UNIVERSIDAD DE GRANADA

FACULTAD DE CIENCIAS

Departamento de Química Analítica

Programa de Doctorado en Química

Grupo de Investigación AGR-274 “Bioactive ingredients”

Centro Tecnológico de Investigación y Desarrollo del Alimento Funcional (CIDAF)

DESARROLLO Y APLICACIÓN DE ESTRATEGIAS METABOLÓMICAS MEDIANTE TÉCNICAS ANALÍTICAS AVANZADAS EN MUESTRAS BIOLÓGICAS

Memoria presentada por

Álvaro Fernández Ochoa

Para optar al grado de

Doctor Internacional en Química por la Universidad de Granada

Tesis doctoral dirigida por

Dr. Antonio Segura Carretero

Dra. Isabel Borrás Linares

Granada, Noviembre de 2019

Editor: Universidad de Granada. Tesis Doctorales
Autor: Álvaro Fernández Ochoa
ISBN: 978-84-1306-478-9
URI: http://hdl.handle.net/10481/62229

Esta tesis doctoral ha sido realizada gracias a la ayuda para la formación de profesorado universitario (FPU) concedida por el Ministerio de Educación, Cultura y Deporte (FPU14/03992), al proyecto europeo PRECISESADS (Molecular Reclassification to Find Clinically Useful Biomakers for Systemic Autoimmune Diseases) financiado por la Iniciativa sobre Medicamentos innovadores (IMI) en virtud del acuerdo de subvención nº[115,565], con la contribución financiera del Séptimo Programa Marco de la Unión Europea (FP7/2007-2013) y la aportación en especie de las empresas EFPIA; y a la financiación con cargo a fondos del grupo de investigación AGR-274 “Bioactive Ingredients” y del Centro de Investigación y Desarrollo del Alimento Funcional (CIDAF), procedentes de diferentes proyectos, contratos y subvenciones de las Administraciones Central y Autonómica, Plan Propio de investigación de la UGR, así como de empresas interesadas en los resultados de investigación.

AGRADECIMIENTOS

Durante mis primeros años de la carrera en la Universidad de La Rioja, empecé a conocer el ámbito que rodea a la carrera científica, consiguiendo una alta motivación para intentar dedicarme a la investigación una vez que acabase la carrera, marcándome por tanto la meta y el sueño de acabar realizando el doctorado. Ahora, una vez que está llegando el momento de terminar esta etapa de mi vida, me gustaría dejar plasmadas en esta memoria unas palabras de agradecimiento a todas las personas que han contribuido con su granito de arena a que haya podido llegar hasta este punto. Análogamente al perfil de una etapa ciclista, el doctorado ha estado llena de subidas, y alguna que otra bajada, y no tengo ninguna duda de que sin todos vosotros no hubiera sido posible llegar hasta esta meta. ¡Muchas Gracias!

Me gustaría comenzar agradeciendo a mi director Antonio Segura Carretero, por abrirme las puertas de tu grupo de investigación allá por 2015, por tu gran labor como director de esta tesis y por tu capacidad para despertarme el interés por la investigación y para motivarme para continuar por el camino de la carrera científica. Y a Isabel, por la codirección de esta tesis, por tu apoyo y el seguimiento durante este período y por haber compartido varios momentos juntos como el “inolvidable” viaje a Ascot.

También agradecer a Rikard Landberg y especialmente a Carl Brunius (Universidad de Chalmers, Suecia) por acogerme en vuestro centro de investigación y darme la oportunidad de formarme en técnicas de procesamiento y análisis de datos metabolómicos mediante programas basados en lenguaje R. Esta estancia de tres meses en Suecia fue sin lugar a dudas una de las experiencias más duras y a la vez más

inolvidables que he vivido durante este período predoctoral. Me gustaría agradecer a todos los compañeros que me acogisteis en “The EpiHub” del departamento de Ciencias Quirúrgicas de la Universidad de Uppsala, y en especial a Anna-Karin, Liisa y Erika por vuestra amabilidad y por haberme acercado vuestra cultura en las charlas durante la “Fika”. Por otra parte, agradecer a Noelia, Javi, Clara y Ramón, porque me hicisteis este período mucho más ameno gracias al tiempo que pasamos en las “Nations” o en los viajes como el de *Gamla Uppsala* o *Sigtuna*.

Me siento agradecido a todas las personas que he conocido durante todos estos años en el CIDAF y por diferentes motivos ya no estáis aquí. Sois muchos y no me gustaría dejarme a ninguno pero sí que me gustaría agradecer especialmente a Hakim, Nassima, Arancha, Celia, María del Mar y Patri vuestra buena acogida que me disteis durante mi primera etapa en el centro. A José Antonio, por todos los momentos que pasamos en el laboratorio de metabolómica, por los molletes de medicina, y sobre todo porque espero que sigamos compartiendo más rutas Hiponova y medias maratones juntos.

Y cómo no, agradecer a todos los que me habéis y seguís aguantado con vuestra continua paciencia durante tantos años en el CIDAF. A David, por tu alegría y por ayudarme siempre que lo he necesitado tanto en tareas de investigación como docentes. A la tercera del trío *Metabol*, Rosa, por todo tu apoyo durante mis inicios en esto de la metabolómica y por todo lo que nos cuidaste en el congreso de Dublín, demostrando la buena madre que eres ahora. A Raquel, por ser la única que a veces me entiendes cuando hablo de cosas del norte de España. A José Antonio, por tu buen saber estar y por ser la gran persona que se ve que eres. A Elena, porque siempre son bienvenidas tus estancias y así sumarte al bando de los que somos de más allá del muro de Despeñaperros.

A Rafa, por ser una de las personas más buena y con mejor corazón que he conocido nunca a pesar de que te llegué a sacar de tus casillas un día con tu memorable frase “’¡Logroñes cabr...!”. Y sí, durante este tiempo también he tenido la oportunidad de trabajar con la persona más inocente que he conocido nunca, Mª Ángeles. Gracias por tu inocencia y todos los buenos ratos que hemos pasados juntos, especialmente en las clases de spinning y por todas las anécdotas que nos has hecho vivir como el pisotón del “thrombocid”, la que se lío con el pulpo, los chispazos de tu cable o cuando se te escapó la sorpresa de Cagliari.

A Jesús, por todo lo que me has demostrado durante todos estos años y por haber compartido momentos juntos más allá del laboratorio como tu despedida en Almuñécar además de haberme dado la posibilidad de conocer a tu Luis. Gracias a los dos por visitarme en Logroño hace dos veranos y por vivir momentos irrepetibles como la ruta por el río Chillar o vuestro bodorrio. A Sandra, por ser la primera persona que me ayudó desde el primer día que llegué al CIDAF hasta el último. Gracias por ayudarme en todo momento y demostrar tu compañerismo a pesar de que te invada tu espacio de la mesa continuamente con mis cosas. Y sobre todo, por tantos buenos momentos como nuestro primer póster en el Desgranando, tu despedida en Cazorla, el congreso de Valencia o el viaje a Cagliari. A Mari Carmen, porque has sido el mejor fichaje del grupo que podía haber llegado, gracias por ser tan natural y por hacernos reír tanto todos los días. ¡No cambies nunca aunque te saques el B2 de inglés, fuyiah!

Y para terminar con la ronda del CIDAF, a Javi y Mari Luz, que más que mis compañeros podría decir que sois mis hermanos. Gracias a los dos por haber compartido tantos momentos como vuestra visita a Logroño el verano pasado y sobre todo por haber estado a mi lado en mis momentos más difíciles. Javi, que hubiese sido de mí durante estos años sin tus continuas bromas y sin tu buen sentido del humor. No cambies nunca, ¡Eres la alegría del CIDAF! Y Mari Luz, que hubiese sido de mí sin todas las movidas en las que me has metido durante estos años, desde mis primeras JIFFI allá por 2017 hasta una charla en el Lemmon Rock a dos semanas de depositar la tesis. Gracias por hacerme ver que la formación doctoral va más allá de estar metido en el laboratorio y publicar artículos. Y sobre todo, gracias por habernos brindado aquella magnífica actuación en la semifinal de Famelab en Sevilla y de haber podido trabajar codo con codo pensando en "ERCs", que espero que haya sido el principio de un largo camino profesional.

Me gustaría continuar agradeciendo a toda la gente que me ha ayudado a desconectar del doctorado durante estos años y hacer que esta etapa de mi vida haya sido lo más llevadera posible. De esta forma, me gustaría especialmente agradecer a la canaria Clau por todas las rutas en bici, carreras, días de tapeo y tantos buenos momentos. Y a María, por aquellos años de entrenos en el Viva, rutas y porque eres el mejor ejemplo de una luchadora que acaba consiguiendo lo que se propone. ¡Eres muy grande! Además, me gustaría agradecer a toda la buena gente que he podido conocer en el club ciclista BTT300 durante todos estos años, por posibilitarme disfrutar de una de mis mejores aficiones con vosotros y de esta forma haber vivido tantos domingos especiales.

Agradecer a Antonio y Aarón por ser dos grandes amigos desde los primeros meses de que llegué a Granada. Aunque ahora vivamos lejos unos de otros, espero que podamos compartir más momentos juntos. Gracias también a Efrén y Patri, por vuestra continúo apoyo desde vuestra isla. ¡Espero que nos reencontremos pronto!

Formar parte del comité organizador del congreso JIFFI ha sido uno de las actividades formativas más enriquecedoras que he realizado durante este período. Me gustaría agradecer a todas las personas que han formado parte del comité organizador de las JIFFI durante las tres últimas ediciones. No me gustaría dejarme a nadie pero si me gustaría agradecer especialmente a Beñat por haber coordinado juntos las JIFFI3, donde me demostraste la gran persona que eres. ¡Granada te echará mucho de menos!

Mi participación en el concurso "3-Minute Thesis" fue otra de las actividades formativas que más me llenaron al aparecer en uno de mis momentos más complicados que he vivido durante la tesis. Por ello, al final de los agradecimientos he querido dejar reflejado el monólogo que realicé en dicho concurso, y me gustaría agradecer a Carlos, Susana y a todos los participantes el apoyo y compañerismo que demostrasteis durante ese período.

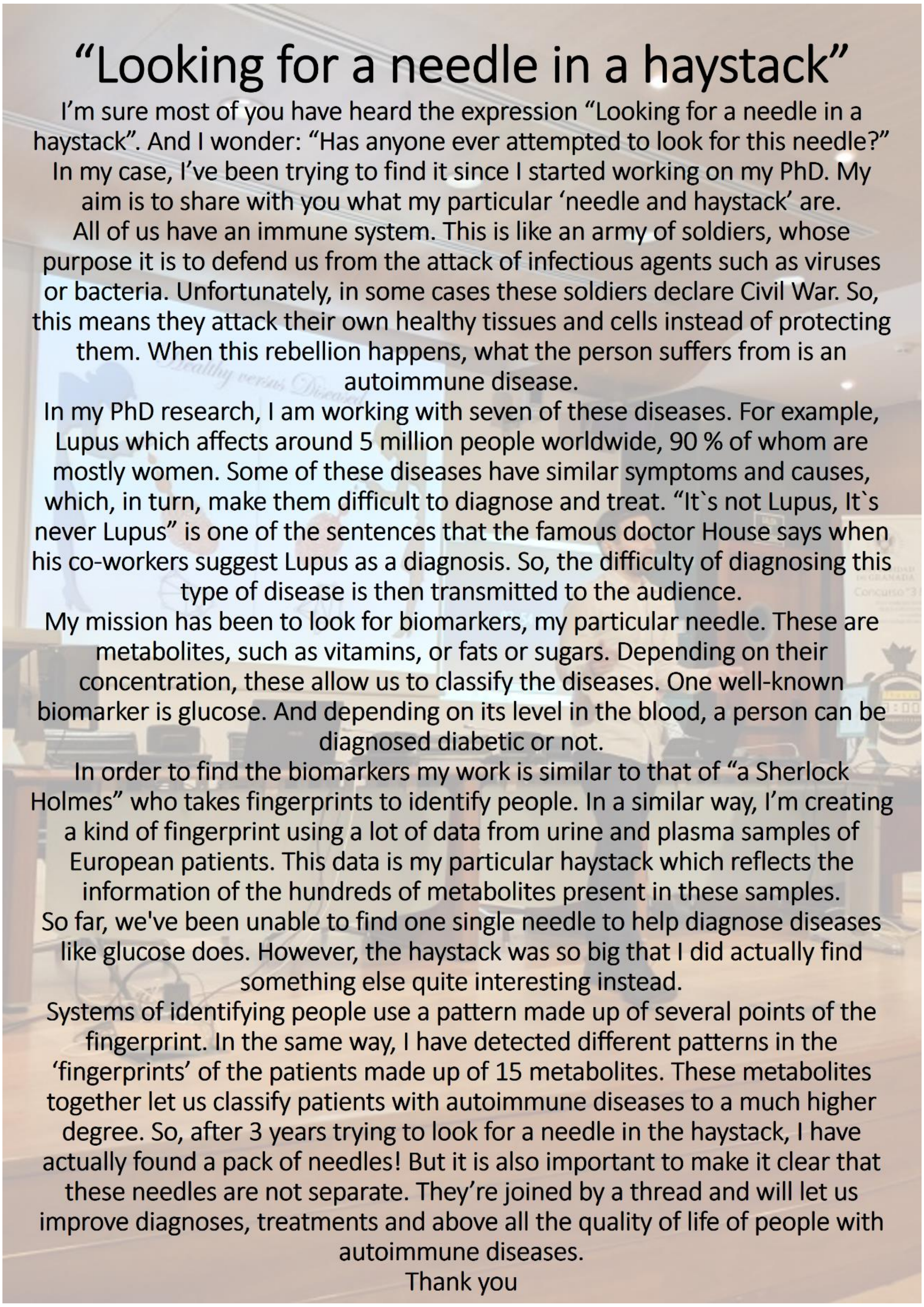
Tampoco me puedo olvidar de mis amigos de Logroño, que a pesar de estar a cientos de kilómetros, siempre estáis ahí cuando se os necesita. Gracias a Martin y Erik, por hacerme reír siempre que nos juntamos y sobre todo por la visita que me hicisteis este agosto, junto con Miriam, justo en el momento cuando más necesitaba desconectar unos días. ¡Mil Gracias!

A Fernando, por ser tan único y por haberte venido dos veces a las JIFFI. ¡No cambies nunca! A Enrique, Sergio, Varela y María, por tantos años de amistad y por vuestro apoyo desde Logroño. A Vanesa, por todo lo que me aguantaste durante tantos años y por los momentos tan especiales que compartimos juntos. A la próxima vendimiadora, Raquel, por tantos años de amistad y por tu alegría ya que siempre que nos vemos, las risas y el buen rollo están asegurados.

A Rubén, por ser "mi puto ídolo". Gracias por estar siempre ahí y porque ya sea en Granada o en Uppsala, siempre eres el primero en visitarme. ¡Eres un grande!

Finalmente, me gustaría agradecer a los responsables principales de que hoy este aquí, mi familia. A mis abuelos Domingo y Dionisia, por todo lo que me cuidasteis y me enseñasteis cuando era pequeño y en definitiva por haber sido los mejores abuelos que he podido tener. A mis tíos, Domingo y Cortijo, por haber sido unos referentes y haberme criado como si fuera vuestro hijo. A mis padres, porque sin vosotros nada en mi vida hubiera tenido sentido. Gracias por la educación que me disteis y por estar siempre orgullosos y pendientes de mí. A mi hermana, por todos los momentos que hemos vivido juntos, por estar ahí siempre que te necesito y sobre todo, por darme la mejor noticia el pasado día de San Mateo que me podrías dar. ¡Espero que el año que viene por mayo me des una noticia todavía mejor!

¡MUCHAS GRACIAS A TODOS DE CORAZÓN!



“Looking for a needle in a haystack”

I’m sure most of you have heard the expression “Looking for a needle in a haystack”. And I wonder: “Has anyone ever attempted to look for this needle?” In my case, I’ve been trying to find it since I started working on my PhD. My aim is to share with you what my particular ‘needle and haystack’ are.

All of us have an immune system. This is like an army of soldiers, whose purpose it is to defend us from the attack of infectious agents such as viruses or bacteria. Unfortunately, in some cases these soldiers declare Civil War. So, this means they attack their own healthy tissues and cells instead of protecting them. When this rebellion happens, what the person suffers from is an autoimmune disease.

In my PhD research, I am working with seven of these diseases. For example, Lupus which affects around 5 million people worldwide, 90 % of whom are mostly women. Some of these diseases have similar symptoms and causes, which, in turn, make them difficult to diagnose and treat. “It`s not Lupus, It`s never Lupus” is one of the sentences that the famous doctor House says when his co-workers suggest Lupus as a diagnosis. So, the difficulty of diagnosing this type of disease is then transmitted to the audience.

My mission has been to look for biomarkers, my particular needle. These are metabolites, such as vitamins, or fats or sugars. Depending on their concentration, these allow us to classify the diseases. One well-known biomarker is glucose. And depending on its level in the blood, a person can be diagnosed diabetic or not.

In order to find the biomarkers my work is similar to that of “a Sherlock Holmes” who takes fingerprints to identify people. In a similar way, I’m creating a kind of fingerprint using a lot of data from urine and plasma samples of European patients. This data is my particular haystack which reflects the information of the hundreds of metabolites present in these samples.

So far, we've been unable to find one single needle to help diagnose diseases like glucose does. However, the haystack was so big that I did actually find something else quite interesting instead.

Systems of identifying people use a pattern made up of several points of the fingerprint. In the same way, I have detected different patterns in the ‘fingerprints’ of the patients made up of 15 metabolites. These metabolites together let us classify patients with autoimmune diseases to a much higher degree. So, after 3 years trying to look for a needle in the haystack, I have actually found a pack of needles! But it is also important to make it clear that these needles are not separate. They’re joined by a thread and will let us improve diagnoses, treatments and above all the quality of life of people with autoimmune diseases.

Thank you

Álvaro Fernández Ochoa

Concurso “3-Minute Thesis” Universidad de Granada

20 de febrero de 2019

Si parece imposible, no descanses

y si pierdes

que te quede la satisfacción

de haberte partido el pecho por algo

Ikeli O’farrell

ÍNDICE

ÍNDICE DE CONTENIDOS

LISTA DE ABREVIATURAS

ANOVA (*analysis of variance*): análisis de la varianza

APCI (*atmospheric-pressure chemical ionization*): sistema de ionización química a presión atmosférica

APPI (*atmospheric pressure photoionization*); fotoionización a presión atmosférica

AUC (*area under the curve*): área bajo la curva

BMI (*body mass index*): índice de masa corporal

CE (*capillary electrophoresis*): electroforesis capilar

cv (cultivated variety): variedad cultivada

CeuMM: *Ceu Mass Mediator*

CRAN (*comprehensive R archive network*): repositorio oficial de paquetes de R

CV (*cross validation*): validación cruzada

dcSSC (*diffuse cutaneous systemic sclerosis):* esclerosis sistémica cutánea difusa

lcSSC (*limited cutaneous systemic sclerosis*): esclerosis sistémica cutánea limitada

DNA (*deoxyribonucleic acid*): ácido desoxirribonucleico

EDTA (*ethylenediaminetetraacetic acid*): ácido etilendiaminotetraacético

EI (*electronic ionization*): ionización electrónica

ESI (*electrospray ionization*): ionización por electrospray

FC (*fold change*): cambio en el incremento

FDR (*false discovery rate*): tasa de descubrimientos falsos

GC (*gas chromatography*): cromatografía de gases

HCA (*hierarchical cluster analysis*): análisis de cluster jerárquico

HILIC (*hydrophilic interaction chromatography*): cromatografía de interacción hidrofilíca

HMDB (*human metabolome database*): base de datos del metaboloma humano

HPLC (*high performance liquid chromatography*): cromatografía de líquidos de alta eficacia

HRMS (*high resolution mass spectrometry*): espectrometría de masas de alta resolución

ICP (*inductively coupled plasma*): plasma acoplado inductivamente

IMS (*ion mobility spectrometry*): espectrometría de movilidad iónica

IS (*internal standard*): patrón interno

KEGG (*Kyoto encyclopedia of genes and genomes*): enciclopedia “Kyoto” de genes y genomas

LC (*liquid chromatography*): cromatografía de líquidos

LC/MS (*mass spectrometry coupled to liquid chromatography*): cromatografía de líquidos acoplada a espectrometría de masas

LOD (*limit of detection*): límite de detección

LOQ (*limit of quantification*): límite de cuantificación

MALDI (*matrix-assited laser desorption/ionization*): desorción/ionización láser asistida por una matriz

MCTD (*mixed connective tissue disease*): enfermedad mixta del tejido conectivo

MetPA (*metabolic pathway analysis*): análisis de rutas metabólicas

MS (*mass spectrometry*): espectrometría de masas

MS/MS (*tandem mass spectrometry*): espectrometría de masas en tándem

MSEA (*metabolite set enrichement analysis*): análisis de enriquecimiento para un grupo de metabolitos

MSI (*Metabolomics Standards Initiative*): Iniciativa de estándares de metabolómica

MSTS (*mass spectrometry total signal*): suma total de las señales detectadas por espectrometría de masas

MSTUS (*mass spectrometry total useful signal*): suma total de las señales comunes detectadas por espectrometría de masas

m/z (*mass-to-charge ratio*): relación masa/carga

NA (*not available, missing values*): valores faltantes

NMR (*nuclear magnetic resonance*): resonancia magnética nuclear

NP-LC (*normal phase liquid chromatography*): cromatografía líquida de fase normal

OPLS (*orthogonal projections to latent structures*): proyecciones ortogonales a estructuras latentes

OPLS-DA (*orthogonal projections to latent structures discriminant analysis*): proyecciones ortogonales a estructuras latentes – análisis discriminante

ORA (*overrepresentation analysis*): análisis de sobrerrepresentación

PAPS (*primary antiphospholipid syndrome*): síndrome antifosfolípido

PCA (*principal component analyses*): análisis de componentes principales

PLS (*partial least squares regression*): regresión de mínimos cuadrados parciales

PLS-DA (*partial least squares regression – discriminant analysis*): regresión de mínimos cuadrados parciales - análisis discriminante

PQN (*probabilistic quotient normalization*): normalización por el cociente probabilístico

pSJS (*primary Sjögren's syndrome*)**:** síndrome de Sjögren primario

Q (*quadrupole*): analizador cuadrupolo

QC (*quality control*): muestra de control de calidad

QEA (*quantitative enrichment analysis*): análisis de enriquecimiento cuantitativo

QTOF (*quadrople time-of-flight*): analizador cuadrupolo-tiempo de vuelo

RA (*rheumatoid arthritis*): artritis reumatoide

RF (*random forest*): bosques aleatorios

RNA (*ribonucleic acid*): ácido ribonucleico

ROC (*receiver operating characteristic*): característica operativa del receptor

RP-LC (*reversed phase liquid chromatography*): cromatografía líquida de fase reversa

RSD (*relative standard deviation*): desviación estándar relativa

RT (*retention time*): tiempo de retención

SADS (*systemic autoimmune diseases*): enfermedades autoinmunes sistémicas

SJS (*Sjögren's syndrome*)**:** síndrome de Sjögren

SLE (*systemic lupus erythematosus*): lupus eritematoso sistémico

SSC (*systemic sclerosis*): esclerosis sistémica.

SSP (*single-sample profiling*): perfil de muestra única

S/N (*signal-to-noise ratio*): relación señal/ruido

TOF (*time of flight*): analizador tiempo de vuelo

UCTD (*undifferentiated connective tissue disease*): enfermedad del tejido conectivo indiferenciado

VP (*volcano plot*): gráfico de volcano

LISTA DE FIGURAS

LISTA DE TABLAS

RESUMEN

SUMMARY

RESUMEN

La presente memoria recoge los resultados alcanzados durante el desarrollo de la tesis doctoral titulada “**Desarrollo y aplicación de estrategias metabolómicas mediante técnicas analíticas avanzadas en muestras biológicas**”. En ella, se han desarrollado metodologías basadas principalmente en estrategias metabolómicas no dirigidas para su aplicación en el estudio de **compuestos bioactivos** así como en el estudio de **enfermedades autoinmunes sistémicas**. Esta memoria se estructura en dos grandes secciones: introducción y parte experimental.

La primera sección de **INTRODUCCIÓN** describe las principales características de la metabolómica, focalizándose principalmente en la metodología basada en **estrategias no dirigidas** al ser la aproximación empleada en el desarrollo de la tesis doctoral. En este sentido, se describe el flujo de trabajo de este tipo de estrategias recorriendo el estado actual de todas las etapas implicadas (diseño del estudio, toma y tratamiento de muestras biológicas, adquisición de datos, pre-procesamiento de datos, análisis estadísticos, identificación de metabolitos e interpretación biológica). El enfoque de la descripción de estas etapas está orientado al empleo de **cromatografía de líquidos acoplada a espectrometría de masas** (HPLC-ESI-QTOF-MS) para estos estudios, al ser la plataforma analítica utilizada en la parte experimental de la memoria. Por último, se detallan los principales campos de aplicación de las estrategias metabolómicas descritas, haciendo hincapié en dos ámbitos, el estudio de enfermedades así como de compuestos bioactivos presentes en alimentos.

Por otro lado, la **PARTE EXPERIMENTAL** incluye los trabajos recopilados en capítulos llevados a cabo a lo largo del periodo predoctoral. Esta sección se divide, a su vez, en

dos grandes bloques diferenciados que describen las aplicaciones metabolómicas llevadas a cabo en el ámbito de los compuestos bioactivos (**Bloque A**) y de las enfermedades autoinmunes sistémicas (**Bloque B**).

El **BLOQUE A**, centrado en el desarrollo y aplicación de estrategias metabolómicas para el estudio de compuestos bioactivos, incluye en primer lugar una breve introducción sobre los **compuestos bioactivos** y sus aplicaciones en las áreas de **alimentación funcional** y **nutracéutica**. Dentro de este bloque, se incluyen tres capítulos.

En el **Capítulo 1**, se realizó un estudio de absorción intestinal y metabolismo a través de un **ensayo de perfusión intestinal** *in situ,* utilizando ratas Wistar como modelo animal. Para ello se utilizó un extracto de romero rico en compuestos bioactivos, el cual ha demostrado poseer un gran potencial anticancerígeno frente a adenocarcinoma de colon.

En los **Capítulos 2** y **3**, se realizaron dos **ensayos de intervención nutricional** para evaluar el efecto en las rutas metabólicas producido por el consumo de extractos ricos en compuestos bioactivos obtenidos de dos fuentes vegetales alimentarias. Concretamente, en el **Capítulo 2**, se empleó un **modelo animal** de ratas diabéticas inducidas por estreptozotocina, a las cuales se les proporcionó durante cinco semanas una dieta rica en compuestos bioactivos extraídos de la pulpa y cáscara de mango cv. “Ataulfo”. La aplicación de una estrategia metabolómica no dirigida en este estudio se llevó a cabo mediante el análisis por HPLC-ESI-QTOF-MS de las muestras de suero sanguíneo e hígado recogidas al final de la intervención.

Por su parte, en el **Capítulo 3**, se llevó a cabo un ensayo de intervención nutricional en **humanos** con la finalidad de conocer los cambios metabólicos que produce la ingesta de un suplemento alimenticio producido a partir de un extracto de ajo. Para ello, se recolectaron muestras de plasma sanguíneo a 30 voluntarios sanos, antes y después de consumir el suplemento de ajo durante el período de un mes. Estas muestras biológicas se analizaron posteriormente mediante HPLC-ESI-QTOF-MS a través de un enfoque metabolómico no dirigido.

El **BLOQUE B** se centra en la aplicación de estrategias metabolómicas no dirigidas para el estudio de **enfermedades autoinmunes sistémicas**. Para ello, en primer lugar, se describen brevemente las características de este tipo de patologías. Concretamente, se han estudiado en este bloque siete de estas enfermedades: lupus eritematoso sistémico, artritis reumatoide, esclerosis sistémica, síndrome de Sjögren, síndrome antifosfolípido, enfermedad mixta del tejido conectivo y enfermedad mixta del tejido conectivo indiferenciado. Este bloque incluye cuatro capítulos, de los cuales dos de ellos (capítulos 6 y 7) se llevaron a cabo en colaboración con la Universidad de Chalmers (Suecia) gracias a una estancia predoctoral de tres meses en el año 2018 financiada por el Ministerio de Educación, Cultura y Deporte (“Estancias Breves FPU”).

En el **Capítulo 4**, se estudiaron las alteraciones metabólicas encontradas en muestras de plasma sanguíneo y orina de 59 pacientes com esclerosis sistémica. Análogamente, el **Capítulo 5** se focalizó en el estudio del síndrome de Sjögren en las muestras tomadas de 43 pacientes. En ambos capítulos, la etapa de pre-procesamiento de datos se realizó mediante el software de la casa comercial de la plataforma analítica utilizada (HPLC-ESI-QTOF-MS), *Agilent Mass Profinder*. Dada la dificultad observada para pre-procesar los datos metabolómicos adquiridos de un elevado número de muestras con este tipo de software, en el **Capítulo 6**, se compararon dos metodologías de pre-procesamiento para su aplicación en estudios de metabolómica no dirigida con un elevado número de muestras. Estas metodologías se basaron en la utilización del software comercial mencionado anteriormente y en programas de acceso libre basados en paquetes desarrollados en lenguaje R.

Finalmente, el **Capítulo 7** se centró en el estudio de las siete enfermedades autoinmunes sistémicas mediante análisis metabolómicos de muestras de orina y de plasma sanguíneo tomadas de 228 voluntarios enfermos y 55 controles sanos. El procesamiento de los datos se realizó mediante la metodología optimizada en el capítulo anterior basada en lenguaje R.

SUMMARY

The current report encompasses all the results achieved during the development of the PhD Thesis entitled "**Development and application of metabolomic strategies through advances analytical techniques in biological samples**". In this thesis, different methodologies have been developed based mainly on untargeted metabolomic strategies for application in the study of **bioactive compounds** as well as in the study of **systemic autoimmune diseases**. This report is divided into two main sections: introduction and experimental part.

The **INTRODUCTION** section describes the main characteristics of the metabolomic field, focusing mainly on **untargeted strategies** since it has been the main approach used in the development of this PhD thesis. In this sense, the workflow of this type of strategy is described, covering the current state of all the stages involved (study design, collection and treatment of biological samples, data acquisition, data pre-processing, statistical analysis, identification of metabolites and biological interpretation). The description of these steps is oriented to the use of **liquid chromatography coupled to mass spectrometry** (HPLC-ESI-QTOF-MS) for these studies, since it has been the analytical platform used in the experimental part of the thesis. Finally, the main fields of application of metabolomic strategies are detailed, emphasizing two different areas, the study of bioactive compounds and diseases.

On the other hand, the **EXPERIMENTAL PART** includes all studies that have been carried out during the PhD. This part is subdivided in two differentiated subsections that describe the metabolomic applications carried out in the field of bioactive compounds (**Section A**) and systemic autoimmune diseases (**Section B**), respectively.

SECTION A, focused on the development and application of metabolomic strategies for the study of bioactive compounds, first includes a brief introduction of **bioactive compounds** and their applications in the areas of **functional foods** and **nutraceuticals**. Three chapters are included within this section.

In **Chapter 1**, a study of intestinal absorption and metabolism was performed through an *in situ* **intestinal perfusion assay** in Wistar rats. In this study, a rosemary extract rich in bioactive compounds was administered, which has been shown to have a great anticancer potential against colon adenocarcinoma.

In **Chapters 2** and **3**, two **nutritional intervention studies** were conducted to assess the effect on metabolic pathways produced by the consumption of extracts rich in bioactive compounds obtained from two food sources. Concretely, in **Chapter 2**, an **animal model** of streptozotocin-induced diabetic rats was used. The animals received for five weeks a diet rich in bioactive compounds extracted from the pulp and peel of mango cv. "*Ataulfo*". The application of an untargeted metabolomic strategy in this study was performed by HPLC-ESI-QTOF-MS analysis of serum and liver samples collected at the end of the intervention.

In **Chapter 3**, a nutritional intervention study was carried out in **humans** in order to know the metabolic changes produced by the intake of a food supplement based on a garlic extract. In this study, blood plasma samples were collected from 30 healthy volunteers, before and after consuming the garlic supplement for a month. These samples were subsequently analysed by HPLC-ESI-QTOF-MS through an untargeted metabolomic approach.

On the other hand, **SECTION B** focuses on the application of untargeted metabolomic strategies for the study of **systemic autoimmune diseases**. Firstly, the main characteristics of this type of illnesses are briefly described. Specifically, seven of these diseases have been studied in this section: systemic lupus erythematosus, rheumatoid arthritis, systemic sclerosis, Sjögren's syndrome, antiphospholipid syndrome, mixed connective tissue disease and undifferentiated connective tissue disease. This section includes four chapters, of which two of them (chapters 6 and 7) were carried out in collaboration with the University of Chalmers (Sweden) thanks to a three-month predoctoral stay in 2018 funded by the Spanish Ministry of Education, Culture and Sports ("Estancias breves FPU").

In **Chapter 4**, the metabolic alterations in plasma and urine samples from 59 systemic sclerosis patients were studied. Similarly, **Chapter 5** focused on the study of Sjögren's syndrome in samples collected from 43 patients. In both chapters, the data pre-processing step was performed using the commercial software of the analytical platform used for the analysis (HPLC-ESI-QTOF-MS), *Agilent Mass Profinder*. Given the observed difficulty in preprocessing the metabolomic data acquired from a large number of samples with this type of software, in **Chapter 6**, two pre-processing methodologies were compared for application in untargeted metabolomic studies comprising a high number of samples. These methodologies were based on the use of the aforementioned commercial software and open sources based on packages developed in R language.

Finally, **Chapter 7** focused on the study of seven systemic autoimmune diseases through metabolomic analysis of urine and plasma samples collected from 228 patients and 55 healthy controls. In this chapter, the data-preprocessing step was performed using the methodology optimized in the previous chapter based on R language.

OBJETIVOS

OBJECTIVES

OBJETIVOS

El empleo de estrategias metabolómicas en diferentes áreas del conocimiento ha experimentado un aumento exponencial a lo largo de la última década, siendo la metabolómica una de las áreas de la ciencia más activas en la actualidad.

Por ello, el **objetivo principal** de la presente tesis doctoral es llevar a cabo el desarrollo, optimización y aplicación de estrategias metabolómicas no dirigidas en el ámbito de la alimentación funcional así como en el estudio de diferentes patologías autoinmunes. Para ello, abordar las diferentes etapas del flujo de trabajo relacionadas con el tratamiento de muestra, la adquisición de datos mediante HPLC-ESI-QTOF-MS, el pre-procesamiento de datos, el análisis estadístico, la identificación de metabolitos así como la interpretación biológica de los resultados.

En concreto, la investigación llevada a cabo en la presente memoria se focaliza, por una parte, en el estudio de los compuestos bioactivos presentes en matrices alimentarias desde un punto de vista metabolómico, y por otra parte, en el estudio de siete enfermedades autoinmunes sistémicas mediante la búsqueda de biomarcadores específicos.

Para abordar este objetivo general, se establecieron los siguientes **objetivos específicos**:

- Estudiar la absorción y el metabolismo de los compuestos fenólicos presentes en un extracto de romero con propiedades bioactivas a través de un estudio de perfusión intestinal llevado a cabo en ratas Wistar.
- Conocer el efecto en el metabolismo que produce el consumo prolongado de dos extractos ricos en compuestos bioactivos obtenidos a partir de mango y

ajo, a través de estudios de intervención nutricional usando un modelo animal y humano, respectivamente.

- Optimizar metodologías de pre-procesamiento de datos adquiridos mediante HPLC-ESI-QTOF-MS para su aplicación en estudios de metabolómica no dirigida con un elevado número de muestras. Así, evaluar y comparar la eficiencia para ejecutar esta etapa de pre-procesamiento de datos de un software comercial y herramientas de acceso libre basadas en R.
- Identificar las alteraciones metabólicas que permitan aumentar el conocimiento de diferentes enfermedades autoinmunes sistémicas que comparten mecanismos patofisiológicos comunes, como el lupus eritematoso sistémico, la esclerosis sistémica progresiva o escleroderma, el síndrome de Sjögren, la artritis reumatoide, el síndrome antifosfolípido, la enfermedad mixta del tejido conectivo y la enfermedad mixta del tejido conectivo indiferenciado mediante análisis de huella dactilar metabólica en plasma y orina empleando HPLC-ESI-QTOF-MS.

OBJECTIVES

The use of metabolomic strategies in different fields of knowledge has experienced an exponential increase over the last decade. In fact, metabolomics is currently one of the most active areas of science.

Therefore, the **main objective** of this PhD thesis is to carry out the development, optimization and application of untargeted metabolomic strategies in the field of functional foods as well as in the study of classification of different autoimmune diseases. In this sense, this objective requires approaching the different steps of the workflow related to sample treatment, data acquisition by HPLC-ESI-QTOF-MS, data pre-processing, statistical analysis, metabolite identification as well as biological interpretation.

Specifically, the research carried out herein focuses, on the one hand, on the study of bioactive compounds present in food matrices from a metabolomic point of view, and on the other hand, on the study of seven systemic autoimmune diseases by searching for specific biomarkers.

To address this general objective, the following **specific objectives** were established:

- To study the absorption and metabolism of the phenolic compounds present in a rosemary extract with bioactive properties through a study of intestinal perfusion carried out in Wistar rats.
- To know the effect on the metabolism produced by the prolonged intake of two extracts rich in bioactive compounds obtained from mango and garlic, through nutritional intervention studies using an animal and human models, respectively.

- To optimize data preprocessing methodologies for application in untargeted metabolomic studies with a large number of samples using HPLC-ESI-QTOF-MS. In this sense, an evaluation and comparison of the efficiency to perform the data pre-processing step has been carried out between a commercial software and an open access pipeline based on R.
- To identify the metabolic alterations that allow increasing the knowledge of seven systemic autoimmune diseases that share common pathophysiological mechanisms (systemic lupus erythematosus, progressive sclerosis or scleroderma, Sjögren's syndrome, rheumatoid arthritis, antiphospholipid syndrome, mixed disease of connective tissue and mixed disease of undifferentiated connective tissue) by HPLC-ESI-QTOF-MS analysis of the metabolic fingerprint in plasma and urine samples.

INTRODUCCIÓN

1. CIENCIAS "ÓMICAS"

Las **ciencias "ómicas"** son un conjunto de técnicas y tecnologías cuya finalidad es la obtención de una visión lo más completa posible acerca de los sistemas biológicos en su conjunto. Es decir, su principal objetivo es identificar, caracterizar y cuantificar todas las moléculas involucradas en todos los procesos biológicos de un organismo. Este conjunto de ciencias están formadas principalmente por la Genómica, la Transcriptómica, la Proteómica y la Metabolómica[1]. Además cabe destacar que ha surgido todo un conjunto de ciencias ómicas relacionadas con ellas, como por ejemplo la Epigenómica, la Microbiómica o la Lipidómica, entre otras[2].

Concretamente, la **Genómica** es la ciencia que se encarga del estudio del conjunto de genes (genotipo) de un organismo. Por su parte, la **Transcriptómica** se focaliza en las moléculas de RNA originadas en la transcripción del DNA y encargadas de la síntesis de proteínas, las cuales son estudiadas por la **Proteómica**[3]. Finalmente, la **Metabolómica** estudia el conjunto de moléculas de bajo peso molecular denominadas metabolitos[4]. Esta ciencia "ómica" será descrita en detalle durante los próximos apartados, al ser la aproximación empleada en el desarrollo de la presente tesis doctoral.

[1] Mario Vailati-Riboni, Valentino Palombo, and Juan J. Loor, "What Are Omics Sciences?," in *Periparturient Diseases of Dairy Cows* (Cham: Springer International Publishing, 2017), 1–7, https://doi.org/10.1007/978-3-319-43033-1_1.

[2] Yehudit Hasin, Marcus Seldin, and Aldons Lusis, "Multi-Omics Approaches to Disease.," *Genome Biology* 18, no. 1 (2017): 83, https://doi.org/10.1186/s13059-017-1215-1.

[3] Claudia Manzoni et al., "Genome, Transcriptome and Proteome: The Rise of Omics Data and Their Integration in Biomedical Sciences," *Briefings in Bioinformatics* 19, no. 2 (March 1, 2018): 286–302, https://doi.org/10.1093/bib/bbw114.

[4] A. Agin et al., "Metabolomics - an Overview. From Basic Principles to Potential Biomarkers (Part 1)," *Medecine Nucleaire*, February 2016, https://doi.org/10.1016/j.mednuc.2015.12.006.

Aunque las distintas "ómicas" estudian los procesos biológicos desde perspectivas distintas, todas ellas están interrelacionadas entre sí gracias a las interacciones existentes entre genes, transcritos, proteínas y metabolitos en dichos procesos, formando la denominada **cascada de las ómicas**[5]. Los procesos estudiados por las distintas ciencias ómicas presentan una mayor relación con la respuesta fenotípica de un organismo a medida que se va descendiendo por dicha cascada tal y como se muestra en la **Figura 1.** Dada la relación existente entre las diferentes "ómicas", y para tener una compresión más holística de los procesos biológicos que tienen lugar en los organismos vivos, cada vez son más comunes los estudios que integran datos obtenidos por diferentes enfoques "ómicos"[6].

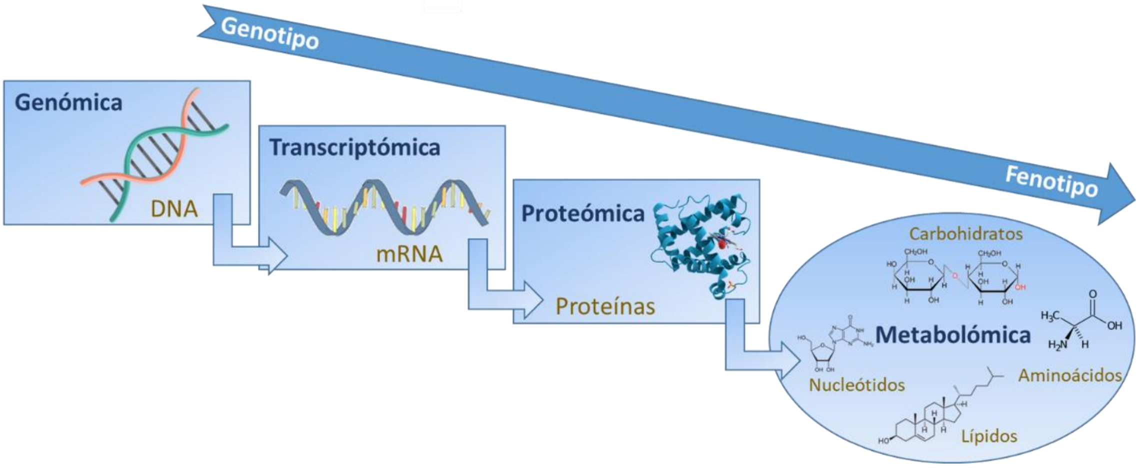


Figura 1. La cascada de las ciencias ómicas.

[5] Kamil Jurowski et al., "Analytical Techniques in Lipidomics: State of the Art," *Critical Reviews in Analytical Chemistry* 47, no. 5 (September 3, 2017): 418–37, https://doi.org/10.1080/10408347.2017.1310613.

[6] Yehudit Hasin, Marcus Seldin, and Aldons Lusis, "Multi-Omics Approaches to Disease," accessed August 29, 2019, https://doi.org/10.1186/s13059-017-1215-1.

1.1. Metabolómica

La **metabolómica** es una disciplina de reciente aparición, desarrollo y aplicación entre todas las ciencias ómicas, cuyas primeras definiciones datan de principios del presente siglo. Concretamente, *Oliver Fiehn* definió en 2001 la metabolómica como la caracterización cualitativa y cuantitativa de todos los metabolitos presentes en un sistema biológico (**metaboloma**)[7]. El término metaboloma, que apareció por primera vez en 1998*,* se define como el conjunto de metabolitos (moléculas de peso molecular inferior a 1500 Dalton) presentes en células, tejidos o fluidos biológicos, que son productos o intermedios de los procesos químicos o enzimáticos resultado del metabolismo celular[8]. El metaboloma se clasifica a su vez en los siguientes subtipos: **metaboloma endógeno**, que engloba los metabolitos intrínsecos del propio organismo, y **metaboloma exógeno**, relacionado con los metabolitos procedentes de factores extrínsecos al sistema biológico tales como la alimentación, medicación o factores ambientales (calidad del agua, aire, contaminación, etc.)[9].

Dada la posición de la metabolómica dentro de cascada de las "ómicas", el metaboloma es capaz de reflejar la respuesta biológica producida por multitud de factores como alteraciones genéticas, estados patológicos, influencias medioambientales o dietéticas, entre otros (**Figura 2**). De esta forma, el estado de una célula, y por tanto la **respuesta fenotípica** exhibida por un organismo, está

[7] O Fiehn, "Combining Genomics, Metabolome Analysis, and Biochemical Modelling to Understand Metabolic Networks.," *Comparative and Functional Genomics* 2, no. 3 (2001): 155–68, https://doi.org/10.1002/cfg.82.

[8] Aline Klassen et al., "Metabolomics: Definitions and Significance in Systems Biology," 2017, 3–17, https://doi.org/10.1007/978-3-319-47656-8_1.

[9] Augustin Scalbert et al., "The Food Metabolome: A Window over Dietary Exposure," *The American Journal of Clinical Nutrition* 99, no. 6 (June 1, 2014): 1286–1308, https://doi.org/10.3945/ajcn.113.076133.

directamente relacionado con el perfil metabólico o nivel de metabolitos del mismo[10]. Por lo tanto, la metabolómica tiene la capacidad de detectar pequeños cambios en rutas metabólicas y la alteración de la homeostasis incluso antes de que sea posible detectar ningún cambio en el organismo a nivel fenotípico. Este hecho es posible gracias a que los metabolitos no son simplemente los productos enzimáticos de las reacciones bioquímicas sino que, de manera integrada, forman parte de la regulación de los procesos bioquímicos que tienen lugar en los sistemas biológicos[11].

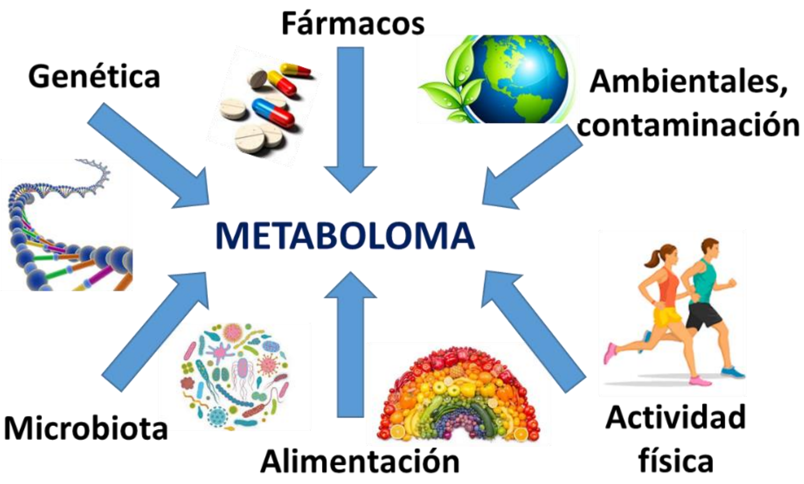


Figura 2. Factores que afectan al metaboloma.

El objetivo principal que se marca la metabolómica es el análisis exhaustivo de muestras biológicas con la finalidad de lograr la completa caracterización y cuantificación de todos los compuestos presentes en el metaboloma. Sin embargo, el estudio del metaboloma es un reto analítico de gran envergadura debido a que este

[10] Stephen J. Bruce et al., “A Plasma Global Metabolic Profiling Approach Applied to an Exercise Study Monitoring the Effects of Glucose, Galactose and Fructose Drinks during Post-Exercise Recovery,” *Journal of Chromatography B* 878, no. 29 (2010): 3015–23, https://doi.org/10.1016/j.jchromb.2010.09.004.

[11] Sastia P. Putri et al., “Current Metabolomics: Technological Advances,” *Journal of Bioscience and Bioengineering* 116, no. 1 (2013): 9–16, https://doi.org/10.1016/j.jbiosc.2013.01.004.

contiene un enorme número de componentes procedentes de diferentes familias de metabolitos (aminoácidos, lípidos, ácidos nucleicos, nucleótidos, etc.), los cuales presentan una alta diversidad en cuanto a sus propiedades físicas y químicas, pudiendo encontrase además en un amplio rango de concentraciones[12]. Dada la complejidad de analizar el metaboloma global de un individuo, la metabolómica se considera como una nueva **área de la ciencia** en lugar de una aproximación analítica[13]. Desde esta nueva perspectiva, la metabolómica se define como el conjunto de ciencias integradas con el objetivo de identificar y cuantificar el conjunto de metabolitos presentes en el metaboloma de un sistema biológico[8].

Debido al potencial de esta "ómica", el número de investigaciones que emplean una aproximación metabolómica ha aumentado exponencialmente en la última década (**Figura 3**). Su mayor interés reside en su contribución en estudios clínicos ya que permite asociar los perfiles metabólicos a distintas situaciones fisiológicas y fisiopatológicas de un organismo. De hecho, está siendo crucial en el estudio y diagnóstico de numerosas enfermedades, habiéndose demostrado ya una gran eficacia en la detección de afecciones respiratorias, diabetes, enfermedades cardiovasculares o cáncer, entre otras patologías. Además, la metabolómica tiene gran aplicabilidad en

[12] Katja Dettmer, Pavel A Aronov, and Bruce D Hammock, "Mass Spectrometry-Based Metabolomics.," *Mass Spectrometry Reviews* 26, no. 1 (January 2007): 51–78, https://doi.org/10.1002/mas.20108.

[13] Silas G Villas-Bôas, Susanne Rasmussen, and Geoffrey A Lane, "Metabolomics or Metabolite Profiles?," *Trends in Biotechnology* 23, no. 8 (August 1, 2005): 385–86, https://doi.org/10.1016/j.tibtech.2005.05.009.

multitud de áreas, como por ejemplo en estudios de toxicología, farmacología, nutrición, biotecnología o tecnología de los alimentos, entre otros[14].

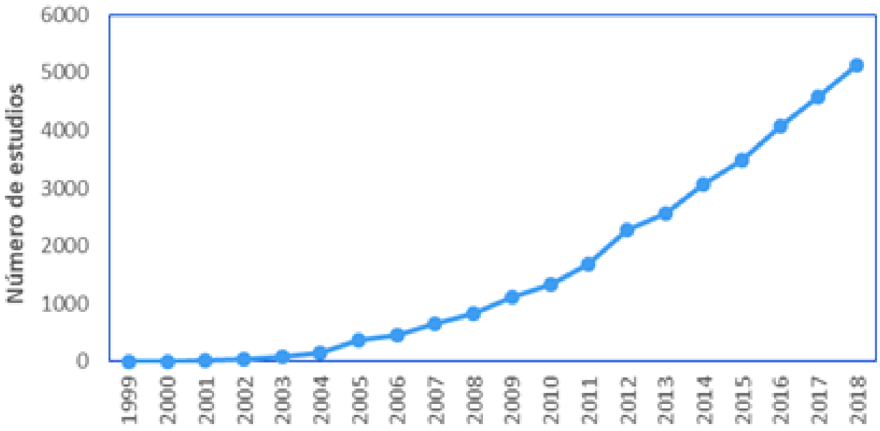


Figura 3. Evolución del número de estudios en el área de metabolómica desde 1999 (Fuente: *Scopus*).

2. ESTRATEGIAS METABOLÓMICAS

Debido a la enorme dificultad de llevar a cabo un estudio metabolómico completo, existen distintas aproximaciones analíticas que pueden ayudar a resolver ciertas cuestiones específicas. Estas aproximaciones en el campo de la metabolómica se engloban principalmente en dos tipos de estudios: **dirigidos** y **no dirigidos**. La principal diferencia entre ambas aproximaciones reside en si los metabolitos son definidos, o no, de manera previa a los análisis, respectivamente. La **Tabla 1** recoge las principales características de ambos enfoques metodológicos[15].

[14] M. V. Gomez-Casati, D. F.; Zanor, M. I.; Busi, "Metabolomics in Plants and Humans: Applications in the Prevention and Diagnosis of Diseases," *BioMed. Res. Int.* 2013 (2013): 1–11, https://doi.org/10.1155/2013/792527.

[15] Alexandra C. Schrimpe-Rutledge et al., "Untargeted Metabolomics Strategies—Challenges and Emerging Directions," *Journal of the American Society for Mass Spectrometry*, 2016, https://doi.org/10.1007/s13361-016-1469-y.

A pesar de tener características diferentes, existen estudios donde se utilizan ambas metodologías de manera complementaria, como por ejemplo, estudios que pretenden obtener una visión mucho más completa del metaboloma de un organismo[16], o la utilización de una metodología dirigida como herramienta de validación de los resultados obtenidos mediante una estrategia no dirigida[17].

Tabla 1. Características de las estrategias metabolómicas dirigidas y no dirigidas.

Estrategias dirigidas	Estrategias no dirigidas
Análisis selectivo de un número específico de metabolitos predefinidos	Análisis global del metaboloma
Metabolitos conocidos previamente	Identificación de las señales de interés mediante MS/MS o NMR
Cuantificación absoluta	Cuantificación relativa
Validación de biomarcadores	Descubrimiento de nuevos biomarcadores
Mayor sensibilidad y sencillez en el procesado de datos	Menos sensibilidad y mayor complejidad en el procesamiento datos

[16] Li Liang et al., "Integrating Targeted and Untargeted Metabolomics to Investigate the Processing Chemistry of Polygoni Multiflori Radix," *Frontiers in Pharmacology*, 2018, https://doi.org/10.3389/fphar.2018.00934.

[17] Farhana R. Pinu et al., "Translational Metabolomics: Current Challenges and Future Opportunities," *Metabolites* 9, no. 6 (June 6, 2019): 108, https://doi.org/10.3390/metabo9060108.

2.1. Estrategias dirigidas

Las **estrategias de metabolómica dirigidas** tienen como objetivo la identificación y cuantificación de un número limitado de metabolitos específicos en muestras biológicas, los cuáles han sido previamente definidos. Generalmente, estos metabolitos son preestablecidos de acuerdo al objetivo del estudio o en función de los compuestos descritos en estudios previos, kits comerciales o librerías de software comerciales. Este tipo de metodologías presentan una alta sensibilidad así como una mayor sencillez en el manejo y procesamiento de datos[18].

Dependiendo del número de metabolitos analizados mediante las estrategias dirigidas, se han establecido las siguientes aproximaciones[19]:

- **Análisis diana** o **dirigido**, cuyo objetivo se centra exclusivamente en determinar un número muy pequeño de metabolitos (<10) concretos que resulten de interés, como puede ser un producto de una reacción enzimática o un biomarcador, los cuales se entienden como cualquier característica útil que puede ser medida y utilizada como un indicador de un proceso patológico o fisiológico particular. Este tipo de metodología utiliza las técnicas más apropiadas en función de los metabolitos que se desean analizar.
- **Perfil metabólico**, se restringe a la identificación y cuantificación de un número predefino de metabolitos (entre 10 y 100 metabolitos) que pueden pertenecer

[18] Lee D Roberts et al., "Targeted Metabolomics.," *Current Protocols in Molecular Biology* Chapter 30 (April 2012): Unit 30.2.1-24, https://doi.org/10.1002/0471142727.mb3002s98.

[19] M S Monteiro et al., "Metabolomics Analysis for Biomarker Discovery: Advances and Challenges.," *Current Medicinal Chemistry* 20, no. 2 (2013): 257–71, https://doi.org/CMC-EPUB-20121126-5 [pii].

a una clase de compuestos específica o intervenir en una ruta metabólica concreta.

2.2. Estrategias no dirigidas

Las **estrategias de metabolómica no dirigidas** tienen como objetivo analizar el perfil metabolómico completo así como la búsqueda de metabolitos diferenciadores entre grupos de estudio. Su principal ventaja reside en la capacidad para caracterizar nuevos metabolitos y en la potencialidad para la identificación de biomarcadores.

Este tipo de metodologías se basan en la detección del mayor número posible de señales, con sentido biológico, obtenidas mediante técnicas analíticas avanzadas de alta resolución, para su posterior identificación mediante el uso de bases de datos metabolómicas. A pesar de que se ha producido un gran avance en la expansión de estas bases de datos durante los últimos años, la etapa de identificación de los metabolitos sigue siendo actualmente la principal limitación de este tipo de estrategias, dado que una parte significativa de las señales no pueden ser identificadas debido a la ausencia de información estructural en dichas bases de datos[20].

Dependiendo del enfoque del estudio, dentro de la metabolómica no dirigida, se diferencian las siguientes aproximaciones:

- **Metabolómica**, *per se*, en la que se realiza el análisis exhaustivo de todos los metabolitos de un sistema biológico para su identificación y cuantificación, revelando de esta forma el metaboloma del sistema biológico bajo estudio.

[20] Kerem Bingol, "Recent Advances in Targeted and Untargeted Metabolomics by NMR and MS/NMR Methods," *High-Throughput* 7, no. 2 (April 18, 2018): 9, https://doi.org/10.3390/ht7020009.

- **Huella dactilar metabólica**, medida global de los metabolitos presentes en un sistema biológico ("*metabolic fingerprint*") y/o consumidos o secretados (exometaboloma, "*metabolomic footprint*") por el mismo con el objetivo de realizar una clasificación de las muestras, de forma rápida, según su origen o relevancia biológica a través del empleo de análisis estadísticos multivariantes, sin necesidad de determinar la concentración individual de cada metabolito[19].

3. FLUJO DE TRABAJO EN ESTUDIOS DE METABOLÓMICA NO DIRIGIDA

Dada la complejidad de los estudios de metabolómica no dirigida, su aplicación requiere llevar a cabo una serie de etapas con una alta precisión y exactitud con el objetivo de obtener resultados de calidad que ayuden a interpretar y resolver las cuestiones biológicas planteadas en este tipo de estudios. La mayor parte de los estudios no dirigidos siguen un **flujo de trabajo** característico (**Figura 4**) formado por las siguientes etapas[21-23]:

1) La primera hace referencia al **diseño del estudio**. Como todo método científico, los primeros pasos para llevar a cabo un estudio en metabolómica, corresponden a la observación de lo que se pretende estudiar para el posterior planteamiento de una hipótesis y unos objetivos. Una vez definidos estos

[21] Arnald Alonso, Sara Marsal, and Antonio Julià, "Analytical Methods in Untargeted Metabolomics: State of the Art in 2015.," *Frontiers in Bioengineering and Biotechnology* 3 (2015): 23, https://doi.org/10.3389/fbioe.2015.00023.

[22] Marynka M. Ulaszewska et al., "Nutrimetabolomics: An Integrative Action for Metabolomic Analyses in Human Nutritional Studies," *Molecular Nutrition & Food Research* 63, no. 1 (January 1, 2019): 1800384, https://doi.org/10.1002/mnfr.201800384.

[23] Yiman Wu and Liang Li, "Sample Normalization Methods in Quantitative Metabolomics," *Journal of Chromatography A* 1430 (January 22, 2016): 80–95, https://doi.org/10.1016/J.CHROMA.2015.12.007.

objetivos, se lleva a cabo el diseño experimental para lograr la consecución de los mismos.

2) Una vez diseñado el estudio, los siguientes pasos corresponden a la **toma y tratamiento de muestras biológicas**. En esta etapa, se recolectan, distribuyen y conservan las muestras biológicas escogidas para el estudio, además de procesarlas de una manera adecuada para hacerlas compatibles con la técnica de análisis que se va a emplear a continuación.
3) Mediante el empleo de sofisticadas plataformas analíticas avanzadas de alta resolución, se lleva a cabo la **adquisición de datos** a partir de los análisis de las muestras biológicas del estudio.
4) A continuación, el **procesamiento de datos** tiene como propósito reducir la complejidad de los datos adquiridos, la extracción de las señales detectadas, así como la transformación de datos para su adecuación a las técnicas estadísticas[24].
5) Posteriormente, los **análisis estadísticos** persiguen la identificación de las señales que presentan efectos significativos de acuerdo a los objetivos formulados en el estudio.
6) El siguiente paso corresponde a la **identificación** de los metabolitos, que tiene como objetivo la asignación de una estructura química a las señales que resultaron estadísticamente significativas en la etapa anterior.

[24] Sandra Castillo et al., "Algorithms and Tools for the Preprocessing of LC–MS Metabolomics Data," *Chemometrics and Intelligent Laboratory Systems* 108, no. 1 (August 15, 2011): 23–32, https://doi.org/10.1016/J.CHEMOLAB.2011.03.010.

7) Finalmente, se lleva a cabo la **interpretación biológica** de los resultados obtenidos pretendiendo darles sentido para que ayuden a resolver la hipótesis y los objetivos planteados en la primera etapa del estudio.

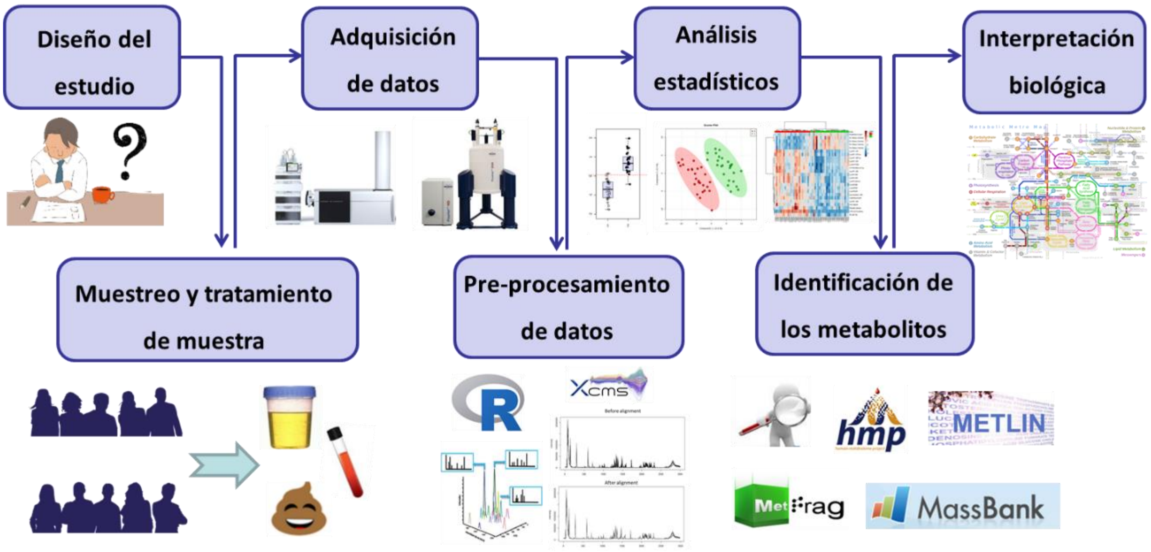


Figura 4. Flujo de trabajo en estudios de metabolómica no dirigida.

