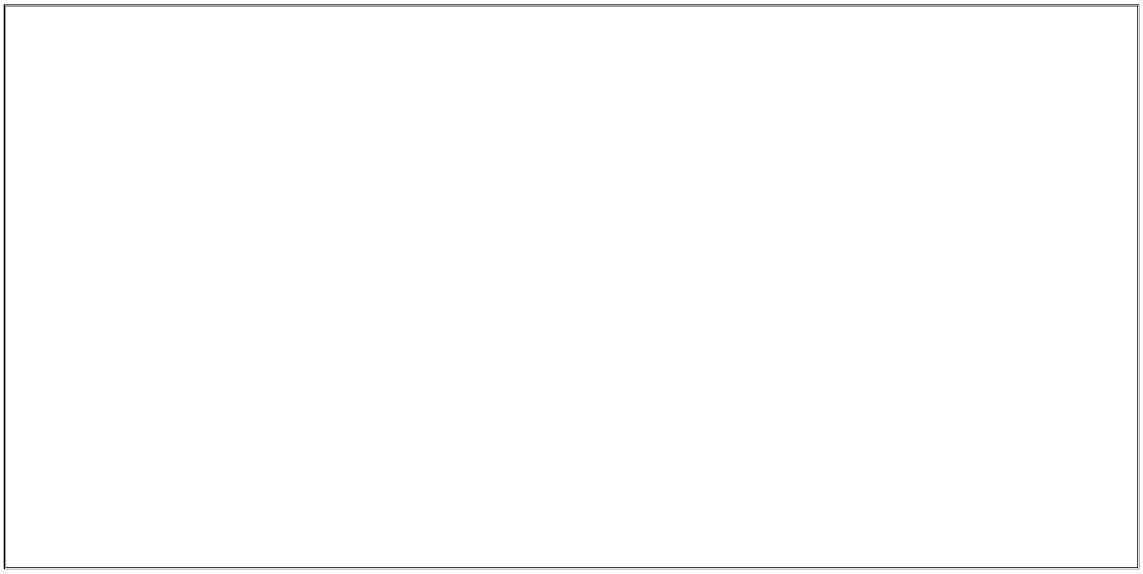
Obtendremos dos informes en formato html con los análisis de ambos repositorios para la comparación de Controles vs EPOC grave.

```
# Ejecutamos los análisis
hhyper <- hyperGTest(hparams)
KEGGhyper <- hyperGTest(hKEGG)
# Creamos el informe de resultados en formato html
comparison <- "topTabCOPD"
hfile <- file.path(resultsDir, paste("GOResults.",comparison,".html",
sep=""))
Kfile <- file.path(resultsDir, paste("KEGGResults.",comparison,".html"
, sep=""))
htmlReport(hhyper, file = hfile, summary.args=list("htmlLinks"=TRUE))
htmlReport(KEGGhyper, file=Kfile, summary.args=list("htmlLinks"=TRUE))
```

Como vemos en el informe obtenido de KEGG, unas de las vías más relacionadas con los genes más diferencialmente expresados en este estudio son la vía del linaje de células hematopoyéticas, la de las interacciones citoquina-receptor y la de la inmunodeficiencia primaria, por lo que realizaremos un análisis de estas vías con el paquete "pathview".

```
library("pathview")

## ################################################################
## Pathview is an open source software package distributed under GNU G
eneral
## Public License version 3 (GPLv3). Details of GPLv3 is available at
## http://www.gnu.org/licenses/gpl-3.0.html. Particullary, users are r
equired to
## formally cite the original Pathview paper (not just mention it) in
publications
## or products. For details, do citation("pathview") within R.
## The pathview downloads and uses KEGG data. Non-academic uses may re
quire a KEGG
## license agreement (details at http://www.kegg.jp/kegg/legal.html).
## ################################################################

KlogFC <- topTab$logFC
names(KlogFC) <- hIds
#Visualizamos la vía de hematopoyesis
pathview(gene.data = KlogFC,
         pathway.id = "hsa04640",
         species = "hsa",
         limit = list(gene=1, cpd=1))

#Vía de las interacciones citoquina-receptor de citoquinas
pathview(gene.data = KlogFC,
         pathway.id = "hsa04060",
         species = "hsa",
         limit = list(gene=1, cpd=1))

#Y la vía de las inmunodeficiencias primarias
pathview(gene.data = KlogFC,
         pathway.id = "hsa05340",
         species = "hsa",
         limit = list(gene=1, cpd=1))
```

Podemos observar en las diferentes vías, los elementos asociados a los genes más diferencialmente sobre o infra expresados en el estudio.



Parte 3: Análisis integrativo

##3.1 Ajuste de los datos

El objetivo de este análisis será crear evaluar el poder de los datos ómicos para la predicción de resultados clínicos como la distancia del test de la marcha, la obstrucción espirométrica medida por FEV1/FVC y porcentaje del FEV1 predicho, además del porcentaje de enfisema y atrapamiento aéreo en la tomografía computarizada.

