TRABAJO DE FIN DE GRADO

Grado en Estadística

Título: Desarrollo de un *pipeline* para el análisis de datos metabólicos

Autor: Aina Fernández Bargalló

Director: Esteban Vegas Lozano y Antonio Miñarro Alonso

Departamento: Departamento de Genética, Microbiología y Estadística

Convocatoria: Primera convocatoria (junio)

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

RESUMEN

El trabajo que se presenta consiste en el desarrollo de un procedimiento para analizar datos metabólicos utilizando el *software* estadístico R. Concretamente, consta de la definición de los pasos necesarios en la fase de preprocesado de los datos y de las técnicas estadísticas más habituales para hacer análisis univariante y multivariante, y concluyendo con la integración de los datos metabólicos con otro tipo de datos. Dicho procedimiento se ha verificado con unos datos reales obtenidos de un repositorio de la red, de forma que se ha estudiado como se pueden describir distintas variedades de vino blanco des del punto de vista metabolómico.

Palabras clave: *biología, metabolómica, preprocesamiento, análisis estadístico, integración, vinos blancos, metabolitos.*

Clasificación AMS: 62P10

RESUM

El treball que es presenta consisteix en el desenvolupament d'un procediment per analitzar dades metabolòmiques utilitzant el *software* estadístic R. Concretament, consta de la definició dels passos necessaris en la fase de preprocessament de les dades i de les tècniques estadístiques més habituals per fer anàlisi univariant i multivariant, i concloent amb la integració de les dades metabolòmiques amb altres tipus de dades. Aquest procediment s'ha verificat amb unes dades reals obtingudes d'un repositori de la xarxa, de manera que s'ha estudiat com es poden descriure diferents varietats de vins blancs des del punt de vista metabolòmic.

Paraules clau: *biologia, metabolòmica, preprocessament, anàlisis estadístic, integració, vins blancs, metabòlits.*

Classificació AMS: 62P10

ABSTRACT

This project consists of the development of a pipeline to analyse metabolomics data using the R software. Specifically, consists on the definition of the necessary steps in the preprocessing phase and the more popular statistic techniques to carry out a univariate and multivariate analysis, and conclude with the integration of the metabolomics data with other types of data. This pipeline has been verified with real data obtained from a network repository, so that we have studied how different varieties of white wines can be described from the metabolomic point of view.

Key words: *biology, metabolomics, preprocessing, statistical analysis, Integration, white wines, metabolites.*

AMS Classification: 62P10

ÍNDICE

INTRODUCCIÓN 1
METODOLOGÍA 4
I. ¿QUÉ SON LOS DATOS METABOLÓMICOS? 6
1. Introducción a la biología 6
1.1. Características de los seres vivos 6
1.2. Niveles de organización de los seres vivos 8
1.3. Diversidad de los organismos 10
2. El metabolismo 12
2.1. ¿Qué es el metabolismo? 12
2.2. Funcionamiento del metabolismo 13
2.3. Biomoléculas principales 15
2.4. Los metabolitos 16
3. Tecnologías ómicas 17
3.1. La genómica 18
3.2. La transcriptómica 21
3.3. La proteómica 23
3.4. La metabolómica 26
3.5. El fenotipo 27
4. Los datos metabolómicos 27
4.1. Plataformas analíticas para datos metabolómicos 28
II. HERRAMIENTAS PARA EL ANÁLISIS DE DATOS METABOLÓMICOS 34
1. Herramientas de análisis más populares 35
2. Funcionalidades de algunas herramientas 37
2.1. Preprocesamiento 37
2.2. Anotación 39
2.3. Postprocesamiento 40
2.4. Análisis estadístico 40
2.5. Flujos de trabajo 40
2.6. Otras herramientas 41
3. Paquetes de R disponibles para el análisis de datos metabolómicos 41

III. DESARROLLO DE UN PIPELINE PARA EL ANÁLISIS DE DATOS METABOLÓMICOS 44
1. Preparación de la muestra y adquisición de la base de datos 45
2. Análisis descriptivo inicial 47
3. Técnicas de preprocessing 48
4. Análisis de los datos 50
5. Interpretación e integración de los datos 57
IV. VERIFICACIÓN DEL *PIPELINE* CON DATOS REALES 59
1. Análisis descriptivo inicial 61
1.1. Principales estadísticos 61
1.2. Estudio de valores faltantes 76
1.3. Boxplots iniciales 79
2. Técnicas de *preprocessing* 83
2.1. Imputación de valores faltantes 83
2.2. Transformaciones 83
2.3. Normalización y escalado 83
3. Análisis de los datos 85
3.1. Prueba ANOVA de una vía para datos independientes 85
3.2. Boxplots 87
3.3. Análisis de Componentes Principales 89
3.4. Heatmap 96
3.5. PLS-DA 98
3.6. Correlaciones 103
4. Integración e identificación 104
4.1. Integración: análisis de factores múltiples 105
4.2. Identificación de metabolitos 114
CONCLUSIONES 116
BIBLIOGRAFIA 117
ANEXO 122
Anexo A. Código en R 122
A.1. Lectura de la base de datos: 122
A.2. Análisis descriptivo inicial 122

A.2. Preprocessing de los datos 125
A.3. Análisis de los datos 125
A.4. Integración de los datos 130
Anexos B. Boxplots iniciales 131
Anexo C. Boxplots finales 131
Anexo D. Matriz de correlaciones entre metabolitos 131
Anexo E. Matriz de correlaciones entre muestras de vinos 132
Anexo F. Análisis de Factores Múltiples 132
Anexo G. Aplicación Shiny 132

ÍNDICE DE FIGURAS

Figura 1. Niveles de organización de los seres vivos 9

Figura 2. Fases del metabolismo 15

Figura 3. Cascada ómica. 18

Figura 4. Ilustración de los cromosomas junto con el ADN y los genes. 19

Figura 5. Proceso de formación de las proteínas 24

Figura 6. Estructura de las proteínas. 25

Figura 7. Esquema de los componentes de un sistema NMR 30

Figura 8. Esquema de un sistema GC-MS. 32

Figura 9. Esquema de un sistema LC-MS. 33

Figura 10. Flujo de un pipeline para el análisis de datos metabolómicos 45

Figura 11. Partes que componen un Boxplot 52

Figura 12. Porcentaje de missings por variable 77

Figura 13. Boxplots de los metabolitos antes del preprocesado 79

Figura 14. Boxplots de algunos de los metabolitos antes del preprocesado 80

Figura 15. Boxplot escalado del metabolito “Talose” 82

Figura 16. Comparación de los boxplots antes y después del preprocesamiento 84

Figura 17. Boxplots de los metbolitos más significativos después del preprocesado 88

Figura 18. Scores Plot del PCA 92

Figura 19. Loadings Plot del PCA 93

Figura 20. Loadings Plot del PCA con las famílias de los metabolitos 96

Figura 21. Gráfico de calor de los metabolitos en función de las muestras de vinos 97

Figura 22. PLS-DA de los metabolitos según las diferentes variedades de vino 101

Figura 23. Gráfico de los ejes parciales de la primera y segunda dimensiones del MFA 107

Figura 24. Gráfico de las variables en la primera y segunda dimensiones del MFA 108

Figura 25. Gráfico de los individuos en la primera y segunda dimensiones del MFA 109

Figura 26. Puntos parciales para las variedades en la primera y segunda dimensiones del MFA 110

Figura 27. Gráfico de los ejes parciales de la primera y tercera dimensiones del MFA 111

Figura 28. Gráfico de las variables en la primera y tercera dimensiones del MFA 112

Figura 29. Gráfico de los individuos en la primera y tercera dimensiones del MFA 113

Figura 30. Puntos parciales para las variedades en la primera y segunda dimensiones del MFA 114

ÍNDICE DE TABLAS

Tabla 1. Características de los cinco reinos 11

Tabla 2. Clasificación de los vinos disponibles en la base de datos 60

Tabla 3. Estadísticos para cada uno de los metabolitos 63

Tabla 4. Estadísticos para cada uno de los metabolitos, separando por variedades..... 76

Tabla 5. Número de missings por variedad..... 78

Tabla 6. P-value y p-value ajustado obtenidos con la prueba ANOVA..... 87

Tabla 7. Loadings de la primera y segunda Componentes del PCA de los metabolitos..... 91

Tabla 8. Clasificación de los metabolitos según su família química 95

Tabla 9. Primera y segunda funciones discriminantes obtenidas con el PLS-DA 100

Tabla 10. Matriz cruzada de las variedades observadas y las predichas con PLS-DA 102

Tabla 11. Temperatura anual media, precipitaciones y altitud de las distintes regiones 105

INTRODUCCIÓN

Los datos metabolómicos son un campo desconocido para muchas personas; de hecho, a menudo puede ser complicado explicar o entender qué es un dato metabolómico o qué relevancia puede tener estudiarlos. Aunque el campo más popular al que se dedica la metabolómica es el estudio de las plantas, lo cierto es que se trata de una disciplina que puede aportar no tan solo conocimiento, sino que también acercan a quienes los estudian a una mayor comprensión de aspectos de la vida que antes no se habían cuestionado, y que está en constante crecimiento de forma que cada vez abarca más tipos de datos.

A pesar de ser un campo en crecimiento, el estudio de este tipo de datos aún no se ha desarrollado completamente, cosa que dificulta su análisis; de hecho, éste no se puede realizar con una sola plataforma analítica, sino que requiere que se combinen varias de ellas para un estudio completo.

Con este trabajo se ha querido implementar un procedimiento (*pipeline*) para analizar datos metabolómicos a partir del *software* estadístico R, con la intención de demostrar que, aunque son un conjunto de datos distintos a los que los estadísticos pueden estar acostumbrados, mediante la aplicación de muchas de las técnicas y métodos estadísticos más comunes se puede realizar un estudio de datos metabolómicos completo y obtener conclusiones acerca de su comportamiento. Además, se ofrece un breve resumen de otras plataformas que se pueden utilizar para el análisis de dichos datos, para que el lector sea consciente de que, a pesar de que no es de los campos más conocidos y desarrollados, existen muchas posibilidades para poder realizar el análisis.

Así pues, el trabajo defiende diversas técnicas que se pueden utilizar para analizar los datos metabolómicos, profundizando en cada una de ellas con la intención de que el lector pueda aplicar el *pipeline* en cualquier conjunto de datos. Se puede considerar un trabajo didáctico, ya que la intención no es analizar una base de datos y extraer conclusiones sobre ella, sino que se busca la comprensión del desarrollo que se lleva a cabo y facilitar así el análisis de futuras bases de datos.

Persiguiendo este objetivo se ha plasmado el *pipeline* en una base de datos sobre los metabolomas en los vinos blancos, de forma que además de la parte más teórica se proporciona un ejemplo de cómo llevar a cabo dicho procedimiento estadístico y qué relevancia tiene cada una de las técnicas que se realizan.

Este trabajo se ha realizado completamente mediante el *software* R, ya que ofrece un largo listado de funciones con las que se pueden desarrollar las técnicas necesarias; por lo tanto, no se ha utilizado ninguna de las plataformas especializadas en datos metabolómicos; aun así, todas las técnicas y métodos son accesibles para cualquier usuario, y son fáciles de realizar e interpretar, por lo que no se necesita ser un experto en R para conseguir realizar el análisis de datos metabolómicos que se propone en este trabajo.

No obstante, merece la pena remarcar que, para el ejemplo de análisis que se ha realizado, se ha obtenido una base de datos de una plataforma de Internet, de modo que los datos ya estaban tomados y preparados para el análisis; así pues, el trabajo tiene la limitación de que, aunque se explica cómo se haría la preparación de la muestra y la adquisición de la base de datos, no se proporciona el ejemplo que muestre cómo se haría.

Este trabajo se divide en cuatro capítulos. El primero es una introducción a la biología y a los datos ómicos, de los que forma parte la metabolómica, y tiene el objetivo de introducir al lector en el mundo de los datos metabolómicos. Se ha creído conveniente empezar des de las bases de la biología y profundizar en distintos aspectos que, aunque están relacionados con la metabolómica, no forman parte explícitamente de ella, con la intención de asegurar la comprensión de cualquier concepto que pueda surgir.

El segundo capítulo se basa en una breve recopilación de distintas herramientas que se pueden usar para el análisis de datos metabolómicos, a pesar de que en el presente trabajo no se va a hacer; esto se hace con la intención de mostrar el amplio abanico de posibilidades para analizar este tipo de datos, de forma que el lector pueda decidir cuál de ellas le es más conveniente para el tipo de análisis que quiere realizar.

En el tercer capítulo, se hace un resumen de los pasos del *pipeline* que se propone en este trabajo, explicando cada una de las técnicas que se puede usar detalladamente.

Por último, en el cuarto capítulo se hace una implementación del *pipeline* detallado en el anterior capítulo sobre una base de datos metabolómicos. Didácticamente, se busca proporcionar ejemplos de algunas de las técnicas explicadas y dar una visión directa y completa del *pipeline* propuesto. Analíticamente, el objetivo es observar en qué se diferencian distintas variedades de uva de vino blanco – Chardonnay, Sauvignon Blanc, Pinot Gris, Riesling y Viognier - des del punto de vista metabolómico, y poder identificarlas en función de los metabolomas que componen el vino, a partir de la

hipótesis de que los metabolitos que componen los vinos blancos se presentan de forma distinta en función de la variedad.

METODOLOGÍA

El presente trabajo se puede separar en dos partes bien diferenciadas: la teórica, que comprendería los tres primeros capítulos, y la práctica, que se corresponde con el cuarto capítulo.

Para las partes teóricas, se ha buscado información en distintos libros y páginas web, aunque en el tercer capítulo también se ha hecho a partir de la investigación personal de las distintas funciones de R y métodos para adecuarlas a los datos metabolómicos. En cuanto a la parte práctica, se ha realizado completamente con R. En concreto, se ha hecho un preprocesado de la base de datos inicial, un análisis univariante y multivariante y una integración de los datos.

El trabajo se basa en distintas técnicas estadísticas para analizar datos metabolómicos. En concreto, se ha hecho uso de los boxplots (o diagramas de cajas), tanto múltiples como individuales – por las distintas variables que tiene la base de datos sobre la que se ha trabajado. Se han usado funciones ya implementadas en R, algunas para el cálculo de estadísticos para hacer un análisis descriptivo (como la media), y otras de preprocesado como la imputación basada en los K vecinos más cercanos (*knnImputation)* o el logaritmo de los datos.

También se han realizado pruebas ANOVA de una vía para datos independientes, con el objetivo de comprobar si las variables – metabolomas – eran significativamente distintos en función de los grupos analizados.

Se ha hecho un análisis de componentes principales, con el que se ha hecho uso de la función *prcomp* para realizar el análisis y del paquete *ggplot2* para graficar los resultados.

Además, se ha realizado un gráfico de calor (*heatmap*), el cálculo de correlaciones juntamente con el estudio de su significación, y un análisis PLS-DA con la función *plsDA*, con el que se ha hecho una predicción mediante la función *predict* y se han presentado los resultados en una matriz de confusión.

Por útlimo, se ha hecho un análisis de factores múltiples mediante la función *MFA* del paquete *factomineR*.

No se ha realizado ningún otro recurso informático, más que el Excel para la extracción de bases de datos con los resultados (es decir, la exportación de algunos datos de R a Excel) y el Word para la elaboración de la memoria.

A pesar de eso, se ha realizado un *dashboard* con el paquete *Shiny* de R con un resumen de los resultados obtenidos durante el análisis que se puede visualizar siguiendo el comando de R que aparece en el Anexo.

Para realizar el estudio se han obtenido los datos de un repositorio de Internet – Metabolomics Workbench[1] – seleccionando una base de datos sobre los metabolomas en los vinos blancos. De hecho, ya se había realizado un estudio sobre dichos datos, pero con un objetivo y unos resultados distintos a los que se han obtenido en este trabajo.

[1]https://www.metabolomicsworkbench.org/data/DRCCMetadata.php?Mode=Project&ProjectID=PR000005

I. ¿QUÉ SON LOS DATOS METABOLÓMICOS?

Definir y entender la metabolómica puede ser una tarea compleja y difícil, ya que está relacionada con una gran variedad de términos que son cruciales para comprenderla. En este capítulo se tratarán aquellos conceptos que pueden acercar al lector al mundo de la metabolómica y a aquellos aspectos que lo componen.

1. Introducción a la biología

Según la Real Academia Española (RAE), la biología se define como *la ciencia que trata de los seres vivos considerando su estructura, funcionamiento, evolución, distribución y relaciones* **(RAE, 2019)**. En la antigua Grecia ya se entendía como un sistema ordenado de conocimientos; de hecho, la palabra deriva de los términos griegos *bio* (vida) y *logia* (ciencia). Aun así, no fue hasta el siglo XIX cuando surgió de manera formal.

Aunque se inició con el objetivo de describir y clasificar el mundo viviente, se ha ido transformando hasta llegar a abarcar otros temas que son fundamentales para el estudio de los organismos, como su desarrollo, la herencia, y la interacción con el medio y con otros organismos, consiguiendo tener una amplia gama de aplicaciones prácticas.

Como consecuencia de esta gran diversidad de enfoques esta ciencia se ha diversificado en numerosas disciplinas, como pueden ser la anatomía, la fisiología, la genética, la medicina, la microbiología o la zoología, que, aunque mantienen una serie de principios y teorías generales, abarcan un amplio conjunto de campos de conocimiento, por lo que la biología se considera un conjunto de subdisciplinas, aunque los límites entre ellas son muy inseguros y frecuentemente se prestan técnicas las unas a las otras. Aun así, la biología no es una ciencia independiente, ya que requiere de otras disciplinas como la física, la química y las matemáticas.

A continuación, se van a detallar algunos de los aspectos más relevantes de la biología, con la finalidad de construir una base sólida para llegar a comprender los términos y aspectos más complicados que aparecerán y que son fundamentales para el seguimiento de este estudio.

1.1. Características de los seres vivos

A menudo se define la biología como la ciencia que estudia la vida. Sin embargo, el hecho de preguntarse qué es la vida propiamente ha sido un tema que ha causado muchas controversias des de hace años, y definirlo ha sido uno de los principales objetivos de la biología. En realidad, cuando se habla de “vida” se hace referencia al

proceso de vivir, pero también había opiniones encontradas sobre el hecho de diferenciar lo que está vivo y lo que no lo está. A pesar de que a lo largo de los años surgieron distintas corrientes que defendían sus ideas al respecto, en el último siglo dominan las ideas de una corriente que combina los principios válidos del fisicismo, que defendía que los organismos vivos no eran diferentes de la materia inanimada, y del vitalismo, que aseguraba que los organismos tenían propiedades que no existían en la materia inerte y que por lo tanto las teorías y conceptos biológicos no se podían reducir a las leyes de la física y la química: el organicismo. A modo muy general, el organicismo defiende que las características exclusivas de los seres vivos no se deben a su composición, sino a su organización.

Así pues, las funciones de los organismos vivos a nivel molecular obedecen las leyes de la física y la química, pero dichos organismos son distintos de la materia inerte; son sistemas ordenados jerárquicamente, con propiedades que no se han observado en la materia inanimada.

Los seres vivos se caracterizan por una serie de propiedades que les confieren ciertas cualidades: tienen una estructura organizada compleja, tienen la capacidad de responder a estímulos, la capacidad de crecer y desarrollarse, la capacidad de autorregulación, la capacidad de adquirir energía y materiales del exterior y transformarlos, la capacidad de reproducirse y la capacidad de evolucionar.

Por un lado, los organismos vivos presentan estructuras ordenadas que se coordinan entre sí, de forma que todas las funciones interactúan unas con las otras para crear un sistema viviente singular y ordenado. De hecho, la unidad de organización de los seres vivos es la célula, ya que todas las funciones vitales tienen su explicación en la estructura y funcionamiento de ésta. Las propiedades de las células están sustentadas en función de sus componentes, responsables del desarrollo y funcionamiento de los seres vivos; los ácidos nucleicos, los aminoácidos, las enzimas, las hormonas y los componentes membranosos. Así pues, la célula es la parte más simple de la materia viva capaz de realizar todas las actividades necesarias para la vida; algunos organismos son unicelulares (constan de una sola célula), pero los más complejos están formados por miles de millones de células y se denominan organismos pluricelulares.

Además, tienen la capacidad de identificar y responder a los estímulos[2] que reciben en su medio ambiente, ya sea interno y externo, gracias al desarrollo de órganos sensoriales y sistemas musculares.

[2] Cualquier agente físico o químico que origina una reacción en un organismo.

Los seres vivos también crecen y se desarrollan. El desarrollo abarca todos los cambios que se producen durante la vida de un organismo, aunque el crecimiento se refiere exclusivamente a los procesos que incrementan la cantidad de sustancia viva en el organismo, así que en realidad es un aumento de la masa celular.

La autorregulación, o homeostasis, se refiere a la capacidad de los organismos para mantener sus condiciones internas en niveles estables y constantes, aun cuando las condiciones externas cambien de forma drástica. Los seres vivos se caracterizan por poseer toda clase de mecanismos de control y regulación, así como múltiples mecanismos de retroalimentación, que mantienen el organismo en este estado estacionario.

Otro hecho fundamental que caracteriza la vida es el hecho de que los seres vivos intercambian sustancias y energía con el medio externo, de forma que funcionan como un sistema abierto. La vida, pues, es una suma de reacciones químicas interconectadas que requieren de energía y nutrientes para ser llevadas a cabo. De hecho, esa suma de reacciones químicas es conocida como metabolismo, concepto en el que se profundizará más adelante.

Por otro lado, la capacidad de auto reproducirse consiste en transmitir información genética a su descendencia y así generar nuevos seres vivos con sus mismas características. Sin ella los seres vivos no podrían persistir en el tiempo, generación tras generación. Dicha reproducción puede ser de tipo asexual, de forma que participa un solo progenitor que se divide, germina o fragmenta para formar descendientes, o de tipo sexual, que requiere la participación de células reproductoras (gametos) femeninas, los óvulos, y masculinas, los espermatozoides, que se unen y forman otra célula llamada cigoto a partir de la cual se formará un nuevo individuo.

Por último, la evolución es la capacidad de los organismos de sobrevivir en el tiempo a través de la adaptación a las condiciones ambientales y a la transmisión de estas características a su descendencia. La fuerza más importante en la evolución es la selección natural, que es el proceso mediante el cual los organismos que poseen rasgos que les ayudan a adaptarse a las condiciones de su medio sobreviven y se reproducen más satisfactoriamente que otros que carecen de tales rasgos.

1.2. Niveles de organización de los seres vivos

Los seres vivos pueden ser estudiados a diferentes niveles, ya que como toda la materia del universo están compuestos de átomos organizados en diferentes niveles de complejidad. Cada nuevo nivel de organización engloba a los niveles inferiores

anteriores. Así pues, al observar las interacciones que ocurren dentro de los grupos de organismos y entre un grupo y otro es posible detectar una jerarquía de complejidad cada vez mayor. En la **Figura 1** se muestran los diversos niveles de organización de los seres vivos.

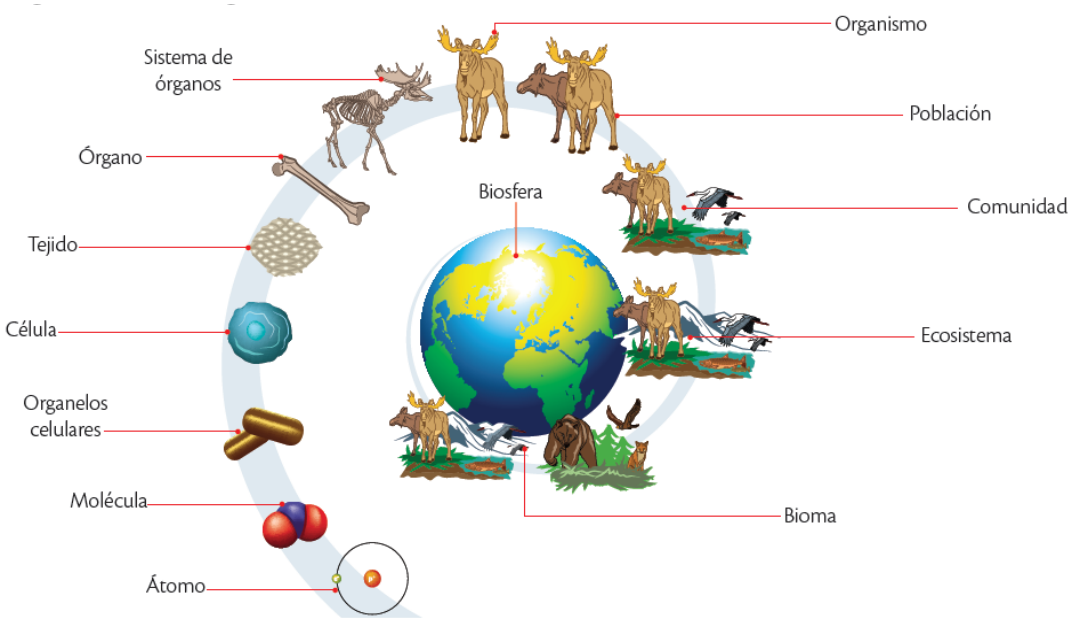

Figura 1[3]. Niveles de organización de los seres vivos

La **Figura 1** se puede dividir en dos partes: la organización a nivel del organismo, y la organización ecológica, que va más allá de los organismos individuales.

A nivel del organismo, a veces se determina como el primer nivel de organización es el subatómico, es decir, las partículas (protones, neutrones y electrones). Aunque en la figura anterior no se han mostrado, estas se organizan para formar el átomo, que a su vez se organiza en moléculas. Las moléculas que se asocian entre sí forman los organelos, que son estructuras complejas y altamente especializadas. En un nuevo nivel surgen las células, que son la unidad básica estructural y funcional de la vida. Cuando las células individuales se organizan formando un organismo multicelular, se crea otro nivel, los tejidos, que pueden constituir órganos. El siguiente nivel son varios órganos que en conjunto realizan una sola función, es decir, el sistema de órganos. Y, por último, todos los sistemas de órganos funcionando coordinadamente constituyen un ser vivo individual: el organismo.

[3] Fuente: https://www.traohh.com/2017/07/niveles-de-organizacion.html

A partir de aquí empieza la organización ecológica. Un grupo de organismos muy parecidos que potencialmente se entrecruzan forman una especie, y los miembros de ésta viven en una población determinada. Las poblaciones diversas que viven en una región determinada y que interactúan entre sí forman una comunidad, que puede estar formada por muchos tipos diferentes de formas de vida. Cuando se engloba la comunidad y el medio no viviente (el suelo, el agua y la atmósfera) se constituye un ecosistema, y toda la superficie de la Tierra que está habitada por seres vivos recibe el nombre de biósfera.

1.3. Diversidad de los organismos

En las poblaciones con reproducción sexual, no existen dos individuos idénticos, así como no existen dos poblaciones o dos especies idénticas. Así pues, la biodiversidad se puede definir como el conjunto de organismos y seres vivos que pueblan determinada zona o ecosistema durante cierto período de tiempo.

No está claro el número de especies distintas que hay; la diversidad biológica se estima en más de tres millones de especies distintas de seres vivos, aunque esta cifra podría llegar a los treinta millones. Aun así, se deben diferenciar tres tipos de diversidad: la específica, que se refiere a la diversidad de especies y es la más aparente y la primera que captamos, la diversidad genética, que facilita la adaptación de la especie a medios cambiantes y su evolución, y la diversidad de ecosistemas, siendo un ecosistema todas las condiciones que rodean a una determinada zona.

Durante años se ha intentado clasificar de alguna forma las especies de seres vivos, pero ha sido una tarea difícil. La unidad básica en que los biólogos clasifican los organismos es la especie, que se puede definir como un grupo de individuos semejantes entre sí, parecidos en sus características estructurales y funcionales, que en la naturaleza puede entrecruzarse libremente y producir descendientes fértiles. Aquellas especies que están íntimamente emparentadas se agrupan para formar la siguiente unidad de clasificación, el género, de forma que cada organismo recibe un nombre científico formado por dos palabras en latín: el género y la especie.

Aunque en el siglo XVIII sólo se reconocían dos grandes grupos de seres vivos (los animales y las plantas) esta clasificación ha ido evolucionando, y de hecho aún en la actualidad la clasificación de organismos está en construcción y constante cambio. Así pues, hay distintos criterios mediante los cuales los organismos pueden ordenarse. No obstante, la que actualmente se usa más es la que propuso el científico Robert Whittaker (1925 – 1980) quien sugirió agruparlos en cinco reinos, diferenciando los organismos de acuerdo con el tipo de célula que los constituyen y agrupándolos en

procariotas, que son las que no tienen núcleo, y eucariotas, en las que hay un citoplasma y un núcleo celular organizado.

Así pues, propuso el *reino monera*, formado por los organismos que poseen células procariotas (como pueden ser las bacterias); el *reino protista*, del cual forman parte los organismos eucariotas unicelulares junto con las algas (aunque algunas de ellas sean procariotas); el *reino fungi* (o hongos), en el que están ubicados los organismos heterótrofos que digieren los alimentos del exterior del cuerpo y luego absorben los nutrientes; el *reino metafita* (o plantas) que contiene a los fotoautótrofos, es decir, aquellos seres vivos que se nutren mediante el proceso fotosintético; y por último el *reino metazoa* (o animales) que ubica a los heterótrofos que ingieren sus alimentos.

En la **Tabla 1**, que se muestra a continuación, se muestra cada uno de los reinos con sus características principales.

		Reinos				
		MONERA	**PROTISTA**	**FUNGI**	**METAFITA**	**METAZOA**
Características	**Tipo de célula**	Procariota	Eucariota	Eucariota	Eucariota	Eucariota
	ADN	Circular	Lineal	Lineal	Lineal	Lineal
	Número de células	Unicelulares	Unicelulares / pluricelulares	Unicelulares / pluricelulares	Pluricelulares	Pluricelulares
	Nutrición	Autótrofos / Heterótrofos	Autótrofos / Heterótrofos	Heterótrofos	Autótrofos	Heterótrofos
	Energía que utilizan	Química / Lumínica	Química / Lumínica	Química	Lumínica	Química
	Reproducción	Asexual	Asexual / Sexual	Asexual / Sexual	Asexual / Sexual	Sexual
	Movilidad	Sí / No	Sí / No	No	No	Sí
	Pared Celular	Existe	Existe / No existe	Existe	Existe	No existe

Tabla 1[4]. Características de los cinco reinos

Para entender mejor el contenido de esta tabla, es adecuado hacer referencia a cada uno de los términos que aparecen. Anteriormente ya se han tratado los tipos de células y el número de células, así como el tipo de reproducción; aun así, no se ha hecho referencia aún a las otras características que se muestran en la tabla.

Por un lado, y en cuanto al tipo de nutrición, los organismos autótrofos son aquellos que sintetizan todas las sustancias esenciales para su metabolismo a partir de sustancias inorgánicas, de forma que pueden producir su propio alimento; en cambio, los heterótrofos se nutren de otros organismos para obtener la materia orgánica ya

[4] Fuente: elaboración propia.

sintetizada, porque no cuentan con un sistema de producción de alimentos independiente.

Por otro lado, mientras la energía química es la energía acumulada en los alimentos y en los combustibles que se produce por la transformación de sustancias químicas que contienen dichos alimentos o elementos, la lumínica es la energía percibida de la energía transportada por la luz.

Por último, la pared celular es una membrana resistente que protege el contenido de las células; es un elemento dinámico que determina qué puede entrar y salir de la célula.

2. El metabolismo

En el apartado anterior se ha hablado brevemente del metabolismo, definiéndolo como la suma de reacciones químicas que forman la vida. No obstante, es un término lo suficientemente complejo y relacionado con los datos metabolómicos como para dedicarle un apartado completo, poniendo énfasis en qué es concretamente y como funciona, así como otros aspectos que se verán a continuación.

2.1. ¿Qué es el metabolismo?

Se entiende por metabolismo el conjunto de transformaciones materiales que se efectúan en las células de los organismos vivos. En realidad, estos procesos de cambios químicos de materia se hacen con la finalidad de obtener energía; así pues, se transforman en el organismo los hidratos de carbono, las proteínas, las grasas y otras sustancias, produciendo calor, dióxido de carbono, agua y detritos[5], que serán los responsables de producir dicha energía que permitirá al organismo realizar transformaciones químicas esenciales y desarrollar actividad muscular.

Las reacciones químicas que integran el metabolismo no tienen lugar de manera independiente unas de otras, sino que están articuladas en lo que se llama rutas metabólicas: secuencias de reacciones consecutivas ligadas entre sí por intermediarios comunes. De esta forma, el producto de cada reacción resulta ser el sustrato o reactivo de la siguiente. Así pues, esta existencia de un intermediario común entre dos reacciones consecutivas hace posible la transferencia de energía química entre ellas.

En cada ruta metabólica suelen participar varias enzimas que catalizan las reacciones químicas que tienen lugar. Dado que las enzimas son proteínas codificadas en el ADN,

[5] Residuos, generalmente sólidos, que provienen de la putrefacción de fuentes orgánicas y minerales.

el daño en el material genético que codifique para una determinada enzima hará que la ruta metabólica en la que está implicada no funcione correctamente, o incluso puede bloquearla por completo.

Todos los seres vivos tienen un metabolismo, aunque puede ser distinto en función de las necesidades nutricionales de cada especie. Dependiendo del funcionamiento metabólico de cada ser vivo algunas sustancias le serán nutritivas y otras tóxicas o letales. A pesar de estas posibles diferencias entre los metabolismos de los seres vivos, lo cierto es que abundan las semejanzas, ya que todos los organismos están compuestos principalmente de carbono.

Se pueden determinar distintos tipos de metabolismo, aunque hay varias clasificaciones posibles. Una primera distinción sería identificar dos grupos: metabolismos autótrofos, que producen su propia energía, y metabolismos heterótrofos, que consumen el carbono de otras materias orgánicas, vivas o no. Sin embargo, si se clasifica teniendo en cuenta la fuente de nutrición, se observan cuatro tipos distintos de metabolismo: los fotolitrótrofos, los fotoorganótrofos, los quimiolitótrofos y los quimioorganótrofos.

Por un lado, los fotolitrótrofos o fotoautótrofos son quienes tienen capacidad de tomar fotones[6] de la luz del Sol como fuente de energía. Por otro lado, los fotoorganótrofos o fotoheterótrofos son los que tienen como fuente de energía la luz y como fuente de carbono, la materia orgánica. En cambio, los quimiolitótrofos o quimioautótrofos son aquellos capaces de utilizar compuestos orgánicos reducidos como substratos para el metabolismo respiratorio. Por último, los quimioorganótrofos obtienen su energía de reacciones de oxidorreducción y utilizan sustratos orgánicos.

Para concluir este subapartado introductorio, es relevante comentar la existencia de problemas que pueden afectar al metabolismo. En un sentido amplio, un trastorno metabólico es cualquier afección provocada por una reacción química anómala en las células del cuerpo. Cuando determinadas sustancias no se pueden metabolizar o lo hacen de forma incorrecta se puede provocar una acumulación de sustancias tóxicas en el cuerpo o una deficiencia de sustancias necesarias para el funcionamiento normal del cuerpo; ambas provocan síntomas graves.

2.2. Funcionamiento del metabolismo

Todo metabolismo se compone de dos procesos distintos pero conjugados, que se llaman catabolismo y anabolismo; estas son sus dos fases principales.

[6] Partículas elementales responsables de las manifestaciones cuánticas del fenómeno electromagnético.

Por un lado, el catabolismo es la fase degradativa del metabolismo. En ella las moléculas orgánicas complejas y relativamente grandes (como pueden ser las proteínas) se degradan para dar lugar a moléculas de estructura más simple y menor tamaño, como el CO2, agua, amoníaco, urea o ácido láctico. Este proceso degenerativo va acompañado de la liberación de la energía química relacionada con la estructura de las moléculas orgánicas que se degradan. Se dice, pues, que es un proceso exergónico, ya que libera energía. Además, muchas de las reacciones del catabolismo suponen una oxidación, es decir, una pérdida de electrones, de los sustratos orgánicos que se degradan.

Por otro lado, el anabolismo es la fase constructiva del metabolismo. En ella tiene lugar la síntesis de los componentes moleculares de las células (como los ácidos nucleicos, las proteínas o los lípidos) a partir de moléculas precursoras de estructura más sencilla y de menor tamaño. Es un proceso endergónico, ya que requiere energía química para poder ser llevado a cabo. Además, la construcción de biomoléculas orgánicas altamente hidrogenadas requiere electrones para reducir a sus precursores relativamente oxidados.

Las rutas metabólicas que forman parte del catabolismo se denominan rutas catabólicas, y las que forman parte del anabolismo, rutas anabólicas. Aquellas rutas que son comunes en ambas fases se llaman rutas anfibólicas.

En resumen, se puede decir que mientras el catabolismo es un proceso degradativo, oxidante y exergónico, el anabolismo es un proceso constructivo, reductor y endergónico. No obstante, y aunque parecen dos procesos que transcurren por separado en el espacio y en el tiempo, lo cierto es que ambos tienen lugar simultáneamente en el citoplasma celular, ya que las células están permanentemente en un proceso de renovación de sus componentes moleculares. En la **Figura 2** se resumen de forma visual ambas fases.

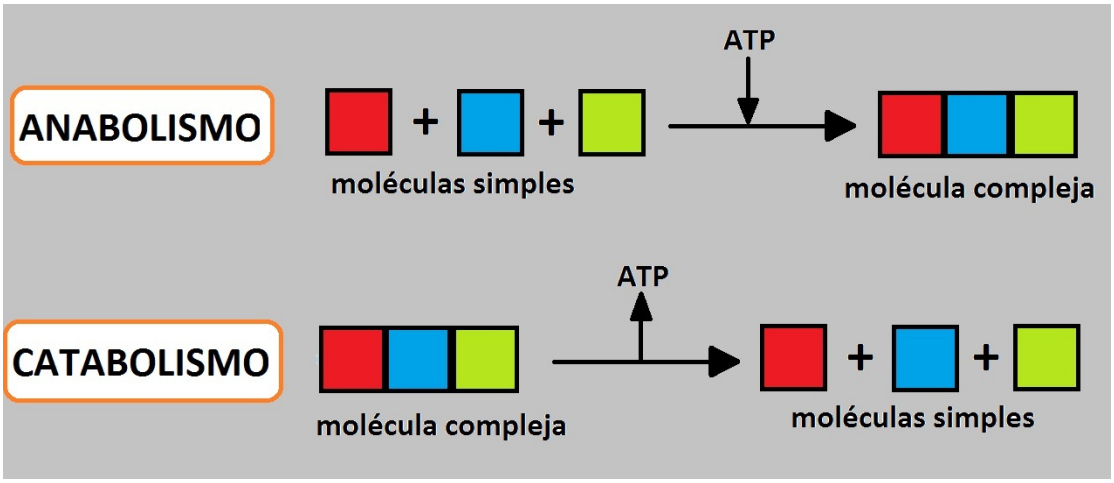

Figura 2[7]. Fases del metabolismo

2.3. Biomoléculas principales

Una de las funciones principales del metabolismo es sintetizar o degradar las moléculas básicas, de las que forman parte la mayor parte de las estructuras constitutivas de los organismos: aminoácidos, glúcidos y lípidos. Así pues, las sintetiza en la construcción de células y tejidos o bien las degrada para utilizarlas como recurso energético en la digestión. El objetivo de este apartado es describir las biomoléculas principales que participan en el funcionamiento del metabolismo.

Por un lado, se encuentran los aminoácidos, que de hecho son compuestos orgánicos que se combinan para formar proteínas. Los aminoácidos se utilizan o bien para producir proteínas con el fin de ayudas a descomponer alimentos, crecer, reparar tejidos corporales y llevar a cabo otras funciones corporales, o bien como una fuente de energía.

Por otro lado, los glúcidos, que también se pueden llamar carbohidratos, son biomoléculas que aportan energía a los seres vivos, ya sea para su uso inmediato o para su almacenamiento. Además, permiten que la temperatura corporal y la presión arterial se mantengan estables. Los carbohidratos básicos se denominan monosacáridos e incluyen galactosa, fructosa y glucosa.

Por último, los lípidos son moléculas orgánicas solubles en solventes diferentes al agua, compuestas principalmente de carbono e hidrógeno. Son las biomoléculas que presentan más biodiversidad, y su función estructural básica consiste en formar parte de membranas biológicas (como puede ser la membrana celular) o servir como recurso energético.

[7] Fuente: https://culturismototal.blogspot.com/2016/08/anabolismo-y-catabolismo.html

Aunque las tres moléculas detalladas son las básicas, también hay otros elementos que forman parte del proceso que lleva a cabo el metabolismo, como pueden ser los nucleótidos, que son la unidad estructural básica y el bloque de construcción del ADN, las coenzimas, pequeñas moléculas orgánicas no proteicas que transportan grupos químicos entre enzimas, y los minerales, esenciales para la actividad de algunas enzimas y de proteínas transportadoras de oxígeno.

2.4. Los metabolitos

Un metabolito es cualquier sustancia producida durante el metabolismo, aunque también puede referirse al producto que queda después de la descomposición de un fármaco por parte del cuerpo.

En este caso, la definición a la que se pone énfasis es la primera. Así pues, los metabolitos son compuestos que participan en las reacciones químicas que tienen lugar a nivel celular, cuyo conjunto, como se ha visto anteriormente, constituye el metabolismo celular, que es la base molecular de la vida.

La sustancia final de una ruta metabólica se conoce como metabolito final, aunque una ruta metabólica puede generar varios de ellos, y las sustancias intermedias como metabolitos intermedios. Un metabolito intermedio o final en una ruta metabólica puede ser el sustrato en otra, lo que hace que la gran mayoría de rutas metabólicas estén interconectadas.

Entre los metabolitos presentes en el cuerpo humano no se encuentran solo compuestos endógenos, sino que también los productos de aquellos que los organismos ingieren o con los que están en contacto, que se llaman metabolismos exógenos.

Los metabolitos se pueden clasificar en dos grandes grupos: los primarios y los secundarios. Los primarios se definen como aquellos que están involucrados de forma directa en el crecimiento, desarrollo y reproducción normal de un organismo con una función fisiológica importante, y suelen ser comunes a amplios grupos de seres vivos. Por otro lado, los secundarios no están involucrados en estos procesos de forma directa, y suelen ser específicos de grupos de especies más reducidos y que generalmente tienen una relación filogénica estrecha entre sí.