Dado que la espectrometría de masas acoplada a cromatografía de líquidos de alta resolución (HPLC-MS) ha sido la plataforma analítica utilizada en la presente tesis doctoral para el análisis metabolómico de las muestras objeto de estudio, los siguientes subapartados describen detalladamente las distintas etapas del flujo de trabajo enfocadas para la utilización de esta plataforma analítica.

3.1. Diseño de estudios

La etapa inicial del flujo de trabajo se centra en la formulación de la **hipótesis** sobre el caso biológico que se quiere estudiar. Este paso es de crucial importancia dado que conlleva la definición de los **objetivos** del estudio así como el **diseño experimental** para lograr la consecución de los mismos. El diseño del estudio implica la definición de las siguientes condiciones: tipo de aproximación metabolómica, tipo de estudio (agudo, longitudinal…), plataformas analíticas, tipo de muestras biológicas (fluidos biológicos, tejidos, células y/o organismos intactos), tipo de muestreo, tamaño muestral (número de muestras a evaluar, réplicas), condiciones de toma, transporte y conservación de las muestras así como los protocolos de preparación de muestra. Dado que existe una interrelación entre todas estas variables (p.ej.: el tratamiento de muestra debe estar en consonancia con el tipo de muestra biológica y la técnica analítica), todas ellas deben definirse en esta primera etapa con el objetivo de garantizar la calidad de los resultados que ayuden a resolver la hipótesis biológica definida[8].

En este sentido, los estudio de metabolómica que utilizan muestras de origen humano o modelos animales, deben estar previamente aprobados por un comité de ética para poder llevar a cabo su realización, garantizándose así el cumplimiento de la legislación vigente sobre los principios éticos a respetar en investigación. Actualmente, la Declaración de Helsinki es el documento internacional más importante que regula la investigación con seres humanos[25].

En general, los estudios de metabolómica, y especialmente los que emplean una estrategia no dirigida, tienen un **carácter comparativo**. Por lo tanto, se requiere la definición de uno o varios conjuntos de muestras que cumplan las condiciones que se quieren investigar, y por otra parte, un grupo de muestras control para poder realizar

[25] "World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects," *Journal of the Korean Medical Association*, 2014, https://doi.org/10.5124/jkma.2014.57.11.899.

la comparación entre los perfiles metabólicos. La etapa de selección de los voluntarios requiere la consideración de la **variabilidad biológica** existente entre individuos debido a distintos factores como el género, la edad, el índice de masa corporal (BMI), el estado de salud, el consumo de fármacos, factores genéticos o el estilo de vida (alimentación, actividad física, hábito de fumar, etc.), entre otros. La consideración de esta variabilidad biológica es fundamental dado que puede producir errores sistemáticos en los resultados que den lugar a interpretaciones erróneas de la hipótesis planteada[22,26].

3.2. Muestreo y tratamiento de muestras biológicas

El metaboloma puede ser explorado en variedad de tejidos y fluidos biológicos (plasma, suero, orina, heces, saliva, sudor, líquido cefalorraquídeo, esperma, fluido gastrointestinal, etc.) cuya elección depende principalmente de la cuestión biológica que se quiere abordar así como de la facilidad de toma de muestra de las diferentes matrices biológicas. Por ejemplo, los fluidos biológicos suelen utilizarse generalmente para identificar biomarcadores, mientras que los tejidos y las células se utilizan para investigar sobre los mecanismos asociados a los procesos fisiopatológicos[27].

[26] Abdul-Hamid Emwas et al., “Standardizing the Experimental Conditions for Using Urine in NMR-Based Metabolomic Studies with a Particular Focus on Diagnostic Studies: A Review.,” *Metabolomics : Official Journal of the Metabolomic Society* 11, no. 4 (2015): 872–94, https://doi.org/10.1007/s11306-014-0746-7.

[27] Andrew J. Chetwynd, Warwick B. Dunn, and Giovanny Rodriguez-Blanco, “Collection and Preparation of Clinical Samples for Metabolomics” (Springer, Cham, 2017), 19–44, https://doi.org/10.1007/978-3-319-47656-8_2.

Las muestras de **sangre** (suero y/o plasma)[28], **orina**[29] y **heces**[30] han sido las más estudiadas y empleadas en metabolómica debido a su facilidad de recolección al ser menos invasivas, además de por la gran cantidad de metabolitos que contienen. Estas muestran proporcionan información complementaria sobre el metaboloma y por tanto, del estado en el que se encuentra un organismo. Por ejemplo, el plasma sanguíneo es un reflejo del estado metabólico que proporciona información directa sobre los procesos catabólicos y anabólicos que ocurren en todo el organismo. Sin embargo, las muestras de heces y orina proporcionan información relativa a xenobióticos y metabolitos excretados como resultado de los procesos catabólicos[31].

La etapa de **recolección de muestras biológicas** debe ser llevada a cabo según el tipo de estudio a realizar y debe tener en cuenta distintos aspectos que pueden influir en la calidad de los resultados, como por ejemplo la hora del muestreo, si los voluntarios están en ayunas, la dieta que han seguido los días previos, el material para recoger las muestras, o el número de muestras por paciente, entre otros.

Tras la etapa de recolección de muestras, estas deben transportarse y almacenarse a temperatura controlada manteniendo la cadena de frío para la correcta **conservación** de todos los metabolitos así como del resto de componentes presentes en la muestra.

[28] Séverine Trabado et al., "The Human Plasma-Metabolome: Reference Values in 800 French Healthy Volunteers; Impact of Cholesterol, Gender and Age," ed. Andrea Motta, *PLOS ONE* 12, no. 3 (March 9, 2017): e0173615, https://doi.org/10.1371/journal.pone.0173615.

[29] Souhaila Bouatra et al., "The Human Urine Metabolome," ed. Petras Dzeja, *PLoS ONE* 8, no. 9 (September 4, 2013): e73076, https://doi.org/10.1371/journal.pone.0073076.

[30] Naama Karu et al., "A Review on Human Fecal Metabolomics: Methods, Applications and the Human Fecal Metabolome Database," *Analytica Chimica Acta* 1030 (November 7, 2018): 1–24, https://doi.org/10.1016/j.aca.2018.05.031.

[31] M.A. Fernández-Peralbo and M.D. Luque de Castro, "Preparation of Urine Samples Prior to Targeted or Untargeted Metabolomics Mass-Spectrometry Analysis," *TrAC Trends in Analytical Chemistry* 41 (December 1, 2012): 75–85, https://doi.org/10.1016/J.TRAC.2012.08.011.

La temperatura de conservación para la mayoría de muestras biológicas debe ser inferior a -18 ºC, siendo generalmente almacenadas para periodos más largos a -80 ºC. Esta etapa es crucial para mantener los metabolitos inalterados en las muestras biológicas, minimizando los procesos de degradación y asegurando la veracidad de los resultados obtenidos en los análisis realizados a posteriori. La temperatura de conservación de la muestra es un parámetro determinante para establecer el tiempo de conservación durante el cual esta se considera óptima para su estudio[32,33].

Por su parte, la etapa de **tratamiento de la muestra** tiene como principal objetivo la adecuación de la misma a la plataforma analítica elegida para el estudio a la vez que debe mantener inalterada en el mayor grado posible la composición original del metaboloma presente en la muestra original. Por ello, el tipo de tratamiento debe ser simple, rápido, no selectivo y reproducible. Además, se puede incluir una etapa de inhibición enzimática (*quenching*) para impedir la reacción de los metabolitos presentes y que los resultados se correspondan con el perfil metabólico en el momento del muestreo[31].

Todas estas etapas tienen un impacto significativo en la calidad de los datos adquiridos en posteriores fases del flujo de trabajo y en consecuencia, en la interpretación de los resultados y en las conclusiones derivadas del estudio. Por lo tanto, todos los aspectos relacionados con el tipo, la recolección y el tratamiento de la muestra, deben ser bien establecidos para garantizar la reproducibilidad y la calidad de los resultados.

[32] Serap Cuhadar et al., “The Effect of Storage Time and Freeze-Thaw Cycles on the Stability of Serum Samples,” *Biochemia Medica*, 2013, https://doi.org/10.11613/BM.2013.009.

[33] Peiyuan Yin et al., “Preanalytical Aspects and Sample Quality Assessment in Metabolomics Studies of Human Blood,” *Clinical Chemistry*, 2013, https://doi.org/10.1373/clinchem.2012.199257.

A continuación se describen las principales características de las muestras biológicas que se han utilizado principalmente en esta memoria, plasma sanguíneo y orina, así como las consideraciones para llevar a cabo sus tratamientos.

3.2.1. Tratamiento de muestras de plasma sanguíneo

El **plasma sanguíneo** es el componente principal de la sangre, representando el 55 % del volumen total. Corresponde a la parte líquida y acelular de la sangre que no contiene los elementos formes (eritrocitos, leucocitos y plaquetas). La diferencia respecto al suero sanguíneo radica en que la composición del plasma sanguíneo contiene los factores de coagulación, como el fibrinógeno.

El plasma está compuesto por un 90 % de agua y múltiples sustancias disueltas en ella, de las cuales las más abundantes son las proteínas. Además de estas, su composición está formada por glúcidos, lípidos, hormonas, sales, gases disueltos y productos de desecho del metabolismo.

El plasma sanguíneo se obtiene tras una etapa de centrifugación de la sangre mezclada con un anticoagulante, como por ejemplo heparina, citrato sódico o ácido etilendiaminotetraacético libre (EDTA) o en forma de sal sódica o potásica (Na_2EDTA, K_2EDTA, K_3EDTA). Tras la etapa de centrifugación, el sobrenadante corresponde al plasma sanguíneo, la fracción intermedia a los leucocitos y la fracción inferior a los eritrocitos. En la **Figura 5** se muestra un esquema de la obtención del plasma sanguíneo[34].

[34] Melissa K. Tuck et al., “Standard Operating Procedures for Serum and Plasma Collection: Early Detection Research Network Consensus Statement Standard Operating Procedure Integration Working Group,” *Journal of Proteome Research*, 2009, https://doi.org/10.1021/pr800545q.

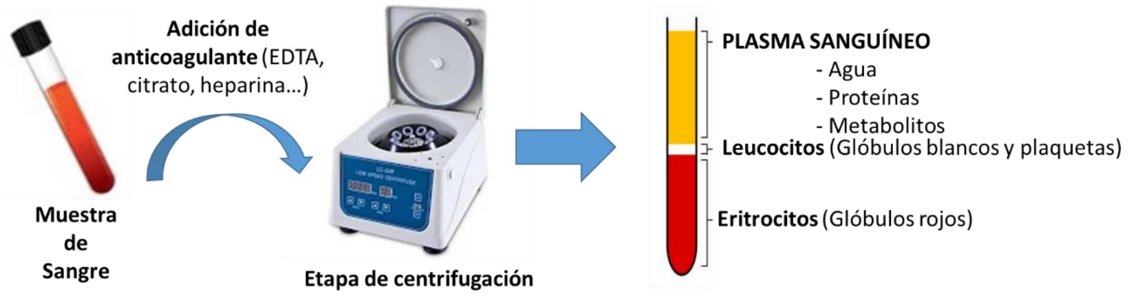

Adición de anticoagulante (EDTA, citrato, heparina...)
Muestra de Sangre
Etapa de centrifugación
PLASMA SANGUÍNEO
- Agua
- Proteínas
- Metabolitos
Leucocitos (Glóbulos blancos y plaquetas)
Eritrocitos (Glóbulos rojos)

Figura 5. Proceso de obtención de plasma sanguíneo.

Dada la composición del plasma sanguíneo, la etapa de preparación de este tipo muestra enfocada a un análisis metabolómico mediante LC-MS tiene como objetivo principal la precipitación y eliminación de las proteínas de la matriz biológica al ser grandes interferentes en este tipo de análisis e incompatibles con ciertos modos de trabajo. Esta etapa es primordial para lograr una eficiente extracción de los metabolitos presentes que asegure una buena calidad de los resultados, así como para alargar el tiempo de vida de la columna cromatográfica, ya que las proteínas pueden quedar retenidas de manera irreversible en el relleno, imposibilitando su uso. Existen diferentes procedimientos para lograr la **precipitación de proteínas,** basados en su desnaturalización, utilizando para ello un aumento de temperatura, una disminución del pH por medio de la adicción de ácidos, o mediante la adicción de disolventes orgánicos. Entre estos procedimientos, la metodología que ha mostrado una mayor eficacia para llevar a cabo esta etapa de precipitación de proteínas, es la basada en la adicción de **disolventes orgánicos**, siendo los más utilizados acetonitrilo, metanol, etanol, acetona o una mezcla de ellos. *Stephen J. Bruce et al.*[10] llevaron a cabo un estudio de optimización de esta etapa donde se demostró que las combinaciones de disolventes orgánicos que mostraron una mayor eficiencia fueron: metanol/etanol (1:1, v/v) o metanol/acetonitrilo/acetona (1:1:1, v/v/v). Otros aspectos a tener en cuenta en esta etapa son la temperatura, el tiempo o la relación muestra:disolvente.

3.2.2. Tratamiento de muestras de orina

La **orina** es un fluido biológico secretado por los riñones como resultado del proceso de depuración y filtrado de la sangre, que es acumulado en la vejiga y expulsado a través de la uretra. La composición química normal de la orina consiste principalmente en un 95 % de agua, e incluye sales inorgánicas (Cl^-, Na^+, K^+, NH_4^+), proteínas, moléculas nitrogenadas (urea, ácido úrico, creatinina), además de otros productos derivados de la degradación del metabolismo de fármacos, alimentos, etc. tras su paso por los riñones y el hígado[29].

Las principales ventajas del empleo de este tipo de muestra radican en que el muestreo no es invasivo, y requiere un menor número de etapas de tratamiento tras la recolección para su análisis mediante LC-MS. Esto es debido a que comparada con otras muestras como suero o plasma, presenta una composición menos compleja al tener un menor contenido en proteínas u otros metabolitos de alto peso molecular. Además, el metaboloma urinario abarca un mayor número de compuestos en comparación con cualquier otra matriz biológica, ya que contiene numerosos metabolitos derivados de la metabolización de alimentos, bebidas, fármacos, contaminantes ambientales, subproductos bacterianos además de productos del metabolismo endógeno[31,35-36].

[35] Aihua Zhang et al., "Urine Metabolomics," *Clinica Chimica Acta* 414 (December 24, 2012): 65–69, https://doi.org/10.1016/J.CCA.2012.08.016.

Existen diferentes modos de recolección para muestras de orina que incluyen muestreos puntuales, en intervalos preestablecidos o durante un período de 24 horas. La principal limitación de las muestras de orina radica en la existencia de grandes **diferencias en los volúmenes** excretados dependiendo del estado de hidratación de cada individuo así como del tiempo desde la última micción, además de otros factores fisiológicos, fisiopatológicos y/o nutricionales. Por consiguiente, la concentración de los metabolitos endógenos en orina pueden variar significativamente entre individuos. De hecho, se han encontrado diferencias de concentración de metabolitos en orina de hasta quince veces en estudios bajo condiciones fisiológicas normales. Este hecho disminuye la capacidad de identificar los compuestos del metaboloma que tienen relación con el objetivo del estudio, e incluso puede dar lugar a falsos descubrimientos que conduzcan a conclusiones erróneas (falsos positivos)[37]. Para solventar este problema, se han desarrollado **estrategias de normalización** que minimizan las diferencias en la dilución entre muestras de orina.

Las estrategias de normalización pueden ser aplicadas como paso previo o posterior a la etapa de adquisición de datos. Los **métodos de normalización previos a la adquisición** se basan en la dilución de las muestras en función de un parámetro intrínseco que sea proporcional a la concentración de la muestra (**factor de normalización**) de modo que todas alcancen el mismo valor de dicho parámetro. De manera alternativa, también se pueden variar los volúmenes de inyección de muestra

[36] Elizabeth J Want et al., "Global Metabolic Profiling Procedures for Urine Using UPLC–MS," *Nature Protocols* 5 (2010), https://doi.org/10.1038/nprot.2010.50.

[37] Bethanne M. Warrack et al., "Normalization Strategies for Metabonomic Analysis of Urine Samples," *Journal of Chromatography B* 877, no. 5–6 (February 15, 2009): 547–52, https://doi.org/10.1016/J.JCHROMB.2009.01.007.

en los análisis de acuerdo con la medida del factor de normalización, sin necesidad de diluir las muestras de manera previa y por tanto sin alterar la muestra original.

Por el contrario, los métodos de **normalización aplicados después de los análisis** se basan en el ajuste de las señales obtenidas en función de un factor de normalización. Por tanto, las muestras son analizadas reflejando su composición original, requiriendo un tratamiento de muestra más simple[23]. Sin embargo, dado que la respuesta en el espectrómetro de masas generalmente no sigue un patrón lineal para amplios rangos de concentración, muchos metabolitos no pueden ser normalizados correctamente mediante la utilización de este tipo de estrategias. Concretamente, las respuestas de los metabolitos no son proporcionales a diferentes niveles de concentración debido a diferencias en la eficiencia de ionización y/o en el grado de supresión iónica en el analizador de masas cuando se emplea ionización por electrospray (ESI) como fuente de ionización. Este hecho, además se puede ver agravado por la presencia de metabolitos en altas concentraciones que exceden el rango de respuesta lineal del detector y saturan su señal en el caso de las muestras más concentradas; o en el caso de las muestras más diluidas, metabolitos que se encuentran en bajas concentraciones y producen señales por debajo de los límites de detección. La mayor parte de estos inconvenientes pueden ser evitados empleando una estrategia de normalización previa a la adquisición de datos, ya que en ellas se ajusta la cantidad de muestra analizada a un mismo nivel de concentración, logrando reducir la contaminación de la fuente de ionización así como el efecto matriz, y por consiguiente, mejorando la eficiencia de ionización de los analitos y consiguiendo una respuesta del analizador mucho más precisa, estable y disminuyendo el riesgo de saturación de la señal. Sin embargo, este tipo de estrategias requieren un mayor tiempo de preparación de muestra, y aumentan la probabilidad de que metabolitos presentes en bajas concentraciones no sean detectados, al tener en muchas ocasiones que diluir la muestra.

A pesar de estas desventajas, varios estudios han comparado ambos tipos de estrategias, declinándose por los métodos de normalización previos a la adquisición de datos como mejor alternativa para la corrección de los diferentes grados de dilución presentes en muestras de orina[38-39].

Por otro lado, ambos tipos de estrategias se basan en diferentes factores de normalización, siendo los más empleados los mostrados de manera esquemática en la **Figura 6**, los cuales se detallan a continuación:

- **Volumen** total de orina. Este parámetro únicamente tiene utilidad cuando se dispone de muestras de orina de 24 horas.
- Concentración de **creatinina**. Este metabolito excretado en orina es un producto del metabolismo de descomposición del fosfato de creatina producido en tejidos musculares. El nivel de concentración de este compuesto se ha utilizado en cantidad de estudios para estandarizar las concentraciones de los metabolitos individuales dado que se ha considerado su nivel de excreción relativamente constante bajo condiciones fisiológicas normales. Sin

[38] William M. B. Edmands, Pietro Ferrari, and Augustin Scalbert, “Normalization to Specific Gravity Prior to Analysis Improves Information Recovery from High Resolution Mass Spectrometry Metabolomic Profiles of Human Urine,” *Analytical Chemistry* 86, no. 21 (November 4, 2014): 10925–31, https://doi.org/10.1021/ac503190m.

[39] Yanhua Chen et al., “Combination of Injection Volume Calibration by Creatinine and MS Signals’ Normalization to Overcome Urine Variability in LC-MS-Based Metabolomics Studies,” *Analytical Chemistry* 85, no. 16 (August 20, 2013): 7659–65, https://doi.org/10.1021/ac401400b.

embargo, esta metodología presenta limitaciones dado que se han observado diferencias de hasta cinco veces en la concentración de creatinina en relación a diferentes factores como la dieta, actividad física, género o estado de salud[40-41].

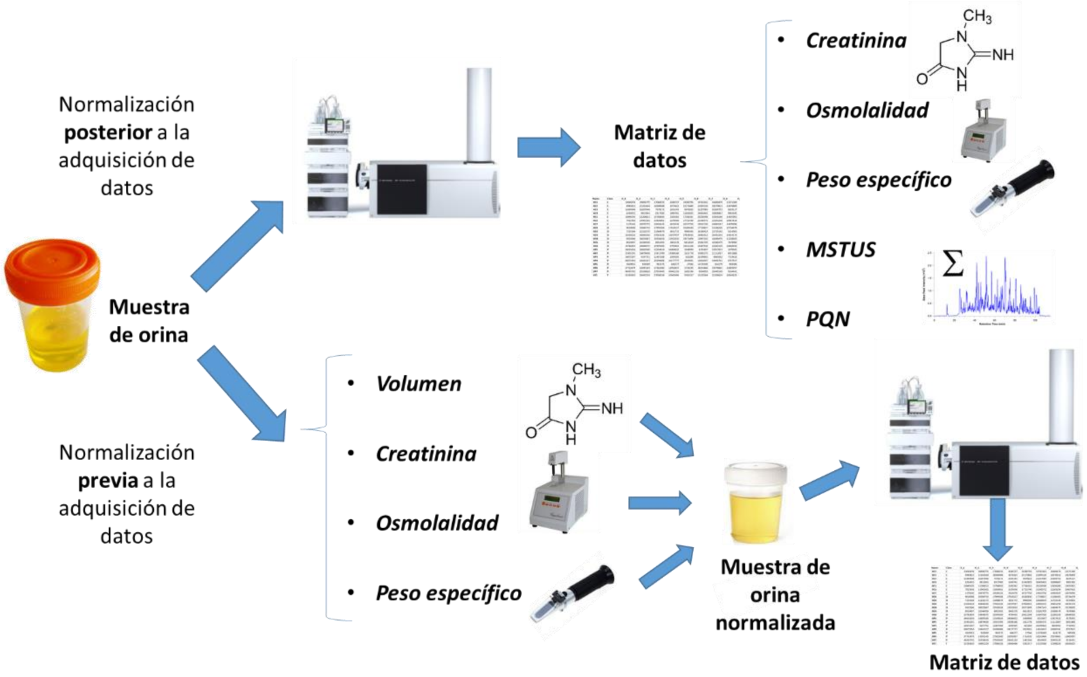


Figura 6. Métodos de normalización aplicados a muestras de orina en estudios de metabolómica no dirigidos.

- **Osmolalidad**. La medida de este parámetro hace referencia a la concentración total de solutos en la muestra de orina, la cual está altamente correlacionada con el grado de dilución de dicha muestra, siendo uno de los parámetros más eficaces en la normalización de esta muestra[27].

[40] David L. Heavner et al., "Effect of Creatinine and Specific Gravity Normalization Techniques on Xenobiotic Biomarkers in Smokers' Spot and 24-h Urines," *Journal of Pharmaceutical and Biomedical Analysis* 40, no. 4 (March 3, 2006): 928–42, https://doi.org/10.1016/J.JPBA.2005.08.008.

[41] Kai Wen Aaron Tang, Qi Chun Toh, and Boon Wee Teo, "Normalisation of Urinary Biomarkers to Creatinine for Clinical Practice and Research--When and Why.," *Singapore Medical Journal* 56, no. 1 (January 2015): 7–10, https://doi.org/10.11622/smedj.2015003.

- **Peso específico**. Este parámetro refleja la relación entre la densidad de la orina y la del agua pura a una temperatura constante, que puede ser medido directamente por gravimetría o indirectamente por refractometría. Este parámetro es un potente estimador de la osmolalidad, por ello su utilización está limitada, realizándose generalmente en caso de no disponer de la medida de osmolalidad[38].
- Suma de las señales comunes (*MS Total Useful Signal,* **MSTUS**). Esta metodología utiliza como factor de normalización el sumatorio de las señales que son comunes en todas las muestras, evitando de esta manera la interferencia de metabolitos exógenos, como es el caso de xenobióticos, que son únicamente detectados en un número reducido de muestras[37].
- Normalización por cociente probabilístico (*Probabilistic Quotient Normalization*, **PQN**)**.** Este método ha sido utilizado generalmente en datos de resonancia magnética nuclear, aunque recientemente se ha utilizado también en datos de espectrometría de masas. Esta metodología se basa en un cociente producto de la relación entre cada espectro y uno de referencia. Estos coeficientes se calculan utilizando la mediana de la relación de cada área de pico dividida por el área de pico correspondiente en la muestra de referencia. Finalmente, todas las áreas obtenidas para cada muestra son divididas por su factor PQN correspondiente[42].

[42] Yoric Gagnebin et al., "Metabolomic Analysis of Urine Samples by UHPLC-QTOF-MS: Impact of Normalization Strategies," *Analytica Chimica Acta* 955 (February 22, 2017): 27–35, https://doi.org/10.1016/J.ACA.2016.12.029.

A modo comparativo, la **Tabla 2** muestra las principales ventajas y desventajas de los distintos factores de normalización[22].

Tabla 2. Ventajas e inconvenientes de los factores de normalización.

Método de Normalización	Ventajas	Inconvenientes
Creatinina	Técnica estándar en laboratorios clínicos	Variabilidad debido a diferentes factores (dieta, género, estado de salud, etc.)
Peso específico	Método de laboratorio de fácil aplicación	El contenido de proteínas afecta a la medida
Volumen	Sencillez	Solo tiene utilidad en muestras de orina de 24 horas. Baja precisión
Osmolalidad	Técnica estándar en laboratorios clínicos. Medida bastante fiable de la concentración de orina	El procedimiento a menudo no está disponible y es sustituido por la medida del peso específico
MSTUS	No necesita medida de ningún parámetro analítico adicional. Evita interferencias de metabolitos exógenos presentes en un grupo reducido de muestras	Su aplicación de manera previa a la adquisición de datos requeriría analizar las muestras dos veces
PQN	Uso complementario a la aplicación de otras metodologías	Aplicación después de la adquisición de datos. No muy utilizado en estudios de MS

Varios estudios han comparado varias de estas metodologías mostrando la medida de osmolalidad, de manera previa a la adquisición, como el factor de normalización que mejor corrige la variabilidad debida a los distintos grados de dilución de las muestras de orina. No obstante, estudios recientes han revelado que se alcanzan mejores resultados cuando se utilizan varias de estas metodologías de forma complementaria. En este sentido, *Warrack et al.*[37] y *Chetwynd et al.*[27] recomiendan la utilización de los métodos basados en osmolalidad y MSTUS de manera complementaria. Por su parte, *Yoric Gagnebin et al.*[42] también llegaron a similares conclusiones dado que obtuvieron mejores resultados cuando combinaron una metodología previa (Osmolalidad) con una posterior (MSTUS o PQN) a la etapa de adquisición de datos.

3.2.3. Muestras de control de calidad

Las muestras de control de calidad (*Quality Control*, QC) se caracterizan por ser representativas de la composición cualitativa y cuantitativa de todas las muestras biológicas objeto de estudio. El propósito principal del empleo de este tipo de muestra es controlar y garantizar la **calidad de los resultados** obtenidos durante el flujo de trabajo de metabolómica. Generalmente, se pueden emplear dos tipos de muestras QC[43]:

- Muestras QC comerciales o sintéticas.
- Muestras QC generadas a partir de las muestras del propio estudio mediante la combinación de alícuotas iguales de cada muestra individual (**Figura 7**). A modo de ejemplo, una alícuota con un volumen comprendido entre 50 y 100 µl por muestra garantiza un volumen suficiente de muestra QC para estudios que analizan un número de muestras inferior a 500. Este tipo de QC presenta la gran ventaja de que representa en mayor medida las características y la composición de las muestras biológicas empleadas en el estudio[22].

[43] Warwick B Dunn et al., "Procedures for Large-Scale Metabolic Profiling of Serum and Plasma Using Gas Chromatography and Liquid Chromatography Coupled to Mass Spectrometry," *Nature Protocols* 6, no. 7 (2011): 1060–83, https://doi.org/10.1038/nprot.2011.335.

Una vez obtenidas las muestras QC, el tratamiento de dicha muestra deberá seguir la misma metodología definida para el conjunto de muestras del estudio. Posteriormente, las muestras QC deberán ser analizadas repetidamente a lo largo de la secuencia analítica para asegurar la reprodubilidad del proceso. A continuación, se describen brevemente los principales usos de este tipo de muestra[44]:

- Optimizar las diferentes etapas del flujo de trabajo (tratamiento de muestra, adquisición de datos, procesamiento de datos…).
- Determinar la precisión de la metodología.
- Estabilizar las condiciones y parámetros instrumentales.
- Controlar la reproducibilidad analítica.
- Normalizar posibles derivas analíticas.
- Ser una muestra representativa sujeta a análisis de masas en tándem (MS/MS) para su uso en la identificación de los metabolitos significativos.

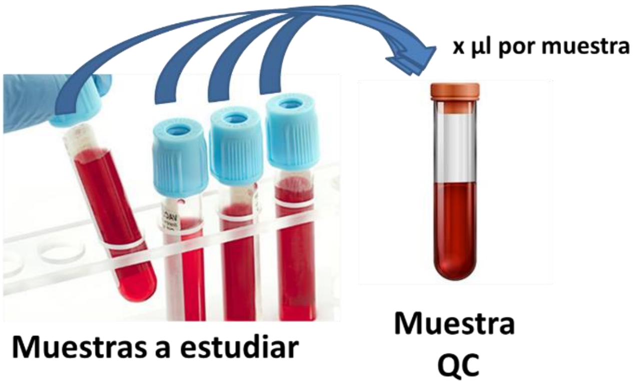


Figura 7. Ejemplo de preparación de una muestra QC.

[44] Warwick B Dunn et al., “The Importance of Experimental Design and QC Samples in Large-Scale and MS-Driven Untargeted Metabolomic Studies of Humans,” *Bioanalysis* 4, no. 18 (September 2012): 2249–64, https://doi.org/10.4155/bio.12.204.

3.3. Plataformas analíticas para la adquisición de datos

La **espectrometría de masas** (**MS**) y la **resonancia magnética nuclear** (**NMR**) son las técnicas generalmente empleadas en la totalidad de los estudios llevados a cabo mediante estrategias de metabolómica no dirigidas.

Las principales ventajas de NMR residen en su alta precisión en la cuantificación y elucidación estructural de analitos así como en su elevada reproducibilidad, mientras que la MS presenta una mayor sensibilidad y selectividad. Aunque existen aplicaciones que utilizan la MS mediante infusión directa[45], la gran potencialidad de esta técnica se alcanza cuando se utiliza acoplada a una **técnica separativa**, como la cromatografía de líquidos (LC), de gases (GC) o electroforesis capilar (CE), permitiendo la detección de cientos de metabolitos en una muestra biológica gracias a la disminución de la complejidad de la muestra así como de los efectos de supresión iónica. A modo comparativo, la **Tabla 3** muestra de manera más detallada las principales características tanto de la MS como de la RNM[46].

En los últimos años, la **espectrometría de movilidad iónica** (IMS) se está desarrollando como una alternativa con un gran potencial en el área de metabolómica, especialmente cuando se utiliza acoplada a un equipo de LC-MS (LC-IMS-MS). La ventaja de este acoplamiento reside en que la IMS aporta una tercera dimensión de

[45] Raúl González-Domínguez, Ana Sayago, and Ángeles Fernández-Recamales, "High-Throughput Direct Mass Spectrometry-Based Metabolomics to Characterize Metabolite Fingerprints Associated with Alzheimer's Disease Pathogenesis," *Metabolites* 8, no. 3 (2018): 1–9, https://doi.org/10.3390/metabo8030052.

[46] Abdul Hamid M. Emwas, "The Strengths and Weaknesses of NMR Spectroscopy and Mass Spectrometry with Particular Focus on Metabolomics Research," *Methods in Molecular Biology* 1277 (2015): 161–93, https://doi.org/10.1007/978-1-4939-2377-9_13.

separación proporcionando una mayor cobertura del metaboloma así como una mejor resolución estructural[47,48].

Tabla 3. Comparación de NMR y MS en el campo de la metabolómica.

	Resonancia Magnética Nuclear (NMR)	Espectrometría de Masas (MS)
Preparación de muestra	Mínima preparación de muestra	Tratamiento de muestra más complejo
Reproducibilidad	Muy alta	Limitada
Sensibilidad	Baja (LOD ≈ 10^{-4} M)	Alta (LOD ≈ 10^{-9} M)
Selectividad	En general se utiliza para análisis no selectivos	Puede ser utilizado para análisis selectivos y no selectivos (dirigidos y no dirigidos)
Número de metabolitos detectables	Inferior a 200, dependiendo de la resolución espectral	Capacidad de detectar más de 500 metabolitos en una muestra
Recuperación de la muestra	Técnica no destructiva	Técnica destructiva, pero la cantidad de muestra necesaria es pequeña
Acoplamiento a técnica separativa	Muy dificultoso	Fácilmente acoplable a LC, GC, CE. Aplicaciones recientes de IMS-MS
Señal obtenida	Los analitos que se encuentran por encima del límite de detección producen una señal medible	Requiere normalmente el acoplamiento a diferentes métodos separativos para poder detectar diferentes familias de metabolitos presentes en las muestras biológicas

[47] Valentina D'Atri et al., "Adding a New Separation Dimension to MS and LC–MS: What Is the Utility of Ion Mobility Spectrometry?," *Journal of Separation Science* (John Wiley & Sons, Ltd, January 1, 2018), https://doi.org/10.1002/jSSC.201700919.

[48] Allison J. Levy et al., "Recent Progress in Metabolomics Using Ion Mobility-Mass Spectrometry," *TrAC Trends in Analytical Chemistry* 116 (July 1, 2019): 274–81, https://doi.org/10.1016/J.TRAC.2019.05.001.

Dado que la LC acoplada a MS ha sido la plataforma analítica utilizada en el desarrollo de la presente tesis doctoral, a continuación se describen detalladamente ambas técnicas así como sus modos de aplicación en el campo de la metabolómica.

3.3.1. Cromatografía de líquidos.

La **cromatografía de líquidos** es una **técnica separativa** donde los componentes de la muestra se distribuyen entre una fase móvil líquida (disolvente o mezcla de disolventes) y una fase estacionaria (relleno de columna). La separación se consigue gracias a la diferente afinidad que presentan los distintos constituyentes de la muestra hacia cada una de estas fases. El gran poder de la LC reside en su **gran versatilidad** debido a las múltiples combinaciones de fases estacionarias, composición de fase móvil, posibilidad de usar elución en gradiente, su idoneidad para trabajar con compuestos no volátiles y térmicamente inestables, su fácil adaptación para permitir determinaciones cuantitativas exactas y su gran aplicabilidad, permitiendo determinar un gran número de especies presentes en muestras biológicas o de otros tipos.

Los componentes principales de un cromatógrafo de líquidos son el sistema de bombeo, el sistema de inyección, el compartimento de columna termostatizado y el detector, que en el caso de aplicaciones metabolómicas no dirigidas corresponde al espectrómetro de masas, considerándose una **técnica híbrida particular**[49] (**Figura 8**). El uso de LC-MS es considerada una de las herramientas analíticas más poderosas, la cual permite identificar los analitos que eluyen de la columna cromatográfica proporcionando una segunda dimensión de separación, ya que tras la separación de

[49] R. Cela, R. A. Lorenzo, and M. C. Casais, “Técnicas de Separación En Química Analítica,” *Intesis*, 2003.

los analitos en el cromatógrafo de líquidos, se produce una segunda separación de los mismos en el espectrómetro de masas en función de su relación masa/carga (m/z)[50,51].

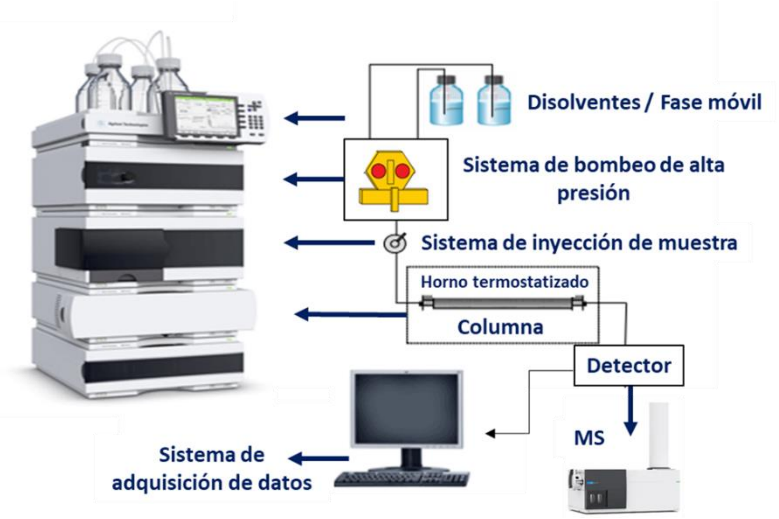

Disolventes / Fase móvil
Sistema de bombeo de alta presión
Sistema de inyección de muestra
Horno termostatizado
Columna
Detector
MS
Sistema de adquisición de datos

Figura 8. Componentes de un cromatógrafo de líquidos.

Básicamente en función del tipo de fase estacionaria utilizada, se distinguen los siguientes tipos de LC: la cromatografía de partición o reparto, la cromatografía iónica, la cromatografía de adsorción y la cromatografía de exclusión por tamaño[50]. De todos ellos, la modalidad más utilizada en el campo de la metabolómica es la **cromatografía de partición**, cuya fase estacionaria se caracteriza por ser un líquido retenido en un soporte sólido. Dependiendo de las polaridades tanto de la fase estacionaria como de la fase móvil, se distinguen las siguientes tres modalidades cromatográficas: cromatografía de líquidos en fase normal (NP-LC), en fase reversa (RP-LC) y de interacción hidrofílica (HILIC-LC).

[50] Douglas a. Skoog et al., “Fundamentos de Química Analítica,” *Fundamentos de Química Analítica*, 2005, https://doi.org/10.1016/S0584-8547.

[51] Mike S. Lee, *Mass Spectrometry Handbook*, ed. Mike S. Lee, Wiley, 2012, https://doi.org/10.1002/9781118180730.

En relación a las aplicaciones de estas modalidades cromatográficas en estudios de metabolómica, cabe destacar que no existe una única modalidad que permita separar y analizar el metaboloma completo de una muestra biológica, tal y como se muestra en la **Figura 9**. Por consiguiente, sería necesario una combinación de las diferentes modalidades de trabajo para obtener una visión lo más completa posible de todo el metaboloma. No obstante, en la mayoría de aplicaciones no es viable utilizar varios métodos cromatográficos debido al incremento de la duración de los análisis así como de los costes económicos derivados. En estos casos, se selecciona una modalidad en función de las familias de metabolitos que se desean estudiar o escogiendo la modalidad que permite una mayor cobertura del metaboloma[52-53].

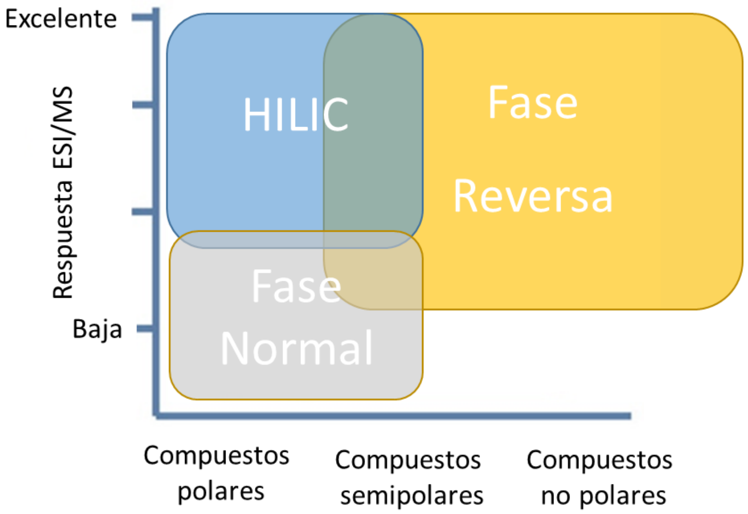


Figura 9. Aplicabilidad de las diferentes modalidades cromatográficas (NP-LC, RP-LC, HILIC-LC en función de la polaridad de los metabolitos a analizar.

[52] Julijana Ivanisevic et al., "Toward 'Omic Scale Metabolite Profiling: A Dual Separation–Mass Spectrometry Approach for Coverage of Lipid and Central Carbon Metabolism," *Analytical Chemistry* 85, no. 14 (July 16, 2013): 6876–84, https://doi.org/10.1021/ac401140h.

[53] Dao-Quan Tang et al., "HILIC-MS for Metabolomics: An Attractive and Complementary Approach to RPLC-MS," *Mass Spectrometry Reviews* 35, no. 5 (September 2016): 574–600, https://doi.org/10.1002/mas.21445.

En LC-ESI-MS, la **cromatografía de fase reversa,** caracterizada por emplear una fase estacionaria de carácter apolar mientras que la fase móvil es de naturaleza polar, es la modalidad más empleada debido a su gran versatilidad y estabilidad, permitiendo el análisis de gran parte del metaboloma. Concretamente, esta modalidad normalmente utiliza columnas con una fase estacionaria de C18 (n-octadecilo), pudiendo separar compuestos semipolares como vitaminas, alcaloides, esteroides glicososilados y otras especies glicosiladas. Por su parte, la cromatografía HILIC está ganando popularidad en los últimos años, la cual es capaz de separar los compuestos más polares como azúcares, aminoazúcares, aminoácidos, vitaminas, ácidos carboxílicos o nucleótidos, entre otros. HILIC se caracteriza por emplear fases estacionarias polares y fases móviles similares a las empleadas en el modo RP-LC. Además, una de las principales ventajas de esta modalidad es que permite el análisis de sustancias cargadas al igual que la cromatografía iónica[53,54]. Por último, la NP-LC, basada en emplear una fase estacionaria polar y una fase móvil apolar, presenta una mayor compatibilidad con otro sistema de ionización, en concreto con el sistema de ionización química a presión atmosférica (APCI-MS). Aunque su uso es menos frecuente, este tipo de modalidad se ha utilizado para el análisis de lípidos no polares como triacilgliceroles, esteroles o esteres de ácidos grasos[55].

[54] Bogusław Buszewski and Sylwia Noga, "Hydrophilic Interaction Liquid Chromatography (HILIC)--a Powerful Separation Technique.," *Analytical and Bioanalytical Chemistry* 402, no. 1 (January 2012): 231–47, https://doi.org/10.1007/s00216-011-5308-5.

[55] Bin Zhou et al., "LC-MS-Based Metabolomics," accessed July 26, 2019, https://doi.org/10.1039/c1mb05350g.

3.3.2. Espectrometría de masas

La **espectrometría de masas** (MS) se basa en la separación a vacío de iones en fase gaseosa de acuerdo a su **relación masa/carga** (m/z). Esta técnica presenta una alta aplicabilidad debido a su gran selectividad y a que es capaz de proporcionar información sobre la composición elemental, la estructura química de las moléculas, la composición cualitativa y cuantitativa de mezclas complejas así como las distribuciones isotópicas de los analitos detectados.

Dentro del espectrómetro de masas, las moléculas se ionizan mediante un **sistema de ionización**, dando lugar a un grupo de iones que se separan en base a su relación m/z y detectan mediante un **analizador de masas** y un **detector de iones**, respectivamente. La señal generada es el correspondiente **espectro de masas** (representación de abundancia relativa de cada ion frente a su m/z), el cual se puede utilizar como la “huella digital” de una sustancia.

Se han desarrollado diferentes tipos de sistemas de ionización, como la ionización por electrospray (ESI), ionización química a presión atmosférica (APCI), fotoionización a presión atmosférica (APPI), plasma acoplado inductivamente (ICP) o desorción/ionización láser asistida por una matriz (MALDI). Entre los diferentes modos, la **ionización por electrospray** (**ESI**) es la técnica de ionización más utilizada para el análisis de metabolitos mediante LC-MS, dado que es una fuente de ionización suave, que produce una mínima fragmentación del analito, detectándose, por tanto, los iones moleculares intactos, lo que es de gran ayuda en la tarea de identificación de compuestos. En este tipo de ionización, la muestra que eluye del cromatógrafo de líquidos pasa a través de una aguja capilar de acero inoxidable de pequeño diámetro cuyo extremo se encuentra sometido a un potencial eléctrico elevado (del orden de varios kV). Con ayuda de un gas nebulizador se produce la formación de una niebla de finas gotas cargadas eléctricamente, que se en encuentran en una cámara de desolvatación, donde se produce la evaporación del disolvente. De este modo, las gotas formadas cada vez tienen un tamaño menor, lo que produce un aumento en su densidad de carga, hasta un momento en el que las fuerzas de repulsión de los iones de signo contrario superan la tensión superficial que mantiene unidas las microgotas, alcanzando el denominado "límite de Rayleigh", momento en el que las gotas se vuelven inestables y comienzan a sufrir el proceso conocido como "explosiones de Coulomb". Este proceso se repite de manera sucesiva hasta que finalmente se forman iones cargados en fase gaseosa con una o más cargas, los cuales son atraídos hacia la entrada del espectrómetro de masas gracias al voltaje aplicado a la entrada del capilar[56]. Esta etapa de ionización se puede llevar a cabo en **modo positivo** o **negativo**, formándose especies iónicas de carga positiva ($[M+nH]^{n+}$, $[M+Na]^{+}$, $[M+Li]^{+}$,…) o negativa ($[M-nH]^{n-}$, $[M+Cl]^{-}$,…), respectivamente. Debido a la diversidad de propiedades químicas de los metabolitos, hay moléculas que únicamente son ionizadas en uno de los dos modos, haciendo necesario en ocasiones el análisis en ambos modos de ionización para una mayor cobertura del metaboloma. Para evitar este inconveniente, muchos instrumentos actuales son capaces de cambiar el modo de

[56] Hanan Awad, Mona M. Khamis, and Anas El-Aneed, "Mass Spectrometry, Review of the Basics: Ionization," *Applied Spectroscopy Reviews* 50, no. 2 (February 7, 2015): 158–75, https://doi.org/10.1080/05704928.2014.954046.

ionización de manera intermitente durante los análisis registrando ambos tipos de datos en el mismo análisis[55,57-58].

Las fuentes de ionización APCI y APPI también son modos de ionización suaves, que se utilizan en metabolómica de manera complementaria a ESI para el análisis de metabolitos no polares y compuestos térmicamente estables, como el caso de los lípidos. La **Figura 10** muestra los rangos de aplicabilidad de los distintos modos de ionización. Actualmente, existen instrumentos con dobles fuentes de ionización (ESI y APCI, o ESI y APPI) con el objetivo de incrementar la cobertura de detección del metaboloma[55].

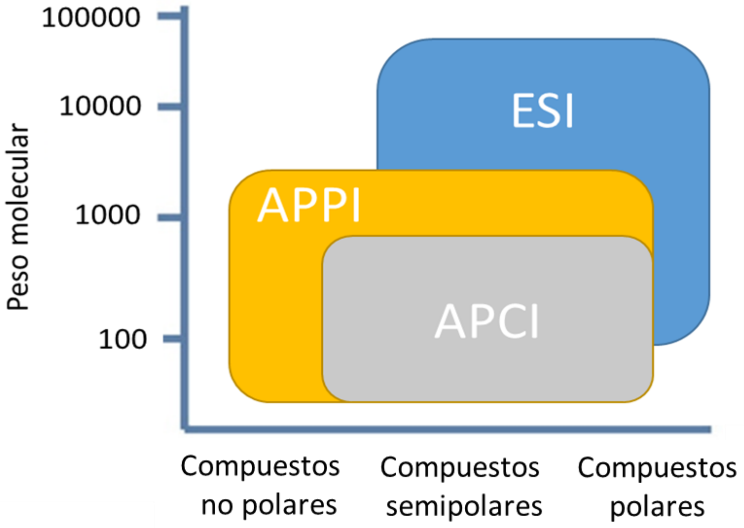


Figura 10. Rangos de aplicabilidad de las fuentes de ionización ESI, APPI, APCI.

[57] G A Nagana Gowda and Danijel Djukovic, "Overview of Mass Spectrometry-Based Metabolomics: Opportunities and Challenges," *Methods Mol Biol* 1198 (2014): 3–12, https://doi.org/10.1007/978-1-4939-1258-2_1.

[58] Hajime Mizuno et al., "The Great Importance of Normalization of LC-MS Data for Highly-Accurate Non-Targeted Metabolomics," *Biomedical Chromatography* 31, no. 1 (2017): 1–7, https://doi.org/10.1002/bmc.3864.

Existen diversos tipos de **analizadores de masas**: cuadrupolo (Q), trampa de iones, orbitrap, sector magnético o tiempo de vuelo (TOF). Entre ellos, los analizadores de MS que presentan una mayor aplicabilidad en estudios de metabolómica no dirigida son los de **alta resolución**, como por ejemplo el analizador de tiempo de vuelo u orbitrap. La alta resolución permite que estos analizadores proporcionen una cuantificación relativa muy precisa así como valores de masa exacta con **cuatro cifras decimales** con una **exactitud inferior a 5 ppm**. Además de la gran precisión en la medida, estos analizadores también proporcionan información sobre las **distribuciones isotópicas** de las especies químicas, lo que es de enorme utilidad junto a los datos de masa exacta para la generación de fórmulas moleculares que ayuden a la identificación de los metabolitos. En cuanto al modo de adquisición para aplicaciones metabolómicas, los datos de MS son generalmente obtenidos en modo *full scan* para su posterior empleo en análisis estadísticos. Concretamente, el **analizador de tiempo de vuelo** (**TOF**), utilizado en los estudios presentados en esta memoria, es uno de los más usados ya que presenta mayor número de ventajas, como su elevada velocidad de adquisición, amplio rango de masas, alta selectividad, simplicidad, robustez y elevada resolución. La separación de los iones en este analizador se consigue gracias a la distinta velocidad que adquieren en el interior del tubo de vuelo en función de su m/z. Los iones al entrar al analizador son acelerados con la misma energía cinética mediante la aplicación de un campo eléctrico pulsante de 10^3 a 10^4 V, de forma que se produce su separación en función de sus diferentes m/z, ya que los iones de mayor m/z “volarán” a menor velocidad que los de menor m/z a lo largo del tubo de vuelo, consiguiéndose su separación[50,59].

Por otro lado, la **espectrometría de masas en tándem** (**MS/MS**) se basa en el empleo simultáneo de dos analizadores, como por ejemplo cuadrupolo-tiempo de vuelo, trampa de iones-tiempo de vuelo o triple cuadrupolo, proporcionando **patrones de fragmentación** de los iones precursores, obteniéndose por tanto un tercer nivel de información, que junto con los datos de tiempos de retención, exactitud de masa y distribuciones isotópicas, son de gran aplicación para la identificación de los metabolitos y la resolución de isómeros[60,61]. La **Figura 11** muestra un ejemplo del diseño de un **espectrómetro de masas cuadrupolo-tiempo de vuelo** (**QTOF**), dado que es el utilizado en el desarrollo experimental de esta memoria.

En los últimos años, se están llevando a cabo aplicaciones de metabolómica utilizando el modo de fragmentación MS^E, el cual se basa en la continua fragmentación de todos los iones precursores mediante la combinación de altas y bajas energías de colisión durante los análisis de las muestras[62]. No obstante, los análisis mediante MS/MS se utilizan generalmente de manera complementaria con el objetivo de conseguir la identificación de las señales que resulten significativas en los análisis estadísticos

[59] Edmond De Hoffmann and Vincent Stroobant, *Mass Spectrometry - Priniples and Applications.*, *Mass Spectrometry Reviews*, 2007, https://doi.org/10.1002/mas.20296.

[60] Sebastian Broecker et al., "Combined Use of Liquid Chromatography–Hybrid Quadrupole Time-of-Flight Mass Spectrometry (LC–QTOF-MS) and High Performance Liquid Chromatography with Photodiode Array Detector (HPLC–DAD) in Systematic Toxicological Analysis," *Forensic Science International* 212, no. 1–3 (2011): 215–26, https://doi.org/10.1016/j.forsciint.2011.06.014.

[61] Aline Soriano Lopes et al., "Metabolomic Strategies Involving Mass Spectrometry Combined with Liquid and Gas Chromatography," in *Advances in Experimental Medicine and Biology*, vol. 965, 2017, 77–98, https://doi.org/10.1007/978-3-319-47656-8_4.

[62] Huan Wu et al., "Untargeted Metabolomics Profiles Delineate Metabolic Alterations in Mouse Plasma during Lung Carcinoma Development Using UPLC-QTOF/MS in MSE Mode.," *Royal Society Open Science* 5, no. 9 (September 2018): 181143, https://doi.org/10.1098/rsos.181143.

realizados en los análisis de MS1. Para ello, un grupo reducido de muestras es sometido a este tipo de análisis de MS/MS, donde se seleccionan los iones precursores de interés para obtener sus patrones de fragmentación, lo que facilita la identificación de los metabolitos correspondientes a dichas señales.

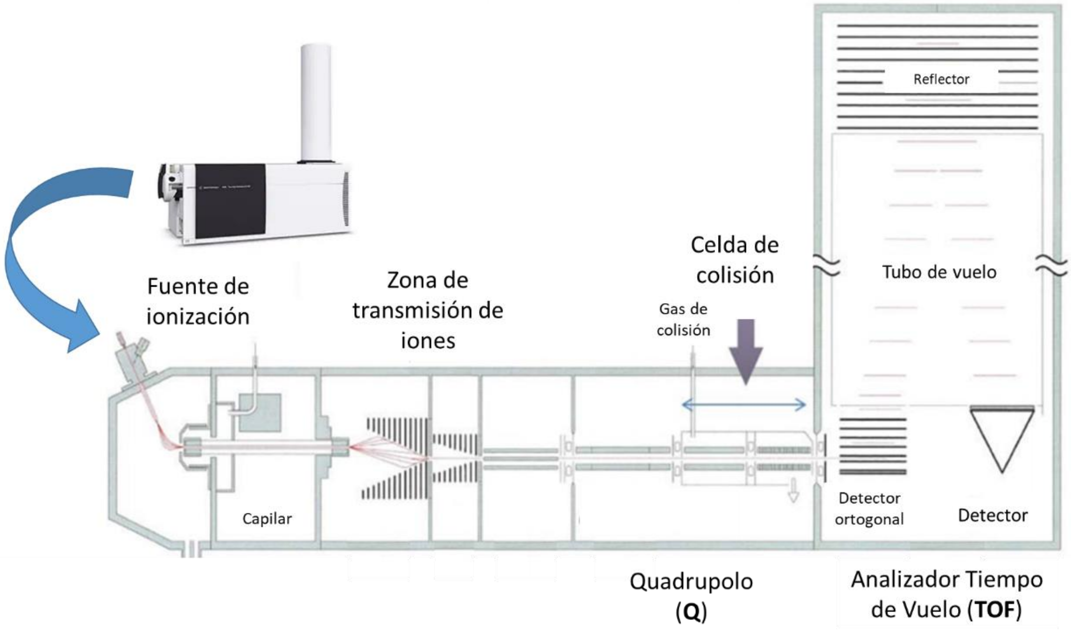


Figura 11. Esquema de las partes principales de las que consta un espectrómetro de masas QTOF.

Actualmente, la última generación de analizadores TOF y QTOF, emplean un software que proporcionan una lista de posibles fórmulas moleculares, utilizando los datos de masa exacta y los valores de distribución isotópica. Este tipo de software usa un algoritmo de ajuste matemático que tiene en cuenta la configuración electrónica, cálculo de dobles enlaces en anillos, además de una sofisticada comparación entre la distribución isotópica teórica y experimental para incrementar la confianza de la fórmula molecular sugerida, eliminando de esta manera más de un 95 % de falsos candidatos[63-64].

3.3.3. Control de la variabilidad analítica

En estudios de metabolómica no dirigida con un alto número de muestras, la etapa de adquisición de datos mediante LC-MS puede durar varias semanas o incluso meses. Por ello, las muestras se distribuyen para su análisis en diferentes secuencias analíticas (***batches***). La larga duración de los análisis puede ocasionar variabilidad en los datos de manera adicional a la propia variabilidad biológica, debido a diferentes factores instrumentales como la eficiencia de ionización por electrospray, composición de las fases móviles, temperatura o rendimiento de la columna cromatográfica[65]. Para controlar esta variabilidad instrumental, es necesario el análisis de una muestra de control de calidad (QC) a lo largo de la secuencia analítica. Además del control de la reproducibilidad analítica, esta muestra es inyectada repetidamente, entre cuatro y seis veces, al inicio de la secuencia analítica con el objetivo de estabilizar las condiciones cromatográficas (acondicionar el sistema cromatográfico para estabilizar los tiempos de retención) y de espectrometría de masas (exactitud en la determinación de m/z e intensidad de la señal). Por otra parte, con fines de identificación de metabolitos, la muestra QC es analizada por MS/MS al ser una muestra representativa

[63] D Arráez-Román et al., "Identification of Phenolic Compounds from Pollen Extracts Using Capillary Electrophoresis-Electrospray Time-of-Flight Mass Spectrometry.," *Analytical and Bioanalytical Chemistry* 389, no. 6 (2007): 1909–17, https://doi.org/10.1007/s00216-007-1611-6.

[64] Tobias Kind and Oliver Fiehn, "Metabolomic Database Annotations via Query of Elemental Compositions: Mass Accuracy Is Insufficient Even at Less than 1 Ppm.," *BMC Bioinformatics* 7 (April 28, 2006): 234, https://doi.org/10.1186/1471-2105-7-234.

[65] J. Kuligowski et al., "Detection of Batch Effects in Liquid Chromatography-Mass Spectrometry Metabolomic Data Using Guided Principal Component Analysis," *Talanta* 130 (December 1, 2014): 442–48, https://doi.org/10.1016/J.TALANTA.2014.07.031.

del conjunto de muestras del estudio, tal y como se describió en la *sección 3.2.3*. En la **Figura 12** se muestra un esquema general de distribución de las muestras en las diferentes secuencias analíticas en un estudio de metabolómica no dirigida. Adicionalmente, se recomienda la inyección al principio y final de las secuencias de blancos analíticos con el objetivo de detectar cualquier artefacto o contaminante procedente de fuentes externas a las muestras biológicas, como por ejemplo de los disolventes de extracción, fases móviles, etc.[66].

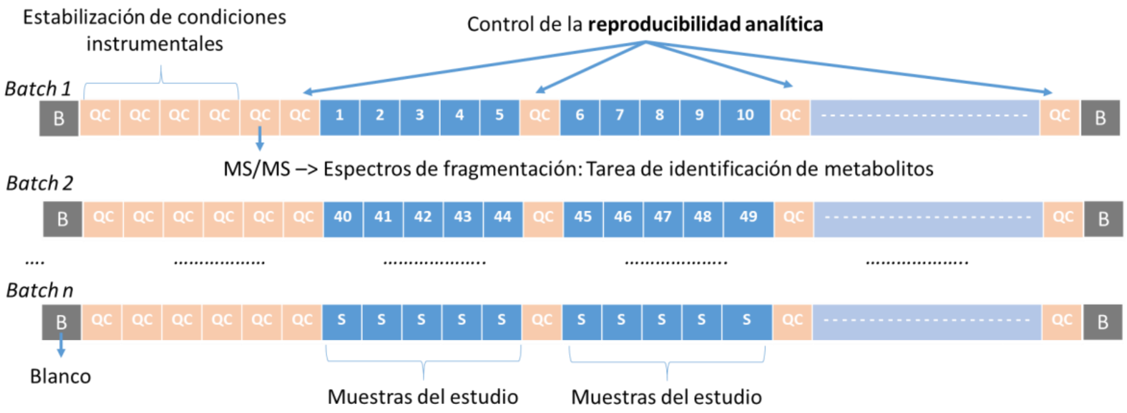

Estabilización de condiciones instrumentales
Control de la reproducibilidad analítica
Batch 1
B QC QC QC QC QC QC 1 2 3 4 5 QC 6 7 8 9 10 QC ---------------------- QC B
MS/MS –> Espectros de fragmentación: Tarea de identificación de metabolitos
Batch 2
B QC QC QC QC QC QC 40 41 42 43 44 QC 45 46 47 48 49 QC ---------------------- QC B
…. ……………… ……………….. ……………….. ………………..
Batch n
B QC QC QC QC QC QC S S S S S QC S S S S S QC ---------------------- QC B
Blanco
Muestras del estudio
Muestras del estudio

Figura 12. Distribución de las muestras experimentales (color azul), QC (color salmón) y blancos (B, color gris) en las secuencias analíticas en estudios de metabolómica no dirigidos.

Otro aspecto a tener en cuenta a la hora de diseñar las distintas secuencias analíticas es el orden aleatorio de inyección de las muestras del estudio a lo largo de las mismas, con la finalidad de que no se produzcan resultados que den lugar a interpretaciones erróneas (falsos positivos). Esta **aleatorización** asegurará que la variabilidad analítica

[66] Mónica Calderón-Santiago et al., “MetaboQC: A Tool for Correcting Untargeted Metabolomics Data with Mass Spectrometry Detection Using Quality Controls,” *Talanta* 174 (November 2017): 29–37, https://doi.org/10.1016/j.talanta.2017.05.076.

afectará en la misma medida a todos los grupos experimentales[36,67]. Como los estudios de metabolómica no dirigida no requieren la cuantificación de los metabolitos, las muestras biológicas del estudio se analizan generalmente únicamente una vez sin la existencia de réplicas analíticas. El motivo de la ausencia de replicados es que en estudios con una gran cantidad de muestras no es viable tanto por tiempo y como por coste económico, además de que desde el punto de vista estadístico los réplicas analíticas no pueden ser tratadas como muestras independientes, y por tanto deben combinarse antes del análisis de datos. En consecuencia, se recomienda aumentar el número de réplicas biológicas obtenidas de diferentes individuos dado que la variabilidad biológica siempre es mayor que la variabilidad analítica[68].

3.4. Pre-procesamiento de datos

El pre-procesamiento de datos es una de las etapas más importantes y críticas dentro del flujo de trabajo de los estudios de metabolómica no dirigida. Los objetivos de esta etapa son: reducir la complejidad de los datos, extraer las señales de interés (*molecular features*) así como su tratamiento, y finalmente obtener una matriz de datos adecuada para su posterior empleo en los análisis estadísticos. Todas las etapas del pre-procesamiento deben ser realizadas mediantes procedimientos claros y

[67] Maya Berg et al., "LC-MS Metabolomics from Study Design to Data-Analysis - Using a Versatile Pathogen as a Test Case.," *Computational and Structural Biotechnology Journal* 4 (2013): e201301002, https://doi.org/10.5936/csbj.201301002.

[68] Lloyd W. Sumner et al., "Proposed Minimum Reporting Standards for Chemical Analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI)," *Metabolomics* 3(3) (2007): 211–21, https://doi.org/10.1007/s11306-007-0082-2.

reproducibles para garantizar la calidad de la matriz de datos y reducir la probabilidad de obtener falsos positivos y/o falsos negativos en los resultados estadísticos[24,69].