Usaremos métodos de reducción de la dimensionalidad, para lo que nos quedaremos solo las variables continuas, algunas de las cuales son las que condicionan la categorización por gravedad de la enfermedad. Mantendremos la variable género y gravedad, para usarlas luego al comparar los datos.

```
#Cargamos los datos clínicos
clindat<- dbgset[,c(6:8,10,12:13,16)]

#Simplificamos los nombres de las columnas a usar
names(clindat) <- c("walkdist", "fev1_fvc", "fev1pp",
                    "gender", "emphCT", "gastrapCT", "severity")

#Volvemos a resumir los datos de este conjunto
summary(clindat)

##     walkdist       fev1_fvc          fev1pp          gender
##  Min.   : 200   Min.   :0.2300   Min.   :  9.00   female:62
##  1st Qu.:1130   1st Qu.:0.4400   1st Qu.: 43.75   male  :74
##  Median :1460   Median :0.6500   Median : 68.00
##  Mean   :1416   Mean   :0.5958   Mean   : 66.80
##  3rd Qu.:1659   3rd Qu.:0.7525   3rd Qu.: 85.25
##  Max.   :2485   Max.   :0.8500   Max.   :135.00
##  NA's   :5
##      emphCT           gastrapCT          severity
##  Min.   : 0.03225   Min.   : 0.3955   Control:42
##  1st Qu.: 0.72632   1st Qu.: 7.9749   MildMod:52
##  Median : 2.50719   Median :16.0263   Severe :42
```

```
##  Mean   : 8.32981   Mean   :25.9467
##  3rd Qu.:15.36130   3rd Qu.:47.6611
##  Max.   :45.06300   Max.   :77.5963
##  NA's   :19         NA's   :21
```

Cargamos los datos ómicos:

```
#Invertimos la matriz de expresión de datos normalizados y filtrados
#para usar en el análisis integrativo. Nos quedaremos con los 20 genes
#más diferencialmente expresados en la comparación de controles y EPOC
grave:
exs0<-exprs(gset_filt[rownames(topTCvsS[1:20,])])
genedat <- t(exs0)

#Usamos los nombres de los genes
colnames(genedat)<-getSYMBOL(row.names(topTCvsS[1:20,]), "hgu133plus2.
db")

#Mantenemos los nombres de las filas iguales entre los datos ómicos y
clínicos (eliminamos los últimos caracteres de los nombres en genedat)
library(stringr)

row.names(genedat)<- str_sub(row.names(genedat), start=1, end = 10)
```

##3.2 Análisis de componentes principales (PCA) inicial

Realizaremos un PCA inicial para ambos conjuntos de datos, clínicos y ómicos:

```
#Cargamos el paquete mixOmics que usaremos para el análisis integrativ
o
library("mixOmics")

#Nos aseguramos que no hay datos ómicos con varianza cercana a cero
var0 <- nearZeroVar(genedat)
print(var0)
```

```
## $Position
## integer(0)
##
## $Metrics
## [1] freqRatio     percentUnique
## <0 rows> (or 0-length row.names)
```

```
#Solo usamos las variables continuas para el análisis
#Fijamos ncomp en 5 ya que es el número de variables continuas
clinpca <- pca(clindat[,-c(4,7)], ncomp = 5, center = TRUE, scale = TR
UE)
clinpca$explained_variance
```

```
##        PC1        PC2        PC3        PC4        PC5
## 0.78795821 0.13371158 0.05348616 0.02161040 0.01139910
```

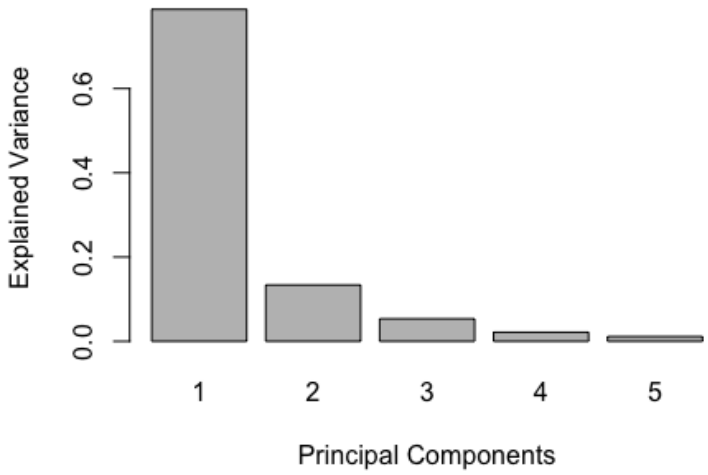
Vemos que la varianza es explicada por los 2 primeros componentes principales (aprox 91%).

```
#Fijamos ncomp en 10 para valorar los primeros 10 PC
genepca <- pca(genedat, ncomp = 10, center = TRUE, scale = TRUE)
genepca$explained_variance
```

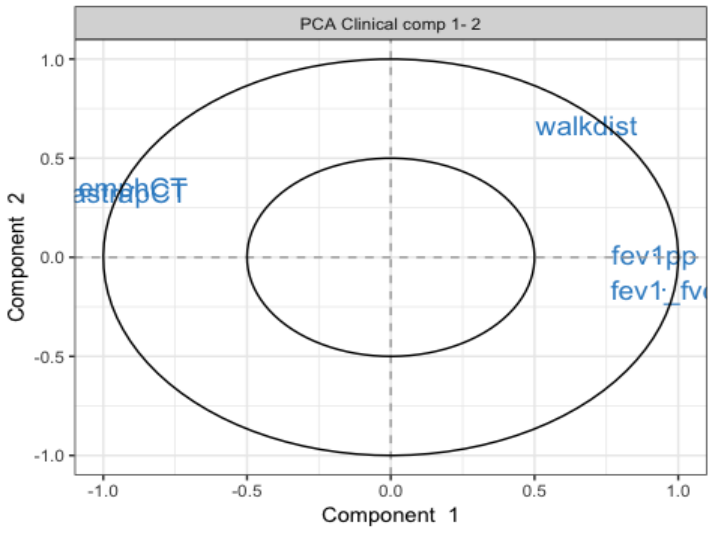
```
##         PC1         PC2         PC3         PC4         PC5         PC6
## 0.56967297 0.07587281 0.05994992 0.05760665 0.03574616 0.03235956
##         PC7         PC8         PC9        PC10
## 0.02780275 0.02196560 0.01885589 0.01579840
```