Los metabolitos finales que no participan en otras rutas metabólicas se conocen como productos de desecho, y no pueden ser utilizados por el organismo para sintetizar

sustancias, así que son excretados a través de distintas vías como pueden ser la orina o el sudor.

La ausencia de un metabolito primario suele conllevar la muerte inmediata o a corto plazo, mientras que la ausencia de uno secundario no. Aun así, ambos pueden tener funciones importantes, como inhibir a una enzima o activar otra ruta metabólica; de hecho, muchos fármacos son metabolitos secundarios producidos por otros seres vivos.

3. Tecnologías ómicas

La llegada de las tecnologías ómicas ha impulsado la expansión de la biología de sistemas. La biología de sistemas es un término más complejo y difícil de definir; de hecho, hay pocos autores que se atreven a hacerlo. Se dice que la biología de sistemas se basa en la comprensión de que el todo es mayor que la suma de las partes, lo que se puede condensar en que es un enfoque para descifrar la complejidad de los sistemas biológicos que parte de la comprensión de que las redes que forman la totalidad de los organismos vivos son más que la suma de sus partes. Al igual que la biología, integra muchas disciplinas científicas, como la propia biología, las ciencias de la computación, la ingeniería, la bioinformática, y otras. Principalmente, pretende predecir como estos sistemas biológicos cambian con el tiempo y bajo diferentes condiciones, y desarrollar soluciones a los principales problemas de salud y ambientales.

Antes de hablar propiamente de las tecnologías ómicas, pero, es importante entender qué quiere decir el término “ómica”. El término se utiliza para definir el estudio de los diferentes sistemas biológicos que conforman el funcionamiento de las células. En el área de la biología molecular se utiliza para definir el estudio de los diferentes sistemas biológicos que conforman el funcionamiento de las células.

Así pues, las tecnologías ómicas son campos de investigación que permiten procesar grandes cantidades de datos, ya que posibilitan el análisis de genomas, proteomas o metabolomas completos. La más conocida hasta el momento ha sido la genómica, gracias al Proyecto Genoma Humano, del cual se hablará más adelante. No obstante, el propósito de descubrir los misterios del fenotipo impulsó el desarrollo de otras tecnologías ómicas. Actualmente las tecnologías ómicas permiten indagar en elementos des de los genes hasta los metabolitos; es gracias a la genómica, la transcriptómica, la proteómica y la metabolómica. La **Figura 3** representa la cascada ómica, mostrando la especialidad de cada una de ellas.

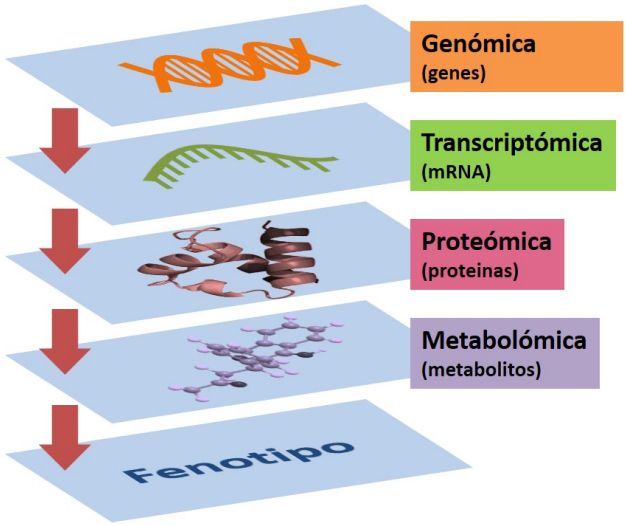

Figura 3[8]. Cascada ómica.

3.1. La genómica

La genómica es un área de la biología cuyo objetivo es el estudio de los genomas a escala global. Se popularizó gracias al Proyecto Genoma Humano, que pretendía la secuenciación total del genoma humano, y fue completado en 2003, ofreciendo a la comunidad científica información detallada sobre la estructura, organización y función del conjunto completo de genes humanos, es decir, el genotipo.

3.1.1. LOS GENES

Previo a detallar aquellos puntos más importantes de la genómica, es de suma relevancia saber qué es exactamente un gen, y para hacerlo, tenemos que hablar antes del ADN, ya que, en rasgos generales, un gen es un segmento corto de ADN.

El ADN (ácido desoxirribonucleico) es un ácido nucleico que tiene la función de almacenar y transmitir la información genética. Además, dirige el proceso de síntesis de proteínas, constituye el material genético y forma los genes. Las órdenes contenidas en el ADN son ejecutadas por el ARN (ácido ribonucleico). El ADN de cada célula consta de miles de millones de bases nucleótidas (A-denina, T-timina, G-guanina y C-citosina), que son las letras con las que se escribe la información genética.

[8] Fuente: https://culturacientifica.com/2019/02/22/metabolomica-el-todo-sobre-la-suma-de-las-partes/

Por lo tanto, los genes son las unidades de almacenamiento de información genética, es decir, segmentos de ADN que contienen la información sobre cómo deben funcionar las células del organismo.

Los genes se encuentran en los cromosomas, estructuras en el interior de la célula que contienen la información genética. Cada cromosoma de las células está formado por una molécula de ADN, asociada a ARN y proteínas. Cada especie tiene un número característico de cromosomas; por ejemplo, en la especie humana cada célula somática tiene 22 pares de cromosomas autosómicos y un par de sexuales.

La **Figura 4** muestra la relación descrita entre los cromosomas, el ADN y los genes.

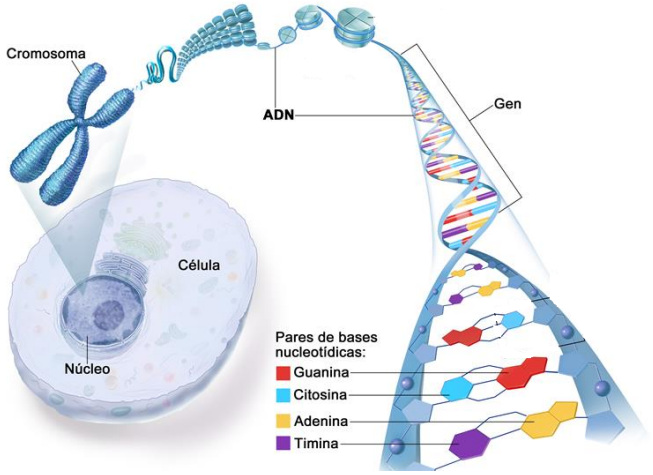

Figura 4[9]. Ilustración de los cromosomas junto con el ADN y los genes.

Como unidad que conserva datos genéticos, el gen se encarga de transmitir la herencia a los descendientes de una especie. El conjunto de genes pertenecientes a una misma especie se define como genoma. Aun así, la tarea de los genes es mucho más compleja; de hecho, son imprescindibles para conseguir que el ARN funcional pueda ser sintetizado: la transcripción genética produce una molécula de ARN que luego se traduce en los ribosomas, que son las macromoléculas responsables de la síntesis o traducción de los aminoácidos del ARNm (en células eucariotas) y de la producción de las proteínas en los seres vivos, para luego generar las proteínas. Sin embargo, hay genes que no son traducidos a proteínas y que cumplen otros roles en forma de ARN.

Además, los genes inciden en el desarrollo de enfermedades, ya que una variación en una secuencia puede provocar las enfermedades genéticas, que se heredan de generación en generación. Esas enfermedades fueron la principal motivación para intentar determinar la secuencia de las bases químicas que forman el ADN e identificar

[9] Fuente: https://www.cancer.gov/espanol/cancer/causas-prevencion/genetica

todos los genes del genoma del ser humano, lo que impulsó el desarrollo de la genómica.

3.1.2. PROFUNDIZACIÓN EN LA GENÓMICA

En concreto, la genómica es el estudio completo de ADN y todos sus genes de un organismo. El estudio del genoma ayuda a entender la interacción entre los genes entre sí y con el entorno, así como la manera en que surgen ciertas enfermedades, como el cáncer, la diabetes y las afecciones del corazón. De hecho, uno de los objetivos principales de la genómica es descubrir nuevas maneras de diagnosticar, tratar y prevenir enfermedades.

La genómica comprende el estudio del contenido, organización, función y evolución de la información genética en un genoma completo. El término es relativamente reciente; se considera que surgió gracias a Thomas Roderick en 1986 para referirse a la subdisciplina de la genética dedicada al estudio de la cartografía, secuenciación y análisis de las funciones de genomas completos.

Para hablar de la genómica se debe hacer referencia al Proyecto del Genoma Humano, ya que como se ha dicho anteriormente es uno de los proyectos más importantes en genómica, si no el más importante, ya que fue el punto de partida de otras investigaciones para entender plenamente cómo funciona el genoma y descubrir las bases genéticas de la salud y las enfermedades. Este proyecto fue dirigido en los Institutos Nacionales de la Salud (NIH) por el Instituto Nacional de Investigación del Genoma Humano. Como ya se ha comentado, en el estudio se produjo una versión de la secuencia del genoma humano de muy alta calidad; es una secuencia combinada derivada de varios individuos, así que es totalmente representativa. Este proyecto generó de esta forma un recurso que puede ser utilizado en una amplia gama de estudios biomédicos.

Así pues, la genómica se ha convertido en un soporte central y unificador de distintas áreas de investigación biológica. No obstante, el hecho de permitir estudiar la información hereditaria de uno o múltiples organismos de forma integral, simultánea y a gran escala provoca que se generen una gran cantidad de datos, así que se requieren grandes capacidades computacionales para poder ser analizados, almacenados y gestionados. Para poder hacerlo se utiliza la bioinformática, otra disciplina con cada vez más importancia.

Dentro de la genómica se pueden encontrar varias subdisciplinas, como la genómica estructural, que se encarga de elucidar la estructura de los genomas completos; la

genómica comparada, relacionada con la anterior, ya que aplica la información que ésta extrae en una mejor comprensión de la evolución y la historia de la vida; la genómica de poblaciones, que permite conocer la variabilidad genética de una especie a nivel genético; la metagenómica, que analiza de forma conjunta el genoma de comunidades microbianas complejas para determinar su biodiversidad y las relaciones entre los organismos presentes; o la genómica funcional, basada en el estudio del funcionamiento global del genoma de los organismos.

Sin embargo, para conseguir los propósitos de la genómica funcional es necesario recurrir a la transcriptómica, la proteómica y la metabolómica, que se tratarán en los siguientes subapartados.

3.2. La transcriptómica

La transcriptómica trata de analizar el conjunto de ARNs transcritos a partir del genoma para un momento y condición determinados. Así pues, antes de profundizar en esta tecnología ómica se hará una introducción al ARN, para comprender mejor sus objetivos.

3.2.1. EL ÁCIDO RIBONUCLEICO (ARN)

El ARN es un ácido encargado de trasladar la información genética del ADN con el fin de sintetizar las proteínas según las funciones y características indicadas. Al igual que el ADN, está presente en el citoplasma de las células eucariotas y procariotas, aunque está compuesto tan solo por una cadena simple, que puede duplicarse en algunas ocasiones.

Dicha cadena está formada por nucleótidos. Cada uno de ellos contiene un grupo fosfato, un azúcar (la ribosa) y una base nucleica o base nitrogenada. En concreto, hay cuatro bases de nucleótidos en el ARN: la adenina, la guanina (ambos presentes en el ADN), la unacitosina y el uracilo. Por su parte, los nucleótidos están unidos por fosfodiéster, un enlace covalente que se da entre dos átomos de oxígeno de un grupo fosfato y los grupos hidroxilo[10] de otras dos moléculas distintas.

En resumen, el ARN copia la información de cada gen del ADN y luego pasa al citoplasma, donde se une al ribosoma para dirigir la síntesis proteica. Aun así, puede realizar muchas otras funciones; en particular, intervenir en reacciones químicas de metabolismo celular.

[10] Grupo funcional compuesto de 1 átomo de oxígeno y 1 átomo de hidrógeno que se une a un hidrocarburo

Hay cuatro tipos distintos de ARN: el ARN mensajero (ARNm), el ARN ribosómico (ARNr), el ARN transferente (ARNt) y el ARN heteronuclear (ARNh).

Por un lado, el ARN mensajero es un ARN lineal que contiene información, copiada del ADN, para sintetizar una proteína. Se forma en el núcleo celular, a partir de una secuencia de ADN; luego, sale del núcleo y se asocia a ribosomas, donde se forma la proteína.

Por otro lado, el ARN ribosómico, que es el tipo de ARN más abundante, es la parte central de los ribosomas, y se encargan de crear los enlaces peptídicos[11] entre los aminoácidos del polipéptido[12] en formación durante la síntesis de proteínas.

El ARN transferente, que a diferencia del mensajero no es lineal, se encarga de llevar los aminoácidos a los ribosomas con el fin de incorporarlos al proceso de síntesis proteica. Además, se encarga de codificar la información que posee el ARN mensajero a una secuencia de proteínas.

Por último, el ARN heteronuclear agrupa todos los tipos de ARN que acaban de ser transcritos (pre-ARN). Se encuentra en el núcleo de las células eucariotas, y en cambio no aparece en las procariotas. Su función consiste en ser el precursor de los distintos tipos de ARN.

Como se verá en la siguiente explicación sobre transcriptómica, el tipo de ARN más presente en los estudios de esta tecnología ómica son los mensajeros, ya que son los encargados de sintetizar las proteínas.

3.2.2. PROFUNDIZACIÓN EN LA TRANSCRIPTÓMICA

Como ya se ha avanzado anteriormente, la transcriptómica es una tecnología ómica que se encarga de estudiar y comparar los conjuntos de ARN mensajeros (transcriptomas) presentes en una célula, tejido u organismo. Así pues, se usa para aprender más de cerca la manera en que los genes se transforman en diferentes tipos de células y cómo esto puede ayudar a la presentación de ciertas enfermedades, como puede ser el cáncer. La transcriptómica se ha utilizado en campos muy diversos, como la medicina reproductiva para la evaluación de la receptividad endometrial.

[11] Enlace químico que se establece entre el grupo carboxilo de un aminoácido y el grupo amino de otro aminoácido.

[12] Secuencia de aminoácidos que están vinculados a través de enlaces peptídicos.

Hay que tener en cuenta, pero, que mientras en algunas células la actividad transcripcional es constante durante toda su vida en otras depende de estados fisiológicos o patológicos y estímulos específicos.

La transcriptómica es, por lo tanto, el estudio de los perfiles de expresión de todos los genes presentes en el genoma. El método transcriptómico más utilizado es el de microarrays de ADN, que permite el análisis simultaneo de la expresión de miles de genes y analizar su expresión bajo distintas condiciones experimentales. Estos microarrays surgen de la necesidad de analizar la información procedente de los grandes proyectos de secuenciación de genomas, y permiten elaborar mapas finos de transcripción, además de proporcionar información indirecta de los niveles de proteínas.

Existe una amplia gama de técnicas utilizadas en transcriptómica, las cuales permiten cuantificar millones de moléculas de ARN al mismo tiempo. Igual que la genómica, la transcriptómica se vale de la bioinformática, además de usar las micromatrices (o microarreglos). La idea de éstas últimas es construir sobre una membrana o lámina de vidrio arreglos de muestras que contienen fragmentos de ADN, marcando por otro lado el ARN o ADN copia de una población con fluorescencia o radioactividad, y usando esta preparación para hibrida con el ADN de la micromatriz. Generalmente se hibrida simultáneamente la misma micromatriz con una muestra de ARN o ADN copia de referencia para facilitar la comparación.

Uno de los objetivos más importantes de la transcriptómica es identificar qué porción del genoma es transcrito en cada tipo celular y en qué condiciones.

3.3. La proteómica

La proteómica es una tecnología en desarrollo que investiga la estructura y función del conjunto de proteínas que conforman el proteoma.

Aunque el término “proteína” ya ha aparecido en numerosas ocasiones hasta el momento, no se ha detallado qué son las proteínas en su totalidad. Así pues, igual que se ha hecho en los apartados anteriores, primero se estudiarán las proteínas y proteomas individualmente, y luego se investigarán más aspectos sobre esta tecnología ómica.

3.3.1. LAS PROTEÍNAS

Una proteína es *una molécula compuesta por aminoácidos que el cuerpo necesita para funcionar de forma adecuada. Las proteínas son la base de las estructuras del cuerpo,*

tales como la piel y el cabello, y de sustancias, como las enzimas, las citosinas y los anticuerpos **(Instituto Nacional del Cáncer)**. Están codificadas en el material genético de cada organismo, donde se especifica su secuencia de aminoácidos, y luego son sintetizadas por los ribosomas, como se ha comentado anteriormente. Este proceso está resumido en la **Figura 5**, que se muestra a continuación.

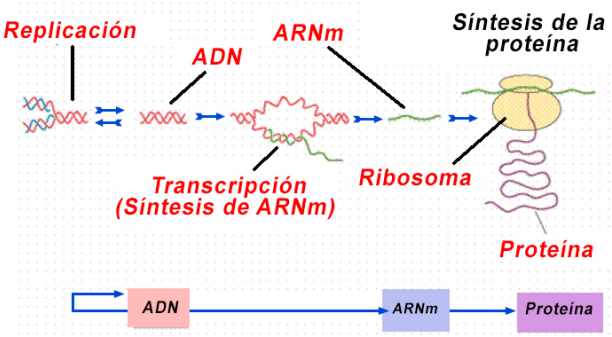

Figura 5[13]. Proceso de formación de las proteínas

Todas las proteínas poseen la misma estructura química central, que consiste en una cadena lineal de aminoácidos. De hecho, podemos distinguir cuatro niveles en la estructuración de las proteínas: la estructura primaria, que es la secuencia de aminoácidos de la proteína e indica los aminoácidos que componen la cadena polipeptídica y el orden en que se encuentran; la estructura secundaria, que se trata de la disposición de la cadena polipeptídica en el espacio y puede ser de tres tipos distintos en función de la forma de su plegamiento (hélice, lámina plegada o hélice de colágeno); la estructura terciaria, que informa sobre la disposición de la estructura secundaria de un polipéptido al plegarse sobre sí mismo originando una conformación globular, cosa que facilita a la proteína su solubilidad en agua y el desarrollo de funciones como el transporte o hormonales; y la estructura cuaternaria, que informa de la unión de varias cadena polipeptídicas con estructura terciaria para formar un complejo proteico.

La **Figura 6** resume esta explicación sobre la estructura de las proteínas de forma más sencilla.

[13] Fuente: https://www.paxala.com/el-proceso-de-sintesis-de-las-proteinas/

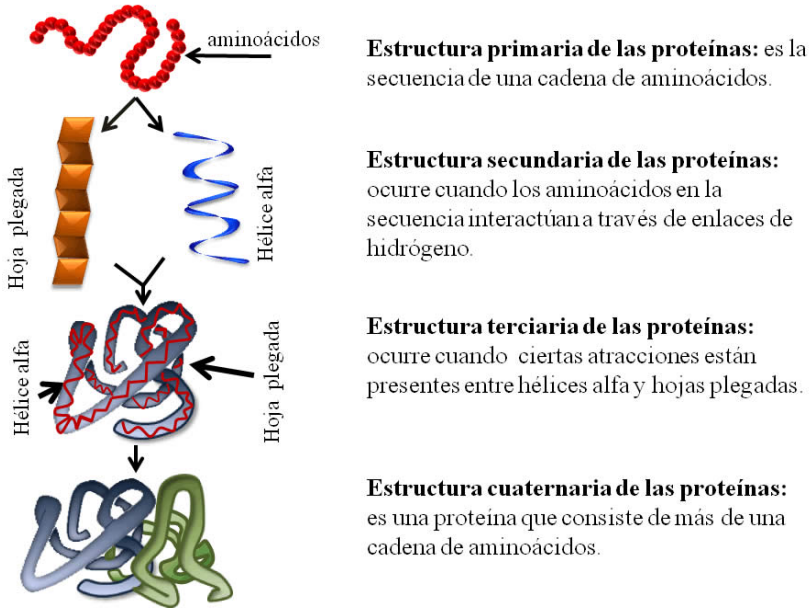

Figura 6[14]. Estructura de las proteínas.

Las proteínas tienen dos propiedades básicas: la especificidad y la desnaturalización. La especificidad se refiere a su función: cada una lleva a cabo una determinada función y lo realiza porque posee una determinada estructura primaria y una conformación espacial propia. Así pues, un cambio en la estructura de la proteína puede significar una pérdida de la función. En cambio, la desnaturalización consiste en la pérdida de la estructura terciaria por romperse los puentes que la forman. La desnaturalización puede ser producida por cambios de temperatura, por ejemplo. En algunos casos, si las condiciones se restablecen, una proteína desnaturalizada puede volver a su anterior plegamiento o conformación. Este proceso es denominado renaturalización.

Las funciones de las proteínas son de gran importancia, varias y bien diferenciadas. Las proteínas determinan la forma y la estructura de las células y dirigen casi todos los procesos vitales. Estas funciones son específicas de cada tipo de proteína, y permiten a las células defenderse de agentes externos, mantener su integridad o controlar y regular funciones. Las principales funciones de las proteínas son la estructural (o de resistencia), la enzimática, la hormonal, la defensiva, la de transporte, la de reserva, la de regulación, la de contracción muscular y la función homeostática.

[14] Fuente: https://www.uaeh.edu.mx/scige/boletin/icbi/n1/e7.html

3.3.2. PROFUNDIZACIÓN EN LA PROTEÓMICA

Como se ha comentado anteriormente, la proteómica enfoca sus estudios en el proteoma, que describe el conjunto completo de proteínas que se expresan según el genoma y factores externos durante la vida de una célula.

El interés de la proteómica reside en el conocimiento del conjunto de las interacciones entre proteínas para constituir la red de interacciones que caracteriza el funcionamiento de los organismos vivos; así pues, de forma resumida se puede decir que la proteómica es el estudio del proteoma. Se debe poner especial énfasis en que esta tecnología no estudia una determinada proteína, sino que se centra en el estudio del funcionamiento del conjunto de todas ellas.

Las tecnologías clásicas más utilizadas en proteómica son la electroforesis y la espectrometría de masas, pero al igual que las tecnologías ómicas vistas anteriormente, se apoya en la bioinformática, ya que ofrece programas capaces de determinar las reacciones metabólicas que tiene una determinada proteína con el resto del proteoma.

3.4. La metabolómica

La metabolómica es una disciplina en rápido crecimiento que se centra en el estudio global de metabolitos de moléculas pequeñas en sistemas biológicos. Gracias a la caracterización de su dinámica, las interacciones y las respuestas a las perturbaciones genéticas o ambientales, proporciona un panorama completo de la fisiología de referencia y las respuestas bioquímicas globales a factores genéticos (los relacionados con la herencia biológica), abióticos (es decir, que no tienen vida) y bióticos (aquellos organismos vivos que influencian la forma de un ecosistema, como la flora y la fauna de un lugar y sus interacciones).

Es importante recordar que los metabolitos, como ya se ha comentado en el apartado que se les ha dedicado, son cualquier molécula utilizada, capaz o producida durante el metabolismo.

La metabolómica, pues, permite identificar y determinar el conjunto de metabolitos o metabolitos específicos en muestras biológicas, y compararlos en condiciones normales contra estados alertados (por ejemplo, enfermedades, modulación ambiental o intervención dietética). Como ya se ha visto, es una herramienta de fenotipado y proporciona la imagen de los metabolitos de un organismo en el curso de un proceso biológico.

En resumen, como los metabolitos indican los puntos finales de la expresión genética y celular, la metabolómica tiene un papel fundamental en la biología de sistemas, ya que puede ayudar a comprender el fenotipo de un organismo.

Esta tecnología se estudiará con más profundidad en el siguiente apartado de este capítulo.

3.5. El fenotipo

Un fenotipo es cualquier característica o rasgo observable de un organismo, como su morfología, desarrollo, propiedades bioquímicas, fisiológicas y comportamiento. Los fenotipos resultan de la expresión de los genes de un organismo, así como de la influencia de los factores ambientales, y de las posibles interacciones entre ambos.

Así pues, el fenotipo no es simplemente un producto del genotipo, sino que se ve influido por el medio ambiente en mayor o menor medida. De hecho, no es un elemento que esté establecido des de un principio, sino que puede ser modificado por las relaciones que el organismo mantiene con el ambiente que lo rodea y que lo hacen un producto de un complejo número de vínculos.

Es importante remarcar la diferencia básica con el genotipo de un organismo: mientras el genotipo está compuesto solamente por los rasgos genéticamente adquiridos, el fenotipo es lo que, sumado a estos rasgos, contiene también a los posibles cambios y variaciones que ese conjunto genético observa a partir de las interacciones con el medio.

4. Los datos metabolómicos

Como se ha comentado anteriormente, la metabolómica es una disciplina amplia y compleja que requiere de diversos pasos para llegar desde la cuestión biológica, es decir, el planteamiento del problema, hasta la interpretación de los resultados.

Aunque el metaboloma no es tan fácilmente definible como el genoma o el proteoma, la metabolómica busca estudiarlo por completo.

Hay dos enfoques analíticos generales para realizar un análisis metabolómico: el enfoque dirigido y el no dirigido. El primero se refiere a la detección y cuantificación precisa de un pequeño conjunto de compuestos conocidos, y está impulsado por una pregunta o hipótesis bioquímica específica en la que el conjunto de metabolitos relacionados con una o más vías ya está definido; así pues, requiere que los compuestos de interés se conozcan a priori y estén disponibles de forma purificada. En

cambio, en enfoque no dirigido (que de hecho también es llamado “huella digital del metabolito”, término que ha aparecido anteriormente) no está impulsado por una hipótesis a priori y se usa para la comparación completa del metaboloma (se miden y comparan la mayor cantidad posible de metabolitos entre muestra), por lo que este no intenta cuantificar con precisión todos los metabolitos medibles en una muestra, sino que proporciona su cuantificación relativa. Los dos enfoques tienen limitaciones que se deben tener en cuenta: por un lado, el enfoque dirigido se basa en la hipótesis a priori que ya hemos comentado, y además no se puede utilizar solo para un análisis exhaustivo del metaboloma; y por otro lado, el enfoque no dirigido tiene las barreras de las limitaciones tecnológicas, como el posible sesgo hacia la detección de las moléculas más abundantes, y además la misma molécula puede fragmentarse de manera diferente según el instrumento o técnica utilizada, cosa que dificulta la comparación entre espectros de los metabolitos (es decir, hay variabilidad dependiente del instrumento que hace aumentar también la variabilidad en la identificación de compuestos).

En cuanto a los campos de aplicación de los datos metabolómicos, des en inicio hasta la actualidad el más común ha sido la metabolómica de las plantas. No obstante, y con la llegada de la medicina de precisión, la metabolómica clínica ha ido ganando peso, y se aplica cada vez más en múltiples casos, como por ejemplo el diagnóstico de enfermedades o nuevos farmacológicos.

Como se ha descrito al hablar de los metabolitos, hay metabolismos endógenos y exógenos; ambos constituyen una familia muy heterogénea de moléculas, con estructuras, propiedades fisicoquímicas y concentraciones muy diversas. Esta heterogeneidad hace que sea imposible medir simultáneamente todo el metaboloma usando una sola técnica. Es por ello por lo que para poder cubrir el máximo rango posible de metaboloma se deben usar diferentes plataformas analíticas, a las que va dedicado el siguiente subapartado.

4.1. Plataformas analíticas para datos metabolómicos

Hay diversas plataformas analíticas disponibles para analizar datos metabolómicos. Especialmente durante los inicios de la metabolómica primaba el uso de la resonancia magnética nuclear (NMR); sin embargo, la metabolómica basada en la espectrometría de masas (MS) ha ido ganando popularidad con el tiempo gracias al desarrollo de instrumentos de alta resolución, como la resonancia ciclotrónica con transformada de Fourier (FTICR), el Orbitrap o el tiempo de vuelo (TOF), así como a los bajos límites de detección y la rapidez del análisis.

Aunque existen algunos estudios en los que la muestra se introduce al MS por infusión directa (DI-MS) lo más común es que se acople al espectrómetro una técnica de separación que ayude a reducir la complejidad de los espectros y a disminuir la supresión iónica debida a la competición por la ionización de los miles de moléculas presentes simultáneamente en la muestra. Dependiendo de los analitos de interés se usarán distintas técnicas. Para estudiar compuestos volátiles la técnica elegida es la cromatografía de gases acoplada a MS (GC-MS), pero tiene el inconveniente de que para analizar metabolitos no volátiles se deben tratar; así pues, la técnica más utilizada es la cromatografía líquida acoplada a MS (LC-MS).

Sin embargo, y dado que ninguna de las técnicas es capaz por si sola de analizar el metaboloma completo, se recomienda el uso de técnicas complementarias. A continuación, se estudiarán en detalle las tres técnicas principales mencionadas: NMR, LC-MS y GC-MS.

4.1.1. RESONANCIA MAGNÉTICA NUCLEAR (NMR)

La espectroscopia de resonancia magnética nuclear es una técnica de química analítica utilizada en el control de calidad y en la investigación para determinar el contenido y la pureza de una muestra, así como su estructura molecular. La NMR puede analizar cuantitativamente mezclas que contienen compuestos conocidos; para compuestos desconocidos, puede usarse para comparar con bibliotecas espectrales o para inferir la estructura básica de manera directa.

Una vez se conoce la estructura básica, la NMR se puede usar para determinar la conformación molecular en la solución y para estudiar las propiedades físicas a nivel molecular, como los cambios de fase o la solubilidad, entre otros.

Así pues, estudia el comportamiento de ciertos núcleos atómicos, concretamente aquellos que poseen un *spin* nuclear distinto a cero en presencia de un campo magnético externo.

Un *spin* es una propiedad fundamental de la naturaleza como la carga eléctrica o la masa. Es una propiedad física de las partículas subatómicas, por la cual toda partícula elemental tiene un momento angular intrínseco de valor fijo. De esta forma, el *spin* proporciona una medida de este momento angular intrínseco de toda partícula.

Este campo magnético aplicado comentado produce un desdoblamiento de los niveles degenerados de energía del *spin* nuclear, de modo que pueden inducirse transiciones entre ellos como consecuencia de la absorción de una radiación electromagnética

adecuada. Como la disposición de los niveles de energía es una propiedad tanto de los núcleos de una molécula como de su entorno electrónico y de las interacciones entre ambos, la intensidad, forma y posición de las señales en el espectro de un núcleo determinado están íntimamente relacionadas con su estructura molecular, así que un análisis detallado del espectro proporciona valiosa información acerca de la estructura del compuesto que lo origina.

Es por eso por lo que esta técnica resulta ser de las más eficientes y útiles para el estudio de la estructura y dinámica de moléculas en disolución.

A continuación, en la **Figura 7** se muestra de forma esquemática los principales componentes de un equipo para medidas de resonancia magnética nuclear.

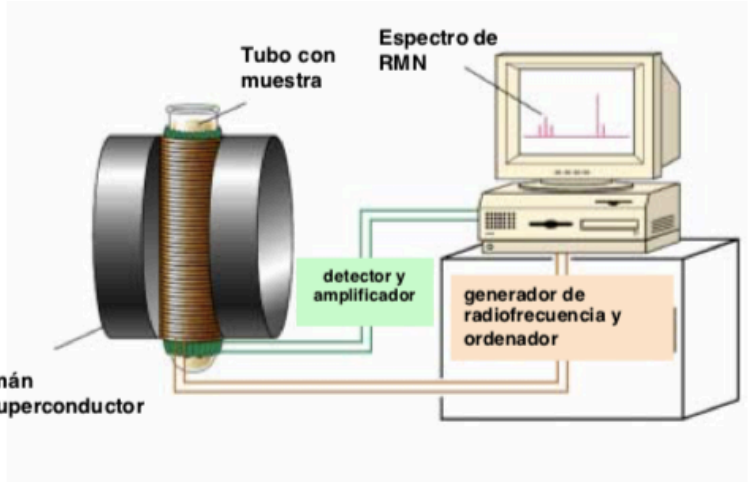

Figura 7[15]. Esquema de los componentes de un sistema NMR

Como se puede observar, consta de cuatro partes: un imán estable, con un controlador que produce un campo magnético preciso; un transmisor de radiofrecuencias, capaz de emitir frecuencias precisas; un detector para medir la absorción de energía de radiofrecuencia de la muestra; y un ordenador y un registrador para realizar las gráficas que constituyen el espectro de NMR.

Para obtener un espectro de NMR, se coloca una pequeña cantidad del compuesto orgánico disuelto en medio mililitro de disolvente en un tubo de vidrio largo que se sitúa dentro del campo magnético del aparato. El tubo con la muestra se hace girar alrededor de su eje vertical.

[15] Fuente: https://www.ehu.eus/documents/1468013/5943652/RMN

En los aparatos modernos el campo magnético se mantiene constante mientras un breve pulso de radiación excita a todos los núcleos simultáneamente. Como el corto pulso de radiofrecuencia cubre un amplio rango de frecuencias los protones individualmente absorben la radiación de frecuencia necesaria para entrar en resonancia (cambiar de estado de *spin*). A medida que dichos núcleos vuelven a su posición inicial emiten una radiación de frecuencia igual a la diferencia de energía entre estados de *spin*. La intensidad de esta frecuencia disminuye con el tiempo a medida que todos los núcleos vuelven a su estado inicial.

Un ordenador recoge la intensidad respecto al tiempo y convierte dichos datos en intensidad respecto a frecuencia, esto es lo que se conoce con el nombre de transformada de Fourier (FT-RMN).

4.1.2. CROMATOGRAFÍA DE GASES ACOPLADA A MS (GC-MS)

Es una técnica que combina la capacidad de separación que presenta la cromatografía de gases con la sensibilidad y capacidad selectiva del detector de masas. Esta combinación permite analizar y cuantificar compuestos trazas en mezclas complejas con un alto grado de efectividad.

Su principio básico es generar iones de compuestos tanto orgánicos como inorgánicos mediante un método adecuado. Posteriormente, estos iones se separan según su relación carga/masa y se detectan de manera cualitativa o cuantitativa. La ionización de la muestra puede producirse por acción de la temperatura, por campos eléctricos, por acción de un reactivo químico o por impacto electrónico.

Como ya se ha comentado, el GC-MS se compone de dos bloques principales: el cromatógrafo de gases y el espectrómetro de masas. El primero utiliza una columna, por la que recorre la muestra a medida que se separan las moléculas, que serán retenidas por la misma columna y luego saldrán en distintos momentos, llamado tiempo de retención, permitiendo que el espectrómetro de masas capture, ionice, acelere, desvíe y detecte las moléculas ionizadas por separado. El espectrómetro de masas hace este procedimiento rompiendo cada molécula en fragmentos ionizados, y detectando estos fragmentos utilizando su relación masa-carga.

La **Figura 8** representa de forma esquemática este procedimiento.

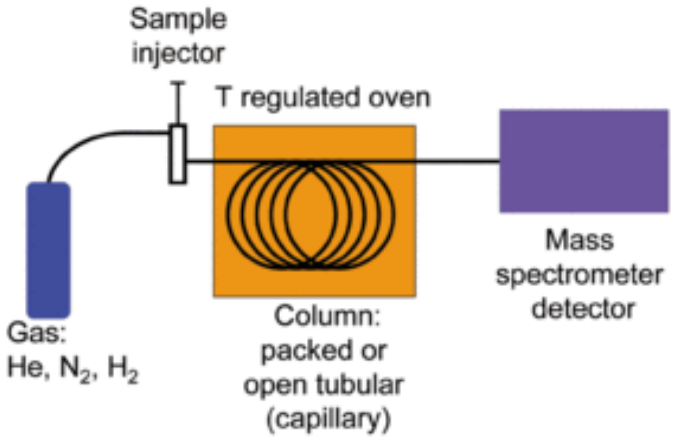

Figura 8[16]. Esquema de un sistema GC-MS.

4.1.3. CROMATOGRAFÍA LÍQUIDA ACOPLADA A MS (LC-MS)

Para hablar de la LC-MS se debe hablar primero de la LC. La LC es un proceso de separación utilizado para aislar los componentes individuales de una mezcla. Esto implica la transferencia de masa de una muestra a través de una fase móvil polar y una fase estacionaria no polar. Cuando la cromatografía líquida es de alto rendimiento se conoce como HPLC (cromatografía líquida de alta presión). La diferencia entre ambas es que el solvente en la LC se desplaza por la fuerza de la gravedad, y en cambio en la HPLC el disolvente viaja a alta presión obtenida por medio de una bomba, lo que reduce el tiempo de separación.

Así pues, la LC-MS es una técnica combinada entre la espectrometría de masas y la cromatografía líquida de alta presión. La combinación de los dos métodos reduce el error experimental y mejora la precisión.

La LC-MS implica la separación de mezclas de acuerdo con sus propiedades físicas y químicas, la identificación de los componentes dentro de cada pico y la detección en función de su espectro de masas.

Como se muestra en la **Figura 9**, los compuestos solubilizados (la fase móvil) se pasan a través de una columna empaquetada con una fase estacionaria (sólida). Esto separa efectivamente los compuestos en función de su peso y afinidad para las fases móviles y estacionarias de la columna, lo cual también conduce a la fragmentación de la muestra y su anionización. Después, la muestra pasa a la cámara de vacío del espectrómetro de masas, donde puede ser analizada y detectada.

[16] Fuente: https://en.wikipedia.org/wiki/Gas_chromatography–mass_spectrometry

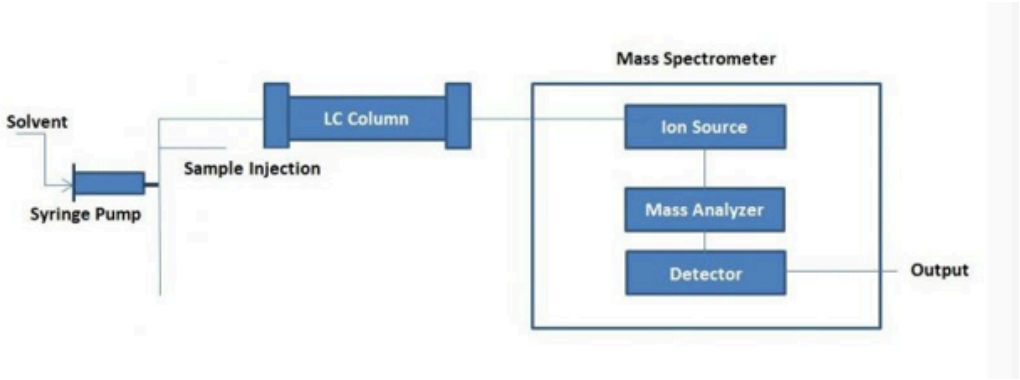

Figura 9[17]. Esquema de un sistema LC-MS.

La LC es la técnica de separación elegida para moléculas más grandes y no volátiles, como las proteínas y los péptidos complejos. Cuando se combina con MS, la LC-MS ofrece una amplia cobertura de la muestra porque se pueden usar diferentes químicas de columna, como la cromatografía líquida de fase inversa.

La LC también es un método ideal para separar isómeros, que tienen la misma masa y, de lo contrario, no se diferenciarán (es decir, se resolverán) mediante un espectrómetro de masas. De hecho, debido a su poder de resolución superior y amplio rango de masa, LC ha reemplazado en gran medida la electroforesis en gel para la separación molecular. Finalmente, la LC ayuda a reducir la supresión de iones, que ocurre cuando las moléculas interactúan unas con otras e impiden el proceso de ionización completa.

[17] Fuente: https://www.chemyx.com/support/knowledge-base/applications/basic-principles-hplc-ms-lc-ms/

II. HERRAMIENTAS PARA EL ANÁLISIS DE DATOS METABOLÓMICOS

Aunque el conocimiento sobre los metabolitos es crucial para la comprensión de la mayoría de los fenómenos celulares, esta información por sí sola no es suficiente para obtener una visión completa de todos los procesos biológicos involucrados. Es por eso por lo que se necesitan enfoques integrados que combinen la metabolómica con la transcriptómica y la proteómica, de forma que se pueda obtener información más profunda que con cualquiera de las técnicas por separado. Aun y poder tener disponible esta información, pero, la integración multinivel de diferentes datos ómicos sigue siendo un desafío, ya que su manejo, procesamiento, análisis e integración requieren herramientas matemáticas, estadísticas y bioinformáticas especializadas, y existen varios problemas técnicos que dificultan el proceso en el campo. Es por eso por lo que se han propuesto distintas herramientas para el análisis de datos metabolómicos y la integración con otros datos ómicos; algunas de ellas se estudiaran en este capítulo.

Es importante remarcar que las capacidades de software requeridas para estas herramientas incluyen el procesamiento de datos espectrales sin procesar, el análisis estadístico para encontrar metabolitos significativamente expresados, la conexión a bases de datos de metabolitos para su identificación, la integración y análisis de múltiples datos ómicos heterogéneos y el análisis bioinformático y visualización de redes de interacción molecular, pero no todas ellas proporcionan la misma información, así que también trataremos de compararlas para obtener una visión más completa de las distintas herramientas y sus funcionalidades.

Antes de desarrollar este capítulo, es importante recordar que los datos producidos requieren un manejo distinto según el método analítico que se ha utilizado (es decir, se requieren diferentes herramientas de análisis de datos y flujos de trabajo); las tecnologías más usadas en el estudio de datos metabolómicos son la cromatografía de líquidos – espectrometría de masas (LC-MS), la cromatografía de gases – espectrometría de masas (GC-MS) y la resonancia magnética nuclear (NMR), que ya se han visto anteriormente.

Las herramientas que se proponen para cada función deben ser intuitivas, fáciles de usar e idealmente de código abierto. Aun así, el hecho de que sean de código abierto hace que puedan contener errores, aunque son superficiales y pueden solucionarse sin demasiados problemas. Por otro lado, el software comercial de código cerrado puede tener la ventaja de ser fácil de usar, estar bien probado y documentado, y de poder adaptarse a usuarios individuales; a pesar de eso, consideramos que es mejor que sean

de código abierto porque en los de código cerrado hay una falta de transparencia en la forma precisa en que se realiza el análisis y además pueden ser muy caros. Es por eso por lo que en este apartado se considerarán aquellas herramientas de uso libre (es decir, de código abierto y gratuitas).

A continuación, se presentarán aquellas herramientas más populares a la hora de analizar datos metabolómicos, y seguidamente se profundizará en cada una de las partes de este análisis para determinar cuáles de ellas se pueden considerar en cada una de las fases. Para cerrar el capítulo, se hará un resumen de los distintos paquetes de R disponibles para este tipo de análisis.