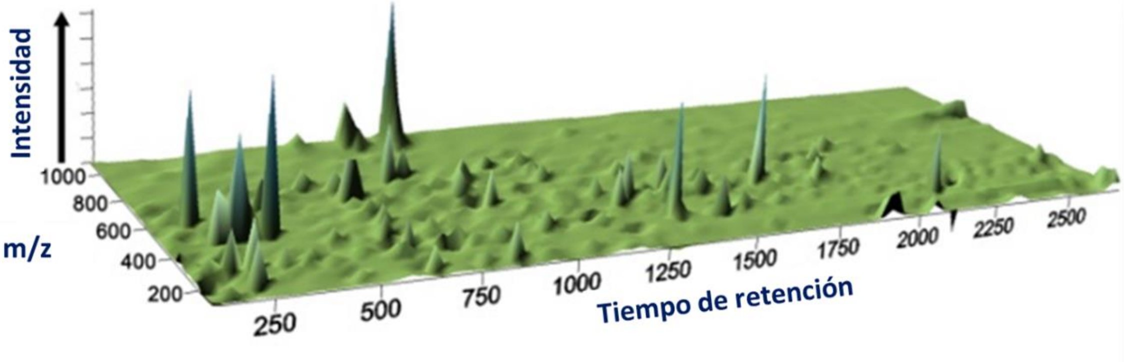


Figura 13. Esquema del tipo de datos (RT, m/z, I) adquiridos por LC-MS.

Las señales adquiridas por LC-MS contienen una gran cantidad de datos reflejados en tres dimensiones: **tiempo de retención** (RT), **relación masa/carga** (m/z) e **intensidad de señal** (I) (**Figura 13**). La gran complejidad de estos datos (RT, m/z, I) se debe a la variabilidad analítica responsable de las fluctuaciones en los RT y en las relaciones m/z entre muestras, así como las derivas de intensidad a lo largo o entre las distintas secuencias analíticas, etc. Por otra parte, un mismo compuesto puede tener asociadas varias señales con diferentes m/z debido a isótopos, aductos o fragmentos. Todo ello, sumado a la variabilidad biológica de las muestras, hace que la etapa de pre-

[69] Ibrahim Karaman, Rui Climaco Pinto, and Gonçalo Graça, "Metabolomics Data Preprocessing: From Raw Data to Features for Statistical Analysis," *Comprehensive Analytical Chemistry* 82 (January 1, 2018): 197–225, https://doi.org/10.1016/BS.COAC.2018.08.003.

procesamiento sea crucial para generar una matriz de datos de calidad para su análisis estadístico[70].

El pre-procesamiento de los datos adquiridos por LC-MS conlleva las siguientes etapas: corrección de la línea base, eliminación del ruido, detección y deconvolución de las señales (*peak picking*), alineamiento de los tiempos de retención y m/z, normalización de la intensidad de señal, estimación de los valores faltantes (*missing values*) o agrupación de las señales correspondientes a un mismo compuesto (*annotation*)[71] (**Figura 14**). La secuenciación de las distintas etapas de pre-procesamiento de datos puede seguir un orden diferente dependiendo de los algoritmos o software empleados. En las siguientes secciones, se describen las características principales de cada una de estas etapas.

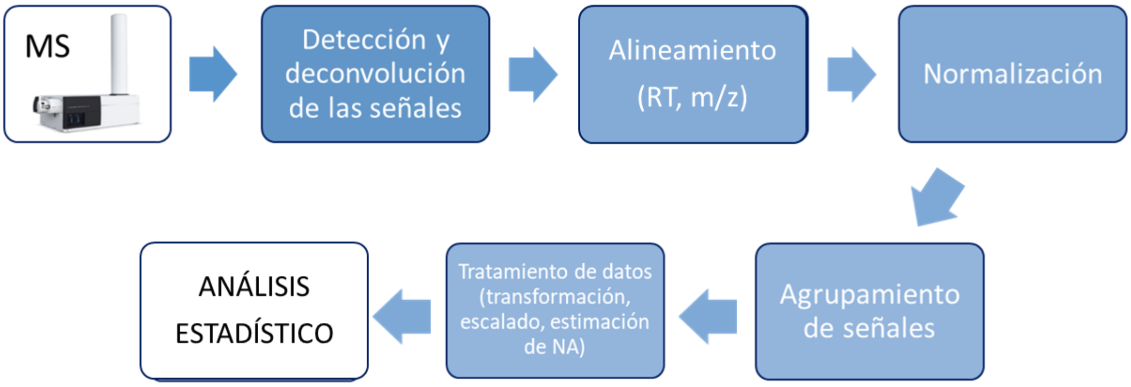

MS
Detección y deconvolución de las señales
Alineamiento (RT, m/z)
Normalización
Agrupamiento de señales
Tratamiento de datos (transformación, escalado, estimación de NA)
ANÁLISIS ESTADÍSTICO

Figura 14. Principales etapas del pre-procesamiento de datos adquiridos por LC-MS.

[70] Carl Brunius, Lin Shi, and Rikard Landberg, “Large-Scale Untargeted LC-MS Metabolomics Data Correction Using between-Batch Feature Alignment and Cluster-Based within-Batch Signal Intensity Drift Correction,” *Metabolomics* 12, no. 11 (2016): 1–13, https://doi.org/10.1007/s11306-016-1124-4.

[71] Mikko Katajamaa and Matej Orešič, “Data Processing for Mass Spectrometry-Based Metabolomics,” *Journal of Chromatography A* (Elsevier, July 27, 2007), https://doi.org/10.1016/j.chroma.2007.04.021.

3.4.1. Detección de señales (*Peak Picking*)

El objetivo de esta etapa es la extracción de las señales detectadas en cada una de las muestras asignándoles su RT, m/z e intensidad o área de pico. Para ello, se utilizan algoritmos que sean capaces de detectar las diferentes señales en cada cromatograma e integrar sus áreas de pico para proporcionar un resultado semi-cuantitativo de las mismas.

Los métodos más utilizados exploran cada espectro de manera individual a través de la combinación de dos algoritmos. El primero de ellos utiliza algoritmos de **suavizado** (*smoothing*) con el objetivo de corregir la línea base reduciendo el nivel de ruido instrumental y por tanto, mejorando las relaciones señal/ruido (S/N)[69] (**Figura 15**).

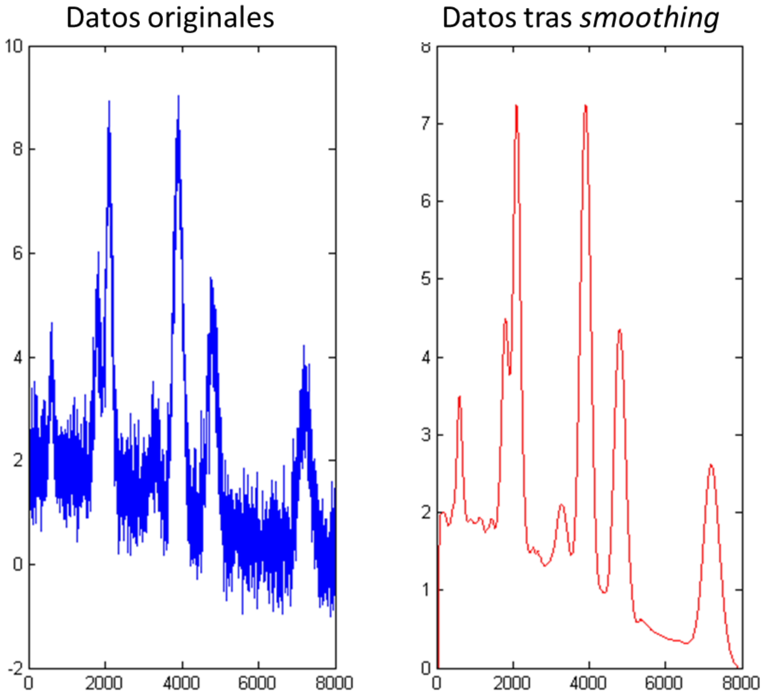


Figura 15. Ejemplo de aplicación de algoritmo de suavizado (smoothing).

En la segunda etapa, los diferentes picos son seleccionados utilizando una acotación para las condiciones de búsqueda en esta etapa. Estas condiciones acotadas, también denominadas filtros, son aplicadas a diferentes parámetros, como por ejemplo la relación S/N, la intensidad o el área integrada de cada pico. Por lo tanto, las señales que se encuentren por debajo de los límites fijados para estos parámetros no son extraídas, evitando así la obtención de señales que no contienen información biológica relevante. Además, se puede establecer que las señales extraídas se encuentren al menos en un porcentaje determinado del número total de muestras estudiadas, y de esta forma descartar los metabolitos que están presentes únicamente en un bajo número de muestras, los cuales estarán generalmente relacionados con xenobióticos[21,72].

3.4.2. Alineamiento

Los valores de RT y m/z pueden sufrir fluctuaciones dependiendo de varios factores como la temperatura, la abundancia de los iones que son detectados de manera simultánea, la velocidad de adquisición del detector, entre otros. Afortunadamente, las mejoras tecnológicas alcanzadas en los equipos de HRMS actuales permiten la calibración de los datos adquiridos garantizado una buena reproducibilidad en la determinación de los valores de m/z. Este hecho se consigue gracias a la inyección, de manera continua o puntual, de una disolución de calibración externa durante el proceso de adquisición de datos por HRMS[73].

[72] Colin A. Smith et al., "XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification," *Analytical Chemistry* 78, no. 3 (2006): 779–87, https://doi.org/10.1021/ac051437y.

[73] Masahiro Sugimoto et al., "Bioinformatics Tools for Mass Spectroscopy-Based Metabolomic Data Processing and Analysis," *Current Bioinformatics*, vol. 7, 2012, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3299976/pdf/CBIO-7-96.pdf.

Respecto a los RT, se producen frecuentemente ligeras variaciones a lo largo de las secuencias analíticas debido a diferentes factores, como por ejemplo el deterioro de la columna cromatográfica, variaciones en la temperatura y/o presión del sistema, modificaciones en la composición de las fases móviles, etc. Este problema tiene especial relevancia en estudios de metabolómica no dirigida dado que un mismo metabolito puede eluir a tiempos de retención ligeramente diferentes entre las distintas muestras.

Por todo ello, se deben emplear algoritmos de alineamiento para corregir estas diferencias tanto en RT como en la relación m/z, y en consecuencia, poder agrupar las señales encontradas en las diferentes muestras que correspondan a la señal generada por un mismo analito[74].

Estos métodos de alineamiento se pueden dividir en las siguientes dos categorías, aunque existen aproximaciones que utilizan una combinación de ambas[71]:

1) Métodos que alinean los RT y m/z tras la realización del *Peak Picking*. Estos métodos se basan en la búsqueda de las mismas características entre todas las muestras en un rango establecido de RT y de m/z.
2) Métodos que alinean los RT y m/z de los datos adquiridos de manera previa a la detección y deconvolución de las señales. Estos métodos se basan en la transformación del eje de RT de cada muestra a un eje común para todas las muestras. Una vez se realiza esta transformación, se procede a la etapa del *Peak Picking*.

[74] Ibrahim Karaman, "Preprocessing and Pretreatment of Metabolomics Data for Statistical Analysis," in *Advances in Experimental Medicine and Biology*, vol. 965, 2017, 145–61, https://doi.org/10.1007/978-3-319-47656-8_6.

3.4.3. Normalización

La etapa de normalización persigue la eliminación de la variabilidad presente en los datos debida a factores externos a la propia variabilidad biológica de las muestras. En estudios de metabolómica no dirigida las principales fuentes de variabilidad que interfieren en la calidad de los datos proceden de las etapas de tratamiento de muestra y adquisición de datos. En estudios que analizan un gran número de muestras, la variabilidad instrumental produce habitualmente efectos sistemáticos en los datos debido a derivas en las intensidades producidas a lo largo (***within-batch effect***) y entre las diferentes secuencias analíticas (***between-batch effect***). Estos efectos son fácilmente observables mediante la distribución de las muestras QC en un Análisis de Componentes Principales (PCA). Por ello, la corrección de estas derivas es uno de los principales objetivos de la etapa de normalización[22].

Para la corrección de dichas fluctuaciones, se han desarrollado una gran cantidad de estrategias de normalización. Una de las aproximaciones se basa en la utilización de patrones internos (IS), pero su empleo no está recomendado en estudios de metabolómica no dirigida ya que en ellos se detectan gran multitud de analitos y las fluctuaciones pueden depender de las características químicas de cada compuesto. Por tanto, un número limitado de IS no representa la naturaleza química de todos los metabolitos detectados, pudiendo producirse errores en la normalización al considerar señales de IS que presentan una respuesta de detección diferente a los analitos[75]. Otras metodologías de normalización se basan en el uso de la intensidad de las señales

[75] Xiaotao Shen et al., "Normalization and Integration of Large-Scale Metabolomics Data Using Support Vector Regression," *Metabolomics* 12, no. 5 (May 26, 2016): 89, https://doi.org/10.1007/s11306-016-1026-5.

más estables, la suma total de las intensidades (MSTS) o de las intensidades comunes (MSTUS). Estos métodos son cuestionables dado que consideran que todas las señales presentan un comportamiento similar[76]. Por otra parte, otros métodos se basan en hacer comparables las distribuciones estadísticas de las intensidades entre las distintas muestras, como por ejemplo la normalización por la mediana, cuantiles o cociente probabilístico (PQN). Sin embargo, estos métodos suponen que todos los metabolitos experimentan el mismo patrón de deriva en el transcurso de los análisis, lo cual no es siempre acertado[77].

Entre las diferentes aplicaciones de las muestras QC descritas en la *sección 3.2.3*, destaca su potencialidad para monitorizar y normalizar este tipo de efectos. De hecho, se han desarrollado estrategias basadas en la normalización mediante QC que permiten emplear distintos algoritmos para normalizar individualmente las señales de los metabolitos según su tendencia de deriva. Este tipo de enfoque se está convirtiendo en la estrategia más utilizada en los estudios de metabolómica llevados a cabo en los últimos años[70,78].

Una vez realizada la etapa de normalización, las señales que continúan presentando una alta variabilidad deben ser descartadas. Para ello se aplica generalmente un filtro basado en la desviación estándar relativa (RSD) de las variables en las muestras QC,

[76] Muhammad Anas Kamleh et al., “Optimizing the Use of Quality Control Samples for Signal Drift Correction in Large-Scale Urine Metabolic Profiling Studies,” *Analytical Chemistry* 84, no. 6 (2012): 2670–77, https://doi.org/10.1021/ac202733q.

[77] Joomi Lee et al., “Quantile Normalization Approach for Liquid Chromatography-Mass Spectrometry-Based Metabolomic Data from Healthy Human Volunteers.,” *Analytical Sciences : The International Journal of the Japan Society for Analytical Chemistry* 28, no. 8 (2012): 801–5, http://www.ncbi.nlm.nih.gov/pubmed/22878636.

[78] Chanisa Thonusin et al., “Evaluation of Intensity Drift Correction Strategies Using MetaboDrift, a Normalization Tool for Multi-Batch Metabolomics Data,” *Journal of Chromatography. A* 1523 (2017): 265, https://doi.org/10.1016/J.CHROMA.2017.09.023.

utilizando un límite comprendido en un rango entre el 20 y 30 %, dependiendo del estudio[79].

Como se ha comentado en la *sección 3.2.2*, existen otros tipos de métodos de normalización que tienen como objetivo corregir la **variabilidad biológica** de las muestras de orina relacionadas con los diferentes grados de dilución. En el caso de no utilizar un método de normalización de manera previa a la adquisición de datos, esta etapa presentaría una mayor dificultad ya que los datos registrados reflejarían la influencia de ambos factores de variabilidad, que deberían ser normalizados.

3.4.4. Agrupación de señales (*annotation*)

Un metabolito puede producir varias especies iónicas en la fase de ionización del MS, cuyo número se ve incrementado en caso de utilizar una fuente de ionización "dura", como por ejemplo la ionización electrónica (EI), ampliamente utilizada en GC-MS. En el caso de ESI, a pesar de ser una fuente de ionización suave, también es frecuente la detección de diferentes especies, que incluyen isótopos, fragmentos, multímeros y/o aductos. De estos últimos, los aductos más frecuentes suelen ser: $[M+H]^+$, $[M+Na]^+$, $[M+H-H_2O]^+$, $[M+K]^+$, $[M+NH_4]^+$, en modo de ionización positivo; y $[M-H]^-$, $[M-H-H_2O]^-$, $[M+Cl]^-$, $[M+HCOO]^-$, $[2M-H]^-$, en modo negativo.

Debido al incremento del número de señales detectadas en relación al número de analitos, esta etapa persigue la agrupación de todos las especies iónicas derivadas de un mismo analito en una única especie molecular (*Molecular Feature*), reduciendo de esta manera el número de señales que contienen información redundante de cara a

[79] Courtney Schiffman et al., "Filtering Procedures for Untargeted LC-MS Metabolomics Data," *BMC Bioinformatics* 20 (2019), https://doi.org/10.1186/S12859-019-2871-9.

los análisis estadísticos. Además, otro objetivo de esta etapa es la asignación de la masa molecular neutra o monoisotópica a las distintas especies moleculares detectadas, que es fundamental para conseguir una adecuada identificación de los metabolitos en etapas posteriores[80-81].

3.4.5. Pre-tratamiento de datos

La matriz de datos obtenida tras las etapas de pre-procesamiento ya descritas presentará una gran variabilidad en los valores relativos a las áreas de integración entre las distintas variables debido a las grandes diferencias de concentración que existen entre los distintos metabolitos. Este hecho puede producir que los análisis estadísticos no sean capaces de detectar la información biológica relevante usando dicha matriz de datos. Por ejemplo, los metabolitos más abundantes mostrarán mayores diferencias absolutas entre muestras en comparación con los metabolitos presentes en bajas concentraciones. Por este motivo, los datos deben tratarse de manera previa a los análisis estadísticos, especialmente cuando se utilizan métodos multivariantes, con el objetivo de normalizar la escala de las distintas variables para que puedan ser comparables entre ellas sin eliminar la información biológica que

[80] Xavier Domingo-Almenara et al., "Annotation: A Computational Solution for Streamlining Metabolomics Analysis," *Analytical Chemistry* 90, no. 1 (January 2, 2018): 480–89, https://doi.org/10.1021/acs.analchem.7b03929.

[81] Carsten Kuhl et al., "CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets," *Analytical Chemistry* 84, no. 1 (January 3, 2012): 283–89, https://doi.org/10.1021/ac202450g.

contienen. Estas técnicas de pre-tratamiento se pueden clasificar en: centrado y escalado, y transformación de datos (**Figura 16**)[74,82].

Las aproximaciones de **centrado** y **escalado**, usadas de manera conjunta, persiguen hacer las variables comparables entre ellas, siendo su empleo necesario cuando las variables difieren entre ellas en distintos órdenes de magnitud. El centrado se basa en fijar en el valor cero las medias obtenidas para todas las variables. Por su parte, el escalado divide los datos por un factor, que es diferente para cada variable, con el objetivo de minimizar las diferencias entre las distintas variables mediante la conversión de los datos en función de un factor de escala. Los métodos de escalado más utilizados en metabolómica son el **autoescalado** y el **escalado de Pareto**, los cuales emplean como factor de escala la desviación estándar o la raíz cuadrada de esta, respectivamente.

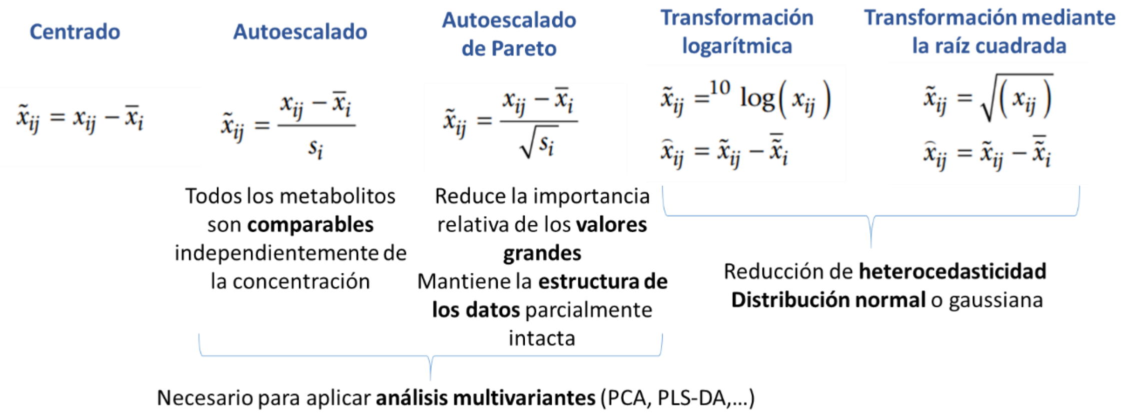

Centrado
Autoescalado
Autoescalado de Pareto
Transformación logarítmica
Transformación mediante la raíz cuadrada
$\tilde{x}_{ij} = x_{ij} - \bar{x}_i$
$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$
$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$
$\tilde{x}_{ij} = {}^{10}\log(x_{ij})$
$\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$
$\tilde{x}_{ij} = \sqrt{(x_{ij})}$
$\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$
Todos los metabolitos son comparables independientemente de la concentración
Reduce la importancia relativa de los valores grandes
Mantiene la estructura de los datos parcialmente intacta
Reducción de heterocedasticidad
Distribución normal o gaussiana
Necesario para aplicar análisis multivariantes (PCA, PLS-DA,…)

Figura 16. Esquema de las principales metodologías de centrado, escalado y transformación utilizadas en estudios de metabolómica.

[82] Robert A van den Berg et al., "Centering, Scaling, and Transformations: Improving the Biological Information Content of Metabolomics Data.," *BMC Genomics* 7 (2006): 142, https://doi.org/10.1186/1471-2164-7-142.

Por su parte, los **métodos de transformación** son conversiones no lineales de los datos que tienen como objetivo la corrección de la heterocedasticidad y conseguir una distribución normal o gaussiana de los datos. La transformación mediante la raíz cuadrada o la transformación logarítmica son las aproximaciones más utilizadas comúnmente en este tipo de estudios. Una limitación de la transformación logarítmica es que no es capaz de transformar el valor 0[74,82]. Este hecho, junto con los problemas que causan los **datos faltantes** (*missing values*, NA) en los análisis estadísticos, hace necesario el empleo de una etapa de tratamiento de dichos valores.

Una matriz de datos de metabolómica, por lo general, contiene un 20 % de *missing values* en al menos el 80 % de las variables, los cuales pueden deberse a las siguientes razones[83,84]:

- La molécula no está presente en una determinada muestra (p.ej. metabolitos exógenos derivados de fármacos, alimentos, etc.).
- La molécula está presente en la muestra pero en una concentración inferior al límite de detección (LOD) del instrumento.
- Efecto matriz (co-elución de compuestos, supresión iónica).
- Pérdida de la capacidad de separación de la columna cromatográfica y/o de la sensibilidad del espectrómetro de masas.
- Limitación en alguna etapa del pre-procesamiento de datos (*peak picking*, alineamiento, etc.).

[83] Olga Hrydziuszko and Mark R. Viant, "Missing Values in Mass Spectrometry Based Metabolomics: An Undervalued Step in the Data Processing Pipeline," *Metabolomics* 8, no. S1 (June 8, 2012): 161–74, https://doi.org/10.1007/s11306-011-0366-4.

[84] Kieu Trinh Do et al., "Characterization of Missing Values in Untargeted MS-Based Metabolomics Data and Evaluation of Missing Data Handling Strategies," *Metabolomics* 14, no. 10 (October 20, 2018): 128, https://doi.org/10.1007/s11306-018-1420-2.

Una primera aproximación para el tratamiento de este tipo de datos es el filtrado donde las variables, o muestras, que contienen un alto porcentaje de *missing values* son eliminadas. En este sentido, una estrategia común es filtrar las variables que tienen al menos un 75 % de NA en alguno de los grupos experimentales, sin embargo deben tenerse en consideración las características del diseño experimental. Por ejemplo, en los estudios de intervención nutricional muchos metabolitos de interés pueden aparecer únicamente en uno de los grupos experimentales, por lo que el uso de este filtro debe ser evaluado[85].

Para estimar los *missing values* de las variables que no han sido filtradas se han desarrollado diferentes estrategias, siendo las más populares la imputación por un valor constante (LOD/2, media, mediana,…), por bosques aleatorios (*random forest,* RF) o por los k vecinos más cercanos (*k-nearest neighbors*, kNN), entre otras[86].

3.4.6. Herramientas para el pre-procesamiento de datos

Para llevar a cabo las distintas etapas del pre-procesamiento de datos, existen diferentes softwares clasificados en dos grandes grupos: softwares comerciales y fuentes de código abierto[71].

Las principales **plataformas comerciales** son desarrolladas por los fabricantes de los propios instrumentos analíticos: Mass Profinder/Profiler (*Agilent Technologies*), Progenesis QI (*Waters Corporation*), SIEVE, Compound Discoverer (*Thermo Fisher Scientific*) o MetaboScape (*Bruker*). Estas plataformas se caracterizan por tener la

[85] Emily Grace Armitage et al., “Missing Value Imputation Strategies for Metabolomics Data,” *ELECTROPHORESIS* 36, no. 24 (December 1, 2015): 3050–60, https://doi.org/10.1002/elps.201500352.

[86] Nishith Kumar et al., “Metabolomic Biomarker Identification in Presence of Outliers and Missing Values,” *BioMed Research International* 2017 (February 14, 2017): 1–11, https://doi.org/10.1155/2017/2437608.

capacidad de llevar a cabo la mayor parte de las etapas del pre-procesamiento de datos en un único entorno, e incluso la realización de análisis estadísticos, identificación de metabolitos e interpretación biológica. Además destacan por ser intuitivos y fáciles de usar a través una interfaz sencilla[87].

Respecto a las **programas de código abierto**, su utilización en el ámbito de la metabolómica está ganando popularidad en los últimos años, siendo la mayoría de ellos desarrollados en lenguaje R.

R es un entorno y lenguaje de programación, que fue creado en 1993, orientado principalmente al análisis estadístico. Entre sus características destaca la gran cantidad de herramientas estadísticas que ofrece, su capacidad gráfica así como la posibilidad de integrarse con bases de datos. La clave de su éxito se debe principalmente a que es un proyecto abierto, donde los usuarios pueden elaborar sus programas (paquetes) y compartirlos públicamente. De hecho, R cuenta con un repositorio oficial (CRAN) que cuenta actualmente con 2697 paquetes (agosto, 2019). Además, existen multitud de paquetes que no están incluidos en dicho repositorio[88].

Los paquetes de trabajo desarrollados en R relacionados con en el pre-tratamiento de datos en metabolómica, generalmente se centran en una etapa concreta del flujo de trabajo. La **Tabla 4** muestra algunos de los principales paquetes desarrollados para cada etapa del pre-procesamiento[87]. Además de su uso gratuito, estos programas

[87] Rachel Spicer et al., "Navigating Freely-Available Software Tools for Metabolomics Analysis.," *Metabolomics : Official Journal of the Metabolomic Society* 13, no. 9 (2017): 106, https://doi.org/10.1007/s11306-017-1242-7.

[88] R Development Core Team and R R Development Core Team, "R: A Language and Environment for Statistical Computing," *R Foundation for Statistical Computing*, 2016, https://doi.org/10.1007/978-3-540-74686-7.

destacan por su versatilidad y adaptación a las necesidades concretas del usuario. Entre todos los paquetes desarrollados en R, **XCMS** es el más ampliamente utilizado por los usuarios en el área de metabolómica[89]. Las funciones de este paquete también están disponibles mediante una interfaz online orientada a usuarios principiantes (https://xcmsonline.scripps.edu/), abarcando todas las etapas del pre-procesamiento de datos así como análisis estadísticos y visualización de datos[90]. Este tipo de plataforma basada en un servidor online de acceso libre está ganando popularidad en los últimos años con la aparición de varias herramientas de esta tipología como Workflow4Metabolomics (https://workflow4metabolomics.org/), o PhenoMeNal (http://phenomenal-h2020.eu/home/), entre otras[87].

Tabla 4. Principales paquetes basados en R.

Etapa de procesamiento	Paquetes de R
***Peak Picking* y alineamiento**	XCMS, IPO, MZmine, OpenMS, MetAlign
Normalización	MetNormalizer, BatchCorr, MixNorm, Normalyzer, NormalizeMets
Annotation	CAMERA, AMDIS, RAMClustR, MSClust
Análisis estadísticos	MetabolAnalyze, Ionwinze, MetabolAnalyze, muma, rolps, MUVR,

Las características que se deben considerar a la hora de elegir un software para llevar a cabo la etapa de pre-procesamiento son: facilidad de uso (intuitivo), coste/uso libre, calidad y rendimiento y la cobertura de todas las etapas del pre-procesamiento[71]. Desafortunadamente, no hay una única plataforma, ni de acceso libre ni comercial,

[89] Ralf J. M. Weber et al., "Computational Tools and Workflows in Metabolomics: An International Survey Highlights the Opportunity for Harmonisation through Galaxy," *Metabolomics* 13, no. 2 (February 27, 2017): 12, https://doi.org/10.1007/s11306-016-1147-x.

[90] Ralf Tautenhahn et al., "XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data.," *Analytical Chemistry* 84, no. 11 (June 5, 2012): 5035–39, https://doi.org/10.1021/ac300698c.

que cumpla todas estas características. Las principales desventajas de las plataformas comerciales son su coste y la limitada oferta en cuanto a las soluciones que ofrecen en relación con las herramientas de acceso libre. Por ejemplo, las metodologías de normalización que ofrecen estos software comerciales se resumen en métodos basados en el uso de patrones internos, MSTS o estadísticos basados en las distribuciones probabilísticas (cuantiles, percentil, PQN, etc.). Por el contrario, las metodologías de normalización basadas en las muestras QC han sido desarrolladas principalmente en programas de acceso libre. Respecto a estos, sus principales desventajas radican en la necesidad de transformar el formato de los datos adquiridos, de conectar diferentes paquetes para poder abarcar todas las etapas del pre-procesamiento de datos, además de que su utilización requiere por parte del usuario un cierto nivel de conocimientos en programación.

3.5. Análisis quimiométricos

El objetivo principal de esta etapa es conocer qué variables muestran un efecto significativo en relación a la hipótesis planteada en el estudio. Para ello se aplican diferentes análisis quimiométricos en la matriz de datos obtenida tras las distintas etapas de pre-procesamiento de datos descritas en la sección anterior.

La **quimiometría** se define como la disciplina de la química que utiliza las matemáticas y métodos estadísticos para el diseño de experimentos y la obtención de información relevante a partir de datos químicos. El avance de las herramientas quimiométricas ha permitido el desarrollo de los estudios de metabolómica no dirigida, gracias a que permiten visualizar, explorar y analizar estadísticamente las matrices de datos obtenidas[91].

Los métodos estadísticos utilizados en metabolómica se pueden dividir en dos grandes grupos: análisis **univariantes** y **multivariantes**. Dado el alto número de variables y la complejidad de las matrices de datos, las técnicas multivariantes son las más empleadas en este tipo de estudios. La **Figura 17** muestra las principales metodologías estadísticas tanto univariantes como multivariantes empleadas en metabolómica.

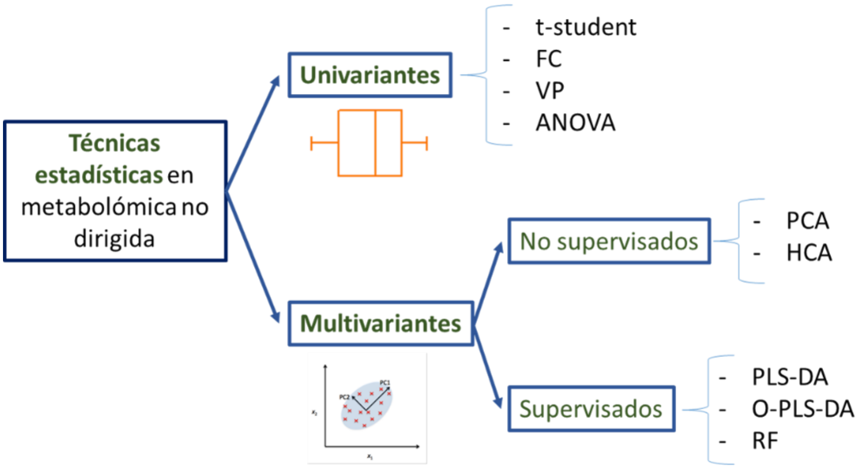


Figura 17. Principales técnicas estadísticas utilizadas en metabolómica.

Al igual que las herramientas de pre-procesamiento de datos, los análisis estadísticos pueden llevarse a cabo tanto en **software comerciales** como en **plataformas libres**. Además de los software de las casas comerciales de los equipos, SIMCA-P es uno de los software comerciales más utilizados en metabolómica. Por otra parte, existen multitud de programas basados en R, varios de ellos mencionados en la **Tabla 4** de la *sección*

[91] Rui Climaco Pinto, “Chemometrics Methods and Strategies in Metabolomics,” in *Advances in Experimental Medicine and Biology*, vol. 965, 2017, 163–90, https://doi.org/10.1007/978-3-319-47656-8_7.

3.4.6, y varias páginas web, siendo MetaboAnalyst[92] (https://www.metaboanalyst.ca/) la más destacada, que han sido desarrolladas como plataformas de acceso libre para el análisis estadístico de los datos obtenidos en estudios de metabolómica.

3.5.1. Análisis estadístico univariante

Los **métodos univariantes** se caracterizan por el análisis de cada variable de manera individual e independiente. Estas técnicas no tienen en cuenta la presencia de posibles interacciones entre los distintos metabolitos, siendo esta la principal desventaja de este tipo de test estadístico en el campo de la metabolómica. Además, tampoco tienen la capacidad de evaluar el efecto de otros factores, como el género, la nacionalidad, el estilo de vida (dieta, tabaco, actividad física…), etc., que influyen en los resultados incrementando la probabilidad de obtener resultados falsos positivos y/o negativos[21].

Existen diferentes métodos estadísticos univariantes, siendo los más utilizados la prueba **t de student** y el **análisis de la varianza** (**ANOVA**), los cuáles evalúan si hay diferencias significativas en las medias obtenidas para una variable entre dos o varias clases de muestras del estudio, respectivamente. La utilización de estos análisis estadísticos requiere que los datos utilizados presenten una distribución normal.

El empleo de análisis univariantes tiene asociado una alta probabilidad de encontrar resultados estadísticamente significativos por azar (**falso positivo**) debido al elevado número de variables que se comparan en estudios de metabolómica no dirigida. Para evitar este problema, existen diferentes métodos de ajuste del p-valor, como la corrección de Bonferroni, el ajuste de Holm o la corrección de Benjamini-Hochberg

[92] Jasmine Chong et al., "MetaboAnalyst 4.0: Towards More Transparent and Integrative Metabolomics Analysis," *Nucleic Acids Research* 46, no. W1 (July 2, 2018): W486–94, https://doi.org/10.1093/nar/gky310.

(*false-discovery rate* o *FDR*). Estos métodos se diferencian en el balance particular que otorgan entre la relación de errores de tipo I (falsos positivos) y tipo II (falsos negativos). Así, la **corrección de Bonferroni** es la metodología más conservativa que evita en un mayor grado la aparición de falsos positivos a expensas de incrementar el número de falsos negativos. Sin embargo, la **aproximación FDR** es menos conservativa y se centra en controlar la proporción de falsos positivos con el objetivo de no superar un determinado valor[93].

El test ***Fold Change* (FC)** se utiliza en estudios que comparan dos grupos de muestras (p.ej. casos vs controles), para determinar la relación entre las medias de ambos grupos. Los resultados de este test, junto con el test estadístico de la t-student, suelen representarse mediante el **gráfico de Volcano** (VP), con el objetivo de identificar y visualizar las variables significativas en ambos (**Figura 18**)[94].

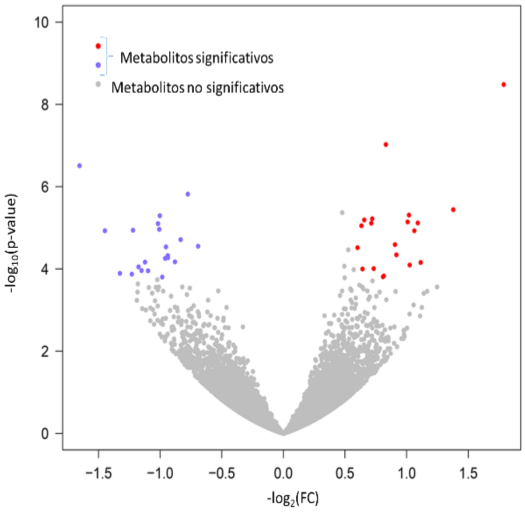


Figura 18. Ejemplo de un gráfico de Volcano.

[93] Shi-Yi Chen, Zhe Feng, and Xiaolian Yi, "A General Introduction to Adjustment for Multiple Comparisons.," *Journal of Thoracic Disease* 9, no. 6 (June 2017): 1725–29, https://doi.org/10.21037/jtd.2017.05.34.

[94] Manhoi Hur et al., "A Global Approach to Analysis and Interpretation of Metabolic Data for Plant Natural Product Discovery.," *Natural Product Reports* 30, no. 4 (April 2013): 565–83, https://doi.org/10.1039/c3np20111b.

3.5.2. Modelos estadísticos multivariantes

Los **métodos multivariantes** se caracterizan por tener en cuenta todas las variables obtenidas en la etapa de pre-procesamiento de datos de manera simultánea. Por consiguiente, su principal ventaja reside en la posibilidad de identificar relaciones e interacciones entre los distintos metabolitos, algo que es imposible con los análisis univariantes. Las metodologías multivariantes se clasifican en dos grupos: **análisis no supervisados** y **supervisados**. Los métodos no supervisados no tienen en cuenta el tipo o clase a la que pertenecen las distintas muestras, en cambio, esta información si es considerada por los métodos supervisados con la finalidad de detectar qué metabolitos o combinación de estos están más relacionados con el fenotipo estudiado[21].

3.5.2.1. Métodos no supervisados

Los métodos no supervisados son generalmente empleados con el objetivo de visualizar como se distribuyen los datos, y de este modo, poder detectar tendencias relacionadas con las condiciones experimentales o biológicas del estudio. La técnica no supervisada más empleada en estudios no dirigidos es el **análisis de componentes principales** (**PCA**). Este tipo de análisis se utiliza también con fines de **control de calidad de los datos**, permitiendo la detección de derivas analíticas así como de datos anómalos[95]. PCA se basa en la transformación de los datos en un conjunto de variables ortogonales denominadas componentes principales, donde se maximiza la varianza explicada por cada una de las componentes.

[95] Bradley Worley and Robert Powers, "Multivariate Analysis in Metabolomics.," *Current Metabolomics* 1, no. 1 (2013): 92–107, https://doi.org/10.2174/2213235X11301010092.

Como resultado de un análisis PCA se obtienen dos tipos de parámetros denominados *loadings* y *scores*. Los ***loadings*** hacen referencia a las contribuciones de cada variable a la componente principal, mientras que los ***scores*** representan la proyección de cada muestra en dichas componentes (**Figura 19**). La representación de estos *scores* permite observar la distribución de la muestras en un espacio uni-, bi- o tridimensional, dependiendo del número de componentes principales seleccionado en función de la varianza explicada por ellas.

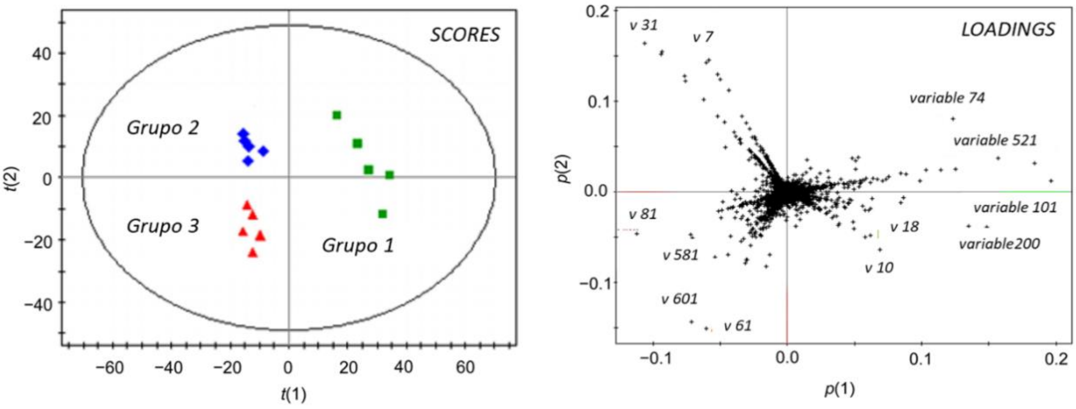


Figura 19. Ejemplo de scores y loadings (Análisis PCA).

Los **análisis de clusters** no supervisados también son ampliamente utilizados en el ámbito de la metabolómica. Estos métodos se basan en la determinación de distancias entre las variables o muestras en función de la similitud entre ellas. De esta manera, las variables o muestras que estén más correlacionadas, presentarán valores de distancia pequeños y se agruparán formando clusters. Estos agrupamientos son fácilmente observables mediante la representación gráfica de las distancias por medio de **dendrogramas**[96]. Además, los clusters de las muestras y de las variables pueden ser representados de manera conjunta mediante un gráfico denominado ***heatmap,*** donde

[96] Joshua Heinemann, "Cluster Analysis of Untargeted Metabolomic Experiments" (Humana Press, New York, NY, 2019), 275–85, https://doi.org/10.1007/978-1-4939-8757-3_16.

la abundancia relativa de los metabolitos detectados en cada muestra se representa en una escala de intensidad de color (**Figura 20**)[97].

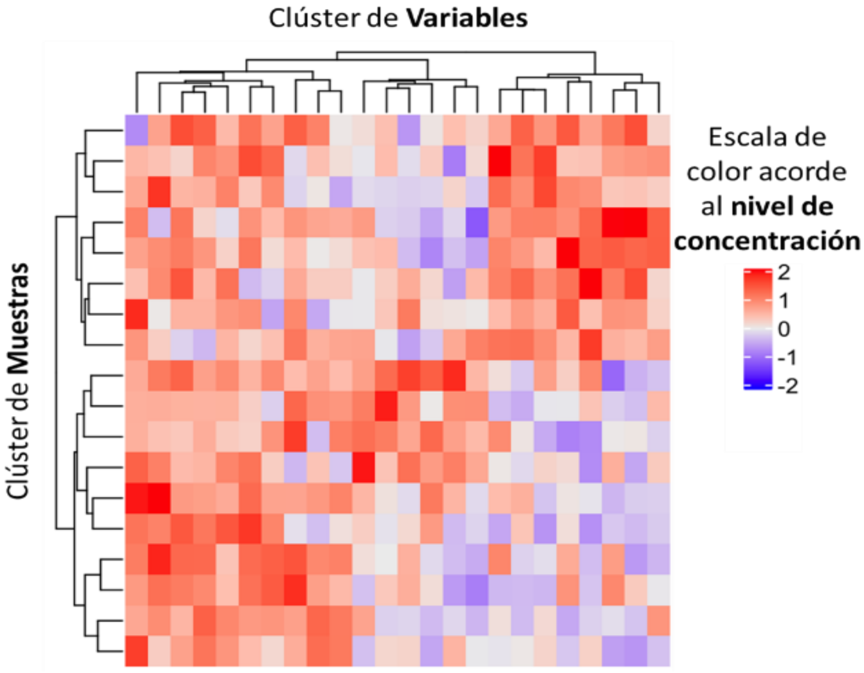


Figura 20. Ejemplo de análisis de cluster vía heatmap.

3.5.2.2. Métodos supervisados

Los **modelos supervisados** se basan en identificar qué variables presentan mayor relación con la cuestión biológica del estudio, y son la base para la construcción de modelos clasificatorios de muestras[98].

La **regresión de mínimos cuadrados parciales** (*Partial Least Squares Regression*, **PLS**) es el método supervisado más utilizado en metabolómica. Se puede emplear como un análisis de regresión o combinado con un **análisis discriminante** entre dos o más clases

[97] Paul H Benton et al., "An Interactive Cluster Heat Map to Visualize and Explore Multidimensional Metabolomic Data," *Metabolomics : Official Journal of the Metabolomic Society* 11, no. 4 (August 1, 2015): 1029–34, https://doi.org/10.1007/s11306-014-0759-2.

[98] Jianguo Xia et al., "Translational Biomarker Discovery in Clinical Metabolomics: An Introductory Tutorial," *Metabolomics* 9, no. 2 (2013): 280–99, https://doi.org/10.1007/s11306-012-0482-9.

(**PLS-DA**). El fundamento de este modelo es similar al de PCA, pero en vez de maximizar la varianza explicada por el conjunto de datos, es maximizada la covarianza entre la variable dependiente de interés (p. ej: clases, grupos de muestra, etc.) y las variables independientes correspondientes al conjunto de datos obtenidos a través del flujo de trabajo[99]. De este modo, los ***loadings*** de un modelo PLS representan la contribución de una variable a la separación entre los diferentes grupos de muestras.

El método de **proyecciones ortogonales a estructuras latentes** (*Orthogonal Projections to Latent Structures*, **OPLS**) es una variante del método PLS que está siendo frecuentemente aplicado en estudios de metabolómica[100]. Al igual que el método PLS-DA, OPLS también se utiliza combinado a un análisis discriminante entre grupos (**OPLS-DA**). Este método divide la variabilidad de los datos en dos componentes: una primera componente que incluye la información correlacionada con la variable independiente (variación predictiva) y la segunda componente de información no correlacionada a dicha variable independiente (ortogonal). Tanto el método PLS como el OPLS presentan la misma capacidad predictiva, siendo el tipo de interpretación la diferencia existente entre ellos (**Figura 21**)[101].

[99] Piotr S. Gromski et al., "A Tutorial Review: Metabolomics and Partial Least Squares-Discriminant Analysis – a Marriage of Convenience or a Shotgun Wedding," *Analytica Chimica Acta* 879 (June 2015): 10–23, https://doi.org/10.1016/j.aca.2015.02.012.

[100] Mohamed N Triba et al., "PLS/OPLS Models in Metabolomics: The Impact of Permutation of Dataset Rows on the K-Fold Cross-Validation Quality Parameters.," *Molecular BioSystems* 11, no. 1 (January 2015): 13–19, https://doi.org/10.1039/c4mb00414k.

[101] Max Bylesjö et al., "OPLS Discriminant Analysis: Combining the Strengths of PLS-DA and SIMCA Classification," *Journal of Chemometrics* 20, no. 8–10 (August 1, 2006): 341–51, https://doi.org/10.1002/cem.1006.

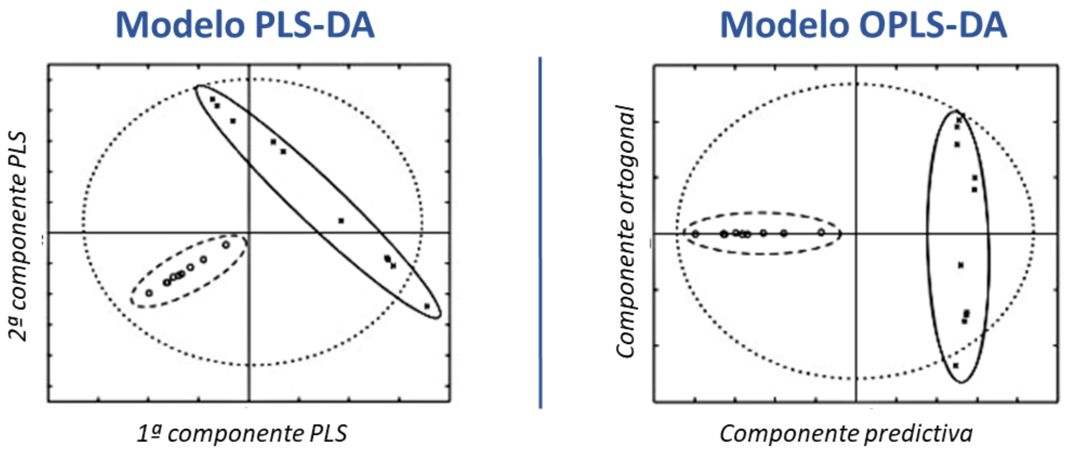


Figura 21. Diferencias entre los modelos PLS-DA y OPLS-DA.

Estos métodos se basan en la búsqueda de un patrón de respuesta lineal en los datos metabolómicos. Sin embargo existen otros métodos supervisados que no se basan en esta característica, dado que entienden que los procesos biológicos, debido a su elevada complejidad, no siguen comúnmente procesos que se pueden caracterizar de manera lineal. En este sentido, se han propuesto técnicas estadísticas basadas en el reconocimiento de patrones o de aprendizaje automático (*machine learning*) dentro del ámbito de la inteligencia artificial. Los métodos de este tipo más utilizados en metabolómica son el método de bosques aleatorios (***RanfomForest***, ***RF***), de máquinas de vectores de soporte (*Support Vector Machines*, *SVMs*) y el método *Kernel* de mínimos cuadrados parciales (*Kernel-PLS*)[102].

[102] Lunzhao Yi et al., "Chemometric Methods in Data Processing of Mass Spectrometry-Based Metabolomics: A Review," *Analytica Chimica Acta* 914 (March 31, 2016): 17–34, https://doi.org/10.1016/J.ACA.2016.02.001.

3.5.2.3. Validación de modelos multivariantes

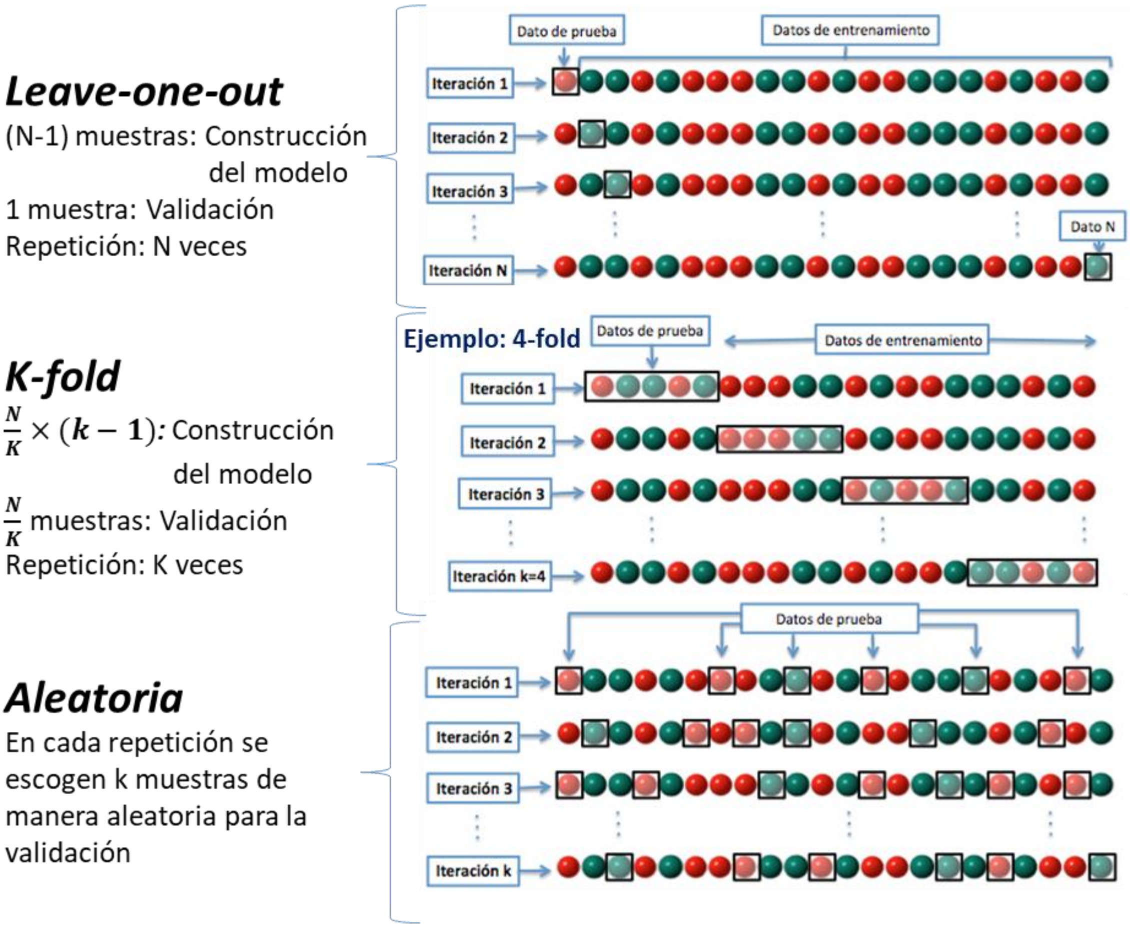
Los métodos multivariantes se caracterizan por la posibilidad de sobreajustar los datos experimentales a las condiciones del modelo (***overfitting***). Por este motivo, la validación de los modelos es un paso necesario para garantizar la calidad de los resultados. En este sentido, existen principalmente tres métodos de validación[95,98]:

- **Validación cruzada** (*cross-validation*, *CV*). Estos métodos se basan en la división de los datos en dos subconjuntos para usarlos en la construcción y validación de los modelos, de manera separada. Para llevar a cabo esta división de la matriz de datos se han desarrollado varias estrategias: validación cruzada dejando una muestra fuera (*leave-one-out*), K iteraciones (K-fold), aleatoria o doble CV (**Figura 22**)[103].

  En este sentido se pueden seleccionar diferentes parámetros estadísticos para determinar el grado de validez del método escogido, siendo los más habituales el **coeficiente de determinación** (**R^2**), de **predicción** (**Q^2**), o el **área bajo la curva ROC** (*Receiver Operating Characteristic*). En estudios biológicos se consideran generalmente aceptables valores de Q^2 y AUC superiores a 0.4 y 0.7, respectivamente[104].

[103] Ewa Szymańska et al., “Double-Check: Validation of Diagnostic Statistics for PLS-DA Models in Metabolomics Studies,” *Metabolomics* 8, no. S1 (June 8, 2012): 3–16, https://doi.org/10.1007/s11306-011-0330-3.

[104] Jayawant N. Mandrekar, “Receiver Operating Characteristic Curve in Diagnostic Test Assessment,” *Journal of Thoracic Oncology* 5, no. 9 (September 1, 2010): 1315–16, https://doi.org/10.1097/JTO.0B013E3181EC173D.



Leave-one-out

(N-1) muestras: Construcción del modelo
1 muestra: Validación
Repetición: N veces

K-fold

$\frac{N}{K} \times (k-1)$: Construcción del modelo
$\frac{N}{K}$ muestras: Validación
Repetición: K veces

Aleatoria

En cada repetición se escogen k muestras de manera aleatoria para la validación

Figura 22. Tipos de división del conjunto de datos para la validación cruzada de los modelos multivariantes.

La **curva ROC** es considerada la metodología estándar para la evaluación y validación de modelos clasificatorios, tanto multivariantes como univariantes, entre dos clases de muestras. Esta curva representa gráficamente la relación entre los porcentajes de verdaderos positivos (sensibilidad) y el de falsos positivos (1-especifidad) obtenidos por un sistema clasificatorio (PLS-DA, RF, biomarcador, etc.) entre dos clases según se va modificando el umbral de discriminación[105]. Un modelo ideal que clasifique correctamente todas las muestras (p.ej.: pacientes enfermos, valores positivos, respecto a individuos sanos, valores negativos), obtendrá valores de sensibilidad, especificidad (% de verdaderos negativos) y AUC de 1. En cambio, un modelo que obtiene un valor de AUC de 0.5 indica que el resultado es equivalente a la clasificación aleatoria de las muestras (**Figura 23**)[98].

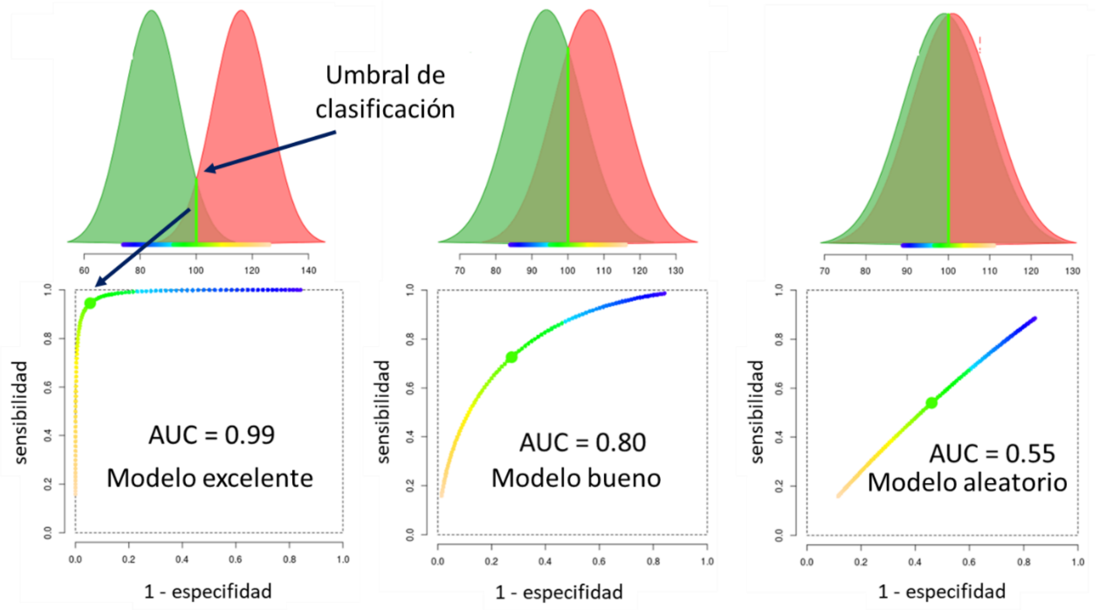
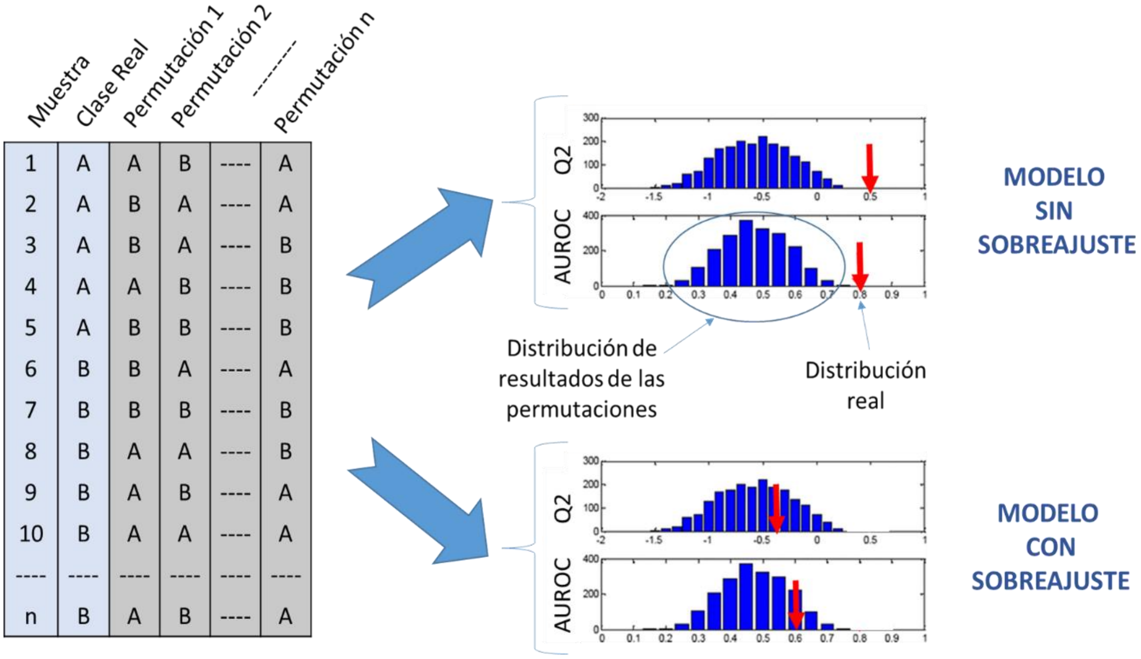


Figura 23. Ejemplos de 3 curvas ROC obtenidas para tres distribuciones de datos diferentes (https://www.bioestadistica.uma.es/analisis/roc1/).

- **Test de permutaciones.** Este test se utiliza con el objetivo de verificar que el modelo estadístico no presenta sobreajuste. Para ello, el test de permutación se basa en utilizar el modelo clasificatorio a validar con las clases de las muestras asignadas de manera aleatoria. De esta forma, se compara estadísticamente los resultados obtenidos por el modelo original con los

[105] Rolf Bünger and Robert T Mallet, “Metabolomics and Receiver Operating Characteristic Analysis: A Promising Approach for Sepsis Diagnosis.,” *Critical Care Medicine* 44, no. 9 (2016): 1784–85, https://doi.org/10.1097/CCM.0000000000001795.

obtenidos por medio de las permutaciones. Si los resultados del modelo original presentan diferencias estadísticamente significativas con la distribución de resultados obtenida a partir de los modelos permutados, se verificará que no existe *overfitting* en el modelo (**Figura 24**)[106,107].



Muestra	Clase Real	Permutación 1	Permutación 2	--------	Permutación n
1	A	A	B	----	A
2	A	B	A	----	A
3	A	B	A	----	B
4	A	A	B	----	B
5	A	B	B	----	B
6	B	B	A	----	A
7	B	B	B	----	B
8	B	A	A	----	B
9	B	A	B	----	A
10	B	A	A	----	A
----	----	----	----	----	----
n	B	A	B	----	A

Figura 24. Test de permutaciones. Ejemplo de modelos con y sin sobreajuste.

- **Validación externa**. Estos métodos se basan en la validación de los resultados mediante el análisis de un nuevo conjunto de muestras independientes a los utilizados en la creación de los modelos estadísticos, con la finalidad de verificar y generalizar los resultados obtenidos. Este procedimiento de validación se utiliza generalmente cuando se quiere dar una aplicación a los resultados obtenidos (p.ej.: aplicación clínica). Para ello, la validación externa

[106] Triba et al., "PLS/OPLS Models in Metabolomics: The Impact of Permutation of Dataset Rows on the K-Fold Cross-Validation Quality Parameters."

[107] Johan A. Westerhuis et al., "Assessment of PLSDA Cross Validation," *Metabolomics* 4, no. 1 (2008): 81–89, https://doi.org/10.1007/s11306-007-0099-6.

se realiza después de la etapa de identificación, utilizando una metodología metabolómica dirigida para el análisis y cuantificación de los compuestos que se desean validar[108-110].

3.6. Identificación de los metabolitos

El objetivo de esta etapa es la asignación de una identidad (nomenclatura y estructura química) a las variables que resultaron significativas en los análisis estadísticos para poder interpretar biológicamente los resultados de acuerdo a los objetivos del estudio.