En este caso, la varianza se explica fundamentalmente por los primeros 2 PC (aprox 63%). Graficamos:

```
plot(clinpca)
```



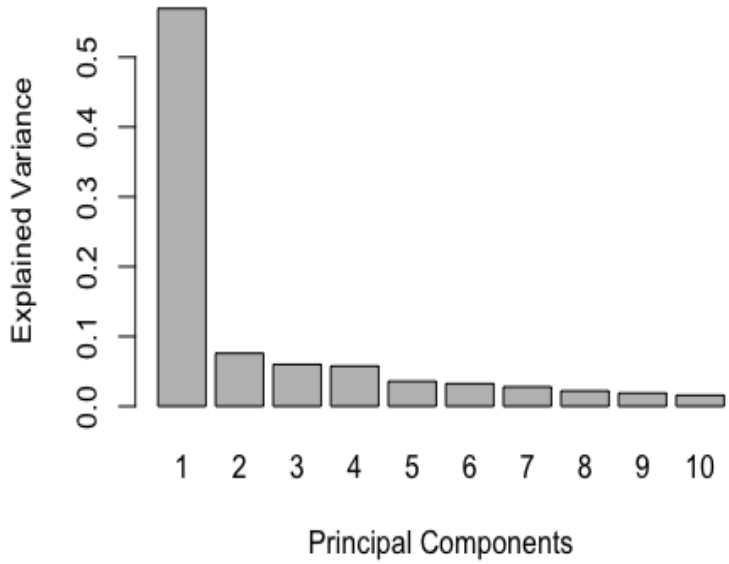

```
plotVar(clinpca, comp = c(1, 2), var.names = T, title = "PCA Clinical
comp 1- 2")
```



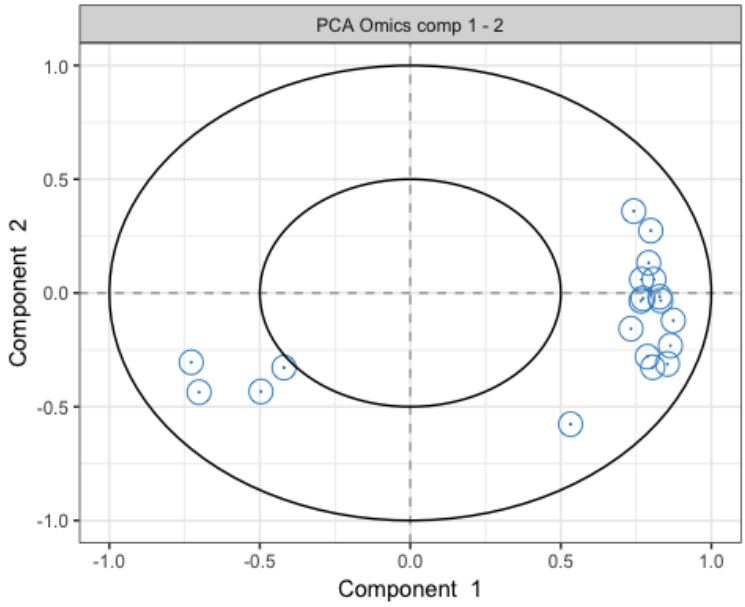
Vemos que los datos clínicos se alejan bastante del 0, y que el primer PC parece explicar gran parte de la varianza.

Seguimos con los datos ómicos:

```
plot(genepca)
```



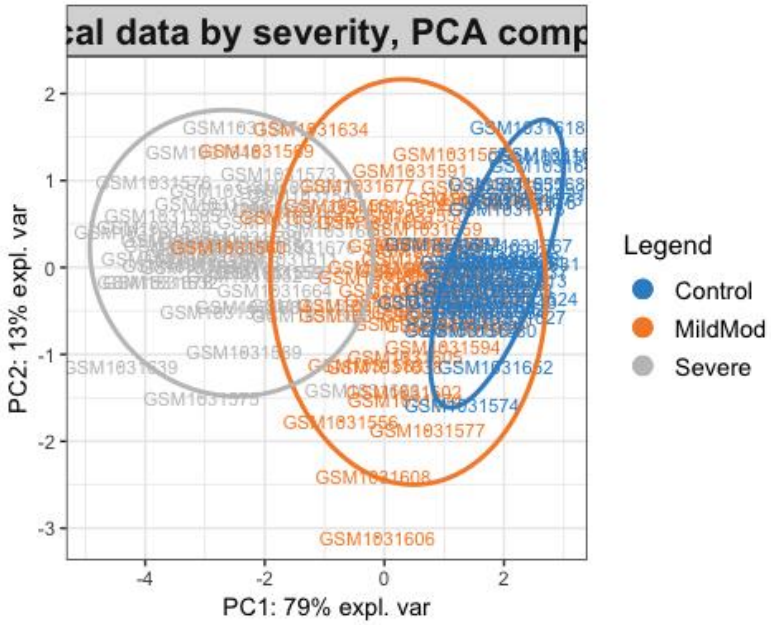

```
plotVar(genepca, comp = c(1:2), var.names = FALSE, title = "PCA Omics
comp 1 - 2")
```



En este caso la mayor parte de la varianza de los datos se explica por aproximadamente los primeros 2 PC. Compararemos ahora por grupo de severidad o control sano:

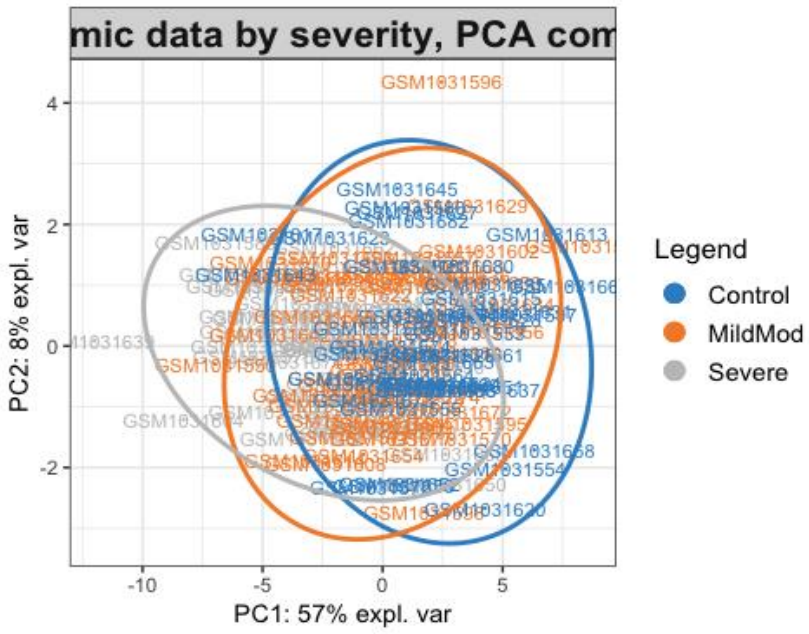
Clínicos:

```
plotIndiv(clinpca, comp = c(1, 2), group = clindat$severity, ellipse =
TRUE,
          legend = TRUE, title = "Clinical data by severity, PCA comp
1 - 2")
```



Ómicos:

```
plotIndiv(genepca, comp = c(1, 2), group = clindat$severity, ellipse =
TRUE,
          legend = TRUE, title = "Genomic data by severity, PCA comp 1
-2")
```



En el conjunto de datos clínicos parecen existir grupos claramente separados, principalmente entre los controles y los EPOC graves. Menos evidente parece en el caso de los datos ómicos.

##3.3 Análisis de mínimos cuadrados parciales (PLS) y mínimos cuadrados parciales dispersos (sPLS)

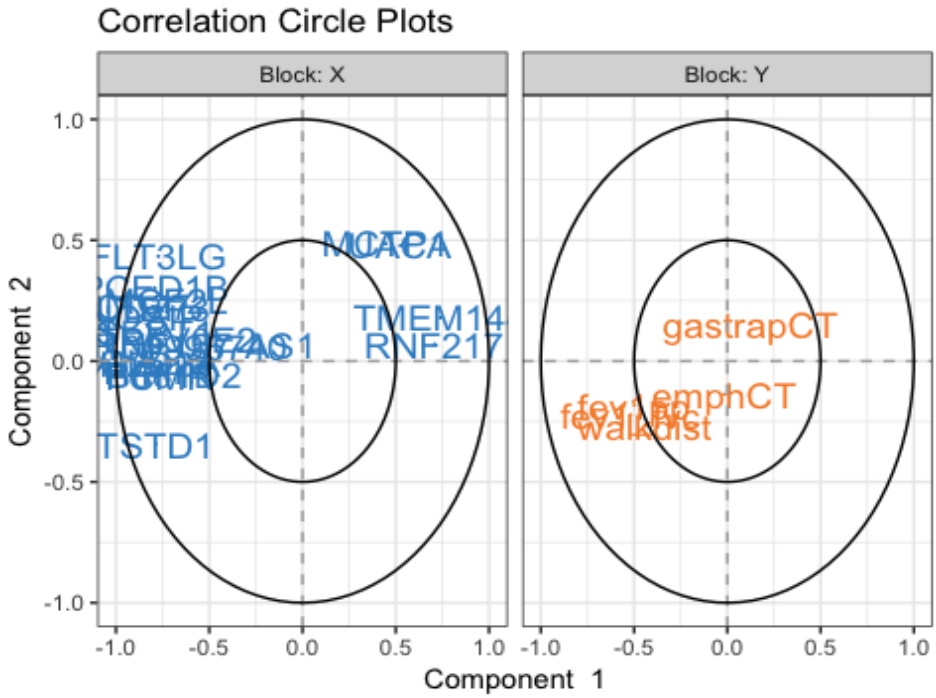
En vista de que teníamos algunos valores faltantes (NAs) en algunas de las variables (walkdist, porcentaje de enfisema y de atrapamiento aéreo), debemos especificar el manejo de los mismos. Una opción es imputarlos mediante el análisis iterativo de mínimos cuadrados parciales no lineales con la función nipals().

```
#Especificamos la función y el número de componentes, en este caso sig
ue siendo 5
clin_nipals <- nipals(clindat[,-c(4,7)], reconst = T, ncomp = 5)$rec

#Objetivamos los valores de imputación
id.na <- is.na(clindat[,-c(4,7)])
clin_nipals[id.na]
```