1. Herramientas de análisis más populares

Hay una gran cantidad de software específicamente diseñados para el análisis de datos metabolómicos; de hecho, hay disponibles aproximadamente doscientas herramientas; las más utilizadas por los investigadores que trabajan con datos metabolómicos son *MetaCore* y *MetaboAnalyst*, ambas implementadas a principios de los años 2000, pero más recientemente aparecieron dos herramientas más, *InCroMAP* y *3Omics*, que, a pesar de su facilidad de uso, no han llegado a superar a las anteriores. Por otro lado, también se dispone de otras herramientas bastante populares, como *XCMS* y *XCMS Online*, *Haystack*, *MZMine2*, o *MET-IDEA*. A continuación, se hará una breve descripción de cada una de ellas.

MetaCore es una herramienta comercial disponible como una aplicación independiente y basada en la web; es un software utilizado para el análisis funcional de diferentes tipos de datos moleculares de alto rendimiento. Se trata de un sistema integrado, consistente en una base de datos de alta calidad, herramientas de análisis genómico, herramientas de minería de datos, y herramientas para el ensamblaje personalizado de redes funcionales, y un conjunto de analizadores para cargar y manipular diferentes tipos de datos moleculares de alto rendimiento.

MetaboAnalyst es una plataforma basada en web de libre acceso integrada. Ofrece un conjunto de herramientas en línea para el análisis de datos metabolómicos que combinan el análisis estadístico con la interpretación y visualización de los datos, con gráficos interactivos (aunque con opciones limitadas para su realización), tanto funcional como biológica; de hecho, dispone de una amplia variedad de pruebas estadísticas. Sus módulos funcionales (ocho en total) se pueden agrupar en tres categorías: análisis estadístico exploratorio, análisis funcional y métodos avanzados para estudios de traslación. Por lo tanto, y dependiendo del tipo de datos que se analizan, también se puede utilizar para el análisis de biomarcadores, análisis de

enriquecimiento y análisis de rutas, entre otros. Uno de sus inconvenientes, pero, es el hecho de que se basa en datos preprocesados.

InCroMAP es un software Java independiente; es una herramienta adecuada para estudios completos de biología de sistemas, ya que admite datos metabolómicos anotados. El software realiza análisis de enriquecimiento de conjuntos de metabolitos y genera mapas globales interactivos del metabolismo celular.

3Omics es una herramienta basada en web, independiente de la plataforma, y desarrollada para el análisis, la integración y la visualización de datos humanos transcriptómicos, proteómicos y metabolómicos. Admite el análisis de correlación, el perfil de coexpresión, el mapeo de fenotipos, el análisis de enriquecimiento de rutas y el análisis de enriquecimiento basado en ontología genética.

XCMS es un software basado en R que se utiliza para el procesamiento de datos LCMS. Utiliza la corrección de tiempo de retención no lineal, la filtración combinada, la detección de picos y la concordancia de picos para extraer información relevante de este tipo de datos en bruto. Este software se puede combinar con funciones y paquetes de R para procesar de manera completa los datos. Entre sus ventajas están que los parámetros son ajustables, y el flujo de trabajo está optimizado; aun así, requiere conocimientos del lenguaje R y está basado en línea de comando.

XCMS Online es una versión en web del XCMS que ofrece muchas de las ventajas del paquete de R sin el uso de un entorno basado en línea de comandos (uno de los inconvenientes de XCMS). Permite un control limitado sobre los parámetros y proporciona gráficos interactivos de análisis univariados y multivariados, así como almacenar datos en la nube y compartirlos, pero no es tan personalizable como la versión en R.

Haystack es una herramienta de procesamiento basada en servidor web que utiliza depósitos masivos para filtrar y extraer información de datos LCMS en bruto, cosa que hace que los valores faltantes debido a picos redundantes o faltantes estén ausentes de los datos procesados (de forma que no se tienen valores “cero”). Proporciona herramientas gráficas para visualizar datos, ya sean sin procesar como procesados, e incorpora algunas herramientas de análisis estadístico exploratorio. Es importante remarcar que, aunque no depende de la calidad de la cromatografía y es imparcial, sus gráficos no son personalizables, y no se tiene en cuenta el tiempo máximo de retención.

MZmine2 es una plataforma basada en Java que permite el procesamiento flexible de datos de MS a través de una interfaz gráfica fácil de usar y parámetros personalizables para el procesamiento y visualización de datos. A pesar de eso, y de estar basado en un proyecto de lotes, las opciones para customizar los gráficos son limitadas, y además tiene numerosas opciones que pueden ser abrumadoras para el usuario.

Por último, *MET-IDEA* (*Metabolomics Ion-Based Data Extraction Algorithm*) es un programa de procesamiento de datos metabolómicos a gran escala que generalmente se usa para datos GCMS, aunque puede usarse también para los datos LCMS. Realiza la alineación de picos, la anotación y la integración de datos de espectrometría de masas con guion y permite la visualización de picos integrados junto con sus espectros de masas adjuntos. Funciona bien con bases de datos grandes, y permite hacer la integración de los datos manualmente, pero sus gráficos son de baja calidad.

2. Funcionalidades de algunas herramientas

En este segundo apartado, se clasificarán los softwares en distintas categorías según su funcionalidad principal. Se ha decidido clasificar las distintas fases en preprocesamiento, anotación, postprocesamiento, análisis estadístico y flujos de trabajo, ya que como se verá más adelante, corresponden en cierto modo con los pasos del pipeline que se va a desarrollar. Además, se explicará que hay otras herramientas que no se pueden clasificar en ninguno de los grupos anteriores.

2.1. Preprocesamiento

Un requerimiento de la mayoría de software disponibles es que los datos estén en un formato abierto; así pues, antes de empezar con el preprocesamiento a menudo es conveniente convertirlos a ese tipo de datos.

Las etapas en los distintos pasos a seguir pueden ser distintas según la tecnología usada. En este caso, las etapas iniciales del preprocesamiento son similares para la metabolómica LC-MS y GC-MS; en general, el proceso consiste en la selección de picos, la deconvolución (es decir, aquellas operaciones matemáticas que se emplean para recuperar datos que han sido degradados por un proceso físico), la coincidencia de picos y la alineación de picos en todas las muestras. Esta alineación de picos sí que varía un poco entre LC-MS y GC-MS: mientras en LC-MS se realiza utilizando los tiempos de retención, en la segunda tecnología los tiempos de retención generalmente se convierten en índices de retención independientes del instrumento. En cuanto a la metabolómica de la RMN sí que difiere bastante de las otras dos, a

pesar de que algunas etapas son las mismas que para la MS, aunque los algoritmos precisos utilizados pueden variar.

En cuanto al preprocesamiento para LC-MS, muchas de las herramientas establecidas se implementan como paquetes R, incluido *XCMS*, que se ha descrito anteriormente. LC-MS es la técnica analítica más utilizada en metabolómica, y es por eso por lo que se ha desarrollado un número mucho mayor de software para ese tipo de datos. Entre ellos se encuentra, como se ha comentado, *XCMS*, pero también *XCMS Online*, *MZmine2* y *MetaboAnalyst*, a los que nos hemos centrado en el apartado anterior. Aun así, hay tres softwares más disponibles de forma gratuita diseñado específicamente para el preprocesamiento de datos: *OpenMS*, *MetAlign* y *MS-DIAL*. El primero es una biblioteca para el análisis de datos LC-MS, e incorpora la selección de picos, el filtrado de ruido, la alineación del tiempo de retención y la cuantificación e identificación de metabolitos. Por otro lado, *MetAlign* también proporciona una serie de funciones de preprocesamiento similares a los del *OpenMS*, aunque incluye la corrección de línea de base y el llenado de valores faltantes. Por último, *MS-DIAL* (*Mass Spectrometry-Data Independent AnaLysis*) proporciona la deconvolución de datos de MS y MS de adquisición independiente de datos no dirigidos.

La GC-MS es una técnica analítica menos utilizada para la metabolómica que la LC-MS ya que se pueden detectar muchos menos analitos, lo cierto es que es una técnica más robusta y reproducible con bibliotecas y bases de datos establecidas para la identificación de metabolitos. La herramienta más ampliamente disponible y gratuita en bases de datos amplias es *AMDIS* (*Automated Mass Spectral Deconvolution and Identification System*), y se puede aplicar a todos los datos de GC-MS, aunque no incluye la alineación espectral, así que se debe usar un software adicional. Por otro lado, y a pesar de estar diseñado para el análisis LC-MS y no tener funciones específicas para GC-MS, también se utiliza mucho XCMS. Otros software ampliamente utilizados son *MetaboliteDetector*, un software para el procesamiento especifico de la cromatografía de gases y la espectrometría de masas que incorpora corrección de línea de base, suavizado, detección de picos y deconvolución; *MET-IDEA*, que se ha explicado en el apartado anterior y está diseñado para tomar la salida de *AMDIS* como entrada y puede cuantificar los resultados; *MeltDB*, el cual proporciona un conjunto de herramientas modulares e incluye una serie de algoritmos para la selección de picos; y *metaMS*, basado en *CAMERA* y *XCMS* pero adaptado para este tipo de análisis, entre otros.

Finalmente, para el análisis de datos NMR no se han desarrollado tantos softwares de código abierto; el software *TopSpin* es la más utilizada para el procesamiento previo de

datos metabolómicos de NMR. Gran parte del software de uso gratuito disponible está escrito en MATLAB, pero la encuesta de *Metabolomics Society* informó de un software abierto para el preprocesamiento de NMR, rNMR, que utiliza un enfoque basado en regiones de interés para el análisis de los espectros en una y dos dimensiones.

2.2. Anotación

Se debe tener en cuenta que la identificación de metabolitos es la etapa más costosa en cuanto al tiempo; concretamente, es complicado identificar las características de LC-MS, del que sólo se puede obtener información estructural limitada de la espectrometría de masas. Una solución, aunque parcial, es el análisis de GC-MS, donde se pueden usar extensas bibliotecas comerciales para la identificación; así pues, no hay herramientas diseñadas específicamente para la identificación de metabolitos GC-MS, sino que se tratarán las categorías de MS y NMR, dividiendo así las herramientas entre las que sirven para la espectrometría de masas y las que sirven para la identificación y cuantificación de metabolitos de NMR.

Para la espectrometría de masas la herramienta más utilizada es *CAMARA*, que puede interactuar directamente con XCMS y acceder directamente a la base de datos de *MZedDB* desde R. Otros softwares utilizados son *Rdiop, SIRUS, MI-PACK,* y *PUTMEDID-LCMS*. Además, hay otros paquetes de R para esta función: *MetAssign*, que utiliza el agrupamiento bayesiano para asignar probabilidades posteriores a la probabilidad de la anotación, y *ProbMetab*, que calcula la probabilidad de la asignación de cada compuesto a la característica objetivo utilizando información bioquímica, de la precisión de masa y, en caso de que esté disponible, del patrón de carbono isotópico. Por último, hay ciertos softwares que realizan una comparación automática de bases de datos, lo que permite buscar múltiples bases de datos de MS/MS simultáneamente; se trata de *FingerID, MAGMa* y *MEtFrag*.

En cuanto a la identificación y cuantificación de metabolitos de NMR, y al igual que con el preprocesamiento, los softwares más usados son los comerciales, siendo *Chenomx NMR Suite* y *AMIX* los más populares. Por otro lado, existen herramientas gratuitas que proporcionan tanto la identificación como la cuantificación de los metabolitos (que a menudo se realizan de forma simultánea); se trata de *BATMAN*, un analizador de metabolitos automatizado bayesiano para espectros de NMR, y *Bayesil*. Otras herramientas alternativas podrían ser *MetaboMiner* y *COLMAR*.

2.3. Postprocesamiento

Los métodos de postprocesamiento, que incluyen el filtrado de datos, la imputación, la normalización, el centrado, el escalado y la transformación, se utilizan antes de cualquier análisis estadístico. Existen tantas técnicas para la imputación, normalización y escalado que no se ha encontrado el método ideal que sea apropiado para todos los datos. Aun así, la mayor parte de las herramientas para esta etapa están disponibles como paquetes de R, sí que tanto el postprocesamiento como el análisis estadístico se podrán realizar en un mismo entorno. Algunos de esos paquetes son *batchCorr*, que se utiliza especialmente para los datos LC-MS; *crmn*, que trabaja con datos LC-MS y GC-MS; *KMIDA*, ideal para todos los datos MS; o *metabolomics*, con el que se pueden analizar datos MS y también NMR.

2.4. Análisis estadístico

Las técnicas que se utilizan para el análisis estadístico son apropiadas tanto para datos MS como NMR, ya que están en el mismo formato: matrices de datos. Todos los métodos estadísticos utilizados en esta etapa se pueden encontrar implementados en muchas aplicaciones de software de análisis estadístico general que no están específicamente diseñadas para el análisis metabolómico, estando la mayoría de ellos implementados en R. Aun así, es relevante comentar que hay una técnica estadística adicional exclusiva de la NMR: *STOCSY* (la espectroscopia de correlación estadística).

Algunos de los paquetes de R que se pueden utilizar para llevar a cabo esta etapa son *Ionwinze*, *MetabolAnalyze* o *metabolomics*, entre otros.

2.5. Flujos de trabajo

La diferencia entre los flujos de trabajo y los softwares mencionados es que los primeros proporcionan múltiples herramientas interconectadas que abarcan todas las etapas de análisis (preprocesamiento, anotación y análisis estadístico). Son programas fáciles de usar, y permiten realizar todo el análisis con una sola herramienta; la mayoría de ellos se proporcionan como aplicaciones web y están diseñados principalmente para el análisis de datos LC-MS.

Algunos de estos flujos de trabajo son *Galaxy*, que proporciona una plataforma de flujo de trabajo biológico; *XCMS Online, MZmine2* y *MetaboAnalyst*, que ya se han comentado con profundidad anteriormente; *MAVEN* (*Metabolomics Analysis and Visualisation ENgine*); y *MAIT* (*Metabolite Automatic Identification Toolkit*), que

proporciona una envoltura de *XCMS* y *CAMERA* para un análisis de datos LC-MS fácil de usar.

2.6. Otras herramientas

Como se ha comentado, hay programas que no se pueden clasificar en ninguna de las categorías vistas anteriormente, ya que proporcionan otras funcionalidades. Éstas están relacionadas con la mejora del diseño experimental y la optimización de parámetros. Por lo tanto, hay herramientas que están diseñadas para optimizar la detección de características de los datos de LC-MS (/MS) y otras para estimar el tamaño de muestra requerido para lograr suficiente potencia. Algunas de estas otras herramientas son *MetShot, MetSizeR, MSClust*, y *msPurity*, entre muchas otras.

3. Paquetes de R disponibles para el análisis de datos metabolómicos

Aunque en este trabajo no se va a usar ningún paquete específico, ya que los datos sobre los que se va a trabajar ya están preparados para el análisis y usando algunas funciones y paquetes generales ya se pueden extraer las conclusiones necesarias, es conveniente saber que existen algunos paquetes que están específicamente implementados para el análisis de este tipo de datos. En este apartado se van a comentar algunos de ellos, poniendo énfasis en la explicación de sus funcionalidades.

Por un lado, hay el paquete *omu*. *Omu* es un paquete de R que permite un análisis de bases de datos metabolómicos, además de crear gráficos bastante intuitivos. Además, es un paquete muy sencillo de usar, ya que fue creado pensando en usuarios de R sin experiencia. Una de las funcionalidades más interesantes del paquete es el hecho de poder recopilar los nombres de los genes de la base de datos KEGG (*Kyoto Encyclopedia of Genes and Genomes)* que están asociados con los metabolitos en un conjunto de datos.

Con el paquete se puede hacer un análisis univariante completo, obteniendo los estadísticos univariantes más relevantes o con un test ANOVA, además de realizar distintos gráficos, como gráficos de barras, de pastel o gráficos de volcán. También se puede hacer un análisis de componentes principales, obteniendo un gráfico muy bonito visualmente y sencillo de interpretar.

Por otro lado, el paquete *MetaboDiff* es un paquete para el análisis metabolómico diferencial, que ofrece la exploración de rasgos de muestra en una red de correlación de datos metabolómicos. Dentro del paquete se puede realizar el procesamiento de los datos, empleando el método KNN para la imputación de valores faltantes, además

de combinar el clustering jerárquico y *k-means* para determinar los *outliers*, con la opción de excluir muestras individuales o una agrupación de muestras.

Además, el análisis estadístico se puede hacer mediante este paquete, teniendo disponibles el análisis de componentes principales y la incrustación de vecinos estocásticos distribuidos en t (*tSNE*). También se hacen pruebas univariantes, como la t-Student o la ANOVA, gráficos de volcán, o la identificación y exploración de módulos de correlación metabólica.

Otro paquete para enriquecer los datos metabolómicos es *FELLA*, que utiliza las entradas KEGG para dicho enriquecimiento. Dado un conjunto de compuestos, el paquete sugiere reacciones afectadas, enzimas, módulos y vías que utilizan la propagación de etiquetas en una red de modelos de conocimiento, de forma que la red resultante se puede visualizar y exportar para un posterior análisis.

El paquete *MAIT* (*Metabolite Automatic Identification Toolkit)*, por otro lado, sirve para procesar datos metabolómicos de LC/MS. Contiene funciones para realizar un análisis estadístico de extremo a extremo de los datos de este tipo, haciendo hincapié en la anotación de picos y en el diseño funcional modular de las funciones.

Para el análisis univariante y multivariante de datos metabolómicos, un paquete muy completo es el *muma.* Este paquete proporciona un canal simple paso a paso para los análisis estadísticos de metabolómica, proporcionando herramientas fáciles de usar para todo el proceso de análisis de datos. Adicionalmente, se ha implementado dentro del paquete una sección dedicada a la interpretación de datos metabolómicos, que proporciona técnicas específicas para las asignaciones moleculares y la interpretación bioquímica de los patrones metabólicos.

Otro paquete disponible es el *Metab*, que se utiliza para el procesamiento de alto rendimiento de datos metabolómicos. Este paquete analiza los datos mediante el Sistema Automatizado de Deconvolución e Idetificación Espectral de Masas (AMDIS), además de realizar pruebas de hipótesis estadística y de análisis de varianza (ANOVA).

Como se ha comentado en apartados anteriores, hay algunas plataformas que tienen su versión en R para superar las limitaciones que puedan tener para los usuarios. Estos son, por ejemplo, *MetaboAnalystR* y *XCMS*.

Por un lado, *MetaboAnalystR* consta de más de quinientas funciones organizadas por módulos. El paquete está basado en otros paquetes de R, como *caret* para la clasificación y evaluación del rendimiento o *ROCR* para visualizar el rendimiento de

biomarcadores. Además, utiliza la base de conocimientos de su plataforma madre, *MetaboAnalyst*, que incluye bibliotecas de compuestos, de vías y de conjuntos de metabolitos. El paquete permite hacer análisis estadísticos, de biomarcadores, de series de tiempo, de enriquecimiento de rutas, etcétera, de forma que es un paquete muy completo y funcional.

Por último, *XCMS* se especializa en el procesamiento y visualización de datos espectrales de masas separados por cromatografía y espectros individuales. Dentro del procesamiento incluye el filtrado e identificación de picos (con la función *xcmsSet)*, la agrupación de picos entre muestras (*group*), llena los valores faltantes en los picos (*fillPeaks),* la extracción de resultados estadísticos *(difffreport)* y la visualización de los picos más importantes (*getEIC)*.

III. DESARROLLO DE UN PIPELINE PARA EL ANÁLISIS DE DATOS METABOLÓMICOS

Como dice el propio título del trabajo, el objetivo de este estudio es realizar un *pipeline* para el análisis de datos metabolómicos. Para ello es importante definir los pasos que dicho procedimiento de análisis va a tener. Hay diversas revisiones que han discutido en detalle la variedad de técnicas analíticas y los métodos de recolección y almacenamiento de datos disponibles, pero ningún artículo ha establecido de manera sistemática la forma en que se pueden convertir los datos metabolómicos en conocimiento biológico. En este trabajo, y aunque no todos los estudios sobre datos metabolómicos convergen en la determinación del proceso, se dividirá el procedimiento en cuatro grandes etapas: la preparación de la muestra y adquisición de la base de datos, el preprocesamiento de dichos datos, su análisis y, por último, la interpretación e integración de los resultados. Aun así, antes de entrar en la etapa de preprocesamiento es importante hacer un análisis descriptivo inicial.

En este capítulo se ofrece una descripción detallada de cómo se realiza un análisis metabolómico, explicando las etapas principales, sus aspectos principales, y los pasos que configuran cada una de estas etapas, que van desde el diseño de buenos experimentos hasta métodos de validación para dar conclusiones, pasando por la manipulación de los datos, métodos de preprocesamiento y técnicas analíticas.

En la **Figura 10** se muestra el flujo del pipeline que se va a comentar. A partir de ésta, se detallará cada una de las etapas que participan en este análisis de datos metabolómicos, concretando por cada una de ellas los apartados que se pueden tener en cuenta.

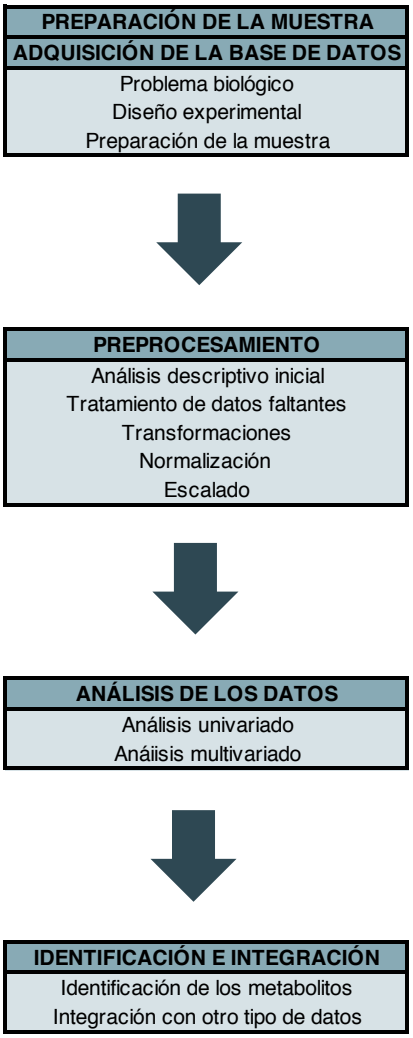

Figura 10[18]. Flujo de un pipeline para el análisis de datos metabolómicos

1. Preparación de la muestra y adquisición de la base de datos

Para empezar cualquier estudio metabolómico, es fundamental formular de forma clara y directa el problema biológico que se va a abordar, paso que regirá el diseño experimental que sigue. En este punto se deben definir una serie de puntos que de alguna manera estarán interrelacionados, como el tipo de enfoque metabólico (dirigido o no dirigido), el tipo de muestra que tendremos (fluidos biológicos, tejidos, células y/u organismos intactos), o el tamaño de la muestra, entre otros. En este diseño experimental, en caso de que el estudio sea de carácter comparativo, se debe definir también cuál es el grupo control y el grupo de prueba. Además de establecer el tipo de variación en las variables independientes que se requieren y cómo optimizarlas

[18] Fuente: elaboración propia

dentro de las limitaciones del material experimental, recursos y tiempo limitados, también se deben considerar cuestiones como la variación biológica, la preparación de la muestra, el número de muestras y réplicas a analizar y de qué tipo son, y el rango y la sensibilidad del método analítico.

Así pues, una vez se ha definido el problema biológico y se han establecido las condiciones experimentales adecuadas, se considera la preparación de la muestra, que está íntimamente relacionada con el tipo de muestra, el enfoque de la metabolómica seleccionado y la plataforma analítica elegida. Mientras que para la metabolómica dirigida el procedimiento de extracción generalmente se optimiza para los metabolitos específicos o las clases químicas de metabolitos en consideración, para la metabolómica no dirigida la preparación de la muestra suele ser mínima, y depende estrictamente del tipo de datos que trabajamos (tejidos, células, gases, etc.).

Otra área relacionada con el diseño experimental es la de la optimización instrumental. En el caso de los espectrómetros de masas (denominado "sintonización") a menudo se asume que se ha realizado satisfactoriamente, cuando en realidad no es así, ya que los experimentadores no pueden probar todas las combinaciones de configuraciones posibles; así pues, se considera apropiado un método heurístico[19], en los cuales se buscan soluciones buenas, pero no óptimas. Para fines de optimización serían adecuados los algoritmos evolutivos.

El último paso de esta etapa es la recopilación de los datos, o adquisición de la muestra. Este paso ha sido un gran desafío analítico, ya que hay una gran variedad de composiciones químicas dentro de cada muestra biológica, abarcando desde compuestos con distintas propiedades químicas, características estructurales y funcionalidad, hasta niveles de concentración discrepantes; de hecho, ninguna plataforma analítica única permite identificar y cuantificar completamente todo el conjunto de metabolitos de un sistema biológico, de manera que se deben combinar distintas técnicas analíticas para generar resultados. Las más comunes, que ya se han visto en profundidad en el primer capítulo, son la cromatografía de líquidos – espectrometría de masas (LC-MS), la cromatografía de gases – espectrometría de masas (GC-MS) y la resonancia magnética nuclear (NMR).

Aun así, existen otras técnicas que, a pesar de no ser tan populares, también se utilizan, como la electroforesis capilar acoplada a la espectrometría de masas (CE-MS),

[19] Parte práctica del concepto de heurística, que es cualquier enfoque para la resolución de problemas, aprendizaje o descubrimiento que emplea un método práctico no garantizado para ser óptimo o perfecto, pero suficiente para los objetivos inmediatos.

la espectrometría de masas por infusión directa (DI), la ionización de desorción con láser asistida por matriz acoplada a MS (MALDI-MS) o la imagen de espectrometría de masas MALDI (MALDI-MSI), entre otras.

2. Análisis descriptivo inicial

Antes de entrar en el preprocesamiento, es conveniente obtener un análisis descriptivo inicial para cada una de las variables. De este modo se tendrá una visión de la base de datos y se podrá determinar qué pasos del *preprocesing* es imprescindible seguir y cómo se debe hacer para obtener unas conclusiones del estudio acuradas.

Para empezar, es importante hacer un análisis descriptivo global, obteniendo algunos estadísticos que puedan ser informativos. Los estadísticos básicos que pueden aportar más información y que por lo tanto deberían aparecer en este punto son la media, la mediana, la desviación estándar, y el coeficiente de variación. La media y la moda son medidas de posición o tendencia central, es decir, indican donde está el centro de la distribución de los datos; la primera es el valor medio de las muestras, y la segunda es un valor que deja por debajo el 50% de los datos y el otro 50% por encima. En cambio, la desviación estándar y el coeficiente de variación son medidas de dispersión, es decir, miden la variabilidad de los datos, aunque la primera es una medida de dispersión absoluta y la segunda relativa, cosa que significa que es adimensional y permite comparar la dispersión de variables expresadas en distintas unidades.

En R se puede hacer este análisis descriptivo de forma muy sencilla, ya que dispone de funciones para cada uno de los estadísticos (*mean, median, sd)* excepto para el coeficiente de variación, aunque se puede calcular rápidamente dividiendo la desviación estándar entre la media. Aun así, vale la pena remarcar que, como es probable que se tengan valores faltantes, conviene especificar dentro de la función el parámetro *na.rm=TRUE*, ya que de este modo calculará el estadístico obviando los valores faltantes y la función no devolverá un valor *NA.*

También se debe hacer una descriptiva de los valores faltantes. De hecho, hay dos tipos de valores faltantes: los “reales”, es decir, aquellos que no constan debido a acontecimientos como errores en la transcripción de los datos o fallos en el instrumento de medida, entre otros, y los que son iguales a cero. Estos últimos se deben tener en cuenta porque, aunque puede ser que el valor sea realmente un cero, también es posible que se haya considerado como cero porque el valor real quedaba por debajo del nivel de detección; son los de este último caso los que se deberán tratar como se verá más adelante. Se deben separar porque la imputación se debe hacer de forma distinta en función del tipo de dato faltante que sea. En este punto del análisis

descriptivo es conveniente considerar qué porcentaje de valores faltantes de cada tipo hay en la base de datos. Además, es importante observar el porcentaje de valores faltantes por variable, ya que, si alguna de ellas tiene muchos *missings*, por ejemplo, un porcentaje superior al 20%, se puede considerar eliminarla de la base de datos.

Para tener una base descriptiva visual, a menudo se realizan análisis gráficos, como por ejemplo los boxplots. Lo ideal es graficar boxplots múltiples; de este modo se puede observar si conviene transformar los datos o hacer un escalado. Adicionalmente, se pueden graficar boxplots por cada una de las variables, y hasta separarlas por los grupos que luego se quieran analizar.

3. Técnicas de preprocessing

Independientemente de la técnica analítica utilizada a la hora de recolectar los datos, una vez se dispone de la base de datos y se ha hecho el análisis descriptivo inicial se debe pasar por la etapa de preprocesamiento para convertir los datos sin procesar a una forma adecuada, y además también puede ser conveniente someter estos datos modificados a una reducción de datos para que solo se utilicen las variables de entrada más relevantes en el análisis de datos posterior.

Los pasos incluidos normalmente en esta fase son la imputación de valores faltantes, la transformación, la normalización de los datos y el escalado.

Los valores faltantes, como ya se ha comentado, son celdas vacías donde no se ha asignado ningún valor numérico a un pico de metabolito respectivo; pueden surgir porque están por debajo del límite de detección (en este caso el valor aparecerá en la base de datos como un cero) o porque no se recopilaron (lo que se denomina como un *missing* real). Es importante lidiar con eso porque muchos métodos multivariados requieren una matriz completamente definida o se convierten en ineficientes computacionalmente cuando hay datos incompletos. A pesar de que hay varias estrategias para solucionarlo, como eliminar la variable con valores faltantes que exceden un cierto umbral, normalmente se realiza una imputación de valores faltantes, que consiste en reemplazarlos con un valor pequeño, asumiendo así que la característica en cuestión está por debajo del límite de detección.

En función del tipo de dato faltante con el que se está lidiando (es decir, si es un dato no recopilado o un dato que estaba por debajo del límite de detección) la imputación se realizará de una forma u de otra. Por un lado, si el valor faltante es un dato no recopilado (casos en los que suele aparecer como NA, *Not Avilable*) la forma más usual de imputar el dato es con la imputación por KNN (*K Nearest Neighbours*). En concreto,

en R hay una función llamada *KNNimputation* que llena todos los valores que aparecen como NA en la base de datos utilizando los k vecinos más cercanos de cada caso. Por defecto, utiliza los valores de los vecinos y obtiene un promedio ponderado por la distancia al caso de sus valores para completar los desconocidos. Por otro lado, los casos en los que se trata de valores *missing* iguales a cero, es decir, valores faltantes por nivel de detección, se suele buscar el valor mínimo que no sea cero de la base de datos y se divide por un número, usualmente 2 o la raíz cuadrada de 2, y se reemplazan todos los valores faltantes por nivel de detección con este valor.

Se debe tener en cuenta, pues, que mientras los datos NA sí que se reemplazan todos ellos con valores distintos ya que se consideran los valores más cercanos de cada uno de ellos, los valores faltantes por nivel de detección se reemplazarán todos con el mismo número.

Por otro lado, la transformación ayuda a eliminar la heterocedasticidad[20] de los datos y a corregir una distribución de datos sesgada. Normalmente se calcula el logaritmo natural de los datos, y se hace de forma sencilla con la función de R *log.*

Es importante tener en cuenta que la transformación se debe hacer después de la imputación de los valores faltantes, ya que si se hace antes todos aquellos valores faltantes por nivel de detección aparecerán como NA después de la transformación, ya que el logaritmo de cero no existe.

En cuanto a la normalización, su propósito es reducir la variación sistemática y separar la variación biológica de la no biológica, que es la introducida por el proceso experimental. Aun así, la normalización a una señal total constante (para tener en cuenta tamaños de muestra variables) introduce dependencias entre las variables que sin este paso no existirían. Puede ser muestral, característica o ambas. La normalización inteligente de las muestras hace que estas sean más comparables entre sí.

La normalización inteligente de las funciones implica centrar los datos alrededor de la media combinada con varios tipos de escalado: es aquí donde entra el último paso. En cuanto a la escala, hay diversos enfoques posibles; el más común es la escala de unidad o auto escalado, que consiste en centrar cada variable en la media y luego dividirla por la desviación estándar, de forma que todas las variables se vuelven igualmente importantes. En R se puede utilizar la función *scale,* y hace directamente el centrado y el autoescalado. Aun así, para datos metabolómicos se recomienda utilizar

[20] Variancia no constante.

la escala de Pareto, que usa la raíz cuadrada de la desviación estándar como factor de escalado.

No obstante, hay otros métodos disponibles, como la escala extensa, que usa la desviación estándar y el coeficiente de variación como factor de escalado, o la escala de rango, donde el factor de escalado es el rango.

4. Análisis de los datos

El objetivo de esta etapa es encontrar patrones dentro de los datos que proporcionen información biológica útil que se pueda usar para generar hipótesis que puedan ser verificadas y refinadas.

Hay distintas técnicas que se pueden llevar a cabo en el análisis de los datos; la elección de las que se realizarán se basa en el objetivo del estudio que se esté llevando a cabo.

En cuanto al análisis univariante, lo más común es realizar un ANOVA, que sirve para comparar las medias de dos o más grupos o para estudiar los posibles efectos de los factores sobre la varianza de una variable. Aunque existen distintos tipos de ANOVA, todos parten de la misma hipótesis nula, que es que la media de la variable estudiada es la misma en los diferentes grupos, en contraposición a la hipótesis alternativa de que al menos dos medias difieren de forma significativa. Aunque permite comparar las medias, lo hace mediante el estudio de las varianzas, ya que consiste en comparar la varianza entre las medias extraídas de cada uno de los grupos frente a la varianza promedio dentro de los grupos.

Como ya se ha comentado, existen distintos tipos de ANOVA: la ANOVA entre sujetos, que se usa para datos independientes; la ANOVA de una vía para datos independientes, que se emplea cuando los datos no están pareados y se quiere estudiar si existen diferencias significativas entre las medias de una variable aleatoria continua en los diferentes niveles de otra variable cualitativa o factor; la ANOVA de dos vías para datos independientes, que sirve para estudiar la relación entre una variable dependiente cuantitativa y dos variables independientes cualitativas (factores) cada uno con varios niveles, y que además puede ser aditivo o de interacción dependiendo de si los factores son o no independientes; y la ANOVA con variables dependientes, que considera que los datos son pareados y por lo tanto se usa cuando las variables a comparar son mediciones distintas pero sobre los mismos sujetos.

La elección del tipo de ANOVA, pues, dependerá de los datos que se estén estudiando; no obstante, lo más habitual será encontrar datos que requieren usar una ANOVA de una vía. En R es muy sencillo realizar este análisis univariante: con la función *aov* y especificando la variable cuantitativa y la variable factor se puede obtener un resumen que informa del valor del estadístico F, que es el ratio entre la varianza de las medias de los grupos y el promedio de la varianza dentro de los grupos y sigue una distribución F de Fisher-Snedecor, y del *p-value* del test, que informa sobre si las diferencias entre los grupos son significativas (cuando el *p-value* es inferior a 0,05, con un nivel de significación del 95%) o no lo son, además de otros valores de interés como los grados de libertad u otros relacionados con los residuos.

Para el análisis de datos metabolómicos conviene extraer también el *p-value* ajustado; este ajuste normalmente se hace por FDR (*False Discovery Rate*), que consiste en corregir la inflación del error de tipo I, es decir, la probabilidad de rechazar la hipótesis nula cuando es cierta. Aunque se informe del *p-value* obtenido con el test ANOVA, es más fiable utilizar el ajustado.

Una vez se han detectado diferencias significativas entre grupos cabe la posibilidad de que se quiera saber entre qué grupos hay diferencias más notables; eso se puede hacer mediante las comparaciones dos a dos. El método más recomendado para hacerlas es el *TukeyHSD*, que tiene una función disponible en R con ese mismo comando, y compara dos a dos cada uno de los grupos del factor, informando de cuales de ellas tienen diferencias significativas.

En el caso de que el estudio incluya una comparación entre dos grupos, por ejemplo, un grupo de tratamiento y uno de control, es habitual usar métodos univariantes para obtener un resumen de los datos e identificar variables potencialmente importantes antes de aplicar los métodos multivariantes. La herramienta más utilizada son los diagramas de volcán o *Volcano Plot*, un tipo de gráfico de dispersión que muestra las diferencias de cambio de pliegue y la importancia estadística para cada variable. El registro del cambio de pliegue se traza en el eje X para que los cambios en ambas direcciones (arriba y abajo) aparezcan equidistantes del centro. El eje Y muestra el registro negativo del *p-value* de una prueba *t* de dos muestras. Los puntos de datos que están lejos del origen, es decir, cerca de la parte superior de la parcela y en uno de los extremos, se consideran variables importantes con una relevancia biológica potencialmente alta.

Los gráficos de volcán se pueden realizar des del software R usando distintas funciones, aunque no existe ninguna función que lo haga directamente. Primero se

tiene que buscar la media de cada grupo, y dividirlas para obtener la ratio y hacer el logaritmo de éste, de forma que los cambios en las dos direcciones aparezcan equidistantes del centro. Luego, con la función *t.test* se obtienen los *p-values* correspondientes a un test apareado de dos muestras independientes, aunque es conveniente utilizar una corrección de prueba múltiple al realizar la prueba *t* en múltiples variables, cosa que se puede hacer con la función *p.adjust*. A continuación, se toma el logaritmo en base 10 negativo, con la finalidad que las variables con menor ajuste de los *p-values* aparezcan cerca de la cima del gráfico. Finalmente, los datos se unen en un solo *data frame* que puede ser graficado con el paquete *ggplot2.*

Por otro lado, se grafican de nuevo los boxplots múltiples para comprobar que la distribución de los datos es más acertada después del *preprocessing*, además de aquellos boxplots que se crean más convenientes, habitualmente separándolos por los grupos que se están estudiando. En este caso es conveniente graficar los boxplots de aquellas variables que en el análisis ANOVA han salido más significativas, es decir, con diferencias más significativas entre los grupos, y servirán para ver gráficamente estas diferencias.

Los boxplots, o diagramas de cajas, es una forma estandarizada de mostrar la distribución de una base de datos basada en cinco estadísticos: el mínimo, los tres primeros cuartiles y el máximo. Además, da información sobre los *outliers*, es decir, valores que difieren significativamente del resto. Las partes que componen un boxplot se pueden ver en la **Figura 11**.

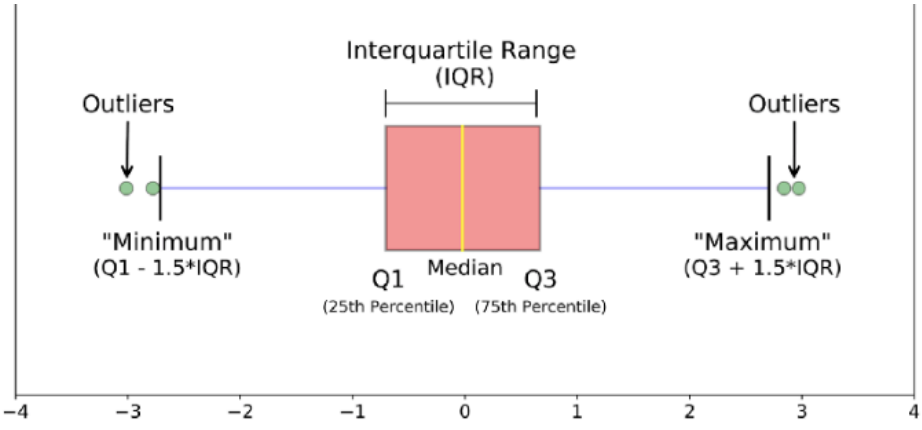

Figura 11[21]. Partes que componen un Boxplot

Como ya se ha comentado, se muestran los valores *outliers* en forma de bolas, que aparecerán siempre fuera de la caja y el bigote (que es la línea que une el máximo y

[21] Fuente: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

mínimo con la caja, y que se muestra en azul). El mínimo aparece en el extremo izquierdo del bigote, y se trata del valor observado más pequeño. Yendo hacia la derecha, aparece el primer cuartil (o percentil 25), que es el número medio entre el mínimo y la mediana; así pues, el 25% de los datos están por debajo de este valor. La mediana, que es la línea amarilla de dentro de la caja, indica el valor que está en la mitad de los datos, es decir, aquel que deja el 50% de datos por debajo y el 50% por encima. Cierra la caja el tercer cuartil, o percentil 75, que deja el 75% de los datos por debajo y es el valor que está en la mitad entre la mediana y el máximo. El máximo aparece en el extremo derecho, y es el valor observado más grande que no se puede determinar cómo outlier.

Aunque en ocasiones el determinar que un valor es *outlier* (o atípico) y no un máximo o un mínimo es una cuestión subjetiva, normalmente se considera que es un valor atípico cuando se encuentra a más de 1,5 veces la distancia del rango intercuartílico (que es la amplitud de la caja) del primer o del tercer cuartil (atípico leve), como se muestra en el gráfico, o a 3 veces esa distancia (atípico extremo).

Así pues, gracias a los boxplots se puede tener información sobre si los datos están bien distribuidos, es decir, la dispersión de los datos, y sobre los valores atípicos que aparecen, pero también sobre la simetría de los datos, así como de qué tan bien se agrupan y si están sesgados (y, en el caso que lo estén, cómo se sesgan). La dispersión de los datos viene dada por el tamaño de la caja; la simetría, por la posición de la mediana dentro de la caja; y el sesgo, por la largada de las colas (bigotes) de los boxplots.

En ocasiones, y dependiendo del objetivo del estudio, es conveniente separar los datos en distintos grupos, para así poder comparar los boxplots de un mismo metabolito entre los grupos realizados.

Los boxplots se pueden graficar de forma sencilla en R gracias a la función *boxplot*, especificando la variable que se quiere estudiar y pudiendo cambiar distintos parámetros como los colores o la dirección de los boxplots (horizontal o vertical). En el caso de que se quieran hacer en función de otra variable, para por ejemplo comparar entre grupos, basta con especificar ambas variables como se haría en un modelo de regresión.

Otro método que se realiza habitualmente es el Análisis de Componentes Principales (PCA). Es un método estadístico que permite simplificar la complejidad de espacios muestrales con muchas dimensiones a la vez que conserva su información. Esto quiere decir que, con un menor número de factores subyacentes, consigue explicar

aproximadamente lo mismo que todas las variables originales. Estos factores son los que se nombran como "componente principal". Pertenece a la familia de los métodos de aprendizaje no supervisado, que son aquellos que tienen como objetivo extraer información empleando variables predictoras para, por ejemplo, identificar subgrupos.