En esta etapa, se utilizan diferentes tipos de información obtenida a lo largo del flujo de trabajo para poder identificar los metabolitos significativos, como por ejemplo los tiempos de retención, los valores de masa exacta, las distribuciones isotópicas, los aductos formados así como los patrones de fragmentación obtenidos por análisis de MS/MS.

Los **tiempos de retención** son de gran utilidad en el caso de que se comparen los resultados con el de un patrón comercial analizado bajo las mismas condiciones instrumentales. No obstante, los tiempos de retención ofrecen una información orientativa sobre la familia de compuestos al que pueden pertenecer las señales de interés debido a su alta relación con la polaridad y por tanto estructura de los mismos.

[108] S.E Bleeker et al., “External Validation Is Necessary in Prediction Research:: A Clinical Example,” *Journal of Clinical Epidemiology* 56, no. 9 (September 1, 2003): 826–32, https://doi.org/10.1016/S0895-4356(03)00207-5.

[109] Calena R Marchand et al., “A Framework for Development of Useful Metabolomic Biomarkers and Their Effective Knowledge Translation.,” *Metabolites* 8, no. 4 (September 30, 2018), https://doi.org/10.3390/metabo8040059.

[110] Shama Naz et al., “Method Validation Strategies Involved in Non-Targeted Metabolomics,” *Journal of Chromatography A* 1353 (August 2014): 99–105, https://doi.org/10.1016/j.chroma.2014.04.071.

Los datos de **masa exacta** así como de las **distribuciones isotópicas** obtenidos por MS son de gran utilidad en la tarea de identificación, dado que permite la obtención de las fórmulas moleculares más probables de las moléculas de interés. Esta predicción de las fórmulas moleculares es obtenida por los software comerciales de tratamiento de datos de MS, aunque también existen herramientas de acceso libre como *Sirius*[111]. El tipo de **aductos** detectados también puede proporcionar información orientativa de la familia de metabolitos a identificar, ya que algunas de ellas son generalmente detectadas formando un tipo de aducto determinado.

Finalmente, los **espectros de fragmentación** obtenidos por análisis de **MS/MS** son generalmente la información que posibilita la asignación de una identidad química a las señales candidatas por medio de su comparación con los disponibles en librerías comerciales o bases de datos online de acceso libre, como *Human Metabolome Database* (HMDB) (http://www.hmdb.ca/), *METLIN* (https://metlin.scripps.edu/), *Kyoto Encyclopedia of Genes and Genomes* (KEGG) (https://www.genome.jp/kegg/), LipidMaps (https://www.lipidmaps.org/) o *MassBank* (http://www.massbank.jp/) (**Figura 25**). Además, recientemente se han desarrollado herramientas como *CEU Mass Mediator* (CeuMM) (http://ceumass.eps.uspceu.es/) que aúnan información de dichas plataformas con el objetivo de facilitar la búsqueda a través de las diferentes bases de datos mencionadas[112].

[111] Kai Dührkop et al., "SIRIUS 4: A Rapid Tool for Turning Tandem Mass Spectra into Metabolite Structure Information," *Nature Methods*, 2019, https://doi.org/10.1038/s41592-019-0344-8.

[112] Alberto Gil de la Fuente et al., "Differentiating Signals to Make Biological Sense – A Guide through Databases for MS-Based Non-Targeted Metabolomics," *Electrophoresis* 38, no. 18 (2017): 2242–56, https://doi.org/10.1002/elps.201700070.

A pesar de que la información que contienen estas bases de datos se está incrementado en los últimos años gracias a las contribuciones de los investigadores, en la mayor parte de estudios todavía existe un porcentaje significativo de señales que no logran ser identificadas. Por ello, esta etapa es actualmente el principal cuello de botella de los estudios de metabolómica no dirigida llevados a cabo mediante MS. No obstante, se han desarrollado herramientas *in silico*, *como MetFrag* (https://ipb-halle.github.io/MetFrag/) o *Metabolomics In silico Network Expansion Databases* (MINE) (https://minedatabase.mcs.anl.gov/), que predicen teóricamente por medio de algoritmos los patrones de fragmentación más probables de una determinada molécula[113]. La comparación de los datos experimentales con datos *in silico* debe realizarse con precaución, dado que tienen asociado un riesgo de asignación de falsos positivos.

Figura 25. Principales plataformas de acceso libre para la identificación de metabolitos.

[113] Christoph Ruttkies et al., “MetFrag Relaunched: Incorporating Strategies beyond in Silico Fragmentation,” *Journal of Cheminformatics* 8, no. 1 (December 29, 2016): 3, https://doi.org/10.1186/s13321-016-0115-9.

3.6.1. Niveles de identificación

La interpretación biológica de los resultados presentará una mayor o menor validez en función de los parámetros utilizados para identificar los metabolitos. Con el objetivo de establecer unos niveles de confianza en esta etapa, se han establecido criterios para asignar niveles de identificación. En 2007, el grupo de trabajo de análisis químico (*Metabolomics Standards Initiative (MSI)*) propuso los siguientes niveles de identificación[68]:

- **Nivel 1: Metabolito identificado.** Se necesita que al menos dos o más propiedades (p. ej.: RT, m/z, MS/MS) de un patrón comercial analizado bajo las mismas condiciones experimentales coincidan con las obtenidas para la señal que se desea identificar.
- **Nivel 2: Metabolito anotado** (*putatively annotated compound*). En este caso la comparación se realiza con las bibliotecas y/o bases de datos públicas o comerciales, en vez de con patrones comerciales.
- **Nivel 3: Clase de compuesto caracterizada** (*putatively characterized compound class*)**.** El metabolito es asignado a una familia de compuestos en base a sus propiedades fisicoquímicas y/o su similitud espectral con metabolitos de dicha familia.
- **Nivel 4: Compuesto desconocido.** La señal de interés no ha podido ser identificada ni clasificada en ninguna familia de metabolitos. No obstante este compuesto puede diferenciarse y cuantificarse en función de los resultados espectrales.

Estos niveles de identificación han sido ampliamente utilizados, aunque en los últimos años se han propuesto distintas revisiones para definir con mayor precisión el tipo de identificación llevada a cabo en cada caso. Estos cambios principalmente afectan al segundo nivel de identificación, en base a si la información comparada con las bases de datos corresponde al espectro de MS1 (identificación tentativa) o MS/MS (estructura tentativa)[15]. La **Figura 26** muestra un esquema de los 4 niveles de identificación propuestos por *Sumner et al.*[68] incluyendo estas modificaciones.

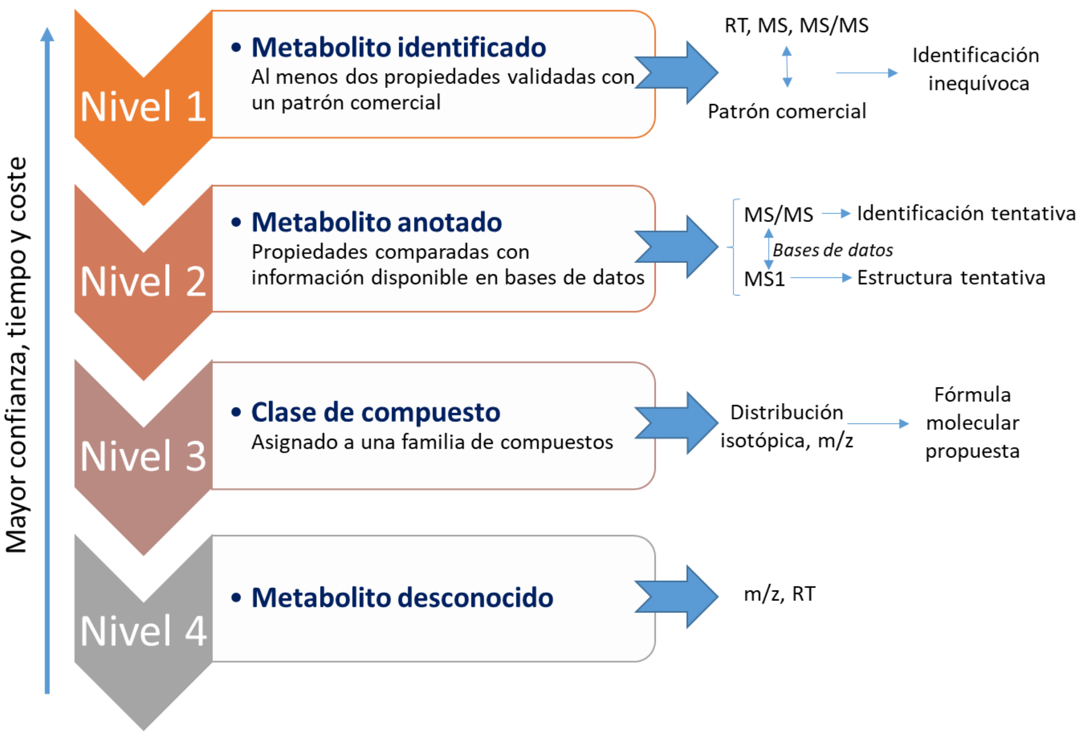

Mayor confianza, tiempo y coste
Nivel 1
• Metabolito identificado
Al menos dos propiedades validadas con un patrón comercial
RT, MS, MS/MS
Patrón comercial
Identificación inequívoca
Nivel 2
• Metabolito anotado
Propiedades comparadas con información disponible en bases de datos
MS/MS
Identificación tentativa
Bases de datos
MS1
Estructura tentativa
Nivel 3
• Clase de compuesto
Asignado a una familia de compuestos
Distribución isotópica, m/z
Fórmula molecular propuesta
Nivel 4
• Metabolito desconocido
m/z, RT

Figura 26. Niveles de identificación de acuerdo a MSI.

3.7. Interpretación biológica

Una vez identificados los metabolitos de interés, la última etapa del flujo de trabajo corresponde a la interpretación biológica de los resultados. Esta interpretación permite resolver la hipótesis planteada en el estudio y/o la generación de nuevas hipótesis. Se debe tener en cuenta que la mayoría de los metabolitos pueden presentar múltiples roles en el metabolismo dentro de un sistema biológico, además de que varios de ellos pueden estar implicados en una misma ruta metabólica.

Las **bases de datos** generalmente proporcionan información sobre las rutas metabólicas y los procesos bioquímicos en los que están implicados los metabolitos, lo que es de gran utilidad para la interpretación biológica de los resultados.

A partir de la información disponible en bases de datos (KEGG, *Small Molecule Pathway DataBase* (SMPDB), WikiPathways, MetaCyc, o *The Edinburgh Human Metabolic Network* (EHMN)), se han desarrollado herramientas que permiten evaluar directamente las **rutas metabólicas** alteradas significativamente en función de los resultados obtenidos. Estos métodos se denominan análisis de enriquecimiento para un grupo de metabolitos (*Metabolite Set Enrichement Analysis*, MSEA)[114] y análisis de rutas metabólicas (*Metabolic Pathway Analysis*, MetPA)[115]. A su vez, existen diferentes modalidades de análisis de tipo MSEA: análisis de sobrerrepresentación (*Overrepresentation Analysis*, ORA), perfil de muestra única (*Single-Sample Profiling*, SSP) y análisis de enriquecimiento cuantitativo (*Quantitative Enrichment Analysis*, QEA). ORA utiliza únicamente una lista de metabolitos para identificar las rutas metabólicas relacionadas con estos, QEA requiere además las concentraciones obtenidas de la lista de metabolitos para todas las muestras y SSP se centra únicamente en una muestra de manera individual. Por su parte, los métodos MetPA

[114] J. Xia and D. S. Wishart, “MSEA: A Web-Based Tool to Identify Biologically Meaningful Patterns in Quantitative Metabolomic Data,” *Nucleic Acids Research* 38, no. Web Server (July 1, 2010): W71–77, https://doi.org/10.1093/nar/gkq329.

[115] J. Xia and D. S. Wishart, “MetPA: A Web-Based Metabolomics Tool for Pathway Analysis and Visualization,” *Bioinformatics* 26, no. 18 (September 15, 2010): 2342–44, https://doi.org/10.1093/bioinformatics/btq418.

son una extensión de las herramientas MSEA que incluyen la medida del impacto de un metabolito desregulado en una ruta metabólica determinada[21].

En la actualidad existen diversas plataformas que permiten la realización de este tipo de análisis, como por ejemplo MetaboAnalyst, Paintomics, Cytoscape, MeltDB, BioCyc/HumanCyc, IMPaLA, MetScape2 o Metabox. La principal limitación de estas herramientas está relacionada con el número de metabolitos y rutas metabólicas registradas en las distintas bases de datos[116]. La **Figura 27** muestra dos ejemplos de este tipo de análisis.

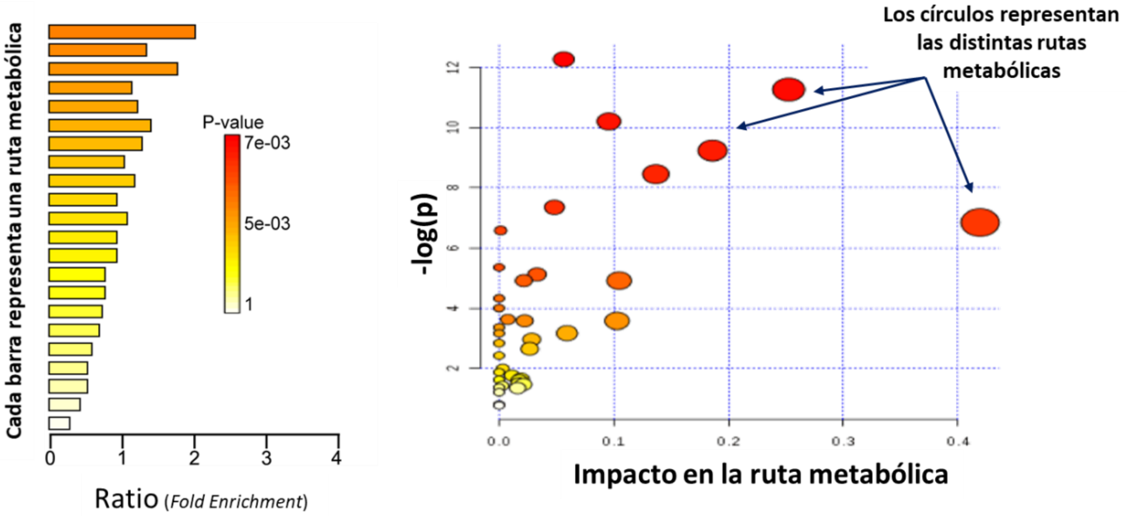


Figura 27. Ejemplos de análisis MSEA y MetPA obtenidos por MetaboAnalyst.

Con la finalidad de mejorar la interpretación de los mecanismos y relaciones biológicas que tienen lugar en el organismo, el número de estudios que correlacionan datos de metabolómica con los obtenidos por otras ciencias "*ómicas*" (Genómica,

[116] Anna Marco-Ramell et al., "Evaluation and Comparison of Bioinformatic Tools for the Enrichment Analysis of Metabolomics Data," *BMC Bioinformatics* 19, no. 1 (December 2, 2018): 1, https://doi.org/10.1186/s12859-017-2006-0.

Transciptómica, Proteómica, Microbiómica) está aumentando considerablemente en los últimos años[117].

4. CAMPOS DE APLICACIÓN DE LAS ESTRATEGIAS METABOLÓMICAS

Las estrategias metabolómicas están siendo utilizadas actualmente en un amplio rango de aplicaciones en diferentes ámbitos del conocimiento tal y como se muestran en la **Figura 28**. En las siguientes subsecciones, se detallan las potencialidades de la metabolómica en el ámbito de los compuestos bioactivos y la nutrición así como en el estudio de patologías, al ser los campos de aplicación en los que se focalizan los bloques A y B de la presente tesis doctoral, respectivamente.

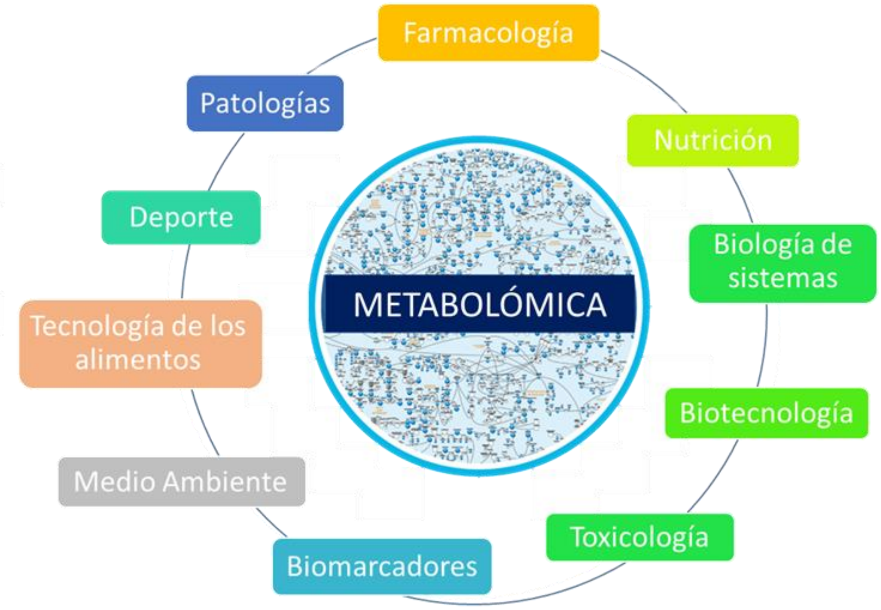


Figura 28. Campos de aplicación de la metabolómica.

[117] Su Chu et al., "Integration of Metabolomic and Other Omics Data in Population-Based Study Designs: An Epidemiological Perspective," *Metabolites* 9, no. 6 (June 18, 2019): 117, https://doi.org/10.3390/metabo9060117.

4.1. Aplicación de la metabolómica en el ámbito de los compuestos bioactivos y la nutrición

El campo de la nutrición y alimentación es un área de conocimiento donde las estrategias metabolómicas han presentado una alta aplicabilidad en los últimos años debido a su potencial para la detección de compuestos gracias a su alta sensibilidad, su capacidad de reflejar el estado actual de un individuo y su alta relación con la respuesta fenotípica. Además, hay que considerar que la alimentación es uno de los principales factores que afectan a la composición del metaboloma, tanto al formado por los metabolitos endógenos como exógenos[22,118]. Las estrategias metabolómicas en este campo de aplicación se han utilizado generalmente persiguiendo alguno de los siguientes objetivos[119,120]:

- Caracterizar la composición de los metabolitos presentes en las matrices alimentarias. La información exhaustiva del perfil metabolómico de los alimentos permite el **descubrimiento de nuevos compuestos**, entre los cuales encontramos **compuestos bioactivos** (fitoquímicos) o tóxicos, lo que garantiza un mayor conocimiento de los efectos saludables de los alimentos y una mayor **seguridad alimentaria**, respectivamente. Además, las herramientas metabolómicas permiten la detección de adulteraciones alimentarias gracias a

[118] Marta Guasch-Ferré, Shilpa N Bhupathiraju, and Frank B Hu, "Use of Metabolomics in Improving Assessment of Dietary Intake.," *Clinical Chemistry* 64, no. 1 (2018): 82–98, https://doi.org/10.1373/clinchem.2017.272344.

[119] Lorraine Brennan, "Metabolomics in Nutrition Research: Current Status and Perspectives: Figure 1," *Biochemical Society Transactions* 41, no. 2 (2013): 670–73, https://doi.org/10.1042/BST20120350.

[120] Helena Gibbons, Aoife O'Gorman, and Lorraine Brennan, "Metabolomics as a Tool in Nutritional Research," *Current Opinion in Lipidology* 26, no. 1 (2015): 30–34, https://doi.org/10.1097/MOL.0000000000000140.

la identificación de la huella dactilar de distintas variedades o productos, pudiendo distinguir entre ellos y detectar la posibilidad de fraude[121,122].

- Identificar los **metabolitos** originados tras la ingesta de determinados alimentos que ayuden a conocer las reacciones metabólicas que sufren los compuestos originales en el organismo tras ser ingeridos. Dichos metabolitos pueden ser utilizados como **biomarcadores** nutricionales indicadores de la adhesión a determinadas dietas o la ingesta de ciertos alimentos. Estos estudios ayudan a conocer los compuestos responsables de los efectos de la alimentación en el bienestar del individuo así como su relación con la prevención o desarrollo de diferentes enfermedades[123].
- Conocer el **impacto** en el **metabolismo** endógeno que produce el consumo prolongado de determinados alimentos, nutrientes, microorganismos o compuestos bioactivos. Esta perspectiva permite identificar las alteraciones producidas en las rutas metabólicas con el objetivo de conocer los mecanismos de acción que ejercen los compuestos procedentes de la dieta en el organismo, y profundizar en el conocimiento de los efectos saludables que estos poseen[22,118].

Dado los distintos objetivos que presenta la metabolómica en esta área, existen diferentes aproximaciones metodológicas para abordarlos. A continuación, nos

[121] Farhana R. Pinu, "Metabolomics—The New Frontier in Food Safety and Quality Research," *Food Research International* 72 (June 1, 2015): 80–81, https://doi.org/10.1016/J.FOODRES.2015.03.028.

[122] Elena Cubero-Leon, Rosa Peñalver, and Alain Maquet, "Review on Metabolomics for Food Authentication," *Food Research International* 60 (June 1, 2014): 95–107, https://doi.org/10.1016/J.FOODRES.2013.11.041.

[123] Linda H Münger et al., "Biomarker of Food Intake for Assessing the Consumption of Dairy and Egg Products.," *Genes & Nutrition* 13 (2018): 26, https://doi.org/10.1186/s12263-018-0615-5.

centramos en los **estudios de intervención nutricional**, al ser los empleados en la presente memoria.

Los **ensayos de intervención nutricional** tienen como objetivo conocer el efecto en el organismo del consumo de un determinado producto (nutriente, compuesto bioactivo, extracto alimentario, alimento, suplemento dietético, dieta, etc.). Este tipo de estudios, a su vez, se clasifican principalmente en dos grupos dependiendo del tiempo durante el que se lleva a cabo la intervención. Por un lado, los ensayos de intervención aguda investigan los efectos de la ingesta durante las horas inmediatamente posteriores a la misma, hasta un máximo de 48 h. En este grupo se engloban los estudios de **absorción**, **farmacocinética**, **biodisponibilidad** o **metabolismo** de los compuestos ingeridos. Por otro lado, los estudios **intervención nutricional crónicos** estudian el efecto que causa en el organismo la ingesta de esa sustancia, alimento o dieta bajo estudio, prolongada en el tiempo, durante semanas, meses o años[22].

4.1.1. Estudios de absorción, biodisponibilidad y metabolismo

La **biodisponibilidad** se puede definir como la fracción de un ingrediente ingerido que es absorbido, se encuentra disponible y consigue alcanzar el sistema circulatorio a través del cual puede llegar a tejidos diana donde ejercer su acción biológica. Un aspecto a tener en cuenta es que no existe una relación directa entre la cantidad ingerida y la fracción biodisponible de una sustancia. Es decir, los compuestos que son ingeridos de manera mayoritaria no siempre son los que presentan mayor biodisponibilidad, de hecho algunos de los compuestos más abundantes en la dieta no necesariamente son los absorbidos en mayor cantidad o concentración.

Por su parte, el **metabolismo** hace referencia al conjunto de transformaciones y reacciones químicas que sufren los compuestos, tanto endógenos como exógenos, dentro del organismo. Este tipo de estudios pueden ser llevados a cabo de manera *in vitro* o *in vivo*, utilizando tanto modelos animales como humanos.

Los ensayos ***in vitro*** se basan en la simulación de las distintas reacciones que sufren los compuestos durante el proceso digestivo mediante el uso de digestores artificiales o modelos celulares (p.ej. células Caco-2). Las ventajas de estos modelos residen en su alta reproducibilidad, rapidez y simplicidad, dado que permiten un control estricto de todas las condiciones experimentales[124].

Por su parte, los ensayos ***in vivo*** se centran en el análisis de muestras biológicas (sangre, orina, heces, etc.) recogidas en humanos o distintos modelos animales tras la ingesta del alimento o dieta bajo estudio (**Figura 29**). El uso de **modelos animales** posibilita la obtención de diferentes tejidos (hígado, riñón, intestino, etc.) donde los metabolitos derivados de los alimentos se acumulan o pueden ejercer su acción biológica, además de una mayor versatilidad en los diseños experimentales, como por ejemplo la posibilidad de realizar **estudios de perfusión intestinal**. En concreto, estos estudios de perfusión consisten en la creación de un pequeño compartimiento intestinal aislado con la ayuda de jeringas y válvulas con la finalidad de introducir la matriz que se quiere estudiar directamente en el intestino, así como recoger muestras de contenido intestinal a diferentes tiempos para el estudio de la absorción y

[124] Juana M. Carbonell-Capella et al., “Analytical Methods for Determining Bioavailability and Bioaccessibility of Bioactive Compounds from Fruits and Vegetables: A Review,” *Comprehensive Reviews in Food Science and Food Safety* 13, no. 2 (March 1, 2014): 155–71, https://doi.org/10.1111/1541-4337.12049.

metabolismo de los compuestos de interés en la región intestinal[125]. No obstante, los **modelos** en **humanos** son considerados particularmente útiles, dado que son los que proporcionan los resultados más precisos de manera más acorde a la realidad[126].

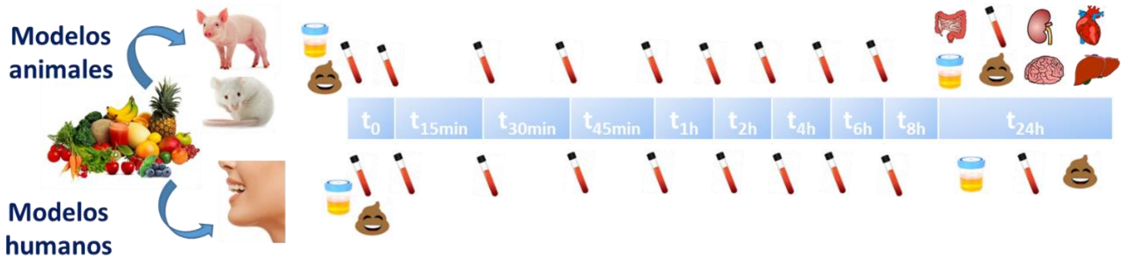


Figura 29. Ejemplos de diseño de estudios de biodisponibilidad y metabolismo en modelos animales y humanos.

4.1.2. Estudios de intervención nutricional longitudinales

Los estudios de intervención nutricional longitudinales se basan en la observación y detección de los cambios metabólicos debidos a la ingesta prolongada, durante semanas o meses, de una determinada dosis diaria de un alimento, compuesto, extracto o dieta. Este tipo de ensayos pueden ser aplicados con diferentes objetivos, como estudiar la influencia que posee una determinada dieta en los factores de riesgo de distintas patologías (p.ej.: enfermedades cardiovasculares, diabetes, cáncer, etc.)[127];

[125] Isabel Lozoya-Agullo et al., "In Situ Perfusion Model in Rat Colon for Drug Absorption Studies: Comparison with Small Intestine and Caco-2 Cell Model," *Journal of Pharmaceutical Sciences* 104, no. 9 (2015): 3136–45, https://doi.org/10.1002/jps.24447.

[126] Antonio Cilla, Reyes Barberá, and Amparo Alegría, eds., "Overview of In Vivo and In Vitro Methods for Assessing Bioavailability of Bioactive Food Compounds," in *Frontiers in Bioactive Compounds* (BENTHAM SCIENCE PUBLISHERS, 2017), 54–98, https://doi.org/10.2174/9781681084299117020007.

[127] Jordi Salas-Salvadó et al., "Reduction in the Incidence of Type 2 Diabetes with the Mediterranean Diet: Results of the PREDIMED-Reus Nutrition Intervention Randomized Trial.," *Diabetes Care* 34, no. 1 (January 2011): 14–19, https://doi.org/10.2337/dc10-1288.

o determinar el impacto del consumo de determinados alimentos o la ingesta de compuestos bioactivos o nutrientes en las rutas metabólicas[128,129].

Este tipo de estudios pueden realizarse tanto con modelos animales como en humanos, donde se recolectan las muestras biológicas al principio y al final del ensayo clínico para la identificación de los metabolitos endógenos alterados tras la intervención nutricional (**Figura 30**). Los participantes deben ser seleccionados a través de unos criterios de inclusión y exclusión previamente establecidos, dado que numerosos factores pueden afectar a los resultados del estudio, como por ejemplo la edad, el género, el estilo de vida, el estado de salud, el estado metabólico, etc. Para verificar los resultados obtenidos y garantizar su calidad, se requiere el empleo de estudios aleatorios de doble ciego mediante la utilización de grupos control que consuman una dieta placebo, donde los participantes no conozcan su pertenencia al grupo experimental asignado.

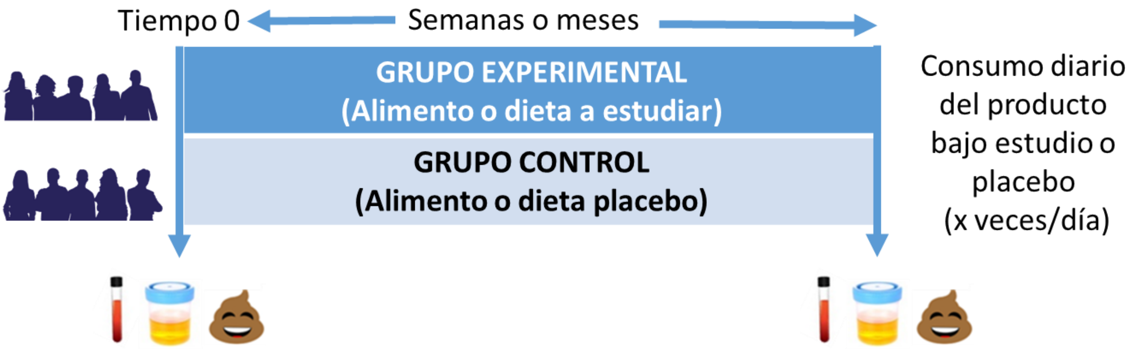

Tiempo 0
Semanas o meses
GRUPO EXPERIMENTAL
(Alimento o dieta a estudiar)
GRUPO CONTROL
(Alimento o dieta placebo)
Consumo diario del producto bajo estudio o placebo (x veces/día)

Figura 30. Ejemplo de diseño de un ensayo de intervención nutricional longitudinal en humanos.

[128] Hong Zheng et al., "Metabolomics to Explore Impact of Dairy Intake," *Nutrients* 7, no. 6 (2015): 4875–96, https://doi.org/10.3390/nu7064875.

[129] Olha Khymenets et al., "Metabolic Fingerprint after Acute and under Sustained Consumption of a Functional Beverage Based on Grape Skin Extract in Healthy Human Subjects," *Food Funct.* 6, no. 4 (2015): 1288–98, https://doi.org/10.1039/C4FO00684D.

4.2. Aplicación de la metabolómica en el estudio de enfermedades

La capacidad de la metabolómica para la identificación de cientos de metabolitos presentes en muestras biológicas, los cuales están estrechamente relacionados con la respuesta fenotípica y los procesos bioquímicos de un organismo, permite detectar cambios en el metabolismo celular originados por una patología o cualquier otro estímulo externo. Por lo tanto, el estudio de patologías y procesos fisiológicos es uno de los campos de mayor aplicación de las estrategias metabolómicas en la actualidad. Entre las patologías más estudiadas mediante aproximaciones metabolómicas se encuentran enfermedades como el cáncer[130], la diabetes[131], enfermedades cardiovasculares[132], afecciones respiratorias[133] o enfermedades neurodegenerativas como el Alzheimer[134].

Entre las diferentes posibilidades de la metabolómica en el estudio de patologías, destaca su gran potencial para mejorar **diagnósticos** y **pronósticos** de las mismas, su alta capacidad para comprender la **patogénesis de enfermedades**, así como para el desarrollo y evaluación de **fármacos, terapias y tratamientos** para combatirlas[8]. Otro de los grandes retos de la metabolómica en el ámbito de la salud se centra en su contribución en el desarrollo de la medicina personalizada desde una perspectiva

[130] Leonor Puchades-Carrasco and Antonio Pineda-Lucena, "Metabolomics Applications in Precision Medicine: An Oncological Perspective.," *Current Topics in Medicinal Chemistry* 17, no. 24 (2017): 2740–51, https://doi.org/10.2174/1568026617666170707120034.

[131] Rigoberto Pallares-Méndez et al., "Metabolomics in Diabetes, a Review," *Annals of Medicine* 48, no. 1–2 (January 8, 2016): 89–102, https://doi.org/10.3109/07853890.2015.1137630.

[132] Robert W. McGarrah et al., "Cardiovascular Metabolomics," *Circulation Research* 122, no. 9 (April 27, 2018): 1238–58, https://doi.org/10.1161/CIRCRESAHA.117.311002.

[133] Darryl J Adamko, Brian D Sykes, and Brian H Rowe, "The Metabolomics of Asthma: Novel Diagnostic Potential.," *Chest* 141, no. 5 (May 2012): 1295–1302, https://doi.org/10.1378/chest.11-2028.

[134] Alejandro Botas et al., "Metabolomics of Neurodegenerative Diseases," in *International Review of Neurobiology*, vol. 122, 2015, 53–80, https://doi.org/10.1016/bs.irn.2015.05.006.

holística con el resto de ciencias ómicas. Este concepto de **medicina personalizada** se basa en la adaptación del tratamiento médico en función de las características individuales de cada paciente permitiendo elegir el principio activo y la dosis más adecuada y efectiva[135].

Para llevar a cabo las distintas aplicaciones se requiere la utilización de indicadores de diagnóstico de las enfermedades, denominados biomarcadores, que también pueden ser utilizados para evaluar la aplicabilidad de las intervenciones terapéuticas. Estos **biomarcadores** son definidos como características medibles y evaluables objetivamente como indicadores de procesos biológicos normales, procesos patogénicos o respuestas farmacológicas a una intervención terapéutica[136].

Un biomarcador ideal debe cumplir una serie de requisitos, como por ejemplo ser específico ante una enfermedad particular, predictivo, estable, presente en muestras biológicas no invasivas (sangre, orina, etc.) y fácilmente cuantificable mediante métodos rápidos, simples y económicos. Muchos de los biomarcadores metabólicos son usados para el diagnóstico de enfermedades de manera aislada o en combinación con otros datos o análisis. Un ejemplo de este tipo de biomarcadores se basa en el perfil de aminoácidos obtenido por estrategias metabolómicas con el objetivo de predecir el riesgo de padecer diabetes mellitus tipo 2. En este sentido, el alto riesgo de sufrir esta enfermedad está asociado con altos niveles de isoleucina, leucina, valina, fenilalanina, tirosina y ácido aminoadípico en suero, los cuales pueden ser detectados

[135] Nadia Koen, Ilse Du Preez, and Du Toit Loots, “Metabolomics and Personalized Medicine,” *Advances in Protein Chemistry and Structural Biology* 102 (January 1, 2016): 53–78, https://doi.org/10.1016/BS.APCSB.2015.09.003.

[136] Drupad K. Trivedi, Katherine A. Hollywood, and Royston Goodacre, “Metabolomics for the Masses: The Future of Metabolomics in a Personalized World,” *New Horizons in Translational Medicine*, 2017, https://doi.org/10.1016/j.nhtm.2017.06.001.

incluso 15 años antes de padecer la enfermedad[137,138]. Otro ejemplo, son los estudios que han sido capaces de detectar varios metabolitos, 2-hidroxibutarato, sarcosina, colina, succinato, lactato, fumarato o glucosa, relacionados con distintos tipos de cáncer (leucemia, cáncer de riñón, mama, próstata o cerebro)[139].

Uno de los principales retos de la metabolómica en relación con el descubrimiento de biomarcadores es el diseño de **herramientas de diagnóstico** para su aplicación clínica de manera rutinaria. Las características ideales de estos dispositivos de diagnóstico son su capacidad de miniaturización, comercialización así como su sencillez y reproducibilidad. Un ejemplo de estas herramientas son los dispositivos portátiles para la medida de glucosa en sangre en personas diabéticas[17].

A pesar de que se han realizado numerosas investigaciones con la finalidad de identificar biomarcadores en distintas enfermedades de difícil diagnóstico, la principal limitación se encuentra en la **validación** de los mismos. Para llegar a encontrar una aplicación clínica a los biomarcadores identificados se necesita que estos estén correctamente validados por medio de un nuevo conjunto significativo de muestras. De esta forma, la validación de biomarcadores así como el diseño de aplicaciones clínicas son dos de los principales retos actuales de la metabolómica[17,136,140].

[137] Thomas J. Wang et al., “2-Aminoadipic Acid Is a Biomarker for Diabetes Risk,” *Journal of Clinical Investigation* 123, no. 10 (October 1, 2013): 4309–17, https://doi.org/10.1172/JCI64801.

[138] Thomas J Wang et al., “Metabolite Profiles and the Risk of Developing Diabetes,” *Nature Medicine* 17, no. 4 (April 20, 2011): 448–53, https://doi.org/10.1038/nm.2307.

[139] D.S. Wishart et al., “Cancer Metabolomics and the Human Metabolome Database,” *Metabolites* 6, no. 1 (2016), https://doi.org/10.3390/metabo6010010.

[140] Calena R Marchand et al., “A Framework for Development of Useful Metabolomic Biomarkers and Their Effective Knowledge Translation.,” *Metabolites* 8, no. 4 (September 30, 2018), https://doi.org/10.3390/metabo8040059.

En cuanto al diseño del estudio metabolómico en este ámbito de aplicación, se distinguen principalmente dos tipos de estudios: estudios de **cohorte transversal** o **longitudinal** (**Figura 31**).

Los estudios de **cohorte transversal** se caracterizan por llevar a cabo la recogida de muestras biológicas en un único momento temporal. El objetivo de este tipo de estudio es la asignación de un perfil metabólico a un fenotipo determinado o la identificación de diferencias metabólicas entre diferentes fenotipos (p.ej.: identificar diferencias entre individuos sanos y enfermos, entre varias enfermedades, o en función del género, edad u otras características).

Por su parte, en los **estudios longitudinales** se recogen repetidamente muestras biológicas de los mismos individuos durante un período de tiempo prolongado. De esta forma, este tipo de estudios permiten conocer la evolución de una enfermedad a lo largo del tiempo, o la respuesta a un determinado fármaco o tratamiento, entre otras aplicaciones[117].

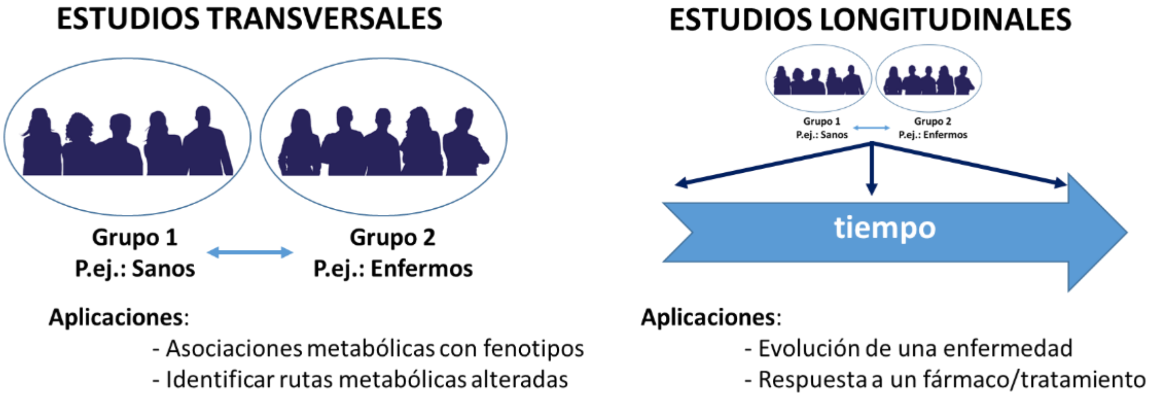


Figura 31. Estudios transversales y longitudinales en metabolómica.

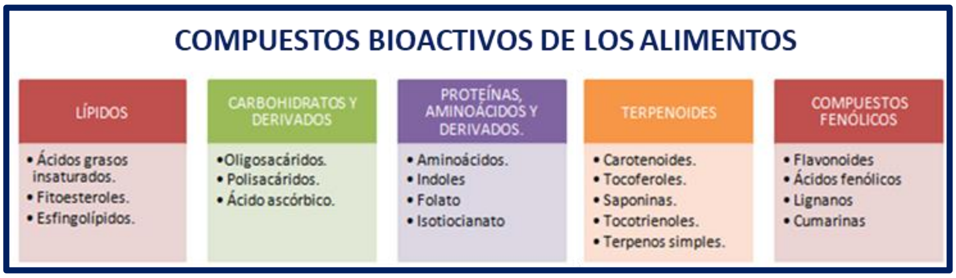
BLOQUE A

Desarrollo y aplicación de estrategias metabolómicas para el estudio de compuestos bioactivos

1. Compuestos bioactivos: alimentación funcional y nutracéutica

Los **compuestos bioactivos** se definen como aquel conjunto de sustancias naturales que ejercen una actividad biológica que conduce a alteraciones metabólicas asociadas a efectos beneficiosos sobre la salud humana, como pueden ser la mejora de ciertas funciones fisiológicas o la reducción del riesgo de padecer diversas enfermedades.

Estos compuestos bioactivos se encuentran generalmente en productos de **origen vegetal**, entre otras fuentes, siendo una amplia variedad de sustancias con diferentes estructuras químicas y actividades biológicas. Algunos ejemplos de estos compuestos que presentan beneficios en la salud humana son los minerales (Hierro, Zinc, Calcio, Selenio), las vitaminas (C, E, K, vitaminas del grupo B) y otros compuestos no nutrientes presentes principalmente en plantas denominados fitoquímicos, como los carotenoides, los compuestos fenólicos, los glucosinatos o los fitoesteroles, entre otros (**Figura 32**)[141,142].



COMPUESTOS BIOACTIVOS DE LOS ALIMENTOS

LÍPIDOS	CARBOHIDRATOS Y DERIVADOS	PROTEÍNAS, AMINOÁCIDOS Y DERIVADOS.	TERPENOIDES	COMPUESTOS FENÓLICOS
• Ácidos grasos insaturados. • Fitoesteroles. • Esfingolípidos.	•Oligosacáridos. • Polisacáridos. • Ácido ascórbico.	• Aminoácidos. • Indoles • Folato • Isotiocianato	• Carotenoides. • Tocoferoles. • Saponinas. • Tocotrienoles. • Terpenos simples.	• Flavonoides • Ácidos fenólicos • Lignanos • Cumarinas

Figura 32. Principales familias de compuestos bioactivos.

[141] Saleh Hosseinzadeh et al., "The Application of Medicinal Plants in Traditional and Modern Medicine: A Review of Thymus Vulgaris," *International Journal of Clinical Medicine* 06, no. 09 (2015): 635–42, https://doi.org/10.4236/ijcm.2015.69084.

[142] Singh R., "Medicinal Plants: A Review," *Journal of Plant Sciences* 3, no. 1 (2015): 50–55, https://doi.org/10.11648/j.jps.s.2015030101.18.

El avance en el conocimiento de los compuestos bioactivos está permitiendo su utilización en el desarrollo de nuevos productos alimentarios, como alimentos funcionales y nutraceúticos, con el objetivo de que su consumo aporte efectos beneficiosos para la salud[143]. Un **alimento funcional** es definido como aquel alimento que aporta algún efecto beneficioso en el organismo más allá de los propios efectos nutricionales que todos poseen ya que contienen compuestos bioactivos. Por su parte, un **nutracéutico** es considerado aquel compuesto bioactivo aislado o extracto en formato farmacéutico (p.ej.: cápsula, pastillas, etc.) que aporta beneficios en la prevención o el tratamiento de enfermedades[144,145].

Entre las diferentes familias de compuestos bioactivos, esta memoria se ha centrado en los **compuestos fenólicos.** Estos compuestos han despertado un gran interés en los últimos años por parte de la comunidad científica para el desarrollo de alimentos funcionales y/o nutracéuticos. Este hecho es debido a su amplia diversidad estructural, siendo uno de los grupos de especies fitoquímicas más numeroso, y a las diferentes actividades biológicas que presentan. Estos compuestos conforman uno de los grupos de sustancias naturales más numeroso y ampliamente distribuido en el reino vegetal, con más de 8000 estructuras fenólicas conocidas actualmente. La mayoría de los compuestos fenólicos contienen un esqueleto básico común formado por al menos un anillo aromático con uno o más sustituyentes hidroxilo. Debido a la gran variedad

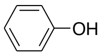
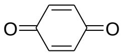
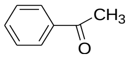
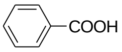
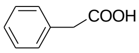
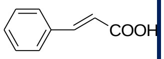
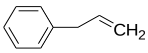
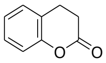
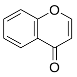
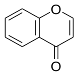
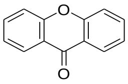
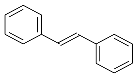
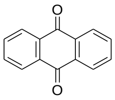
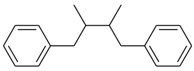
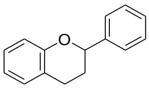
[143] Rosa Perez-Gregorio and Jesus Simal-Gandara, “A Critical Review of Bioactive Food Components, and of Their Functional Mechanisms, Biological Effects and Health Outcomes,” *Current Pharmaceutical Design* 23, no. 19 (July 20, 2017): 2731–41, https://doi.org/10.2174/1381612823666170317122913.

[144] S. El Sohaimy, “Functional Foods and Nutraceuticals-Modern Approach to Food Science,” *World Applied Sciences Journal*, 2012, https://doi.org/10.5829/idosi.wasj.2012.20.05.66119.

[145] Khalid Gul, A. K. Singh, and Rifat Jabeen, “Nutraceuticals and Functional Foods: The Foods for the Future World,” *Critical Reviews in Food Science and Nutrition* 56, no. 16 (December 9, 2016): 2617–27, https://doi.org/10.1080/10408398.2014.903384.

estructural que presentan, se pueden clasificar en diferentes familias, las cuales se reflejan en la **Tabla 5**[146-148].

Tabla 5. Principales familias de compuestos fenólicos y su estructura básica.

FENOLES SIMPLES	Fenoles		Benzo- quinonas		Acetofenonas	
	Ácidos benzoicos		Ácidos fenilacéticos		Ácidos cinámicos	
	Fenil- propenos		Cumarinas Isocumarinas		Cromonas	
	Nafto- quinonas					
POLIFENOLES	Ligninos	Polímero aromático altamente entrecruzado	Xantonas		Estilbenos	
	Antra- quinonas		Lignanos Neolignanos		Taninos hidrolizables	Polímero heterogéneo formado por ácidos fenólicos y azúcares simples
	Flavonoides	**Subfamilias:** Flavandioles, Flavanoles, Dihidrochalconas, Proantocianidinas/Taninos condensados, antocianidinas, Isoflavonoides, dihidroflavonoles, biflavonoles,flavonoles, auronas, flavonas, chalconas, flavanonas.				

Los compuestos fenólicos se encuentran en una amplia variedad de alimentos, como por ejemplo en frutas, verduras, cereales e incluso en bebidas como el té, el café o el vino. A continuación, se describen brevemente las matrices vegetales ricas en compuestos fenólicos que han sido estudiadas en la presente memoria (**Figura 33**) mediante el uso de herramientas metabolómicas:

[146] Daniele Del Rio et al., "Dietary (Poly)Phenolics in Human Health: Structures, Bioavailability, and Evidence of Protective Effects Against Chronic Diseases," *Antioxidants & Redox Signaling* 18, no. 14 (2013): 1818–92, https://doi.org/10.1089/ars.2012.4581.

[147] Rong Tsao, "Chemistry and Biochemistry of Dietary Polyphenols," *Nutrients* (Molecular Diversity Preservation International, December 10, 2010), https://doi.org/10.3390/nu2121231.

[148] Xiuzhen Han, Tao Shen, and Hongxiang Lou, "Dietary Polyphenols and Their Biological Significance," *International Journal of Molecular Sciences* 8, no. 9 (2007): 950–88, https://doi.org/10.3390/i8090950.

- **Romero** (*Rosmarinus officinalis* L.): Esta especie vegetal despierta gran interés dentro de la comunidad científica gracias a la gran cantidad de compuestos bioactivos presentes en su composición, como por ejemplo, diterpenos fenólicos tipo abietano, triterpenos, ácidos fenólicos o flavonoides. Esta variedad de compuestos hace que el romero presente de forma natural un gran número de actividades farmacológicas, entra las que destaca su actividad antiinflamatoria, antitrombótica, diurética, antidiabética, antidepresiva, analgésica, hepatoprotectora, antioxidante y anticancerígena[149,150].
- **Mango** (*Mangifera indica* L.): Esta fruta tropical es una excelente fuente de compuestos bioactivos, entre los que destacan los carotenoides y compuestos fenólicos. Entre los efectos de estos últimos destaca su capacidad antimutagénica, antiinflamatoria, antioxidante, antidiabética e inmunomoduladora[151,152].
- **Ajo** (*Allium sativum*): Esta planta bulbosa de la familia de las liliáceas ha sido desde la antigüedad de gran interés debido a la atribución de numerosas propiedades beneficiosas, relacionadas con su composición rica en ácidos fenólicos, dipéptidos, ácidos grasos, flavonoides y compuestos organosulfurados. Entre sus propiedades bioactivas destaca su capacidad

[149] Naisheng Bai et al., “Flavonoids and Phenolic Compounds from Rosmarinus Officinalis,” *Journal of Agricultural and Food Chemistry* 58, no. 9 (2010): 5363–67, https://doi.org/10.1021/jf100332w.

[150] Isabel Borrás-Linares et al., “Rosmarinus Officinalis Leaves as a Natural Source of Bioactive Compounds.,” *International Journal of Molecular Sciences* 15, no. 11 (2014): 20585–606, https://doi.org/10.3390/ijms151120585.

[151] H. Palafox-Carlos, E.M. Yahia, and G.A. González-Aguilar, “Identification and Quantification of Major Phenolic Compounds from Mango (Mangifera Indica, Cv. Ataulfo) Fruit by HPLC–DAD–MS/MS-ESI and Their Individual Contribution to the Antioxidant Activity during Ripening,” *Food Chemistry* 135, no. 1 (November 1, 2012): 105–11, https://doi.org/10.1016/J.FOODCHEM.2012.04.103.

[152] Masibo Martin and Qian He, “Mango Bioactive Compounds and Related Nutraceutical Properties-A Review,” *Food Reviews International*, September 2009, https://doi.org/10.1080/87559120903153524.

antioxidante, anticancerígena, antibacteriana, antienvejecimiento, antimicrobiana, antiplaquetaria y antidiabética, así como su capacidad inmunomoduladora[153,154].

Figura 33. *Rosmarinus officinalis*, *Magnifera indica* y *Allium sativum*.

Las **estrategias metabolómicas** han permitido el estudio de las relaciones existentes entre la ingesta de compuestos fenólicos y sus efectos beneficiosos sobre la salud del consumidor, describiéndose en bibliografía numerosas aplicaciones relacionadas con estos compuestos. Dichas investigaciones abarcan desde estudios metabolómicos de compuestos fenólicos desarrollados con la finalidad de conocer el perfil metabólico presente en matrices alimenticias así como estudios de absorción, farmacocinética, biodisponibilidad o metabolización de dichos compuestos[155]. Además, también se han aplicado ensayos de intervención nutricional longitudinales para la evaluación de los efectos metabólicos causados por el consumo prolongado de estas sustancias. Por

[153] Harunobu Amagase et al., “Intake of Garlic and Its Bioactive Components,” *The Journal of Nutrition* 131, no. 3 (2001): 955–62, https://doi.org/https://doi.org/10.1093/jn/131.3.955S.

[154] Aneta Kopec et al., “Healthy Properties of Garlic,” *Current Nutrition & Food Science* 9, no. 4 (2013): 59–64, https://doi.org/10.2174/1573401311309010010.

[155] Claudine Manach et al., “The Complex Links between Dietary Phytochemicals and Human Health Deciphered by Metabolomics,” *Molecular Nutrition & Food Research* 53, no. 10 (2009): 1303–15, https://doi.org/10.1002/mnfr.200800516.

ejemplo, se han estudiado los efectos metabólicos que causa el consumo prolongado de resveratrol en pacientes con síndrome metabólico[156], o de una bebida funcional basada en un extracto rico en compuestos fenólicos obtenido de la piel de uva en sujetos sanos[129].

El estudio de la **biodisponibilidad** de los compuestos fenólicos es una tarea ardua debido a la existencia de distintos factores que pueden afectar su absorción en el organismo. Uno de los factores que afectan a la biodisponibilidad de un compuesto es su estructura química. La gran mayoría de los compuestos fenólicos existen en la naturaleza en forma glicosilada, las cuales no pueden ser absorbidas directamente a través de las células del epitelio y deben ser transformadas en otras estructuras más sencillas mediante las enzimas presentes en el intestino delgado o en el colon. Así, cuando los compuestos fenólicos llegan al intestino delgado, se producen reacciones de hidrólisis enzimática que transforman las estructuras glicosiladas en las agliconas correspondientes.

Tras esta primera etapa, las agliconas, antes de incorporarse al torrente sanguíneo, pueden sufrir otro proceso de transformación a través del metabolismo de fase I y de fase II, en los cuales se producen reacciones de oxidación-reducción y de conjugación (sulfatación, metilación y glucuronización), respectivamente.

Aquellos compuestos que no son absorbidos en el intestino delgado pueden ser posteriormente transformados por la microflora del colon y ser absorbidos en esta parte del tracto gastrointestinal, o bien ser excretados en las heces. En el primer caso,

[156] Anne Sofie Korsholm et al., “Comprehensive Metabolomic Analysis in Blood, Urine, Fat, and Muscle in Men with Metabolic Syndrome: A Randomized, Placebo-Controlled Clinical Trial on the Effects of Resveratrol after Four Months’ Treatment,” *International Journal of Molecular Sciences* 18, no. 3 (2017), https://doi.org/10.3390/ijms18030554.

la microflora del colon se encarga de hidrolizar ciertos compuestos fenólicos, como por ejemplo los que se encuentran formando esteres, produciendo las agliconas correspondientes e incluso pudiendo llegar a degradar estas formando estructuras más simples como los ácidos fenólicos.

Por su parte, los metabolitos que pasan al torrente sanguíneo pueden ser transportados hasta el hígado, lugar donde pueden sufrir más transformaciones correspondientes al metabolismo de fase I y II. Estos compuestos pueden tomar diferentes rutas: ser transportados a los riñones y ser excretados en la orina, ser enviados de nuevo al intestino a través del líquido biliar donde pueden ser reabsorbidos; o alcanzar células y tejidos a través del torrente sanguíneo. En la **Figura 34** se muestra un esquema del metabolismo de los compuestos fenólicos que se produce a partir del intestino delgado[157-160].

[157] G R Velderrain-Rodríguez et al., "Phenolic Compounds: Their Journey after Intake.," *Food & Function* 5, no. 2 (February 2014): 189–97, https://doi.org/10.1039/c3fo60361j.

[158] Urszula Lewandowska et al., "Overview of Metabolism and Bioavailability Enhancement of Polyphenols," *Journal of Agricultural and Food Chemistry* 61, no. 50 (2013): 12183–99, https://doi.org/10.1021/jf404439b.

[159] Alan Crozier, Daniele Del Rio, and Michael N. Clifford, "Bioavailability of Dietary Flavonoids and Phenolic Compounds," *Molecular Aspects of Medicine* 31, no. 6 (2010): 446–67, https://doi.org/10.1016/j.mam.2010.09.007.

[160] Laura Marín et al., "Bioavailability of Dietary Polyphenols and Gut Microbiota Metabolism : Antimicrobial Properties," *Biomed Research International* 2015 (2015) https://doi.org/ 10.1155/2015/905215.

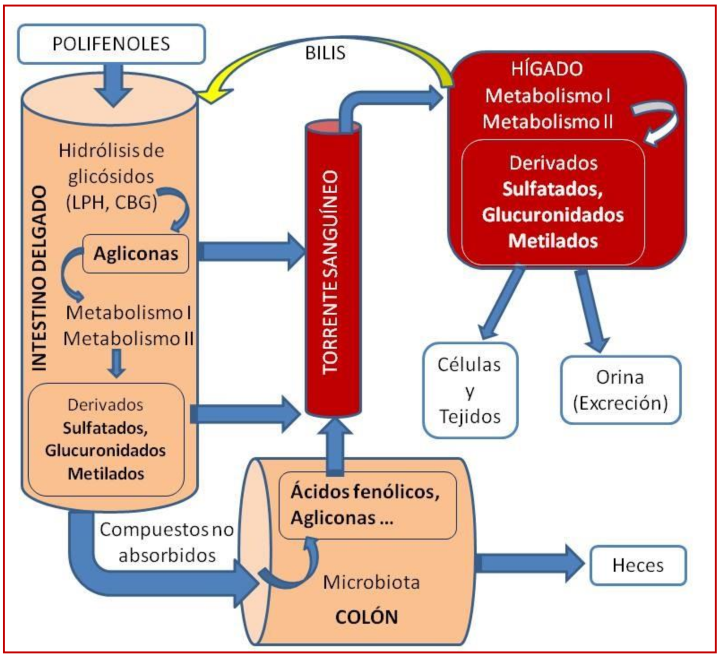
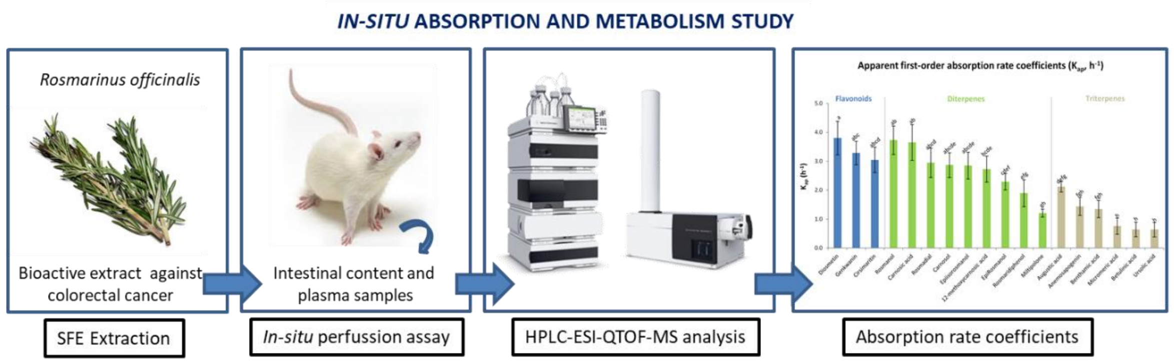


Figura 34. Principales reacciones de metabolización de los compuestos fenólicos.

Capítulo 1

Estudio de perfusión *in situ* de un extracto de romero rico en compuestos compuestos bioactivos



IN-SITU **ABSORPTION AND METABOLISM STUDY**

Rosmarinus officinalis

Bioactive extract against colorectal cancer

SFE Extraction

Intestinal content and plasma samples

In-situ perfussion assay

HPLC-ESI-QTOF-MS analysis

Absorption rate coefficients

Journal of Functional Foods. 2017, 33, 202-210

Álvaro Fernández-Ochoa, Isabel Borrás-Linares, Almudena Pérez-Sánchez, Enrique Barrajón-Catalán, Isabel González-Álvarez, David Arráez-Román, Vicente Micol, Antonio Segura-Carretero

DOI:10.1016/j.jff.2017.03.046

Phenolic compounds in rosemary as potential source of bioactive compounds against colorectal cancer: *in situ* absorption and metabolism study

ABSTRACT

Phenolic compounds in rosemary have shown antiproliferative/cytotoxic activity against colorectal cancer cells. The aim of this work was to study in depth the absorption and metabolism of the compounds present in a rosemary extract. An *in-situ* perfusion assay was performed in mice, for which samples of gastrointestinal liquid taken at different times and plasma obtained at the end of the experiment were analysed by HPLC-ESI-QTOF-MS. The absorption-rate coefficients showed that flavonoids and diterpenes were highly absorbed compared to triterpenes. Several diterpenes and their metabolites were also bioavailable in plasma, highlighting the higher concentrations of glucuronide metabolites compared to non-metabolised phenolic compounds. The antiproliferative/cytotoxic properties could be attributed to the absorbed diterpenes and metabolites, which could reach the colon through bloodstream. On the other hand, compounds poorly absorbed, such as triterpenes and some diterpenes, could exhibit their bioactivity in the large intestine by a mechanism of direct interaction with the microbiota.

Keywords. *Rosmarinus officinalis*, HPLC-ESI-QTOF-MS, phenolic compounds, absorption, metabolism, *in situ* assay.

1. INTRODUCTION

Rosemary (*Rosmarinus officinalis* L., Lamiaceae) is a shrubby plant which has a high content in bioactive compounds with different pharmacological activities, especially antimicrobial (Bozin, Mimica-Dukic, Samojlik, & Jovin, 2007; Jiang et al., 2011), antithrombotic (Naemura, Ura, Yamashita, Arai, & Yamamoto, 2008), diuretic (Haloui, Louedec, Michel, & Lyoussi, 2000), anti-inflammatory (Altinier et al., 2007), hepatoprotective (Sotelo-Félix et al., 2002), anti-oxidant (Pérez-Fons, Garzón, & Micol, 2010), anticancer (Srancikova, Horvathova, & Kozics, 2013), anti-diabetic (Sedighi, Zhao, Yerke, & Sang, 2015) and anti-obesity (Sedighi et al., 2015). Numerous studies have highlighted that the majority of these biological activities are correlated with the phenolic composition (Borrás Linares et al., 2011). Primarily, antioxidant and anticancer activities have been linked to the presence of abietan-type diterpenes, such as carnosic acid or carnosol (Birtić, Dussort, Pierre, Bily, & Roller, 2015; Petiwala & Johnson, 2015), as well as triterpenes, phenolic acids, and flavonoids (Bai et al., 2010; Borrás-Linares et al., 2014).

In recent years, several works have demonstrated the antiproliferative activity of rosemary phenolics in different cancer-cell models, as in prostate cancer (Petiwala, Puthenveetil, & Johnson, 2013), neuroblastome (Tsai, Lin, Lin, & Chen, 2011), colorectal cancer (Borrás-Linares et al., 2015; Valdés et al., 2013) or ovarian cancer (Tai, Cheung, Wu, & Hasman, 2012). Colorectal cancer was responsible for 1.4 million cases and 693,000 deaths in 2012 (Torre et al., 2015), bringing to light that this disease should be contained. In this sense, in recent decades, many *in vitro* and *in vivo* assays have shown diverse protective effects of rosemary regarding chemoprevention and tumour reduction (Ngo, Williams, & Head, 2011; Valdés et al., 2013). Therefore, there is an urgent need to extend our knowledge concerning metabolism, absorption, bioavailability and the mechanism of action of these phenolic compounds due to their bioactive properties. To date, numerous studies concerning the gut absorption of phenolic compounds have been performed although most have used *in vitro* models, such as intestinal absorption through Caco-2 cell simulations. Therefore, several of these studies have focused only on specific compounds such as apigenin, resveratrol, emodin, crissofal (Teng et al., 2012) or resveratrol oligomers (Willenberg et al., 2015). Similar studies have been reported on vegetable matrices with high phenolic contents, such as tea (Tenore, Campiglia, Giannetti, & Novellino, 2015), cocoa (Kosińska & Andlauer, 2012), and artichoke heads (D'Antuono, Garbetta, Linsalata, Minervini, & Cardinali, 2015).