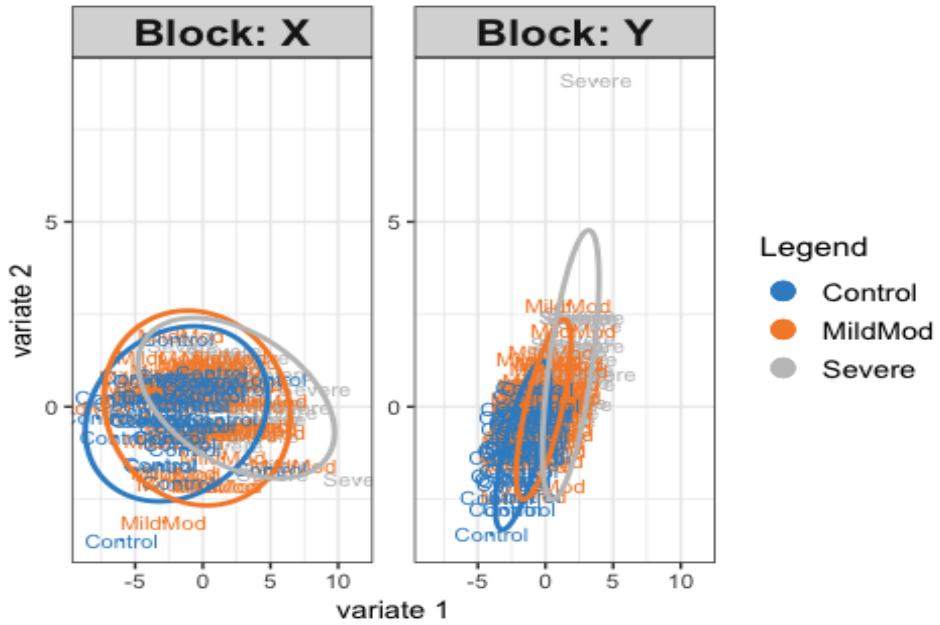
```
##  [1]   771.067277   723.959328   912.991676  1691.182288   749.5355
81
##  [6]   168.518052   138.466436   105.375205    69.897436  -107.9396
97
## [11]  -143.716937    68.883666    86.510468    87.093685   -79.7450
69
## [16]    -1.228132    12.122484     4.731378   108.443005  -222.0053
25
## [21]    13.262429    22.959491   110.518775 -2723.286701    33.4722
55
## [26]    34.521067    29.700159    18.873131    13.902411    -1.2974
53
## [31]    19.915036    24.514260     3.814233    20.219720    18.1868
23
## [36]    25.526588    21.115950     3.492110     4.964621    23.9986
12
## [41]     4.103741    15.102701    24.975815    28.101990  1161.7918
11
```

Ahora aplicamos el PLS y sPLS a estos datos ajustados. Para el PLS nos quedaremos con 5 PC, y para el sPLS los fijaremos en 10 para los ómicos y 5 para los clínicos:

```
clinpls <- pls(genedat, clin_nipals, ncomp = 5, mode = "regression")
clinspls <- spls(genedat, clin_nipals, keepX = c(10,10,10),
                 keepY = c(5,5,5), mode = "regression")

#Graficamos los datos por separado
plotVar(clinpls, overlap = F)
```



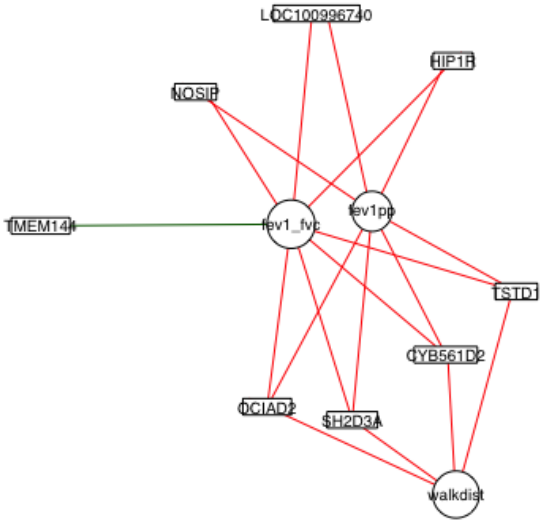

```
#Y agrupados por gravedad-control.
plotIndiv(clinpls, group = clindat$severity, ind.names = clindat$sever
ity,
          legend = T, ellipse = T)
```



Observamos dos grupos grandes de genes que se comportan de forma opuesta, mientras que las variables clínicas parecen estar más correlacionadas entre sí, excepto el atrapamiento aéreo.

Creamos finalmente un gráfico con la red de relevancia con los dos primeros PC en función de la expresión génica y la información clínica, analizados con sPLS. Las líneas verdes representan las correlaciones positivas, y las rojas, negativas.

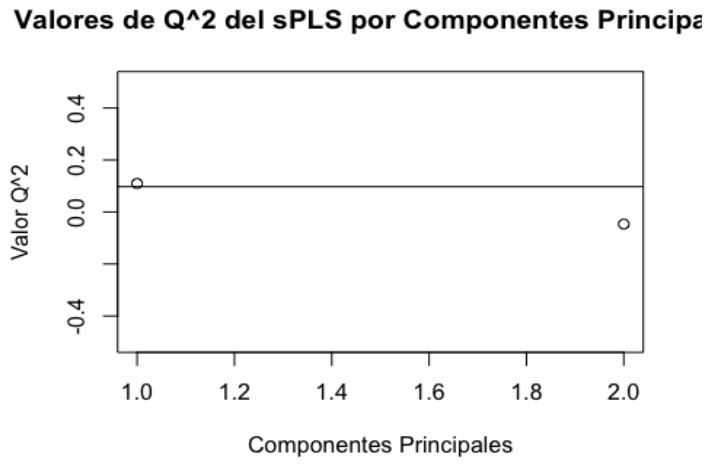
```
relev.net <- network(clinspls, comp = 1:2,
                     color.edge = c("darkgreen", "red"),
                     shape.node = c("rectangle", "circle"),
                     show.color.key = FALSE,
                     cex.node.name = 0.6)
```



##3.4 Evaluación inicial

Emplearemos la función perf para calcular los criterios de evaluación para el modelo sPLS. Realizaremos una validación cruzada (Mfold) y estableceremos la evaluación de 10 PC repitiendo el análisis 50 veces (nrepeat = 50).

```
set.seed(123)
splseval<- perf(clinspls, validation= "Mfold", folds= 10, nrepeat= 50,
                progressBar = FALSE)
plot(splseval$Q2.total, ylim = c(-0.5, 0.5),
    main = "Valores de Q^2 del sPLS por Componentes Principales",
    xlab = "Componentes Principales",
    ylab = "Valor Q^2")
abline(h = 0.0975)
```