Aunque se puede escoger el número de componentes principales sobre el que se trabaja, y normalmente se hace en base a la inercia acumulada (es decir, la proporción de variancia explicada acumulada), en el caso del análisis de datos metabolómicos se suele trabajar con solo dos componentes principales, ya que un gráfico de estas dos componentes normalmente ya proporciona un buen resumen de las muestras y puede revelar si hay diferencias entre los grupos y detectar *outliers*.

La primera componente principal es aquella cuya dirección refleja o contiene la mayor variabilidad en los datos; así pues, es la que contiene más información. Por otro lado, la segunda componente principal será una combinación lineal de las variables que recoje la segunda dirección con mayor variancia de los datos, pero sin estar correlacionada con la primera. La segunda componente, pues, será perpendicular u ortogonal respecto la primera.

El PCA da dos tipos de información valiosos: una matriz de *scores* y otra de *loadings*. La primera contiene las coordenadas de las muestras para cada componente y proporciona un resumen de las observaciones en un espacio dimensional más pequeño. Por otro lado, la matriz de *loadings* describe la relación entre las variables medidas por cada componente principal. Así pues, un gráfico de puntos de los dos primeros vectores de *loadings* puede revelar la influencia de variables individuales en el modelo. Las direcciones en ambas matrices coinciden, así que la comparación de los dos gráficos se puede usar para identificar qué variables (*loadings*) son más importantes para separar las distintas muestras (*scores*). En concreto, las direcciones de los *loadings* indican las variables que tienen la mayor influencia en la separación de clases.

Aunque en R hay muchas formas distintas de calcular el PCA, la más conocida es la que usa la función *prcomp*; este método usa la descomposición singular de valores para calcular los *eigenvalues*, es decir, los valores propios. Una vez hecho el PCA, se puede graficar mediante el paquete *ggplot2,* así como presentar aquellos metabolitos que son más importantes para cada una de las componentes buscando el argumento de la función *$rotation*.

Además, puede ser relevante algún otro método de análisis multivariante. Por ejemplo, a menudo se realizan gráficos de calor, o *heatmaps.* Estos son una

herramienta eficaz para mostrar la variación de características entre grupos de muestras, y representan las relaciones entre variables con imágenes en color. En ellos, las filas y las columnas se reordenan para que las variables y/o las muestras con perfiles similares se localicen cerca entre ellas, haciendo los perfiles más visibles. Cada valor de la matriz es representado con un color distinto, de forma que es muy fácil identificar los patrones gráficamente.

Los gráficos de calor utilizan un algoritmo de agrupamiento jerárquico aglomerado para ordenar y mostrar los datos en forma de dendograma, un tipo de ilustración gráfica o diagrama de datos en formato de árbol que organiza los datos en subcategorías que se van dividiendo en otras hasta llegar al nivel deseado.

En R hay cinco funciones que permiten hacer gráficos de calor: *heatmap*, que dibuja gráficos de calor simples; *heatmap.2*, que dibuja un mapa de calor mejorado en comparación con la función anterior; *pheatmap*, del paquete de R con el mismo nombre, que dibuja mapas de calor bonitos y proporciona más control para cambiar su apariencia; *d3heatmap*, que de nuevo pertenece a un paquete con el mismo nombre, que dibuja mapas de calor interactivos y pulsables; y *Heatmap*, del paquete *ComplexHeatmap*, el cual dibuja, anota y organiza mapas de calor complejos y es muy útil para el análisis de datos genómicos.

Además, también se puede realizar alguna técnica de métodos supervisados, ya que como se ha visto, el análisis de componentes principales forma parte de los no supervisados. En los métodos de aprendizaje supervisado, los algoritmos trabajan con datos "etiquetados", intentando encontrar una función que, dadas las variables de entrada, asigne la etiqueta de salida adecuada, es decir, predice el valor de salida. En cambio, el aprendizaje no supervisado se usa para describir la estructura de los datos e intentar encontrar algún tipo de organización que simplifique el análisis, así que en lugar de predecir tiene un carácter exploratorio.

Entre las técnicas de aprendizaje supervisado más usadas se encuentra el *Partial least squares-discriminant analysis* (PLS-DA) y el *Orthogonal projection to latent structures-discriminant analysis* (OPLS-DA), que de hecho es una extensión del primero, como ahora se verá.

El PLS-DA se hace para afinar la separación entre grupos de observaciones, normalmente con una rotación de los componentes del PCA de tal forma que se obtiene una separación máxima entre las clases, con la finalidad de entender qué variables son más responsables de la separación de clases.

Consiste en una regresión por mínimos cuadrados parciales clásica donde la variable respuesta es una variable categórica reemplazada por el conjunto de variables ficticias que describen las categorías y que expresa la pertenencia a una clase en unidades estadísticas. Así pues, PLS-DA no permite otras variables de respuesta que la de definir los grupos de individuos. Como consecuencia, todas las variables medidas desempeñan el mismo papel con respecto a la asignación de clase. En realidad, los componentes PLS se construyen al tratar de encontrar un compromiso entre dos fines: describir el conjunto de variables explicativas y la predicción de las respuestas. Aun así, los métodos supervisados como este tienden a sobre ajustar los modelos, de forma que este tipo de métodos deben revisarse y validarse, garantizando así que las clasificaciones de variables son realmente informativas.

En R existen distintos paquetes con los que se puede realizar esta técnica; *pls, plsdepot* y *muma* son algunos de ellos. Normalmente se extraen las funciones discriminantes para visualizar aquellos metabolitos que son más relevantes a la hora de discriminar entre grupos, que de hecho tendrían que coincidir con aquellos que están más cerca del círculo graficado.

Además, es conveniente calcular la calidad del método, es decir, el porcentaje de muestras clasificadas en el grupo correcto, que de hecho es la validación del modelo. Esto se puede hacer de forma sencilla con el paquete *carat*, con el que se pueden hacer una serie de predicciones y proporciona la matriz de confusión, que identifica en qué grupo situa la predicción cada muestra y de qué grupo forma parte realmente, y la exactitud de la precisión, es decir, el porcentaje de muestras identificadas correctamente, entre otros.

Por otro lado, OPLS-DA busca maximizar la varianza explicada entre grupos en una sola dimensión o la primera variable latente, es decir, aquella que no se puede observar directamente y que se tiene que deducir mediante un modelo matemático de otras variables que sí que son observables o medibles, y separar la varianza dentro del grupo en variables latentes ortogonales. Las cargas (*loadings*) de las variables y/o las ponderaciones de los coeficientes de un modelo OPLS-DA validado se pueden utilizar para clasificar todas las variables y discriminar así entre grupos. De nuevo, la validación del modelo es un requisito crítico.

El OPLS-DA se usa normalmente cuando solo hay dos clases a comparar. Se considera mejor que el PLS-DA porque discrimina mejor entre clases y es más robusto en cuando la identificación de características importantes.

Dentro del paquete *muma* en R se puede realizar esta técnica con dos funciones simples: *explore.data* para leer la base de datos y *oplsda* para obtener los resultados numéricos y gráficos, que se guardarán en una carpeta dentro del directorio de trabajo.

Es importante reiterar que el uso de una u otra técnica dependerá del objetivo del estudio y de los datos de que se disponga, y que está en manos del investigador decidir cuáles de ellas son más convenientes en cada caso, si bien aquí se han descrito las más importantes y, en cierto modo, se ha dado una referencia sobre cómo aplicarlas en R.

Adicionalmente, se puede calcular la correlación entre variables (metabolitos) y/o entre grupos de estudio, para comprobar cuáles son los que tienen más correlación. Existen diferentes técnicas para hacerlo, desde el cálculo de la matriz de correlaciones con su *p-value* asociado para saber cuáles de ellas son significativas hasta métodos gráficos más complejos que pueden, por ejemplo, representar en un gráfico la correlación de las dos variables y el valor de dicha correlación, marcando con asteriscos su significación, aunque hay muchas otras opciones. El problema con este tipo de datos es que suelen haber muchas variables, así que a menudo los métodos gráficos no son del todo representativos, porque aparecen demasiado pequeños y no se visualizan bien. Así pues, en muchas ocasiones la mejor opción será el cálculo simple de las correlaciones, que ya aportan información para saber cuáles son los metabolitos o los grupos que se relacionan entre sí.

Aun así, en el cálculo de correlaciones también es importante investigar cuáles de ellas se pueden considerar estadísticamente diferentes de cero. Esto se puede hacer mediante un contraste de hipótesis, donde la hipótesis nula será que no existe relación entre los dos coeficientes (y, por lo tanto, que el coeficiente no se puede considerar distinto de cero), de forma que se concluirá que hay relación si se rechaza la hipótesis nula. En R, se puede realizar el contraste entre dos variables con la función *corr.test*, aunque si se le añade el parámetro *use = "complete"* se obtiene una matriz completa con todos los *p-value*. Se puede escoger también el método que se usará para encontrar el nivel de significancia; aunque se puede escoger el tipo de correlación de Pearson, de Spearman o de Kendall, para datos metabolómicos normalmente se preferirá la correlación de Spearman.

5. Interpretación e integración de los datos

Esta última etapa, como su propio nombre indica, se compone de dos partes: la interpretación y la integración de los datos.

Por un lado, la identificación, en rasgos generales, consiste en buscar el tipo químico de los metabolitos que sirven para describir cada uno de los grupos y dar información relacionada con los elementos químicos encontrados. Esta parte se puede hacer buscando la información en PubChem, ya que cada metabolito tiene su número identificador (PubChemID) de forma que la búsqueda es sencilla y la información proporcionada muy completa.

Explicado de forma más técnica, los metabolitos seleccionados están vinculados al contexto biológico en estudio a través del enriquecimiento, que apunta a investigar la sobreexpresión o subexpresión de grupos predefinidos de metabolitos relacionados funcionalmente para identificar cambios significativos y coordinados de expresión entre ellos, y del análisis de ruta, que implica la descripción y visualización de las interacciones entre genes, proteínas o metabolitos dentro de células, tejidos u órganos con el objetivo de identificar las vías que impactan significativamente en un proceso biológico. Por lo tanto, con la identificación se busca encontrar esos cambios de expresión entre metabolitos e identificar las vías metabolómicas.

Por otro lado, la integración consiste en integrar la información proporcionada por las vías metabólicas con datos proteómicos y transcriptómicos para obtener una visión completa de todos los procesos biológicos involucrados. Además, y para alcanzar una evaluación más confiable del proceso en estudio, también se puede realizar la integración con el conocimiento biológico derivado de la literatura o de datos experimentales anteriores.

La integración puede realizarse con un Análisis de Factores Múltiples, que se puede hacer con la función *MFA* de R, que está incluida en el paquete *FactoMineR*. Esta técnica consiste en estudiar simultáneamente varios conjuntos de variables, ya sean continuas o categóricas, y es útil en muchos campos donde las variables se estructuran en grupos, como la genómica o los análisis sensoriales. Teniendo en cuenta la estructura de los datos, la función permite equilibrar la influencia de cada grupo de variables, estudiar los enlaces entre estos conjuntos, y dar los gráficos clásicos y los específicos, con una representación parcial (individuos vistos por un grupo de variables) o por grupos de variables.

IV. VERIFICACIÓN DEL *PIPELINE* CON DATOS REALES

Para la verificación del desarrollo del pipeline se ha seleccionado una base de datos que contiene información sobre metabolitos obtenidos a partir de 102 muestras de 17 variedades de vinos blancos. Para entender mejor los datos que se van a trabajar, se va a hacer una descripción del contenido de la base de datos, así como una explicación de ciertos conceptos teóricos que pueden ayudar a comprenderla.

Los vinos blancos escogidos para el estudio representan cinco variedades, tres añadas y múltiples regiones de vinificación distintas. Las variedades de uva recogidas en el estudio son Chardonnay, Pinot gris, Riesling, Sauvignon blanc y Viognier. A continuación, se describirán brevemente las características de cada una de ellas.

El Chardonnay es probablemente la variedad más conocida de uva blanca, y también una de las de un cultivo más ampliamente extendido por su buena adaptación climática. Es una variedad originaria de Francia (concretamente de la región de la Borgoña). Su brotación es muy precoz y tiene un periodo corto de maduración. Su vinificación produce un mosto suave y aromático, cosa que hace que se utilice mayoritariamente para la elaboración de vinos espumosos (tanto en cava como en champagne). Además, los vinos blancos monovarietales de la variedad Chardonnay son muy apreciados por su elegancia y refinamiento.

El Pinot gris se considera originario del noreste de Francia, aunque se extendió a Suiza, a Alemania, a Italia, a Rumanía y a Hungría. El apelativo “gris” se debe al color azulado de la baya. Se distinguen por una fragancia floral, siendo cremosos, especiados y con matices cítricos. El cuerpo del vino es distinto en función de su madurez y vinificación, pudiendo ser ligeros o completos, dulces o secos.

El Riesling, originario de Alemania, es una variedad de climas fríos. Ofrece vinos de aromas frescos muy distinguibles cuando todavía son jóvenes, y con el envejecimiento (de hasta los diez años) consiguen tener un aroma sublime.

El Sauvignon Blanc es originario de Burdeos. Produce unos vinos elegantes, secos y ácidos; sus aromas se caracterizan por marcados aromas herbáceos.

Finalmente, el origen de la variedad Viognier es incierto; sus vinos pueden alcanzar un buen grado alcohólico y presentan una acidez baja. Ofrece aromas frutales y florales de gran intensidad, elegancia y complejidad. En boca, los vinos son untuosos, y con un largo final.

Además, la base de datos caracteriza dos vinos blancos concretos: el Elevage Blanc y el Fume Blanc. El primero es una mezcla de Sauvignon Blanco y Sauvignon Gris, mientras el segundo es un vino de autor creado a partir de uvas de la variedad de Sauvignon Blanc. Para este estudio se identificarán ambos como Sauvignon Blanc, a pesar de que en el Elevage también se encuentra Sauvignon Gris.

Como se ha comentado anteriormente, los vinos son de tres añadas distintas: 2001, 2003 y 2004. De hecho, hay tan sólo un vino que sea del año 2001, y cinco que sean del 2003, mientras que once de ellos son del 2004. Por otro lado, la mayor parte de los vinos estudiados son de California, aunque de distintas regiones: Carneros, Napa Valley, Monterey, Lake Country y Dunningan Hills (Yolo Country). Aun así, también están representados el suroeste de Australia, Oregón y los Finger Lakes (en New York). La **Tabla 2** muestra la clasificación de los distintos vinos del estudio que se acaba de comentar.

		Identificador	Año	Región
Variedad de uva	Chardonnay	**CH01**	2003	Carneros, CA
		CH02	2003	Carneros, CA
		CH03	2004	Carneros, CA
		CH04	2003	Monterey, CA
		CH05	2003	Napa Valley, CA
		CH06	2004	Sudeste Australia
	Pinot Gris	**PG01**	2003	Oregon
		PG02	2004	Napa Valley, CA
	Riesling	**R01**	2004	Napa Valley, CA
		R02	2004	Finger Lakes, NY
	Sauvignon Blanc	**SB01**	2001	Lake Country, CA
		SB02	2004	Napa Valley, CA
		SB03	2004	Napa Valley, CA
		SB04	2004	Napa Valley, CA
		SB05	2004	Napa Valley, CA
	Viognier	**V01**	2004	Dunningan Hills, CA
		V02	2004	Napa Valley, CA

Tabla 2[22]. Clasificación de los vinos disponibles en la base de datos

El objetivo que se persigue en el desarrollo de este *pipeline* es observar en qué se diferencian las distintas variedades des del punto de vista metabolómico, y poder identificarlas en función de los metabolomas que componen el vino.

[22] Fuente: elaboración propia.

Se debe remarcar que en este estudio ya se ha hecho el diseño experimental y se han puesto los datos en una base de datos, de modo que están listos para el análisis. Así pues, no se puede hacer la primera parte de un *pipeline*, que sería, como ya se ha comentado, la extracción de la muestra, entre otros. Tampoco se va a poder hacer la representación gráfica de la espectrometría, que sería el punto de partida de todo análisis de datos metabolómicos.

1. Análisis descriptivo inicial

Como se ha visto anteriormente, el primer paso del *pipeline* cuando ya se tiene una base de datos sobre la que trabajar es hacer un análisis descriptivo inicial; así pues, este se hará antes de cualquier procesamiento de los datos. Esta etapa consta de distintas fases: el cálculo de los principales estadísticos, el estudio de valores faltantes, y una representación de las variedades gráfica, en este caso en forma de boxplots, tanto múltiples como por algunos de los metabolitos.

1.1. Principales estadísticos

Antes de empezar cualquier estudio, es importante hacer un análisis descriptivo basado en el cálculo de los principales estadísticos relevantes. Como ya se ha visto en el anterior capítulo, en los estudios de metabolómica, los estadísticos que aportan más información son la media, la mediana, la desviación estándar y el coeficiente de variación.

Metabolito	Media	Mediana	SE	CV
X1.2.anhydro.myo.inositol	13160,4	11815,0	5554,2	0,4
X1.hexadecanol	2627,5	2369,0	1220,5	0,5
X2.hydroxyglutaric.acid	90153,8	77671,5	43451,5	0,5
X2.isopropylmalic.acid	5509,0	4742,5	2933,4	0,5
X2.ketoisocaproic.acid	684,4	654,5	235,6	0,3
X226091.0	1354,8	1248,0	507,7	0,4
X3.6.anhydrogalactose	3109,0	2788,0	1634,9	0,5
X3.hydroxy.3.methylglutaric.acid	2614,6	2374,5	1187,4	0,5
X4.hydroxyproline	12208,9	11166,0	4554,4	0,4
alanine	236113,4	206682,0	106044,9	0,4
alpha.ketoglutaric.acid	2956,8	2126,5	2342,5	0,8
arabinose	65277,7	49343,5	39528,4	0,6
arabitol	5257,6	4762,0	2146,5	0,4
arachidic.acid	1861,3	1723,0	600,7	0,3
asparagine	12638,5	10214,0	7527,6	0,6

Metabolito	Media	Mediana	SE	CV
isoleucine	19537,9	17785,5	8021,9	0,4
isothreonic.acid	4182,7	3588,0	2122,4	0,5
lactic.acid	699732,4	300066,0	568932,0	0,8
lauric.acid	9642,1	8526,5	6082,0	0,6
leucine	67102,2	65804,0	23988,1	0,4
lysine	46864,2	46361,5	17099,2	0,4
lyxitol	4855,7	4420,0	3424,6	0,7
malate	227369,3	255470,0	172884,4	0,8
maleic.acid	164784,1	123076,0	93079,8	0,6
maltose	9351,9	6187,5	9088,4	1,0
mannitol	183297,6	147835,0	127109,5	0,7
melibiose	8469,6	7015,5	7804,5	0,9
methionine	4651,1	4550,0	1741,6	0,4
N.acetyl.D.hexosamine	12668,6	11949,0	4434,9	0,4
N.acetyl.D.mannosamine	12192,1	11704,5	3054,6	0,3

aspartic.acid	30506,2	28347,5	9467,8	0,3
behenic.acid	2535,3	2452,5	848,9	0,3
benzoic.acid	6758,1	5447,0	4153,4	0,6
beta.alanine	8845,9	8760,0	4626,1	0,5
butyl.stearate	4426,9	2898,0	4430,4	1,0
caffeic.acid	2300,5	1921,0	1887,9	0,8
cellobiose	23741,2	15360,0	23496,5	1,0
cis.caffeic.acid	1530,7	1097,5	1122,0	0,7
citramalate	20315,1	18443,5	7618,1	0,4
citric.acid	106777,5	111466,0	65986,3	0,6
citrulline	7968,0	7711,0	5796,6	0,7
conduritol.beta.epoxide	46560,5	45542,0	19349,2	0,4
cysteine	2131,1	1398,0	2469,3	1,2
dehydroascorbate	13714,2	11182,0	9075,5	0,7
erythritol	88431,8	83640,5	35295,9	0,4
fructose	263528,5	218265,0	158831,1	0,6
fucose	16076,2	14757,0	6882,1	0,4
GABA	261663,2	236487,0	149064,6	0,6
galactinol	2543,2	2153,0	1574,2	0,6
galactonic.acid	1657,5	1346,5	1120,5	0,7
galactose	116782,8	33654,5	175234,9	1,5
galacturonic.acid	146575,1	123217,0	65635,5	0,4
glucoheptulose	3491,0	3359,0	843,4	0,2
gluconic.acid	3382,9	3167,0	2404,1	0,7
glucose	370041,6	256995,0	251371,9	0,7
glucuronic.acid	3451,4	2094,5	4939,4	1,4
glutamic.acid	11757,0	11719,0	3861,1	0,3
glyceric.acid	4470,6	4030,5	2288,7	0,5
glycerol	1025760,0	956914,0	425635,5	0,4
glycerol.3.galactoside.NIST	11168,5	10081,0	5010,6	0,4
glycerol.alpha.phosphate	1690,3	1481,5	819,3	0,5
glycine	71271,1	66209,0	18440,5	0,3
guanine	761,6	671,5	530,9	0,7
heptadecanoic.acid	2531,0	2371,0	762,0	0,3
homoserine	3317,9	3225,0	855,4	0,3
indole.3.lactate	1025,6	831,0	643,8	0,6
inositol.myo.	313796,3	300576,0	116737,0	0,4
inulobiose	4492,1	1251,5	8005,6	1,8
isocitric.acid	6166,5	5748,0	2042,9	0,3
isogalactinol	3784,7	3545,0	1631,6	0,4

nicotianamine	568,1	533,0	168,1	0,3
octadecanol	1072,1	929,0	528,6	0,5
oleic.acid	328,7	285,0	213,3	0,6
ornithine	66292,1	63645,0	48643,5	0,7
oxoproline	189142,6	169596,0	99507,8	0,5
palmitic.acid	12850,9	11852,5	5377,9	0,4
pelargonic.acid	19876,4	15641,5	15892,4	0,8
phenylalanine	17032,0	16458,0	5243,4	0,3
pipecolic.acid	10355,3	10475,5	3530,8	0,3
proline	1692466,9	1434206,0	804220,8	0,5
pseudo.uridine	2005,9	1814,5	1172,0	0,6
putrescine	24535,9	19958,5	18410,4	0,8
quinic.acid	3736,6	2550,0	2831,4	0,8
ribonic.acid	2639,1	2396,5	1093,9	0,4
ribose	7231,6	6940,5	1917,6	0,3
saccharopine	609,7	593,5	252,7	0,4
serine	21769,8	18590,5	9429,5	0,4
shikimic.acid	36445,8	32745,0	20524,7	0,6
sophorose	12608,7	10497,0	7831,5	0,6
sorbitol	101896,2	93268,0	41396,0	0,4
spermidine	1327,2	1132,0	793,0	0,6
stearic.acid	87540,6	70134,5	44583,5	0,5
suberyl.glycine	1882,9	1472,0	1195,2	0,6
succinic.acid	964304,2	881536,5	421844,9	0,4
talose	50506,3	3190,5	170938,2	3,4
tartaric.acid	524453,5	491953,0	167391,7	0,3
threitol	12699,7	9320,0	6851,7	0,5
threonic.acid	9091,0	7408,0	5148,1	0,6
threonine	14120,4	13622,5	5354,4	0,4
trehalose	15885,1	6705,0	21392,0	1,3
tryptophan	1798,5	1726,0	1063,8	0,6
tyrosine	29906,7	28835,5	9260,3	0,3
uracil	10725,1	9570,0	6649,8	0,6
urea	3683,0	2944,0	2607,6	0,7
uridine	200,0	187,0	103,1	0,5
valine	38912,8	37595,5	9914,0	0,3
xanthine	955,3	783,5	593,4	0,6
xylitol	3089,6	3054,5	1238,8	0,4
xylose	24465,1	23280,0	6041,4	0,2

Tabla 3[23]. Estadísticos para cada uno de los metabolitos

La **Tabla 3** muestra los estadísticos comentados anteriormente calculados para cada uno de los metabolitos. Aun así, en este estudio no se quieren estudiar las diferencias entre metabolitos, sino que se quiere observar si un mismo metabolito es distinto en función de las variedades y, por lo tanto, si se pueden perfilar las distintas variedades en función de los metabolitos. Es por eso por lo que, aunque la **Tabla 3** es útil para tener una visión general del comportamiento de los metabolitos, se han calculado los mismos estadísticos para los metabolitos, pero separando por variedades. La **Tabla 4** muestra estos resultados.

Metabolito	Variedad	Media	Mediana	SE	CV
X1.2.anhydro.myo.inositol	Chardonnay	10699,50	10271,50	2010,60	0,19
	Pinot_Gris	11905,00	12085,00	1327,25	0,11
	Riesling	22971,92	20422,00	8679,13	0,38
	Sauvignon_Blanc	11869,87	11537,00	3066,54	0,26
	Viognier	15213,17	12172,00	5728,68	0,38
X1.hexadecanol	Chardonnay	2003,97	1993,00	776,10	0,39
	Pinot_Gris	3675,75	3616,00	1349,43	0,37
	Riesling	2402,92	2430,50	1185,92	0,49
	Sauvignon_Blanc	3064,69	2852,00	1276,44	0,42
	Viognier	2726,83	2674,50	1139,95	0,42
X2.hydroxyglutaric.acid	Chardonnay	61666,39	66494,00	14498,93	0,24
	Pinot_Gris	86266,08	86450,00	17947,75	0,21
	Riesling	31848,17	31619,50	4213,00	0,13
	Sauvignon_Blanc	141591,37	133253,00	28811,08	0,20
	Viognier	109215,17	108915,50	12049,04	0,11
X2.isopropylmalic.acid	Chardonnay	4165,36	3836,50	1179,99	0,28
	Pinot_Gris	4874,42	4961,00	390,42	0,08
	Riesling	2044,17	1927,00	455,74	0,22
	Sauvignon_Blanc	8354,20	9665,50	3451,29	0,41
	Viognier	6526,08	6403,50	405,40	0,06
X2.ketoisocaproic.acid	Chardonnay	510,67	524,50	118,22	0,23
	Pinot_Gris	891,92	886,50	157,89	0,18
	Riesling	657,17	680,50	205,34	0,31
	Sauvignon_Blanc	824,36	807,00	259,09	0,31
	Viognier	698,42	697,00	162,58	0,23
X226091.0	Chardonnay	1066,33	968,50	285,51	0,27

[23] Fuente: elaboración propia.

	Pinot_Gris	1528,92	1466,50	240,78	0,16
	Riesling	1178,58	1036,50	538,67	0,46
	Sauvignon_Blanc	1857,27	1766,00	530,23	0,29
	Viognier	1133,50	1090,00	192,88	0,17
X3.6.anhydrogalactose	Chardonnay	2170,03	1694,50	1258,73	0,58
	Pinot_Gris	3269,92	3297,00	624,36	0,19
	Riesling	6178,67	5912,50	1412,91	0,23
	Sauvignon_Blanc	3197,23	2887,00	977,87	0,31
	Viognier	2474,58	2366,50	970,31	0,39
X3.hydroxy.3.methylglutaric.acid	Chardonnay	1857,83	1945,50	539,09	0,29
	Pinot_Gris	3008,33	3161,00	660,49	0,22
	Riesling	1772,17	1872,50	679,86	0,38
	Sauvignon_Blanc	3110,46	3212,00	1039,08	0,33
	Viognier	4342,17	4182,50	1218,63	0,28
X4.hydroxyproline	Chardonnay	16036,14	16002,50	4136,61	0,26
	Pinot_Gris	9808,50	10066,50	1401,18	0,14
	Riesling	11071,42	10985,50	1857,21	0,17
	Sauvignon_Blanc	9716,79	8973,50	4325,81	0,45
	Viognier	9249,58	9101,00	1189,97	0,13
alanine	Chardonnay	276246,75	269818,50	110724,96	0,40
	Pinot_Gris	244789,75	217804,00	85325,28	0,35
	Riesling	286539,08	280845,00	60244,45	0,21
	Sauvignon_Blanc	206577,30	154330,00	101790,76	0,49
	Viognier	130451,83	125796,50	57734,49	0,44
alpha.ketoglutaric.acid	Chardonnay	2119,42	1995,50	599,20	0,28
	Pinot_Gris	2068,42	2121,00	602,98	0,29
	Riesling	1072,50	1187,50	301,42	0,28
	Sauvignon_Blanc	5250,23	4183,50	3180,22	0,61
	Viognier	2508,00	2340,50	726,07	0,29
arabinose	Chardonnay	45469,14	39595,50	20022,27	0,44
	Pinot_Gris	82105,67	69241,00	23164,53	0,28
	Riesling	63434,17	61450,50	30655,66	0,48
	Sauvignon_Blanc	75951,53	45603,50	59444,51	0,78
	Viognier	83034,42	83217,00	5211,33	0,06
arabitol	Chardonnay	6641,61	6547,00	1642,52	0,25
	Pinot_Gris	4115,17	4008,50	903,86	0,22
	Riesling	3807,42	3492,00	1537,38	0,40
	Sauvignon_Blanc	4248,50	4017,00	1434,26	0,34
	Viognier	6052,33	6037,50	3374,17	0,56
arachidic.acid	Chardonnay	1448,50	1410,00	317,74	0,22
	Pinot_Gris	2305,17	2343,00	515,70	0,22

	Riesling	1537,58	1600,50	365,79	0,24
	Sauvignon_Blanc	2328,81	2292,00	581,63	0,25
	Viognier	1927,25	2057,50	516,06	0,27
asparagine	Chardonnay	9699,72	9477,50	3327,31	0,34
	Pinot_Gris	16627,92	16179,50	7500,33	0,45
	Riesling	10551,08	9823,00	2733,60	0,26
	Sauvignon_Blanc	10530,00	10697,50	1638,44	0,16
	Viognier	23769,92	23597,00	13768,24	0,58
aspartic.acid	Chardonnay	36361,19	34036,00	12574,15	0,35
	Pinot_Gris	30862,92	30552,50	2362,39	0,08
	Riesling	31024,75	30760,50	6751,33	0,22
	Sauvignon_Blanc	24859,17	24434,00	4211,11	0,17
	Viognier	26183,50	26334,50	1446,62	0,06
behenic.acid	Chardonnay	2072,19	1920,00	586,03	0,28
	Pinot_Gris	3113,58	3236,00	997,95	0,32
	Riesling	2052,50	2151,00	497,08	0,24
	Sauvignon_Blanc	2980,96	2672,00	860,49	0,29
	Viognier	2789,42	2909,00	659,93	0,24
benzoic.acid	Chardonnay	3814,08	3359,50	1342,61	0,35
	Pinot_Gris	11773,83	11338,00	6282,64	0,53
	Riesling	6331,67	6639,00	1886,63	0,30
	Sauvignon_Blanc	8862,46	7915,50	3768,99	0,43
	Viognier	6090,67	5708,50	1964,34	0,32
beta.alanine	Chardonnay	13348,72	12685,00	3371,58	0,25
	Pinot_Gris	6539,75	6763,00	1901,55	0,29
	Riesling	3294,50	3097,00	1640,01	0,50
	Sauvignon_Blanc	6864,03	8333,00	3294,40	0,48
	Viognier	8149,67	7937,50	2896,86	0,36
butyl.stearate	Chardonnay	3326,81	2526,00	2815,05	0,85
	Pinot_Gris	3580,25	2672,00	2623,76	0,73
	Riesling	4460,75	2431,50	4524,15	1,01
	Sauvignon_Blanc	6062,66	3602,00	6444,97	1,06
	Viognier	4587,25	5592,50	2883,89	0,63
caffeic.acid	Chardonnay	1890,03	1961,50	685,48	0,36
	Pinot_Gris	4060,75	4052,50	891,99	0,22
	Riesling	3293,58	3416,00	604,73	0,18
	Sauvignon_Blanc	2230,46	1552,50	3088,97	1,38
	Viognier	930,58	925,00	148,83	0,16
cellobiose	Chardonnay	13329,11	12701,50	8260,71	0,62
	Pinot_Gris	18027,00	18080,50	4685,48	0,26
	Riesling	12741,83	12200,00	3908,49	0,31

	Sauvignon_Blanc	49560,70	50250,50	28227,42	0,57
	Viognier	7142,67	3789,50	8549,46	1,20
cis.caffeic.acid	Chardonnay	1179,25	1097,50	578,06	0,49
	Pinot_Gris	3147,50	3140,00	591,51	0,19
	Riesling	2780,00	2668,50	532,07	0,19
	Sauvignon_Blanc	1221,11	969,00	1129,51	0,92
	Viognier	440,92	442,00	57,76	0,13
citramalate	Chardonnay	18174,72	17055,00	5932,94	0,33
	Pinot_Gris	18624,50	18007,00	3722,18	0,20
	Riesling	22348,50	19836,50	11999,89	0,54
	Sauvignon_Blanc	22913,73	20565,00	8832,23	0,39
	Viognier	19896,58	19409,50	4016,54	0,20
citric.acid	Chardonnay	80735,44	87586,00	35366,05	0,44
	Pinot_Gris	137409,50	131930,00	16123,69	0,12
	Riesling	45183,58	42977,00	9464,84	0,21
	Sauvignon_Blanc	120430,72	159437,00	90676,32	0,75
	Viognier	182870,25	186246,00	11634,43	0,06
citrulline	Chardonnay	7107,00	7081,00	5199,90	0,73
	Pinot_Gris	9224,08	8721,50	5781,18	0,63
	Riesling	8911,08	9399,50	3133,06	0,35
	Sauvignon_Blanc	5518,28	2027,00	5937,49	1,08
	Viognier	14271,92	13131,00	4708,05	0,33
conduritol.beta.epoxide	Chardonnay	47376,56	46747,50	6072,79	0,13
	Pinot_Gris	46480,50	45848,00	4537,78	0,10
	Riesling	28757,00	26104,50	17309,51	0,60
	Sauvignon_Blanc	40282,50	39393,00	10338,81	0,26
	Viognier	77690,42	78590,00	34721,87	0,45
cysteine	Chardonnay	1223,94	938,50	1001,70	0,82
	Pinot_Gris	1806,33	1371,00	2176,13	1,20
	Riesling	1142,25	941,50	830,63	0,73
	Sauvignon_Blanc	4199,29	3891,00	3439,82	0,82
	Viognier	2029,83	511,50	2466,53	1,22
dehydroascorbate	Chardonnay	9581,61	9101,50	2780,63	0,29
	Pinot_Gris	16376,17	17288,00	2591,27	0,16
	Riesling	5269,67	5349,00	708,19	0,13
	Sauvignon_Blanc	22084,30	15299,00	12061,38	0,55
	Viognier	10969,00	11073,50	450,34	0,04
erythritol	Chardonnay	84846,19	78987,50	24224,61	0,29
	Pinot_Gris	61588,25	63355,50	6670,34	0,11
	Riesling	50410,25	47471,00	22683,56	0,45
	Sauvignon_Blanc	105755,70	101388,50	34074,70	0,32

	Viognier	120743,67	114664,00	39859,20	0,33
fructose	Chardonnay	222463,97	187655,00	128745,15	0,58
	Pinot_Gris	418726,25	382797,00	176048,24	0,42
	Riesling	459630,25	430861,00	129313,93	0,28
	Sauvignon_Blanc	208722,83	159926,50	117584,44	0,56
	Viognier	172436,92	159590,50	61128,74	0,35
fucose	Chardonnay	13184,81	13749,50	2555,93	0,19
	Pinot_Gris	18001,50	18389,50	2407,65	0,13
	Riesling	12839,25	12542,50	2893,19	0,23
	Sauvignon_Blanc	21436,90	21029,00	9953,92	0,46
	Viognier	12660,00	12706,00	2184,64	0,17
GABA	Chardonnay	229775,72	227281,50	101645,25	0,44
	Pinot_Gris	300543,00	292333,00	126059,70	0,42
	Riesling	173280,33	169470,50	58888,34	0,34
	Sauvignon_Blanc	260428,97	234903,00	207531,11	0,80
	Viognier	409914,08	409245,50	27654,65	0,07
galactinol	Chardonnay	1588,31	1587,00	530,29	0,33
	Pinot_Gris	2323,33	2466,50	742,05	0,32
	Riesling	2691,75	2445,00	1175,02	0,44
	Sauvignon_Blanc	2926,71	2949,00	853,46	0,29
	Viognier	4712,33	3554,50	2921,96	0,62
galactonic.acid	Chardonnay	1863,61	1346,50	1490,63	0,80
	Pinot_Gris	1483,67	1458,00	257,86	0,17
	Riesling	1901,92	1289,00	1641,45	0,86
	Sauvignon_Blanc	1360,07	1196,00	571,63	0,42
	Viognier	1662,25	1700,00	451,67	0,27
galactose	Chardonnay	71400,97	25479,50	128776,49	1,80
	Pinot_Gris	232509,42	55678,00	247812,38	1,07
	Riesling	334013,17	335830,00	202569,48	0,61
	Sauvignon_Blanc	45639,57	27777,50	91584,65	2,01
	Viognier	97829,50	31693,00	132664,69	1,36
galacturonic.acid	Chardonnay	124636,97	117714,50	37757,97	0,30
	Pinot_Gris	150506,17	148545,00	23916,39	0,16
	Riesling	233705,67	223080,50	38253,31	0,16
	Sauvignon_Blanc	152485,17	108166,00	89148,26	0,58
	Viognier	106552,42	104050,00	17752,61	0,17
glucoheptulose	Chardonnay	3387,50	3305,00	654,58	0,19
	Pinot_Gris	4360,83	4358,50	768,64	0,18
	Riesling	2767,83	2644,00	1040,86	0,38
	Sauvignon_Blanc	3340,03	3338,00	628,95	0,19
	Viognier	4032,25	3815,00	798,53	0,20

gluconic.acid	Chardonnay	3970,00	3980,00	2815,98	0,71
	Pinot_Gris	2558,25	2590,50	784,67	0,31
	Riesling	5032,75	4603,50	2151,04	0,43
	Sauvignon_Blanc	2183,07	1296,00	1706,17	0,78
	Viognier	3795,58	3507,50	2443,34	0,64
glucose	Chardonnay	305658,97	254593,00	225436,81	0,74
	Pinot_Gris	585665,00	594262,00	224126,97	0,38
	Riesling	506608,83	519410,50	282690,99	0,56
	Sauvignon_Blanc	283163,00	212332,00	244818,08	0,86
	Viognier	406475,58	381348,50	157543,32	0,39
glucuronic.acid	Chardonnay	2982,97	2178,50	2541,14	0,85
	Pinot_Gris	3045,50	2169,50	2183,83	0,72
	Riesling	7337,83	2200,50	12014,36	1,64
	Sauvignon_Blanc	3230,03	2068,00	3143,48	0,97
	Viognier	1929,33	911,00	2875,75	1,49
glutamic.acid	Chardonnay	12530,56	12362,50	3620,49	0,29
	Pinot_Gris	11998,42	11794,00	778,05	0,06
	Riesling	15106,83	15161,00	2096,08	0,14
	Sauvignon_Blanc	10795,63	8322,00	4677,63	0,43
	Viognier	8248,25	8000,00	1510,17	0,18
glyceric.acid	Chardonnay	3333,53	3040,00	1484,15	0,45
	Pinot_Gris	3952,58	3951,50	759,47	0,19
	Riesling	3355,83	3475,00	782,19	0,23
	Sauvignon_Blanc	5519,77	5299,50	2105,97	0,38
	Viognier	6891,50	5515,00	3596,16	0,52
glycerol	Chardonnay	812101,42	729162,00	356444,02	0,44
	Pinot_Gris	1178330,42	1102932,00	210163,57	0,18
	Riesling	810031,08	776930,50	202899,51	0,25
	Sauvignon_Blanc	1380013,57	1353853,50	436886,77	0,32
	Viognier	844260,58	845445,50	258594,70	0,31
glycerol.3.galactoside.NIST	Chardonnay	9279,11	9613,50	2237,79	0,24
	Pinot_Gris	9637,58	9549,50	1004,08	0,10
	Riesling	20522,75	19662,50	6012,18	0,29
	Sauvignon_Blanc	11849,21	11440,50	2767,53	0,23
	Viognier	7652,17	5278,50	4850,73	0,63
glycerol.alpha.phosphate	Chardonnay	1299,11	1274,00	533,15	0,41
	Pinot_Gris	1119,92	1138,50	292,09	0,26
	Riesling	1560,42	1561,00	367,44	0,24
	Sauvignon_Blanc	2248,75	2377,00	918,87	0,41
	Viognier	2260,83	2058,50	906,33	0,40
glycine	Chardonnay	75695,50	71800,50	18991,24	0,25

	Pinot_Gris	66236,50	65937,50	5613,15	0,08
	Riesling	59966,08	59333,00	7199,18	0,12
	Sauvignon_Blanc	71697,47	64356,50	23855,85	0,33
	Viognier	73271,42	71876,00	12450,23	0,17
guanine	Chardonnay	965,31	696,50	772,38	0,80
	Pinot_Gris	663,25	727,50	182,63	0,28
	Riesling	480,00	461,00	158,07	0,33
	Sauvignon_Blanc	636,92	691,50	297,92	0,47
	Viognier	800,75	766,50	236,66	0,30
heptadecanoic.acid	Chardonnay	2101,56	2192,00	367,67	0,17
	Pinot_Gris	3100,83	3269,00	668,11	0,22
	Riesling	2233,75	2318,00	546,97	0,24
	Sauvignon_Blanc	2955,96	2949,50	956,71	0,32
	Viognier	2625,67	2771,00	611,17	0,23
homoserine	Chardonnay	3313,50	3163,00	979,47	0,30
	Pinot_Gris	3447,67	3211,50	964,04	0,28
	Riesling	2493,83	2356,50	503,07	0,20
	Sauvignon_Blanc	3601,83	3648,00	708,43	0,20
	Viognier	3339,50	3464,50	408,78	0,12
indole.3.lactate	Chardonnay	1120,42	899,00	724,63	0,65
	Pinot_Gris	1664,33	1672,00	773,53	0,46
	Riesling	1384,08	1425,00	339,96	0,25
	Sauvignon_Blanc	771,69	719,00	187,32	0,24
	Viognier	357,92	323,50	151,72	0,42
inositol.myo.	Chardonnay	192729,39	189660,50	24085,29	0,12
	Pinot_Gris	319440,25	313176,00	53811,96	0,17
	Riesling	296821,08	292629,50	38234,43	0,13
	Sauvignon_Blanc	398659,93	409112,50	71638,07	0,18
	Viognier	476168,92	440825,50	93118,82	0,20
inulobiose	Chardonnay	4618,81	1027,00	8166,99	1,77
	Pinot_Gris	1327,08	1236,50	301,23	0,23
	Riesling	1808,17	1508,50	1045,15	0,58
	Sauvignon_Blanc	1343,03	1220,50	694,89	0,52
	Viognier	17833,75	14044,00	11563,49	0,65
isocitric.acid	Chardonnay	4732,56	4901,50	956,37	0,20
	Pinot_Gris	6853,25	6700,00	472,92	0,07
	Riesling	4764,50	4899,50	774,44	0,16
	Sauvignon_Blanc	7610,07	8231,00	2247,94	0,30
	Viognier	7694,83	7735,00	1977,95	0,26
isogalactinol	Chardonnay	3276,67	3295,00	1061,51	0,32
	Pinot_Gris	3545,83	3519,00	932,79	0,26