On the other hand, a few *in situ* and *in vivo* studies have also been performed through intestinal perfusion, bioavailability, and pharmacokinetic assays using animal or human models (Clifford, van der Hooft, & Crozier, 2013; Teng et al., 2012). Regarding phenolic compounds from *Rosmarinus officinalis*, a majority of these studies have been focused on the absorption, distribution, and elimination of carnosic acid, which is considered the most bioactive compound in rosemary (Doolaege, Raes, De Vos, Verhe, & De Smet, 2011; Pelillo et al., 2004; Yan et al., 2009). Only one *in situ* study is available in the literature on bioavailability and metabolism of diterpenes from a rosemary extract (Romo Vaquero et al., 2013). Nevertheless, although the number of these types of studies has recently increased, there is a lack of information on absorption and bioavailability of complex compounds for a thorough understanding of the action mechanism of these molecules.

Therefore, as a continuation of our previous work regarding the bioactivity of rosemary compounds (Borrás-Linares et al., 2015) against colorectal cancer, the aim of this work was to examine absorption, bioavailability, and metabolism at *in situ* level, through a small-intestine perfusion assay, of the phenolic compounds present in a rosemary-leaf extract with proven antiproliferative activity. The results could help to clarify the absorption and metabolism of rosemary bioactive compounds, which in turn would contribute to a fuller understanding of the mechanisms of action of these compounds against colorectal cancer.

2. MATERIAL AND METHODS

2.1. Chemicals

All chemical were of analytical reagent-grade purity and used as received. Formic acid and LC-MS grade acetonitrile were purchased from Fluka, Sigma-Aldrich (Steinheim, Germany) and Fisher Scientific (Madrid, Spain), respectively. Water was purified by a Milli-Q system from Millipore (Bedford, MA, USA). Standard compounds luteolin, diosmetin, genkwanin, and ursolic acid were purchased from Extrasynthese (Genay, France), and apigenin, carnosol and carnosic acid were obtained from Fluka, Sigma-Aldrich (Steinheim, Germany). Standard solutions were prepared in dimethyl sulfoxide (DMSO) and methanol (Fisher Scientific, Madrid, Spain) and stored at -20°C until used. For plasma treatment, ethanol and methanol (Fisher Scientific) were used.

2.2. Rosemary-leaf extract

The extract used for the *in situ* assay, was taken from dried rosemary leaves from Herboristería Murciana (Murcia, Spain) as in previous works (Borrás Linares et al., 2011; Herrero, Plaza, Cifuentes, & Ibáñez, 2010). In brief, a supercritical fluid extraction (SFE) system (Suprex Prep Master, Supres Corporation, Pittsburg, PA, USA) was used to obtain the extract under the follow conditions: a flow of neat CO_2 of 60 g/min at 150 bar and 40ºC using 6.6% of ethanol as the modifier. An extraction time of 5 h was applied to the rosemary sample in order to ensure high recovery efficiency. After the extraction, the remaining solvent was evaporated using a Rotavapor R-210 (Buchi Labortechnik AG, Flawil, Switzerland) and the solid rosemary extract was stored at -20ºC in darkness. This extract had been previously characterized (Borrás-Linares et al., 2015; Borrás Linares et al., 2011) (**Table S1**). Mainly, the extract composition was made up of triterpenes, diterpenes, and flavonoids with abundance values of 49.5, 44.3 and 2.1%, respectively, with carnosic acid being the main compound in the extract.

2.3. Perfusion assay

The perfusion assay was developed in whole small intestine and was performed by *in situ* closed-loop perfusion method based on Doluisio's Technique (Doluisio, Billups, Dittert, Sugita, & Swintosky, 1969). Briefly, male Wistar rats (body weight, 250–300 g) which were food deprived for 4 h and provided free access to water, were used for these studies. Rats were anaesthetized using a mixture of pentobarbital (40 mg/kg) and butorphanol (0.5 mg/kg). A midline abdominal incision was made, the intestinal segment was manipulated in order to minimize any intestinal blood-supply disturbances and the bile duct was tied off in order to avoid drug enterohepatic circulation and the presence of bile salts in the lumen. All intestinal contents were washed and flushed with a physiologic isotonic saline solution (pH 6.9) with 1% Sörensen phosphate buffer (v/v) at 37ºC. Once the system was set up, rosemary extract with a concentration of 500 mg/mL in Sörensen phosphate buffer (pH 7.0) supplemented with 1% v/v of dimethyl sulphoxide (DMSO) was injected into the intestinal segment. Afterwards, samples were collected with the aid of the syringes and stopcock valves every 5 min up to a period of 30 min. Blood samples were taken by cardiac puncture in heparinized tubes and immediately maintained over orbital stirring until plasma separation. At the end of the experiments, the animals were euthanized. Rats were used in accordance with 2010/63/EU European and RD-53/2013 Spanish directives regarding the protection of animals used for scientific experimentation. The Ethics Committee for Animal Experimentation of Miguel Hernández University approved the experimental protocols (IBM.VMM.007-11 code).

2.4. Sample treatments

Firstly, gastrointestinal liquid and plasma samples taken over the course of the assay were thawed on ice and spiked with 5 ppm of luteolin as an internal standard assisted with vortex mixing. Subsequently, gastrointestinal liquid samples were centrifuged in a Sorvall ST 16R centrifuge (Thermo Scientific, Waltham, MA, USA) in order to remove solid interferences for 10 min at 25830 g and 4ºC, after which 40 µl of supernatant were transferred into HPLC vials and stored at -80ºC until the HPLC-ESI-QTOF-MS analysis.

Regarding plasma samples, an additional step was necessary in order to remove the protein content present in the matrix. Therefore, protein precipitation was carried out with a mixture of organic solvent, in particular 100 µl of the spiked plasma were mixed with 500 µl of ethanol-methanol (50:50, v/v) for 2 h at -20ºC in order to avoid possible degradation. After the protein-precipitation step, the sample was centrifuged for 10 min at 25830 g and 4ºC, and the supernatant was evaporated to dryness under vacuum in a centrifugal evaporator (Concentrator Plus, Eppendorf, Hamburg, Germany) for 3 h. Afterwards, the dry residue was reconstituted in 50 µl of mobile phase A and centrifuged under the same conditions as above to remove any interference. Finally, a 40-µl aliquot was transferred into HPLC vials and stored at -80ºC prior to the analysis.

2.5. HPLC-ESI-QTOF-MS analysis

Analyses were performed using an Agilent 1260 HPLC instrument (Agilent Technologies, Palo Alto, CA, USA) coupled to an Agilent 6540 Ultra High Definition (UHD) Accurate Mass Q-TOF equipped with a Jet Stream dual ESI interface.

The compounds were separated using a reversed-phase C18 analytical column (Agilent Zorbax Eclipse Plus, 1.8 µm, 4.6×150 mm) protected by a guard cartridge of the same packing. The mobile phases were water containing 0.1% of formic acid and acetonitrile as solvent A and B, respectively. The following gradient of these mobile phases was used in order to achieve efficient separation: 0.0 min [A:B 95/5], 5.0 min [A:B 38/62], 10.0 min [A:B 32/68], 19.0 min [A:B 20/80], 34.0 min [A:B 5/95], and 37.0 min [A:B 95/5]. Finally, initial conditions were kept for 5 min at the end of each analysis to equilibrate the analytical column before the next analysis. The column temperature and auto-sampler compartment were set at 25ºC and 4ºC, respectively, whereas the flow rate and the injection volume were 0.8 ml/min and 5 µl.

Detection was performed in negative-ion mode over a range from 100 to 1700 m/z. All spectra were corrected by means of continuous infusion of two reference masses: trifluoroacetate anion (m/z 112.985587) and an adduct of hexakis (^{1}H, ^{1}H, ^{3}H-tetrafluoropropoxy) phosphazine or HP-921 (m/z 1033.988109). Both reference ions provided accurate mass measurements typically better than 2 ppm.

Ultra-high pure nitrogen was used as the drying and nebulizer gas at temperatures of 325 and 400ºC with flows of 10 and 12 L/min, respectively. Other optimised parameters included: capillary voltage +4000V; nebuliser, 20 psi; fragmentor, 130 V; nozzle voltage, 500 V; skimmer, 45 V; and octopole 1 RF Vpp, 750 V.

The MS data were processed through Qualitative Analysis of MassHunter workstation software version B.06.00 (Agilent Technologies).

2.6. Statistical analysis

Data were analyzed using Statgraphics Centurion (version XVI) to perform one-way analysis of variance (ANOVA) with Duncan’s test at a 95% confidence level ($p \leq 0.05$) to identify significant differences among the estimated absorption-rate coefficients for phenolic compounds in rosemary.

3. RESULTS

3.1. Qualitative characterization of phenolic compounds and their metabolites

Phenolic compounds from the extract and several metabolites were qualitatively characterized in gastrointestinal liquid and plasma samples (**Table 1**). This characterization was made by comparing their retention times and mass spectra with commercial standards, when these compounds were commercially available, and/or, when not, with data from the literature (Borrás-Linares et al., 2015; Borrás Linares et al., 2011; Romo Vaquero et al., 2013). The main flavonoids, diterpenes, and triterpenes of the rosemary extract were identified in gastrointestinal liquid samples. Furthermore, the analysis of these samples revealed the presence of 6 metabolites from reactions of carnosic acid, carnosol, and rosmanol. Regarding the plasma samples, the results revealed that 7 compounds from the rosemary extract together with 4 metabolites were bioavailable at the end of the assay. These compounds were identified as rosmanol and its isomers, carnosol, rosmadial, carnosic acid, and 12-methoxycarnosic acid; meanwhile, the metabolites found were the glucuronide forms of carnosic acid, carnosol, and rosmanol as well as 5,6,7,10-tetrahydro-7-hydroxy rosmariquinone.

Table 1. Qualitative and quantitative results for phenolic compounds and their metabolites identified in gastrointestinal liquid and plasma samples (Value = X ± IC(95%)). ND: Not detected.

	ANALYTE	Retention time (min)	Molecular formula	INTESTINAL CONTENT (µg/ml)						K_{ap}, (h^{-1})	PLASMA CONTENT (µg/ml)
				5 min	10 min	15 min	20 min	25 min	30 min		
FLAVONOIDS	Apigenin	7.73	$C_{15}H_{10}O_5$	0.022 ± 0.006	0.015 ± 0.005	<LOQ	<LOQ	<LOQ	<LOQ	-	ND
	Hispidulin	7.82	$C_{16}H_{12}O_6$	0.040 ± 0.007	0.029 ± 0.004	0.022 ± 0.008	<LOQ	<LOQ	<LOQ	-	ND
	Diosmetin	8.45	$C_{16}H_{12}O_6$	0.14 ± 0.03	0.096 ± 0.006	0.065 ± 0.009	0.05 ± 0.01	0.035 ± 0.007	0.031 ± 0.006	3.8 ± 0.6^{a}	ND
	Cirsimaritin	8.63	$C_{17}H_{14}O_6$	0.36 ± 0.05	0.31 ± 0.02	0.21 ± 0.04	0.16 ± 0.01	0.13 ± 0.02	0.13 ± 0.02	3.1 ± 0.4^{abcd}	ND
	Genkwanin	9.46	$C_{16}H_{12}O_5$	2.2 ± 0.2	2.0 ± 0.2	1.27 ± 0.07	1.0 ± 0.2	0.70 ± 0.05	0.71 ± 0.12	3.3 ± 0.4^{abc}	ND
DITERPENES	Rosmanol	8.83	$C_{20}H_{26}O_5$	1.8 ± 0.3	1.4 ± 0.1	0.84 ± 0.06	0.64 ± 0.07	0.42 ± 0.03	0.37 ± 0.04	3.7 ± 0.5^{ab}	0.051 ± 0.007
	Epiisorosmanol	9.16	$C_{20}H_{26}O_5$	0.3 ± 0.1	0.18 ± 0.05	0.12 ± 0.05	0.11 ± 0.04	0.08 ± 0.03	0.07 ± 0.02	2.9 ± 0.5^{abcde}	<LOQ
	Epirosmanol	9.56	$C_{20}H_{26}O_5$	0.15 ± 0.07	0.09 ± 0.02	0.06 ± 0.01	0.060 ± 0.006	0.043 ± 0.005	0.045 ± 0.003	2.3 ± 0.3^{cdef}	<LOQ
	Miltipolone	11.79	$C_{19}H_{24}O_3$	0.051 ± 0.009	0.041 ± 0.006	0.035 ± 0.004	0.034 ± 0.004	0.029 ± 0.004	0.025 ± 0.002	1.2 ± 0.1^{gh}	ND
	Carnosol	12.79	$C_{20}H_{26}O_4$	3.6 ± 0.8	2.6 ± 0.4	1.7 ± 0.2	1.7 ± 0.2	1.1 ± 0.2	1.2 ± 0.2	2.9 ± 0.4^{abcde}	0.054 ± 0.007
	Rosmadial	13.82	$C_{20}H_{24}O_5$	0.43 ± 0.04	0.33 ± 0.04	0.23 ± 0.03	0.20 ± 0.03	0.14 ± 0.02	0.14 ± 0.01	2.9 ± 0.5^{abcd}	0.13 ± 0.01
	Rosmaridiphenol	14.66	$C_{20}H_{28}O_3$	0.30 ± 0.08	0.31 ± 0.06	0.21 ± 0.05	0.22 ± 0.04	0.15 ± 0.03	0.16 ± 0.03	1.9 ± 0.5^{efg}	ND
	Carnosic acid	16.11	$C_{20}H_{28}O_4$	39 ± 19	30 ± 5	15 ± 3	13 ± 4	8.3 ± 0.3	8.5 ± 0.4	3.7 ± 0.6^{ab}	20 ± 5
	12-methoxycarnosic acid	18.41	$C_{21}H_{30}O_3$	4.8 ± 0.9	3.9 ± 0.6	3.0 ± 0.6	2.2 ± 0.3	1.6 ± 0.2	1.5 ± 0.5	2.7 ± 0.5^{bcde}	0.58 ± 0.06
TRITERPENES	Anemosapogenin	14.28	$C_{30}H_{48}O_4$	8 ± 2	7 ± 1	6 ± 1	6 ± 1	4.9 ± 0.9	4.0 ± 0.4	1.4 ± 0.3^{fgh}	ND
	Augustic acid	14.55	$C_{30}H_{48}O_4$	6 ± 2	5 ± 1	4 ± 1	3.3 ± 0.7	3.1 ± 0.8	2.5 ± 0.7	2.1 ± 0.2^{defg}	ND
	Benthamic acid	15.15	$C_{30}H_{48}O_4$	5 ± 1	4 ± 1	4 ± 1	3.3 ± 0.7	3.3 ± 0.8	2.8 ± 0.7	1.4 ± 0.3^{fgh}	ND
	Micromeric acid	21.64	$C_{30}H_{46}O_3$	12 ± 2	9 ± 3	10 ± 3	9 ± 2	11 ± 2	9 ± 1	0.8 ± 0.3^{h}	ND
	Betulinic acid	22.94	$C_{30}H_{48}O_3$	116 ± 11	116 ± 16	102 ± 14	101 ± 14	117 ± 8	106 ± 11	0.6 ± 0.3^{h}	ND
	Ursolic acid	24.32	$C_{30}H_{48}O_3$	11 ± 3	9 ± 3	10 ± 3	9 ± 3	11 ± 2	9 ± 2	0.6 ± 0.3^{h}	ND
METABOLITES	Carnosic acid sulfate	6.46	$C_{20}H_{28}O_7S$	0.54 ± 0.05	0.59 ± 0.05	0.55 ± 0.05	0.55 ± 0.03	0.53 ± 0.04	0.53 ± 0.03		ND
	Carnosic cysteine	6.84	$C_{23}H_{33}NO_6S$	0.7 ± 0.1	0.60 ± 0.05	0.61 ± 0.07	0.57 ± 0.05	0.59 ± 0.05	0.57 ± 0.06		ND
	Rosmanol glucuronide	7.47	$C_{26}H_{34}O_{11}$	0.041 ± 0.004	0.08 ± 0.01	0.09 ± 0.02	0.11 ± 0.01	0.11 ± 0.02	0.15 ± 0.02		0.39 ± 0.06
	Carnosol glucuronide	8.51	$C_{26}H_{34}O_{10}$	0.031 ± 0.004	0.034 ± 0.005	0.035 ± 0.005	0.05 ± 0.01	0.044 ± 0.009	0.06 ± 0.01		0.4 ± 0.1
	Carnosic acid glucuronide	9.08	$C_{26}H_{36}O_{10}$	0.48 ± 0.02	0.69 ± 0.08	0.7 ± 0.1	1.1 ± 0.2	1.0 ± 0.2	1.6 ± 0.3		26 ± 4
	5,6,7,10-tetrahydro-7-hydroxy rosmariquinone	15.31	$C_{19}H_{26}O_3$	0.71 ± 0.09	0.62 ± 0.04	0.54 ± 0.01	0.52 ± 0.01	0.48 ± 0.01	0.48 ± 0.02		0.34 ± 0.02

3.2. Quantitative characterization

For quantitative purposes, standard calibration curves of apigenin, diosmetin, genkwanin, carnosol, carnosic acid, and ursolic acid were prepared using luteolin as the internal standard. The validation of the proposed method was performed with linearity, sensitivity, and precision parameters. **Table 2** shows the limits of detection (LODs) and quantification (LOQs), calibration range, calibration equations, and regression coefficient (r^2) for all standards used. Calibration curves showed good linearity between different concentrations depending on the analytes studied. The LODs and LOQs for individual compounds in standard solutions were calculated as S/N = 3 and S/N = 10, respectively.

Table 2. Calibration parameters for the six commercial standards used.

Analyte	LOD (µg/ml)	LOQ (µg/ml)	Calibration range (µg/ml)	Calibration equation	r^2
Carnosic acid	0.006	0.02	LOQ – 75	y = 0.783x - 0.340	0.97
Genkwanin	0.003	0.01	LOQ – 10	y = 0.233x + 0.083	0.988
Diosmetin	0.003	0.01	LOQ – 5	y = 0.250x + 0.002	0.998
Apigenin	0.004	0.014	LOQ – 5	y = 0.474x + 0.010	0.9997
Carnosol	0.004	0.012	LOQ – 10	y = 1.037x - 0.513	0.989
Ursolic acid	0.05	0.15	LOQ – 125	y = 0.0123x + 0.026	0.98

Intraday and interday precision values were measured to evaluate the repeatability of the method. The rosemary extract was injected several times (n = 6) on the same day (intraday precision) and 3 times on 2 consecutive days (interday precision, n=12). The relative standard deviation in terms of concentration was determined. The intraday repeatability of the method developed for all the analytes ranged from 0.18 to 2.46%, whereas the interday repeatability ranged from 2.1 to 9.63%.

The compound concentrations were calculated by interpolation of the corrected area for each compound (3 replicates) in its appropriate calibration curve. Nevertheless, compounds which were commercially unavailable were tentatively quantified by calibration curves from commercial standards with structural analogies.

Thus, rosmanol, its isomers (epirosmanol and epiisorosmanol) and its glucuronide derivate, miltipolone, rosmadial, rosmaridiphenol, and carnosol glucuronide concentrations were estimated using the carnosol calibration curve. Betulinic, micromeric, augustic, and benthamic acids together with anemosapogenin, were quantified using ursolic acid standard. Carnosic acid curve was used to quantify 12-methoxycarnosic acid, carnosic acid glucuronide, 5,6,7,10-tetrahidro-7-rosmariquinone, carnosic acid sulphate, and carnosic cysteine. Finally, cirsimaritin and hispidulin were tentatively quantified using genkwanin and diosmetin standards, respectively.

Unfortunately, the absorption of apigenin and hispidulin could not be quantified because their concentration values in the advanced stages of the assay were below the LOQ in the appropriate calibration curve. These results were expected due to their corresponding low concentrations in the rosemary extract supplied. Analogously, the signals of epirosmanol and epiisorosmanol in plasma samples were below the corresponding LOQ.

3.3. Absorption-rate coefficients

Once the concentration for each compound was determined in the different gastrointestinal liquid samples, the absorption-rate coefficients (k_a, h^{-1}) of rosemary compounds were estimated by a nonlinear regression analysis of the corrected concentrations (C_t) found in the gut vs. time (t) by adjusting to a first-order kinetic model (Eq. 1). The correction of the concentration took into account the volume reduction during the course of the assay and the water-reabsorption procedure. This model has been applied mainly in studies to evaluate drug absorption (I. Lozoya-Agullo et al., 2016; Oltra-Noguera et al., 2015) and the results have shown a good correlation with literature values of fraction dose absorbed in humans (Isabel Lozoya-Agullo et al., 2015).

$$C_t = C_o e^{-k_a t} \quad \textbf{Eq. (1)}$$

Table 1 lists the rosemary compounds and metabolites found in the different samples collected during the assay, together with their retention times, molecular formulas, absorption-rate coefficients in addition to their concentrations in the gastrointestinal liquid and plasma samples.

4. DISCUSSION

4.1. Monitoring of phenolic compounds and their metabolites in the gastrointestinal tract

The trends found are shown schematically in **Fig. 1**, while **Fig. 2** represents the values of apparent first-order absorption-rate coefficients for each phenolic compound joined to the information concerning the statistical analysis. For a better understanding, the results are presented by compound families, in particular, flavonoids, diterpenes, triterpenes, and metabolites.

4.1.1. Flavonoids

Regarding flavonoids, a similar trend was observed for genkwanin, cirsimaritin, and diosmetin in small intestine. This is depicted in **Fig. 1a**, where showing that the contents of these flavonoids were in decline over the time points of the assay. Hence, for these compounds, the highest concentrations were found at the beginning of the assay, when a few or none of the compounds had been absorbed or metabolised by the gut.

The apparent first-order absorption-rate coefficients were 3.8, 3.3, and 3.1 h^{-1} for diosmetin, genkwanin, and cirsimaritin, respectively. These values were very high compared to those of the other phenolic compounds of the extract; in fact, the absorption coefficient found for diosmetin was the highest.

However, the statistical analysis reflected no significant differences, either among flavonoids or with among the majority of the diterpenes (except rosmaridiphenol and miltipolone), whereas there were significant differences with triterpenes with the exception of augustic acid and cirsimaritin constants.

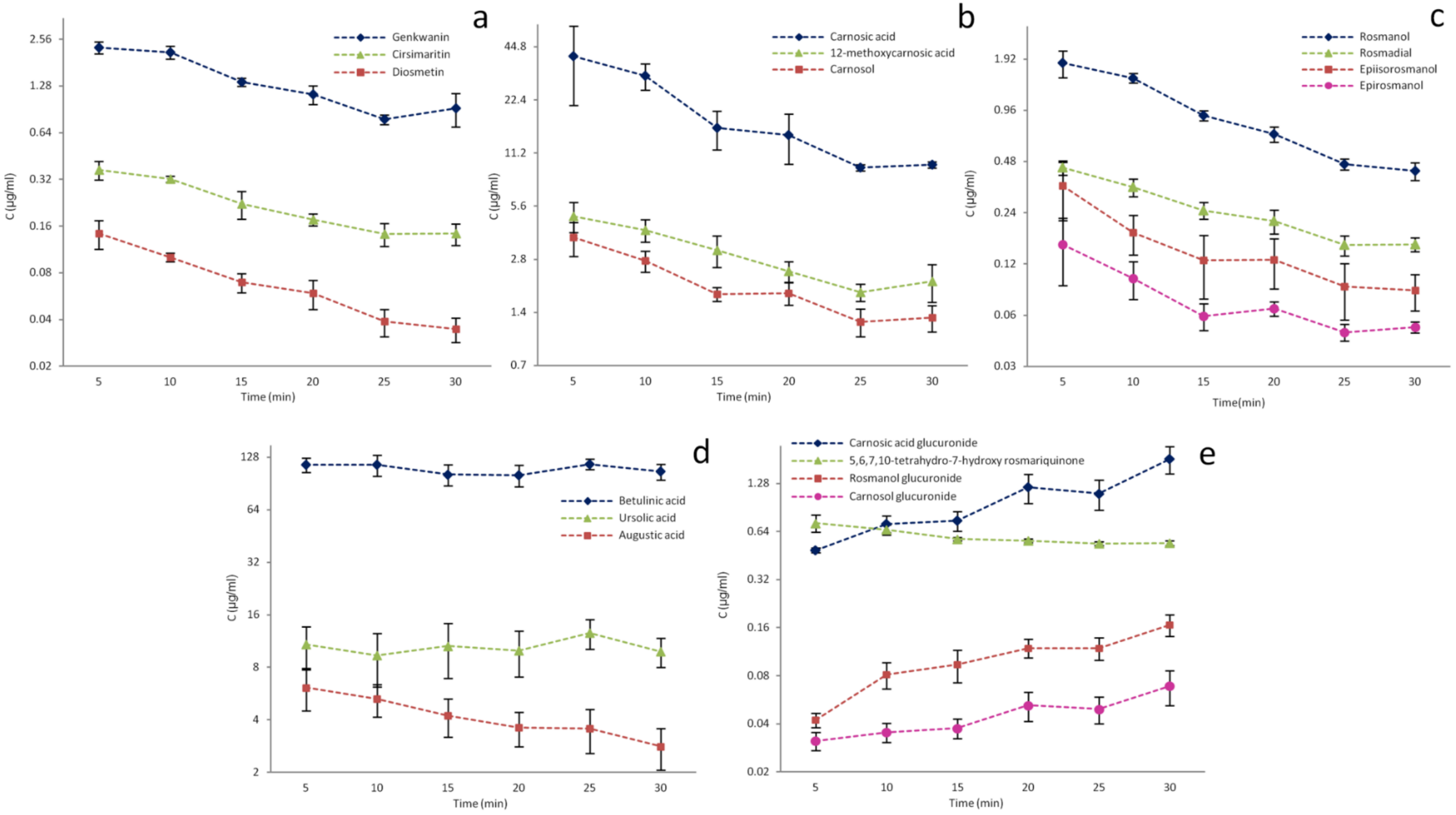


Fig. 1. Monitorization of phenolic compounds and their metabolites in the small intestine over time. 1a: Flavonoids (Genkwanin, Cirsimaritin and Diosmetin); 1b: Diterpenes (Carnosic acid, 12-methoxycarnosic acid and Carnosol); 1c: Diterpenes (Rosmanol, Rosmadial, Epiisorosmanol and Epirosmanol); 1d: Triterpenes (Betulinic, Ursolic and Augustic acids); 1e: Metabolites (Carnosic acid glucuronide, 5,6,7,10-tetrahydro-7-hydroxy rosmariquinone, rosmanol glucuronide, and carnosol glucuronide)

4.1.2. Diterpenes

The analysis of the gastrointestinal-liquid samples revealed the presence of the rosemary diterpenes. **Fig. 1b** shows that carnosic acid presented the highest concentration in the small intestine over time. Regarding this compound, it should be taken into account that carnosic acid could undergo several degradation reactions due to a certain instability by the action of several factors, mainly the presence of oxygen or light. Therefore, it could enter a decomposition process leading to the formation of other diterpenes, such as rosmanol and its isomers, 12-methoxycarnosic acid or carnosol (Zhang et al., 2012). Therefore, this possibility has been considered in discussing the results in the current research, although light and temperature were not taken into account during the assay. As observed for the flavonoids, carnosic acid concentration tended to decline in the small intestine over the assay.

Meanwhile, carnosol, 12-methoxycarnosic acid (**Fig. 1b**), rosmadial, rosmanol, and its isomers (**Fig. 1c**) also showed patterns similar to those of carnosic acid. The highest absorption-rate coefficients of this family of compounds during the assay followed the order: carnosic acid and rosmanol (3.7 h^{-1}) > epiisorosmanol, carnosol, and rosmadial (2.9 h^{-1}) > 12-methoxycarnosic acid (2.7 h^{-1}), although no significant differences were found between them.

However, epirosmanol (2.3 h^{-1}), rosmaridiphenol (1.9 h^{-1}), and miltipolone (1.2 h^{-1}) showed different trends, with absorption-rate coefficients significantly lower than found for the rest of diterpenes. Consequently, the intestinal absorption of these compounds proved to be the lowest found for diterpenes; in addition, they showed no significant differences with respect to some triterpenes (anemosapogenin, augustic acid, and benthamic acid) or with all of them in the case of miltipolone. These results could be interpreted as indicating that these compounds were neither absorbed nor metabolised by the gut. This effect could be explained by their chemical structures, which could hinder absorption across the intestinal barrier. However, these compounds reach the colon in high proportion, where they could exhibit their bioactivity against colorectal cancer by means of interactions with the microbiota. This possibility would likely have a direct effect on the progression of cancerous colonocytes. Testing this hypothesis, studies have selected some of these compounds as having potential antiproliferative properties against colon adenocarcinoma cells, as in the case of rosmaridiphenol (Sánchez-Camargo, García-Cañas, Herrero, Cifuentes, & Ibáñez, 2016).

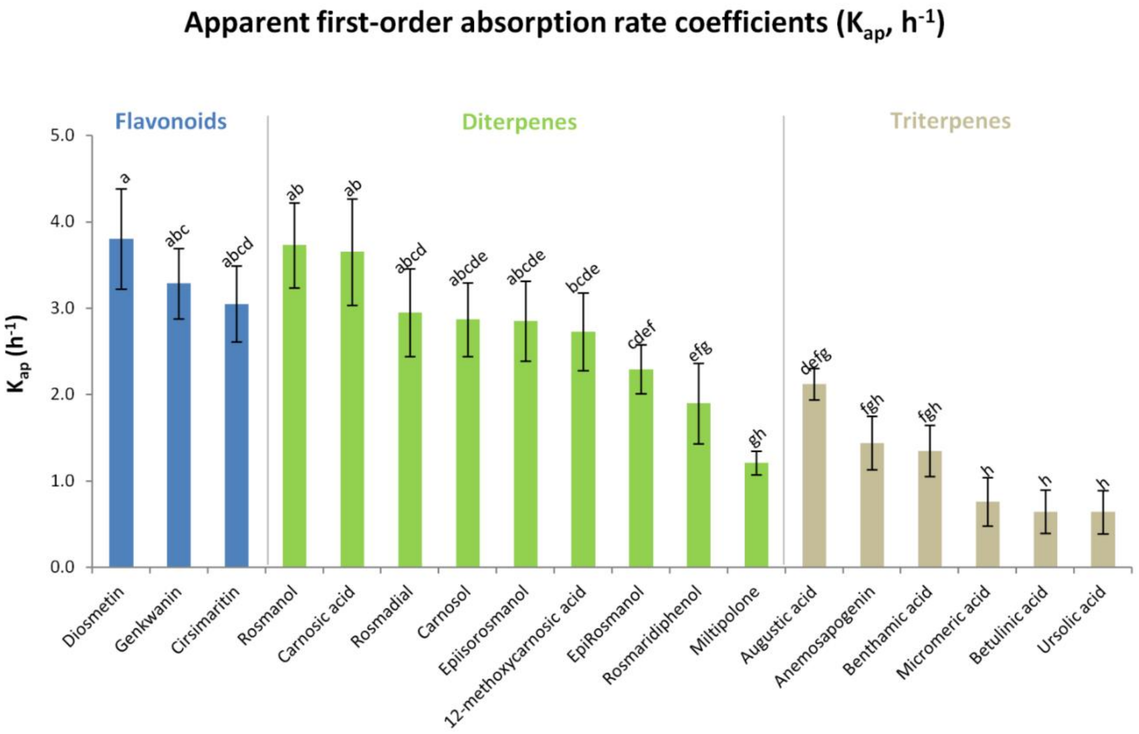


Fig. 2. Apparent first-order absorption-rate coefficients of the phenolic compounds from the rosemary-leaf extract in the small intestine. (The same letter means no significant differences among them, p>0.05).

4.1.3. Triterpenes

In the gastrointestinal-liquid samples, triterpenes presented the highest concentrations. These results agreed with the greatest percentage of this family in the extract supplied. Among the compounds studied, betulinic acid was the most concentrated triterpene characterized in the gastrointestinal liquid, followed by ursolic and micromeric acids. However, anemosapogenin, augustic, and benthamic acids were found to a lesser extent.

Nevertheless, despite the high concentration of these kinds of compounds found in the extract and in the gut, they exhibited poor absorption and metabolism during the assay according to the absorption-rate coefficients. This trend is represented in **Fig. 1d** for ursolic, betulinic, and augustic acids, the most representative ones. Augustic acid was the most absorbed rosemary triterpene according to the absorption-rate coefficient (2.1 h^{-1}), although it did not show statistically significant differences with anemosapogenin or benthamic acid, the flavonoid cirsimaritin or certain diterpenes. Benthamic acid and anemosapogenin (1.4 h^{-1}) followed identical patterns, without significant differences with triterpenes, rosmaridiphenol, epirosmanol or miltipolone. On the other hand, micromeric, betulinic, and ursolic acids registered the lowest absorption-rate coefficients among all phenolic compounds (0.8, 0.6, and 0.6 h^{-1}, respectively). The hampered absorption of these compounds could be attributed to the great molecular weights and complexity of the structures of these types of compounds. In fact, an inverse relation between structure complexity and absorption has been reported in the literature for polyphenols (Rein et al., 2013). On the other hand, triterpenes are the most hydrophobic compounds among the families studied. Hence, their gut absorption is more hindered due to their greater difficulty to cross the water layer on the gut membranes. Indeed, several studies have highlighted a low intestinal absorption for ursolic and oleanic acids, which possess similar structures to those of the aforementioned triterpenes (Jeong et al., 2007; Liao et al., 2005). Nevertheless, it is important to consider that these poorly absorbed compounds may exhibit their bioactive properties on the colon by means of their interactions with the gut microbiota and through a direct effect on colonocytes. This suggestion was supported by the high concentrations of these compounds found in the gastrointestinal samples as well as the antiproliferative properties in colon cancer evidenced in previous research for ursolic and betulinic acids (Kim et al., 2014; Shan, Xuan, Zheng, Dong, & Zhang, 2009).

4.1.4. Metabolites from rosemary phenolic compounds

The results suggest that several diterpenes were quickly metabolised in the small intestine due to the presence of their metabolites in gastrointestinal samples at 5 min after the administration of the extract.

The glucuronidation reaction appeared to be the main metabolism pathway for these diterpenes, which occurs through uridine diphosphate glucuronosyltransferares (UGT) in the small intestine and/or liver (Phase II metabolism) (Crozier, Del Rio, & Clifford, 2010; Marín, Miguélez, Villar, & Lombó, 2015). Furthermore, the presence of carnosic acid sulphate suggests that a metabolisation reaction by the action of 3’-phosphoadenosine-5’-phosphosulfate with sulfotransferases (Levsen et al., 2005) also took place in the small intestine. Moreover, carnosic acid underwent another metabolisation process, leading to the formation of carnosic cysteine, probably by the loss of glycine and glutamic acid from the carnosic glutathione molecule (Romo Vaquero et al., 2013). In addition, 5,6,7,10-

tetrahydro-7-hydroxy rosmariquinone was described in bibliography as a degradation product of carnosic acid due to light exposure (Zhang et al., 2012), although the presence of this compound in perfusion samples was probably due to an oxidation reaction of carnosic acid (Satoh et al., 2008) through phase-I metabolism, favoured by oxygen and radicals of the cells and cytochrome P450, as has been proposed in the work (Romo Vaquero et al., 2013).

The results demonstrate that these molecules were metabolised inside the small intestine and, consequently, these metabolites could remain in the gut and reach the colon, where they could exhibit their bioactive effects and/or could be absorbed into the bloodstream.

Intestinal activity with respect to glucuronide metabolites of carnosic acid, carnosol, and rosmanol showed an trend opposite to that of the phenolic compounds. Thus, concentrations for these compounds increased over time, as reflected in **Fig. 1e**. For the compound 5,6,7,10-tetrahydro-7-hydroxy rosmariquinone a slightly lower concentration was observed in gastrointestinal samples, corresponding to the content of this compound found in plasma. On the contrary, the other metabolised forms of carnosic acid (carnosic acid sulphate and carnosic cysteine) maintained their concentrations almost constant during the experiment, perhaps indicating that these transformations were less favoured than glucuronidation reactions.

4.2. Bioavailability of phenolic compounds and their metabolites

As mentioned above, phenolic compounds and metabolites could be absorbed across the gut barrier to reach the blood stream. For verification of the absorption and an in-depth study of the bioavailability of these compounds, plasma samples taken at the end of the *in situ* assay were also analysed.

These results revealed that carnosic acid was the phenolic compound which presented the highest concentration in plasma, and therefore it could reach target tissues more readily. In fact, the contents of its major metabolite (carnosic acid glucuronide) and carnosic acid were by far the most abundant compounds in plasma samples, in agreement with the evidence found by other authors suggesting that carnosic acid is the main bioavailable and bioactive compound from *Rosmarinus officinalis* (Zhao et al., 2015).

On the other hand, the rest of the quantified compounds were found in much lower quantities than was carnosic acid and its glucuronide metabolite. It is also important to highlight that glucuronide metabolites in plasma were found at higher concentrations than their corresponding phenolic compounds (carnosic acid, carnosol, and rosmanol). Therefore, the glucuronidation reaction for these three phenolic compounds appears to be highly favoured, as mentioned above. This favoured absorption could be related to the increase in the hydrophilic nature of the metabolites with respect to the phenolic compounds. In fact, bibliographic studies of polyphenols detailed that the predominant structures in plasma are conjugates (sulphate, glucuronide or methylated) (Kroon et al., 2004). In addition, the presence of these metabolites in both samples evidenced that these analytes were effectively absorbed in their metabolised form through gastrointestinal barrier, in all likelihood.

Despite that absorption-rate constants of flavonoids were the highest, these compounds were not detected in plasma samples. This finding could be the result of the strong tendency of these kinds of compounds to bind with plasma proteins, mainly albumin, as reported elsewhere (Xiao & Kai, 2012). Moreover, studies have shown that this sort of binding with proteins could compromise the antioxidant capacity of flavonoids (Arts et al., 2002). Therefore, it is highly likely that the antiproliferative activity is largely related to diterpenes, their metabolites, and triterpenes instead of flavonoids.

CONCLUSIONS

According with the results, it can be concluded that the previously demonstrated antiproliferative/cytotoxic effects of rosemary extract on colorectal cancer appear to be highly correlated with these bioavailable compounds and their metabolites, which could reach the colon by means of the small intestine and bloodstream. Additionally, the non-absorbed compounds (triterpenes and some diterpenes) may also exhibit their biological activity in the large intestine through their interactions with the gut microbiota and by a direct effect on colonocytes with respect to the onset of cancer or its progression. These compounds with antiproliferative properties were neither absorbed nor metabolised, reaching the colon unaltered. These findings suggest that several bioactive compounds from rosemary could exert an antitumor effect by various complementary modes of action, and they may be considered promising supplementary chemopreventive agents against colorectal cancer. In this sense, further studies on the evolution of compounds in large intestine and the bioactive properties of glucuronides metabolites should be conducted in order to increase the knowledge about the assuring bioactivity of rosemary.

Acknowledgements

This work was supported by the Spanish Ministry of Science and Innovation (grant numbers AGL2015-67995-C3-2-R, AGL2015-67995-C3-1-R, AGL2011-29857-C03-02, AGL2011-29857-C03-03), Andalusian Regional Government Council of Innovation and Science (grant numbers P10-FQM-6563, P11-CTS-7625) Generalitat Valenciana (grant numbers PROMETEO/2012/007, PROMETEO/2016/006, ACOMP/2013/093) and CIBER (grant number CB12/03/30038). The author IBL gratefully acknowledges the Spanish Ministry of Economy and Competitiveness (MINECO) in association with the European Social Fund (FSE) for the contract PTQ-13-06429, the author AFO received support from the Spanish Ministry of Education, Culture and Sports (FPU grant 14/03992) and APS from the Spanish Ministry of Science and Innovation (FPU grant AP2007-03246). Finally, we thank David Nesbitt for reviewing the English language of the text.

Conflict of interest.

All authors declare that they have no conflict of interest.

Chemical compounds studied in this article

Carnosol (PubChem CID 442009); Carnosic Acid (PubChem CID: 65126); Rosmanol (PubChem CID: 13966122); Rosmadial (PubChem CID: 15801061); Ursolic Acid (PubChem CID: 64945); Betulinic Acid (PubChem CID: 64971); Micromeric Acid (PubChem CID: 73242194); Genkwanin (PubChemCID: 5281617); Cirsimaritin (PubChem CID: 188323); Diosmetin (PubChemCID: 5281612)

References.

Altinier, G., Sosa, S., Aquino, R. P., Mencherini, T., Della Loggia, R., & Tubaro, A. (2007). Characterization of topical antiinflammatory compounds in Rosmarinus officinalis L. *Journal of Agricultural and Food Chemistry*, *55*(5), 1718–23. https://doi.org/10.1021/jf062610+

Arts, M. J. T. J., Haenen, G. R. M. M., Wilms, L. C., Beetstra, S. A. J. N., Heijnen, C. G. M., Voss, H. P., & Bast, A. (2002). Interactions between flavonoids and proteins: Effect on the total antioxidant capacity. *Journal of Agricultural and Food Chemistry*, *50*(5), 1184–1187. https://doi.org/10.1021/jf010855a

Bai, N., He, K., Roller, M., Lai, C.-S., Shao, X., Pan, M.-H., & Ho, C.-T. (2010). Flavonoids and Phenolic Compounds from Rosmarinus officinalis. *Journal of Agricultural and Food Chemistry*, *58*(9), 5363–5367. https://doi.org/10.1021/jf100332w

Birtić, S., Dussort, P., Pierre, F.-X., Bily, A. C., & Roller, M. (2015). Carnosic acid. *Phytochemistry*, *115*, 9–19. https://doi.org/10.1016/j.phytochem.2014.12.026

Borrás-Linares, I., Pérez-Sánchez, A., Lozano-Sánchez, J., Barrajón-Catalán, E., Arráez-Román, D., Cifuentes, A., … Carretero, A. S. (2015). A bioguided identification of the active compounds that contribute to the antiproliferative/cytotoxic effects of rosemary extract on colon cancer cells. *Food and Chemical Toxicology*, *80*, 215–222. https://doi.org/10.1016/j.fct.2015.03.013

Borrás-Linares, I., Stojanović, Z., Quirantes-Piné, R., Arráez-Román, D., Švarc-Gajić, J., Fernández-Gutiérrez, A., & Segura-Carretero, A. (2014). Rosmarinus officinalis leaves as a natural source of bioactive compounds. *International Journal of Molecular Sciences*, *15*(11), 20585–606. https://doi.org/10.3390/ijms151120585

Borrás Linares, I., Arráez-Román, D., Herrero, M., Ibáñez, E., Segura-Carretero, A, & Fernández-Gutiérrez, A. (2011). Comparison of different extraction procedures for the comprehensive characterization of bioactive phenolic compounds in Rosmarinus officinalis by reversed-phase high-performance liquid chromatography with diode array detection coupled to electrospray time. *Journal of Chromatography. A*, *1218*(42), 7682–90. https://doi.org/10.1016/j.chroma.2011.07.021

Bozin, B., Mimica-Dukic, N., Samojlik, I., & Jovin, E. (2007). Antimicrobial and antioxidant properties of rosemary and sage (Rosmarinus officinalis L. and Salvia officinalis L., Lamiaceae) essential oils. *Journal of Agricultural and Food Chemistry*, *55*(19), 7879–85. https://doi.org/10.1021/jf0715323

Clifford, M., van der Hooft, J., & Crozier, A. (2013). Human studies on the absorption , distribution , metabolism , and excretion of tea polyphenols. *American Journal of Clinical Nutrition*, *98*, 1619S–30S. https://doi.org/10.3945/ajcn.113.058958.1

Crozier, A., Del Rio, D., & Clifford, M. N. (2010). Bioavailability of dietary flavonoids and phenolic compounds. *Molecular Aspects of Medicine*, *31*(6), 446–467. https://doi.org/10.1016/j.mam.2010.09.007

D’Antuono, I., Garbetta, A., Linsalata, V., Minervini, F., & Cardinali, A. (2015). Polyphenols from artichoke heads (Cynara cardunculus (L.) subsp. scolymus Hayek): in vitro bio-accessibility, intestinal uptake and bioavailability. *Food Funct.*, *6*(4), 1268–1277. https://doi.org/10.1039/C5FO00137D

Doluisio, J. T., Billups, N. F., Dittert, L. W., Sugita, E. T., & Swintosky, J. V. (1969). Drug absorption I: An in situ rat gut technique yielding realistic absorption rates. *Journal of Pharmaceutical Sciences*, *58*(10). https://doi.org/10.1002/jps.2600581006

Doolaege, E. H., Raes, K., De Vos, F., Verhe, R., & De Smet, S. (2011). Absorption, distribution and elimination of carnosic acid, a natural antioxidant from Rosmarinus officinalis, in rats. *Plant Foods Hum Nutr*, *66*(2), 196–202. https://doi.org/10.1007/s11130-011-0233-5

Haloui, M., Louedec, L., Michel, J.-B., & Lyoussi, B. (2000). Experimental diuretic effects of Rosmarinus officinalis and Centaurium erythraea. *Journal of Ethnopharmacology*, *71*(3), 465–472. https://doi.org/10.1016/S0378-8741(00)00184-7

Herrero, M., Plaza, M., Cifuentes, A, & Ibáñez, E. (2010). Green processes for the extraction of bioactives from Rosemary: Chemical and functional characterization via ultra-performance liquid chromatography-tandem mass spectrometry and in-vitro assays. *Journal of Chromatography. A*, *1217*(16), 2512–20. https://doi.org/10.1016/j.chroma.2009.11.032

Jeong, D. W., Kim, Y. H., Kim, H. H., Ji, H. Y., Yoo, S. D., Choi, W. R., … Lee, H. S. (2007). Dose-linear pharmacokinetics of oleanolic acid after intravenous and oral administration in rats. *Biopharmaceutics & Drug Disposition*, *28*(2), 51–7. https://doi.org/10.1002/bdd.530

Jiang, Y., Wu, N., Fu, Y.-J., Wang, W., Luo, M., Zhao, C.-J., … Liu, X.-L. (2011). Chemical composition and antimicrobial activity of the essential oil of Rosemary. *Environmental Toxicology and Pharmacology*, *32*(1), 63–8. https://doi.org/10.1016/j.etap.2011.03.011

Kim, J.-H., Kim, Y. H., Song, G.-Y., Kim, D.-E., Jeong, Y.-J., Liu, K.-H., … Oh, S. (2014). Ursolic acid and its natural derivative corosolic acid suppress the proliferation of APC-mutated colon cancer cells through promotion of β-catenin degradation. *Food and Chemical Toxicology : An International Journal Published for the British Industrial Biological Research Association*, *67*, 87–95. https://doi.org/10.1016/j.fct.2014.02.019

Kosińska, A., & Andlauer, W. (2012). Cocoa polyphenols are absorbed in Caco-2 cell model of intestinal epithelium. *Food Chemistry*, *135*(3), 999–1005. https://doi.org/10.1016/j.foodchem.2012.05.101

Kroon, P. A., Clifford, M. N., Crozier, A., Day, A. J., Donovan, J. L., Manach, C., & Williamson, G. (2004). How should we assess the effects of exposure to dietary polyphenols in vitro? *American Journal of Clinical Nutrition*, *80*(1), 15–21.

Levsen, K., Schiebel, H.-M., Behnke, B., Dötzer, R., Dreher, W., Elend, M., & Thiele, H. (2005). Structure elucidation of phase II metabolites by tandem mass spectrometry: an overview. *Journal of Chromatography A*, *1067*(1–2), 55–72. https://doi.org/10.1016/j.chroma.2004.08.165

Liao, Q., Yang, W., Jia, Y., Chen, X., Gao, Q., & Bi, K. (2005). LC-MS determination and pharmacokinetic studies of ursolic acid in rat plasma after administration of the traditional chinese medicinal preparation Lu-Ying extract. *Yakugaku Zasshi : Journal of the Pharmaceutical Society of Japan*, *125*(6), 509–515. https://doi.org/10.1248/yakushi.125.509

Lozoya-Agullo, I., Zur, M., Beig, A., Fine, N., Cohen, Y., González-Álvarez, M., … Dahan, A. (2016). Segmental-dependent permeability throughout the small intestine following oral drug administration: Single-pass vs. Doluisio approach to in-situ rat perfusion. *International Journal of Pharmaceutics*, *515*(1–2). https://doi.org/10.1016/j.ijpharm.2016.09.061

Lozoya-Agullo, I., Zur, M., Wolk, O., Beig, A., González-Álvarez, I., González-Álvarez, M., … Dahan, A. (2015). In-situ intestinal rat perfusions for human Fabs prediction and BCS permeability class determination: Investigation of the single-pass vs. the Doluisio experimental approaches. *International Journal of Pharmaceutics*, *480*(1), 1–7. https://doi.org/10.1016/j.ijpharm.2015.01.014

Marín, L., Miguélez, E. M., Villar, C. J., & Lombó, F. (2015). Bioavailability of Dietary Polyphenols and Gut Microbiota Metabolism : Antimicrobial Properties. *Biomed Research International*, *2015*.

Naemura, A., Ura, M., Yamashita, T., Arai, R., & Yamamoto, J. (2008). Long-term intake of rosemary and common thyme herbs inhibits experimental thrombosis without prolongation of bleeding time. *Thrombosis Research*, *122*(4), 517–22. https://doi.org/10.1016/j.thromres.2008.01.014

Ngo, S. N. T., Williams, D. B., & Head, R. J. (2011). Rosemary and cancer prevention: preclinical perspectives. *Critical Reviews in Food Science and Nutrition*, *51*(10), 946–54. https://doi.org/10.1080/10408398.2010.490883

Oltra-Noguera, D., Mangas-Sanjuan, V., González-Álvarez, I., Colon-Useche, S., González-Álvarez, M., & Bermejo, M. (2015). Drug gastrointestinal absorption in rat: Strain and gender differences. *European Journal of Pharmaceutical Sciences*, *78*, 198–203. https://doi.org/10.1016/j.ejps.2015.07.021

Pelillo, M., Cuvelier, M. E., Biguzzi, B., Gallina Toschi, T., Berset, C., & Lercker, G. (2004). Calculation of the molar absorptivity of polyphenols by using liquid chromatography with diode array detection: The case of carnosic acid. *Journal of Chromatography A*, *1023*(2), 225–229. https://doi.org/10.1016/S0021-9673(03)01206-8

Pérez-Fons, L., Garzón, M. T., & Micol, V. (2010). Relationship between the antioxidant capacity and effect of rosemary (Rosmarinus officinalis L.) polyphenols on membrane phospholipid order. *Journal of Agricultural and Food Chemistry*, *58*(1), 161–71. https://doi.org/10.1021/jf9026487

Petiwala, S. M., & Johnson, J. J. (2015). Diterpenes from rosemary (Rosmarinus officinalis): Defining their potential for anti-cancer activity. *Cancer Letters*, *367*(2), 93–102. https://doi.org/10.1016/j.canlet.2015.07.005

Petiwala, S. M., Puthenveetil, A. G., & Johnson, J. J. (2013). Polyphenols from the Mediterranean herb rosemary (Rosmarinus officinalis) for prostate cancer. *Frontiers in Pharmacology*, *4*, 29. https://doi.org/10.3389/fphar.2013.00029

Rein, M. J., Renouf, M., Cruz-Hernandez, C., Actis-Goretta, L., Thakkar, S. K., & da Silva Pinto, M. (2013). Bioavailability of bioactive food compounds: A challenging journey to bioefficacy. *British Journal of Clinical Pharmacology*, *75*(3), 588–602. https://doi.org/10.1111/j.1365-2125.2012.04425.x

Romo Vaquero, M., García Villalba, R., Larrosa, M., Yáñez-Gascón, M. J., Fromentin, E., Flanagan, J., … García-Conesa, M. T. (2013). Bioavailability of the major bioactive diterpenoids in a rosemary extract: Metabolic profile in the intestine, liver, plasma, and brain of Zucker rats. *Molecular Nutrition and Food Research*, *57*, 1834–1846. https://doi.org/10.1002/mnfr.201300052

Sánchez-Camargo, A. P., García-Cañas, V., Herrero, M., Cifuentes, A., & Ibáñez, E. (2016). Comparative study of green sub- and supercritical processes to obtain carnosic acid and carnosol-enriched rosemary extracts with in vitro anti-proliferative activity on colon cancer cells. *International Journal of Molecular Sciences*, *17*(12). https://doi.org/10.3390/ijms17122046

Satoh, T., Kosaka, K., Itoh, K., Kobayashi, A., Yamamoto, M., Shimojo, Y., … Lipton, S. a. (2008). Carnosic acid, a catechol-type electrophilic compound, protects neurons both in vitro and in vivo through activation of the Keap1/Nrf2 pathway via S-alkylation of targeted cysteines on Keap1. *Journal of Neurochemistry*, *104*(4), 1116–1131. https://doi.org/10.1111/j.1471-4159.2007.05039.x

Sedighi, R., Zhao, Y., Yerke, A., & Sang, S. (2015). Preventive and protective properties of rosemary (Rosmarinus officinalis L.) in obesity and diabetes mellitus of metabolic disorders: a brief review. *Current Opinion in Food Science*, *2*, 58–70. https://doi.org/10.1016/j.cofs.2015.02.002

Shan, J., Xuan, Y., Zheng, S., Dong, Q., & Zhang, S. (2009). Ursolic acid inhibits proliferation and induces apoptosis of HT-29 colon cancer cells by inhibiting the EGFR/MAPK pathway. *Journal of Zhejiang University. Science. B*, *10*(9), 668–74. https://doi.org/10.1631/jzus.B0920149

Sotelo-Félix, J. ., Martinez-Fong, D., Muriel, P., Santillán, R. ., Castillo, D., & Yahuaca, P. (2002). Evaluation of the effectiveness of Rosmarinus officinalis (Lamiaceae) in the alleviation of carbon tetrachloride-induced acute hepatotoxicity in the rat. *Journal of Ethnopharmacology*, *81*(2), 145–154. https://doi.org/10.1016/S0378-8741(02)00090-9

Srancikova, A., Horvathova, E., & Kozics, K. (2013). Biological effects of four frequently used medicinal plants of Lamiaceae. *Neoplasma*, *60*(6), 585–97. https://doi.org/10.4149/neo_2013_076

Tai, J., Cheung, S., Wu, M., & Hasman, D. (2012). Antiproliferation effect of Rosemary (Rosmarinus officinalis) on human ovarian cancer cells in vitro. *Phytomedicine : International Journal of Phytotherapy and Phytopharmacology*, *19*(5), 436–43. https://doi.org/10.1016/j.phymed.2011.12.012

Teng, Z., Yuan, C., Zhang, F., Huan, M., Cao, W., Li, K., … Mei, Q. (2012). Intestinal Absorption and First-Pass Metabolism of Polyphenol Compounds in Rat and Their Transport Dynamics in Caco-2 Cells. *PLoS ONE*, *7*(1), e29647.

https://doi.org/10.1371/journal.pone.0029647

Tenore, G. C., Campiglia, P., Giannetti, D., & Novellino, E. (2015). Simulated gastrointestinal digestion, intestinal permeation and plasma protein interaction of white, green, and black tea polyphenols. *Food Chemistry*, *169*, 320–6. https://doi.org/10.1016/j.foodchem.2014.08.006

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., & Jemal, A. (2015). Global cancer statistics, 2012. *CA: A Cancer Journal for Clinicians*, *65*(2), 87–108. https://doi.org/10.3322/caac.21262

Tsai, C.-W., Lin, C.-Y., Lin, H.-H., & Chen, J.-H. (2011). Carnosic acid, a rosemary phenolic compound, induces apoptosis through reactive oxygen species-mediated p38 activation in human neuroblastoma IMR-32 cells. *Neurochemical Research*, *36*(12), 2442–51. https://doi.org/10.1007/s11064-011-0573-4

Valdés, A., García-Cañas, V., Rocamora-Reverte, L., Gómez-Martínez, A., Ferragut, J. A., & Cifuentes, A. (2013). Effect of rosemary polyphenols on human colon cancer cells: transcriptomic profiling and functional enrichment analysis. *Genes & Nutrition*, *8*(1), 43–60. https://doi.org/10.1007/s12263-012-0311-9

Willenberg, I., Michael, M., Wonik, J., Bartel, L. C., Empl, M. T., & Schebb, N. H. (2015). Investigation of the absorption of resveratrol oligomers in the Caco-2 cellular model of intestinal absorption. *Food Chemistry*, *167*, 245–50. https://doi.org/10.1016/j.foodchem.2014.06.103

Xiao, J., & Kai, G. (2012). A review of dietary polyphenol-plasma protein interactions: characterization, influence on the bioactivity, and structure-affinity relationship. *Critical Reviews in Food Science and Nutrition*, *52*(1), 85–101. https://doi.org/10.1080/10408398.2010.499017

Yan, H., Wang, L., Li, X., Yu, C., Zhang, K., Jiang, Y., … Tu, P. (2009). High-performance liquid chromatography method for determination of carnosic acid in rat plasma and its application to pharmacokinetic study. *Biomedical Chromatography : BMC*, *23*(7), 776–81. https://doi.org/10.1002/bmc.1184

Zhang, Y., Smuts, J. P., Dodbiba, E., Rangarajan, R., Lang, J. C., & Armstrong, D. W. (2012). Degradation study of carnosic acid, carnosol, rosmarinic acid, and rosemary extract (rosmarinus officinalis L.) assessed using HPLC. *Journal of Agricultural and Food Chemistry*, *60*(36), 9305–9314. https://doi.org/10.1021/jf302179c

Zhao, Y., Sedighi, R., Wang, P., Chen, H., Zhu, Y., & Sang, S. (2015). Carnosic acid as a major bioactive component in rosemary extract ameliorates high-fat-diet-induced obesity and metabolic syndrome in mice. *Journal of Agricultural and Food Chemistry*, *63*(19), 4843–52. https://doi.org/10.1021/acs.jafc.5b01246

SUPLEMENTARY MATERIAL

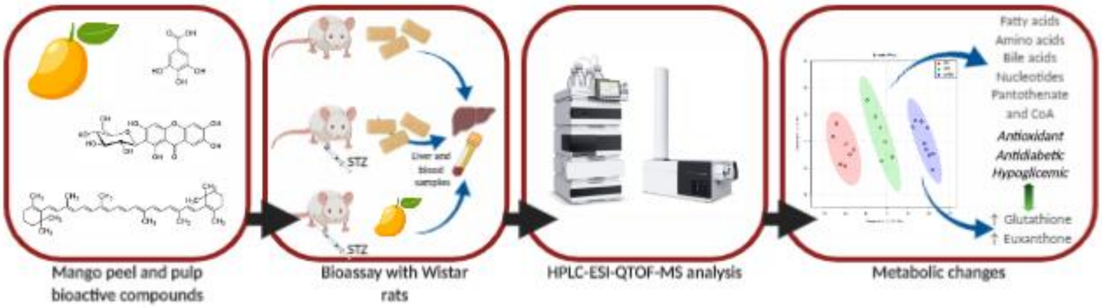
Table 1S. Composition of supplied rosemary extract.

PHENOLIC COMPOSITION OF ROSEMARY-LEAF EXTRACT			
Compound	**m/z**	**Molecular formula**	**Concentration (mg/g)**
Apigenin	269.0461	$C_{15}H_{10}O_5$	0.50 ± 0.02
Hispidulin	299.0565	$C_{16}H_{12}O_6$	0.31 ± 0.01
Diosmetin	299.0553	$C_{16}H_{12}O_6$	0.62 ± 0.04
Cirsimaritin	313.0721	$C_{17}H_{14}O_6$	0.78 ± 0.07
Rosmanol	345.1714	$C_{20}H_{26}O_5$	4.4 ± 0.1
Epiisorosmanol	345.1709	$C_{20}H_{26}O_5$	0.80 ± 0.05
Epirosmanol	345.1709	$C_{20}H_{26}O_5$	0.38 ± 0.02
Genkwanin	283.0620	$C_{16}H_{12}O_5$	2.61 ± 0.05
Miltipolone	299.1652	$C_{19}H_{24}O_3$	0.32 ± 0.04
Carnosol	329.1770	$C_{20}H_{26}O_4$	10 ± 1
Rosmadial	343.1548	$C_{20}H_{24}O_5$	1.36 ± 0.06
Anemosapogenin	471.3471	$C_{30}H_{48}O_4$	6.5 ± 0.5
Augustic acid	471.1960	$C_{30}H_{48}O_4$	6.5 ± 0.5
Rosmaridiphenol	315.1969	$C_{20}H_{28}O_3$	0.25 ± 0.05
Bentamic acid	471.3480	$C_{30}H_{48}O_4$	2.1 ± 0.2
Carnosic acid	331.1930	$C_{20}H_{28}O_4$	83 ± 4
12-methoxycarnosic acid	345.2091	$C_{21}H_{30}O_3$	7.20 ± 0.01
Micromeric acid	453.3375	$C_{30}H_{46}O_3$	47 ± 2
Betulinic acid	455.3548	$C_{30}H_{48}O_3$	38 ± 3
Ursolic acid	455.3540	$C_{30}H_{48}O_3$	21.5 ± 0.6

Published in Food and Chemical Toxicology. Permission from the journal: license number 4031310597485

Capítulo 2

Evaluación de los cambios metabólicos en muestras de hígado y suero de ratas con diabetes inducida con streptozocina tras la ingesta de una dieta rica en mango



Álvaro Fernández-Ochoa, Rosario Cázares-Camacho, Isabel Borrás-Linares, J. Abraham Domínguez-Avila, Antonio Segura-Carretero, Gustavo Adolfo González-Aguilar.

Evaluation of metabolic changes in liver and serum of streptozotocin-induced diabetic rats after Mango diet supplementation

ABSTRACT

The composition of mango has shown bioactive properties against several diseases such as diabetes mellitus. Due to these effects, we evaluated how a diet based on compounds from mango affects metabolic pathways in diabetic rats. Serum and liver samples were collected from 26 rats divided into 3 groups (healthy, untreated diabetic and mango-treated diabetic) after dietary intervention. These were analyzed by an LC-MS untargeted metabolomic strategy. Twenty-six and 29 metabolites in serum and liver were potentially annotated showing significant differences between groups. The most affected pathways were related to fatty acid metabolism, bile acid homeostasis and amino acid degradation. It is also remarkable the detection of euxanthone and glutathione in liver. Both metabolites are related to mangiferin, one of the most important bioactive compounds from mango peel. These results suggest that enhancement of the antioxidant status in the liver of diabetic rats is promoted by mango-diet rich in phenolic compounds.