Imprimimos los valores de Q^2

```
splseval$Q2.total
```

```
##           Q2.total
## 1 comp  0.10964554
## 2 comp -0.04638512
```

… y los de R^2

```
print(splseval$R2)
```

```
##                 1 comp       2 comp
## walkdist  1.509647e-01 1.188729e-01
## fev1_fvc  1.793817e-01 1.695390e-01
## fev1pp    1.497119e-01 1.508646e-01
## emphCT    2.907220e-02 4.578146e-05
## gastrapCT 8.648024e-05 1.196637e-04
```

Vemos como solo el Q^2 del primer componente está por encima de 0.0975, por lo que es el que tiene mayor poder de predicción.

##3.5 Evaluación de las predicciones

Primero determinamos las predicciones con ambos conjuntos de datos y por cada componente por separado:

```
predspls <- as.data.frame(predict(clinspls, genedat)$predict)
predspls_pc1 <- predspls[, grepl("dim 1", colnames(predspls))]
predspls_pc2 <- predspls[, grepl("dim 2", colnames(predspls))]
```

Luego evaluamos las predicciones iniciales con la raíz del error cuadrático medio (RMSE).

```
#Primero creamos la función para calcular el RMSE.
rmse <- function(measured, predicted){
  return(sqrt(mean((measured-predicted)^2)))
}

#Y luego una función para evaluar las predicciones
metcalc <- function(measured, predicted){
    #Crea un data frame para guardar los valores
    resdf <- data.frame()
    #Crea un vector para guardar los valores
    resob <- vector()
    for (i in 1:ncol(measured)){
        resob[length(resob)+1] <- rmse(measured[,i], predicted[,i])
    }
    #Se unen ambos elementos
    resdf <- rbind(resdf, resob)
    colnames(resdf) <- colnames(measured)
    return(resdf)
}
```

Unimos las dos predicciones con los dos primeros componentes:

```
pred1 <- rbind(metcalc(clin_nipals, predspls_pc1),
               metcalc(clin_nipals, predspls_pc2))
row.names(pred1) <- c("sPLS Comp 1", "sPLS Comp 2")
print(pred1)

##             walkdist  fev1_fvc   fev1pp  emphCT gastrapCT
## sPLS Comp 1 383.4563 0.1717404 24.63936 236.448  98.54509
## sPLS Comp 2 381.7444 0.1717072 24.60861 226.546  95.71014
```

Vemos que el error no es muy grande. Comparamos las predicciones del primer componente con los datos reales:

```
#Seleccionamos solo las predicciones de la primera dimensión, luego
#limpiamos el conjunto de datos y asignamos los nombres de las columna
s
colnames(predspls_pc1) <- colnames(clin_nipals)
pred_frame <- data.frame()
pred_frame <- rbind(pred_frame, cbind(row.names(predspls_pc1), predspl
s_pc1))
```

Buscamos los datos con las predicciones de la dimensión 1:

```
#Creamos el objeto para las predicciones
respred <- list()
respred[["pred_frame"]] <- pred_frame
respred <- respred$pred_frame

#Renombramos la columna de los nombres de sujetos
colnames(respred)[1] <- "ID"
respred$ID <- as.character(respred$ID)

#Agregamos las medias
clin_respred <- aggregate(respred[,-1], by=list(respred$ID), FUN=mean)

#Fijamos los nombres de las filas y eliminamos columnas no usadas
row.names(clin_respred) <- clin_respred[,1]
clin_respred <- clin_respred[,2:ncol(clin_respred)]

#Observamos resultados
cat("Predicted:\n")

## Predicted:

print(round(head(clin_respred),2))

##            walkdist fev1_fvc fev1pp emphCT gastrapCT
## GSM1031549  1390.01     0.58  66.15 -10.16     33.84
## GSM1031550  1073.91     0.42  44.88 -27.91     54.87
## GSM1031551  1407.96     0.59  67.35  -9.15     32.64
## GSM1031552   990.41     0.37  39.27 -32.61     60.43
```

```
## GSM1031553  1596.75     0.68  80.05   1.46     20.07
## GSM1031554  1662.56     0.72  84.48   5.15     15.70
```

Y ahora los datos reales (con la penalización del método nipals)

```
cat("Measured:\n")
```

```
## Measured:
```

```
print(round(head(clin_nipals),2))
```

```
##            walkdist fev1_fvc fev1pp  emphCT gastrapCT
## GSM1031549  1600.00     0.46  36.96  168.52     33.47
## GSM1031550  1805.01     0.58  53.97  138.47     34.52
## GSM1031551  1620.00     0.54  52.98  105.38     29.70
## GSM1031552  1010.01     0.33  31.99   69.90     18.87
## GSM1031553  1641.00     0.78  93.99    1.58     19.11
## GSM1031554  1460.00     0.73  96.03 -107.94     13.90
```

```
summary(clin_nipals)
```

```
##     walkdist       fev1_fvc          fev1pp           emphCT      
##  Min.   : 200   Min.   :0.08361   Min.   :  9.001   Min.   :-2723.2867  
##  1st Qu.:1093   1st Qu.:0.42669   1st Qu.: 43.742   1st Qu.:   0.7005  
##  Median: 1445   Median:0.59343   Median: 67.999   Median :  2.8424  
##  Mean   :1400   Mean   :0.58267   Mean   : 66.806   Mean   :  -9.6069  
##  3rd Qu.:1658   3rd Qu.:0.72247   3rd Qu.: 85.248   3rd Qu.:  17.1583  
##  Max.   :2485   Max.   :1.09362   Max.   :134.996   Max.   :  168.5181  
##    gastrapCT      
##  Min.   :  -1.298  
##  1st Qu.:   7.977  
##  Median :  18.294  
##  Mean   :  33.183  
##  3rd Qu.:  41.604  
##  Max.   :1161.792
```

Podemos observar como algunas predicciones son bastante buenas, pero algunas (especialmente en aquellas con más valores faltantes, como el porcentaje de enfisema), no lo es tanto. Por lo tanto, concluimos que a través de este conjunto de datos ómicos (expresión génica) es posible predecir con cierta precisión, algunas variables clínicas como la distancia caminada en el test de la marcha y los valores espirométricos.