	Riesling	2472,67	2321,50	882,59	0,36
	Sauvignon_Blanc	4215,86	3988,00	1612,31	0,38
	Viognier	5817,33	5231,50	2136,92	0,37
isoleucine	Chardonnay	24294,61	22396,50	9532,84	0,39
	Pinot_Gris	16399,50	16294,00	3347,43	0,20
	Riesling	14855,25	14944,50	1768,92	0,12
	Sauvignon_Blanc	18173,43	17308,00	7043,08	0,39
	Viognier	16499,75	16698,00	5721,40	0,35
isothreonic.acid	Chardonnay	3679,86	3683,00	409,44	0,11
	Pinot_Gris	2664,33	2577,00	664,40	0,25
	Riesling	2575,17	2540,50	726,72	0,28
	Sauvignon_Blanc	3998,63	3683,00	888,71	0,22
	Viognier	9277,50	8784,00	1789,96	0,19
lactic.acid	Chardonnay	896714,08	1106024,50	556758,99	0,62
	Pinot_Gris	195772,08	164641,00	100977,62	0,52
	Riesling	1316416,58	1191983,00	313536,33	0,24
	Sauvignon_Blanc	618376,93	235224,00	539210,78	0,87
	Viognier	199452,17	198545,50	25962,47	0,13
lauric.acid	Chardonnay	8471,44	6573,00	5031,76	0,59
	Pinot_Gris	8696,00	8388,00	2032,61	0,23
	Riesling	9444,08	6612,50	9369,37	0,99
	Sauvignon_Blanc	9457,00	9482,00	1763,85	0,19
	Viognier	14760,58	9320,00	11170,58	0,76
leucine	Chardonnay	80118,61	73945,50	26513,32	0,33
	Pinot_Gris	62502,17	61086,50	6725,25	0,11
	Riesling	43685,17	44319,00	4339,99	0,10
	Sauvignon_Blanc	66988,17	73169,00	23404,74	0,35
	Viognier	56345,83	56576,50	15298,37	0,27
lysine	Chardonnay	57286,44	52857,00	18310,25	0,32
	Pinot_Gris	47373,83	48642,50	5982,02	0,13
	Riesling	32429,58	31413,50	5296,81	0,16
	Sauvignon_Blanc	43893,73	49171,50	16856,84	0,38
	Viognier	36948,67	36923,00	9512,38	0,26
lyxitol	Chardonnay	3902,97	3813,00	827,38	0,21
	Pinot_Gris	4481,42	4467,50	225,81	0,05
	Riesling	4184,00	4194,50	1402,86	0,34
	Sauvignon_Blanc	6309,70	5357,00	5796,98	0,92
	Viognier	5124,50	4852,00	2329,80	0,45
malate	Chardonnay	145299,42	93018,50	121849,06	0,84
	Pinot_Gris	391317,83	398793,50	42458,65	0,11
	Riesling	103873,58	96252,00	77928,52	0,75

	Sauvignon_Blanc	251765,93	283759,50	218100,64	0,87
	Viognier	372134,67	373356,50	37730,00	0,10
maleic.acid	Chardonnay	249199,69	272504,00	78779,00	0,32
	Pinot_Gris	98506,42	87873,50	28299,38	0,29
	Riesling	119984,42	91598,00	117070,54	0,98
	Sauvignon_Blanc	119967,31	116445,00	39654,31	0,33
	Viognier	130922,33	120880,00	60476,85	0,46
maltose	Chardonnay	4430,69	3505,50	2820,29	0,64
	Pinot_Gris	5760,08	5601,00	1420,71	0,25
	Riesling	4106,08	3936,50	1363,62	0,33
	Sauvignon_Blanc	21159,04	16795,50	10801,90	0,51
	Viognier	9339,25	8479,00	3560,94	0,38
mannitol	Chardonnay	242539,28	169771,00	157503,17	0,65
	Pinot_Gris	71583,42	75021,50	14923,77	0,21
	Riesling	254172,67	232791,50	124735,90	0,49
	Sauvignon_Blanc	156668,67	159732,50	68880,21	0,44
	Viognier	112983,92	108515,50	56736,32	0,50
melibiose	Chardonnay	4946,92	4759,50	3060,43	0,62
	Pinot_Gris	6806,42	6949,50	3512,31	0,52
	Riesling	7334,83	7226,50	2676,08	0,36
	Sauvignon_Blanc	8682,88	8998,00	4419,61	0,51
	Viognier	21373,67	16102,50	14718,51	0,69
methionine	Chardonnay	5635,58	5354,50	2031,97	0,36
	Pinot_Gris	4156,25	4114,00	563,94	0,14
	Riesling	3054,25	3235,00	504,00	0,17
	Sauvignon_Blanc	4429,04	4886,50	1513,16	0,34
	Viognier	4307,67	4263,00	1222,68	0,28
N.acetyl.D.hexosamine	Chardonnay	11901,56	11591,00	1982,32	0,17
	Pinot_Gris	12264,92	12774,00	2704,66	0,22
	Riesling	8194,08	7988,00	2950,71	0,36
	Sauvignon_Blanc	12422,34	12338,00	2697,34	0,22
	Viognier	20442,67	20572,00	6459,58	0,32
N.acetyl.D.mannosamine	Chardonnay	11555,08	11587,00	1979,75	0,17
	Pinot_Gris	17311,83	17068,50	1542,29	0,09
	Riesling	8422,42	8591,50	1479,26	0,18
	Sauvignon_Blanc	12511,37	11974,00	2805,61	0,22
	Viognier	11955,08	11758,50	1455,56	0,12
nicotianamine	Chardonnay	549,42	520,00	142,71	0,26
	Pinot_Gris	611,92	615,00	138,10	0,23
	Riesling	413,92	381,00	117,64	0,28
	Sauvignon_Blanc	546,88	533,00	109,53	0,20

	Viognier	777,25	718,50	208,30	0,27
octadecanol	Chardonnay	822,33	853,50	240,89	0,29
	Pinot_Gris	1592,50	1635,00	625,90	0,39
	Riesling	711,25	639,00	317,59	0,45
	Sauvignon_Blanc	1313,00	1354,50	541,43	0,41
	Viognier	1059,33	870,50	561,98	0,53
oleic.acid	Chardonnay	251,50	219,00	119,66	0,48
	Pinot_Gris	360,75	372,00	79,87	0,22
	Riesling	357,08	328,50	202,15	0,57
	Sauvignon_Blanc	433,48	317,00	316,31	0,73
	Viognier	246,58	216,00	92,10	0,37
ornithine	Chardonnay	60602,08	62002,50	41424,70	0,68
	Pinot_Gris	97664,92	83876,50	69095,11	0,71
	Riesling	62425,58	64925,50	17283,18	0,28
	Sauvignon_Blanc	41378,17	16553,00	42387,80	1,02
	Viognier	116064,75	110196,50	29488,75	0,25
oxoproline	Chardonnay	224572,78	164470,00	140587,90	0,63
	Pinot_Gris	174132,08	176880,00	44118,54	0,25
	Riesling	220263,58	219926,50	37393,65	0,17
	Sauvignon_Blanc	147834,79	149489,00	68968,91	0,47
	Viognier	166569,00	167878,50	34786,43	0,21
palmitic.acid	Chardonnay	8595,42	8622,50	2063,66	0,24
	Pinot_Gris	17591,75	18012,50	3656,80	0,21
	Riesling	10986,17	11402,00	4033,68	0,37
	Sauvignon_Blanc	17517,90	16987,50	4949,46	0,28
	Viognier	11073,33	11042,00	2712,02	0,24
pelargonic.acid	Chardonnay	12093,89	10619,00	7519,10	0,62
	Pinot_Gris	17814,50	17969,00	3811,72	0,21
	Riesling	29172,75	21163,00	24897,18	0,85
	Sauvignon_Blanc	26616,63	20069,50	19792,23	0,74
	Viognier	19139,25	17653,50	7097,93	0,37
phenylalanine	Chardonnay	19634,56	18156,00	6530,47	0,33
	Pinot_Gris	16083,33	15663,00	2107,91	0,13
	Riesling	14714,00	14951,00	2677,04	0,18
	Sauvignon_Blanc	16768,67	17837,50	4290,69	0,26
	Viognier	13149,75	13139,00	3309,63	0,25
pipecolic.acid	Chardonnay	9708,69	9344,50	2841,08	0,29
	Pinot_Gris	6174,17	6347,50	784,06	0,13
	Riesling	10753,33	9637,50	5178,86	0,48
	Sauvignon_Blanc	11963,97	11642,00	3062,01	0,26
	Viognier	12056,92	12124,50	2408,08	0,20

proline	Chardonnay	2566507,67	2687400,00	449275,49	0,18
	Pinot_Gris	1161848,08	1124120,50	260290,82	0,22
	Riesling	753707,42	696244,00	412620,00	0,55
	Sauvignon_Blanc	1096838,17	1052200,00	229671,94	0,21
	Viognier	2028794,67	2038970,50	243975,99	0,12
pseudo.uridine	Chardonnay	2140,89	2291,50	1112,95	0,52
	Pinot_Gris	1585,50	1790,00	711,15	0,45
	Riesling	2174,75	1950,50	882,12	0,41
	Sauvignon_Blanc	1471,19	1473,50	445,10	0,30
	Viognier	3011,17	3421,00	2091,73	0,69
putrescine	Chardonnay	16831,58	16095,00	4555,34	0,27
	Pinot_Gris	17091,17	17199,50	1134,08	0,07
	Riesling	13063,75	12106,00	8413,50	0,64
	Sauvignon_Blanc	38968,13	33632,00	26558,28	0,68
	Viognier	30484,92	29396,00	10464,24	0,34
quinic.acid	Chardonnay	2721,50	2223,00	1359,12	0,50
	Pinot_Gris	8018,75	8090,00	4930,85	0,61
	Riesling	2317,33	2231,00	432,79	0,19
	Sauvignon_Blanc	4131,83	3255,00	2078,34	0,50
	Viognier	2964,00	2933,50	2207,08	0,74
ribonic.acid	Chardonnay	2339,81	2332,00	503,82	0,22
	Pinot_Gris	1572,25	1526,50	664,52	0,42
	Riesling	1675,67	1647,00	223,12	0,13
	Sauvignon_Blanc	3007,73	2734,00	857,19	0,28
	Viognier	4645,42	4488,50	665,28	0,14
ribose	Chardonnay	5835,47	5692,50	910,51	0,16
	Pinot_Gris	8332,33	8020,00	1636,92	0,20
	Riesling	6870,75	6965,50	840,99	0,12
	Sauvignon_Blanc	9255,10	9137,50	1456,07	0,16
	Viognier	5621,67	5516,50	558,82	0,10
saccharopine	Chardonnay	589,56	563,00	195,67	0,33
	Pinot_Gris	554,92	505,50	222,33	0,40
	Riesling	356,17	296,00	164,08	0,46
	Sauvignon_Blanc	796,46	833,50	248,63	0,31
	Viognier	604,75	588,00	270,41	0,45
serine	Chardonnay	30429,06	30449,50	9181,86	0,30
	Pinot_Gris	18118,33	17968,00	2188,77	0,12
	Riesling	23858,67	22082,50	5801,60	0,24
	Sauvignon_Blanc	15567,43	14637,00	3066,91	0,20
	Viognier	12860,50	14523,50	5357,97	0,42
shikimic.acid	Chardonnay	42750,53	43909,50	9434,68	0,22

	Pinot_Gris	14329,00	14181,50	628,54	0,04
	Riesling	61387,17	56051,00	28451,09	0,46
	Sauvignon_Blanc	36819,67	26976,00	18072,36	0,49
	Viognier	13772,58	13518,50	3209,80	0,23
sophorose	Chardonnay	6398,08	6378,50	2133,20	0,33
	Pinot_Gris	12029,92	12252,00	2294,81	0,19
	Riesling	7579,17	7051,50	3364,64	0,44
	Sauvignon_Blanc	21330,50	23358,00	7328,38	0,34
	Viognier	15044,08	13683,00	4729,25	0,31
sorbitol	Chardonnay	87171,53	83695,50	17494,94	0,20
	Pinot_Gris	77533,25	80045,00	6357,14	0,08
	Riesling	76251,17	70323,50	39315,30	0,52
	Sauvignon_Blanc	117031,20	110802,50	28836,68	0,25
	Viognier	158240,50	163287,50	69421,94	0,44
spermidine	Chardonnay	1697,39	1508,50	987,01	0,58
	Pinot_Gris	986,42	1089,50	389,33	0,39
	Riesling	1372,83	1354,00	323,34	0,24
	Sauvignon_Blanc	812,88	727,00	316,94	0,39
	Viognier	1582,92	1461,00	814,02	0,51
stearic.acid	Chardonnay	55213,06	53346,50	15813,39	0,29
	Pinot_Gris	114971,50	119660,00	33297,71	0,29
	Riesling	77339,00	65404,00	35906,33	0,46
	Sauvignon_Blanc	126667,73	113920,00	47201,52	0,37
	Viognier	69475,83	67149,00	17241,35	0,25
suberyl.glycine	Chardonnay	1329,67	1229,00	510,14	0,38
	Pinot_Gris	2465,83	2474,00	726,08	0,29
	Riesling	1504,00	1439,00	437,23	0,29
	Sauvignon_Blanc	1780,64	1598,00	715,21	0,40
	Viognier	3551,75	3449,00	2283,17	0,64
succinic.acid	Chardonnay	811826,53	787728,00	173454,98	0,21
	Pinot_Gris	885785,00	886659,50	64676,46	0,07
	Riesling	464936,75	448624,00	77439,36	0,17
	Sauvignon_Blanc	1334952,20	1081027,50	538576,60	0,40
	Viognier	1073004,25	1097538,50	144541,84	0,13
talose	Chardonnay	26711,92	2521,50	77055,09	2,88
	Pinot_Gris	4194,75	3759,50	1790,96	0,43
	Riesling	304145,08	79139,00	390179,71	1,28
	Sauvignon_Blanc	3335,21	2461,50	4382,84	1,31
	Viognier	8904,58	4874,00	7833,54	0,88
tartaric.acid	Chardonnay	374417,92	353324,00	69889,33	0,19
	Pinot_Gris	599653,17	580829,00	70140,28	0,12

	Riesling	489255,33	477697,50	105499,61	0,22
	Sauvignon_Blanc	707568,93	694135,00	137918,28	0,19
	Viognier	476769,83	479991,50	20348,60	0,04
threitol	Chardonnay	7967,11	7930,50	1043,77	0,13
	Pinot_Gris	7740,17	7783,50	1223,95	0,16
	Riesling	6819,17	6134,00	3424,80	0,50
	Sauvignon_Blanc	18187,80	17886,50	3498,14	0,19
	Viognier	24017,25	23401,50	4728,54	0,20
threonic.acid	Chardonnay	5699,22	5731,50	729,92	0,13
	Pinot_Gris	8328,33	8123,50	1275,60	0,15
	Riesling	4618,25	4252,00	1506,73	0,33
	Sauvignon_Blanc	13710,33	11246,50	5427,83	0,40
	Viognier	12953,00	12162,50	5050,25	0,39
threonine	Chardonnay	18972,31	18497,50	4417,98	0,23
	Pinot_Gris	13743,25	14629,00	2193,89	0,16
	Riesling	13498,58	12978,50	2668,26	0,20
	Sauvignon_Blanc	10725,17	11092,00	3427,17	0,32
	Viognier	9051,75	8818,50	4598,58	0,51
trehalose	Chardonnay	4879,44	2104,50	5561,54	1,14
	Pinot_Gris	7246,50	6093,00	3428,57	0,47
	Riesling	4549,83	3809,00	2910,45	0,64
	Sauvignon_Blanc	40833,67	29223,50	25159,94	0,62
	Viognier	6504,25	6548,50	1652,20	0,25
tryptophan	Chardonnay	1382,33	1501,00	883,50	0,64
	Pinot_Gris	2033,00	1972,50	459,80	0,23
	Riesling	2308,00	2153,50	545,92	0,24
	Sauvignon_Blanc	1883,00	1547,00	1196,53	0,64
	Viognier	2127,08	1916,50	1647,88	0,77
tyrosine	Chardonnay	35646,08	36168,50	9802,73	0,28
	Pinot_Gris	28395,42	27404,00	2232,75	0,08
	Riesling	19723,58	18315,00	9694,66	0,49
	Sauvignon_Blanc	28584,90	28409,50	5472,71	0,19
	Viognier	27687,50	26642,00	8374,70	0,30
uracil	Chardonnay	8021,56	6761,00	4946,97	0,62
	Pinot_Gris	12110,67	11723,50	3318,37	0,27
	Riesling	7368,75	7064,00	1812,01	0,25
	Sauvignon_Blanc	16156,92	12845,00	8800,57	0,54
	Viognier	9489,83	10162,50	4213,18	0,44
urea	Chardonnay	2768,86	2411,50	1109,68	0,40
	Pinot_Gris	2829,33	2563,50	758,21	0,27
	Riesling	2656,17	2719,00	1024,62	0,39

	Sauvignon_Blanc	3825,71	3989,50	1288,66	0,34
	Viognier	7972,67	7643,50	5173,99	0,65
uridine	Chardonnay	238,81	203,00	142,91	0,60
	Pinot_Gris	205,33	202,50	48,86	0,24
	Riesling	128,58	124,00	36,59	0,28
	Sauvignon_Blanc	190,72	194,00	63,45	0,33
	Viognier	169,33	171,00	59,13	0,35
valine	Chardonnay	44968,56	43194,00	11899,41	0,26
	Pinot_Gris	37128,83	35572,50	3519,57	0,09
	Riesling	39343,58	38987,50	4564,04	0,12
	Sauvignon_Blanc	36378,53	36530,50	5608,77	0,15
	Viognier	28434,17	29403,50	8647,95	0,30
xanthine	Chardonnay	735,17	653,50	327,25	0,45
	Pinot_Gris	776,17	839,00	147,17	0,19
	Riesling	738,17	683,00	154,39	0,21
	Sauvignon_Blanc	1539,42	1504,00	835,32	0,54
	Viognier	746,33	723,50	184,11	0,25
xylitol	Chardonnay	2320,36	2336,00	608,13	0,26
	Pinot_Gris	2979,58	2777,00	430,89	0,14
	Riesling	2147,08	1860,50	1305,56	0,61
	Sauvignon_Blanc	3894,67	3558,50	865,61	0,22
	Viognier	4437,17	4192,50	1543,28	0,35
xylose	Chardonnay	28636,36	30697,00	6888,57	0,24
	Pinot_Gris	23391,08	23999,50	1906,33	0,08
	Riesling	23889,25	22244,50	6573,75	0,28
	Sauvignon_Blanc	20054,90	20053,00	2611,32	0,13
	Viognier	24626,67	23927,00	2770,41	0,11

Tabla 4[24]. Estadísticos para cada uno de los metabolitos, separando por variedades

En la **Tabla 4** se puede observar como los valores de los estadísticos difieren por cada metabolito en función de la variedad que se esté tratando, lo que lleva a pensar que sí que hay diferencias entre variedades en un mismo metabolito. Los resultados de la tabla se van a ver más adelante gráficamente gracias a los boxplots.

1.2. Estudio de valores faltantes

El segundo punto que se tendrá en cuenta es el recuento de valores faltantes que hay en la base de datos; en concreto, se hará un breve estudio sobre los valores que no

[24] Fuente: Elaboración propia.

están disponibles en la base de datos, y que por lo tanto aparecen como *Not Available (NA).*

En la base de datos hay 155 valores faltantes; por lo tanto, y teniendo en cuenta que se dispone de un total de 11.220 datos representan un 1,38% del total de datos. Este porcentaje es muy pequeño, así que, aunque se hará la imputación para evitar posibles problemas en fases que no aceptan valores faltantes, en realidad estos no distorsionarían demasiado los resultados del estudio.

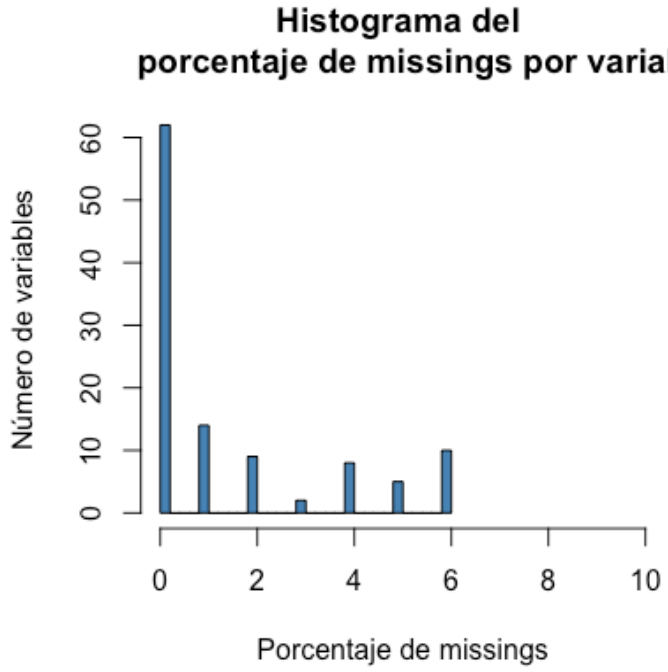

Figura 12[25]. Porcentaje de missings por variable

El histograma de la **Figura 12** complementa esta información. En él se puede observar el porcentaje de *missings* por variable; de este modo se puede observar como la mayoría de las variables de la base de datos no tienen valores faltantes. Aun así, es cierto que algunas de ellas tienen un porcentaje de *missings* más elevado; por ejemplo, más de diez de las variables tienen un porcentaje de valores faltantes que se aproxima al 6%. De hecho, en las variables con más porcentaje de valores faltantes estos representan concretamente el 5,88% de los datos de la variable, y son las correspondientes a *Hydrowy 3 methyglutaric acid, hidroxyproline, asparagine, cysteine, galactinol, glicerol 3 galactoside NIST, maltose, nicotianamine, saccharopine* y *talose.* De todos modos, un porcentaje de valores faltantes inferior al 6% se considera

[25] Fuente: elaboración propia.

muy pequeño; así pues, no se contempla la posibilidad de eliminar ninguna de las variables, ya que todas ellas tienen pocos *missings*.

Por otro lado, se ha estudiado también qué variedades presentan más *missings;* la **Tabla 5** muestra, por cada una de las variables, el número de valores faltantes que se han encontrado.

	Variedades				
	CHARDONNAY	PINOT GRIS	RIESLING	SAUVIGNON BLANC	VIOGNIER
Número de missings	0	0	0	155	0

Tabla 5[26]. Número de missings por variedad

De este modo se puede observar como la mayoría de las variedades no presentan valores faltantes, y sin embargo una de ellas los contiene todos: el Sauvignon Blanc. A la hora de presentar los resultados, es importante saber que en esta variedad había la todos los valores faltantes, aunque teniendo en cuenta que se tienen un total de 30 observaciones de Sauvignon Blanc, y por lo tanto 3.270 metabolitos, los valores faltantes tan solo representan un 4,7% de los datos disponibles de esta variedad.

Hasta ahora se ha hablado de los valores faltantes reales, es decir, aquellos que no están disponibles en la base de datos. No obstante, hay otro tipo de valores faltantes, y son aquellos que se han determinado como cero en la base de datos porque estaban por debajo del límite de detección del instrumento que analizó el metabolito.

Como se ha visto en el capítulo anterior, los valores faltantes se deben tratar de forma distinta en función de su tipología; así pues, es importante determinar cuántos valores iguales a cero hay en la base de datos y estudiar su origen para comprobar si son ceros reales o no.

Haciendo un *summary* sencillo en R, ya se puede observar que no hay ningún metabolito en el que el mínimo sea cero. Aun así, y para confirmarlo, se ha analizado cada variedad por separado, buscando cuantos valores iguales a cero se pueden encontrar en cada una de ellas. No obstante, no se han encontrado valores cero en ninguna de las variedades; por lo tanto, se puede confirmar que en la base de datos no hay valores de este tipo.

[26] Fuente: elaboración propia.

1.3. Boxplots iniciales

Para empezar, se ha hecho un gráfico con boxplots múltiples que comprende los 109 metabolitos disponibles en la base de datos.

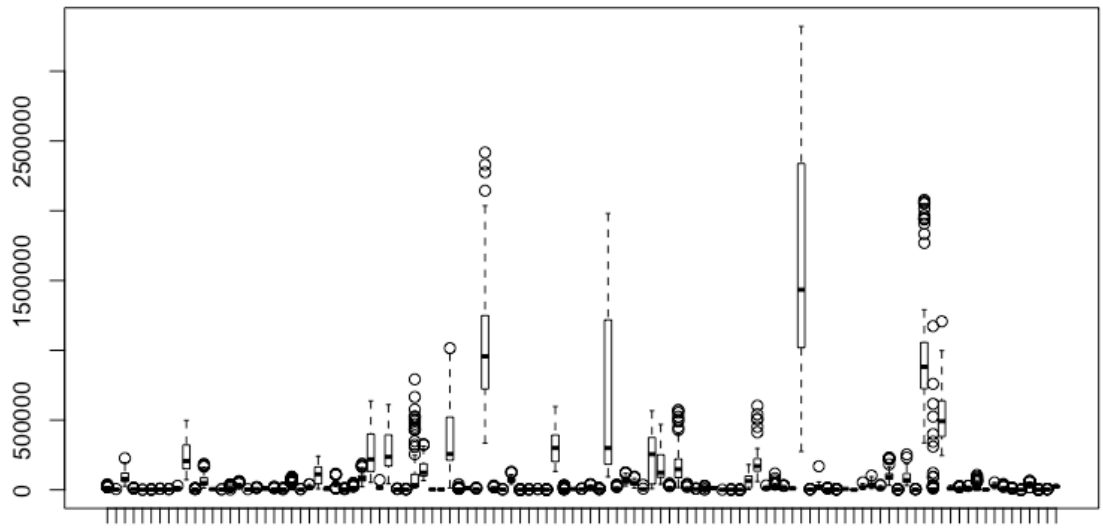

Figura 13[27]. Boxplots de los metabolitos antes del preprocesado

En la **Figura 13** se puede observar cómo los metabolitos presentan bastante variabilidad entre ellos, cosa que indica que puede ser necesaria una transformación y un escalado de los datos.

También se han graficado en forma de boxplot cada uno de los metabolitos encontrados discriminando por cada una de las variedades, con la intención de estudiar el sesgo, la simetría y la dispersión de los datos. Aunque inicialmente se han graficado todos ellos, y se pueden ver en el Anexo[28], se ha hecho una selección de algunos de ellos que serán los que se presentarán en el informe, y que se mostrarán a continuación.

[27] Fuente: elaboración propia.

[28] https://drive.google.com/drive/folders/1nbjoDt7UCsV3Gio0N-bBzivRgEP6XjwU

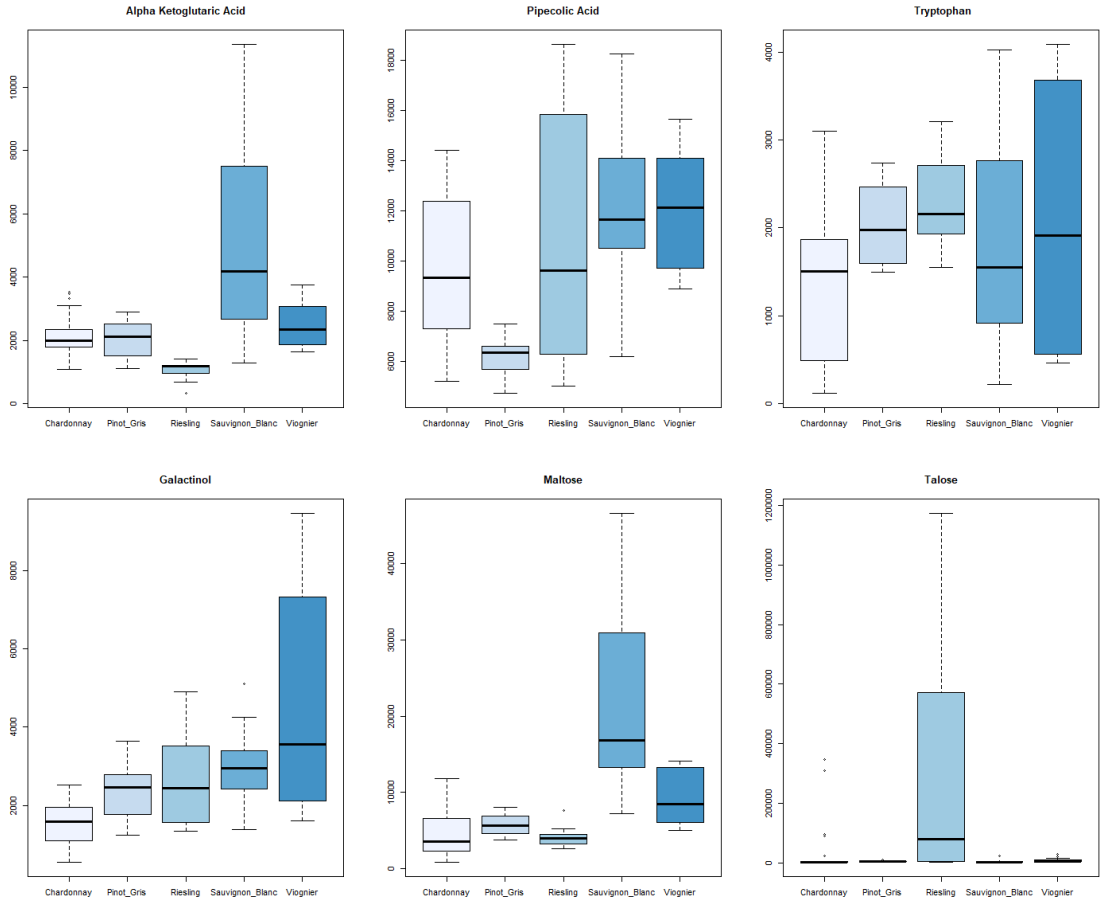

Figura 14[29]. Boxplots de algunos de los metabolitos antes del preprocesado

Los primeros boxplots que se muestran son los relacionados con el metabolito *Alpha Ketoglutaric Acid*, uno por cada variedad de vino. En el Sauvignon se observa una gran dispersión y valores más altos comparado con el resto de las variedades, en especial con el Riesling, que tiene valores bajos del metabolito y poca dispersión. El Pinot Gris parece que tiene mucha simetría, aunque en el resto de las variedades la asimetría no es muy pronunciada. La variedad donde se intuye más asimetría es el Riesling, con una asimetría hacia la derecha, de forma que la mediana está mucho más cerca del tercer cuartil que del primero. Se observan valores *outliers* en el Chardonnay y el Riesling, pero no en el resto de las variedades. En cuanto al sesgo, se puede determinar que en el Sauvignon y el Viognier hay sesgo positivo, ya que la cola de la caja es más larga hacia los valores más altos.

En el caso del *Pipecolic Acid* también destaca el Pinot Gris, que muestra unos valores bastante inferiores a los del resto de variedades. Se detectan variables con mucha dispersión, sobre todo el Riesling o el Chardonnay; de hecho, la única variedad en la

[29] Fuente: elaboración propia.

que no parece ser tan grande es el Pinot Gris. Además, destaca la simetría casi perfecta del Viognier, ya que el resto de las variedades presentan una distribución algo asimétrica. En este caso no se detectan valores *outlier*, y el sesgo no parece ser pronunciado en ninguna de las variedades.

Del metabolito *Tryptophan* se debe destacar la gran dispersión de la variedad Viognier y del Sauvignon y el Chardonnay, aunque en menor medida. Destaca la asimetría hacia la derecha del Chardonnay, así como la asimetría hacia la izquierda del Riesling, aunque las otras variedades tampoco son muy simétricas. Se observa también sesgo positivo en la variedad Chardonnay y Sauvignon Blanc: la cola superior es bastante más larga hacia los valores más altos. En este caso tampoco se observan valores atípicos.

En los boxplots referentes al *galactinol* se observa como todas las variedades tienen más o menos la misma dispersión a excepción del Viognier, que muestra una dispersión de este metabolito muy superior al resto de variedades. El Viognier presenta valores muy superiores al resto de variedades, aun teniendo el Sauvignon un *outlier*, aunque las medianas no difieren en exceso. También se observa sesgo positivo en las variedades Riesling y Viognier.

En cuanto a la *maltose*, se observan variedades con muy poca dispersión y en general bastante asimetría. En este caso destaca el boxplot del Sauvignon Blanc, donde la dispersión es mucho más grande, y presenta un sesgo hacia la derecha, y asimetría hacia la izquierda, cosas que también se observan en el Chardonnay. Se concluiría que hay una diferencia notable sobre todo entre el Sauvignon y el resto de las variedades.

Por otro lado, y quizás el caso más curioso, hay el metabolito *talose*. En él se observan valores bajísimos, muy cercanos al cero, en todas las variedades, excepto en el Riesling. De hecho, no se puede interpretar el boxplot de ninguna otra variedad, ya que lo único que se observa en ellos son algunos valores *outliers* en el Chardonnay, el Pinot Gris y el Sauvignon Blanc. Esto quiere decir que los valores del metabolito en las distintas variedades son muy pequeños y están muy concentrados, de forma que no hay dispersión. El gráfico, además, viene condicionado por los grandes valores que toma el metabolito en el Riesling, lo que hace que el eje vertical llegue a valores muy altos y no se concentre en los inferiores; en esta variedad hay un notable sesgo positivo. En cuanto al Riesling, se sospecha como el mínimo casi coincide con el primer cuartil (aunque se tendría que observar más de cerca para poder confirmarlo), y tiene una asimetría hacia la izquierda bastante pronunciada.

Para poder interpretar correctamente los boxplots del metabolito se ha modificado la escala del eje vertical, de forma que no se observa por completo el boxplot de la variedad Riesling, pero se pueden visualizar perfectamente los de las otras variedades.

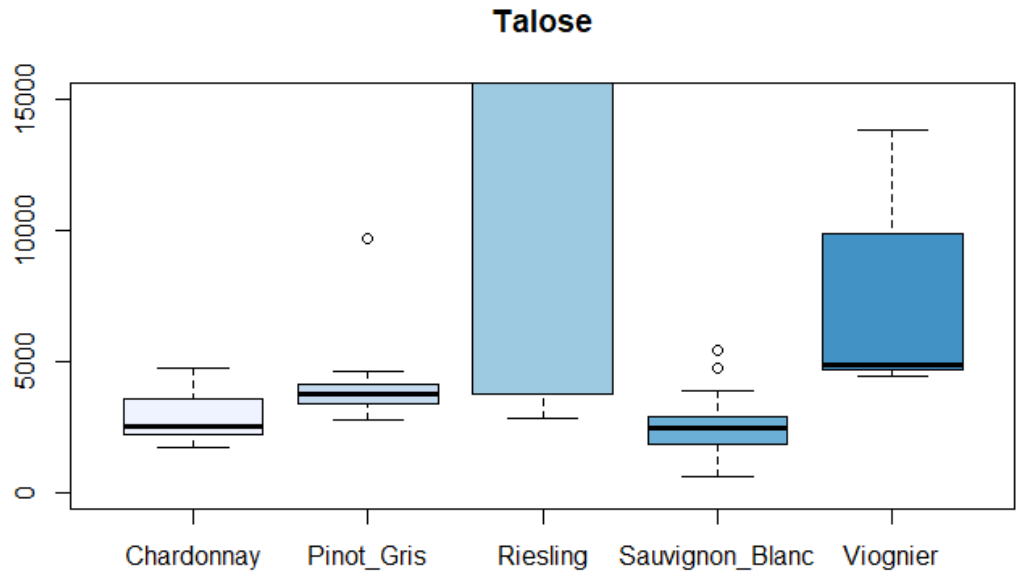

Figura 15[30]. Boxplot escalado del metabolito "Talose"

Primero se debe rechazar la sospecha de que el mínimo del Riesling coincide con el primer cuartil, ya que en la **Figura 15** se muestra que no es así. Por otra parte, se observa que las otras variedades tienen poca dispersión, a excepción de la Viognier, que muestra más que las otras, además de sesgo positivo y una gran asimetría hacia la derecha, cosa que también se detecta en el Chardonnay. En cambio, parece que el Pinot Gris y el Sauvignon son bastante simétricos en cuanto a los valores de este metabolito, y tienen poca dispersión, aunque se identifican uno y dos valores atípicos, respectivamente.

Así pues, y aunque a priori parecía que a excepción del Riesling todas las variables tenían valores casi iguales a cero y no presentaban dispersión ni asimetría, profundizando se ha visto como eso no era así; no obstante, sí que es cierto que las medianas del resto de variedades están bastante cerca, a excepción de la del Viognier, que está un poco más elevada.

Se debe poner énfasis en que estos gráficos están hechos con valores faltantes, así que tan solo sirven como apoyo para ver cómo se comportan los datos y como deberán

[30] Fuente: elaboración propia.

tratarse, pero no se pueden extraer otro tipo de conclusiones del estudio a partir de ellos; así pues, después de la imputación se deberán realizar de nuevo.

2. Técnicas de *preprocessing*

Una vez hecho el análisis descriptivo inicial, se procede a hacer el *preprocessing* de los datos. En esta etapa se realiza la imputación de los valores faltantes, la transformación de los datos y la normalización y escalado, en caso de ser necesario.

2.1. Imputación de valores faltantes

En el apartado anterior se ha visto como, a pesar de que sí que hay valores faltantes porque no se recogieron, no hay valores que se deban tratar porque estuviesen debajo del nivel de detección.

Como se ha explicado en el desarrollo del *pipeline*, los valores faltantes que aparecen en la base de datos como NA se tratan con la función *KNNimputation*. Así pues, se ha aplicado esta función a la base de datos con el software R, y se ha podido comprobar como la base de datos resultante no tiene valores faltantes.

Al no haber valores que fuesen cero porque quedaban por debajo del nivel de detección, no se debe hacer más tratamiento en este paso.

2.2. Transformaciones

En el análisis descriptivo inicial, sobre todo gracias a los boxplots múltiples, se ha visto que hay dispersión en los metabolitos, en algunos de ellos muy significativa, y también se han observado datos sesgados. La transformación ayuda a eliminar la variancia no constante y a corregir una distribución sesgada de los datos; así pues, es conveniente realizarla.

Siguiendo las recomendaciones que se han comentado en el capítulo anterior, se ha aplicado la transformación logarítmica a la base de datos con la función *log* del software R.

2.3. Normalización y escalado

Como se ha visto en el anterior capítulo, la normalización sirve para reducir la variación sistemática y separar la variación biológica de la no biológica introducida por el proceso. Por otro lado, el escalado se hace para centrar los datos alrededor de la media. Así pues, es conveniente realizarla en la mayoría de los estudios, ya que la variación en los estudios experimentales es notable. Además, gracias al estudio

descriptivo inicial se ha visto que en los datos hay variación; así pues, se ha decidido implementar una función para hacer un escalado de Pareto, que es el más recomendado para datos metabolómicos.

Para comprobar que el *preprocessing* se ha hecho de forma conveniente y que se ha conseguido arreglar los problemas de variación que se habían visto en el análisis descriptivo inicial, se ha hecho una comparación entre los boxplots múltiples obtenido antes y después del preprocesado.

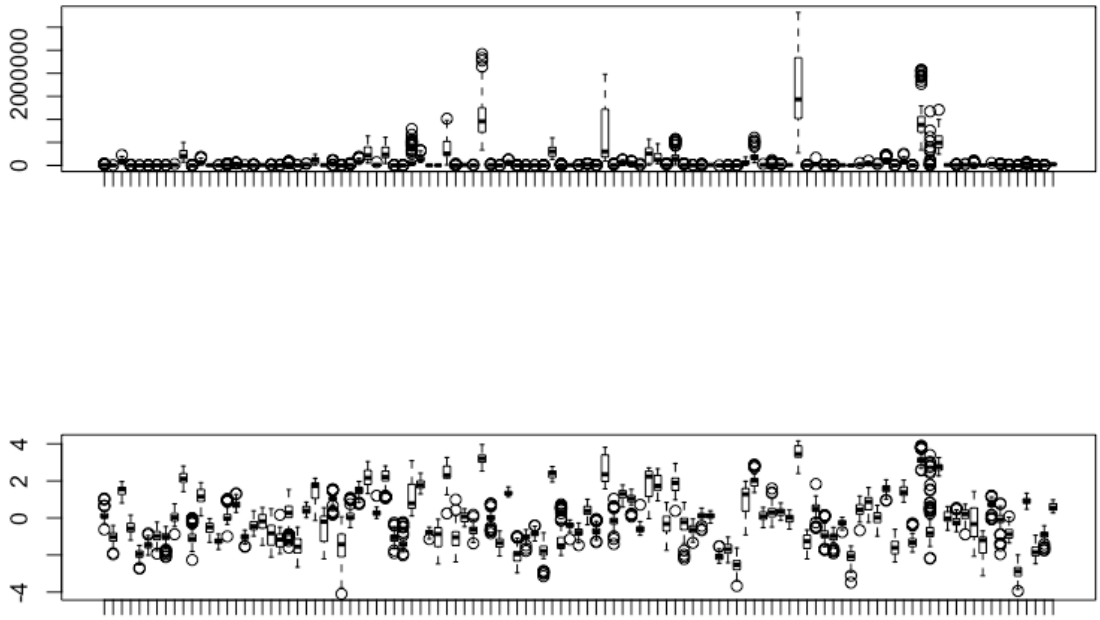

Figura 16[31]. Comparación de los boxplots antes y después del preprocesamiento

En la **Figura 16** se muestra cómo, mientras inicialmente había variabilidad en los datos y estaban descentrados, con las transformaciones que se han hecho durante el *preprocessing* se ha conseguido solucionar estos problemas, de modo que ahora se tienen datos centrados y con una menor variabilidad, detalles importantes para realizar un análisis de los datos más cuidadoso.