Keywords: ‘Ataulfo’ mango, *Mangifera indica* L, diabetes mellitus, metabolomics, mass spectrometry.

1. INTRODUCTION

Diabetes mellitus (DM) is among the most prevalent chronic diseases in the world, as stated by the International Diabetes Federation (IDF). In 2017, there were about 451 million diagnosed cases and 5 million deaths worldwide (Cho et al., 2018). DM alludes to a group of metabolic disorders characterized by chronic hyperglycemia, resulting from impaired insulin production by pancreatic β-cells and/or insulin resistance by peripheral tissues (Goyal & Jialal, 2018). Genetic, environmental and lifestyle (diet and exercise) factors are among the different causes that promote DM development. Moreover, oxidative stress has been strongly associated with DM pathogenesis and progression of its comorbidities.

Recent studies have reported that vegetable byproducts, which are often discarded in industrial processing (e.g. bark, peel, leaves, roots, seeds, flowers, etc.), are potential sources of bioactive compounds that can exert antidiabetic and antioxidant actions, similar to that of commercially-available drugs (Naveen & Baskaran, 2018). Fenugreek, avocado, garlic, pomegranate, grapes, guava and papaya are crops whose components have been extensively studied as antidiabetics, mainly due to their powerful antioxidant activity and hypoglycemic effects, as proven in animal models and clinical trials (Beidokhti, ethnopharmacology, & 2017, n.d.; Naveen & Baskaran, 2018). Among these crops, mango (*Mangifera indica* L.) cv. 'Ataulfo' is an excellent source of nutritional (vitamins, minerals, dietary fiber) and bioactive compounds (carotenoids and phenolic compounds) with functional properties (Palafox-Carlos, Yahia, & González-Aguilar, 2012).

Mango consumption, or that of its main bioactive compounds, has been linked to antimutagenic, anti-inflammatory, immunomodulatory, antioxidant, hypoglycemic and antidiabetic effects (Martin & He, 2009). Previous studies have been carried out to validate the effectiveness of consuming mango peel extract to modulate oxidative stress and ameliorate various biochemical parameters in streptozotocin (STZ)-induced diabetic rats, which takes place by different mechanisms like improving serum glucose uptake and catabolism, increasing antioxidant system activity and inhibiting digestive starch-hydrolyzing enzymes, among others (Gandhi et al., 2014; Gondi & Rao, 2015; Sellamuthu, Arulselvan, Muniappan, Fakurazi, & Kandasamy, 2013).

There is substantial evidence correlating diabetic complications to disrupted common metabolic pathways. In this way, different metabolites, such as some lipids, monosaccharides and amino acids have been identified as altered in individuals with DM (Sas, Karnovsky, Michailidis, & Pennathur, 2015). These metabolites could be used as promising biomarkers that could lead to improving early diagnosis and generate novel and effective treatments against the disease. They have been identified thanks to the recent development of metabolomic approaches, which aim to study a substantial number of low molecular weight compounds present in biological systems (Agin et al., 2016). There are also more specific subfields of metabolomics, such as nutrimetabolomics, which aims to identify food-related metabolites that can be correlated to certain beneficial health effects. These studies focus on evaluating the metabolic impact of a specific compound, diet, food or nutrient (Ulaszewska et al., 2019). Due to the promising beneficial effects of mango consumption in DM, the aim of the present study is to evaluate metabolic changes in serum and liver after a prolonged intake of bioactive compounds from 'Ataulfo' mango peel and pulp in STZ-induced diabetic Wistar rats. Our data will contribute to identify potential changes to diabetes-related biomarkers after consumption of a mango extract rich in bioactive compounds.

2. MATERIAL AND METHODS

2.1 Animals and samples

All experiments involving living organisms were reviewed and approved by the Bioethics Committee of the Research Center for Food and Development, where they were performed (CE/013/2018). Animals were cared for and manipulated according to applicable local and international rules and regulations, such as the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health and the Mexican NOM-062-ZOO-1999. Twenty-six male Wistar rats (280 ± 40 g initial weight) were obtained from the University of Sonora, Mexico, and housed in individual ventilated metal cages under standard conditions (12 h light/dark cycles at 24±1 °C). After seven days of acclimatization with free access to food and water, rats were randomly divided into three groups: healthy control group (HC, n=7), untreated diabetic group (UD, n=6) and mango-treated diabetic group (MTD, n=13). Diabetes was induced on UD and MTD groups under overnight fasted conditions by an intraperitoneal dose of STZ (60 mg/kg body weight) dissolved in 0.9% NaCl. Three days after induction, rats with glycemia ≥200 mg/dL were considered diabetic and used for the experiment. Initial and final glycemia and body weight gain were registered (**Table 1S**).

HC and UD groups were fed a diet based on the standard A5001 diet; MTD group was fed a diet supplemented with 5% and 10% of lyophilized 'Ataulfo' mango peel and pulp, respectively. Diets were isoenergetic (3.5 kcal/g), with a macro- and micronutrient content adequate for rodents (**Table 2S**). The phytochemical content of mango peel and flesh has been reported in previous papers (see **Supplemental Material**) (Pacheco-Ordaz, Antunes-Ricardo, Gutiérrez-Uribe, & González-Aguilar, 2018; Quirós-Sauceda et al., 2017; Velderrain-Rodríguez et al., 2018).

Animals were maintained for five weeks after diabetes induction, during which food and water were freely available and replenished daily. After this period, they were fasted overnight (10 h), anesthetized with a single intraperitoneal dose of sodium pentobarbital (120 mg/kg body weight; Pisabental, PISA Agropecuaria, Atitalaquia, Hidalgo, Mexico) and euthanized. Whole blood samples were collected in gold top serum separator Vacutainer tubes (Becton-Dickinson, Franklin Lakes, NJ, USA) to obtain serum, while liver samples (0.5 g) were lyophilized. Both types of samples were stored at -80°C until processed.

2.2 Sample treatment

Biological samples were thawed on ice before analysis. A quality control sample (QC) was prepared for each matrix by combining equal amounts of each case-study sample. This QC sample was treated like the study samples as described below.

In the case of serum, proteins were precipitated with a mixture of organic solvents (methanol:ethanol, 50:50 v/v) in a 1:2 ratio. The mixture was kept during 30 min at -20 °C and then centrifuged (14800 rpm, 4 °C, 10 min). Supernatant (200 µL) was evaporated to dryness. Afterwards, the dry residue was reconstituted in 100 µL of H_2O:methanol: (95:5, v/v) and centrifuged to collect an aliquot for analysis in HPLC vials. Samples were stored at -80 °C, until analysis.

Regarding liver samples, 100 mg of lyophilized tissue were weighed and mixed with 500 µL of methanol. The mixture was vortex-mixed and then introduced into an ultrasonic bath for 3 min under refrigerated conditions. The mixture was kept during 30 min at -20 °C and then centrifuged (14800 rpm, 4 °C, 10 min). The supernatant was separated, and the extraction process was repeated two more times using 250 µL of methanol:H_2O (80:20, v/v). Obtained supernatants were combined, and a 700 µL aliquot was evaporated to dryness. Finally, the residue was reconstituted in 150 µL of methanol:H_2O (35:65, v/v), taking an aliquot in an HPLC vial and stored at -80 °C, until analysis.

2.3 HPLC-MS analysis

Both biological samples were analyzed with the same HPLC-ESI-QTOF-MS methodology. Metabolites were separated using an Agilent Zorbax Eclipse Plus column (3.5 µm particle size, 150 mm x 2.1 mm). The column was maintained at 25 ºC in an Agilent 1260 HPLC instrument. Mobile phases were water containing 0.1% formic acid (A) and methanol (B), used with a flow rate of 0.4 mL/min with the following gradient: 0 min, 5 % B; 5 min, 10 % B; 15 min, 85 % B; 32-40 min, 100 % B; 45-50 min, 5 % B. Injection volume was 5 µL, and for avoiding possible degradation, samples were maintained at 4 °C in the auto-sampler compartment. The QC sample was repeatedly injected every five study samples throughout the analytical sequence.

MS analyses were performed using an Agilent 6540 UHD Accurate Mass Q-TOF analyzer. Data was acquired in negative-ion mode over a 50 to 1700 *m/z* range. For identification purposes, MS/MS analyses were performed in QC samples with different collision energies (10, 20 and 40 eV). Ultrahigh purity nitrogen was used as drying (200 ºC, 10 L/min) and nebulizer gas (350 ºC, 12 L/min). All masses were calibrated by means of continuous infusion of the following substances: trifluoroacetate anion ($C_2F_3O_2^-$, *m/z* 112.985587) and an adduct of HP-921 (*m/z* 1033.988109). Both reference ions provided accurate mass measurements, typically better than 2 ppm. The analytical methodology is described in detail elsewhere (Fernández-Ochoa et al., 2019).

2.4 Data processing and statistical analysis

Batch Recursive Feature Extraction for small molecules was performed using MassHunter Profinder software (B.06.00, Agilent Technologies). Parameters for peak picking were an intensity threshold of 1000 counts, an RT window of ± 0.25 min, a mass window of 20 ppm ± 2 mDa and Agile2 as the integration method. Possible adducts with a maximum charge of 2 were the following: $[M\text{-}H]^-$, $[M+Cl]^-$, $[M\text{-}H_2O^-H]^-$ and $[M+HCOO]^-$.

Regarding statistical analyses, Principal Component Analysis (PCA) was performed first to check the reproducibility and detect possible outliers. In order to explore the metabolic differences in relation to the different study groups, multivariate statistical test (Partial Least Squares Discriminant Analysis (PLS-DA), and a hierarchical clustering via heatmap) and univariate statistical tests (ANOVA) were performed. Logarithmic transformation and Pareto scaling were applied to the data before performing multivariate analyses. All statistical tests were carried out in MetaboAnalyst 4.0 software (Chong et al., 2018).

2.5 Metabolite identification

Identification was carried out by comparing accurate mass, isotopic distribution and fragmentation patterns obtained in MS/MS analysis with available online metabolomic databases. Databases were searched by CEU Mass Mediator tool (Gil de la Fuente, Grace Armitage, Otero, Barbas, & Godzien, 2017). This tool allowed simultaneous metabolite search in several databases such as METLIN, LipidMaps, KEGG as well as Human Metabolome Database. The MS/MS patterns were also compared with *in silico* MS/MS fragmentation resources, concretely, MetFrag (https://ipb-halle.github.io/MetFrag/).

3. RESULTS

3.1 Data quality assessment

After data pre-processing, 770 and 508 molecular features were extracted from liver and serum samples, respectively. After that, 19 and 32 features were discarded in both datasets, respectively, due to a Relative Standard Deviation (RSD) in QC samples higher than 30 %. PCA was performed with the purpose of detecting outliers and checking analytical reproducibility (Ulaszewska et al., 2019). In this sense, well grouping of the QC samples in the PCA scores plot (see **Figure 1S**, supplemental material) indicated a good reproducibility of the obtained data. However, the first main component (PC1) of both models explained great differences between specific samples (one serum and three liver samples) and the rest of samples. This fact implied that the separated samples were assigned as outliers, and, therefore, were discarded for the rest of the statistical analyses. Despite this, differences between groups (HC, UD and MTD) were detected thanks to the information explained by the second component (PC2) in both

PCA analyses. These groupings were more clearly observed in the PCA of the liver samples. Therefore, PCA analyses were performed again without the outliers and QC samples (**Figure 1**). As mentioned above, higher differences between HC, UD and MTD groups were observed in the case of liver samples. The variability explained by PC1 (46.2 %) showed a clear separation between all groups (**Figure 1a**). Regarding serum samples, no clear separations between the three groups were observed as in the case of liver samples, however, a clear separation between diabetic and healthy rats was detected, as well as a good grouping of the HC samples (**Figure 1b**).

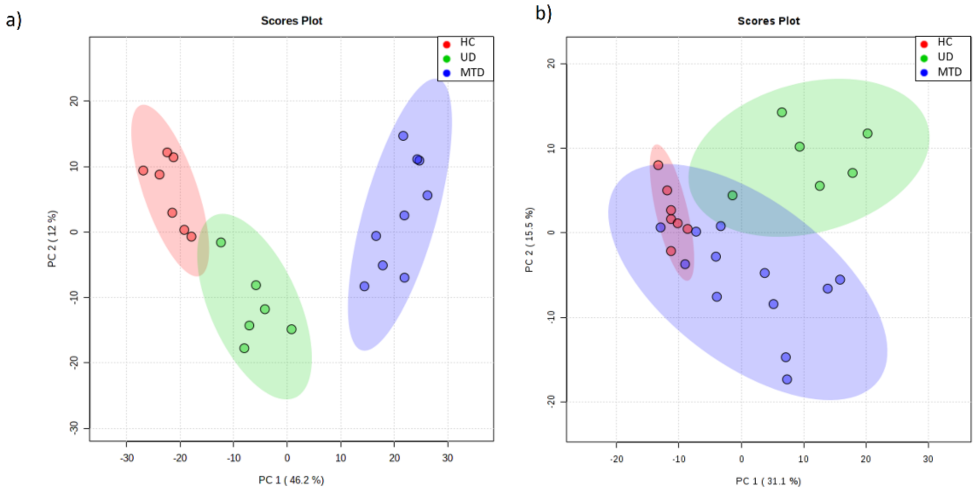


Figure 1. PCA scores plots from data obtained in the analyses from liver (a) and serum samples (b) after the removal of outliers and QC samples. (red dots, healthy controls, HC; green dots, untreated diabetic group, UD; blue dots, mango-treated diabetic group, MTD).

3.2. PLS-DA models and Analysis of Variance (ANOVA)

PLS-DA models were constructed with data from serum and liver samples, separately, in order to discriminate samples according to different classes (HC, UD and MTD). **Figure 2** shows the main results of both models. Analogous to PCA results, the three classes were better separated by the PLS-DA model created with liver data. Clearly, all three classes were separated by PC1 (38.9 %) in the PLS-DA model. On the other hand, PC1 (25%) in the model created with serum data separated the healthy controls and diabetic samples. In addition, PC2 (18.1%) differentiated rats from the UD group, to those that consumed mango-treated diets (MTD).

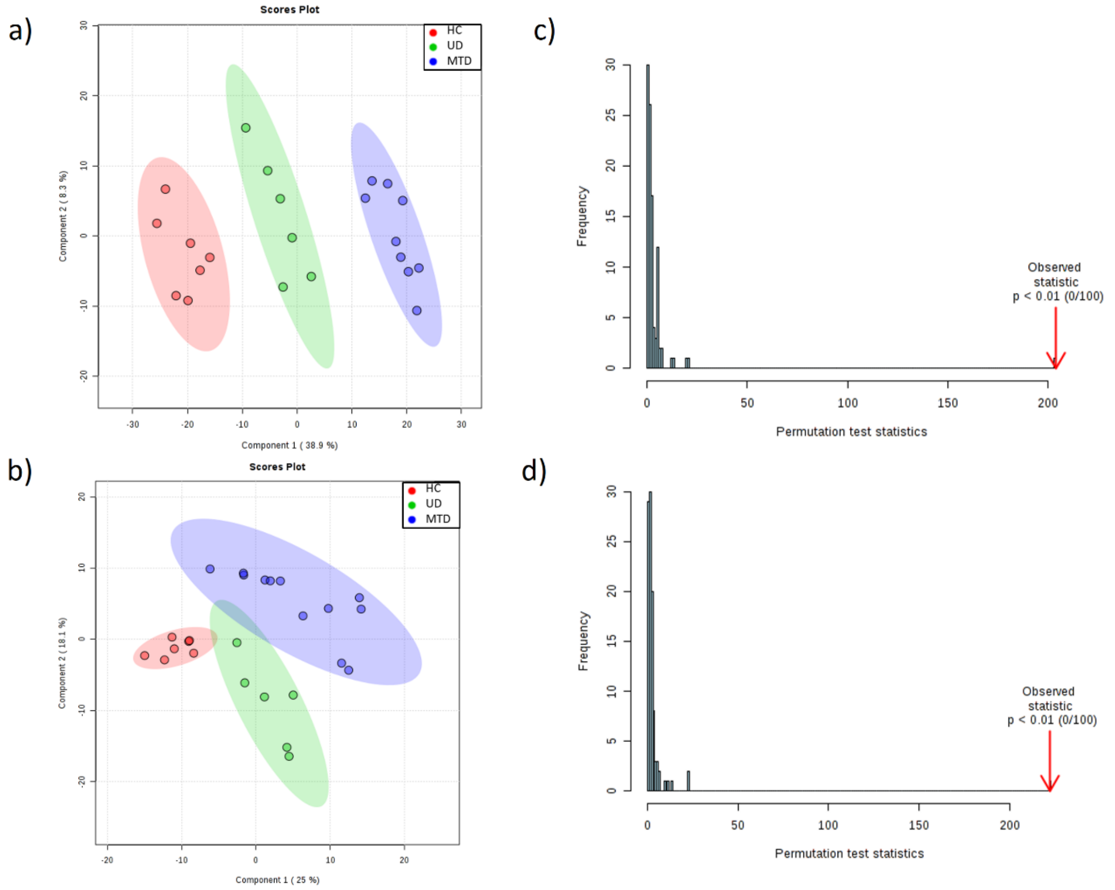


Figure 2. A Supervised Partial Least Squares Discriminant Analyses (PLS-DA). (a,b): PLS-DA score plot (2a: liver model; 2b: serum model). (red dots, healthy controls, HC; green dots, untreated diabetic group, UD; blue dots, mango-treated diabetic group, MTD). Permutation test results using separation distance (B/W) (2c: liver model, 2d: serum model).

For model assessment, accuracy, R^2 and Q^2 values were obtained by 10-fold cross validation (CV) method for the PLS-DA models with two components. Analogous to the sample distribution observed in score plots, better results were obtained with the model created with liver data (Accuracy: 0.95, R^2: 0.98 and Q^2: 0.95), as compared to the model with serum data (Accuracy: 0.92, R^2: 0.84 and Q^2: 0.64). Permutation tests were also performed, in which p-values lower than 0.01 were obtained in both models (**Figures 2c**, **2d**), indicating a lack of overfitting in these PLS-DA models. Molecular features with VIP values higher than 1.50 were selected for identification as potential biomarkers, since these were responsible for most differences found between groups in the models. Furthermore, analysis of variance (ANOVA) was applied to these features, and those that were not significant in this statistical test were discarded for identification. In this sense, 51 and 39 molecular features from liver and serum, respectively, were selected for identification. As a result of identification, 26 and 29 metabolites for serum and liver, respectively, could be annotated (**Table 1** and **Table 2)**. According to the Metabolomics Standards Initiative (MSI), proposed metabolites were putatively annotated (Sumner et al., 2007) using the MS/MS fragments (**Tables 3S-4S**). Regarding unknown features, **Tables 5S-6S** show their corresponding parameters.

Table 1. Retention times, masses, VIP, FDR and the results of the Tukey's HSD from annotated metabolites present in serum samples.

RT (min)	Mass (Da)	VIP value	FDR	Tukey's HSD	Molecular Formula	Score (%)	Metabolite
22.58	302.2258	3.86	2.85E-08	T-C; T-D	$C_{20}H_{30}O_2$	97.82	Eicosapentanoic acid (EPA)
19.19	318.2227	3.56	2.57E-08	T-C; T-D	$C_{20}H_{30}O_3$	95.54	12-HEPE
17.95	465.3130	3.16	5.76E-03	D-C; T-C	$C_{26}H_{43}NO_6$	97.27	Glycocholic acid
1.5	270.0931	2.93	9.12E-07	D-C; T-C	$C_{10}H_{14}N_4O_5$	87.3	Histidinyl-Aspartate (His-Asp)
17.71	408.294	2.62	3.19E-02	T-C	$C_{24}H_{40}O_5$	76.36	Cholic acid
1.42	279.1320	2.56	2.69E-03	D-C; T-C	$C_{11}H_{21}NO_7$	58.12	N-(1-Deoxy-1-fructosyl)valine
24.58	330.2550	2.55	3.77E-07	T-C; T-D	$C_{22}H_{34}O_2$	97.93	Docosapentaenoic acid (DPA)
12.71	284.0918	2.55	2.52E-02	T-C	$C_{13}H_{16}O_7$	91.16	p-Cresol glucuronide
23.71	678.4681	2.28	1.80E-05	D-C; T-C; T-D	$C_{39}H_{67}O_7P$	87.31	PA(P-36:5)
12.01	86.0735	2.13	3.86E-04	D-C; T-C	$C_5H_{10}O$	85.06	Iso-valeraldehyde
6.79	118.1311	2.05	6.05E-04	D-C; T-C	$C_5H_{10}O_3$	84.99	3-hydroxyisovaleric acid
23.13	254.2265	2.05	4.81E-03	D-C; T-C	$C_{16}H_{30}O_2$	98.52	Palmitoleic acid
23.14	276.4137	2.02	3.93E-03	D-C; T-C	$C_{18}H_{28}O_2$	80.27	Stearidonic acid
34.7	583.5189	1.94	1.65E-04	D-C; T-C	$C_{34}H_{67}NO_3$	78.75	Cer(d18:1/16:0)
11.37	122.0373	1.91	5.77E-03	D-C; T-C	$C_7H_6O_2$	98.88	4-hydroxybenzaldehyde
25.99	332.2735	1.87	8.95E-04	T-C; T-D	$C_{22}H_{36}O_2$	72.81	Adrenic Acid
1.51	129.0406	1.87	3.00E-03	T-C; T-D	$C_5H_7NO_3$	87.4	Pyroglutamic acid
23.7	396.2277	1.84	1.13E-06	T-C; T-D	$C_{18}H_{37}O_7P$	94.52	PA(15:0)
1.76	129.0419	1.83	4.66E-04	D-C; T-C	$C_5H_7NO_3$	93.43	Pyrrolidonecarboxylic acid
23.71	396.2320	1.77	1.13E-06	D-C; T-C; T-D	$C_{18}H_{37}O_7P$	99.21	(9S,10S)-10-hydroxy-9-(phosphonooxy)octadecanoate
23.72	328.2431	1.77	3.13E-06	D-C; T-C; T-D	$C_{22}H_{32}O_2$	96.02	Docosahexaenoic acid (DHA)
22.7	300.4351	1.71	2.14E-02	T-C	$C_{20}H_{28}O_2$	65.90	Retinoic acid
12.79	173.1054	1.65	4.07E-02	T-C	$C_8H_{15}NO_3$	84.38	Hexanoylglycine
22.71	278.2261	1.57	2.56E-02	T-C	$C_{18}H_{30}O_2$	76.03	Linolenic acid
21.17	567.3359	1.54	4.81E-03	D-C; T-C	$C_{30}H_{50}NO_7P$	92.64	LysoPC(22:6)
23.78	654.4694	1.53	1.11E-04	T-C; T-D	$C_{44}H_{62}O_4$	72.04	14-DHAHDHA

Table 2. Retention times, masses, VIP, FDR and the results of the Tukey’s HSD from annotated metabolites present in liver samples.

RT (min)	Mass (Da)	VIP value	FDR	Tukey’s HSD	Molecular Formula	Score (%)	Metabolite
7.61	358.1014	3.54	1.04E-10	T-C, T-D	$C_{11}H_{23}N_2O_7PS$	91.76	Pantetheine 4’’-phosphate
17.34	228.0429	3.53	5.00E-17	T-C, T-D	$C_{13}H_8O_4$	95.11	Euxanthone
34.15	825.5523	3.38	6.74E-25	T-C, T-D	$C_{45}H_{80}NO_{10}P$	96.77	PS(39:4)
15.79	320.1476	3.35	3.26E-07	D-C, T-C	$C_{14}H_{28}O_8$	97.85	Octanoylglucoronide
1.47	345.0379	3.24	2.23E-16	T-C, T-D	$C_{10}H_{12}N_5O_7P$	53.27	Cyclic GMP
1.47	307.0847	3.21	1.56E-06	T-C, T-D	$C_{10}H_{17}N_3O_6S$	92.63	Glutathione
3.05	216.1133	1.68	3.04E-07	D-C, T-C, T-D	$C_9H_{16}N_2O_4$	86.73	Pro-Thr
7.25	282.1076	1.64	1.58E-09	D-C, T-C, T-D	$C_{10}H_{14}N_6O_4$	88.44	2-Aminoadenosine
2.75	176.0691	1.59	1.88E-06	D-C, T-C, T-D	$C_7H_{12}O_5$	78.56	2-Isopropylmalic acid
27.94	676.5406	1.57	7.57E-07	D-C, T-C, T-D	$C_{38}H_{77}O_7P$	97.54	PA(O-18:0/17:0)
1.41	347.0596	1.57	3.45E-08	D-C, T-C, T-D	$C_{10}H_{14}N_5O_7P$	99.15	Adenosine monophosphate
3.28	267.0994	1.55	1.27E-06	D-C, T-C, T-D	$C_{10}H_{13}N_5O_4$	66.33	Adenosine
7.25	89.0468	1.54	1.27E-06	D-C, T-C, T-D	$C_3H_7NO_2$	94.58	beta-Alanine
25.08	330.255	1.54	1.51E-06	D-C, T-C, T-D	$C_{22}H_{34}O_2$	76.84	Docosapentaenoic acid (DPA)
12.79	173.1059	1.54	2.58E-06	D-C, T-C, T-D	$C_8H_{15}NO_3$	86.99	N-Acetylisoleucine
14.39	172.1103	1.54	1.27E-06	D-C, T-C, T-D	$C_9H_{16}O_3$	99.9	9-oxo-nonanoic acid
5.43	327.1332	1.53	2.58E-06	D-C, T-C, T-D	$C_{15}H_{21}NO_7$	97.34	N-(1-Deoxy-1-fructosyl)phenylalanine
2.23	343.1256	1.53	1.90E-06	D-C, T-C, T-D	$C_{15}H_{21}NO_8$	70.01	N-(1-Deoxy-1-fructosyl)tyrosine
2.38	181.0721	1.53	2.84E-06	D-C, T-C, T-D	$C_9H_{11}NO_3$	70.05	L-Tyrosine
2.64	293.1509	1.52	2.08E-06	D-C, T-C, T-D	$C_{12}H_{23}NO_7$	70.28	N-(1-Deoxy-1-fructosyl)leucine/N-(1-Deoxy-1-fructosyl)isoleucine
1.53	541.0576	1.52	6.47E-09	T-C, T-D	$C_{15}H_{21}N_5O_{13}P_2$	91.89	Cyclic ADP-ribose
22.67	521.3481	1.51	7.83E-07	T-C, T-D	$C_{26}H_{52}NO_7P$	89.79	LysoPC(18:1)
5.42	237.1027	1.51	5.48E-06	D-C, T-C, T-D	$C_{12}H_{15}NO_4$	97.49	N-lactoyl-phenylalanine
2.64	203.1178	1.41	2.08E-06	T-C, T-D	$C_9H_{17}NO_4$	91.58	N-lactoyl-Leucine
2.39	131.095	1.50	5.78E-06	D-C, T-C, T-D	$C_6H_{13}NO_2$	76.63	Beta-Leucine
9.72	366.1447	1.50	7.72E-06	D-C, T-C, T-D	$C_{17}H_{22}N_2O_7$	84.06	N-(1-Deoxy-1-fructosyl)tryptophan
22.67	507.3321	1.50	9.84E-07	T-C, T-D	$C_{25}H_{50}NO_7P$	94.44	LysoPE(20:1)
6.41	284.0768	1.50	5.87E-06	D-C, T-C, T-D	$C_{10}H_{12}N_4O_6$	89.96	Xanthosine
7.25	219.1161	1.50	7.49E-06	D-C, T-C, T-D	$C_9H_{17}NO_5$	82.88	Pantothenic acid

3.3. Hierarchical clustering analysis via heatmap

Hierarchical clustering analyses via heatmap were performed to the annotated metabolites using a Pearson distance measure and Ward clustering algorithm. **Figure 3** shows results obtained for these analyses. In both models, sample clustering shows three clear clusters (S1, S2 and S3), corresponding to the three classes of samples (MTD, HC and UD), respectively. Regarding variable clustering, different clusters are clearly observed. In liver model, four clusters (L1, L2, L3 and L4) were assigned (**Figure 3a**). L1 and L2 clusters were clearly influenced by the mango-supplemented diet, while L3 and L4 seemed to be more related to diabetes. Regarding serum clustering (**Figure 3b**), three clusters were assigned (P1, P2 and P3). The first two were composed of altered metabolites due to diet, while the third cluster corresponded to metabolites related to diabetes.

4. DISCUSSION

A detailed understanding of the pathophysiology of DM by identifying its metabolic alterations is imperative for early diagnosis and the development of possible preventive strategies and effective treatments. In the present study, we attempted to inspect the metabolic profile of serum and liver of STZ-induced diabetic rats using HPLC-ESI-QTOF-MS, and explore the alterations generated by a phenolic-supplemented diet. The most affected pathways by the pathology were related to fatty acid metabolism, bile acid homeostasis and amino acid catabolism, while metabolites enhanced by the mango treatment were associated with the antioxidant system.

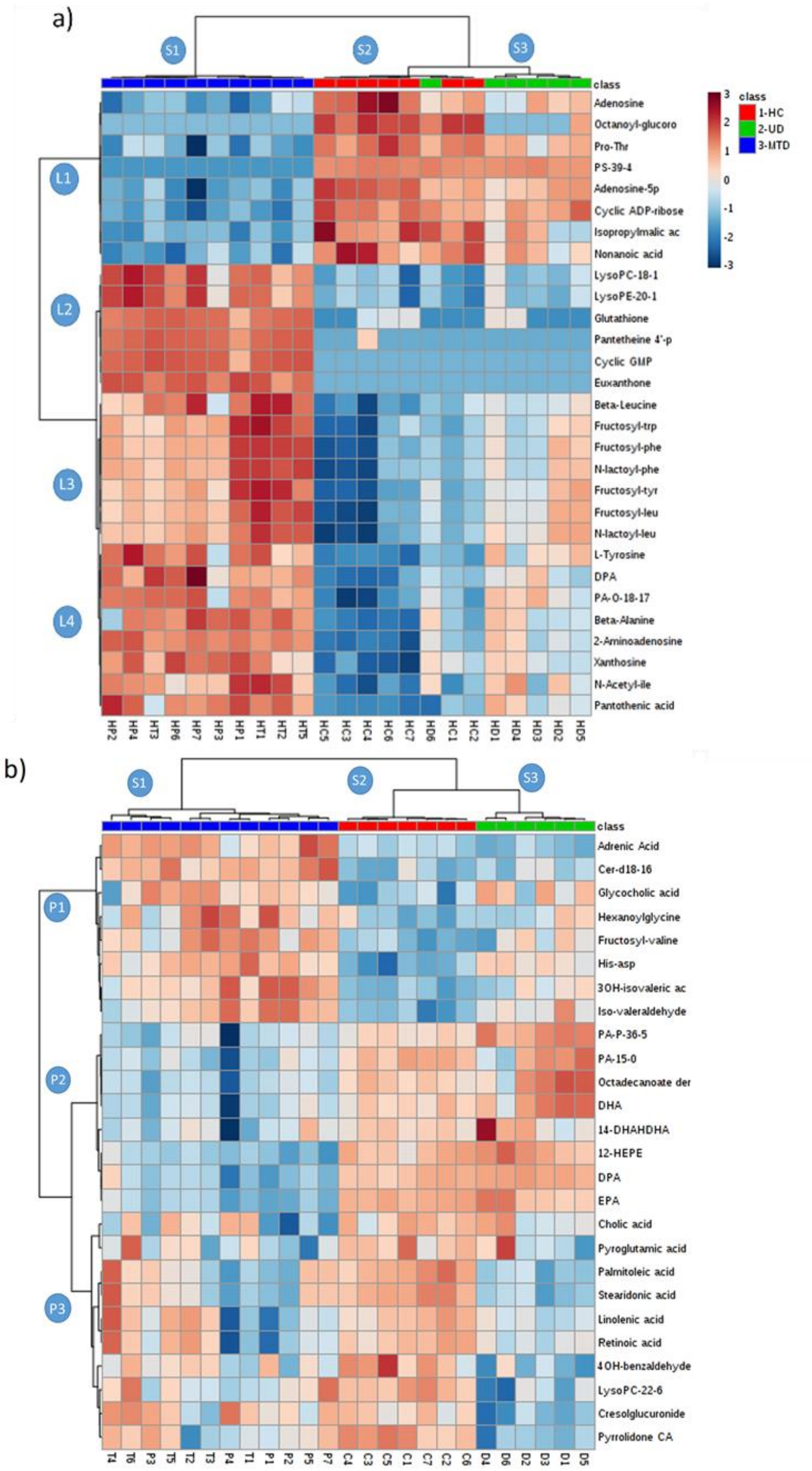


Figure 3. Hierarchical clustering via heatmap using Pearson as distance measure and Ward as clustering algorithm, of the significant annotated metabolites. (3a: liver model; 3b: serum model.

4.1 *Metabolic changes related to molecular mechanisms in response to mango-supplemented diet.*

The effects of mango peel and pulp intake on the serum and liver metabolome in STZ-induced rats were studied. The major findings associated to diet consumption were the presence of euxanthone and glutathione, molecules that are related to the enhancement of the antioxidant status of the liver of diabetic rats, among other biological properties, which are discussed in further detail below.

4.1.1 *Euxanthone.*

Euxanthone was only detected and positively identified in mango-treated group (L2). Mango contains numerous bioactives, however, the only metabolite that bioaccumulated/biotransformed in liver was euxanthone. This is a metabolite from mangiferin, a bioactive compound found in mango and mangosteen (*Garcinia mangostana* Linn.) (Pedraza-Chaverri, Cárdenas-Rodríguez, Orozco-Ibarra, & Pérez-Rojas, 2008). When mangiferin is orally consumed, intestinal bacteria first deglycosylate it to norathyriol, leaving the core phenolic scaffolding intact (Sanugul et al., 2005). Norathyriol is then further metabolized into euxanthone by removing two hydroxyl groups at positions 3 and 6 (Derese, Guantai, Yaouba, & Kuete, 2017). Mangiferin and its metabolites are of high interest due to numerous health-promoting effects documented, such as antioxidant, anti-inflammatory, antidiabetic and others (Martin & He, 2009).

It has been shown that mangiferin is able of modulate different signaling molecules including mitogen-activated protein kinases (MAPKs), nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB) and protein kinase C (PKC) isoforms (Saha, Sadhukhan, & Sil, 2016), the latter related to the non-enzymatic formation of advanced glycation end products (AGEs) in hyperglycemic organisms (Das Evcimen & King, 2007). Additionally, mangiferin suppresses some pro-inflammatory cytokines (Saha et al., 2016) and inhibits enzymes associated with carbohydrate metabolism such as α-amylase and α-glucosidase (Sekar, Chakraborty, Mani, Sali, & Vasanthi, 2019), exerting beneficial effects in STZ-induced diabetic rats (Gondi & Rao, 2015).

Our results suggest that euxanthone is bioaccumulated in the liver of diabetic rats after consuming a diet rich in 'Ataulfo' mango peel and pulp for a period of five weeks, possibly contributing to the anti-diabetic effects of this diet.

4.1.2 *Glutathione*

Glutathione was found in higher concentration in the liver of MTD rats, as compared to HC and UD animals. This finding suggests that mango consumption exerted a profound antioxidant effect by stimulating the endogenous antioxidant system. Oxidative stress in diabetic organisms is well documented and strongly correlates with diabetic complications in most affected organs (Tangvarasittichai, 2015). Various authors have found a protective effect of phenolic-rich foods or extracts. For example, *Arafat et al.*, (Arafat et al., 2016) administered *Momordica charantia* fruit pulp to alloxan-induced diabetic rats and found a restored hepatic glutathione concentration. *Abdel-Moneim et al.*, (Abdel-Moneim, Yousef, Abd El-Twab, Abdel Reheim, & Ashour, 2017) report that gallic acid and *p*-coumaric acid (both present in mango pulp) prevent brain glutathione depletion in STZ-induced diabetic rats. Finally, it has been shown that a mangiferin pretreatment increases glutathione levels and enhances the activity of glutathione-related enzymes in heart tissue of rats (Prabhu, Jainu, Sabitha, & Devi, 2006), which could be related to the findings of the present work. Therefore, our data suggests that mango supplementation apparently prevents glutathione depletion in livers of diabetic rats. This is likely due to the many phenolic compounds present in 'Ataulfo' mango peel (mangiferin) and pulp (gallic acid, chlorogenic acid, protocatechuic acid and vanillic acid), which countered some of the pro-oxidative effects of diabetes.

4.2 *Metabolic changes related to molecular pathophysiological mechanisms of diabetes*

Further to the increased hyperglycemia (**Table 1S**), changes in serum and liver metabolites were mainly related to alterations of fatty acids, bile acids, amino acids, pantothenate, and nucleotide metabolism. The above-mentioned changes are consistent with some known disrupted mechanisms linked to DM and its complications in STZ-induced rat models.

4.2.1 *Fatty acid metabolism*

Most changes documented in serum were related to fatty acids, their metabolites and phospholipids, indicating that lipid metabolism was significantly altered. Interestingly, changes were found on monounsaturated fatty acids (MUFAs) and polyunsaturated fatty acids (PUFAs).

Therefore, palmitoleic acid decreased in both diabetic groups, as compared to the control. This MUFA is a lipokine that can promote peripheral insulin sensitivity in muscle cells via p-38 MAPK signaling (Talbot, Wheeler-Jones, & Cleasby, 2014), and in animal models by up-regulating expression of carbohydrate- and lipid-catabolizing genes (Duckett, Volpi-Lagreca, Alende, & Long, 2014). Furthermore, linolenic acid and stearidonic acid were also significantly altered, with a decreasing tendency in both diabetic groups, while eicosapentaenoic acid (EPA) diminished in the MTD group. These ω-3 fatty acids have shown bioactivities related to insulin resistance, and their consumption has been explored to be useful in type 2 diabetes (Jovanovski et al., 2017).

Adrenic acid, docosapentaenoic acid (DPA), docosahexaenoic acid (DHA) and 14-(docosahexaenoyloxy) docosahexaenoic acid (14-DHADHA) also showed significant changes. Increased levels of adrenic acid have been correlated with decreased insulin sensitivity and with increased area under the glucose curve in a type 2 diabetic male population (Lankinen et al., 2015). The effects of DHA have been shown to oppose those of previously-mentioned ω-6 fatty acids, for example, dietary supplementation with ω-3 fatty acids (including DHA) can improve some endocrine and anthropometric parameters in a type-2 diabetic population (Jacobo-Cejudo et al., 2017). In addition, branched fatty acid esters of hydroxy fatty acids (FAHFAs) such as 14-DHADHA has been reported to exert anti-inflammatory and anti-diabetic properties mainly in type 2 diabetes in humans and rats (Balas, Durand, & Feillet-Coudray, 2018).

Taken together, these results showed an increased level of pro-inflammatory ω-6, concurrent with a decreased level of anti-inflammatory ω-3 fatty acids. An increased ω-6-to-ω-3 ratio has been correlated with cardiac events (Takahashi et al., 2017), suggesting that dietary supplementation of ω-3 could be beneficial in diabetic organisms.

In addition to free fatty acids, phospholipids [PA(P-36:5), PA(15:0/0:0), LPC(22:6)] and sphingolipids [ceramide (d18:1/16:0)] were also significantly altered in serum of diabetic rats. The behavior of these compounds (see **Figure 3**) suggests that phospholipase activity increased in the absence of insulin (Lin et al., 2016). This fact has also been found to be related to oxidative stress (Yui et al., 2015). Both phospholipids and sphingolipids are key metabolic mediators that modulate intracellular signaling related to insulin, as well as numerous other processes, which makes them important targets to prevent, mitigate or revert metabolic anomalies (Meikle & Summers, 2017).

4.2.2 *Bile acids*

Significant changes were documented on the serum concentration of cholic acid and glycocholic acid. The upregulation of glycocholic acid has been considered a circulating metabolomic biomarker of the disease. The regulation of hepatic bile acid metabolism suggests the prevention of insulin resistance as observed in previous studies. In fact, a role as a ligand of the farnesoid X receptor (FXR), also known as bile acid receptor, has been assigned to bile acids (Cipriani, Mencarelli, Palladino, & Fiorucci, 2010; Renga, Mencarelli, Vavassori, Brancaleone, & Fiorucci, 2010). FXR is a transcription factor that, among other functions, stimulates insulin transcription and secretion from pancreatic β-cells.

4.2.3 *Branched-chain amino acids (BCAAs) and aromatic amino acids metabolism*

Main changes in liver amino acids were related to branched and aromatic. BCAAs and some of their derivatives like leucine, N-(1-deoxy-1-fructosyl)-leucine, N-lactoyl-leucine and N-(1-deoxy-1-fructosyl)-isoleucine were all increased in the liver of diabetic groups (UD and MTD), whereas, N-(1-deoxy-1-fructosyl)-valine was increased in serum, demonstrating an impaired amino acid metabolic control as revealed in one-week STZ-induced diabetic rats (Diao et al., 2014).

It is well known that STZ-diabetic induced rats are susceptible to increased protein catabolism in tissues like skeletal muscle, which leads to an alteration in diverse amino acid metabolic pathways, increasing circulating BCAA levels (Rodríguez et al., 1997). Since the liver is one of the principal tissues where BCAA oxidation takes place (60% to 83%), it is possible that excess of α-ketoacids produced from BCAA in muscles by proteolysis in diabetic organisms, are metabolized here and modulated by the branched-chain α-keto acid dehydrogenase complex (BCKDH) in this organ. This would lead to the formation of acetyl-CoA or succinyl-CoA, which can enter the tricarboxylic acid cycle (TCA); once there , they can be converted into pyruvate or any gluconeogenic TCA intermediate (Neinast et al., 2019). However, hepatic BCKDH activity has been shown to decrease in diabetic rat models (Doisaki et al., 2010; Kuzuya et al., 2008), which could increase hepatic BCAA concentration. We also found a significant increase in alanine levels (tentatively produced by BCAA transamination with pyruvate), which is related to the glucose-alanine cycle from muscle to liver, which apparently supply gluconeogenic substrates due to a lack of inhibitory insulin signaling (Suhre et al., 2010).

Finally, our results show the upregulation of three aromatic amino acids or its fructosyl derivatives (phenylalanine, tyrosine, and tryptophan) in both diabetic groups (UD and MTD), which is consistent with data reported by *Lanza et al.* (Lanza et al., 2010). Changes in aromatic amino acids have been associated with perturbations in gut microbiota (Neis, Dejong, & Rensen, 2015), which has been reported as altered in patients with impaired glucose metabolism (Chen et al., 2016). In diabetic organisms, AGEs are formed when reducing sugars (such as glucose and fructose) spontaneously react with proteins, which has been linked to the development of diabetic neuropathy, nephropathy and retinopathy, among other complications (Karachalias, Babaei-Jadidi, Ahmed, & Thornalley, 2003).

As the concentration of fructose in serum is much lower than that of glucose, this molecule has not been a central focus on glycation research. However, it is more reactive than glucose, and has been reported to be elevated via polyol pathway activation in diabetic tissues (Gallagher, LeRoith, Stasinopoulos, Zelenko, & Shiloach, 2016).

Non-enzymatic reactions between reducing sugars and amines depend on their circulating levels, promoting the spontaneous formation of early glycation products (EGPs), including N-terminal fructosyl compounds, which are precursors of AGEs (Beisswenger, 2012). As seen in cluster L3 of the heatmap, our results show that all BCAAs and aromatic amino acids in liver and serum of diabetic rats (UD and MTD) were glycated with fructosamine residues (L3). The AGE levels and N-fructosyl-lysine were described to be increased in retinal, renal glomeruli, sciatic nerve and serum proteins in STZ-induced diabetic rats, which was related to high glycemia (Karachalias, Babaei-Jadidi, Rabbani, & Thornalley, 2010). In previous study, increased glycated hemoglobin (HbA_1 = 17.7 ± 1.6%) was associated with the presence of N-fructosyl-valine in diabetic rats after a 24-week period (Karachalias et al., 2003). We can argue that in organisms with altered glucose homeostasis, glycation reactions of amino acids and other molecules increase, causing protein dysfunction and disruption in amino acid metabolism.

4.2.4 *Pantothenate and CoA biosynthesis*

According to our results, different precursors of CoA showed changes between groups. Pantothenic acid levels were higher in the liver of both diabetic groups, as compared to the control (L4), while pantetheine 4’’-phosphate was upregulated in the MTD group (L2), suggesting an increased synthesis of CoA with predicted impact on acetyl-CoA production. An increased generation of acetyl-CoA is linked to excessive fatty acid oxidation in the liver of fasted and diabetic rats (due to the lack of carbohydrate utilization), which are metabolized in TCA or HMG-CoA pathways to produce ATP or ketone bodies, respectively (McGarry & Foster, 1980; Wieland, Weiss, & Eger-Neufeldt, 1964).

4.2.5 *Nucleotide metabolism*

This study showed a decreased in adenine and AMP in the liver of both diabetic groups (UD and MTD), while there was an increase of 2-aminoadenosine and xanthosine in the liver of both diabetic groups, as compared to the HC (L1, and L4, respectively). Meanwhile, cGMP was upregulated in the MTD group (L2).

Nucleotides play a pivotal role in most metabolic pathways, acting as part of nucleic acids, coenzymes and signaling molecules. It has been reported that adenosine, AMP, guanosine, GMP, GTP, inosine and IMP are increased in type 1 and type 2 diabetes (Dudzinska, 2014; Huang et al., 2011). Incremented nucleoside levels could be a result of an accelerated nucleotide degradation under diabetic conditions, changes in the enzymes and transporters responsible for their metabolism, and alterations at the genetic level (Dudzinska, 2014). Xanthosine elevation could suggest an accelerated breakdown of the xanthine in purine metabolism, which is associated with oxidative stress in the liver of diabetic rats (Kristal, Vigneau-Callahan, Moskowitz, & Matson, 1999). Regarding the increment of cGMP in MTD group, there are several studies that relate the effect of phenolic compounds, mainly on the prevention and treatment of cardiovascular diseases and hypertension, which are linked to the increment of the endothelial nitric oxide (NO) and vasorelaxation via cGMP pathway (Benito et al., 2002). cGMP is a second messenger that promotes glucose-stimulated insulin secretion, prevents β-cells apoptosis and promote its differentiation, however, modulation of guanosine metabolites remains unclear in diabetic organisms.

5. CONCLUSIONS

In this study, potential biomarkers were identified related to diabetic disorders involving significant changes in the fatty acid, amino acid, bile acid, pantothenate and nucleotide pathways. Furthermore, the metabolomic approach found two main metabolites in the liver linked to mango-based dietary intervention, suggesting that some bioactive compounds present therein are bioavailable and can reach target tissues such as liver. These findings also demonstrated that ‘Ataulfo’ mango consumption improves antioxidant status of diabetic rats. These results may contribute to a better understanding of the metabolic changes that take place in diabetic organisms, and the effects of bioactive compounds from mango in STZ-induced diabetic rats. Nonetheless, further research is still needed in order to verify or refute the associations of the possible bioactivity of phenolic compounds in other altered animal tissues or in a nutritional intervention in humans.

Conflict of interest

All authors declare that they have no conflict of interest.

Acknowledgements

The author AFO received support from the Spanish Ministry of Education, Culture and Sports (FPU grant 14/03992). RCC is thankful to CIAD-CONACYT for the scholarship received. This study is considered as part of the collaboration that ASC and GAG have in the network (ALSUB-CYTED, 118RT0543).

REFERENCES

Abdel-Moneim, A., Yousef, A. I., Abd El-Twab, S. M., Abdel Reheim, E. S., & Ashour, M. B. (2017). Gallic acid and p-coumaric acid attenuate type 2 diabetes-induced neurodegeneration in rats. *Metabolic Brain Disease*, *32*(4), 1279–1286. https://doi.org/10.1007/s11011-017-0039-8

Agin, A., Heintz, D., Ruhland, E., Chao de la Barca, J. M., Zumsteg, J., Moal, V., … Namer, I. J. (2016, February). Metabolomics - an overview. From basic principles to potential biomarkers (part 1). *Medecine Nucleaire*. https://doi.org/10.1016/j.mednuc.2015.12.006

Arafat, S. Y., Nayeem, M., Jahan, S., Karim, Z., Reza, H. M., Hossain, M. H., … Alam, M. A. (2016). Ellagic acid rich Momordica charantia fruit pulp supplementation prevented oxidative stress, fibrosis and inflammation in liver of alloxan induced diabetic rats. *Oriental Pharmacy and Experimental Medicine*, *16*(4), 267–278. https://doi.org/10.1007/s13596-016-0242-x

Balas, L., Durand, T., & Feillet-Coudray, C. (2018). Branched FAHFAs, appealing beneficial endogenous fat against obesity and type-2 diabetes. *Chemistry - A European Journal*, *24*(38), 9463–9476. https://doi.org/10.1002/chem.201800853

Beidokhti, M., ethnopharmacology, A. J.-J. of, & 2017, undefined. (n.d.). Review of antidiabetic fruits, vegetables, beverages, oils and spices commonly consumed in the diet. *Elsevier*.

Beisswenger, P. J. (2012, April). Glycation and biomarkers of vascular complications of diabetes. *Amino Acids*. Springer Vienna. https://doi.org/10.1007/s00726-010-0784-z

Benito, S., Lopez, D., Sáiz, M. P., Buxaderas, S., Sánchez, J., Puig-Parellada, P., & Mitjavila, M. T. (2002). A flavonoid-rich diet increases nitric oxide production in rat aorta. *British Journal of Pharmacology*, *135*(4), 910–916. https://doi.org/10.1038/sj.bjp.0704534

Chen, T., Zheng, X., Ma, X., Bao, Y., Ni, Y., Hu, C., … Jia, W. (2016). Tryptophan Predicts the Risk for Future Type 2 Diabetes. *PLoS ONE*, *11*(9), e0162192. https://doi.org/10.1371/journal.pone.0162192

Cho, N. H., Shaw, J. E., Karuranga, S., Huang, Y., da Rocha Fernandes, J. D., Ohlrogge, A. W., & Malanda, B. (2018). IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Research and Clinical Practice*, *138*, 271–281.

Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., … Xia, J. (2018). MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. *Nucleic Acids Research*, *46*(W1), W486–W494. https://doi.org/10.1093/nar/gky310

Cipriani, S., Mencarelli, A., Palladino, G., & Fiorucci, S. (2010). FXR activation reverses insulin resistance and lipid abnormalities and protects against liver steatosis in Zucker (fa / fa) obese rats. *Journal of Lipid Research*, *51*(4), 771–784. https://doi.org/10.1194/jlr.m001602

Das Evcimen, N., & King, G. L. (2007, June). The role of protein kinase C activation and the vascular complications of diabetes. *Pharmacological Research*. https://doi.org/10.1016/j.phrs.2007.04.016

Derese, S., Guantai, E. M., Yaouba, S., & Kuete, V. (2017). Mangifera indica L. (Anacardiaceae). In *Medicinal Spices and Vegetables from Africa: Therapeutic Potential Against Metabolic, Inflammatory, Infectious and Systemic Diseases* (pp. 451–483). Academic Press. https://doi.org/10.1016/B978-0-12-809286-6.00021-2

Diao, C., Zhao, L., Guan, M., Zheng, Y., Chen, M., Yang, Y., … Gao, H. (2014). Systemic and characteristic metabolites in the serum of streptozotocin-induced diabetic rats at different stages as revealed by a ^{1} H-NMR based metabonomic approach. *Mol. BioSyst.*, *10*(3), 686–693. https://doi.org/10.1039/C3MB70609E

Doisaki, M., Katano, Y., Nakano, I., Hirooka, Y., Itoh, A., Ishigami, M., … Shimomura, Y. (2010). Regulation of hepatic branched-chain α-keto acid dehydrogenase kinase in a rat model for type 2 diabetes mellitus at different stages of the disease. *Biochemical and Biophysical Research Communications*, *393*(2), 303–307. https://doi.org/10.1016/j.bbrc.2010.02.004

Duckett, S. K., Volpi-Lagreca, G., Alende, M., & Long, N. M. (2014). Palmitoleic acid reduces intramuscular lipid and restores insulin sensitivity in obese sheep. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, *7*, 553–563. https://doi.org/10.2147/DMSO.S72695

Dudzinska, W. (2014). Purine nucleotides and their metabolites in patients with type 1 and 2 diabetes mellitus. *Journal of Biomedical Science and Engineering*, *07*(01), 38–44. https://doi.org/10.4236/jbise.2014.71006

Fernández-Ochoa, Á., Quirantes-Piné, R., Borrás-Linares, I., Gemperline, D., Alarcón Riquelme, M. E., Beretta, L., & Segura-Carretero, A. (2019). Urinary and plasma metabolite differences detected by HPLC-ESI-QTOF-MS in systemic sclerosis patients. *Journal of Pharmaceutical and Biomedical Analysis*, *162*, 82–90. https://doi.org/10.1016/j.jpba.2018.09.021

Gallagher, E. J., LeRoith, D., Stasinopoulos, M., Zelenko, Z., & Shiloach, J. (2016). Polyol accumulation in muscle and liver in a mouse model of type 2 diabetes. *Journal of Diabetes and Its Complications*, *30*(6), 999–1007. https://doi.org/10.1016/j.jdiacomp.2016.04.019

Gandhi, G. R., Jothi, G., Antony, P. J., Balakrishna, K., Paulraj, M. G., Ignacimuthu, S., … Al-Dhabi, N. A. (2014). Gallic acid attenuates high-fat diet fed-streptozotocin-induced insulin resistance via partial agonism of PPARγ in experimental type 2 diabetic rats and enhances glucose uptake through translocation and activation of GLUT4 in PI3K/p-Akt signaling pathway. *European Journal of Pharmacology*, *745*, 201–216.

Gil de la Fuente, A., Grace Armitage, E., Otero, A., Barbas, C., & Godzien, J. (2017). Differentiating signals to make biological sense – A guide through databases for MS-based non-targeted metabolomics. *Electrophoresis*, *38*(18), 2242–2256. https://doi.org/10.1002/elps.201700070

Gondi, M., & Rao, U. J. S. P. (2015). Ethanol extract of mango (Mangifera indica L.) peel inhibits α-amylase and α-glucosidase activities, and ameliorates diabetes related biochemical parameters in streptozotocin (STZ)-induced diabetic rats. *Journal of Food Science and Technology*, *52*(12), 7883–7893.

Goyal, R., & Jialal, I. (2018). Diabetes mellitus, type 2. In *StatPearls [Internet]*. StatPearls Publishing.

Huang, Q., Yin, P., Wang, J., Chen, J., Kong, H., Lu, X., & Xu, G. (2011). Method for liver tissue metabolic profiling study and its application in type 2 diabetic rats based on ultra performance liquid chromatography-mass spectrometry. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*, *879*(13–14), 961–967. https://doi.org/10.1016/j.jchromb.2011.03.009

Jacobo-Cejudo, M. G., Valdés-Ramos, R., Guadarrama-López, A. L., Pardo-Morales, R. V., Martínez-Carrillo, B. E., & Harbige, L. S. (2017). Effect of n-3 polyunsaturated fatty acid supplementation on metabolic and inflammatory biomarkers in type 2 diabetes mellitus patients. *Nutrients*, *9*(6), 1–11. https://doi.org/10.3390/nu9060573

Jovanovski, E., Li, D., Thanh Ho, H. V., Djedovic, V., De Castro Ruiz Marques, A., Shishtar, E., … Vuksan, V. (2017, May). The effect of alpha-linolenic acid on glycemic control in individuals with type 2 diabetes. *Medicine (United States)*. Wolters Kluwer Health. https://doi.org/10.1097/MD.0000000000006531

Karachalias, N., Babaei-Jadidi, R., Ahmed, N., & Thornalley, P. J. (2003). Accumulation of fructosyl-lysine and advanced glycation end products in the kidney, retina and peripheral nerve of streptozotocin-induced diabetic rats. *Biochemical Society Transactions*, *31*(6), 1423–1425. https://doi.org/10.1042/bst0311423

Karachalias, N., Babaei-Jadidi, R., Rabbani, N., & Thornalley, P. J. (2010). Increased protein damage in renal glomeruli, retina, nerve, plasma and urine and its prevention by thiamine and benfotiamine therapy in a rat model of diabetes. *Diabetologia*, *53*(7), 1506–1516. https://doi.org/10.1007/s00125-010-1722-z

Kristal, B. S., Vigneau-Callahan, K. E., Moskowitz, A. J., & Matson, W. R. (1999). Purine catabolism: Links to mitochondrial respiration and antioxidant defenses? *Archives of Biochemistry and Biophysics*, *370*(1), 22–33. https://doi.org/10.1006/abbi.1999.1387

Kuzuya, T., Katano, Y., Nakano, I., Hirooka, Y., Itoh, A., Ishigami, M., … Shimomura, Y. (2008). Regulation of branched-chain amino acid catabolism in rat models for spontaneous type 2 diabetes mellitus. *Biochemical and Biophysical Research Communications*, *373*(1), 94–98. https://doi.org/10.1016/j.bbrc.2008.05.167

Lankinen, M. A., Stančáková, A., Uusitupa, M., Ågren, J., Pihlajamäki, J., Kuusisto, J., … Laakso, M. (2015). Plasma fatty acids as predictors of glycaemia and type 2 diabetes. *Diabetologia*, *58*(11), 2533–2544. https://doi.org/10.1007/s00125-015-3730-5

Lanza, I. R., Zhang, S., Ward, L. E., Karakelides, H., Raftery, D., & Sreekumaran Nair, K. (2010). Quantitative metabolomics by 1H-NMR and LC-MS/MS confirms altered metabolic pathways in diabetes. *PLoS ONE*, *5*(5), e10538. https://doi.org/10.1371/journal.pone.0010538

Lin, X. hong, Xu, M. tong, Tang, J. ying, Mai, L. fang, Wang, X. yi, Ren, M., & Yan, L. (2016). Effect of intensive insulin treatment on plasma levels of lipoprotein-associated phospholipase A2 and secretory phospholipase A2 in patients with newly diagnosed type 2 diabetes. *Lipids in Health and Disease*, *15*(1), 1–10. https://doi.org/10.1186/s12944-016-0368-3

Martin, M., & He, Q. (2009, September). Mango bioactive compounds and related nutraceutical properties-A review. *Food Reviews International*. https://doi.org/10.1080/87559120903153524

McGarry, J. D., & Foster, D. W. (1980). Regulation of Hepatic Fatty Acid Oxidation and

Ketone Body Production. *Annual Review of Biochemistry*, *49*(1), 395–420. https://doi.org/10.1146/annurev.bi.49.070180.002143

Meikle, P. J., & Summers, S. A. (2017, February). Sphingolipids and phospholipids in insulin resistance and related metabolic disorders. *Nature Reviews Endocrinology*. Nature Publishing Group. https://doi.org/10.1038/nrendo.2016.169

Naveen, J., & Baskaran, V. (2018). Antidiabetic plant-derived nutraceuticals: a critical review. *European Journal of Nutrition*, *57*(4), 1275–1299.

Neinast, M. D., Jang, C., Hui, S., Murashige, D. S., Chu, Q., Morscher, R. J., … Arany, Z. (2019). Quantitative Analysis of the Whole-Body Metabolic Fate of Branched-Chain Amino Acids. *Cell Metabolism*, *29*(2), 417-429.e4. https://doi.org/10.1016/j.cmet.2018.10.013

Neis, E. P. J. G., Dejong, C. H. C., & Rensen, S. S. (2015, April). The role of microbial amino acid metabolism in host metabolism. *Nutrients*. Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/nu7042930

Pacheco-Ordaz, R., Antunes-Ricardo, M., Gutiérrez-Uribe, J. A., & González-Aguilar, G. A. (2018). Intestinal permeability and cellular antioxidant activity of phenolic compounds from mango (Mangifera indica cv. ataulfo) peels. *International Journal of Molecular Sciences*, *19*(2). https://doi.org/10.3390/ijms19020514

Palafox-Carlos, H., Yahia, E. M., & González-Aguilar, G. A. (2012). Identification and quantification of major phenolic compounds from mango (Mangifera indica, cv. Ataulfo) fruit by HPLC–DAD–MS/MS-ESI and their individual contribution to the antioxidant activity during ripening. *Food Chemistry*, *135*(1), 105–111. https://doi.org/10.1016/J.FOODCHEM.2012.04.103

Pedraza-Chaverri, J., Cárdenas-Rodríguez, N., Orozco-Ibarra, M., & Pérez-Rojas, J. M. (2008, October). Medicinal properties of mangosteen (Garcinia mangostana). *Food and Chemical Toxicology*. Pergamon. https://doi.org/10.1016/j.fct.2008.07.024

Prabhu, S., Jainu, M., Sabitha, K. E., & Devi, C. S. S. S. (2006). Role of mangiferin on biochemical alterations and antioxidant status in isoproterenol-induced myocardial infarction in rats. *Journal of Ethnopharmacology*, *107*(1), 126–133. https://doi.org/10.1016/j.jep.2006.02.014

Quirós-Sauceda, A. E., Oliver Chen, C. Y., Blumberg, J. B., Astiazaran-Garcia, H., Wall-Medrano, A., & González-Aguilar, G. A. (2017). Processing 'ataulfo' mango into juice preserves the bioavailability and antioxidant capacity of its phenolic compounds. *Nutrients*, *9*(10), 1–12. https://doi.org/10.3390/nu9101082

Renga, B., Mencarelli, A., Vavassori, P., Brancaleone, V., & Fiorucci, S. (2010). The bile acid sensor FXR regulates insulin transcription and secretion. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, *1802*(3), 363–372. https://doi.org/10.1016/j.bbadis.2010.01.002

Rodríguez, T., Alvarez, B., Busquets, S., Carbó, N., López-Soriano, F. J., & Argilés, J. M. (1997). The increased skeletal muscle protein turnover of the streptozotozin diabetic rat is associated with high concentrations of branched-chain amino acids. *Biochemical and Molecular Medicine*, *61*(1), 87–94. https://doi.org/10.1006/bmme.1997.2585

Saha, S., Sadhukhan, P., & Sil, P. C. (2016, September). Mangiferin: A xanthonoid with multipotent anti-inflammatory potential. *BioFactors*. https://doi.org/10.1002/biof.1292

Sanugul, K., Akao, T., Li, Y., Kakiuchi, N., Nakamura, N., & Hattori, M. (2005). Isolation of a Human Intestinal Bacterium That Transforms Mangiferin to Norathyriol and Inducibility of the Enzyme That Cleaves a C-Glucosyl Bond. *Biological & Pharmaceutical Bulletin*, *28*(9), 1672–1678. https://doi.org/10.1248/bpb.28.1672

Sas, K. M., Karnovsky, A., Michailidis, G., & Pennathur, S. (2015). Metabolomics and diabetes: analytical and computational approaches. *Diabetes*, *64*(3), 718–732. https://doi.org/10.2337/db14-0509

Sekar, V., Chakraborty, S., Mani, S., Sali, V. K., & Vasanthi, H. R. (2019). Mangiferin from Mangifera indica fruits reduces post-prandial glucose level by inhibiting α-glucosidase and α-amylase activity. *South African Journal of Botany*, *120*, 129–134.