Anexo 2: Algunas de las vías biológicas en las que participan los genes más diferencialmente expresados, según la biblioteca Gene Ontology (GO).

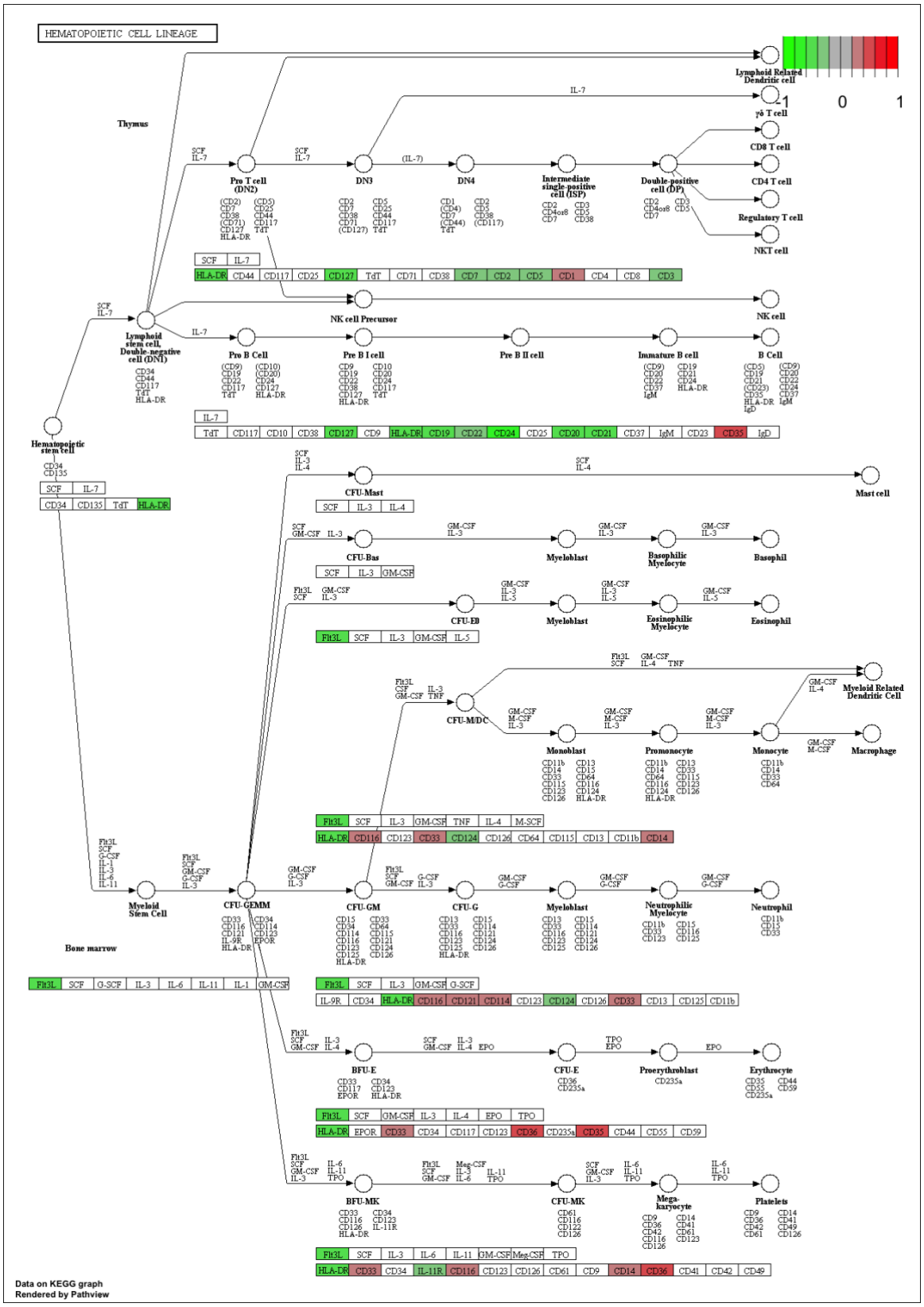
GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0045321	0.000	2.253	92	168	854	leukocyte activation
GO:0001775	0.000	2.142	102	178	944	cell activation
GO:0006955	0.000	1.863	136	212	1265	immune response
GO:0002252	0.000	2.020	84	143	778	immune effector process
GO:0002443	0.000	2.205	59	109	546	leukocyte mediated immunity
GO:0002274	0.000	2.284	49	94	455	myeloid leukocyte activation
GO:0002376	0.000	1.636	195	270	1810	immune system process
GO:0002263	0.000	2.167	54	99	500	cell activation involved in immune response
GO:0002366	0.000	2.154	54	98	497	leukocyte activation involved in immune response
GO:0002682	0.000	1.776	101	155	934	regulation of immune system process
GO:0002275	0.000	2.227	42	80	392	myeloid cell activation involved in immune response
GO:0043299	0.000	2.234	42	79	386	leukocyte degranulation
GO:0002696	0.000	2.695	24	54	226	positive regulation of leukocyte activation
GO:0050867	0.000	2.654	25	55	233	positive regulation of cell activation
GO:0043312	0.000	2.272	38	73	351	neutrophil degranulation
GO:0032940	0.000	1.766	96	148	893	secretion by cell
GO:0036230	0.000	2.231	39	75	366	granulocyte activation
GO:0002283	0.000	2.247	38	73	354	neutrophil activation involved in immune response
GO:0042119	0.000	2.231	39	74	361	neutrophil activation