[31] Fuente: elaboración propia.

3. Análisis de los datos

Como se ha visto en el capítulo anterior, en este apartado se pueden realizar distintos métodos de análisis en función del objetivo del estudio. Es relevante recordar, pues, que el objetivo principal de este es observar en qué se diferencian distintas variedades de vinos blancos des del punto de vista metabolómico.

A continuación, se desarrollarán una serie de técnicas de las que se han visto en el capítulo anterior, que son las que se han creído más relevantes para el estudio que se está realizando.

3.1. Prueba ANOVA de una vía para datos independientes

Para empezar con el análisis de los datos, se ha hecho un análisis estadístico univariado mediante la prueba ANOVA. Dada la estructura de los datos y el objetivo del estudio se ha hecho una ANOVA de una vía para datos independientes para cada uno de los metabolitos, ya que lo que se desea ver es si en cada uno de ellos existen diferencias significativas entre las distintas variedades o no. La **Tabla 6** muestra el *p-value* que se ha obtenido en la prueba para cada uno de los metabolitos, además del *p-value* ajustado, que en realidad será el que se usará a la hora de extraer las conclusiones convenientes.

METABOLITO	P-VALUE	P-VALUE AJUSTADO	METABOLITO	P-VALUE	P-VALUE AJUSTADO
threitol	5,33E-39	5,80E-37	sorbitol	1,83E-08	9,86E-07
inositol.myo.	1,41E-38	1,52E-36	phenylalanine	2,89E-08	1,53E-06
isothreonic.acid	9,78E-36	1,05E-33	methionine	7,26E-08	3,78E-06
X2.hydroxyglutaric.acid	8,18E-33	8,67E-31	fructose	9,11E-08	4,64E-06
proline	1,38E-31	1,45E-29	lysine	1,58E-07	7,90E-06
serine	2,24E-28	2,33E-26	ornithine	2,76E-07	1,35E-05
threonic.acid	2,34E-26	2,41E-24	talose	5,55E-07	2,67E-05
shikimic.acid	2,07E-25	2,11E-23	leucine	6,98E-07	3,28E-05
cis.caffeic.acid	4,31E-25	4,35E-23	citrulline	7,24E-07	3,33E-05
beta.alanine	1,20E-22	1,20E-20	X3.hydroxy.3.methylglutaric.acid	8,19E-07	3,68E-05
trehalose	7,01E-21	6,94E-19	quinic.acid	8,86E-07	3,90E-05
threonine	9,37E-21	9,18E-19	isoleucine	1,35E-06	5,80E-05
ribonic.acid	2,12E-20	2,05E-18	isocitric.acid	1,42E-06	5,98E-05
sophorose	5,70E-20	5,47E-18	urea	1,61E-06	6,59E-05
X4.hydroxyproline	2,49E-19	2,36E-17	nicotianamine	2,16E-06	8,64E-05
maltose	4,81E-18	4,52E-16	glycerol.alpha.phosphate	3,96E-06	0,000154335
tartaric.acid	1,59E-17	1,48E-15	melibiose	4,22E-06	0,000160185

caffeic.acid	1,86E-17	1,71E-15	oxoproline	4,26E-06	0,000160185
ribose	1,07E-16	9,78E-15	X226091.0	5,90E-06	0,000212341
glycerol.3.galactoside.NIST	1,18E-16	1,06E-14	glyceric.acid	6,17E-06	0,000215965
xylose	2,58E-16	2,30E-14	glycerol	9,02E-06	0,000306694
valine	2,76E-16	2,43E-14	glycine	1,02E-05	0,00033678
conduritol.beta.epoxide	8,72E-16	7,59E-14	pelargonic.acid	1,31E-05	0,000420116
putrescine	1,46E-15	1,26E-13	arachidic.acid	2,94E-05	0,000884221
indole.3.lactate	3,35E-15	2,85E-13	isogalactinol	2,85E-05	0,000884221
X2.isopropylmalic.acid	5,26E-15	4,42E-13	gluconic.acid	4,66E-05	0,001352464
maleic.acid	1,56E-14	1,29E-12	citric.acid	8,57E-05	0,002400423
dehydroascorbate	4,28E-14	3,51E-12	malate	0,00010333 9	0,002790161
X1.2.anhydro.myo.inositol	7,26E-14	5,88E-12	asparagine	0,00028764 3	0,007478723
inulobiose	8,46E-14	6,77E-12	uracil	0,00104096	0,026023994
glutamic.acid	1,68E-13	1,33E-11	suberyl.glycine	0,00113321	0,027197044
palmitic.acid	1,80E-13	1,40E-11	GABA	0,00125033 7	0,028757757
cellobiose	2,96E-13	2,28E-11	glucose	0,00127730 3	0,028757757
fucose	5,78E-13	4,40E-11	uridine	0,00131644 7	0,028757757
arabitol	9,31E-13	6,98E-11	glucuronic.acid	0,00149032 9	0,02907218
stearic.acid	1,25E-12	9,23E-11	octadecanol	0,00145360 9	0,02907218
N.acetyl.D.hexosamine	1,38E-12	1,01E-10	pseudo.uridine	0,00177041 9	0,031867546
spermidine	1,68E-12	1,21E-10	xanthine	0,00206832 8	0,035161577
X3.6.anhydrogalactose	1,76E-12	1,25E-10	saccharopine	0,00267494 7	0,042799151
benzoic.acid	2,63E-12	1,84E-10	cysteine	0,0028932	0,043398005
alanine	3,12E-12	2,15E-10	arabinose	0,00596991 9	0,077742952
aspartic.acid	3,38E-12	2,30E-10	galactonic.acid	0,00555306 8	0,077742952
alpha.ketoglutaric.acid	6,01E-12	4,03E-10	homoserine	0,00726323 6	0,087158834
tyrosine	6,89E-12	4,55E-10	X2.ketoisocaproic.acid	0,00959438 7	0,10553826
N.acetyl.D.mannosamine	4,08E-11	2,65E-09	guanine	0,01978268 4	0,197826839
erythritol	4,29E-11	2,74E-09	tryptophan	0,02363413 4	0,212707203
galactinol	3,31E-10	2,09E-08	X1.hexadecanol	0,03017620 5	0,21507037
glucoheptulose	3,57E-10	2,22E-08	behenic.acid	0,02688379 6	0,21507037
galactose	5,42E-10	3,31E-08	oleic.acid	0,12803781 1	0,768226866
xylitol	7,14E-10	4,28E-08	heptadecanoic.acid	0,19677317	0,851795714
galacturonic.acid	1,13E-09	6,66E-08	lyxitol	0,17035914 3	0,851795714
mannitol	1,47E-09	8,54E-08	butyl.stearate	0,92913625 7	1
succinic.acid	2,12E-09	1,21E-07	citramalate	0,52516200 5	1
lactic.acid	2,71E-09	1,51E-07	lauric.acid	0,51545394 1	1

pipecolic.acid	9,25E-09	5,09E-07

Tabla 6[32]. P-value y p-value ajustado obtenidos con la prueba ANOVA

En color rosado se han marcado aquellos metabolitos en los que no se han encontrado diferencias significativas entre los diferentes grupos de variedades. Como se puede comprobar, pues, la mayoría de ellos encuentran diferencias estadísticamente significativas en las medias de al menos dos variedades, ya que el *p-value* es superior a 0.05 en tan solo catorce de ellos, que representan algo menos del 13% de los metabolitos.

La **Tabla 6** muestra los metabolitos ordenados de más a menos significativos, teniendo en cuenta el *p-value* ajustado; por lo tanto, un *p-value* más pequeño induce a pensar que las diferencias en aquel metabolito son más notables. Además, vale la pena comentar que el ajuste hace que haya menos metabolitos significativos, y en general todos ellos son inferiores al *p-value* sin ajustar; eso es porque, como se ha visto, al ajustar se está siendo más estricto a la hora de testear la significación.

Este análisis ANOVA sirve para corroborar que las bases del objetivo del estudio están bien fundadas, ya que si existen diferencias entre variedades en casi todos los metabolitos seguramente se podrán identificar las variedades con cada uno de ellos y se podrá observar en qué se diferencian des del punto de vista metabolómico. Además, las próximas técnicas de análisis que se realizarán deberían coincidir con esta clasificación, destacando como metabolitos más importantes a la hora de identificar las distintas variedades aquellos que en el ANOVA son más significativos.

3.2. Boxplots

Una vez estudiada la significación de los metabolitos en relación con sus diferencias entre las distintas variedades, se representarán algunos de los más significativos en forma de Boxplot.

Se ha decidido representar los seis metabolitos más significativos, aunque en el Anexo[33] del trabajo se pueden encontrar todos los que forman parte de la base de datos. En el test ANOVA, se ha comprobado que aquellos metabolitos con un *p-value* ajustado más pequeño son los siguientes: *threitol, inisitol myo, isothreonic acid, X2 hydroglutaric acid, proline* y *serine.*

[32] Fuente: elaboración propia.

[33] https://drive.google.com/open?id=1nbjoDt7UCsV3Gio0N-bBzivRgEP6XjwU

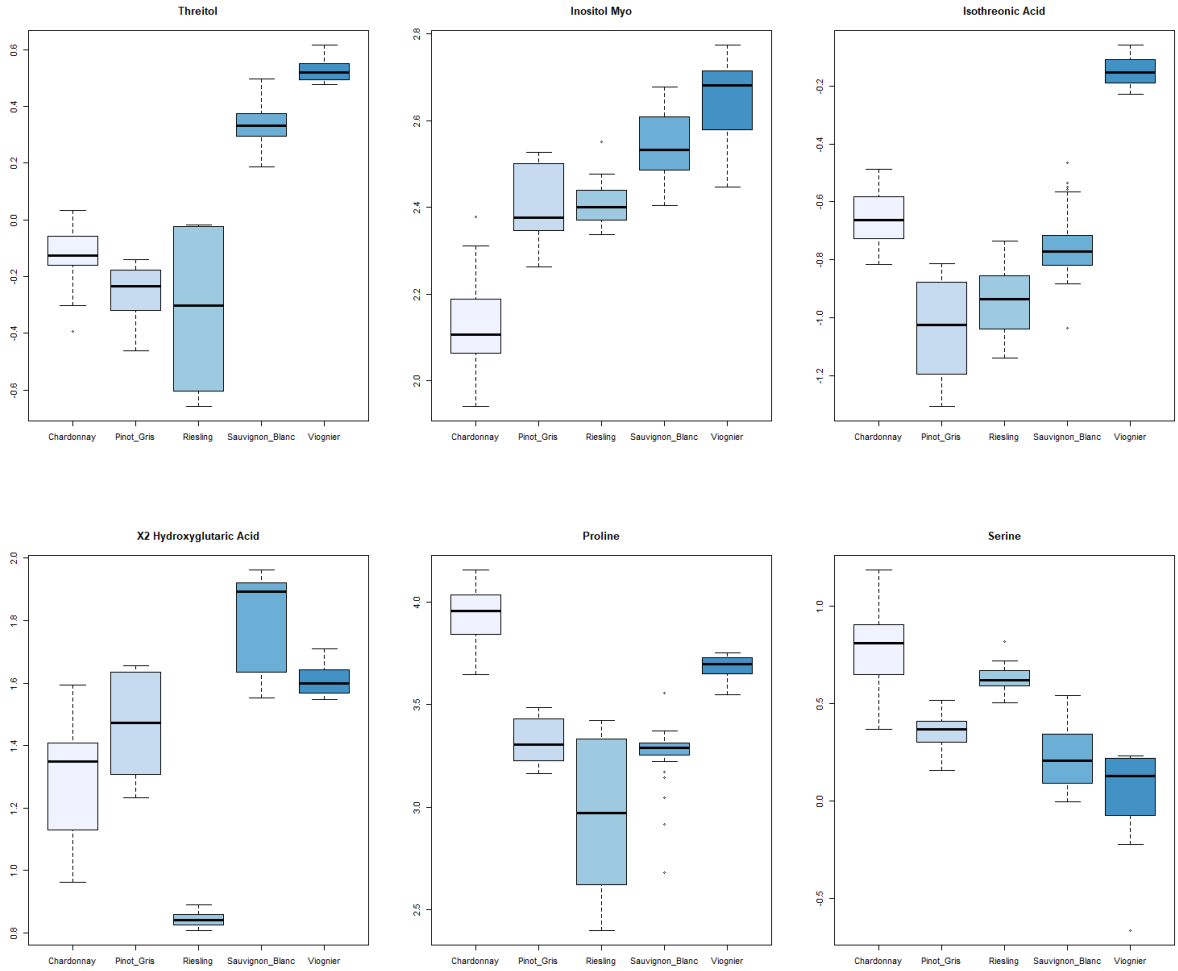

Figura 17[34]. Boxplots de los metbolitos más significativos después del preprocesado

En todos ellos se observan diferencias entre variedades. En cuanto al *threitol*, el metabolito más significativo, toma valores bastante altos en el Viognier y el Sauvignon, donde además se observa poca dispersión; por el contrario, las otras tres variedades toman valores más bajos del metabolito, y destaca la gran dispersión en el Riesling. Aun así, no se detecta asimetría en ninguna de las variedades.

Por otro lado, en el *inisitol myo* destaca el Chardonnay por tener valores bajos del metabolito, y aunque las otras variedades toman valores más altos, son bastante distantes entre ellas. En este caso sí que se observa asimetría hacia la izquierda en todas las variedades, a excepción de en el Viognier, donde esta es hacia la derecha, coincidiendo con ser la variedad con valores más altos del metabolito.

En el *isothreonic acid* el patrón es similar al observado en el primer metabolito comentado, ya que destaca el Viognier por tener valores muy superiores a los del resto

[34] Fuente: elaboración propia.

de las variedades, aunque en este caso es el Chardonnay la que más se le acerca, y por la simetría en todas las variedades.

Observando los boxplots correspondientes a *X2 hydroglutaric acid*, en este caso destaca el Riesling por sus valores bajos, su asimetría y su poca dispersión, y el Sauvignon por todo lo contrario: valores altos, con gran asimetría hacia la derecha y gran dispersión. El resto de las variables también se distribuyen de forma distinta entre ellas.

En el *proline* se observa que hay bastantes similitudes entre el Pinot Gris y el Sauvignon Blanc, pero que en las otras variables se ven muchas diferencias, contrastando la alta dispersión y valores más bajos del Riesling con la menor dispersión del Viognier y los valores altos del Chardonnay.

Por último, el *serine* también se comporta de forma distinta según la variedad. En este caso no hay ninguna variedad en la que se presente con mucha dispersión, aunque sí con poca el Pinot Gris y el Riesling. No obstante, donde se observan más diferencias es entre valores altos y bajos del metabolito, ya que de nuevo varían entre variedades, destacando el Chardonnay por tenerlos más altos.

En resumen, en los metabolitos más significativos se ve claramente que en función de la variedad que se esté estudiando el metabolito se comporta de una forma o de otra, cosa que corrobora que hay diferencias entre variedades y que los metabolitos pueden ayudar a explicarlas.

3.3. Análisis de Componentes Principales

Para identificar las variedades en función de los metabolitos mediante un método gráfico se ha realizado un estudio de análisis de componentes principales. Como se ha comentado en el capítulo anterior, este estudio proporcionará dos tipos de gráficos que pueden ayudar a explicar qué metabolitos son más importantes en cada una de las variedades.

En este caso, la primera componente principal consigue explicar un 23,2% de la variación de los datos, mientras que la segunda explica el 14,4% de esta variación. Se determina que esta proporción de variancia explicada acumulada ya proporciona un buen resumen de las muestras, así que se trabajará con tan solo dos componentes.

Antes de la representación gráfica, se han buscado aquellos metabolitos que tienen más peso en cada una de las dos componentes principales en las que se basa el estudio. Un metabolito será más relevante para explicar la componente cuanto más

grande sea en valor absoluto. Para hacerlo, se ha buscado el valor de los *loadings* para cada componente (*eigenvector)*.

Primera componente principal	
METABOLITO	**loadings**
citrulline	-0,330384051
ornithine	-0,325372517
malate	-0,322210025
talose	-0,267882606
cellobiose	0,254143602
citric.acid	-0,236899431
gluconic.acid	-0,203310569
trehalose	0,202896887
galactose	-0,186776736
GABA	-0,165474376
lactic.acid	0,155759796
cysteine	0,151657092
dehydroascorbate	0,137248854
inulobiose	-0,132135611
alpha.ketoglutaric.acid	0,130299126
arabinose	-0,123133678
xanthine	0,122256395
oxoproline	-0,122201813
maltose	0,116981233
sophorose	0,111774589
succinic.acid	0,110715553
mannitol	0,101726744
X2.hydroxyglutaric.acid	0,099102761
uracil	0,08488352
stearic.acid	0,082671724
X2.isopropylmalic.acid	0,078912161
galactonic.acid	-0,076699101
glycerol	0,076191186
galacturonic.acid	-0,07578583
pelargonic.acid	0,067474756
erythritol	0,066989872
tryptophan	0,064050377
leucine	0,063360773
tartaric.acid	0,060889094
glucose	-0,058561753
X4.hydroxyproline	-0,058463784

Segunda componente principal	
METABOLITO	**loadings**
trehalose	0,331836813
lactic.acid	-0,330962891
malate	0,276921975
maltose	0,222414271
mannitol	-0,183393782
threonine	-0,164668827
serine	-0,164136189
citric.acid	0,163728905
indole.3.lactate	-0,151171865
shikimic.acid	-0,148325373
sophorose	0,148165599
threitol	0,141204145
spermidine	-0,139354913
isoleucine	-0,133074276
threonic.acid	0,128271883
putrescine	0,124916318
maleic.acid	-0,124365541
inulobiose	0,123663391
X2.hydroxyglutaric.acid	0,122957471
cis.caffeic.acid	-0,121066543
lysine	-0,118924602
guanine	-0,118431039
talose	-0,115754482
cysteine	0,111324659
valine	-0,102878618
arabinose	0,100873019
methionine	-0,100687216
leucine	-0,09979295
phenylalanine	-0,098331985
X4.hydroxyproline	-0,097571722
inositol.myo.	0,095334275
caffeic.acid	-0,094598192
aspartic.acid	-0,093128062
alanine	-0,091360482
melibiose	0,090611744
pseudo.uridine	-0,088725969

isoleucine	0,057479913	xylitol	0,088709161
asparagine	-0,057377183	arabitol	-0,087471683
glutamic.acid	-0,056236026	glucuronic.acid	-0,086003023
saccharopine	0,055668777	GABA	0,083586168
alanine	-0,055550808	uridine	-0,082439033
lysine	0,055129761	gluconic.acid	-0,075160764
palmitic.acid	0,054660184	X3.hydroxy.3.methylglutaric.acid	0,072988171
threitol	0,05433341	urea	0,071938357
beta.alanine	-0,053756454	xylose	-0,07026734
X1.2.anhydro.myo.inositol	-0,053648463	glutamic.acid	-0,069018496
		sorbitol	0,068868061
		isocitric.acid	0,068820151
		glyceric.acid	0,066016473
		X2.isopropylmalic.acid	0,065426287
		galactinol	0,064198687
		proline	-0,062602097
		benzoic.acid	0,060038105
		glycerol.3.galactoside.NIST	-0,059924153
		galactose	-0,057460839
		isothreonic.acid	0,055602656
		stearic.acid	0,054652035
		citrulline	0,053694587
		palmitic.acid	0,052341913
		suberyl.glycine	0,052081013

Tabla 7[35]. Loadings de la primera y segunda Componentes del PCA de los metabolitos

En la **Tabla 7** se muestran los valores de los *loadings* para la primera y la segunda componentes principales. En este caso, se ha decidido mostrar tan solo aquellos metabolitos con un valor superior a 0,05 en valor absoluto, ya que se ha considerado que estos serán los más importantes para explicar cada una de las componentes. Aunque en orden distinto, la mayoría de los metabolitos que son importantes en la primera componente también lo son en la segunda. De hecho, revisando la **Tabla 6** con los *p-values* ajustados se observa que tan solo cuatro de los catorce metabolitos que no eran significativos en la ANOVA aparecen en alguna de las componentes como metabolito importante, concretamente el *arabinose,* el *galactaric acid*, la *guanine* y el *tryptophan*. Así pues, se puede determinar que los dos métodos dan unos resultados similares, ya que el resto de los metabolitos de la **Tabla 7** aparecen como muy significativos en la ANOVA, aun y ajustando el *p-value*.

[35] Fuente: elaboración propia.

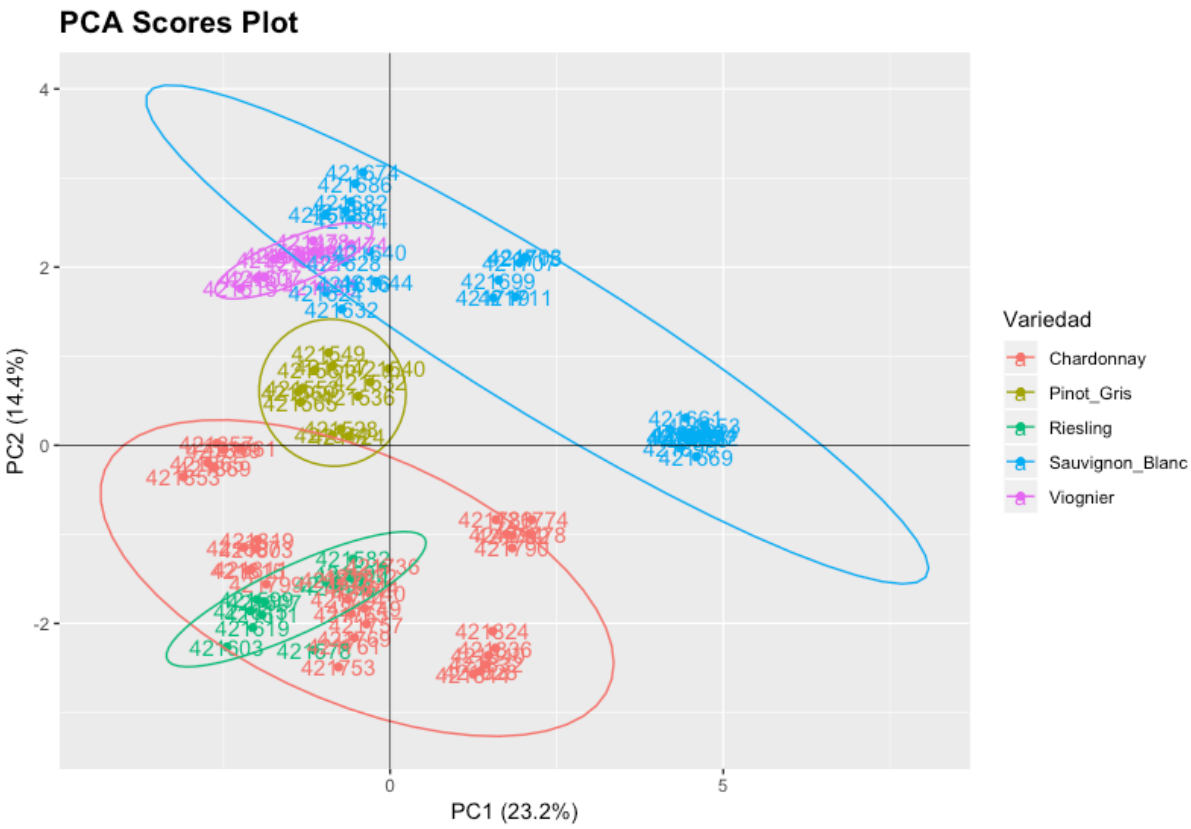

Figura 18[36]. Scores Plot del PCA

La **Figura 18** muestra el *Scores Plot* que se ha obtenido utilizando las dos primeras componentes principales. En el eje horizontal aparece la primera componente principal y en el vertical la segunda. Por lo tanto, los grupos estarán bien identificados por la primera componente si se separan bien de la parte izquierda del gráfico a la segunda, y por la segunda componente si los grupos de la parte superior del gráfico quedan bien separados de la inferior.

Lo primero que se observa es que casi no hay valores atípicos, ya que el análisis de componentes principales consigue representar todas las muestras dentro de la variedad correspondiente, aunque sí que se observan dos puntos en la variedad del Sauvignon Blanc que quedan fuera, a pesar de no estar a mucha distancia del límite del óvalo. Cada uno de los colores representa una de las variedades, como indica la leyenda; por ejemplo, dentro del óvalo azul hay todos los vinos que corresponden a la variedad Sauvignon Blanc, y dentro del naranja los que son Chardonnay.

En cuanto a la primera componente principal, no consigue identificar del todo bien los grupos, ya que el Sauvignon y el Chardonnay quedan divididos entre la parte izquierda del gráfico y la derecha; aun así, se puede considerar que el Sauvignon queda en la

[36] Fuente: elaboración propia.

parte derecha del gráfico en su mayor parte, y en cambio el resto de las variedades en la parte izquierda, incluido el Chardonnay, aunque como ya se ha comentado algunas de sus muestras quedan en la otra división.

En cambio, la segunda componente principal consigue identificar muy bien las variedades: por un lado, en la parte superior se encuentran el Sauvignon, el Viognier y el Pinot Gris, y, por otro lado, el Chardonnay y el Riesling se localizan en la parte inferior del gráfico.

Con la visualización del *plot* de *Loadings* se tendrá una visión de cuáles son los metabolitos que más influyen en cada una de las componentes analizadas, de forma que se podrá descubrir qué metabolitos son más relevantes para agrupar las muestras de las distintas variedades.

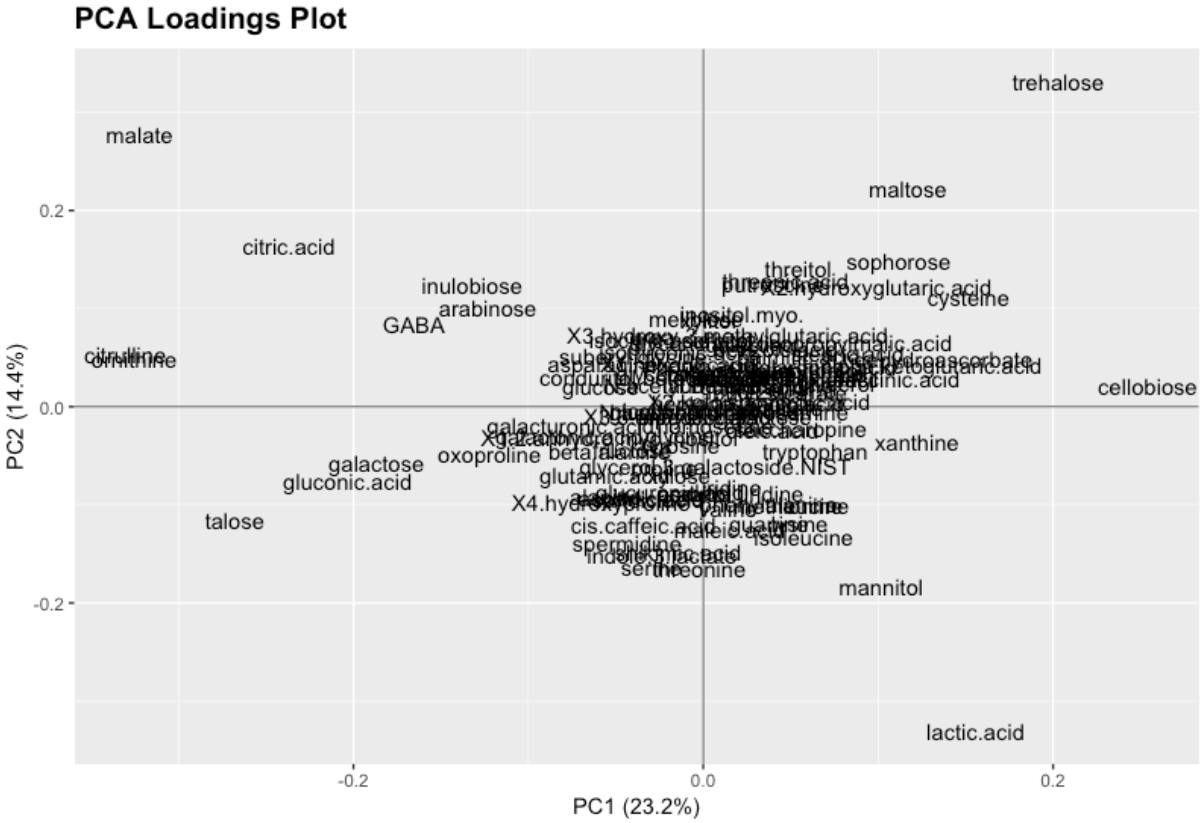

Figura 19[37]. Loadings Plot del PCA

La **Figura 19** representa el *plot* de *loadings* del PCA, y permite averiguar cuáles son los metabolitos que contribuyen a agrupar las muestras de las distintas variedades y separarlas de las demás; dicho de otra manera, cuáles son los metabolitos que

[37] Fuente: elaboración propia.

contribuyen a agrupar las entradas de Sauvignon a la derecha del gráfico y las del resto de variedades a la izquierda.

Los metabolitos más importantes son los que quedan más alejados del centro del gráfico, que como se puede observar es donde se concentran la mayoría de los metabolitos. Por lo tanto, los metabolitos que más explican la primera componente principal, y que por lo tanto separan las variedades de derecha a izquierda, son los *citrulline, ornithine, cellobiose* y podríamos incluir también *talose, galactose y gluconic acid.* Estos son los metabolitos que están más alejados del centro, pero a la vez más cerca de la línea horizontal que marca el cero del eje vertical. En cambio, aquellos que son más relevantes para la segunda componente serian *malalte, trehalose* y *lactic acid,* porque son los que se encuentran en los extremos más alejados del gráfico.

Si se revisa la **Tabla 7** con los *loadings*, se puede comprobar que estos metabolitos nombrados son, de hecho, aquellos que tenían los valores más grandes en valor absoluto para cada una de las componentes, y que además tenían un *p-value* ajustado significativo al realizar el test ANOVA.

Adicionalmente, se ha decidido representar este mismo gráfico separando los distintos metabolitos por su clase, es decir, separándolos en función de si son aminoácidos, ácidos orgánicos, azúcares, ácidos azucarados, azúcares alcohólicos o misceláneos, para ver si hay algún tipo de patrón en función de la clase de los metabolitos.

La **Tabla 8** muestra la clasificación en las diversas familias químicas que se ha considerado en el análisis.

FAMÍLIA	METABOLITOS		
Aminoácidos	4-hydroxyproline alanine asparagine aspartic acid beta alanine citrulline cysteine GABA glutamic acid glycine	homoserine isoleucine leucine lysine methionine N-acetyl-D-hexosamine N-acetyl-D-mannosamine ornithine phenylalanine	pipecolic acid proline saccharopine serine suberyl glycine threonine tryptophan tyrosine valine
Ácidos grasos	2-ketoisocaproic acid arachidic acid	heptadecanoic acid lauric acid	oleic acid palmitic acid
Alcoholes grasos	octadecanol		
Misceláneos	guanine	nicotianamine	xanthine
Ácidos orgánicos	1,2-anhydro-myo-inositol 2-hydroxyglutaric acid 2-isopropylmalic acid 226091.0 3-hydroxy-3-methylglutar 3,6-anhydrogalactose alpha ketoglutaric acid arabinose behenic acid benzoic acid caffeic acid cis-caffeic acid citramalate citric acid	conduritol-beta-epoxide dehydroascorbate glucoheptulose glucuronic acid glycerol-3-galactoside NIST glycerol-alpha-phosphate indole-3-lactate inulobiose isocitric acid lactic acid malate maleic acid melibiose oxoproline	pelargonic acid pseudo uridine putrescine quinic acid ribonic acid shikimic acid spermidine stearic acid succinic acid threonic acid uracil urea uridine
Azúcares	cellobiose fructose fucose galactose	glucose maltose ribose sophorose	talose trehalose xylose
Ácidos azucarados	galactonic acid galacturonic acid gluconic acid	glyceric acid isogalactinol	isothreonic acid tartaric acid
Azúcares alcohólicos	1-hexadecanol arabitol butyl stearate erythritol	galactinol glycerol inositol myo- lyxitol	mannitol sorbitol threitol xylitol

Tabla 8[38]. Clasificación de los metabolitos según su família química

Como se puede observar, la clase con más presencia son los ácidos orgánicos, seguidos por los aminoácidos, aunque también hay un número considerable de azúcares de distintos tipos.

[38] Fuente: elaboración propia.

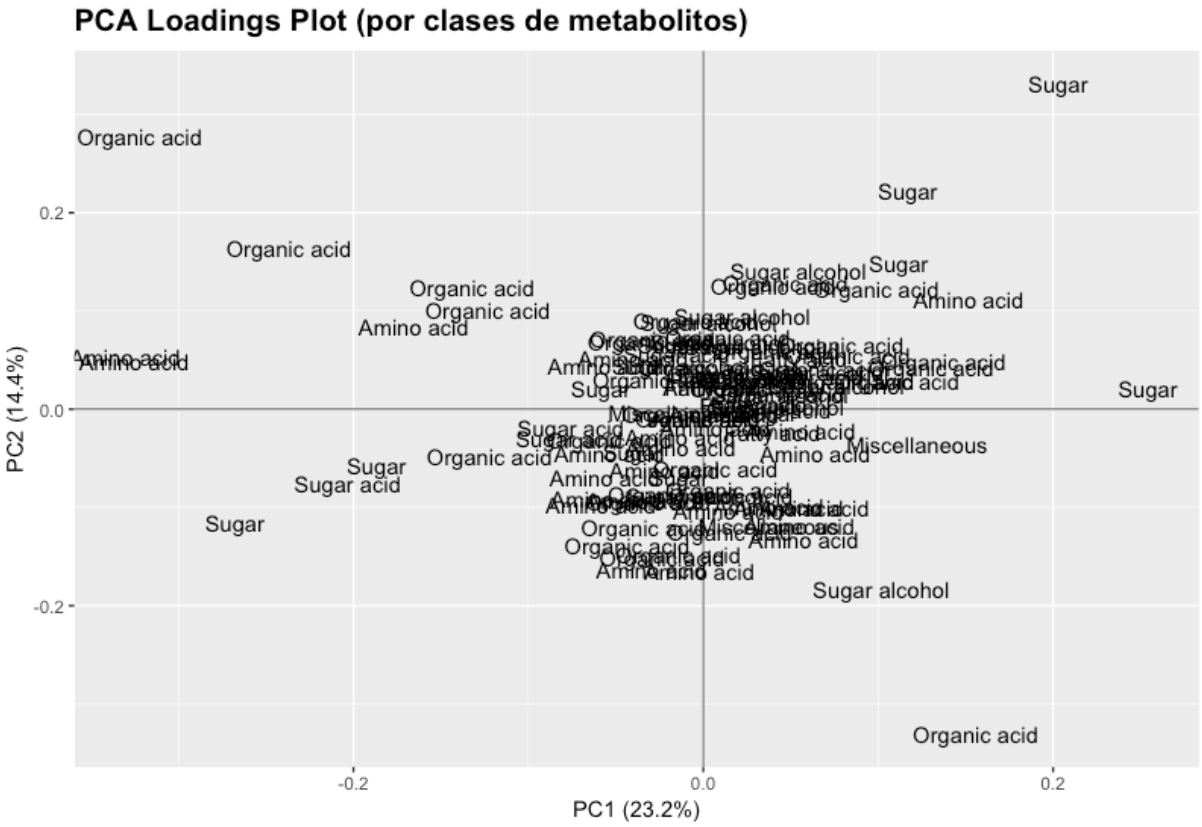

Figura 20[39]. Loadings Plot del PCA con las familias de los metabolitos

La **Figura 20** muestra el gráfico de *loadings* hecho teniendo en cuenta la clasificación anterior de los metabolitos. Como se puede observar, en realidad no hay ningún patrón, salvo el hecho de que aquellos metabolitos más significativos son o bien ácidos orgánicos, aminoácidos o azúcares. Aun así, no se puede extraer ninguna conclusión acerca de este gráfico, ya que también hay metabolitos de esas mismas clases que no son nada relevantes para la identificación de las variedades.

3.4. Heatmap

Otro método para analizar las diferencias entre variedades en base al contenido de los metabolitos es con un mapa de calor. Este tipo de gráfico muestra la cantidad de un determinado metabolito en una muestra. Los valores de la escala van de -6 (azul) a 6 (rojo). Cuanto más intenso sea el color, más relacionado está el metabolito con la muestra en cuestión. Se representan los valores de todas las réplicas hechas por cada uno de los vinos. Adicionalmente, se señala la variedad de la muestra que se está tratando con la barra superior, donde el color amarillo corresponde a los vinos Chardonnay, el color rosa a los Pinot Gris, el verde al Riesling, el coral al Sauvignon Blanc y el cian al Viognier.

[39] Fuente: elaboración propia.

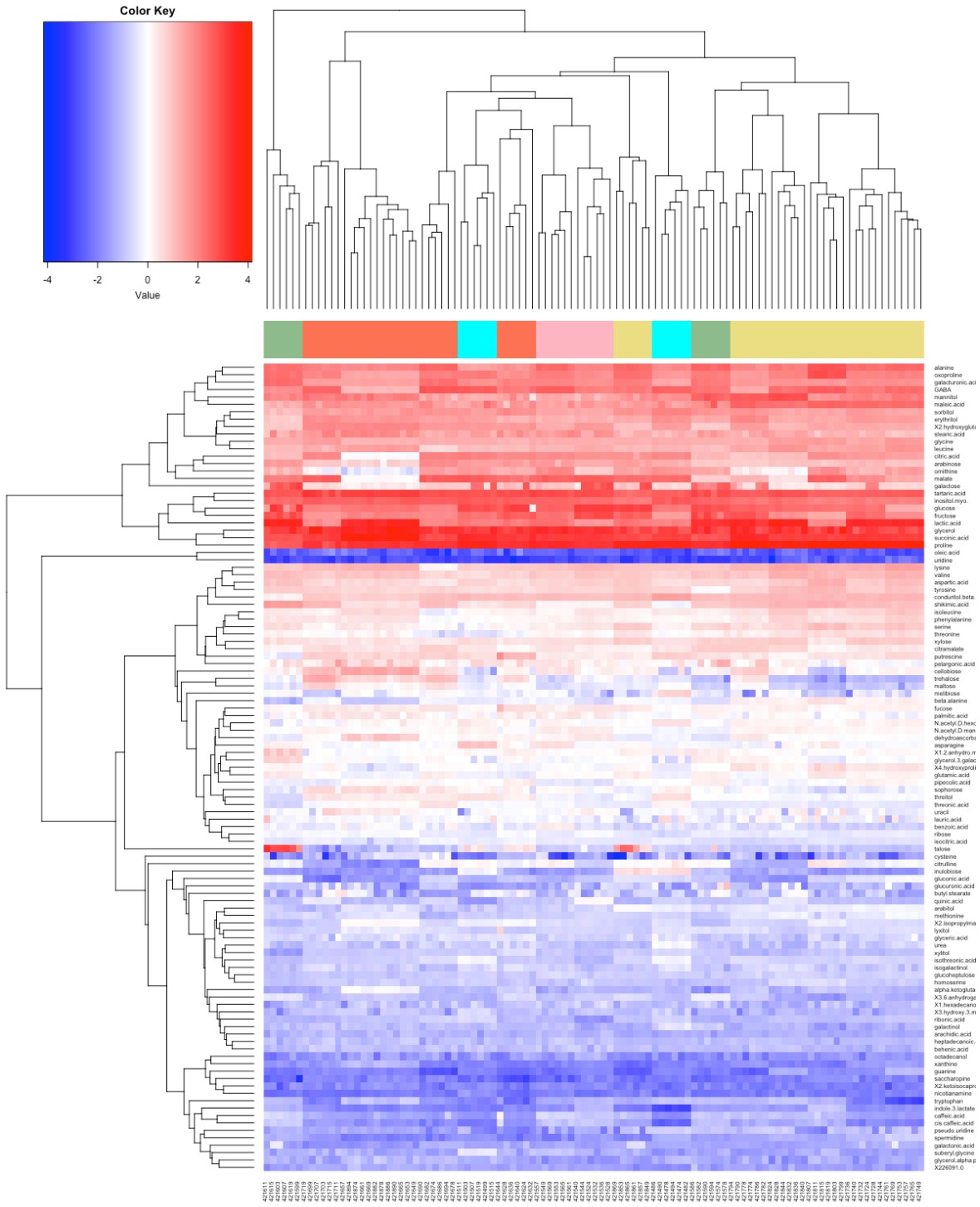

Figura 21[40]. Gráfico de calor de los metabolitos en función de las muestras de vinos

En la **Figura 21**, los metabolitos aparecen en el eje vertical del gráfico y las distintas muestras en el horizontal. Lo primero que se observa es que el método consigue juntar bastante bien casi todas las variedades, teniendo en cuenta que las agrupaciones se han realizado a partir de la distancia euclídea. En concreto, todas las muestras de Pinot

[40] Fuente: elaboración propia.

Gris han quedado completamente agrupadas en el gráfico, y las de Chardonnay y Sauvignon también han quedado muy cerca (en los dos casos tan solo un grupo pequeño de muestras ha quedado un poco más alejado del resto). Las muestras que no han quedado tan bien agrupadas por variedades son las de Viognier y, sobretodo, las de Riesling. En ambos casos se identifican dos grupos de muestras que están bastante alejados entre ellos.

Por otro lado, y en cuanto a los metabolitos, se identifican algunos metabolitos con colores muy intensos en todas las muestras, como pueden ser el *lactic acid,* el *glicerol*, el *succinic acid* y el *proline*, donde presentan un color rojo muy intenso, o el *oleic acid* y el *uridine* donde se presenta un azul muy fuerte. Además, hay otros metabolitos que presentan un color intenso tan solo en algunas de las muestras, como por ejemplo el *cysteine* o el *talose.*

Relacionándolo con el PCA, ambos coinciden en que metabolitos como el *lactic acid* y el *talose,* entre otros, están altamente relacionadas con algunas de las variedades.

3.5. PLS-DA

Mediante esta técnica de análisis supervisado, se puede analizar qué metabolitos son más responsables de la separación entre clases. En este estudio esto es muy relevante, ya que con los métodos planteados anteriormente se ha analizado qué metabolitos son más importantes para identificar cada una de las clases, pero no la separación entre ellas. Con el PLS-DA, además, se va a poder hacer un análisis discriminatorio con el objetivo de intentar predecir de qué tipo de variedad es un vino con unas características concretas. Además, y al ser una rotación de los componentes del PCA, se van a poder comparar los resultados obtenidos con ambos métodos con la finalidad de tener una visión completa de la influencia de los metabolitos en las variedades de vinos blancos estudiadas.