Sellamuthu, P. S., Arulselvan, P., Muniappan, B. P., Fakurazi, S., & Kandasamy, M. (2013). Mangiferin from Salacia chinensis prevents oxidative stress and protects pancreatic β-cells in streptozotocin-induced diabetic rats. *Journal of Medicinal Food*, *16*(8), 719–727.

Suhre, K., Meisinger, C., Döring, A., Altmaier, E., Belcredi, P., Gieger, C., … Illig, T. (2010). Metabolic footprint of diabetes: A multiplatform metabolomics study in an epidemiological setting. *PLoS ONE*, *5*(11), e13953. https://doi.org/10.1371/journal.pone.0013953

Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., … Viant, M. R. (2007). Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics*, *3(3)*, 211–221. https://doi.org/10.1007/s11306-007-0082-2

Takahashi, M., Ando, J., Shimada, K., Nishizaki, Y., Tani, S., Ogawa, T., … Komuro, I. (2017). The ratio of serum n-3 to n-6 polyunsaturated fatty acids is associated with diabetes mellitus in patients with prior myocardial infarction: A multicenter cross-sectional study. *BMC Cardiovascular Disorders*, *17*(1), 41. https://doi.org/10.1186/s12872-017-0479-4

Talbot, N. A., Wheeler-Jones, C. P., & Cleasby, M. E. (2014). Palmitoleic acid prevents palmitic acid-induced macrophage activation and consequent p38 MAPK-mediated skeletal muscle insulin resistance. *Molecular and Cellular Endocrinology*, *393*(1–2), 129–142. https://doi.org/10.1016/j.mce.2014.06.010

Tangvarasittichai, S. (2015). Oxidative stress, insulin resistance, dyslipidemia and type 2 diabetes mellitus. *World Journal of Diabetes*, *6*(3), 456. https://doi.org/10.4239/wjd.v6.i3.456

Ulaszewska, M. M., Weinert, C. H., Trimigno, A., Portmann, R., Andres Lacueva, C., Badertscher, R., … Vergères, G. (2019). Nutrimetabolomics: An Integrative Action for Metabolomic Analyses in Human Nutritional Studies. *Molecular Nutrition & Food Research*, *63*(1), 1800384. https://doi.org/10.1002/mnfr.201800384

Velderrain-Rodríguez, G., Torres-Moreno, H., Villegas-Ochoa, M., Ayala-Zavala, J., Robles-Zepeda, R., Wall-Medrano, A., & González-Aguilar, G. (2018). Gallic acid content and an antioxidant mechanism are responsible for the antiproliferative activity of 'Ataulfo'mango peel on LS180 cells. *Molecules*, *23*(3), 695.

Wieland, O., Weiss, L., & Eger-Neufeldt, I. (1964). Enzymatic regulation of liver acetyl-CoA metabolism in relation to ketogenesis. *Advances in Enzyme Regulation*, *2*(C), 85–99. https://doi.org/10.1016/S0065-2571(64)80007-8

Yui, D., Nishida, Y., Nishina, T., Mogushi, K., Tajiri, M., Ishibashi, S., … Yokota, T. (2015). Enhanced phospholipase A2 group 3 expression by oxidative stress decreases the insulin-degrading enzyme. *PLoS ONE*, *10*(12), e0143518. https://doi.org/10.1371/journal.pone.0143518

SUPLEMENTARY MATERIAL

1) Phytochemical content of mango peel and flesh and diet

Main compounds present in peel include mangiferin (2383 mg/kg DW), gallic acid (29,426 mg/kg DW) and 2-hydroxybenzoic acid (700 mg/kg DW); main compounds present in pulp include gallic acid (215 mg/kg DW), vanillic acid (102.4 mg/kg DW) and chlorogenic acid (118.1 mg/kg DW). Thus, the concentration of mangiferin, gallic acid, 2-hydroxybenzoic acid, vanillic acid and chlorogenic acid in the mango-enriched diet were 119.15, 1473.45, 35, 1.02 and 1.18 mg/kg of diet, respectively.

Table 1S. Body weight gain (BWG) and blood glucose levels in fasted STZ-induced diabetic rats.

Parameters	Groups		
	HC	UD	MTD
Initial BWG (g)	46.07 ± 8.12^{a}	1.62b ± 5.75^{b}	41.6a ± 11.30^{a}
Final BWG (g)	66.93 ± 5.60^{a}	11.92 ± 4.08^{b}	47.4 ± 13.42^{ab}
Initial glycemia (mg/dL)	84.57 ± 5.14^{b}	322.8 ± 25.31^{a}	323.9 15.44^{a}
Final glycemia (mg/dL)	83.71 12.63^{b}	434 ± 39.96^{a*}	364.5 36.85^{a}

Rats were classified as diabetic if their fasted blood glucose level ≥200 mg/dL on day 3 after STZ injection. Values of BWG and glycemia are mean ± S.E.M. ($p<0.05$) when applying one-way ANOVA followed by Tukey's post-hoc test. ^{ab} indicate significantly differences between groups; * indicate significantly differences at initial and final measures in the same group.

Table 2S. Composition of the experimental diets (g/kg).

Ingredient	Control diet	Mango-enriched diet
Corn starch[1]	414	410
Sucrose	100	-
Cellulose	50	30
Casein[2]	320	315
Corn oil	27	25
Lard	27	25
Vitamin mix[3]	10	10
Salt mix[4]	30	28
Choline chloride[5]	2	2
Water	20	5
Mango peel	-	50
Mango pulp	-	100

Mango pulp was used to substitute sucrose in the experimental diet. 20 % of cellulose was substituted by mango peel in the experimental diet. Diets were isocaloric (3.5 kcal/g); 57 % of kcal from carbohydrates, 29 % protein and 14 % from lipids. Ingredients [1-5] were obtained from Bio-Serv (Flemington, NJ, USA) #3200[1]; #1100[2]; #F800[3]; #F8505[4]; #6105[5]

Figure 1S. PCA scores plots from raw data obtained in the analyses from liver (a) and serum samples (b) (red dots, healthy controls, HC; green dots, untreated diabetic group, UD; blue dots, mango-treated diabetic group, MTD).

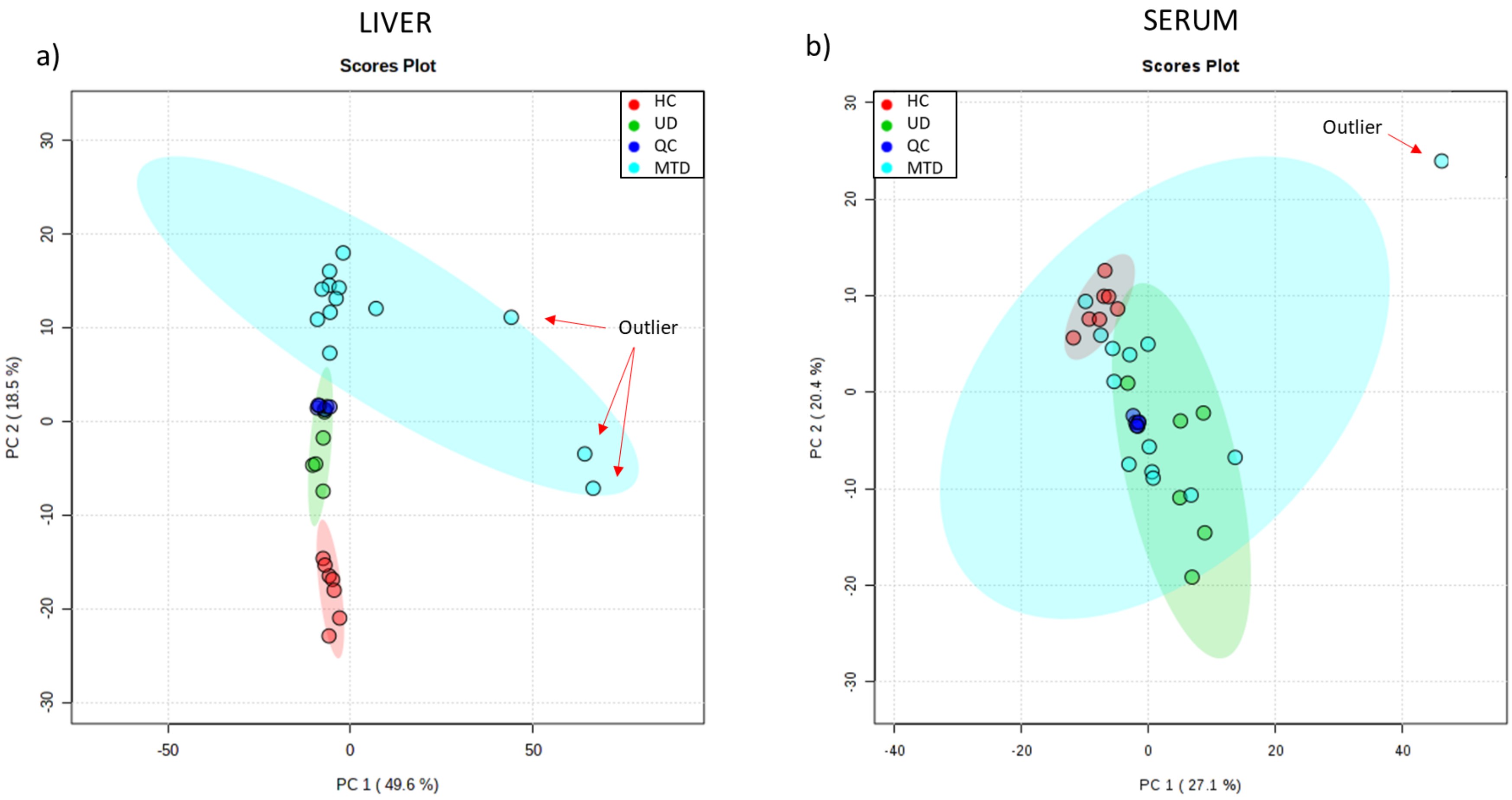


Table 3S Analytical parameters of the annotated metabolites in serum samples and search database id.

RT (min)	Mass (Da)	Molecular Formula	Metabolite	Search Database ID	MS/MS Fragments
22.58	302.2258	$C_{20}H_{30}O_2$	Eicosapentanoic acid	HMDB01999	149.1325/229.1987/257.2275/258.2305
19.19	318.2227	$C_{20}H_{30}O_3$	12-HEPE	HMDB10202	317.2134
17.95	465.3130	$C_{26}H_{43}NO_6$	Glycocholic acid	HMDB00138	74.0237/400.2851/402.3023
1.50	270.0931	$C_{10}H_{14}N_4O_5$	Histidinyl-Aspartate	HMDB28881	269.0893 (Level 2)
17.71	408.294	$C_{24}H_{40}O_5$	Cholic acid	HMDB00619	69.0340/289.2222/343.2685
1.42	279.1320	$C_{11}H_{21}NO_7$	N-(1-Deoxy-1-fructosyl)valine	HMDB37844	116.072/158.0823/159.0859
24.58	330.2550	$C_{22}H_{34}O_2$	Docosapentaenoic acid	HMDB06528	44.9983/59.0133/83.0502/96.9585
12.71	284.0918	$C_{13}H_{16}O_7$	p-Cresol glucuronide	HMDB11686	59.0130/107.0493
23.71	678.4681	$C_{39}H_{67}O_7P$	PA(P-36:5)	LMGP10030022	255.1698/284.2706/462.3059
12.01	86.0735	$C_5H_{10}O$	Iso-valeraldehyde	HMDB0006478	41.0019/43.018
6.79	118.1311	$C_5H_{10}O_3$	3-hydroxyisovaleric acid	HMDB00754	71.0476/117.0528
23.13	254.2265	$C_{16}H_{30}O_2$	Palmitoleic acid	HMDB60082	253.2168
23.14	276.4137	$C_{18}H_{28}O_2$	Stearidonic acid	HMDB06547	133.067
34.70	583.5189	$C_{34}H_{67}NO_3$	Cer(d18:1/16:0)	HMDB04949	582.5099 (Level 2)
11.37	122.0373	$C_7H_6O_2$	4-hydroxybenzaldehyde	HMDB11718	92.0266/121.0276
25.99	332.2735	$C_{22}H_{36}O_2$	Adrenic Acid	HMDB02226	59.0208/61.9872/331.2584
1.51	129.0406	$C_5H_7NO_3$	Pyroglutamic acid	HMDB60262	41.0025/41.9967/52.01784
23.70	396.2277	$C_{18}H_{37}O_7P$	PA(15:0)	HMDB62324	59.0139/211.0365/241.2154/326.1541
1.76	129.0419	$C_5H_7NO_3$	Pyrrolidonecarboxylic acid	HMDB00805	41.0389/41.9978/44.0128/57.0333/58.0302/66.0361/86.06
23.71	396.2320	$C_{18}H_{37}O_7P$	(9S,10S)-10-hydroxy-9-(phosphonooxy)octadecanoate	HMDB59632	44.9968/59.0134/141.0933
23.72	328.2431	$C_{22}H_{32}O_2$	Docosahexaenoic acid	HMDB0002183	121.1019/135.1174 /283.243/284.2474
22.70	300.4351	$C_{20}H_{28}O_2$	Retinoic acid	HMDB12874	44.9968/161.1341/229.1958/301.2222
12.79	173.1054	$C_8H_{15}NO_3$	Hexanoylglycine	HMDB00701	130.0882
22.71	278.2261	$C_{18}H_{30}O_2$	Linolenic acid	HMDB30962	44.9964/71.014
21.17	567.3359	$C_{30}H_{50}NO_7P$	LysoPC(22:6)	HMDB10404	59.0128/78.9584/152.9967/168.0431/224.0713
23.78	654.4694	$C_{44}H_{62}O_4$	14-DHAHDHA	HMDB0112171	653.4559 (Level 2)

Table 4S Analytical parameters of the annotated metabolites in liver samples and search database id.

RT (min)	Mass (Da)	Molecular Formula	Metabolite	Search Database ID	MS/MS Fragments
7.61	358.1014	$C_{11}H_{23}N_2O_7PS$	Pantetheine 4''-phosphate	HMDB0001416	113.0245/136.0338/ 208.0358
17.34	228.0429	$C_{13}H_8O_4$	Euxanthone	HMDB30724	41.002/62.0147/65.0032/73.0079/79.0186/91.018/181.0306
34.15	825.5523	$C_{45}H_{80}NO_{10}P$	PS(39:4)	LMGP03010795	44.9985/253.2164/303.2328/764.5223/824.5473
15.79	320.1476	$C_{14}H_{28}O_8$	Octanoylglucoronide	HMDB0010347	41.0027/43.0184/59.0133/71.0133/87.0082/99.1174/143.1072
1.47	345.0379	$C_{10}H_{12}N_5O_7P$	Cyclic GMP	HMDB0001314	101.0244/107.0354/128.035/134.0474/135.0308/148.0437
1.47	307.0847	$C_{10}H_{17}N_3O_6S$	Glutathione	HMDB0000125	41.9947/71.0091/74.0028/102.0498/128.0353/148.0437
3.05	216.1133	$C_9H_{16}N_2O_4$	Pro-Thr	HMDB0029027	44.9952/215.1057
7.25	282.1076	$C_{10}H_{14}N_6O_4$	2-Aminoadenosine	CHEBI:1014	41.998/44.0132/59.0128/71.0139/71.05/72.0085/74.0229/88.0402/146.0843
2.75	176.0691	$C_7H_{12}O_5$	2-Isopropylmalic acid	HMDB0000402	41.0013/59.0114/69.0328/87.0055
27.94	676.5406	$C_{38}H_{77}O_7P$	PA(O-18:0/17:0)	LMGP10020025	283.2647/284.2677/ 285.2705/351.251/657.5225
1.41	347.0596	$C_{10}H_{14}N_5O_7P$	Adenosine monophosphate	HMDB0000045	107.0347/148.0436/134.0468/211.0010/346.0552
3.28	267.0994	$C_{10}H_{13}N_5O_4$	Adenosine	METLINID86	92.0250/107.0369/134.0457
7.25	89.0468	$C_3H_7NO_2$	beta-Alanine	HMDB0000056	43.0184/44.0500/41.0027
25.08	330.255	$C_{22}H_{34}O_2$	Docosapentaenoic acid	HMDB06528	44.9983/59.0133/83.0502/96.9585
12.79	173.1059	$C_8H_{15}NO_3$	N-Acetylisoleucine	HMDB61684	42.9851/44.9973/74.0241/ 130.0855
14.39	172.1103	$C_9H_{16}O_3$	9-oxo-nonanoic acid	HMDB0094711	109.0655/111.0811/113.0961/153.0921
5.43	327.1332	$C_{15}H_{21}NO_7$	N-(1-Deoxy-1-fructosyl)phenylalanine	HMDB0037846	101.0243/103.0553/104.0586/118.0662/132.0812/147.0447/ 148.0485/149.06/164.0717
2.23	343.1256	$C_{15}H_{21}NO_8$	N-(1-Deoxy-1-fructosyl)tyrosine	HMDB0037845	101.0236/ 103.0397/110.0244/117.0188/151.0259/152.0339/180.0663
2.38	181.0721	$C_9H_{11}NO_3$	L-Tyrosine	HMDB0000158	52.0156/72.0045/74.0203/93.0299/119.0446
2.64	293.1509	$C_{12}H_{23}NO_7$	N-(1-Deoxy-1-fructosyl)leucine	HMDB37840	59.0107/71.0092/73.0256/82.0625/101.0188/113.0553/115.0721/130.0812
1.53	541.0576	$C_{15}H_{21}N_5O_{13}P_2$	Cyclic ADP-ribose	C13050	71.0127/78.9593/101.0244/128.0343/158.9245/161.0439/179.0542/199.0717/272.955/290.0868
22.67	521.3481	$C_{26}H_{52}NO_7P$	LysoPC(18:1)	HMDB0010385	44.9971/78.9583/281.2519/566.3461
5.42	237.1027	$C_{12}H_{15}NO_4$	N-lactoyl-phenylalanine	HMDB0062175	103.0518/147.0451/148.0445/164.0671
2.64	203.1178	$C_9H_{17}NO_4$	N-lactoyl-Leucine	HMDB62176	55.0149/59.01/68.01/73.0252/84.0753/130.0819
2.39	131.095	$C_6H_{13}NO_2$	Beta-Leucine	HMDB03640	41.0006/43.016/56.9953/59.0089/71.0093/88.0355
9.72	366.1447	$C_{17}H_{22}N_2O_7$	N-(1-Deoxy-1-fructosyl)tryptophan	FDB016998	116.0504/203.0841/365.1343
22.67	507.3321	$C_{25}H_{50}NO_7P$	LysoPE(20:1)	HMDB11512	59.013/68.9959/83.0498/87.0449/103.0389/279.2361
6.41	284.0768	$C_{10}H_{12}N_4O_6$	Xanthosine	HMDB0000299	108.0202/109.0239/122.0354/133.0157/150.018/151.0259/152.0289/165.0416
7.25	219.1161	$C_9H_{17}NO_5$	Pantothenic acid	HMDB0000210	41.0017/41.9978/44.013/44.9968/59.0126/71.0494/72.0089/88.0386

Table 5S. Retention times, masses, VIP, FDR and the results of the Tukey's HSD from unknown metabolites present in serum samples.

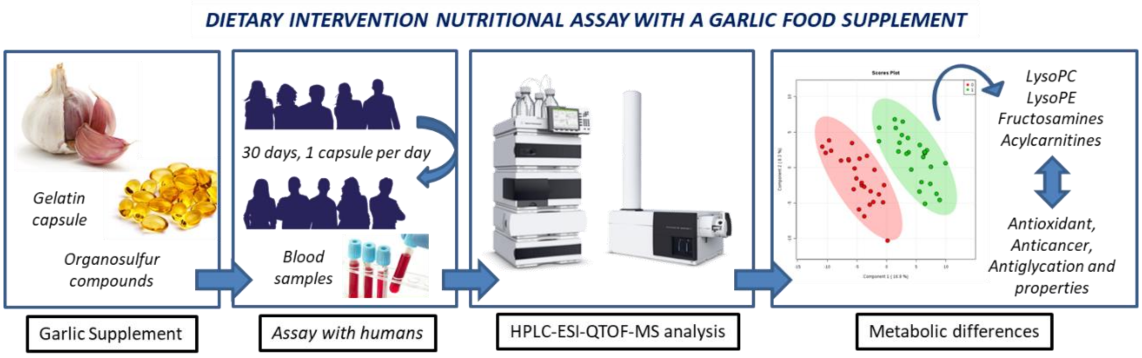
Mass (Da)	RT (min)	VIP value	FDR	Tukey's HSD
261.9855	1.19	2.93	2.44E-07	T-C; T-D
216.0366	1.08	2.13	6.23E-03	D-C; T-C
218.0335	1.07	2.05	4.81E-03	D-C; T-C
600.1933	23.69	1.86	2.02E-05	T-C; T-D
532.2063	23.70	1.84	6.98E-06	T-C; T-D
428.1697	23.69	1.83	3.85E-07	T-C; T-D
284.2536	23.72	1.74	1.03E-06	D-C; T-C; T-D
1099.5202	35.14	1.67	3.13E-05	D-C; T-C
307.0667	2.16	1.62	7.53E-04	D-C; T-C
970.4830	15.81	1.61	3.06E-02	D-C; T-C
1038.4655	15.81	1.59	4.84E-02	T-C
627.2129	15.80	1.54	1.48E-02	D-C; T-C
790.3542	15.80	1.51	3.62E-02	D-C; T-C

Table 6S. Retention times, masses, VIP, FDR and the results of the Tukey's HSD from unknown metabolites present in liver samples.

Mass (Da)	RT (min)	VIP value	FDR	Tukey's HSD
432.2527	18.69	3.34	2.54E-20	D-C, T-C, T-D
193.0426	4.75	3.28	3.89E-26	T-C, T-D
651.4653	27.92	3.28	1.27E-19	D-C, T-C, T-D
309.0845	1.46	3.26	2.68E-16	T-C, T-D
341.1267	11.11	3.23	1.15E-19	T-C, T-D
329.0648	1.47	3.15	1.68E-16	T-C, T-D
422.2667	21.35	2.46	5.14E-06	T-C, T-D
710.5087	36.13	2.39	2.97E-06	T-C, T-D
218.0902	1.57	1.94	1.14E-09	D-C, T-C, T-D
663.3408	20.82	1.64	1.37E-06	D-C, T-C, T-D
127.0520	0.99	1.61	2.18E-08	D-C, T-C, T-D
420.2465	27.94	1.61	1.15E-07	D-C, T-C, T-D
402.2297	5.30	1.58	1.13E-07	D-C, T-C, T-D
390.1303	5.42	1.55	4.12E-06	D-C, T-C, T-D
326.1345	0.99	1.54	6.71E-07	D-C, T-C, T-D
586.4947	33.81	1.54	1.56E-06	D-C, T-C, T-D
369.2506	27.93	1.53	1.32E-06	D-C, T-C, T-D
454.3303	25.84	1.53	2.43E-06	D-C, T-C, T-D
112.0873	14.39	1.53	1.37E-06	D-C, T-C, T-D
356.1431	2.64	1.53	1.51E-06	D-C, T-C, T-D
351.1164	1.25	1.52	1.37E-06	D-C, T-C
201.0467	11.87	1.52	5.78E-06	D-C, T-C, T-D

Capítulo 3

Efecto del consumo de un suplemento alimenticio de ajo en el metabolismo humano


DIETARY INTERVENTION NUTRITIONAL ASSAY WITH A GARLIC FOOD SUPPLEMENT
Gelatin capsule
Organosulfur compounds
Garlic Supplement
30 days, 1 capsule per day
Blood samples
Assay with humans
HPLC-ESI-QTOF-MS analysis
Scores Plot
LysoPC
LysoPE
Fructosamines
Acylcarnitines
Antioxidant, Anticancer, Antiglycation and properties
Metabolic differences

Journal of Functional Foods. 2018, 49, 137-145

Álvaro Fernández-Ochoa, Isabel Borrás-Linares, Alberto Baños, J. David García López, Enrique Guillamón, Cristina Nuñez-Lechado, Rosa Quirantes-Piné, Antonio Segura-Carretero

DOI:10.1016/j.jff.2018.08.003

A fingerprinting metabolomic approach reveals deregulation of endogenous metabolites after the intake of a bioactive garlic supplement

ABSTRACT

Garlic (Allium sativum) has been described as containing phytonutrients with healthy properties. In this study, the effect of a bioactive garlic food supplement intake on human plasma metabolome was examined with the aim of understanding the mechanisms of action and involved pathways responsible for beneficial effects. With this purpose, a dietary intervention assay was performed in thirty healthy volunteers collecting plasma samples before intake and after one month of daily supplement consumption. Plasma samples were analysed by a fingerprinting metabolomic strategy based on HPLC-ESI-QTOF-MS., Our results revealed a total of 26 metabolites affected by supplement intake. In general, alterations in phospholipid metabolism were shown, detecting an increase in lysophosphatidylcholines, lysophosphatidylethanolamines and acylcarnitines. It is also remarkable that the level of four fructosamines decreased after the assay. These results are according with the antioxidant and antiglycation properties that have been previously associated with garlic extracts.

Keywords: Food supplement; Fructosamines; Garlic; HPLC-ESI-QTOF-MS; Lysophosphatidylcholines; Metabolomics.

1. INTRODUCTION

Metabolomics is an ‘omics’ technology that aims to study all low molecular weight molecules present in biological systems, which are known as metabolites. In this way, this tool allows to find alterations and interactions in the organism due to different conditions or causes (Agin et al., 2016). Currently, the main analytical techniques able to detect the greatest number of metabolites used in metabolomics studies are ^{1}H nuclear magnetic resonance spectroscopy (^{1}H-NMR) and mass spectrometry (MS) (Mumtaz et al., 2017).

Most metabolomics studies have been focused on human diseases, in order to know the pathways involved in their development and also to find biomarkers that allow the improvement of their diagnosis, prognosis and treatments (Johnson, Ivanisevic, & Siuzdak, 2016; X. Wang, Chen, & Jia, 2016). On the other hand, metabolomics studies have also been reported in other areas with different aims such as, classifying species, studying toxicity (Farag, Fekry, et al., 2017), or in the field of nutrition, mainly distinguished into three types of studies: dietary biomarker discovery, relation of diet and diseases and dietary intervention studies (Brennan, 2013; Gibbons, O’Gorman, & Brennan, 2015).

The last ones try to understand how certain foods or diets impact in the metabolic pathways focusing on both endogenous and exogenous metabolites. In this way, metabolomics has been widely applied to dietary intervention studies performed with foods highly consumed daily in the human diet such as butter, milk, cheese, tea, chocolate, cocoa, vitamins or fish oils, among others (Brennan, 2013; Zheng, Clausen, Dalsgaard, & Bertram, 2015).

Nevertheless, due to consumer concerns and demands, other types of food have appeared in the market whose effects in metabolome deserve further attention. In recent years there is a great interest in new nutritional products such as nutraceuticals, functional foods and food supplements. This kind of products has beneficial properties in the human health due to their high content in bioactive compounds, as the case of polyphenols. The dietary intake of phenolic compounds has presented beneficial properties in several diseases such as neurodegenerative diseases, cancer, hypertension or cardiovascular diseases (Del Rio et al., 2013; Rodriguez-Mateos et al., 2014). One example of supplement food containing these type of compounds has been detailed by Letizia Bresciani et al. who characterized 119 phenolic compounds in three food supplements which contained 36 different vegetables, fruits and berries (Bresciani et al., 2015).

Some dietary intervention studies have been also found in literature regarding specific compounds or food supplements. For instance, the effects of vitamin E supplementation (Wong & Lodge, 2012), intake of a functional beverage based on a grape skin extract (Khymenets et al., 2015) or grape extracts or wine supplementation (Jacobs et al., 2012) on human metabolism have been studied.

Among different products with bioactive compounds, garlic (*Allium sativum*) is one of the most famous since antiquity that has gained a great interest due to its varied composition including vitamins, phenolic acids, dipeptides, fatty acids, flavonoids and organosulfur compounds. The combination of these compounds makes this matrix has excellent properties such as anticancer, antioxidant, antibacterial, antimutagenic, antiplatelet, antimicrobial, antiaging and antihyperlipidemic activities, as well as immunomodulatory capacity and being able to modulate glucose and insulin levels. In this way, *Allium* present health properties for treatment of hypercholesterolemia, cancer hypertension, diabetes type 2, cataract, obesity and disturbances of the gastrointestinal tract (Amagase, Petesch, Matsuura, Kasuga, & Itakura, 2001; Farag, Ali, et al., 2017; Kopec, Piatkowska, Leszczynska, & Sikora, 2013).

Despite the number of dietary intervention studies has recently increased, there is still a lack of information on how food matrices, mainly new nutritional products, affect human metabolism. In this way, there is an urgent need to study the effect of these products in the metabolism due to their bioactive properties, which may help to understand their beneficial effects and the mechanisms of action and involved pathways in the human organism. Due to its composition in bioactive compounds and health benefits, garlic extracts are currently being used as nutraceutical or dietary supplement despite their impact in the human metabolome has not been deeply studied.

In this context the present study aims to examine the human metabolism changes due to a prolonged intake of a bioactive garlic supplement by means of a dietary intervention assay. The importance of this study is that it allows knowing what metabolic pathways are mainly altered in healthy individuals due to garlic consumption. The expected results can be related to the health benefits of garlic.

2. MATERIAL AND METHODS

2.1. Garlic supplement

Aliocare ®, a product containing 14.5% of organosulfur compounds, was provided by DOMCA S.A. (Granada, Spain).

2.2. Chemicals

All chemicals were of analytical reagent grade and used as received. Formic acid and LC-MS grade methanol for mobile phases were purchased from Fluka, Sigma-Aldrich (Steinheim, Germany) and Fisher Scientific (Madrid, Spain), respectively. Water was purified by a Milli-Q system from Millipore (Bedford, MA, USA). For plasma treatment, ethanol and methanol (Fisher Scientific Madrid, Spain) were used.

2.3. Dietary intervention nutritional assay

Thirty healthy volunteers (15 men and 15 women), age range of 20-40 years, were recruited in the city of Granada (Spain) to participate in the intervention nutritional assay. Each volunteer signed a consent form after receiving a detailed explanation of the study.

Exclusion criteria was based on current physical status and history of conditions including chronic severe diseases, current infection and antibiotic treatment or anti-inflammatory drugs within the previous two months, and any diseases or medications that could interfere with study outcome measures. Participants were withdrawn if they ingested food containing alliaceae or if they suffered diseases that require treatment with antibiotics or anti-inflammatory drugs during the study period.

Participants were informed to abstain from the intake of garlic, onion, leek and nutritional supplements (prebiotics, fitobiotics, vitamins or minerals) within the previous three weeks. The ethic committee of the University of Granada approved the study. During the study, the volunteers ingested one gelatin capsule contained 70 mg of garlic supplement per day. At the beginning and at the end of the study, blood samples were collected from participants into citrate containers. Plasma samples were obtained by centrifugation of containers for 15 min at 2000 *g* at 4 °C, then rapidly frozen and stored at −80 °C until further treatment and analysis.

2.4. Sample treatment

Plasma samples, which were stored at -80 °C, were thawed on ice. A plasma aliquot of 100 µl was mixed with 200 µl methanol:ethanol (50:50, v/v) in order to remove the protein content (Bruce et al., 2009). Afterwards, the mixture was vortex-mixed and then was kept at -20 °C during 30 min in order to achieve an efficient protein precipitation and avoid possible degradations. Next, the sample was centrifuged during 10 min at 14800 r.p.m. and 4 °C, and the supernatant was evaporated to dryness under vacuum in a centrifugal evaporator (Concentrator Plus, Eppendorf, Hamburg, Germany) during 2 h. Afterwards, the dry residue was reconstituted in 100 µl of initial mobile phase conditions (0.1% aqueous formic acid:methanol, 95:5, v/v) and centrifuged as mentioned above in order to remove solid particles. Finally, a 40 µl aliquot was transferred into HPLC vials and stored at -80 °C prior to analysis. A quality control sample (QC) was prepared by mixing equal volumes (20 µl) from each sample and treated as described above (Dettmer, Aronov, & Hammock, 2007).

2.5. HPLC-ESI-QTOF-MS analysis

Analyses were performed using an Agilent 1260 HPLC instrument (Agilent Technologies, Palo Alto, CA, USA) coupled to an Agilent 6540 Ultra High Definition (UHD) Accurate Mass Q-TOF equipped with a Jet Stream dual ESI interface.

The compounds were separated using a reversed-phase C18 analytical column (Agilent Zorbax Eclipse Plus, 1.8 µm, 4.6×150 mm) protected by a guard cartridge of the same packing. The mobile phases were water containing 0.1% of formic acid and methanol as solvent A and B, respectively. The following gradient of these mobile phases was used in order to obtain an efficient separation: 0 min [A:B 95/5], 5 min [A:B 90/10], 15 min [A:B 15/85], 30 min [A:B 0/100], and 35 min [A:B 95/5]. Finally, initial conditions were kept for 5 min at the end of each analysis to equilibrate the analytical column before the next run. The autosampler and column compartment temperatures were set at 4 and 25 ºC, respectively, whereas the flow rate and the injection volume were 0.4 mL/min and 5 µl.

Detection was performed in positive-ion mode over a range from 50 to 1700 m/z. All spectra were corrected by means of continuous infusion of two reference masses: purine (m/z 121.050873) and hexakis (^{1}H, ^{1}H, ^{3}H-tetrafluoropropoxy) phosphazine or HP-921 (m/z 922.009798). Both reference ions provided accurate mass measurement typically better than 2 ppm.

Ultrahigh pure nitrogen was used as drying and nebulizer gas at temperatures of 200 and 350 ºC and flows of 10 and 12 L/min, respectively. Other optimized parameters were as follows: capillary voltage, +4000V; nebuliser, 20 psi; fragmentor, 130 V; nozzle voltage, 500 V; skimmer, 45 V and octopole 1 RF Vpp, 750 V.

The analytical sequence of the samples consisted in: 2 blanks, 5 QCs, 5 randomized samples, 1 blank, 2 QCs, 5 randomized samples, etc. Finally, a MS/MS analysis of the QC sample was performed in order to facilitate the identification of potential biomarkers. This experiment was performed using nitrogen as the collision gas with the following collision energy values: 10 eV, 20 eV and 40 eV.

2.6. Data processing

Recursive Feature Extraction for small molecules was performed by means of MassHunter Profinder software (B.06.00, Agilent Technologies) to generate a list of the representative features present in plasma samples with their integrated areas. This algorithm combines "Molecular Feature Extraction" with "Find by Ion" algorithms (Kitawa et al., 2013). Therefore, the first algorithm finds features which are defined as the combination of co-eluted species that are related by isotopic distribution, presence of adducts, loss of molecules and/or charge-state envelop. Secondly, the features found in the samples are aligned by mass and retention time. Finally, a list with the resulting features is created and used to find them in the same samples more accurately.

Peaks were filtered by intensity threshold of 1250 counts. $[M+H]^+$, $[M+Na]^+$ and $[M-H_2O]$ were the considered species with a maximum charge of 2. Feature alignment parameters were ± 0.25 minutes and 40 ppm ± 4 mDa for retention time and mass windows, respectively. The integration method was Agile2 carrying out an average of spectra at peak start and end to subtract a background spectrum. Nevertheless, integration results were manually supervised to correct defaults.

2.7. Statistical analysis

Initially, the data were explored by unsupervised Principal Component Analysis (PCA) to check the reproducibility according to the distribution of QC samples and to identify any outliers. For multivariate analysis, data were transformed by means of log transformation to get a Gaussian distribution of the data and were set to Pareto scaling to make each variable comparable to each other (van den Berg, Hoefsloot, Westerhuis, Smilde, & van der Werf, 2006).

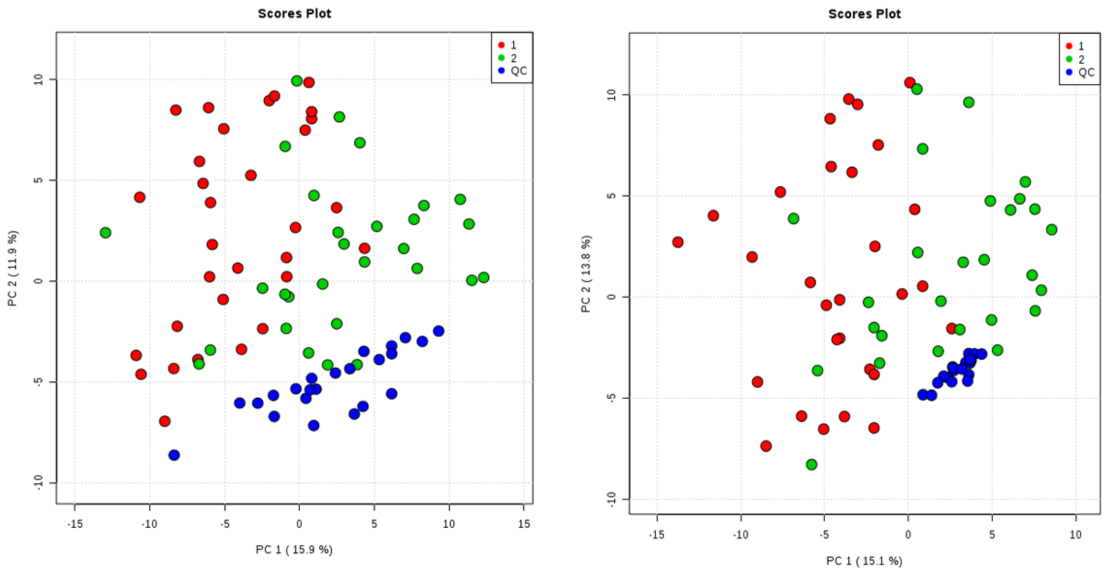
Features were normalized according to the QC samples (more details are described in the results section), and afterwards the features with high variability (RSD>30 %) in the QC samples were removed (Dunn et al., 2011).

After these steps, a supervised Partial Least Squares Discriminant Analysis (PLS-DA), a hierarchical clustering via heatmap and univariate statistical tests (paired t-test and paired fold change analysis) were performed in order to find metabolic differences due to nutritional supplementation. Both univariate and multivariate statistical tests were carried out in Metaboanalyst 3.0 software (Xia et al., 2015; Xia & Wishart, 2016).

3. RESULTS

3.1. Data quality assessment

The data processing described in material and methods section allowed obtaining a total of 306 molecular features. Firstly, PCA was performed for overall data in order to check the analytical reproducibility according to QC samples distribution. An analytical drift was detected due to the dispersion of the QC samples in the PCA scores plot (**Figure 1a**) according to their injection order.



Raw Data **Normalized Data**

Figure 1. PCA scores plot from Raw Data and Normalized Data. (red, pre-treatment; green, post-treatment, blue, QC samples).

This bias is often present in large-scaled non-targeted metabolomic studies and is usually related to fluctuations in the ionization efficiency of the electrospray interface (ESI) throughout the analytical sequence. In order to correct this variability, different strategies have been described in bibliography (Mizuno et al., 2017). Some of them are based on the use of the QC samples to monitor the drift and correct it (Dunn, Wilson, Nicholls, & Broadhurst, 2012; Kamleh, Ebbels, Spagou, Masson, & Want, 2012).

In this case, the integrated areas obtained for each feature in each sample were normalized by the sum of the total useful signal from the nearest QC in order to correct the aforementioned drifts and to get the areas of the samples comparable between them (Gika, Macpherson, Theodoridis, & Wilson, 2008). The improvement of data quality after applying this normalization procedure is shown in **Figure 1b,** where there is a clear clustering of QCs in PCA scores plot. Outliers were not detected and a slight grouping between two groups can be appreciated.

3.2. PLS-DA model and univariate statistical analysis

3.2.1. PLS-DA model

A PLS-DA model was built to discriminate the samples according to the supplementation. **Figure 2a** shows the scores plot of the PLS-DA model where the samples are clearly grouped according to their conditions

The model was established with two components obtaining the following performance parameters by 10-fold cross validation: accuracy, 0.9808; R2, 0.8899 and Q2, 0.7805. In order to test for possible overfitting, a permutation test was performed with 2000 permutations and using the prediction accuracy during training and the separation distance (B/W) as statistics tests. The results of these tests are showed in **Figures 2b** and **2c** resulting p-values under 5 E-4 which means that there is no overfitting in the model (Xia & Wishart, 2011). The PLS-DA model was also validated by means of the Receiver Operating Characteristic (ROC) curve (Steyerberg et al., 2010; Worley & Powers, 2013) (**Figure 2d**), obtaining an area under the curve (AUC) value of 0.995 (95% CI: 0.954-1), showing a perfect discrimination between both groups.

A total of 76 molecular features, whose VIP (variable importance in projection) values were higher than 1.0, were selected as responsible for the sample discrimination.

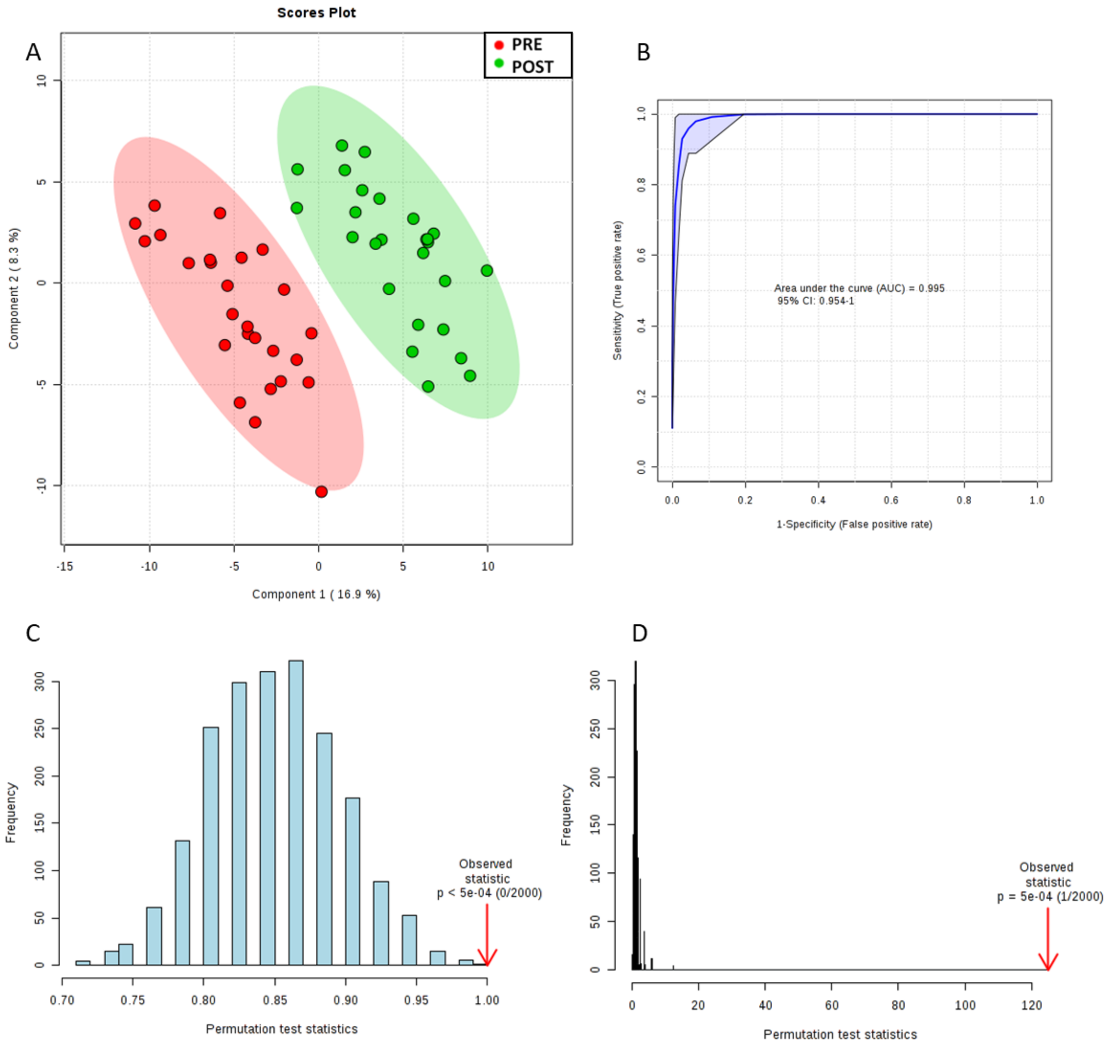


Figure 2. A supervised Partial Least Squares Discriminant Analysis (PLS-DA). Fig2a. PLS-DA scores plot (red and green points represent samples of pre and post-treatment, respectively, the circular areas represent the 95% confidence region of each group); Fig2b-c. Permutation test results (Statistical tests: 2c separation distance (B/W), 2d prediction accuracy during training); Fig2d. ROC curve for PLS-DA model validation.

3.2.2. Univariate statistical tests.

Univariate analyses were performed on the 76 selected features from PLS-DA model. Significant metabolites between the pre and post supplementation were estimated by a paired t-test (p-value ≤ 0.05) and paired fold change analysis (FC>1.5 in at least 75% of pairs). As a result of both tests, 39 significant features were obtained.

ROC curves were also constructed for the significant metabolites. AUROC values were used to evaluate the discriminatory power of each metabolite. A good curve is considered when the AUROC is higher than 0.7-0.8 (Xia, Broadhurst, Wilson, & Wishart, 2013). The AUROC of the selected significant features were higher than 0.75, which means that these metabolites could be considered biomarkers of the garlic extract intake. The top six metabolites with the higher AUROC values (**Figure 3**) were L-palmitoylcarnitine, 3-OH-cis-5-octenoylcarnitine, LysoPC(18:0), N-1-Deoxy-1fructosylTryptophan, Threonine-Methionine-Tryptophan (Thr-Met-Trp) and N-1-Deoxy-1fructosylTryptophan.

3.3. Altered metabolites after garlic supplement intake

The 39 statistically significant features were attempted to identify. This identification was carried out through the comparison of the accurate mass, isotopic distribution and fragmentation patterns obtained in MS/MS analysis with the online available metabolomic databases such as METLIN (http://metlin.scripps.edu), LipidMaps (http://lipidmaps.org), and Human Metabolome Database (http://hmdb.ca), as well as MS/MS fragmentation resources such as MetFrag (http://msbi.ipb-halle.de/MetFrag/).

As a result, 26 metabolites of the 39 candidates could be identified. Within the identified metabolites, four lysophosphatidylcholines, namely LysoPC(14:0), LysoPC(16:0), LysoPC(17:0) and LysoPC(18:0) were identified as two isomeric species. **Table 1** lists the significant metabolites which were identified together with their retention times, vip-values, molecular formulas, scores, p-values, in addition to the

AUROC values. Regarding unknown features, their corresponding parameters are located in the **Table S1.**

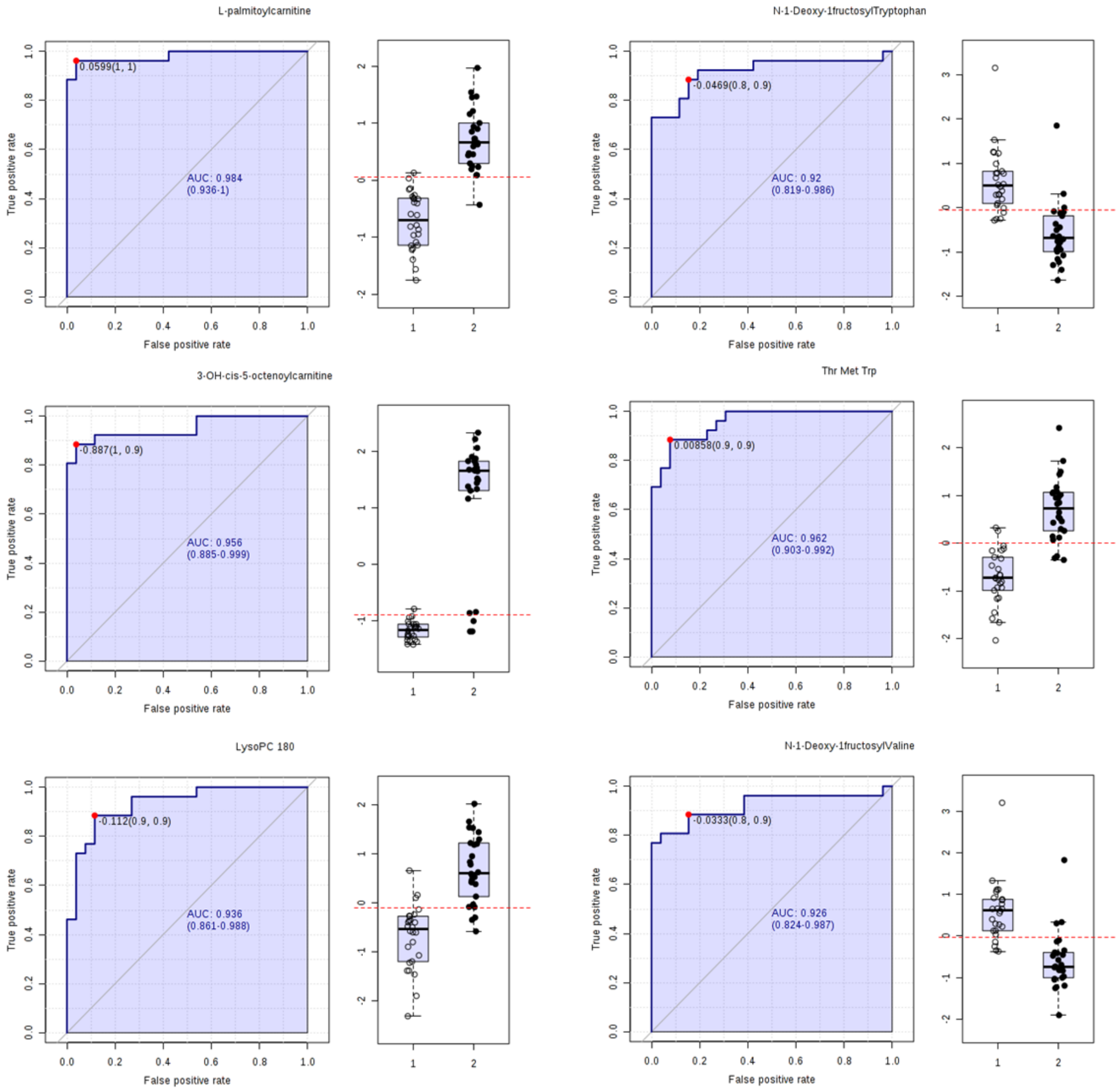


Figure 3. ROC curves corresponding to the six metabolites that present the highest AUROC (L-palmitoylcarnitine, 3-OH-cis-5-octenoylcarnitine, LysoPC(18:0), N-1-Deoxy-1fructosylTryptophan, Thr Met Trp and N-1-Deoxy-1fructosylTryptophan).

Table 1. Molecular and statistical details (Retention times, masses, molecular formulas with their scores, p-values, VIP-values and areas under ROC curve) of identified metabolites that presented significant differences between before and after garlic supplement intake.

	RT (min)	Mass (Da)	p-value	VIP value	AUROC	Molecular Formula	Score	Proposed Metabolite
SIGNIFICANT METABOLITES	5.5	279.1311	1.35 E-10	2.1887	0.9260	$C_{11}H_{27}NO_7$	94.6	N-(1-Deoxy-1fructosyl)Valine
	10.0	293.1473	8.20 E-7	1.8322	0.8036	$C_{12}H_{23}NO_7$	99.7	N-(1-Deoxy-1fructosyl)Isoleucine
	10.6	293.1475	1.74 E-8	1.9199	0.8297	$C_{12}H_{23}NO_7$	98.1	N-(1-Deoxy-1fructosyl)Leucine
	13.9	301.1886	5.24 E-11	4.2654	0.9560	$C_{15}H_{27}NO_5$	83.8	3-OH-cis-5-octenoylcarnitine
	15.5	366.1427	1.46 E-10	2.1661	0.9200	$C_{17}H_{22}N_2O_7$	97.1	N-(1-Deoxy-1fructosyl)Tryptophan
	20.9	213.2434	4.41 E-6	2.2279	0.8956	$C_{14}H_{31}N$	85.6	Tetradecylamine
	22.5	436.1749	1.31 E-9	2.6085	0.9620	$C_{20}H_{28}N_4O_5S$	91.1	Thr Met Trp
	24.2	399.3282	5.08 E-13	2.6069	0.9840	$C_{23}H_{45}NO_4$	77.9	L-palmitoylcarnitine
	24.7	425.3444	1.14 E-7	1.6577	0.8571	$C_{25}H_{47}NO_4$	84.4	L-oleoylcarnitine C18:1
	26.6	467.2935	4.24 E-6	1.7287	0.8187	$C_{22}H_{46}NO_7P$	83.8	LysoPC(14:0) isomers
	27.3	467.2966	9.38 E-6	1.6591	0.8132		95.3	
	27.9	511.3213	9.29 E-9	2.0390	0.8819	$C_{24}H_{50}NO_8P$	95.0	PS(O-18:0/0:0)**
	28.7	481.3396	1.71 E-6	1.6793	0.8558	$C_{23}H_{48}NO_7P$	97.5	LysoPC(15:0)
	28.8	525.2681	6.33 E-7	1.7375	0.8517	$C_{27}H_{44}NO_7P$	98.9	LysoPE (22:6)
	28.9	501.2759	1.05 E-7	1.6691	0.8640	$C_{25}H_{44}NO_7P$	90.0	LysoPE (20:4)
	30.2	495.3265	2.03 E-11	1.1813	0.9148	$C_{24}H_{50}NO_7P$	81.9	LysoPC (16:0) isomers
	29.4	495.3188	2.01 E-11	1.9116	0.8665		96.5	
	30.1	453.2765	3.92 E-10	1.6357	0.8640	$C_{21}H_{44}NO_7P$	82.0	LysoPE(16:0)
	30.8	442.3133	1.61 E-6	1.1305	0.8764	$C_{27}H_{42}N_2O_3$	69.6	N-palmitoyl tryptophan
	31.2	509.3449	2.80 E-6	2.0982	0.8750	$C_{25}H_{52}NO_7P$	97.1	LysoPC (17:0) isomers
	31.7	509.3357	6.75 E-8	2.3073	0.9200		98.7	
	32.0	481.3396	5.86 E-11	2.3294	0.8956	$C_{24}H_{52}NO_6P$	99.3	1-O-Hexadecyl-sn-glycero-3-phosphocholine
	32.3	523.3463	2.90 E-11	2.5224	0.9360	$C_{26}H_{54}NO_7P$	95.6	LysoPC (18:0) isomers
	33.6	523.3463	1.50 E-10	2.2400	0.9320		96.7	
	32.6	507.3567	3.72 E-8	1.7478	0.7912	$C_{26}H_{54}NO_6P$	98.7	LysoPC (P18:0)
	33.0	545.3404	1.08 E-9	2.1003	0.9011	$C_{28}H_{52}NO_7P$	92.6	LysoPC (20:3)

3.4. Hierarchical clustering analysis

Hierarchical clustering analysis was applied to twenty-six metabolites that were identified using a Pearson distance measure and Ward clustering algorithm. **Figure 4** shows the resulting heatmap where the metabolites clustering indicates two separate groups depending on whether the concentration increase or decrease after supplementation. In this way, N-palmitoyl tryphophan and four fructosamines compounds (Valine, Trypthophan, Leucine and Isoleucine) were the metabolites whose concentration decreased after garlic supplementation. On the opposite, the rest of significant identified metabolites concentrations increased after the intervention assay. On the other hand, the sample clustering shows also two groups according to the class. All samples were correctly classified with the exception of the one collected from the male volunteer number 10 after the supplementation (PH10-2). The cause of the wrong cluster of this sample could be that the initial levels of the fructosamines in this person (PH10-1) were very high in comparison with the rest of volunteers as can be observed in the intensity colors of these compounds in the heatmap.

4. DISCUSSION

Among the identified significant metabolites, it should be highlighted the number of lysophosphatidylcholines (LysoPC(14:0), LysoPC(15:0), LysoPC(16:0), LysoPC(17:0), LysoPC(18:0), LysoPC(P18:0), LysoPC(20:3)), which were detected in higher concentration after the nutritional supplementation. Among them, LysoPC(17:0) was the most increased (Fold Change: +2.29). Thus, this is clear evidence that the consumption of the garlic food supplement altered the phospholipid metabolism. In this sense, Lysophosphatidylcholines (LPC) are bioactive phospholipids which are originated

by the hydrolysis of phosphatidylcholines (PC) mediated by the phospholipase A_2 in living cells (Jackson, Abate, & Tonks, 2008). Another way of producing LysoPC, which is highly related to blood concentration, is derived from the reaction of PC and cholesterol in liver by means of lecithin:cholesterol acyltranferase (LCAT) enzyme (Rousset, Vaisman, Amar, Sethi, & Remaley, 2009).

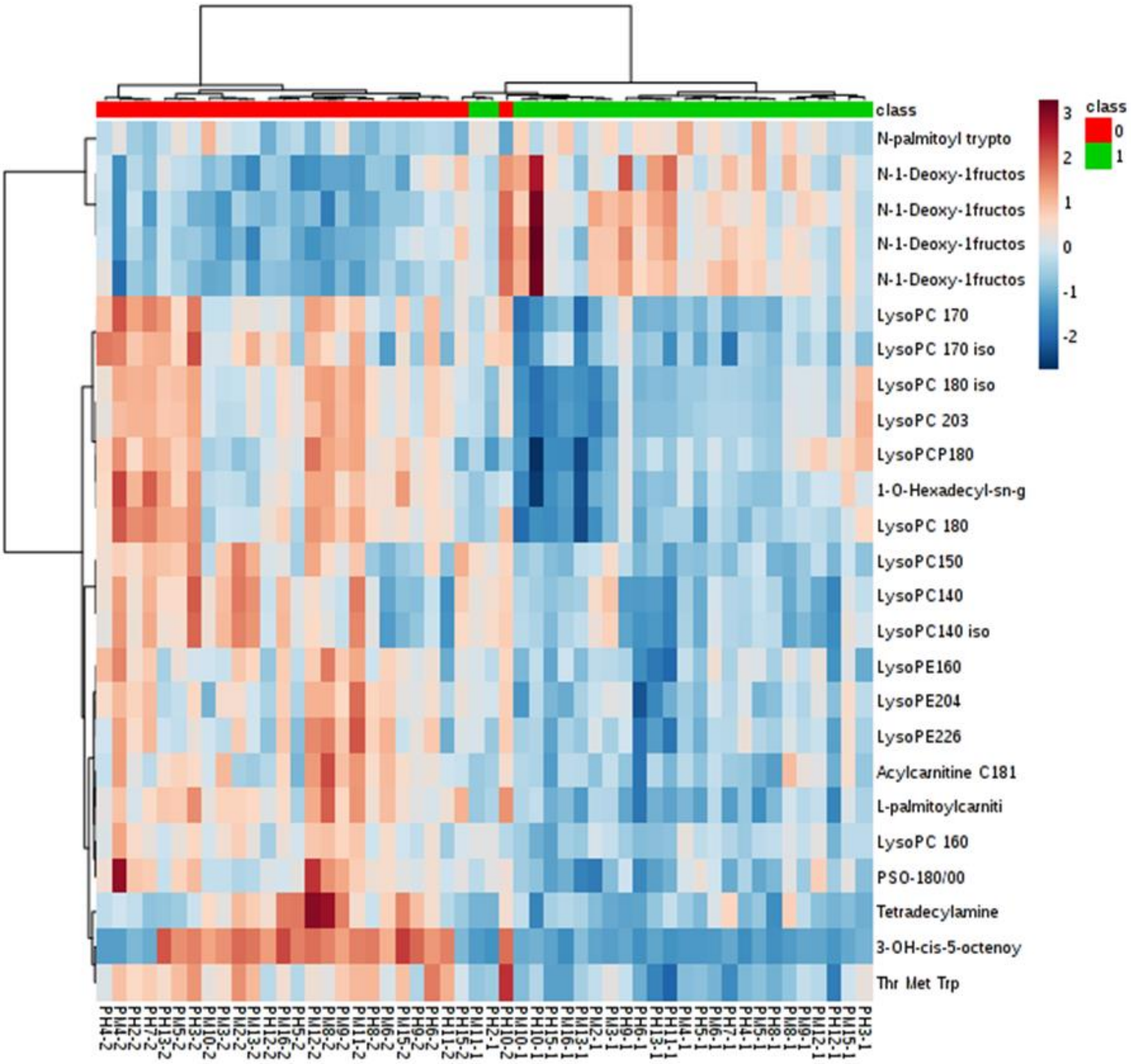


Figure 4. Hierarchical clustering via heatmap (Pearson and Ward as distance measure and clustering algorithm) of the 25 significant identified metabolites. (0: pre-treatment, 1: post-treatment).

The relationship of lysophosphatidylcholines with inflammatory processes and with the modulation of the immune response is well-known (J. H. S. Kabarowski, Xu, & Witte, 2002). The deregulation of LysoPC concentration in plasma has been reported in bibliography in numerous studies mainly focused on diseases. In this sense, it has been found that lower concentrations of LysoPCs are related to a higher risk of several types of cancer, such as prostate, breast or colorectal cancer (Kühn et al., 2016; Zhao et al., 2007). Moreover, same trends have also been reported for other types of diseases such as Alzheimer (Y. et al., 2014), obesity or Type 2 Diabetes (Barber et al., 2012).

On the other hand, there are studies which have related the increase of LysoPC in plasma to the consumption of nutritional supplement in humans, as the case of vitamin E (Wong & Lodge, 2012). In fact, vitamin E has been described as a component of garlic, which has showed bioactive properties with beneficial effects on human health. Among them, it could be highlighted its antioxidant effect and its functions as cardioprotective agent, regulator of specific gene expression and reducer of inflammation and oxidative stress (Rizvi et al., 2014).

In addition, LysoPCs have been implicated in the immune response as an immunoregulation factor. It has been studied how concentration of LysoPCs changes during the immune response, and in these cases a decrease of concentration has been found as the general tendency (Wikoff, Kalisak, Trauger, Manchester, & Siuzdak, 2009). According to that, the observed decrease of this family of compounds has been observed in autoimmune diseases such as multiple sclerosis (Del Boccio et al., 2011). However, there is a bit of controversy due to other studies have shown the opposite trend for several autoimmune diseases such as atherosclerosis and systemic lupus erythematous (J. H. Kabarowski, 2009). Therefore, nowadays there is no clear relationship between these compounds and the immune response. Nevertheless, it seems that the consumption of the garlic supplement could probably cause immunomodulatory effects because of its content in organosulfur compounds, specially enriched in propyl propane thiosulfate (PTSO), which has previously showed properties for modulating the immune response in humans (PCT/ES2014/070928, 2014). In this way, the observed deregulation of LysoPCs could be a metabolic consequence caused by the immunomodulatory capacity of the PTSO derivatives in the organism.