GO:0002444	0.000	2.167	43	79	395	myeloid leukocyte mediated immunity
GO:0002694	0.000	2.247	38	72	349	regulation of leukocyte activation
GO:0045055	0.000	1.992	56	97	522	regulated exocytosis
GO:0046649	0.000	2.040	51	90	474	lymphocyte activation
GO:0046903	0.000	1.717	103	155	958	secretion
GO:0002446	0.000	2.198	39	73	360	neutrophil mediated immunity
GO:0042110	0.000	2.229	36	69	336	T cell activation
GO:0050776	0.000	1.862	66	108	615	regulation of immune response
GO:0002684	0.000	1.821	71	113	656	positive regulation of immune system process
GO:0050865	0.000	2.114	40	73	371	regulation of cell activation

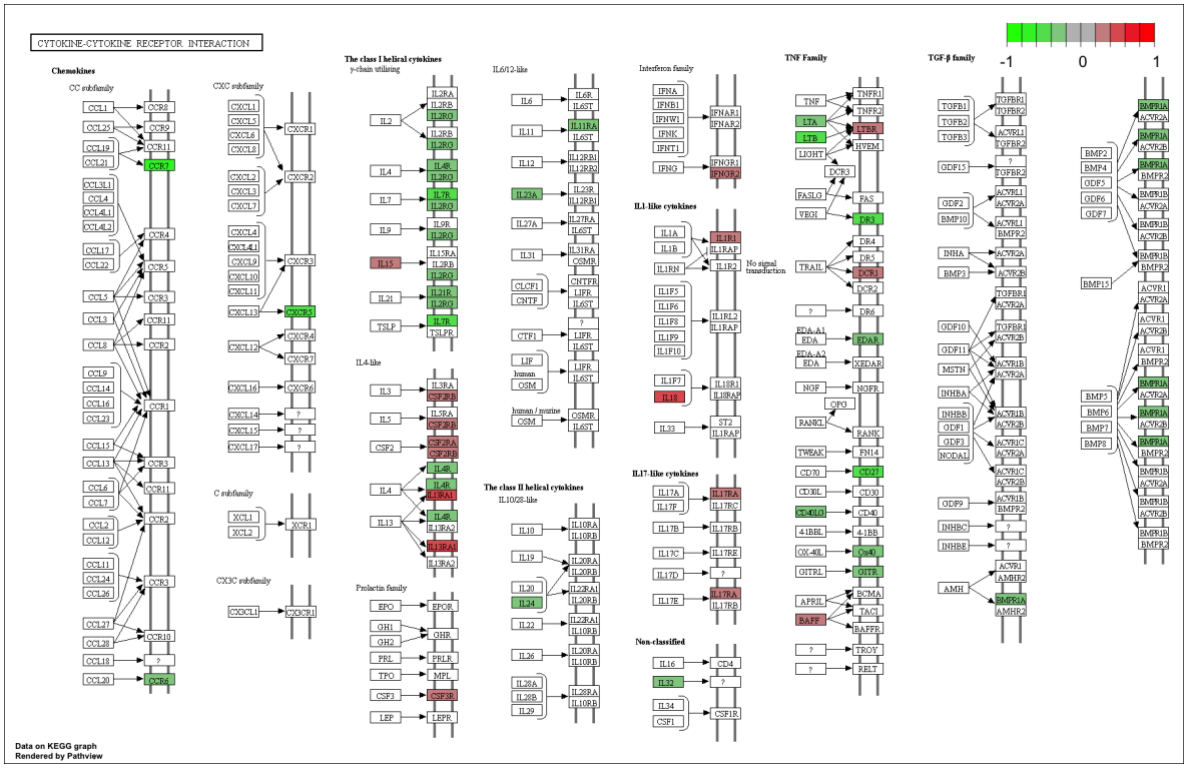
Anexo 3: Algunas de las vías biológicas en las que participan los genes más diferencialmente expresados, según la biblioteca Kyoto Encyclopedia of Genes and Genomes (KEGG).

KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
04640	0.000	3.350	8	21	69	Hematopoietic cell lineage
04060	0.000	2.390	18	35	149	Cytokine-cytokine receptor interaction
05340	0.000	5.523	3	11	26	Primary immunodeficiency
05144	0.002	3.568	4	10	31	Malaria
00010	0.007	3.511	3	8	25	Glycolysis / Gluconeogenesis
00590	0.007	3.511	3	8	25	Arachidonic acid metabolism
04330	0.008	3.055	4	9	31	Notch signaling pathway

Anexo 4: Participación de los genes más diferencialmente expresados en la vía del linaje de células hematopoyéticas. En verde, los genes sobreexpresados, y en rojo los infraexpresados.



Anexo 5: Participación de los genes más diferencialmente expresados en la vía de las interacciones citoquina-receptor. En verde, los genes sobreexpresados, y en rojo los infraexpresados.



Anexo 6: Participación de los genes más diferencialmente expresados en la vía de las inmunodeficiencias primarias. En verde, los genes sobreexpresados, y en rojo los infraexpresados.