Lo primero que interesa de un PLS-DA son sus funciones discriminantes. En este caso, se ha decidido centrarse en las dos primeras, ya que son las que explican más, igual que pasaba con las dos primeras componentes principales del PCA.

	Primera función discriminante		Segunda función discriminante	
	METABOLITO	**COEFICIENTE**	**METABOLITO**	**COEFICIENTE**
	INTERCEPT	-0,350275369	INTERCEPT	0,73054371
	tyrosine	0,291080966	glucoheptulose	0,253491202
	inositol.myo.	-0,206440659	N.acetyl.D.mannosamine	0,222393106

	tartaric.acid	-0,169258223	threitol	-0,17469424
	fucose	0,1449101	tyrosine	-0,165816096
	serine	0,138725722	shikimic.acid	-0,144962928
	glucoheptulose	0,138510136	pipecolic.acid	-0,143720685
	xylose	0,136093355	isothreonic.acid	-0,139047701
	proline	0,13388386	homoserine	-0,129964707
	methionine	0,123447322	quinic.acid	0,10574494
	nicotianamine	0,117930993	aspartic.acid	0,10479487
	beta.alanine	0,109375316	ribose	-0,101362083
	xylitol	-0,109166584	ribonic.acid	-0,099192775
	isothreonic.acid	0,101287989	benzoic.acid	0,094514934
	isocitric.acid	-0,095391113	heptadecanoic.acid	0,092550507
	urea	-0,09206192	caffeic.acid	0,091613154
	lyxitol	-0,087305408	glycine	-0,091274714
	X4.hydroxyproline	0,086471959	tartaric.acid	0,082206887
	arabitol	0,085715421	oxoproline	-0,078579285
	galactinol	-0,084684556	threonic.acid	0,077644753
	glyceric.acid	-0,083400072	glycerol.alpha.phosphate	-0,076262024
	N.acetyl.D.hexosamine	0,074897835	octadecanol	0,075368545
	behenic.acid	0,074508364	lysine	0,07523431
	sophorose	-0,074323761	citramalate	-0,07397494
	asparagine	-0,073049618	fucose	0,071757272
	shikimic.acid	0,069895381	sophorose	0,069933536
	conduritol.beta.epoxide	0,067687625	dehydroascorbate	0,064644965
	alanine	0,065510332	serine	-0,06266065
	glycine	0,064329423	mannitol	-0,061047381
	galactonic.acid	0,061864329	maltose	-0,058486986
	guanine	-0,058661229	glycerol.3.galactoside.NIST	-0,056910456
	X1.2.anhydro.myo.inositol	-0,056550769	glutamic.acid	-0,055623467
	quinic.acid	-0,05426887	lyxitol	0,053851267
	xanthine	-0,051348941	threonine	0,051313099
	suberyl.glycine	-0,050764925	N.acetyl.D.hexosamine	0,046906003
	X2.isopropylmalic.acid	0,048310886	X1.hexadecanol	0,046888407
	glycerol.alpha.phosphate	-0,046740136	xylose	0,04677698
	spermidine	-0,04631531	asparagine	0,043802037
	dehydroascorbate	-0,045820822	galactinol	-0,04326988
	octadecanol	-0,04468803	X2.ketoisocaproic.acid	0,042049607
	X3.hydroxy.3.methylglutaric.acid	-0,044400873	fructose	0,041790443
	valine	0,042170103	cis.caffeic.acid	0,040663151
	ribonic.acid	0,041633786	xanthine	-0,039773755
	X3.6.anhydrogalactose	-0,04109213	putrescine	-0,037975745
	erythritol	0,041017234	butyl.stearate	-0,037610461

	lysine	0,03886677	oleic.acid	-0,037374279
	threonic.acid	-0,037021208	GABA	0,037283919
	uracil	-0,036048899	malate	0,03652496
	benzoic.acid	-0,035695037	arachidic.acid	0,034372375
	X2.ketoisocaproic.acid	-0,034446381	lactic.acid	-0,033553824
	maleic.acid	0,034205602	uridine	0,033492242
	X1.hexadecanol	-0,033213886	inositol.myo.	-0,03316121
	threonine	0,032988306	conduritol.beta.epoxide	0,032549788
	citrulline	-0,030655685	isogalactinol	-0,030885482
			tryptophan	0,030853297
			X4.hydroxyproline	-0,030613785
			isocitric.acid	0,030359316

Tabla 9[41]. Primera y segunda funciones discriminantes obtenidas con el PLS-DA

Los metabolitos son más discriminantes cuanto mayor sea su coeficiente en las funciones discriminantes, en valor absoluto. De este modo, la **Tabla 9** muestra los coeficientes de los distintos metabolitos de las dos primeras funciones discriminantes ordenados de mayor a menor. No obstante, solo se muestran aquellos que tienen un coeficiente mayor de 0,03 en valor absoluto. A pesar de que algunos metabolitos coinciden en los primeros puestos de ambas listas, se ve como la mayoría de ellos son más importantes en una función que en la otra.

En realidad, en la **Tabla 9** se muestra el valor numérico de los coeficientes, pero para su interpretación siempre es más sencillo verlo de forma visual. Es por eso por lo que en la siguiente figura se muestra el círculo de correlaciones de las dos primeras funciones, que son las que aparecen en la tabla superior.

[41] Fuente: elaboración propia.

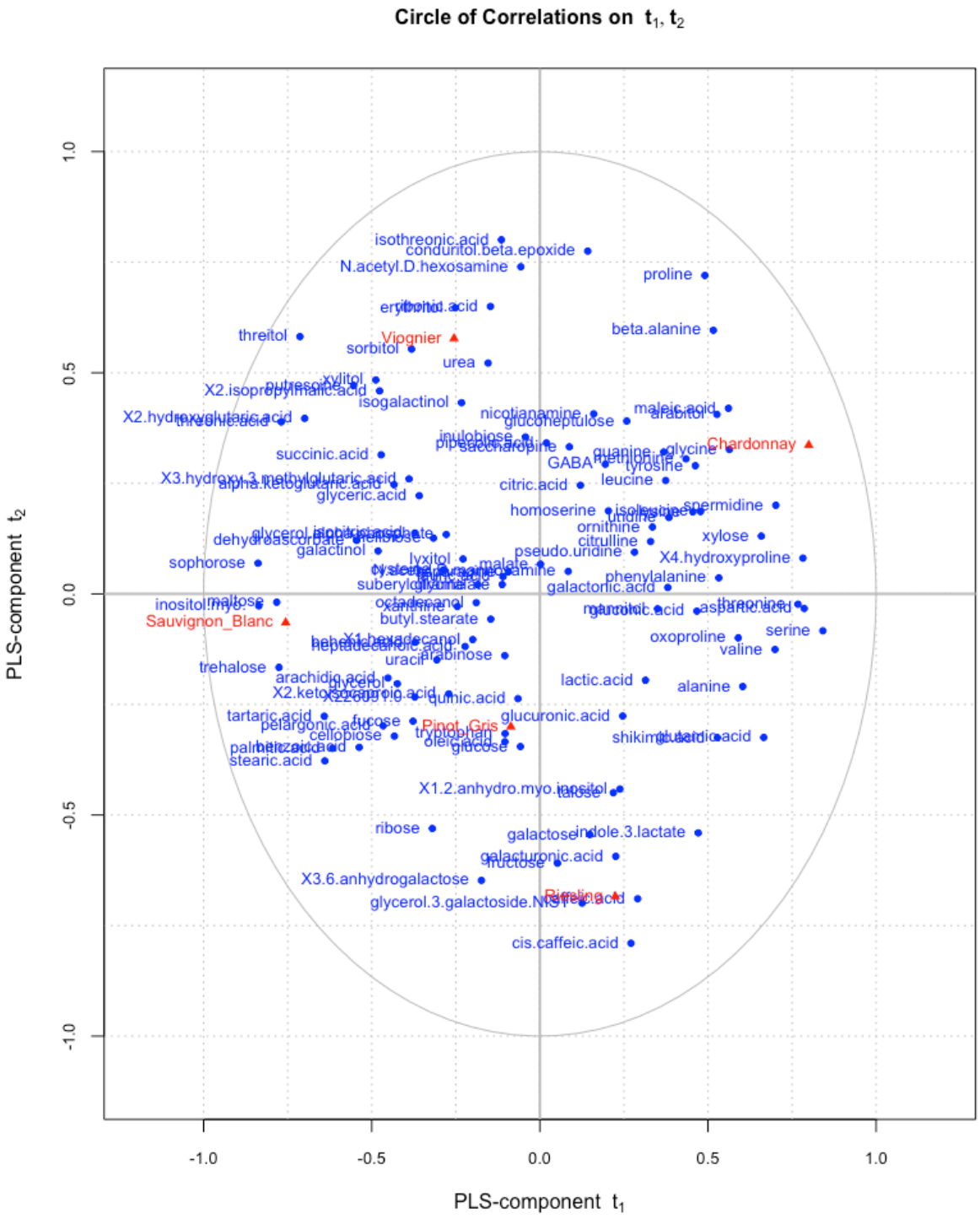

Figura 22[42]. PLS-DA de los metabolitos según las diferentes variedades de vino

Lo que se muestra en la **Figura 22**, en realidad, es un análisis de componentes principales rotado de tal manera que se obtiene una separación máxima entre clases. De hecho, el gráfico pone de manifiesto que las cinco variedades están bastante

[42] Fuente: elaboración propia.

alejadas unas de las otras. En el caso del análisis gráfico, los metabolitos discriminantes son aquellos que quedan más cerca de la línea que marca el círculo (es decir, más lejos del centro). Así pues, se observa que metabolitos como el *isotheronic acid,* el *inositol myo*, *maltose*, *threitol*, o el *serine*, son aquellos más discriminantes. De hecho, deberían de coincidir con los que tienen el coeficiente más grande en la **Tabla 9**, que muestra las funciones discriminantes. Comparando ambos métodos de análisis, se observa que es así, ya que la mayoría de los metabolitos que aparecen cerca de la línea circular también aparecen en las posiciones altas de la tabla, como por ejemplo el *inositol myo* o el *serine*, antes comentados, de la primera función, o el *threitol* de la segunda, entre otros.

Una vez hecho el análisis PLS-DA, es importante validarlo para evitar un sobreajuste de los modelos y asegurar que las clasificaciones que se obtienen con el modelo son informativas. En este caso se ha decidido que la base de entrenamiento esté formada por dos tercios de los datos (es decir, 68 de las muestras) y la base de testeo, por el tercio restante (34 muestras). La predicción ha resultado tener una precisión del 88,2%, es decir, década cien muestras, ochenta-y-ocho están identificadas dentro de la variedad que les corresponde. El intervalo de confianza de dicha precisión es de (0.7255, 0.967), lo que quiere decir que en el 95% de las predicciones que se hagan la precisión estará entre el 73% y el 97% aproximadamente.

		Referencia				
		CHARDONNAY	PINOT GRIS	RIESLING	SAUVIGNON BLANC	VIOGNIER
Predicción	CHARDONNAY	12	0	0	0	0
	PINOT GRIS	0	0	0	0	0
	RIESLING	0	1	4	0	0
	SAUVIGNON BLANC	0	3	0	10	0
	VIOGNIER	0	0	0	0	4

Tabla 10[43]. Matriz cruzada de las variedades observadas y las predichas con PLS-DA

Gracias a la **Tabla 10** se observa que con la predicción hecha todas las muestras de Chardonnay y de Viognier que formaban parte de la prueba se han identificado correctamente; en cambio, una de las muestras de Riesling y tres de las de Sauvignon Blanc se han identificado como Pinot Gris. En cuanto al Pinot Gris, a tener pocas observaciones en la base de datos ninguna de ellas ha llegado a formar parte del *test*, y es por eso por lo que no consta ninguna predicción en la fila de esta variedad.

[43] Fuente: elaboración propia.

Aun así, se comprueba que la mayoría de los vinos se han identificado correctamente, y de hecho una precisión del 88% está bastante bien, por lo que se determina que la predicción es acertada y se puede confiar en los resultados obtenidos con el PLS-DA.

3.6. Correlaciones

En el estudio de las correlaciones interesará ver qué metabolitos están más correlacionados entre ellos, pero también qué muestras se correlacionan más. De este modo, se podrán determinar relaciones tanto entre variables (metabolitos) como entre las distintas muestras de vinos.

En cuanto a las correlaciones entre los metabolitos, se espera que aquellos metabolitos que están más correlacionados se comporten de una forma similar en las técnicas de análisis implementadas anteriormente. Es importante determinar estas correlaciones porque quizá un metabolito que aparece como muy significativo hace que otro metabolito con el que está muy correlacionado sea significativo, pero quizá este metabolito por sí solo no tendría este peso dentro de la clasificación de las variedades. La matriz de correlaciones, a causa de su tamaño (se debe tener en cuenta que se dispone de 109 metabolitos, lo que representan 109 filas y 109 columnas en la matriz de correlaciones) se presenta en el Anexo[44] del trabajo.

En general, los metabolitos están muy poco correlacionados entre ellos; de hecho, la mayoría de las correlaciones obtenidas no superan el 0,5 (en valor absoluto), teniendo en cuenta que una correlación perfecta sería -1 o 1, y que el 0 significa que no hay correlación. Sin embargo, sí que se han encontrado algunos metabolitos que pueden estar correlacionados entre ellos, como por ejemplo el *X2 hydroxyglutoric acid* y el *anhydro myo inistol*, con una correlación negativa de -0,53, el *talose* y el *anhidro myo inistol*, que obtienen una correlación de 0,56, o incluso metabolitos donde la correlación es más grande aún, como el *alpha ketoglutoric acid* y el *X2 isopropylmalic acid* (0,78), el *glutamic acid* y el *alanine* (0,86) y el *leucine* con el *methionine* (0,89).

En cambio, y en cuanto a las correlaciones entre las muestras, que de nuevo se muestran en el Anexo[44], se observa que todas ellas son correlaciones positivas y muy altas; de hecho, ninguna de ellas es menor de 0,8. Esto significa que, a diferencia de los metabolitos, todas las muestras están relacionadas entre ellas, sean de la variedad que sean. Esto lleva a pensar que las diferencias entre variedades no son grandes si no se tiene en cuenta a qué metabolito pertenece cada valor.

[44] https://drive.google.com/open?id=1nbjoDt7UCsV3Gio0N-bBzivRgEP6XjwU

Para descubrir si hay suficiente evidencia basada en la muestra para concluir que las correlaciones en la población difieren de cero, es decir, probar la significancia estadística, se ha hecho una prueba de correlación con el método de Spearman, donde la hipótesis nula es que no hay relación y la alternativa que sí que la hay; de este modo, se aceptará que el coeficiente es distinto de cero si se rechaza la hipótesis nula, es decir, si el *p-value* es inferior al nivel de significación (0.05 en este caso). En el anexo se adjunta la matriz de los *p-values* para todas las correlaciones, tanto entre los metabolitos como entre las muestras de los vinos, al lado de cada una de las correlaciones.

Observando los *p-values* para las correlaciones entre metabolitos, se observa que pocos de ellos son significativos, la cual cosa quiere decir que no se puede rechazar la hipótesis nula y que por lo tanto no existen diferencias entre los dos metabolitos. Así pues, entre metabolitos hay pocas diferencias, aunque en algunos casos sí que se observan. Por ejemplo, en los casos que se han comentado anteriormente, todos ellos son significativos con un *p-value* igual a cero.

En cuanto a la significación de las correlaciones entre muestras de vinos, se observa que todas ellas son significativas con un *p*-value igual a cero y, por lo tanto, todas ellas son significativamente distintas de cero. Así pues, entre muestras de vinos se observa una correlación muy fuerte que además es significativa. Se puede decir, pues, que las diferencias entre variedades vienen muy marcadas por los metabolitos que contienen, ya que como ya se ha comentado, las muestras se relacionan mucho si no se tienen en cuenta qué metabolitos contienen en mayor o menor proporción, de forma que no se pueden observar grandes diferencias entre muestras dependiendo de su variedad.

4. Integración e identificación

En esta última etapa, y como ya se ha descrito en el capítulo anterior, se realizan dos estudios finales con un objetivo común: completar la información obtenida con en el análisis de los datos para proporcionar conclusiones más concretas y cuidadosas. De este modo, primeramente, se van a integrar las variables metabolómicas que se han estudiado hasta ahora con otras variables relacionadas con las localizaciones de cada uno de los vinos (su temperatura, precipitaciones y altitud) de forma que se va a poder relacionar el efecto de los metabolitos en la separación de las distintas variedades con el hecho de que no todos los viñedos tienen las mismas características climáticas. A continuación, con la identificación se va a proporcionar información lo más completa posible de los metabolitos que se han identificado como más significativos a la hora de separar las variedades de vinos blancos estudiadas.

4.1. Integración: análisis de factores múltiples

La integración de los datos se puede realizar con un análisis de factores múltiples, con una función implementada en R que proporciona diversos gráficos con los que extraer conclusiones. Sin embargo, antes de empezar con dicho análisis se ha tenido que buscar información sobre las nuevas variables de interés para la integración, ya que en este caso no aparecían en la base de datos con la que se está trabajando.

Región	Año	Temperatura	Precipitaciones	Altitud
Carneros	2003	24,6º	607 mm	213 m
	2004	24,3º	550 mm	213 m
Monterey	2003	29,0º	424 mm	540 m
Napa Valley	2003	20,8º	510 mm	838 m
	2004	20,4º	520 mm	838 m
Oregon	2003	20,2º	917 mm	1.005 m
Sudeste Australia	2004	20,5º	638 mm	577 m
Finger Lakes	2004	22,2º	860 mm	334 m
Lake Country	2001	19,0º	375 mm	600 m
Dunnigan Hills	2004	15,4º	762 mm	80 m

Tabla 11[45]. Temperatura anual media, precipitaciones y altitud de las distintes regiones

La **Tabla 11** muestra, para cada una de las regiones en las que se elaboran los vinos estudiados, la temperatura anual media y las precipitaciones totales del año del vino, además de la altitud en la que cada una de ellas se sitúa. Vale la pena remarcar que los datos expresados en la tabla no tienen por qué ser reales; aunque se ha buscado una aproximación lo más cuidadosa posible, no en todos los casos ha sido posible encontrar los datos de temperatura y precipitaciones reales para el año concreto que se está estudiando. Además, se debe tener en cuenta que hay vinos que pertenecen a la misma región e incluso que se elaboraron el mismo año; obviamente, los datos de temperatura, precipitaciones y altitud serán iguales para todos ellos.

[45] Fuente: elaboración propia con datos de Climate-Data.org

Una vez concretados estos datos, y añadidos a la base de datos con la que se estaba trabajando (que contiene todos los metabolitos y la variedad a la que corresponde cada una de las muestras) se puede proceder a la realización del análisis de factores múltiples.

Se ha implementado la función *MFA* en R creando tres grupos de variables: un primero que tan solo contiene la variable que indica la variedad de uva a la que corresponde la muestra; un segundo que contiene las características de la localización y año donde se elaboró el vino (temperatura media anual, precipitaciones totales y altitud); y una tercera con todos los metabolitos que estaban en la base de datos inicial. La variedad de uva se ha determinado que actúe como variable suplementaria, la cual cosa significa que este grupo no participará en la construcción de las dimensiones. Además, se ha decidido trabajar con tres dimensiones, ya que las tres primeras explican aproximadamente un 23%, un 16% y un 13% de la variancia, respectivamente, mientras que la cuarta tan solo explica el 9%, por lo que se ha determinado que las tres primeras dimensiones ya son suficientes para el análisis, ya que para explicar un porcentaje superior de variancia se necesitarían un número muy grande de dimensiones, lo que dificultaría la comprensión de los resultados, así que es mejor conformarse con un 51,4% de variancia acumulada.

Haciendo un resumen de los resultados se obtiene la información numérica que después se plasmará en los gráficos: el porcentaje de variancia que explica cada una de las dimensiones; los resultados de los grupos de variables (tanto de los activos como del suplementario) con la coordenada en las dimensiones, la contribución de los grupos en la construcción de las dimensiones y la calidad de representación explicada con el coseno al cuadrado; los resultados para los individuos, con los mismos datos que para los grupos; los resultados para las variables numéricas (en este caso solo para las activas, ya que no se tenían suplementarias numéricas, pero en caso de haberlas tenido también aparecerían); y los resultados para las variables categóricas, que en este caso son todas suplementarias, proporcionando también en este caso el valor de un test que toma valores entre -2 y 2 si la coordinada de la variable no es significativamente distinta de cero.

El resumen completo se muestra en el Anexo[46], aunque se van a comentar los puntos más interesantes. Como ya se ha dicho, entre las tres primeras dimensiones se llega a explicar el 51,4% de la variancia. Por otro lado, respecto los grupos activos, en la primera dimensión ambos contribuyen más o menos igual en la construcción de la dimensión, pero en la segunda contribuye un poco más el grupo de las características y

[46] https://drive.google.com/drive/folders/1nbjoDt7UCsV3Gio0N-bBzivRgEP6XjwU

en la tercera, el de los metabolitos. La calidad de la representación no es especialmente buena en ninguno de los casos. En cuanto a la variable de las variedades, que es suplementaria, no aparece la contribución porque es cero, ya que como ya se ha comentado no participa en la construcción de las dimensiones. La calidad de la representación de esta variable también es pequeña como en los casos anteriores. No se comentarán los resultados para los individuos y las variables, ya que se pueden revisar en el **Anexo** y se pueden explicar de forma más sencilla y visual con los gráficos que se van a realizar. Tan solo se va a poner énfasis en la significación de la variable categórica, ya que se observa que en la mayor parte de los casos la variedad no es significativa; de hecho, tan solo son significativamente iguales a cero las coordenadas del Riesling en la primera dimensión, del Pinot Gris en la segunda y las coordenadas del Chardonnay y del Sauvignon Blanc en la tercera.

Para empezar, se van a estudiar las dos primeras dimensiones, y luego se estudiarán la primera y tercera dimensiones.

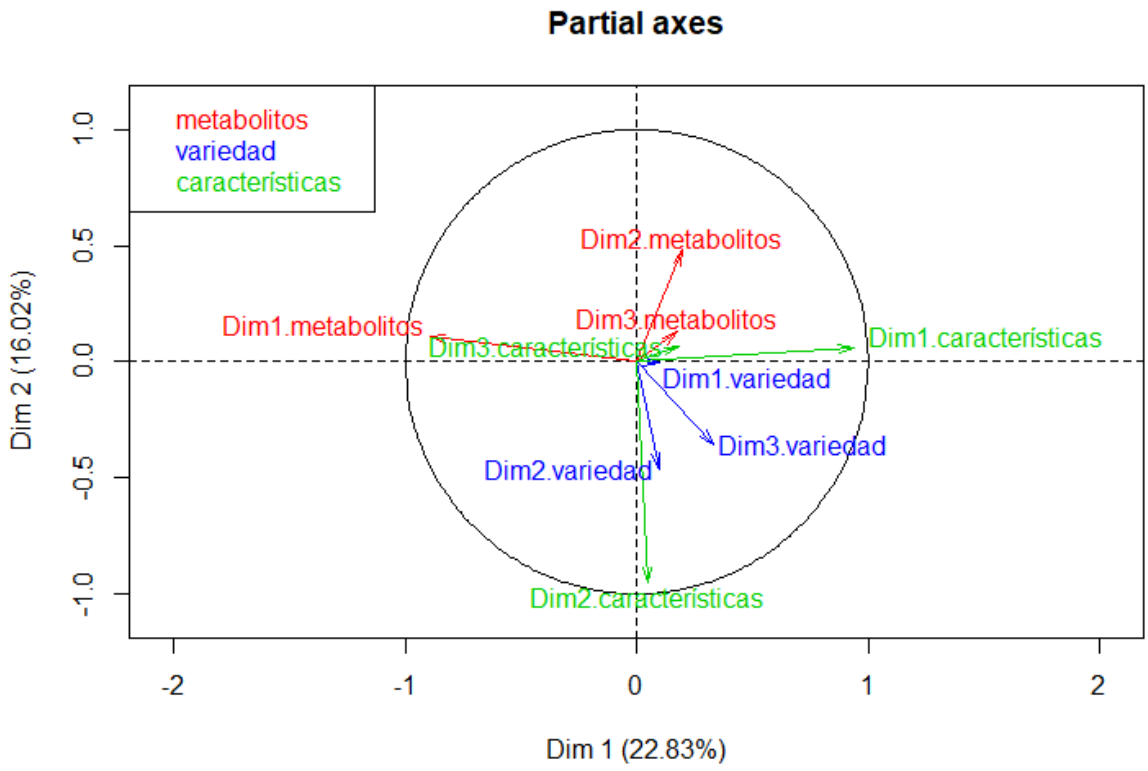

Figura 23[47]. Gráfico de los ejes parciales de la primera y segunda dimensiones del MFA

El primer gráfico relevante que se puede obtener con el análisis de factores múltiples es el que muestra los ejes parciales, es decir, las dimensiones parciales de cada análisis, que se muestra en la **Figura 23**. Para cada grupo de variables se realiza un

[47] Fuente: elaboración propia.

análisis: para las variables numéricas se hace un análisis PCA y se proyectan las dimensiones del PCA como información suplementaria en el MFA. Por ejemplo, se observa que la primera dimensión de las características está muy relacionada con la primera dimensión del MFA, y en cambio la segunda dimensión de este mismo grupo está muy poco relacionada con la segunda dimensión del MFA. Además, la segunda dimensión de los metabolitos aparece relacionada con la segunda dimensión del MFA. Para las variables categóricas se hace un análisis MCA (análisis de correspondencias múltiples), así que se proyectan las dimensiones del MCA en el MFA. En este caso, pero, y como la variedad actúa como variable suplementaria, no parece que ninguna de sus dimensiones tenga relación con el MFA. Además, la tercera dimensión de las características y de los metabolitos están muy cerca del origen, por lo que no tienen relación alguna con ninguna de las dimensiones del MFA.

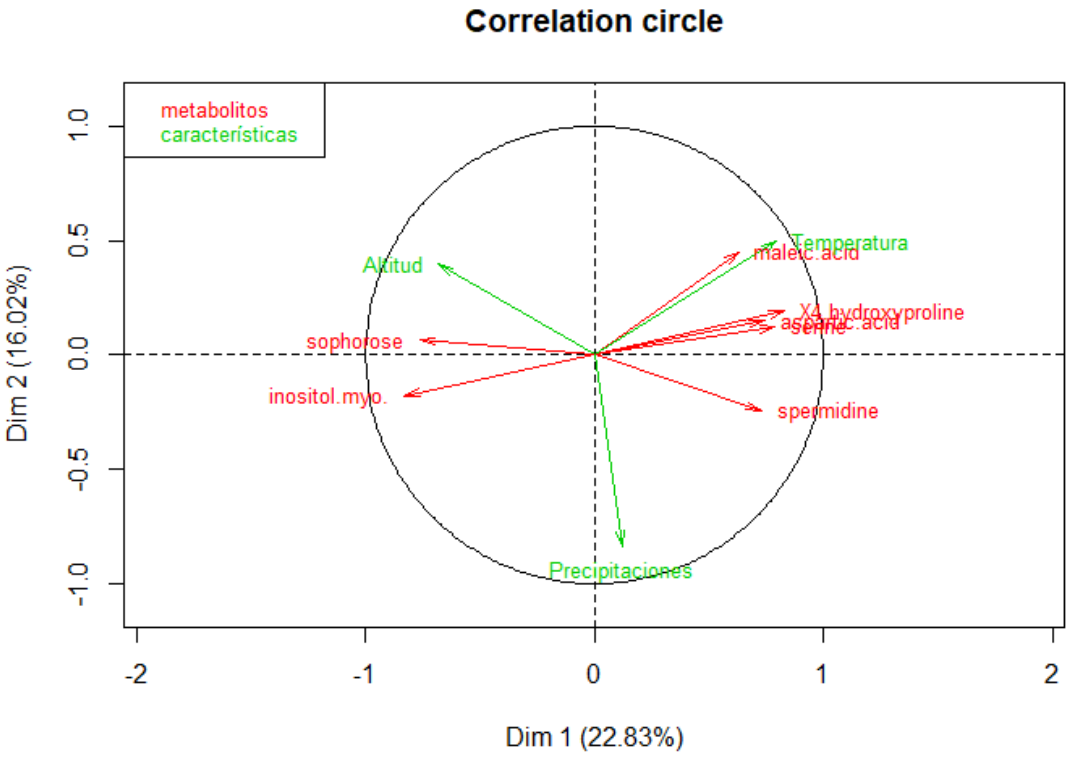

Figura 24[48]. Gráfico de las variables en la primera y segunda dimensiones del MFA

En la **Figura 24** se muestran las diez variables que más contribuyen en la creación de las dimensiones. Se ha decidido graficar tan solo las diez primeras por razones de visualización y porque son las que más información proporcionan. Por un lado, respecto la primera dimensión, se observa que las variedades más relacionadas con ella son la temperatura, el *serine*, el *X4 hydroxyproline* y el *aspartaric acid*, aunque

[48] Fuente: elaboración propia.

también se relaciona bastante con el *maleic acid* y la temperatura. Así pues, los vinos que aparezcan relacionadas con esta dimensión tendrán altos niveles de estas variables y las localizaciones donde se producen, altas temperaturas. Por otro lado, no parece que haya variables muy relacionadas con la segunda dimensión, aunque sí que se puede decir que las precipitaciones están muy poco relacionadas con ella.

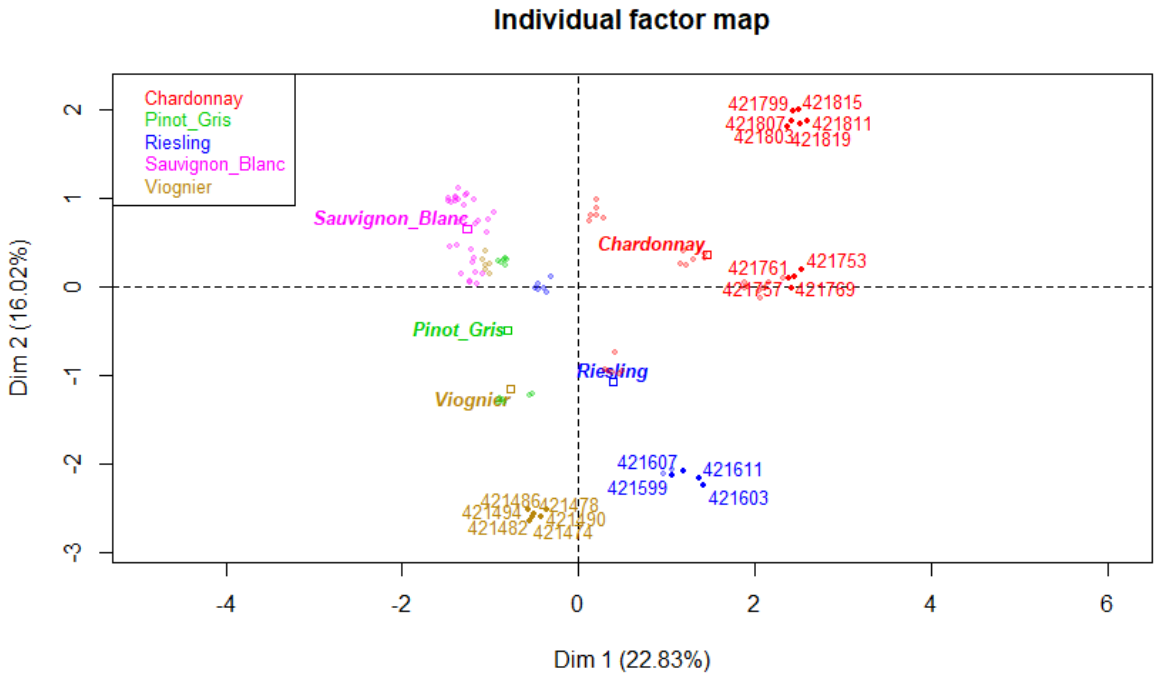

Figura 25[49]. Gráfico de los individuos en la primera y segunda dimensiones del MFA

En la **Figura 25**, por otro lado, se han graficado los individuos (en este caso, las muestras de vino) en las dos primeras dimensiones del análisis de factores múltiples. Para hacer el gráfico más visual, se ha decidido etiquetar tan solo las veinte muestras que más contribuyen en las dimensiones, mientras del resto tan solo aparecen los puntos donde se ubican. Además, se han coloreado las muestras en función de la variedad a la que pertenecen. Se observa, pues, que aquellas muestras que más contribuyen en la formación de las dos primeras dimensiones forman parte de las variedades Chardonnay o Riesling. Todas las muestras se localizan en función de la variedad a la que pertenecen, es decir, aquellas muestras de una misma variedad se agrupan muy bien en el gráfico, igual que pasó en el PCA.

En este caso se puede ver como las muestras de la variedad Chardonnay están muy relacionadas con la primera dimensión, por lo que serán vinos con altos niveles de *serine*, *X4 hydroxyproline* y *aspartaric acid* y además las temperaturas del lugar donde

[49] Fuente: elaboración propia.

se elaboran son altas. En cambio, no se puede relacionar ninguna de las variedades con la segunda dimensión.

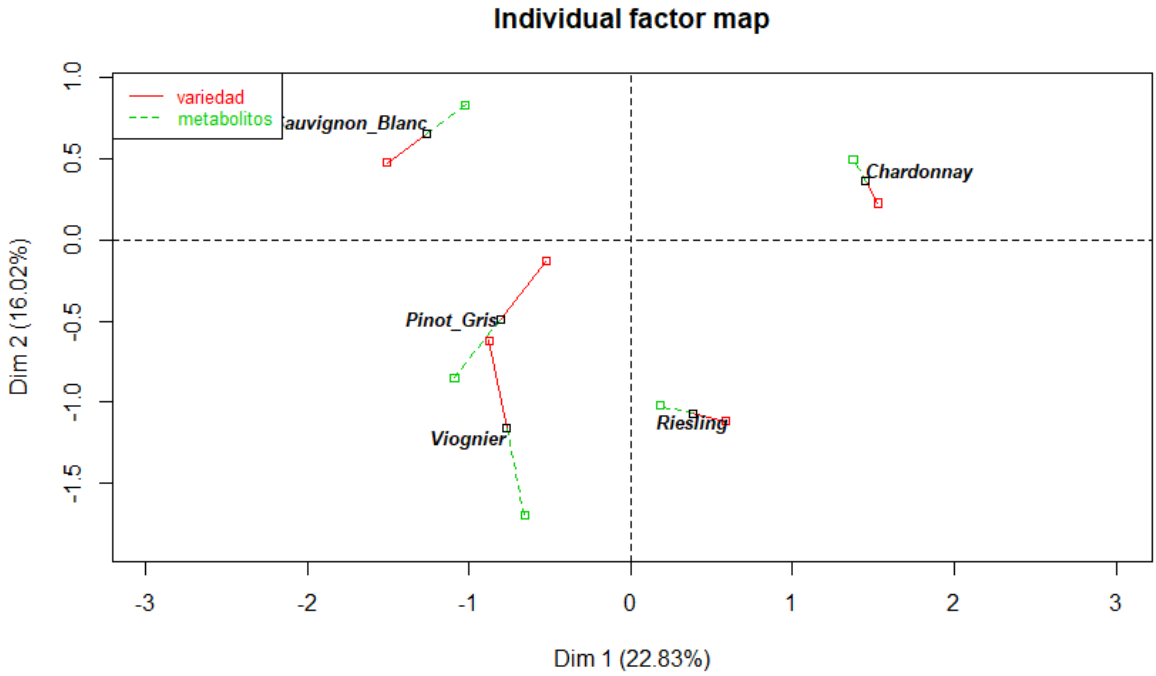

Figura 26[50]. Puntos parciales para las variedades en la primera y segunda dimensiones del MFA

Por último, en la **Figura 26** se muestran las distintas variedades de vino con dos puntos parciales, uno para cada grupo de variables que se ha determinado en el análisis (características y metabolitos). En este gráfico se muestra el baricentro para cada una de las variedades, es decir, el punto en el que está realmente la variedad en relación con la primera dimensión y la segunda; además, se muestran los puntos parciales, que muestran dónde quedaría cada variedad si solo se tuvieran en cuenta las variables del grupo de las características (puntos rojos) o de los metabolitos (puntos verdes). Esto permite saber en qué variedades la inercia es más grande, es decir, que se ven de formas muy distintas en función de los grupos, y en cuales más pequeña. Así pues, en el Chardonnay y el Riesling las diferencias son pequeñas, ya que los tres puntos están muy juntos. En cambio, en el Viognier y el Pinot Gris los puntos están bastante alejados, lo que quiere decir que hay diferencias en las variedades de un grupo de variables a otro; lo mismo pasa en el Sauvignon Blanc, aunque quizás en este caso no están tan alejados.

[50] Fuente: elaboración propia.

A continuación, y teniendo en cuenta que parece que la tercera dimensión juega un papel importante, se van a realizar los mismos gráficos para la primera y tercera dimensiones.

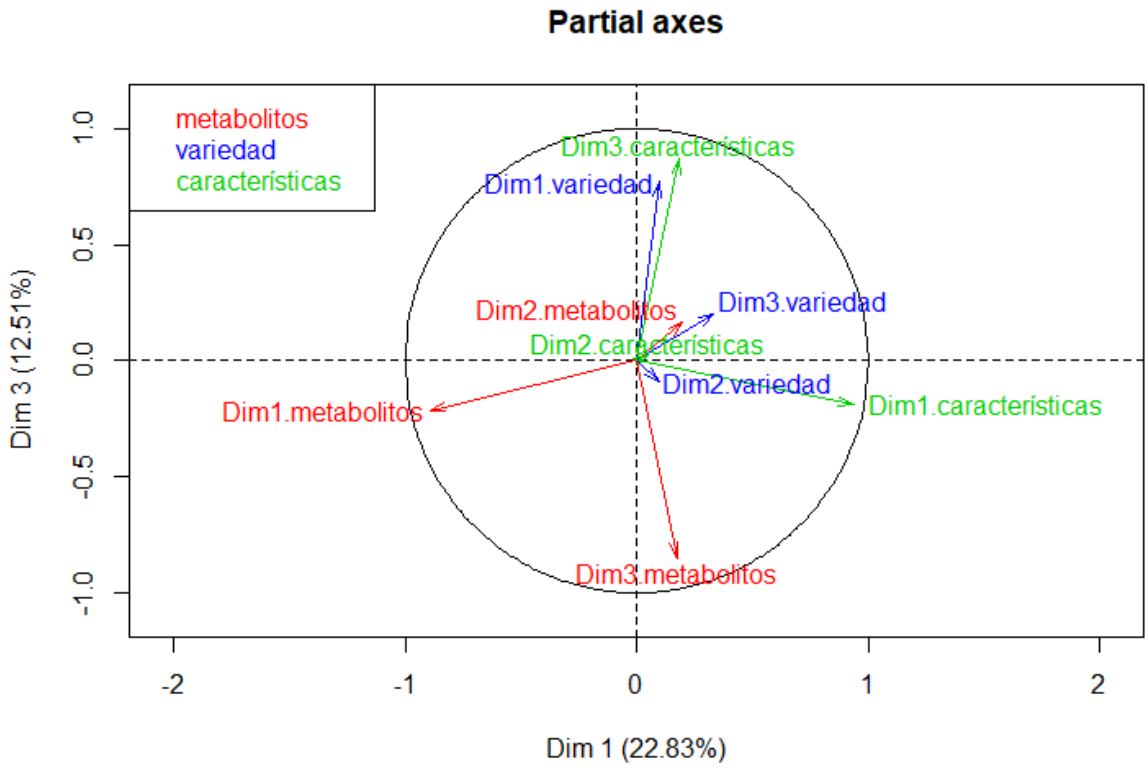

Figura 27[51]. Gráfico de los ejes parciales de la primera y tercera dimensiones del MFA

En este caso, de nuevo se observa como la primera dimensión de las características está relacionada con la primera dimensión del MFA; además, la tercera dimensión de las características en el PCA y la primera de las variedades en el MCA están relacionadas con la tercera dimensión del MFA. En cambio, la primera y tercera dimensiones de los metabolitos están poco relacionadas con la primera dimensión y la tercera del MFA, respectivamente.

[51] Fuente: elaboración propia.

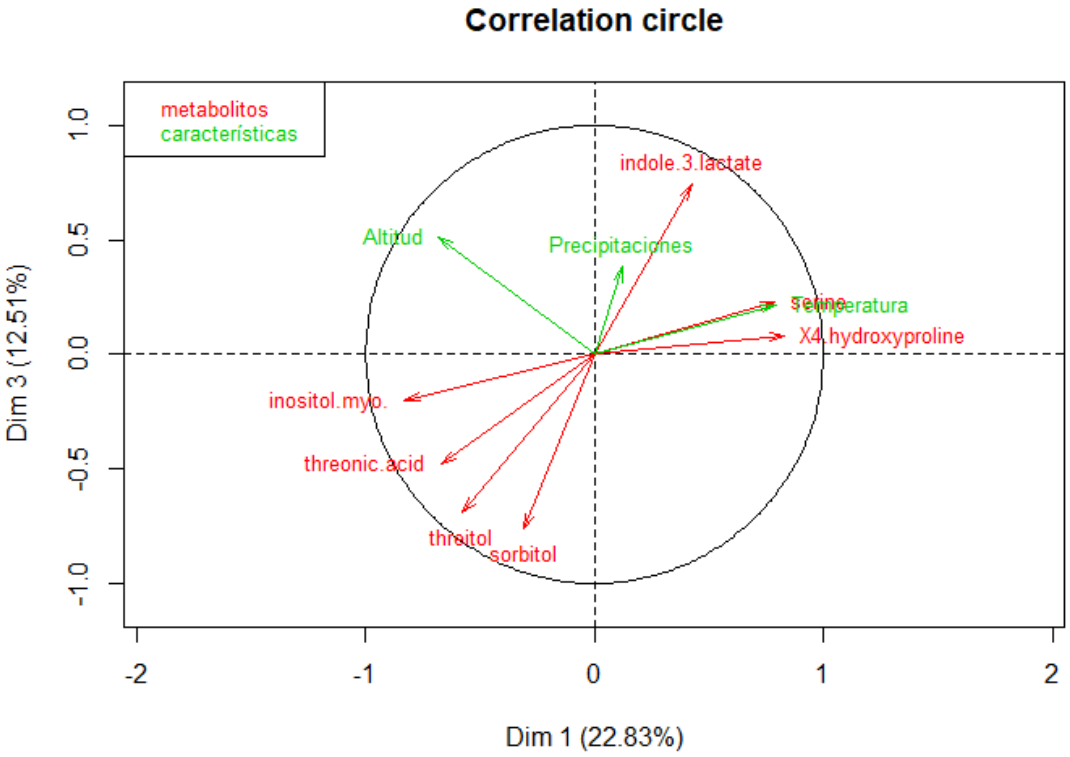

Figura 28[52]. Gráfico de las variables en la primera y tercera dimensiones del MFA

En este caso, se observa como el *serine, X4 hydroxyproline* y la temperatura están relacionados con la primera dimensión (como ya pasaba en el gráfico que mostraba las dos primeras dimensiones), y que el *indole 3 lactate* está muy relacionado con la tercera dimensión, así como las precipitaciones. Recordemos que tan solo se muestran las diez variables que más peso tienen en las dimensiones.

[52] Fuente: Elaboración propia.

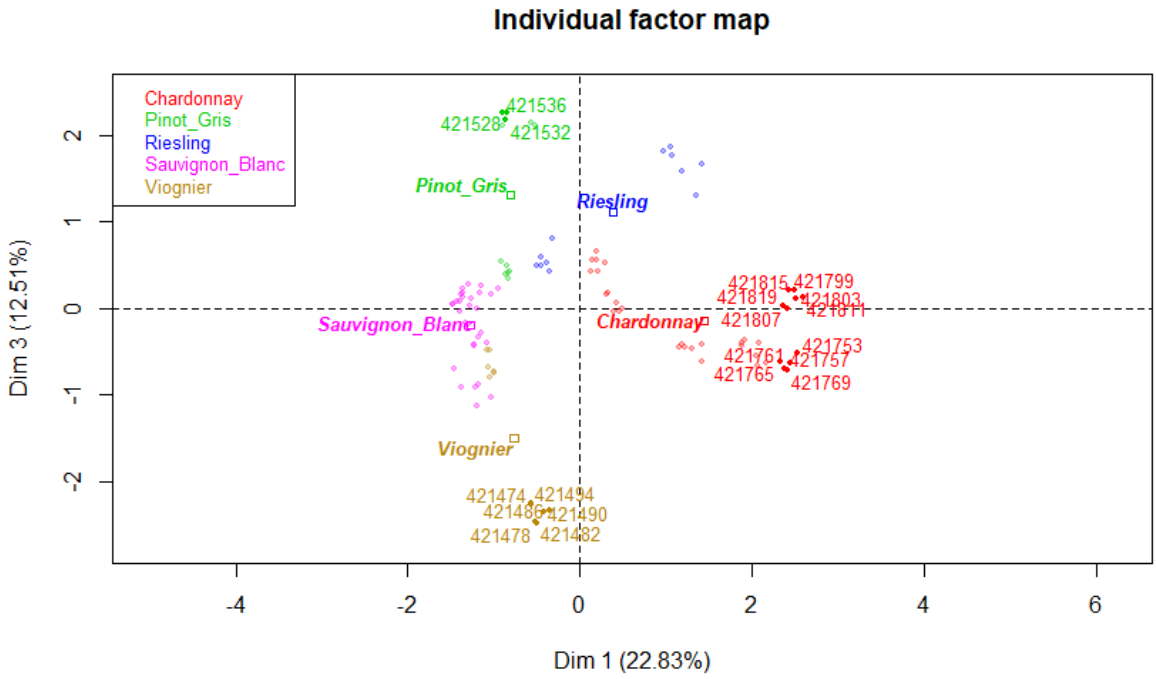

Figura 29[53]. Gráfico de los individuos en la primera y tercera dimensiones del MFA

De nuevo se observa la relación entre la primera dimensión y el Chardonnay, pero en este caso se muestra la relación del Riesling y el Pinot Gris con la tercera dimensión. Por lo tanto, ambas variedades tienen altos niveles de *indole 3 lactate,* y en los lugares donde se producen caen bastantes precipitaciones.

[53] Fuente: elaboración propia.

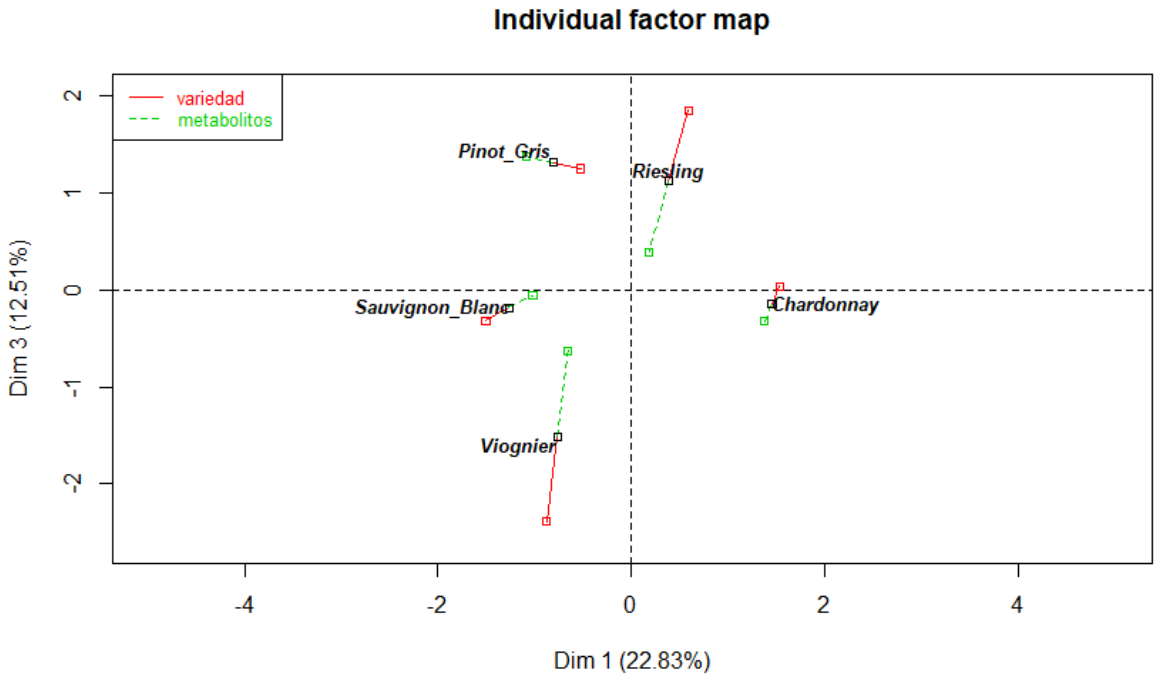

Figura 30[54]. Puntos parciales para las variedades en la primera y segunda dimensiones del MFA

En este caso se observa como las diferencias en función de las variables que se tienen en cuenta son pequeñas en el Pinot Gris, el Sauvingon Blanc y el Chardonnay, pero en cambio son muy grandes en el Riesling y, sobretodo, en el Viognier.

Con todo esto, podemos concluir que, a pesar de que los metabolitos juegan un papel importante en la separación de variedades de vinos blancos, si se tienen en cuenta otras variables como pueden ser las precipitaciones y la temperatura estas diferencias se hacen más notables, ya que estos factores son importantes para separar los vinos de distintas variedades.

4.2. Identificación de metabolitos

Como ya se ha comentado, la identificación de metabolitos consiste en dar más información sobre aquellos metabolitos que han aparecido como significativos en el estudio para separar las distintas variedades de vinos blancos en estudio. En la prueba ANOVA se ha visto que la mayoría de ellos son significativamente distintos en función de las variedades (de hecho, tan solo catorce de los ciento-nueve metabolitos estudiados no lo son), pero en cambio se ha visto que hay pocos metabolitos que tengan relación entre ellos, aunque gracias al análisis de componentes principales, el gráfico de calor y el PLS-DA se han podido concretar algunos de los metabolitos que

[54] Fuente: elaboración propia.

más contribuyen en las variedades o en la separación de dichas variedades, como por ejemplo el *malate*, el *talose*, el *lactic acid*, el *proline* o el *inositol myo*, entre otros.

La información necesaria sobre todos los metabolitos estudiados se puede encontrar en PubChem, un portal web donde, buscando el elemento que interesa, se obtiene una descripción general e información acerca de su estructura, sus nombres e identificadores, sinónimos del mismo metabolito, propiedades físicas y químicas o información espectral, entre otros.

CONCLUSIONES

En este trabajo se ha desarrollado un *pipeline* para analizar datos metabólicos mediante el *software* R con la intención de proporcionar una base comprensible y aplicable a cualquier conjunto de datos metabolómicos para hacer más sencillo su análisis. Adicionalmente, se ha verificado su funcionamiento en una base de datos obtenida de la red, con dos objetivos principales: el de comprobar que las técnicas y métodos propuestos en la implementación del *pipeline* son adecuados y sin errores, además de proporcionar ejemplos para complementar la información ya proporcionada, y el de describir diferentes variedades de vino blanco des del punto de vista metabolómico y poder identificarlas en función de los metabolomas encontrados en un vino.

En cuanto a la implementación del *pipeline*, se ha conseguido hacer un recopilatorio fácil de comprender y accesible a usuarios no tan expertos en R sobre las distintas técnicas que se pueden utilizar para analizar este tipo de datos. Además, se trata de un procedimiento estándar, de forma que se puede aplicar a cualquier base de datos metabolómicos.

Por otro lado, y en cuanto a la verificación de dicho *pipeline*, se ha confirmado que las técnicas explicadas en la implementación responden al objetivo que busca el usuario al realizarlas. Se debe tener en cuenta que no todas las técnicas son convenientes para todos los datos, ya que su elección depende del objetivo del estudio, y es por ello por lo que algunas de las técnicas no se han verificado.

Por último, se ha podido determinar que los distintos metabolitos de los vinos blancos aparecen de forma distinta según la variedad de uva, por lo que se pueden describir dichas variedades en función de sus metabolitos, además de poder identificar de forma bastante precisa de qué variedad será la uva de un vino tan solo mirando los metabolitos que lo componen.

BIBLIOGRAFIA

Breve guía de genómica. A: *National Human Genome Research Institute* [en línea]. [Consulta: 14 de abril de 2019]. Disponible a: < https://www.genome.gov/es/about-genomics/fact-sheets/Breve-guia-de-genomica>

Brown, Marie, Dunn, Warwick B., Ellis, David I., y compañía. A metabolome pipeline: from concept to data to knowledge. *Metabolomics.* Enero 2005, Vol. 1, No. 1. DOI: 10.1007/s11306-005-1106-4.

Cambiaghi, Alica; Ferrario, Manuela y Masseroli, Marco. Analysis of metabolomic data: tools, corrent strategies and future challenges for omics data Integration. *Briefings in Bioinformatics*. Mayo 2017, vol. 18, issue 3, p. 498-510. ISSN 1477-4054. [Consulta: 7 de febrero de 2019]. Disponible en: < https://academic.oup.com/bib/article/18/3/498/2453286 >.

Colmenarez, Luis; Villavicencio, Btglaudia y Rivero, Yliana. Las ciencias ósmicas Genómica, Metagenómica, Transcriptómica, Metabolómica y Genómica [en línea]. Publicado en 7 de diciembre de 2012. [Consulta: 21 de abril de 2019]. Disponible en: < http://lascienciaosmicas.blogspot.com >.

Connor R., Tiffany. Omu, a Metabolomics Analysis R Package. A: *cran-r-project* [en línea]. Publicado en 21 de marzo de 2018. [Consulta: 20 de mayo de 2019]. Disponible en: <https://cran.r-project.org/web/packages/omu/vignettes/Omu_vignette.html>.

Díaz Gimeno, Patricia. Transcriptómica y Genes. A: *Cinabrio Blog* [en línea]. Publicado en 31 de diciembre de 2012. [Consulta: 20 de abril de 2019]. Disponible en: < http://cinabrio.over-blog.es/article-transcriptomica-y-genes-113948467.html >.

Diccionario de cáncer: Proteína. A: *Instituto Nacional del Cáncer* [en línea]. [Consulta: 20 de abril de 2019]. Disponible a: < https://www.cancer.gov/espanol/publicaciones/diccionario/def/proteina>.

Diccionario de cáncer: Transcriptómica. A: *Instituto Nacional del Cáncer* [en línea]. [Consulta: 20 de abril de 2019]. Disponible en: < https://www.cancer.gov/espanol/publicaciones/diccionario/def/transcriptomica >.

El futuro de la biotecnologia. A: *Apuntes de Biotecnología* [en línea]. Publicado en 1 de julio de 2015. [Consulta: 20 de abril de 2019]. Disponible en: < http://apuntesbiotecnologiageneral.blogspot.com >.

Espectroscopía de Resonancia Magnética Nuclear [en línea]. [Consulta: 30 de abril de 2019]. Disponible en: < https://pslc.ws/spanish/nmr.htm >.

Fenotipo. A: *EcuRed* [en línia]. [Consulta: 22 de abril de 2019]. Disponible en: < https://www.ecured.cu/Fenotipo >.

Fernández-Albert, Francesc; Llorach, Rafael; Andrés-Lacueva, Cristina y Perera, Alexandre.An R package to process LC/MS metabolomic data: MAIT (Metabolite Automatic Identification Toolkit). A: *Bioconductor* [en línea]. Publicado en 2 de mayo de 2019. [Consulta: 20 de mayo de 2019]. Disponible en: < https://www.bioconductor.org/packages/release/bioc/vignettes/MAIT/inst/doc/MAIT_Vignette.pdf >.

Floirán Joseph, Diego. Las proteínas: Estructura y función. A: *science Bites* [en línea]. Publicado en 9 de abril de 2016. [Consulta: 20 de abril de 2019]. Disponible a: < https://sciencebitesperu.weebly.com/science-bites/las-proteinas-estructura-y-funcion>.

Gas Chromatography coupled with Mass Spectrometry (GC/MS). A: *SKZ* [en línea]. [Consulta: 30 de abril de 2019]. Disponible en: < https://www.skz.de/en/research/technicalfacilities/pruefverfahren1/spektroskopie1/4870.Gas-Chromatography-coupled-with-Mass-Spectrometry-GCMS.html >.

Gaude, Edoardo; Chignola, Francesca; Spiliotopoulos, Dimitrios y compañía. Muma, an R package for metabolomics univariate and multivariate statistical analysis. *Current Metabolomics* [en línea]. 2013, Vol. 1, Issue 2. DOI: 10.2174/2213235X11301020005. [Consulta: 20 de mayo de 2019]. Disponible en: < http://www.eurekaselect.com/107837>.

GC-MS. A: *Universidad Rey Juan Carlos* [en línea]. [Consulta: 30 de abril de 2019]. Disponible en: < http://www.labte.es/index.php/es/2013-11-03-19-54-23/ensayos-mediante-gc-ms >.

Genómica, transcriptómica y proteómica. A: *Apuntes de Biotecnología* [en línea]. Publicado en 7 de marzo de 2015. [Consulta: 21 de abril de 2019]. Disponible en: < http://apuntesbiotecnologiageneral.blogspot.com/2015/03/genomica-transcriptomica-y-proteomica.html >.

Grace, Stephen C. Y Hudson, Dane A. Processing and Visualization of Metabolomics Data Using R. *Metabolomics – Fundamentals and Applications*. Diciembre 2016, cap. 4. DOI: 10.5772/65405. [Consulta: 7 de febrero de 2019]. Disponible en: <

https://www.intechopen.com/books/metabolomics-fundamentals-and-applications/processing-and-visualization-of-metabolomics-data-using-r >.

Introdución ao RMN. A: *Universidad de Santiago de Compostela* [en línea]. [Consulta: 30 de abril de 2019]. Disponible en: < http://www.usc.es/gl/investigacion/riaidt/rm/rmn/introducion.html >.

Liquid chromatography-mass spectrometry. A: *Wikipedia* [en línea]. Editado en 17 de junio de 2019. [Consulta: 30 de abril de 2019]. Disponible en: < https://en.wikipedia.org/wiki/Liquid_chromatography–mass_spectrometry#Interfaces >.

Metaboloma. A: *Boletin agrario* [en línea]. [Consulta: 20 de abril de 2019]. Disponible en: < https://boletinagrario.com/ap-6,metaboloma,3272.html >.

Mock, Andreas; Warta, Rolf; Dettling, Steffen; Brors, Benedikt; Jäger, Dirk y Herold-Mende Christel. MetaboDiff: an R package for differential metabolomic analysis. A: *Oxford Academy – Bioinformatics* [en línea]. Publicado en 26 de abril de 2018. [Consulta: 20 de mayo de 2019]. Disponible en: <https://academic.oup.com/bioinformatics/article/34/19/3417/4987147>.

Morán, Alberto. ADN, genes, cromosomas… A: *Dciencia para todos* [en línea]. Publicado el 15 de enero de 2013. [Consulta: 14 de abril de 2019]. Disponible a: < http://www.dciencia.es/adn-genes-cromosomas/>

Nucleótidos. A: *Acidosnucleicos* [en línea]. [Consulta: 7 de abril de 2019]. Disponible a: < https://www.acidosnucleicos.net/nucleotidos/>

Pérez Porto, Julián y Gardey, Ana. Definición de Nucleótido. A: *Definición.de* [en línea]. Publicado en 2012. Actualizado en 2015. [Consulta: 7 de abril de 2019]. Disponible a: <https://definicion.de/nucleotido/>.

Pérez Porto, Julián. Definición de Fenotipo. A: *DefinicionDe* [en línea]. Publicado en 2017. [Consulta: 25 de abril de 2019]. Disponible en: < https://definicion.de/fenotipo/ >.

Pérez Porto, Julián. Definición de Glúcidos. A: *Definición.de* [en línea]. Publicado en 2018 [Consulta: 7 de abril de 2019]. Disponible a: <https://definicion.de/glucidos/>.

Pérez, Guillermo. Espectrometría de resonancia magnética nuclear. A: *espectrometria.com* [en linea]. [Consulta: 30 de abril de 2019]. Disponible en: < https://www.espectrometria.com/espectrometra_de_resonancia_magntica_nuclear >.

Perez, Guillermo. Fenotipo. A: *Fenotipo* [en línea]. [Consulta: 22 de abril de 2019]. Disponible en: < https://www.fenotipo.com >.

Picart-Armada, Sergio; Fernández-Albert, Francesc; Vinaixa, Maria; Yanes, Oscar y Perera-Lluna, Alexandre. FELLA: an R package to enrich metabolomics data. A: *Bioconductor* [en línea]. Publicado en 2 de mayo de 2019. [Consulta: 20 de mayo de 2019]. Disponible en: <https://bioconductor.org/packages/devel/bioc/vignettes/FELLA/inst/doc/FELLA.pdf>.

Quintela O., Cruz A., Concheiro M., De Castro A. Y López-Rivadulla M. Metodología LC-MS. Aspectos generales de la técnica y sus aplicaciones en el campo de la toxicología [En línea]. Artículo recibido en 16 de agosto de 2004 y aceptado en 15 de septiembre de 2004. [Consulta: 30 de abril de 2019]. Disponible en: < http://www.redalyc.org/html/919/91922102/ >.

Real Academia Española, A: *RAE* [en línea]. [Consulta: 7 de abril de 2019]. Disponible a: <https://dle.rae.es/?id=5Ykv4ay>.

Serrano, Merche y Vilaseca, Antonia. Del gen a la proteína. A: *Guía metabólica* [en línea]. Publicado en 11 de septiembre de 2012. Actualizado en 03 de junio de 2016. [Consulta: 20 de abril de 2019]. Disponible en: < https://metabolicas.sjdhospitalbarcelona.org/noticia/gen-proteina-0>.

Síntesis de proteïnas. A: *EcuRed* [en línia]. [Consulta: 20 de abril de 2019]. Disponible en: < https://www.ecured.cu/S%C3%ADntesis_de_prote%C3%ADnas >.

Spicer, Rachel, Salek, Reza M., Moreno, Pablo, Cañuato, Daniel y Steinbeck, Christoph. Navigating freely-available software tools for metabolomics analysis [en línea]. *Springer US.* Artículo recibido en 11 de abril de 2017, aceptado el 25 de julio de 2017 y publicado el 9 de agosto de 2017. [Consulta: 10 de marzo de 2019]. Disponible en: <https://link.springer.com/article/10.1007/s11306-017-1242-7#citeas>.

Spin nuclear. A: *spinnucleary* [en línea]. Publicado en 15 de junio de 2012. [Consulta: 30 de abril de 2019]. Disponible en: < https://spinnucleary.wordpress.com/2012/06/15/spin-nuclear/ >.

Sussulini, Alessandra. *Metabolomics: From Fundamentals to Clinical Applications.* Spinger Internetional Publishing AG, 2017. ISBN 978-3-319-47655-1.

Torrades, Sandra. Proteómica. A: *Elsevier* [en línea]. Publicado en abril de 2004 [Consulta: 20 de abril de 2019]. Disponible a: < https://www.elsevier.es/es-revista-offarm-4-articulo-proteomica-13060308>.

Transcriptómica. A: *Boletin agrario* [en línea]. [Consulta: 20 de abril de 2019]. Disponible en: < https://boletinagrario.com/ap-6,transcriptomica,4361.html >.

What is proteomics? A: *European Bioinformatics Institute* [en línea]. [Consulta: 20 de abril de 2019]. Disponible a: < https://www.ebi.ac.uk/training/online/course/proteomics-introduction-ebi-resources/what-proteomics>.

Zita, Ana. Lípidos. A: *TodaMateria* [en línea]. Revisado en 17 de abril de 2019. [Consulta: 20 de abril de 2019]. Disponible a: <https://www.todamateria.com/lipidos/>.

ANEXO

Anexo A. Código en R

A.1. Lectura de la base de datos:

```
setwd("/Users/ainafdz/Desktop/Estadística/TFG/Dades/Vinos_blancos")
datos <- read.csv("summary_421340.csv", sep=";", header=TRUE)
rownames(datos) <- datos[,1]
datos <- datos[ ,-(1:2)]
```

A.2. Análisis descriptivo inicial

A.2.1. Estadísticos por cada metabolito

```
sumstats <- function(z) {
Media <- apply(z, 1, mean, na.rm=TRUE)
Mediana <- apply(z, 1, median, na.rm=TRUE)
SD <- apply(z, 1, sd, na.rm=TRUE)
CV <- apply(z, 1, function(x) sd(x, na.rm=TRUE)/mean(x, na.rm=TRUE))
result <- data.frame(Media, Mediana, SD, CV)
return(result)
}

estadisticos.metab <- sumstats(t(datos[,2:ncol(datos)]))
write.csv(estadisticos.metab, "Estadisticos_Metabolitos.csv")
```

A.2.2. Estadísticos por cada metabolito separando por variedades

```
mitj <- data.frame()
med <- data.frame()
des.st <- data.frame()
metab <- rep(colnames(datos)[-1], each=5)
for(i in 2:ncol(datos)){
  mitj <- rbind(mitj, aggregate(datos[,2],
by=list(datos$Variedad),mean, na.rm=TRUE))
  med <- rbind(med, aggregate(datos[,2],
by=list(datos$Variedad),median, na.rm=TRUE))
  des.st <- rbind(des.st, aggregate(datos[,2],
by=list(datos$Variedad),sd, na.rm=TRUE))
}
res <- cbind(metab, mitj, med, des.st)
res <- res[,-c(4,6)]
colnames(res) <- c("Metabolito", "Variedad", "Media", "Mediana", "SD")
res$CV <- res$SD/res$Media
write.csv(res, "Estadisticos_Por_Variedad.csv")
```

A.2.3. Estudio de missings reales

```
# Cálculo de los missings (cuántos hay y que % representan):
```

```
buscarNA2 <- function(bd){
  s <- 0
  for(i in 1:nrow(bd)){
    for(j in 1:ncol(bd))
      if(is.na(bd[i,j]==TRUE)){
        s <- s  + 1
      }
  }
  return(c(s,((s*100)/(ncol(bd)*nrow(bd)))))
}

round(buscarNA2(datos),2)


# % Missings por variable

buscarNA <- function(x){
  s <- 0
  for(i in 1:length(x)){
    if(is.na(x[i]==TRUE)){
      s <- s  + 1
    }
  }
  return(s*100/length(x))
}

m <- round(apply(datos, 2, buscarNA),2)
sort(m,dec=F)  # Los ordenamos de menos a más missings


# Histograma con el % de missings por variable

hist(m, breaks = 24, main='Histograma del
     porcentaje de missings por variable',
     xlab='Porcentaje de missings',
     ylab='Número de variables', xlim = c(0, 10),col="steelblue")


# Missings por variedad

sum(is.na(subset(datos, datos$Variedad=="Chardonnay")))
sum(is.na(subset(datos, datos$Variedad=="Pinot_Gris")))
sum(is.na(subset(datos, datos$Variedad=="Riesling")))
sum(is.na(subset(datos, datos$Variedad=="Sauvignon_Blanc")))
sum(is.na(subset(datos, datos$Variedad=="Viognier")))
```

A.2.4. Estudio de valores faltantes por nivel de detección

```
sum((subset(datos, datos$Variedad=="Chardonnay"))==0, na.rm=TRUE)
sum((subset(datos, datos$Variedad=="Pinot_Gris"))==0, na.rm=TRUE)
sum((subset(datos, datos$Variedad=="Riesling"))==0, na.rm=TRUE)
sum((subset(datos, datos$Variedad=="Sauvignon_Blanc"))==0, na.rm=TRUE)
sum((subset(datos, datos$Variedad=="Viognier"))==0, na.rm=TRUE)
```

A.2.5. Gráfica de boxplots múltiples inicial

```
library(RColorBrewer)
library(ggplot2)
boxplot(datos[2:ncol(datos)],datos,horizontal=F,names=FALSE )
```

A.2.6. Boxplots iniciales por cada uno de los metabolitos

```
library(RColorBrewer)
library(ggplot2)
vars2=colnames(datos)[-1]
for (va in vars2){
  if (!is.factor(datos[,va])){

boxplot(as.formula(paste0(va,"~Variedad")),datos,main=va,col=brewer.pa
l(7,"Blues"),horizontal=F)
  } else{
    plot(as.formula(paste0("Variedad~",va)),datos,main=va,col=c(2,3))
  }
}
```

Selección de los boxplots a incluir en el informe:

```
par(mfrow=c(1,3))
boxplot(datos$alpha.ketoglutaric.acid ~ Variedad ,datos,main="Alpha
Ketoglutaric Acid",col=brewer.pal(7,"Blues"),horizontal=F)

boxplot(datos$pipecolic.acid ~ Variedad ,datos,main="Pipecolic
Acid",col=brewer.pal(7,"Blues"),horizontal=F)

boxplot(datos$tryptophan ~ Variedad ,datos,main="
Tryptophan",col=brewer.pal(7,"Blues"),horizontal=F)

par(mfrow=c(1,3))

boxplot(datos$galactinol ~ Variedad
,datos,main="Galactinol",col=brewer.pal(7,"Blues"),horizontal=F)

boxplot(datos$maltose ~ Variedad
,datos,main="Maltose",col=brewer.pal(7,"Blues"),horizontal=F)

boxplot(datos$talose ~ Variedad
,datos,main="Talose",col=brewer.pal(7,"Blues"),horizontal=F)

# Con límite
par(mfrow=c(1,1))

boxplot(datos$talose ~ Variedad
,datos,main="Talose",col=brewer.pal(7,"Blues"),horizontal=F,
ylim=c(0,15000))
```

A.2. Preprocessing de los datos

A.2.1. Imputación de valores missing

```
anyNA(datos) # tenemos algun valor NA

library(DMwR)
datos_imp <- knnImputation(datos)  # perform knn imputation.
anyNA(datos_imp) # ya no tenemos valores missing
```

A.2.2. Transformación logaritmica

```
ldatos_imp <- log(datos_imp[,2:ncol(datos_imp)])
```

A.2.3. Normalizacion y escalado

```
# Función:
paretoscale <- function(z) {
            rowmean <- apply(z, 1, mean) # row means
            rowsd <- apply(z, 1, sd)  # row standard deviation
            rowsqrtsd <- sqrt(rowsd) # sqrt of sd
            rv <- sweep(z, 1, rowmean,"-")  # mean center
            rv <- sweep(rv, 1, rowsqrtsd, "/")  # divide by sqrtsd
            return(rv)
            }
# Aplicación:
ldatos_impscale <- paretoscale(ldatos_imp)
```

A.2.4. Datos definitivos:

```
datos_def <- as.data.frame(ldatos_impscale)
Variedad <- datos$Variedad
datos_def <- cbind(datos_def, Variedad)
```

A.2.5. Gráfica de boxplots múltiples después del preprocessing

```
par(mfrow=c(2,1))
boxplot(datos[2:ncol(datos)],datos,horizontal=F,names=FALSE          )
boxplot(datos_def[1:109],datos_def,horizontal=F,names=FALSE )
```

A.3. Análisis de los datos

A.3.1. ANOVA

Extracción de los p-values por cada uno de los metabolitos:

```
v <- vector()
for(i in 1:109){
  aux <- summary(aov(datos_def[,i] ~ datos_def$Variedad))
  v <- c(v, aux[[1]][1,5])
}
```

P-value ajustado:

```
v.adj <- p.adjust(v)
```

Data frame con el metabolito, el p-value y el p-value ajustado:

```
metab <- colnames(datos_def[,-110])
df.anova <- matrix(c(metab, t(v), t(v.adj)), ncol=3)
colnames(df.anova) <- c("Metabolito", "P-value", "Adj. P-value")

setwd("/Users/ainafdz/Desktop/Estadística/TFG/TFG/Figures_DEF")
write.csv(df.anova, "Resultado_ANOVA.csv")
```

A.3.2. Boxplots después del preprocessing por cada uno de los metabolitos

```
vars=colnames(datos_def)[-110]
for (va in vars){
  if (!is.factor(datos_def[,va])){

boxplot(as.formula(paste0(va,"~Variedad")),datos_def,main=va,col=brewe
r.pal(7,"Blues"),horizontal=F)
  } else{

plot(as.formula(paste0("Variedad~",va)),datos_def,main=va,col=c(2,3))
  }
}
```

Selección de los gráficos a incluir en el informe:

```
par(mfrow=c(1,3))
boxplot(datos_def$threitol ~ Variedad
,datos_def,main="Threitol",col=brewer.pal(7,"Blues"),horizontal=F)

boxplot(datos_def$inositol.myo. ~ Variedad ,datos_def,main="Inositol
Myo",col=brewer.pal(7,"Blues"),horizontal=F)

boxplot(datos_def$isothreonic.acid ~ Variedad
,datos_def,main="Isothreonic
Acid",col=brewer.pal(7,"Blues"),horizontal=F)

par(mfrow=c(1,3))

boxplot(datos_def$X2.hydroxyglutaric.acid ~ Variedad
,datos_def,main="X2 Hydroxyglutaric
Acid",col=brewer.pal(7,"Blues"),horizontal=F)

boxplot(datos_def$proline ~ Variedad
,datos_def,main="Proline",col=brewer.pal(7,"Blues"),horizontal=F)

boxplot(datos_def$serine ~ Variedad
,datos_def,main="Serine",col=brewer.pal(7,"Blues"),horizontal=F)
```

2.3.3. Análisis de Componentes Principales (PCA)

```
# Data frame con las variables numericas, a partir de las cuales se
hará el análisis
```

```
dnum_def <- datos_def[,-110]

# PCA
pc2 <- prcomp(dnum_def, retx=TRUE)

# Variabilidad explicada por las dos primeras componentes
PVE2 <- 100*pc2$sdev^2/sum(pc2$sdev^2)
PVE2[1:2]

# Resultados para hacer los gráficos
pcaresults2 <- summary(pc2)
scree.data2 <- as.data.frame(pcaresults2$importance)
score.data2 <- as.data.frame(pcaresults2$x)
loadings.data2 <- as.data.frame(pcaresults2$rotation)

setwd("/Users/ainafdz/Desktop/Estadística/TFG/TFG/Figures_DEF/PCA")
write.csv(scree.data2, "pca_scree_final.csv")
write.csv(score.data2, "pca_score_final.csv")
write.csv(loadings.data2, "pca_loadings_final.csv")

# Buscar los metabolitos más importantes de las dos primeras
componentes

rot.1 <- pc2$rotation[order(-abs(pc2$rotation[,1])), 1]
names(rot.1)
rot.2 <- pc2$rotation[order(-abs(pc2$rotation[,2])), 2]
rot <- matrix(c(names(rot.1), rot.1, names(rot.2), rot.2), ncol=4)
colnames(rot) <- c("Metabolito", "PC1", "Metabolito", "PC2")
setwd("/Users/ainafdz/Desktop/Estadística/TFG/TFG/Figures_DEF/PCA")
write.csv(rot, "Rotacion_componentes.csv")
```

Scores plot:

```
setwd("/Users/ainafdz/Desktop/Estadística/TFG/TFG/Figures_DEF/PCA")
score2 <- read.csv("pca_score_final.csv", header=TRUE)
score2 <- score2[,c(1:3)]  # nos quedamos con la 1a y 2a componentes

Variedad <- datos$Variedad
score2 <- cbind(score2, Variedad) # añadimos la variedad para colorear
los grupos

ggplot(score2, aes(PC1, PC2)) +
  geom_point(aes(color=Variedad)) +
  geom_text(aes(label=score2$X, color=Variedad)) +
  stat_ellipse(aes(color=Variedad)) +
  ggtitle("PCA Scores Plot") +
  theme(plot.title=element_text(size=15, vjust=2, face="bold")) +
  geom_hline(yintercept=0, size=0.25) +
  geom_vline(xintercept=0, size=0.25) +
  xlab((paste0("PC1"," ", "(",round(PVE2[1], 1), "%", ")"))) +
  ylab((paste0("PC2"," ", "(",round(PVE2[2], 1), "%", ")")))
```

Loadings plot:

```
setwd("/Users/ainafdz/Desktop/Estadística/TFG/TFG/Figures_DEF/PCA")
loadings2 <- read.csv("pca_loadings_final.csv", header=TRUE)
loadings2 <- loadings2[,c(1:3)]  # nos quedamos con la 1a y 2a
componentes

ggplot(loadings2, aes(PC1, PC2)) +
  geom_text(aes(label=loadings2$X)) +
  ggtitle("PCA Loadings Plot") +
  theme(plot.title=element_text(size=15, vjust=2, face="bold")) +
  geom_hline(yintercept=0, size=0.25) +
  geom_vline(xintercept=0, size=0.25) +
  xlab((paste0("PC1"," ", "(",round(PVE2[1], 1), "%", ")"))) +
  ylab((paste0("PC2"," ", "(",round(PVE2[2], 1), "%", ")")))


# por clases
loadings3 <- read.csv("pca_loadings_final_clases.csv", header=TRUE,
sep=";", dec=",")
loadings3 <- loadings3[,c(1:3)]  # nos quedamos con la 1a y 2a
componentes

ggplot(loadings3, aes(PC1, PC2)) +
  geom_text(aes(label=loadings3$X)) +
  ggtitle("PCA Loadings Plot (por clases de metabolitos)") +
  theme(plot.title=element_text(size=15, vjust=2, face="bold")) +
  geom_hline(yintercept=0, size=0.25) +
  geom_vline(xintercept=0, size=0.25) +
  xlab((paste0("PC1"," ", "(",round(PVE2[1], 1), "%", ")"))) +
  ylab((paste0("PC2"," ", "(",round(PVE2[2], 1), "%", ")")))
```

2.3.4. Heatmap

```
datos_num <- data.matrix(datos_def[,1:109])

color <- datos_def$Variedad
levels(color) <- c("lightgoldenrod", "lightpink", "darkseagreen",
"coral1", "cyan")
color <- as.character(color)

library(gplots)
heatmap.2(x = t(datos_num), scale = "none", col = bluered(256),
          distfun = function(x){dist(x, method = "euclidean")},
          hclustfun = function(x){hclust(x, method = "average")},
          density.info = "none",
          trace = "none", cexRow = 0.7,
          ColSideColors = color)
```

2.3.5. PLS-DA

```
library(DiscriMiner)
my_pls2 = plsDA(datos_def[,1:109], datos_def$Variedad, autosel=TRUE)

# Gráfico
plot(my_pls2)
```

```
# Funciones discriminantes
func.pls <- as.data.frame(my_pls2$functions)
metabolitos <- c("Intercept", colnames(datos_def[,-110]))
func.pls <- cbind(func.pls,metabolitos)

setwd("/Users/ainafdz/Desktop/Estadística/TFG/TFG/Figures_DEF/PLS-DA")
f.1 <- func.pls[with(func.pls, order(-abs(func.pls$V1))),]
write.csv(f.1[,c(1,6)], "funciones.discriminantes.1.csv")
f.2 <- func.pls[with(func.pls, order(-abs(func.pls$V2))),]
write.csv(f.2[,c(2,6)], "funciones.discriminantes.2.csv")

# Calidad
library(caret)
library(mlbench)

set.seed(1234)
  inTrain <- createDataPartition(
  y = datos_def$Variedad,
  ## the outcome data are needed
  p = 2/3,
  ## The percentage of data in the
  ## training set
  list = FALSE
)

training <- datos_def[ inTrain,]
testing  <- datos_def[-inTrain,]

plsFit <- train(
  Variedad ~ .,
  data = training,
  method = "pls",
  preProc = c("center", "scale")
)

plsVariedad <- predict(plsFit, newdata = testing)
confusionMatrix(data = plsVariedad, testing$Variedad)
```

2.3.6. Correlaciones

```
library("psych")
library("Hmisc")

flattenCorrMatrix <- function(cormat, pmat) {
  ut <- upper.tri(cormat)
  data.frame(
    row = rownames(cormat)[row(cormat)[ut]],
    column = rownames(cormat)[col(cormat)[ut]],
    cor  =(cormat)[ut],
    p = pmat[ut]
    )
}
```

```
# Correlaciones metabolitos
cor.met <- corr.test(datos_def[,1:109],use="complete",
method="spearman")
cor.metab <- flattenCorrMatrix(cor.met$r, cor.met$p)
names(cor.metab) <- c("Metabolito", "Metabolito", "Correlación", "P-
value")
write.csv(cor.metab, "AnexoD.Correlaciones_Metabolitos.csv")

# Correlaciones muestras

# transponemos los datos
t.datos<- as.data.frame(t(datos_def[,-110]))
cor.vin <- corr.test(t.datos, use="complete", method="spearman")
cor.vinos <- flattenCorrMatrix(cor.vin$r, cor.vin$p)
names(cor.vinos) <- c("Muestra", "Muestra", "Correlación", "P-value")
write.csv(cor.vinos, "AnexoE.Correlaciones_Muestras.csv")
```

A.4. Integración de los datos

```
setwd("/Users/ainafdz/Desktop/Estadística/TFG/TFG/Figures_DEF/Integrac
ión")
aux <- read.csv("summary_421340_INT.csv", sep=";", dec=".",
header=TRUE)
datos_int <- cbind(datos_def, aux[4:6])

library(FactoMineR)
res = MFA(datos_int, group=c(109,1,3), type=c("s", "n","s"),
name.group=c("metabolitos","variedad","características"), ncp=3,
num.group.sup = 2, graph = FALSE)
summary(res, nbelements = Inf)


# DIMENSIÓN 1 VS DIMENSIÓN 2

# Partial axes plot
plot(res, choix="axes")

# Individuals plot
plot(res, habillage=110, cex=0.8, select="contrib 20")

# Variables plot
plot(res, choix="var", habillage="group", cex=0.8, select="contrib
10", unselect=1)

# Individual factor map por grupos
plot(res, choix="ind", invisible="ind", habillage="group", cex=0.8,
partial="all")


# DIMENSIÓN 1 VS DIMENSIÓN 3

# Partial axes plot
plot(res, choix="axes", axes=c(1,3))
```

```
# Individuals plot
plot(res, habillage=110, cex=0.8, select="contrib 20", axes=c(1,3))

# Variables plot
plot(res, choix="var", habillage="group", cex=0.8, select="contrib
10", unselect=1, axes=c(1,3))

# Individual factor map por grupos
plot(res, choix="ind", invisible="ind", habillage="group", cex=0.8,
partial="all", axes=c(1,3))
```

Anexos B. Boxplots iniciales

Se pueden encontrar los gráficos de los boxplots obtenidos antes del preprocesamiento siguiendo el siguiente enlace:

https://drive.google.com/open?id=1nbjoDt7UCsV3Gio0N-bBzivRgEP6XjwU

En la carpeta que sigue al enlace, nombrada **Anexos**, se puede encontrar un documento PDF con el nombre de **AnexoB.Boxplots_Iniciales**, que corresponde con este apartado.

El motivo de la creación de dicha carpeta es la gran extensión de los anexos que se han incluido en ella.

Anexo C. Boxplots finales

Se pueden encontrar los gráficos de los boxplots realizados después del preprocesamiento siguiendo el siguiente enlace:

https://drive.google.com/open?id=1nbjoDt7UCsV3Gio0N-bBzivRgEP6XjwU

En la carpeta que sigue al enlace, se puede encontrar un documento PDF con el nombre de **AnexoC.Boxplots_Finales**, que corresponde con este apartado.

Anexo D. Matriz de correlaciones entre metabolitos

En el enlace siguiente se puede encontrar un excel con el nombre de **AnexoD.Correlaciones_Metabolitos** en el que aparece una lista con todas las correlaciones entre metabolitos, así como su *p-value.*

https://drive.google.com/open?id=1nbjoDt7UCsV3Gio0N-bBzivRgEP6XjwU

Anexo E. Matriz de correlaciones entre muestras de vinos

En el enlace siguiente se puede encontrar un excel con el nombre de **AnexoE.Correlaciones_Muestras** en el que aparece una lista con todas las correlaciones entre las distintas muestras de vinos, así como su *p-value.*

https://drive.google.com/open?id=1nbjoDt7UCsV3Gio0N-bBzivRgEP6XjwU

Anexo F. Análisis de Factores Múltiples

En el enlace que figura a continuación (y que corresponde con todos los que han aparecido anteriorment) hay un documento PDF con el nombre de **AnexoF.MFA**, en el que aparece el resumen numérico de los resultados del anàlisis de factores múltiples.

https://drive.google.com/open?id=1nbjoDt7UCsV3Gio0N-bBzivRgEP6XjwU

Anexo G. Aplicación Shiny

Se ha querido mostrar en el trabajo la aplicación Shiny que se ha realizado para la presentación oral. Se puede visualizar usando el siguiente comando en la consola del *software* R:

```
library(shiny)
runGitHub("pipeline_datos_metabolomicos","ainafdz")
```