Moreover, other significant metabolites are also related to phospholipid metabolism, as the case of acylcarnitines (L-palmitoylcarnitine, 3-OH-cis-5-octenoylcarnitine, L-oleoylcarnitine), 1-O-Hexadecyl-sn-glycero-3-phosphocholine and lysophospatidylethanolamines (LysoPE(16:0), LysoPE(20:4), LysoPE(22:6)). Similarly, the level of these metabolites also increases after the supplementation.

Acylcarnitines are metabolites whose function is the transport of the fatty acids into the mitochondria, and therefore, they are associated with the metabolism of fatty acids and amino acids oxidation. The increase of this kind of compounds in plasma has been associated to deregulation in fatty acid oxidation and with higher risk of certain diseases, such as cardiovascular diseases and diabetes (Stephanie J. Mihalik et al., 2010) although this risk depends on the length of the chain, showing no significant differences in long-chain acylcarnitines (S. J. Mihalik et al., 2012). In addition, previous studies have shown that garlic extracts decrease the risk of cardiovascular diseases due to the inhibition of lipid oxidation and oxidation of low density lipoprotein (Iciek, Kwiecień, & Włodek, 2009; Lau, 2006). Nevertheless, in other illnesses, like adult celiac disease, it has been found a decrease of these metabolites with respect to the controls (Bene et al., 2005). In spite of that, it is important to highlight that under normal conditions the up-regulation of long-chain acylcarnitines can be caused by the fatty acid composition of the dietary intake. Therefore, the observed increase could be associated with the fatty acid composition of the food supplement. In fact, palmitic acid has been found as the major fatty acid in garlic (Tsiaganis, Laskari, & Melissari, 2006) and, in addition, studies have showed that the higher concentration of acylcarnitines due to palmitic acid intake is not associated to deregulation of β-oxidation (Kien et al., 2015). On the other hand, studies have revealed bioactive properties of palmitoylcarnitine, which has been described as a local immunomodulatory and antibacterial molecule (Hulme et al., 2017; Wenderska, Chong, McNulty, Wright, & Burrows, 2011). Therefore, it is not clear the impact that the increase of these molecules could have in the human metabolism of healthy people.

On the other hand, it is also remarkable the decrease in the concentration of four fructosamines (1-amino-1-deoxy-D-fructose) metabolites with Valine, Trypthophan, Leucine and Isoleucine as the aminoacid part. These compounds are called Amadori products, which are made up by the early stage of the Maillard reaction. The formation of fructosamines occur both enzymatically and non-enzymatically and is commonly produced in foods and *in-vivo* (Mossine & Mawhinney, 2010). The decrease of concentration of these metabolites is in accordance with the antiglycation properties which have been found for garlic extracts. Thus, these extracts are able to inhibit the formation of early glycation products like fructosamines (Elosta, Slevin, Rahman, & Ahmed, 2017). The concentration of fructosamine derivatives is directly related to the advanced glycation end products (AGE), which have attracted a great interest in recent years due to the relationship between their high concentration and diseases related to aging and diabetes (Gkogkolou & Böhm, 2012). In this way, previous studies have showed the positive effect of garlic supplement in the management of type 2 diabetes mellitus. (Padiya & K. Banerjee, 2013; J. Wang, Zhang, Lan, & Wang, 2017).

According to the obtained results, it can be concluded that the bioactive garlic food supplement alters mainly the phospholipid metabolism in healthy people. Among the deregulate compounds, it is remarkable the concentration increase of acylcarnitines, lysophosphatidylcholines and lysophospathidylethanolamines, joined with the decrease of fructosamines after the supplement intake. These observed alterations could be highly correlated with the antioxidant and antiglycation properties that have been previously described for garlic extracts. In addition, most of the observed tendencies are opposed to disease states. Among them, it should be highlighted the increase in concentration of 7 lysoPCs, whose increase have demonstrated a relationship with the prevention of several diseases, especially cancer. The obtained results suggest that the prolonged intake of a garlic food supplement have an effect in several metabolic pathways in healthy humans. These finding have been related to the bioactive properties of garlic against several diseases. In this sense, further similar studies using patients of these diseases, should be conducted in order to increase the knowledge about assuring the beneficial effects of garlic.

Ethics statement

The authors confirm that any aspect of the work covered in this manuscript that has involved human volunteers has been conducted according to The Code of Ethics of the World Medical Association (Declaration of Helsinki) and with the ethical approval of all relevant bodies and that such approval are acknowledged within the manuscript. In addition, all authors declare that:

- All volunteers signed a consent form after receiving a detailed explanation of the study.
- Confidentiality of all the collected data in the framework of the study is completely ensured.
- The study have been fulfilled all the requirements established on the protection of personal character data by the legislation in force (Statutory law 15/1999 of 13 of December of Protection of Personal Character Data).

Acknowledgements

This work was co-financed by the FEDER program of the Andalusian Innovation and Development Agency (Project 461082). The author IBL gratefully acknowledges the Spanish Ministry of Economy and Competitiveness (MINECO) in association with the European Social Fund (FSE) for the contract PTQ-13-06429, the author AFO received support from the Spanish Ministry of Education, Culture and Sports (FPU grant 14/03992).

Conflict of interest

All authors declare that they have no conflict of interest.

REFERENCES

Agin, A., Heintz, D., Ruhland, E., Chao de la Barca, J. M., Zumsteg, J., Moal, V., … Namer, I. J. (2016, February). Metabolomics - an overview. From basic principles to potential biomarkers (part 1). *Medecine Nucleaire*. https://doi.org/10.1016/j.mednuc.2015.12.006

Amagase, H., Petesch, B., Matsuura, H., Kasuga, S., & Itakura, Y. (2001). Intake of garlic and its bioactive components. *The Journal of Nutrition*, *131*(3), 955–962.

Baños Arjona, A., Galvez Peralta, J. J., Guillamon Ayala, E., Nunez Lechado, C., Maroto Caba, F., & Rodriguez Cabezas, M. E. (2014). *PCT/ES2014/070928*. Retrieved from https://www.google.com/patents/CA2935431A1?cl=en

Barber, M. N., Risis, S., Yang, C., Meikle, P. J., Staples, M., Febbraio, M. A., & Bruce, C. R. (2012). Plasma lysophosphatidylcholine levels are reduced in obesity and type 2 diabetes. *PLoS ONE*, *7*(7), e41456. https://doi.org/10.1371/journal.pone.0041456

Bene, J., Komlósi, K., Gasztonyi, B., Juhász, M., Tulassay, Z., & Melegh, B. (2005). Plasma carnitine ester profile in adult celiac disease patients maintained on long-term gluten free diet. *World Journal of Gastroenterology*, *11*(42), 6671–6675. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16425363

Brennan, L. (2013). Metabolomics in nutrition research: current status and perspectives: Figure 1. *Biochemical Society Transactions*, *41*(2), 670–673. https://doi.org/10.1042/BST20120350

Bresciani, L., Calani, L., Cossu, M., Mena, P., Sayegh, M., Ray, S., & Del Rio, D. (2015). (Poly)phenolic characterization of three food supplements containing 36 different fruits, vegetables and berries. *PharmaNutrition*, *3*(2), 11–19. https://doi.org/10.1016/J.PHANU.2015.01.001

Bruce, S. J., Tavazzi, I., Parisod, V., Rezzi, S., Kochhar, S., & Guy, P. a. (2009). Investigation of human blood plasma sample preparation for performing metabolomics using ultrahigh performance liquid chromatography/mass spectrometry. *Analytical Chemistry*, *81*(9), 3285–3296. https://doi.org/10.1021/ac8024569

Del Boccio, P., Pieragostino, D., Di Ioia, M., Petrucci, F., Lugaresi, A., De Luca, G., … Urbani, A. (2011). Lipidomic investigations for the characterization of circulating serum lipids in multiple sclerosis. *Journal of Proteomics*, *74*(12), 2826–2836. https://doi.org/10.1016/j.jprot.2011.06.023

Del Rio, D., Rodriguez-Mateos, A., Spencer, J. P. E., Tognolini, M., Borges, G., & Crozier, A. (2013). Dietary (Poly)phenolics in Human Health: Structures, Bioavailability, and Evidence of Protective Effects Against Chronic Diseases. *Antioxidants & Redox Signaling*, *18*(14), 1818–1892. https://doi.org/10.1089/ars.2012.4581

Dettmer, K., Aronov, P. A., & Hammock, B. D. (2007). Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, *26*(1), 51–78. https://doi.org/10.1002/mas.20108

Dunn, W. B., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N., … Goodacre, R. (2011). Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature Protocols*, *6*(7), 1060–1083. https://doi.org/10.1038/nprot.2011.335

Dunn, W. B., Wilson, I. D., Nicholls, A. W., & Broadhurst, D. (2012). The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. *Bioanalysis*, *4*(18), 2249–2264. https://doi.org/10.4155/bio.12.204

Elosta, A., Slevin, M., Rahman, K., & Ahmed, N. (2017). Aged garlic has more potent antiglycation and antioxidant properties compared to fresh garlic extract in vitro. *Scientific Reports*, *7*, 39613. https://doi.org/10.1038/srep39613

Farag, M., Ali, S., Hodaya, R., El-Seedi, H., Sultani, H., Laub, A., … Wessjohann, L. (2017). Phytochemical Profiles and Antimicrobial Activities of Allium cepa Red cv. and A. sativum Subjected to Different Drying Methods: A Comparative MS-Based

Metabolomics. *Molecules*, *22*(5), 761. https://doi.org/10.3390/molecules22050761

Farag, M., Fekry, M., Al-Hammady, M., Khalil, M., El-Seedi, H., Meyer, A., … Wessjohann, L. (2017). Cytotoxic Effects of Sarcophyton sp. Soft Corals—Is There a Correlation to Their NMR Fingerprints? *Marine Drugs*, *15*(7), 211. https://doi.org/10.3390/md15070211

Gibbons, H., O'Gorman, A., & Brennan, L. (2015). Metabolomics as a tool in nutritional research. *Current Opinion in Lipidology*, *26*(1), 30–34. https://doi.org/10.1097/MOL.0000000000000140

Gika, H. G., Macpherson, E., Theodoridis, G. A., & Wilson, I. D. (2008). Evaluation of the repeatability of ultra-performance liquid chromatography-TOF-MS for global metabolic profiling of human urine samples. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*, *871*(2), 299–305. https://doi.org/10.1016/j.jchromb.2008.05.048

Gkogkolou, P., & Böhm, M. (2012). Advanced glycation end products: Key players in skin aging? *Dermato-Endocrinology*, *4*(3), 259–270. https://doi.org/10.4161/derm.22028

Hulme, H. E., Meikle, L. M., Wessel, H., Strittmatter, N., Swales, J., Thomson, C., … Wall, D. M. (2017). Mass spectrometry imaging identifies palmitoylcarnitine as an immunological mediator during Salmonella Typhimurium infection. *Scientific Reports*, *7*(1), 2786. https://doi.org/10.1038/s41598-017-03100-5

Iciek, M., Kwiecień, I., & Włodek, L. (2009). Biological properties of garlic and garlic-derived organosulfur compounds. *Environmental and Molecular Mutagenesis*, *50*(3), 247–265. https://doi.org/10.1002/em.20474

Jackson, S. K., Abate, W., & Tonks, A. J. (2008). Lysophospholipid acyltransferases: Novel potential regulators of the inflammatory response and target for new drug discovery. *Pharmacology and Therapeutics*. https://doi.org/10.1016/j.pharmthera.2008.04.001

Jacobs, D. M., Fuhrmann, J. C., Van Dorsten, F. A., Rein, D., Peters, S., Van Velzen, E. J.

J., … Garczarek, U. (2012). Impact of short-term intake of red wine and grape polyphenol extract on the human metabolome. *Journal of Agricultural and Food Chemistry*, *60*(12), 3078–3085. https://doi.org/10.1021/jf2044247

Johnson, C. H., Ivanisevic, J., & Siuzdak, G. (2016). Metabolomics: Beyond biomarkers and towards mechanisms. *Nature Reviews Molecular Cell Biology*, *17*(7). https://doi.org/10.1038/nrm.2016.25

Kabarowski, J. H. (2009, September). G2A and LPC: Regulatory functions in immunity. *Prostaglandins and Other Lipid Mediators*. https://doi.org/10.1016/j.prostaglandins.2009.04.007

Kabarowski, J. H. S., Xu, Y., & Witte, O. N. (2002, July 15). Lysophosphatidylcholine as a ligand for immunoregulation. *Biochemical Pharmacology*. https://doi.org/10.1016/S0006-2952(02)01179-6

Kamleh, M. A., Ebbels, T. M. D., Spagou, K., Masson, P., & Want, E. J. (2012). Optimizing the use of quality control samples for signal drift correction in large-scale urine metabolic profiling studies. *Analytical Chemistry*, *84*(6), 2670–2677. https://doi.org/10.1021/ac202733q

Khymenets, O., Andres-Lacueva, C., Urpi-Sarda, M., Vazquez-Fresno, R., Mart, M. M., Reglero, G., … Llorach, R. (2015). Metabolic fingerprint after acute and under sustained consumption of a functional beverage based on grape skin extract in healthy human subjects. *Food Funct.*, *6*(4), 1288–1298. https://doi.org/10.1039/C4FO00684D

Kien, C. L., Matthews, D. E., Poynter, M. E., Bunn, J. Y., Fukagawa, N. K., Crain, K. I., … Muoio, D. M. (2015). Increased palmitate intake: higher acylcarnitine concentrations without impaired progression of β-oxidation. *Journal of Lipid Research*, *56*(9), 1795–1807. https://doi.org/10.1194/jlr.M060137

Kitawa, N., Fischer, S. M., Roark, J., Samant, M., Sana, T., & Rane, A. (2013). A Novel Two-Pass Feature Statistical Profiling of Mass. In *ASMS*.

Kopec, A., Piatkowska, E., Leszczynska, T., & Sikora, E. (2013). Healthy Properties of Garlic. *Current Nutrition & Food Science*, *9*(4), 59–64.

https://doi.org/10.2174/1573401311309010010

Kühn, T., Floegel, A., Sookthai, D., Johnson, T., Rolle-Kampczyk, U., Otto, W., … Kaaks, R. (2016). Higher plasma levels of lysophosphatidylcholine 18:0 are related to a lower risk of common cancers in a prospective metabolomics study. *BMC Medicine*, *14*(1), 13. https://doi.org/10.1186/s12916-016-0552-3

Lau, B. H. S. (2006). Suppression of LDL Oxidation by Garlic Compounds Is a Possible Mechanism of Cardiovascular Health Benefit. *The Journal of Nutrition*, *136*(3), 765S–768S. https://doi.org/10.1093/jn/136.3.765S

Mihalik, S. J., Goodpaster, B. H., Kelley, D. E., Chace, D. H., Vockley, J., Toledo, F. G. S., & DeLany, J. P. (2010). Increased levels of plasma acylcarnitines in obesity and type 2 diabetes and identification of a marker of glucolipotoxicity. *Obesity (Silver Spring, Md.)*, *18*(9), 1695–1700. https://doi.org/10.1038/oby.2009.510

Mihalik, S. J., Michaliszyn, S. F., de las Heras, J., Bacha, F., Lee, S., Chace, D. H., … Arslanian, S. A. (2012). Metabolomic Profiling of Fatty Acid and Amino Acid Metabolism in Youth With Obesity and Type 2 Diabetes: Evidence for enhanced mitochondrial oxidation. *Diabetes Care*, *35*(3), 605–611. https://doi.org/10.2337/DC11-1577

Mizuno, H., Ueda, K., Kobayashi, Y., Tsuyama, N., Todoroki, K., Min, J. Z., & Toyo'oka, T. (2017). The great importance of normalization of LC-MS data for highly-accurate non-targeted metabolomics. *Biomedical Chromatography*, *31*(1), 1–7. https://doi.org/10.1002/bmc.3864

Mossine, V. V., & Mawhinney, T. P. (2010). *1-Amino-1-deoxy-d-fructose ("Fructosamine") and its Derivatives*. *Advances in Carbohydrate Chemistry and Biochemistry* (Vol. 64). https://doi.org/10.1016/S0065-2318(10)64006-1

Mumtaz, M. W., Hamid, A. A., Akhtar, M. T., Anwar, F., Rashid, U., & AL-Zuaidy, M. H. (2017). An overview of recent developments in metabolomics and proteomics – phytotherapic research perspectives. *Frontiers in Life Science*, *10*(1), 1–37. https://doi.org/10.1080/21553769.2017.1279573

Padiya, R., & K. Banerjee, S. (2013). Garlic as an Anti-diabetic Agent: Recent Progress and Patent Reviews. *Recent Patents on Food, Nutrition & Agriculture*, *5*(2), 105–127. https://doi.org/10.2174/18761429113059990002

Rizvi, S., Raza, S. T., Ahmed, F., Ahmad, A., Abbas, S., & Mahdi, F. (2014). The role of Vitamin E in human health and some diseases. *Sultan Qaboos University Medical Journal*, *14*(2), 157–165.

Rodriguez-Mateos, A., Vauzour, D., Krueger, C. G., Shanmuganayagam, D., Reed, J., Calani, L., … Crozier, A. (2014). Bioavailability, bioactivity and impact on health of dietary flavonoids and related compounds: an update. *Archives of Toxicology*, *88*(10), 1803–1853. https://doi.org/10.1007/s00204-014-1330-7

Rousset, X., Vaisman, B., Amar, M., Sethi, A. A., & Remaley, A. T. (2009). Lecithin:Cholesterol Acyltransferase: From Biochemistry to Role in Cardiovascular Disease. *Curr Opin Endocrinol Diabetes*, *16*, 163–171.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models : A framework for some traditional and novel measures. *Epidemiology*, *21*(1), 128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2.Assessing

Tsiaganis, M. C., Laskari, K., & Melissari, E. (2006). Fatty acid composition of Allium species lipids. *Journal of Food Composition and Analysis*, *19*(6–7), 620–627. https://doi.org/10.1016/j.jfca.2005.06.003

van den Berg, R. A., Hoefsloot, H. C. J., Westerhuis, J. A., Smilde, A. K., & van der Werf, M. J. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, *7*, 142. https://doi.org/10.1186/1471-2164-7-142

Wang, J., Zhang, X., Lan, H., & Wang, W. (2017). Effect of garlic supplement in the management of type 2 diabetes mellitus (T2DM): a meta-analysis of randomized controlled trials. *Food & Nutrition Research*, *61*(1), 1377571. https://doi.org/10.1080/16546628.2017.1377571

Wang, X., Chen, S., & Jia, W. (2016, December 11). Metabolomics in Cancer Biomarker Research. *Current Pharmacology Reports*. Springer International Publishing.

https://doi.org/10.1007/s40495-016-0074-x

Wenderska, I. B., Chong, M., McNulty, J., Wright, G. D., & Burrows, L. L. (2011). Palmitoyl-dl-Carnitine is a Multitarget Inhibitor of Pseudomonas aeruginosa Biofilm Development. *ChemBioChem*, *12*(18), 2759–2766. https://doi.org/10.1002/cbic.201100500

Wikoff, W. R., Kalisak, E., Trauger, S., Manchester, M., & Siuzdak, G. (2009). Response and recovery in the plasma metabolome tracks the acute LCMV-induced immune response. *Journal of Proteome Research*, *8*(7), 3578–3587. https://doi.org/10.1021/pr900275p

Wong, M., & Lodge, J. K. (2012). A metabolomic investigation of the effects of vitamin E supplementation in humans. *Nutrition & Metabolism*, *9*(1), 110. https://doi.org/10.1186/1743-7075-9-110

Worley, B., & Powers, R. (2013). Multivariate Analysis in Metabolomics. *Current Metabolomics*, *1*(1), 92–107. https://doi.org/10.2174/2213235X11301010092

Xia, J., Broadhurst, D. I., Wilson, M., & Wishart, D. S. (2013). Translational biomarker discovery in clinical metabolomics: An introductory tutorial. *Metabolomics*, *9*(2), 280–299. https://doi.org/10.1007/s11306-012-0482-9

Xia, J., Sinelnikov, I. V, Han, B., Wishart, D. S., R., T., J., X., … T., P. (2015). MetaboAnalyst 3.0—making metabolomics more meaningful. *Nucleic Acids Research*, *43*(W1), W251–W257. https://doi.org/10.1093/nar/gkv380

Xia, J., & Wishart, D. S. (2011). Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. *Nature Protocols*, *6*(6), 743–760. https://doi.org/10.1038/nprot.2011.319

Xia, J., & Wishart, D. S. (2016). Using MetaboAnalyst 3.0 for Comprehensive Metabolomics Data Analysis. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]*, *55*(September), 14.10.1-14.10.91. https://doi.org/10.1002/cpbi.11

Y., C., X., L., M., W., L., L., X., S., L., M., … Q., W. (2014). Lysophosphatidylcholine and amide as metabolites for detecting Alzheimer disease using ultrahigh-performance liquid chromatography-quadrupole time-of-flight mass spectrometry-based metabonomics. *Journal of Neuropathology and Experimental Neurology*, *73*(10), 954–963. https://doi.org/10.1097/NEN.0000000000000116

Zhao, Z., Xiao, Y., Elson, P., Tan, H., Plummer, S. J., Berk, M., … Xu, Y. (2007). Plasma lysophosphatidylcholine levels: Potential biomarkers for colorectal cancer. *Journal of Clinical Oncology*, *25*(19), 2696–2701. https://doi.org/10.1200/JCO.2006.08.5571

Zheng, H., Clausen, M. R., Dalsgaard, T. K., & Bertram, H. C. (2015). Metabolomics to explore impact of dairy intake. *Nutrients*, *7*(6), 4875–4896. https://doi.org/10.3390/nu7064875

Supplementary Material

Table 1S. Molecular and statistical details (Retention times, masses, molecular formulas, scores, p-values, Vip-values and Areas under ROC curve) of unknown compounds that presented significant differences between before and after garlic supplement intake.

	RT (min)	Mass (Da)	p-value	VIP value	AUROC	Molecular Formula
Unknown Compounds	21.4	301.2061	4.26 E-7	2.3773	0.9135	$C_{14}H_{27}N_3O_4$
	30.9	624.4307	1.13 E-6	2.1898	0.8420	$C_{28}H_{60}N_6O_9$
	31.5	479.327	4.87 E-8	2.1088	0.8970	$C_{22}H_{41}N_9O_3$
	31.7	547.3468	4.59 E-6	1.6844	0.7775	$C_{22}H_{41}N_{15}O_2$
	33.1	626.4535	4.36 E-12	3.4954	0.9657	$C_{25}H_{54}N_{16}O_3$
	33.4	562.4108	2.82 E-7	1.6413	0.7514	$C_{29}H_{59}N_2O_6P$
	34.5	468.3665	1.10 E-4	1.9808	0.7582	$C_{28}H_{52}O_5$
	27.3	414.2768	8.57 E-4	1.9808	0.7582	$C_{22}H_{34}N_6O_2$
	27.3	557.2541	4.79 E-4	1.1592	0.7710	$C_{25}H_{35}N_9O_6$
	30.2	1508.4322	6.15 E-7	3.1487	0.9753	$C_{46}H_{81}N_{18}O_{37}P$
	30.2	1013.1272	7.31 E-7	1.6635	0.8860	$C_{36}H_{76}N_{28}O_7$
	30.2	1485.9696	2.35 E-13	2.9535	0.9767	$C_{54}H_{18}N_{14}O_{39}$

BLOQUE B

Desarrollo y aplicación de estrategias metabolómicas para el estudio de enfermedades autoinmunes sistémicas

1. Enfermedades autoinmunes sistémicas

Las **enfermedades autoinmunes** se caracterizan por fallos en el reconocimiento de agentes patógenos por parte del sistema inmunitario de un organismo, de tal forma que se produce la identificación de las propias moléculas del organismo como dichos agentes patógenos, desencadenando un ataque contra ellas por parte del sistema inmunológico (**Figura 35**). El origen de estas enfermedades, aunque en muchos casos todavía se desconoce, suele deberse a factores infecciosos, medioambientales, hormonales y/o genéticos, siendo estos últimos los más predominantes. Estas enfermedades, que presentan un significativo índice de morbilidad y mortalidad, afectan aproximadamente al 5 % de la población mundial, correspondiendo el 75 % de los casos a mujeres[161].

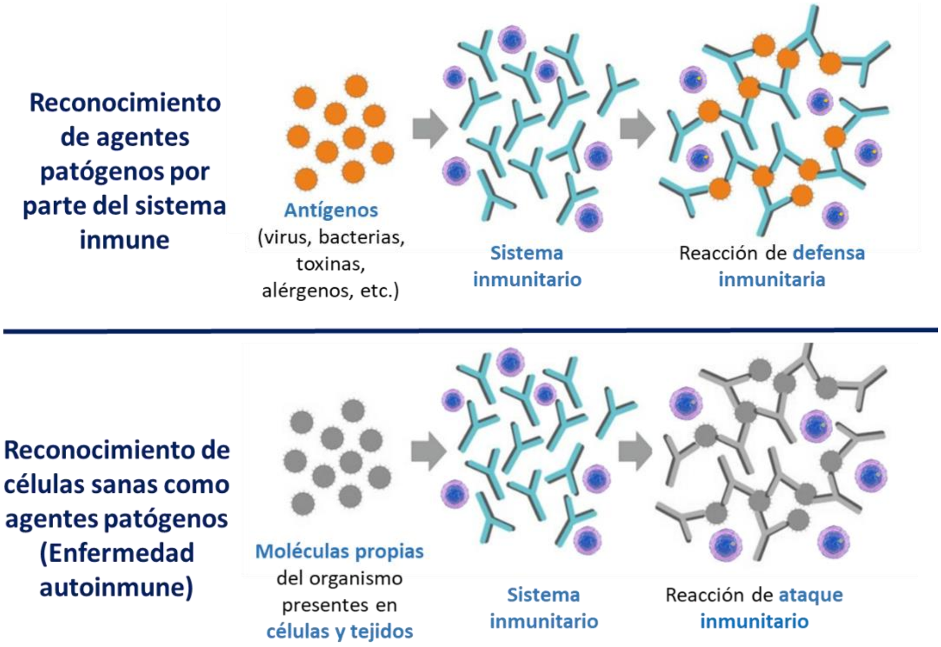

Reconocimiento de agentes patógenos por parte del sistema inmune
Antígenos (virus, bacterias, toxinas, alérgenos, etc.)
Sistema inmunitario
Reacción de defensa inmunitaria
Reconocimiento de células sanas como agentes patógenos (Enfermedad autoinmune)
Moléculas propias del organismo presentes en células y tejidos
Sistema inmunitario
Reacción de ataque inmunitario

Figura 35. Sistema inmunológico frente a agentes patógenos y células sanas (enfermedad autoinmune).

[161] Shashi Singh et al., *UNDERSTANDING AUTOIMMUNE DISEASE: AN UPDATE REVIEW*, *International Journal of Pharmaceutical Technology and Biotechnology*, vol. 3, 2016.

Existen más de 80 enfermedades autoinmunes clasificadas en órgano específicas y sistémicas o multiorgánicas. Las **enfermedades autoinmunes órgano específicas** se caracterizan por el ataque, por parte del sistema inmune, focalizado en los antígenos expresados en un único órgano, mientras que las **enfermedades autoinmunes sistémicas** son aquellas en las que varios órganos y/o tejidos son afectados sin tener muchas veces relación aparente entre ellos. En la **Tabla 6** se muestran varios ejemplos de las enfermedades autoinmunes que más incidencia tienen dentro de la población[162].

Tabla 6. Ejemplos de enfermedades autoinmunes más comunes.

Órgano-específicas	Sistémicas o multiorgánicas
Diabetes mellitus tipo 1	Lupus eritematoso sistémico
Hepatitis autoinmune	Síndrome de Sjögren
Enfermedad de Addison	Artritis reumatoide
Tiroiditis de Hashimoto	Esclerosis sistémica
Enfermedad de Crohn	Polimositis
Anemia perniciosa	Síndrome antifosfolípido
Miastenia grave	Dermatomiositis
Enfermedad de Graves-Basedow	Psoriasis

En concreto, las **enfermedades autoinmunes sistémicas** (**SADS**) son un grupo de patologías inflamatorias crónicas que presentan generalmente grandes dificultades en sus diagnósticos y además existen pocas alternativas para su tratamiento, siendo en su mayoría costosos y con un gran número de efectos secundarios. Las dificultades en el diagnóstico de estas enfermedades provocan la existencia de diagnósticos

[162] Helena Beatriz Ferreira et al., "Lipidomics in Autoimmune Diseases with Main Focus on Systemic Lupus Erythematosus," *Journal of Pharmaceutical and Biomedical Analysis* 174 (September 10, 2019): 386–95, https://doi.org/10.1016/J.JPBA.2019.06.005.

equivocados, siendo los pacientes erróneamente clasificados. Esto ocurre de manera más frecuente de la deseada debido a que los métodos de diagnóstico se basan principalmente en la presencia de síntomas y la detección de autoanticuerpos inespecíficos en muestras de suero[163,164].

Estas enfermedades están representadas principalmente por el **lupus eritematoso sistémico** (SLE), la **artritis reumatoide** (RA) y la **esclerosis sistémica** (SSC). Otros síndromes y enfermedades, como por ejemplo el **síndrome de Sjögren** (SjS), la **enfermedad mixta del tejido conectivo** (MCTD) o el **síndrome antifosfolípido** (PAPS), presentan una superposición clínica con las tres patologías representativas, ocasionando en muchas ocasiones dificultades en el diagnóstico de los pacientes. Este tipo de patologías son catalogadas como enfermedades raras, aunque en su conjunto representan al 1 % de la población. Además, hay pacientes que no cumplen con todos los criterios clínicos representativos de estas afecciones ni presentan los síntomas característicos para el diagnóstico de una patología concreta, y son asignados como casos pertenecientes a la **enfermedad del tejido conectivo indiferenciado** (UCTD) (**Figura 36**)[165,166].

[163] Guixiu Shi et al., "Systemic Autoimmune Diseases.," *Clinical & Developmental Immunology* 2013 (2013): 728574, https://doi.org/10.1155/2013/728574.

[164] Christine Castro and Mark Gourley, "Diagnostic Testing and Interpretation of Tests for Autoimmunity.," *The Journal of Allergy and Clinical Immunology* 125, no. 2 Suppl 2 (February 2010): S238-47, https://doi.org/10.1016/j.jaci.2009.09.041.

[165] Margarida Antunes et al., "Undifferentiated Connective Tissue Disease: State of the Art on Clinical Practice Guidelines," *RMD Open* 4, no. Suppl 1 (February 1, 2019): e000786, https://doi.org/10.1136/rmdopen-2018-000786.

[166] Dan Radulescu et al., "A Rare Case of Systemic Autoimmune Disease with Intricate Features of Systemic Sclerosis, Lupus, Polymyositis and Rheumatoid Arthritis. Overlap Syndrome or Mixed Connective Tissue Disease?," *Acta Reumatologica Portuguesa* 32, no. 3: 292–97, accessed August 28, 2019, http://www.ncbi.nlm.nih.gov/pubmed/17932479.

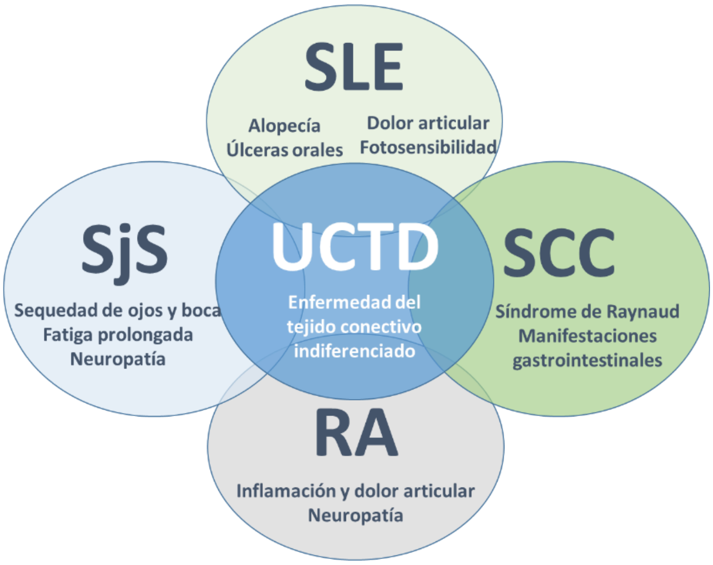


Figura 36. Síntomas de la enfermedad del tejido conectivo indiferenciado en relación con otras enfermedades autoinmunes sistémicas (SLE, RA, SjS y SSC).

Como ya se ha mencionado, las tasas de mortalidad y morbilidad asociadas con estas patologías son altas, además el diseño de terapias y tratamientos está actualmente limitado por el escaso conocimiento acerca de los complejos mecanismos de acción involucrados en las diferentes patologías. Por tanto, existe una necesidad real de incrementar el entendimiento acerca de la patogénesis de estas enfermedades, así como de buscar biomarcadores que permitan una nueva clasificación molecular de las mismas, de forma que se logre crear herramientas de diagnóstico y pronóstico potentes así como el diseño de terapias más efectivas que detengan la progresión de estas enfermedades dentro del concepto de medicina individualizada. Para ello, las ciencias ómicas, y entre ellas la metabolómica, presentan actualmente una alta potencialidad para aumentar el conocimiento de estas enfermedades y afrontar los retos actuales que presentan este tipo de patologías[167,168].

A continuación se describen brevemente las características principales de las siete enfermedades autoinmunes sistémicas estudiadas en la presente tesis doctoral. La **Figura 37** muestra alguno de los signos más característicos de estas enfermedades.

- El **lupus eritematoso sistémico (SLE)** es una enfermedad inflamatoria crónica muy heterogénea que presenta una amplia cantidad de manifestaciones clínicas que incluyen multitud de órganos y tejidos afectados, siendo los más frecuentes las articulaciones, la piel (lupus eritematoso cutáneo), los riñones (nefritis lúpica), el sistema circulatorio, los pulmones o el sistema nervioso central. El desarrollo de esta enfermedad se caracteriza por períodos alternos de aumento y remisión de los síntomas[169].
- La **esclerosis sistémica** progresiva o esclerodermia **(SSC)**, es una enfermedad del tejido conectivo que involucra cambios en la piel, vasos sanguíneos, músculos y órganos internos. Los síntomas se corresponden con la acumulación de colágeno en la piel y otros órganos. La esclerodermia se clasifica

[167] Maria Teruel, Chris Chamberlain, and Marta E. Alarcón-Riquelme, "Omics Studies: Their Use in Diagnosis and Reclassification of SLE and Other Systemic Autoimmune Diseases," *Rheumatology* 56, no. suppl_1 (October 19, 2016): kew339, https://doi.org/10.1093/rheumatology/kew339.

[168] "Precisesads, Investigating Systemic Autoimmune Diseases," accessed August 27, 2019, http://www.precisesads.eu/.

[169] Arvind Kaul et al., "Systemic Lupus Erythematosus," *Nature Reviews Disease Primers* 2, no. 1 (December 16, 2016): 16039, https://doi.org/10.1038/nrdp.2016.39.

principalmente en dos subconjuntos: esclerosis sistémica cutánea difusa (dcSSC) y esclerosis sistémica cutánea limitada (lcSSC)[170].

- La **artritis reumatoide (RA)** es una enfermedad inflamatoria crónica que afecta a muchos tejidos y órganos, pero principalmente ataca a las articulaciones sinoviales, produciendo un excedente de líquido sinovial. Esta patología puede generar daño severo en el cartílago, hueso o ligamentos articulares, provocando la reducción del movimiento de las articulaciones (anquilosis). Además, se puede producir otras manifestaciones extraarticulares que afectan principalmente a los pulmones, el corazón, los vasos sanguíneos o la piel[171].
- El **síndrome de Sjögren (SjS)**, es una enfermedad compleja caracterizada por la infiltración linfocítica de las glándulas exocrinas, como las glándulas salivales y lagrimales, conduciendo a sequedad en la boca y ojos. Además, debido a una activación generalizada del sistema inmunológico, el SjS también puede afectar a varios órganos y/o tejidos como la piel, el corazón, los pulmones, el sistema nervioso o los riñones, entre otros. El SjS puede aparecer de manera aislada (SjS primario) o acompañado de otras enfermedades autoinmunes (SjS secundario), como el lupus, la artritis reumatoide o esclerosis sistémica[172].

[170] Anne Claire Desbois and Patrice Cacoub, "Systemic Sclerosis: An Update in 2016," *Autoimmunity Reviews* 15, no. 5 (2016): 417–26, https://doi.org/10.1016/j.autrev.2016.01.007.

[171] Arsenio Spinillo et al., "Undifferentiated Connective Tissue Diseases and Adverse Pregnancy Outcomes. An Undervalued Association?," *American Journal of Reproductive Immunology* 78, no. 6 (December 2017): e12762, https://doi.org/10.1111/aji.12762.

[172] Frederick B. Vivino et al., "Sjogren's Syndrome: An Update on Disease Pathogenesis, Clinical Manifestations and Treatment," *Clinical Immunology* 203 (June 1, 2019): 81–121, https://doi.org/10.1016/J.CLIM.2019.04.009.

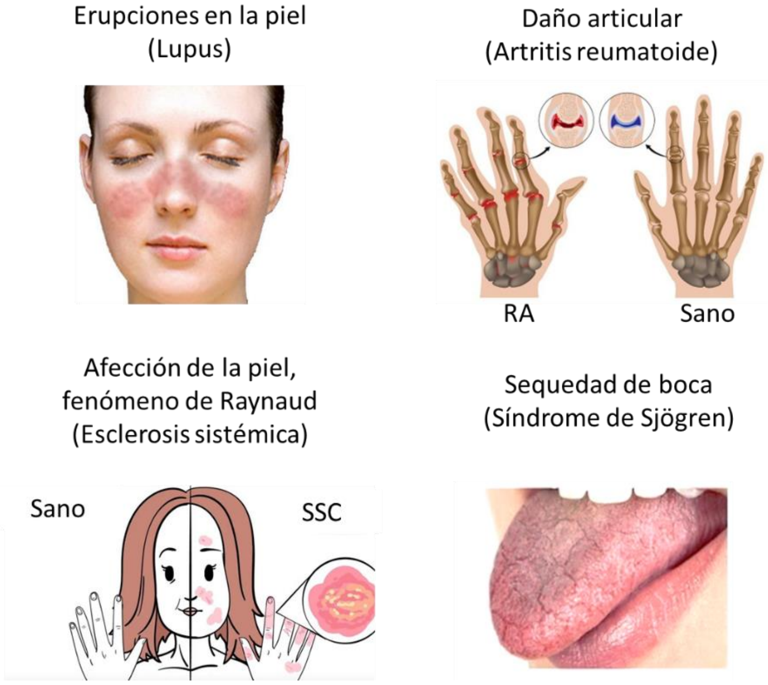


Figura 37. Signos más característicos de varias enfermedades autoinmunes sistémicas.

- El **síndrome antifosfolípido** se caracteriza por un aumento en el riesgo de trombosis recurrente, tanto arterial, venosa o microvascular, y/o complicaciones relacionadas con el embarazo (abortos espontáneos, muerte fetal o partos prematuros) en presencia de anticuerpos antifosfolípidos persistentes. Esta patología se clasifica en síndrome antifosfolípido primario (PAPS), que ocurre en ausencia de otras enfermedades relacionadas; y secundario, que se produce en combinación con otras enfermedades autoinmunes, principalmente lupus eritematoso sistémico[173].
- La **enfermedad mixta del tejido conectivo (MCTD)**, presenta como síntomas principales el fenómeno de Raynaud, artritis, hipertensión pulmonar y manos

[173] Jose A. Gómez-Puerta and Ricard Cervera, “Diagnosis and Classification of the Antiphospholipid Syndrome,” *Journal of Autoimmunity* 48–49 (February 1, 2014): 20–25, https://doi.org/10.1016/J.JAUT.2014.01.006.

edematosas e hinchadas, entre otros. Los pacientes de esta patología presentan síntomas superpuestos de varias enfermedades autoinmunes sistémicas como el lupus, artritis reumatoide o esclerosis sistémica, lo que conlleva dificultades en su diagnóstico. No obstante, a diferencia de los casos indiferenciados, se ha propuesto la presencia de una entidad clínica característica de MCTD, la cual está altamente asociada a la presencia de altos niveles de anticuerpos de ribonucleoproteína[174,175].

- La **enfermedad del tejido conectivo indiferenciado (UCTD)**, corresponde con aquellos pacientes que presentan síntomas característicos de varias enfermedades autoinmunes sistémicas (SLE, SjS, RA o SSC), pero que no reúnen las características necesarias para ser clasificados en una patología concreta. Alrededor de un tercio de los casos se acaban finalmente diagnosticando como una enfermedad reumática concreta conforme a su evolución en el tiempo[171].

La **Tabla 7** recoge información complementaria de estas enfermedades en relación a la prevalencia, el ratio entre géneros, la edad típica de aparición así como los síntomas y signos comunes de cada una de ellas. No obstante, estos datos son orientativos, dado que existen grandes diferencias epidemiológicas en función de la raza o la región geográfica estudiada, al ser los factores genéticos una de las causas principales del desarrollo de estas patologías[176].

[174] Robert W. Hoffman and Eric L. Greidinger, “Mixed Connective Tissue Disease,” 2002, 347–57, https://doi.org/10.1007/978-1-59259-239-5_23.

[175] P J W Venables, “Mixed Connective Tissue Disease.,” *Lupus* 15, no. 3 (January 2006): 132–37, http://www.ncbi.nlm.nih.gov/pubmed/16634365.

[176] | August ; Shapira, “Geoepidemiology of Autoimmune Rheumatic Diseases,” *Nat. Rev. Rheumatol* 6 (2010): 468–76, https://doi.org/10.1038/nrrheum.2010.86.

Tabla 7. Características epidemiológicas de las siete enfermedades autoinmunes sistémicas estudiadas.[165,176-181]

Patología	Prevalencia (/100000)	Ratio Mujeres/ Hombres	Edad de aparición	Principales signos y/o síntomas comunes
RA	200-700	3:1	30-50	Articulaciones
SLE	30-70	9:1	15-44	Afección de la piel y articulaciones
SSC	10-30	5:1	40-50	Daño vascular, fibrosis generalizada.
pSjS	30-100	9:1	40-60	Afección de las glándulas exocrinas
PAPS	40-50	5:1	30-40	Trombosis, abortos espontáneos
MCTD	1-9	10:1	15-35	Combinación de características de SLE, SSC, PM y/o RA
UCTD	No existen datos epidemiológicos rigurosos sobre la prevalencia de esta enfermedad debido principalmente a la dificultad para su definición. De manera orientativa, la incidencia anual de esta patología oscila entre 14 y 140 por 100000 personas de la población general[171]			

[177] Gilberto Cincinelli et al., "Why Women or Why Not Men? Sex and Autoimmune Diseases," *Indian Journal of Rheumatology* 13, no. 1 (2018): 44, https://doi.org/10.4103/injr.injr_1_18.

[178] Xavier Mariette and Lindsey A. Criswell, "Primary Sjögren's Syndrome," ed. Caren G. Solomon, *New England Journal of Medicine* 378, no. 10 (March 8, 2018): 931–39, https://doi.org/10.1056/NEJMcp1702514.

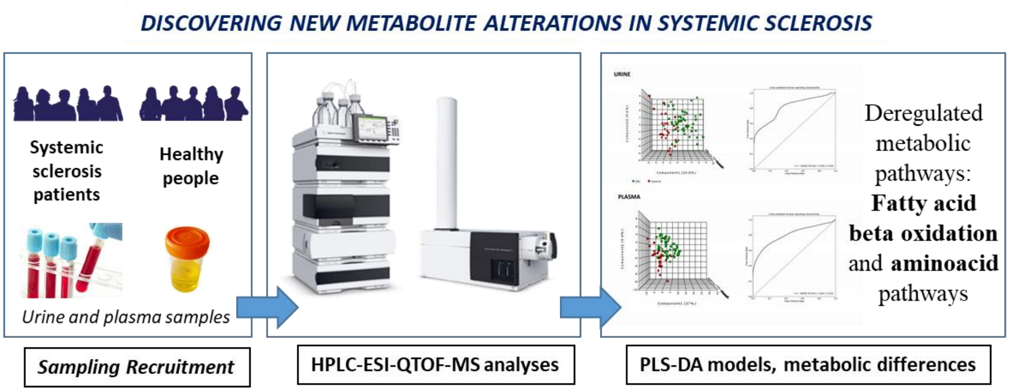
[179] Ricard Cervera, "Antiphospholipid Syndrome," *Thrombosis Research* 151 (March 2017): S43–47, https://doi.org/10.1016/S0049-3848(17)30066-X.

[180] "Orphanet," https://www.orpha.net, accessed August 27, 2019, https://www.orpha.net/.

[181] "NORD (National Organization for Rare Disorders)," accessed August 27, 2019, https://rarediseases.org/.

Capítulo 4

Alteraciones metabólicas en muestras de orina y plasma detectadas en pacientes con esclerosis sistémica mediante HPLC-ESI-QTOF-MS


DISCOVERING NEW METABOLITE ALTERATIONS IN SYSTEMIC SCLEROSIS
Systemic sclerosis patients
Healthy people
Urine and plasma samples
Sampling Recruitment
HPLC-ESI-QTOF-MS analyses
URINE
PLASMA
Deregulated metabolic pathways: Fatty acid beta oxidation and aminoacid pathways
PLS-DA models, metabolic differences

Journal of Pharmaceutical and Biomedical Analysis. 2019, 162, 82-90

Álvaro Fernández-Ochoa, Rosa Quirantes-Piné, Isabel Borrás-Linares, David Gemperline, PRECISESADS Clinical Consortium, Marta E. Alarcón Riquelme, Lorenzo Beretta, Antonio Segura-Carretero. DOI:10.1016/j.jpba.2018.09.021

Urinary and plasma metabolite differences detected by HPLC-ESI-QTOF-MS in systemic sclerosis patients

ABSTRACT

Systemic Sclerosis (SSC) is a chronic autoimmune disease whose origin and pathogenesis are not yet well known. Recent studies are allowing a better definition of the disease. However, few studies have been performed based on metabolomics. In this way, this study aims to find altered metabolites in SSC patients in order to improve their diagnosis, prognosis and treatment. For that, 59 SSC patients and 28 healthy volunteers participated in this study. Urine and plasma samples were analyzed by a fingerprinting metabolomic approach based on HPLC-ESI-QTOF-MS. We observed larger differences in urine than plasma metabolites. The main deregulated metabolic families in urine were acylcarnitines, acylglycines and metabolites derived from amino acids, specifically from proline, histidine and glutamine. These results indicate perturbations in fatty acid beta oxidation and amino acid pathways in scleroderma patients. On the other hand, the main plasma biomarker candidate was 2-arachidonoylglycerol, which is involved in the endocannabinoid system with potential implications in the induction and propagation of systemic sclerosis and autoimmunity.

Keywords. Metabolomics, HPLC-ESI-QTOF-MS, Systemic Sclerosis, biomarker, acylcarnitines, 2-arachidonoylglycerol.

1. INTRODUCTION

Systemic sclerosis (SSC) is a chronic autoimmune disease characterized by immune system activation, endothelial damage and widespread vasculopathy, fibrosis of the skin and of internal organs [1]. Despite recent progress, the pathogenesis of SSC remains elusive and so its treatment, with most patients experiencing long term disability, severe morbidity and increased mortality ratios compared to the general population or to other systemic autoimmune diseases [2].

A better understanding of biological pathways involved in SSC development is mandatory to tackle the processes that lead to disease progression and for the discovery of effective therapies. Ideally, this process should be carried on an individual basis and considering the broad spectrum of alterations that may happen in the organism. Recent technical developments together with the increasing availability of high-throughput methodologies have enabled the detailed description of multiple molecular alterations that coexist in sick individuals [3]. Large scale biology techniques are commonly referred to as "-omics" and include genomics, epigenomics, transcritptomics, proteomics and metabolomics. The study of –omics has gained much attention in many fields of medicine, including systemic autoimmune diseases [4] and SSC is no exception [5]. Nonetheless, among the different -omics, metabolomics has not yet extensively been studied in SSC and to our knowledge, just one study with a limited number of scleroderma samples has been published so far [6]. Here 19 SSC patients were used as a comparison group of 30 systemic lupus erythematosus (SLE) patients along with 20 primary Sjogren's syndrome (SjS) subjects.

Metabolomics can be considered the final step of the biological processes described by –omics techniques, and it concerns the study of the complete set of small molecules intermediates in a biofluid [7]. The description of metabolome provides a portrait of the metabolic state of individuals at a certain point in time and the analysis of the metabolic profile may give insight on the biochemical consequences of disease. The metabolic characterization of patients and the description of metabolite profiling in relation to clinical features and sub-setting may have relevant consequences in understanding disease pathogenesis, in discovering biomarkers and in suggesting individualized therapies [8].

This work aims at characterizing metabolic alterations associated with SSC. To this end, a metabolic fingerprinting strategy based on high performance LC coupled to electrospray ionization quadrupole time-of-flight mass spectrometry (HPLC-ESI-QTOF-MS) is used for the analysis of plasma and urine samples. Metabolic differences between different subtypes of the disease and in relation to major organ involvement are also explored to describe a set of potential metabolic biomarkers.

2. MATERIAL AND METHODS

2.1 Patients and controls

A total of 59 Italian SSC patients were included in the study. All the patients fulfilled the 2013 ACR/EULAR criteria [9] and were categorized into the limited (lcSSC, n=43) or the diffuse cutaneous (dcSSC, n=10) subsets. Patients with definite SSC without skin fibrosis yet with puffy fingers were categorized in the lcSSC subset; the remaining patients with definite disease without fibrosis were retained as a separate group (defSSC, n=6). Interstitial lung disease (ILD) was defined as in Vigone et al. [10], that is involvement of lung parenchyma > 5% on high resolution computed tomography accompanied by a reduced forced vital capacity (FVC) <80% of predicted values or by a reduced diffusing capacity for carbon monoxide (DLco) < 80% of predicted values.

Twenty-eight age- and sex-matched Italian healthy volunteers were included as control group.

Blood samples were collected into tubes with dipotassium ethylenediaminetetraacetic (K_2EDTA) acid and immediately centrifuged at 1500 g for 10 min at room temperature. Plasma were obtained from the supernatant and then the samples were frozen and stored at -80 ºC until sample processing. Random single urine-plot samples were collected and then centrifuged at 2500 g for 10 min at 4ºC. Urine samples were also frozen and stored at -80 ºC until sample processing. This metabolomic analysis is ancillary to the PRECISESADS project (www.precisesads.eu) that was approved by the local ethic committee (comitato etico Area B), and written consent was obtained from each participant.

2.2 Chemicals

All chemicals were of analytical reagent grade and used as received. Formic acid and LC-MS grade methanol for mobile phases were purchased from Fluka, Sigma-Aldrich (Steinheim, Germany) and Fisher Scientific (Madrid, Spain), respectively. Water was purified by a Milli-Q system from Millipore (Bedford, MA, USA). For plasma treatment, ethanol and methanol (Fisher Scientific Madrid, Spain) were used.

2.3 Plasma analysis

Plasma samples, which were stored at -80 °C, were thawed on ice. A plasma aliquot of 100 µl was mixed with 200 µl methanol:ethanol (50:50, v/v) in order to remove the protein content. To achieve an efficient protein precipitation, the mixture was kept at -20 °C during 30 min. Next, the sample was centrifuged during 10 min at 14800 r.p.m. and 4 °C, and the supernatant was evaporated to dryness under vacuum in a centrifugal evaporator (Concentrator Plus, Eppendorf, Hamburg, Germany) during 2 h. Afterwards, the dry residue was reconstituted in 100 µl of 0.1% aqueous formic acid:methanol (95:5, v/v) and centrifuged at the same conditions in order to remove solid particles. Finally, a 40 µl aliquot was transferred into HPLC vials and stored at -80 °C prior to analysis. A quality control sample (QC) was prepared by mixing equal volumes (20 µl) from each sample. This sample was treated as described above.

Analyses were performed using an Agilent 1260 HPLC instrument (Agilent Technologies, Palo Alto, CA, USA) coupled to an Agilent 6540 Ultra High Definition (UHD) Accurate Mass Q-TOF equipped with a Jet Stream dual ESI interface.

The compounds were separated using a reversed-phase C18 analytical column (Agilent Zorbax Eclipse Plus, 1.8 µm, 4.6×150 mm) protected by a guard cartridge of the same packing. The mobile phases used in the analysis were water containing 0.1% of formic acid (Mobile Phase A) and methanol (Mobile Phase B). The following gradient of these mobile phases was used in order to obtain an efficient separation: 0 min (A:B, 95/5), 5 min (A:B, 90/10), 15 min (A:B, 15/85), 32 – 40 min (A:B, 0/100), and 45 min (A:B, 95/5). Finally, initial conditions were kept for 5 min at the end of each analysis to equilibrate the analytical column before the next analysis. The column and autosampler compartment temperatures were set at 25 and 4 ºC, respectively, whereas the flow rate and the injection volume were 0.4 mL/min and 5 µl.

Detection was performed in positive-ion mode over a range from 50 to 1700 m/z. All spectra were corrected by means of continuous infusion of two reference masses: purine (m/z 121.050873) and hexakis (^{1}H, ^{1}H, ^{3}H-tetrafluoropropoxy) phosphazine or HP-921 (m/z 922.009798). Both reference ions provided accurate mass measurement typically better than 2 ppm.

Ultrahigh pure nitrogen was used as drying and nebulizer gas at temperatures of 200 and 350 ºC and flows of 10 and 12 L/min, respectively. Other optimized parameters were as follows: capillary voltage, +4000V; nebuliser, 20 psi; fragmentor, 130 V; nozzle voltage, 500 V; skimmer, 45 V and octopole 1 RF Vpp, 750 V.

The samples were analyzed following this sequence: 2 blanks, 5 QCs, 5 randomized samples, 1 blank, 2 QC, 5 randomized samples, etc. Finally, a MS/MS analysis of the QC sample was performed in order to facilitate the identification of potential biomarkers. This experiment was performed using nitrogen as the collision gas with the following collision energy values: 10 eV, 20 eV and 40 eV.

2.4 Urine analysis

Urine samples, which were stored at -80 °C until treatment, were thawed on ice. In order to correct the concentration variation between samples due to individual's hydration status, a pre-analysis normalization step was performed in these samples by means of osmolality measure [11]. The measurement of urine osmolality was determined by freezing point depression using an OSMOMAT 3000 osmometer

(Gonotec, Berlin, Germany). The samples were diluted with water in order to achieve a final osmolality value of 100 mOsm/Kg.

Afterwards, the samples were centrifuged during 10 min at 14800 r.p.m. and 4 °C in order to remove solid particles, and 40 µl of the supernatant was transferred into HPLC vials and stored at -80 °C prior to analysis. A urine QC sample was also prepared by mixing 20 µl from each sample.

Regarding the HPLC-ESI-QTOF-MS methodology, all conditions were the same as plasma samples with the exception of the mobile phases gradient and injection volume. In this case, the following gradient was performed: 0 min (A:B, 95/5), 30 min (A:B, 70/30), 40 min (A:B, 0/100), 50-60 min (A:B, 95/5); and the injection volume was 3 µl.

2.5 Data processing

Recursive Feature Extraction for small molecules was performed by means of MassHunter Profinder software (B.06.00, Agilent). This algorithm combines "Molecular Feature Extraction" with "Find by Ion" algorithms. Therefore, the first algorithm finds features which are defined as the combination of co-eluted species that are related by isotopic distribution, presence of adducts, loss of molecules and/or charge-state envelope. Secondly, the features found in the samples are aligned by mass and retention time. Finally, a list with the resulting features is created and used to find them in the same samples more accurately.

Peaks were filtered with intensity threshold at 1250 counts. $[M+H]^+$, $[M+Na]^+$ and $[M-H_2O]$ were the considered species with a maximum charge of 2. Feature alignment parameters were ± 0.25 minutes and 20 ppm ± 2 mDa for retention time and mass windows, respectively.

The integration method was Agile2 carrying out an average of spectra at peak start and end to subtract a background spectrum. Nevertheless, integration results were manually supervised to correct defaults.

After Molecular Feature Extraction and their manual supervison, Principal Component Analysis (PCA) was performed for plasma and urine samples separately in order to check the analytical reproducibility according to QC samples distribution and to identify any outliers. PCA analysis was performed with Mass Profiler Professional software (B.14.00, Agilent). A data normalization strategy was applied to the plasma data. This correction was performed for each sample by the MS Total Useful Signal (MSTUS) [12]. The normalization step was performed with Metaboanalyst 3.0 software.

Afterwards, features with high variability in QC samples (RSD>30%) were filtered. Missing values were replaced by a small value (half of the minimum positive value in the original data). The purpose of this filter was to discard metabolites whose concentration throughout the sequence was not reproducible. On the other hand, features with a percentage of missing values among all study samples (SSC and Healthy Controls) higher than 25 %, were removed. In this filter, we did not use the QC samples. The aim of this filter was to discard exogenous metabolites related to drugs and treatments.

2.6 Statistical analysis

Univariate (t-test and fold change [FC] analysis) and multivariate (supervised Partial Least Squares Discriminant Analysis (PLS-DA)) statistical tests were performed in order to find metabolic differences between cases and healthy controls. For univariate analysis a nominal alpha level equal to 0.05 and a FC threshold equal to 1.2 were chosen; univariate t-test was conducted on log-transformed data. P-values were validated with the False Discovery Rate Multiple Testing Correction (FDR). Fold change values were calculated for each metabolite by comparison between the means of the two groups in order to estimate the variation between them. PLS-DA models were validated by means of 10-fold cross-validation; to account for variability 1000 runs of cross-validation were performed. Validation results are expressed as the area under receiver operating characteristics curve (AUROC). The AUROC measures the overall discrimination of a classification algorithm, where 1 represents a perfect test and 0.5 a test doing no better than at random. A permutation-based step-down minP procedure was used to derive empirical p values of cross-validated AUROC [13]. Pareto scaling was used to transform the data before PLS-DA analysis to make variables comparable to each other. Univariate analysis and PLS-DA were performed with the scikit-learn algorithm (http://scikit-learn.org/stable/index.html) [14] along with custom python codes implemented by LB. PLS-DA scores were visualized via the Metaboanalyst 3.0 software [15]. All the statistical tests were also applied to find metabolic differences between SSC subtypes and between patients with or without ILD.

2.7 Metabolite annotation

The software Agilent Mass Hunter finds features and assigns a molecular formula for that specie with a specific error related with the uncertainty in the assignment. The metabolite annotation of the significant features was carried out using MS and MS/MS data (accurate masses, isotopic distributions and fragmentation patterns) in comparison with different metabolomic databases (LipidMaps (http://lipidmaps.org), Human Metabolome Database (http://hmdb.ca) and METLIN (http://metlin.scripps.edu)) as well as MS/MS fragmentation resources such as MetFrag (http://msbi.ipb-halle.de/MetFrag/).

3. RESULTS

3.1 Clinical characteristics

Demographic and clinical characteristics of healthy controls and SSC patients are reported in **Table 1**. Briefly, the majority of our patients were lcSSC (72.8%) with a long disease duration; disease duration was similar for lcSSC (12.2 ± 9.75 years) and dcSSC patients (13.5 ± 10.41 years); these data indicate that our cohort was enriched with late dcSSC subjects. Gastrointestinal involvement was the most prevalent complication of SSC, followed by vasculopathy (digital ulcers) and by ILD that was ascertained in 30.5% of patients according to our criteria.

3.2 Data quality assessment

After Molecular Feature Extraction and their manual supervision, a total of 432 and 625 features were obtained for plasma and urine samples, respectively. **Figure 1** shows the scores plots of the Principal Component Analysis. A drift in the plasma analytical sequence could be observed (**Figure 1a**) because QC samples were accommodated in PCA scores according to their injection order. This problem occurs frequently in LC-MS metabolomic studies due to the long duration of the analysis sequence producing decrease in the detected signal by the loss of the ionization efficiency [16]. However, the injection order effect was not observed in urine samples (**Figure 1b**) where the corresponding QC samples are grouped in the PCA scores. Therefore, a normalization strategy was applied only to the plasma data. The result of this normalization process is shown in **Figure 1c** where it can be observed that QC samples are well grouped.

After filtering steps, metabolites from drugs largely consumed by SSC patients, such as omeprazole or nifedipine, were discarded in this step. As a result, a total of 606 and 421 features for urine and plasma, respectively, were selected for statistical analysis.

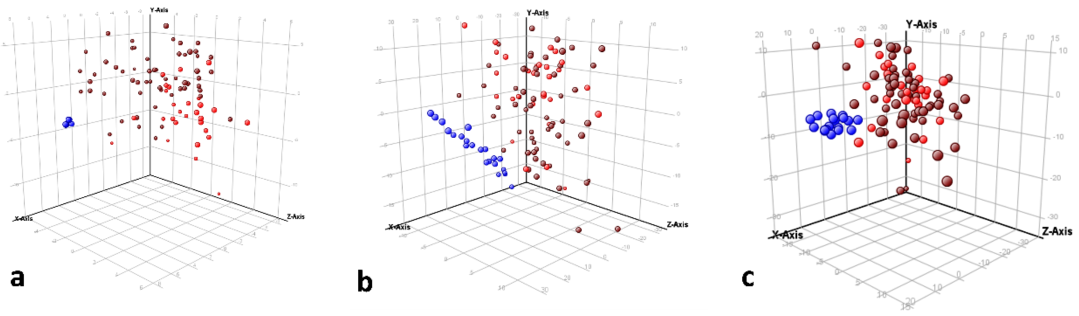


Figure 1. PCA scores plots from raw data (**1a**: Urine, **1b**: Plasma) and normalized data for plasma samples (**1c**). (QC, blue plots; SSC, brown plots; HCs, red plots)

Table 1. Clinical characteristics of study groups. (SSC, systemic sclerosis; lcSSC, limited cutaneous SSC; dcSSC, diffuse cutaneous SSC; ANA, antinuclear antibody; ACA, anticentromere antibody; Topo I, anti-topoisomerase I antibody; FVC, forced vital capacity; DLco, diffusing capacity for carbon monoxyde; ILD, interstitial lung disease; DU, digital ulcers; GERD, gastroesophageal reflux disease; MMF, mycophenolate mophetil; AZA, azathioprin; CYC, cyclophosphamide; TNF, tumor necrosis factor alpha.).

Variable		HC (n = 28)	SSC (n = 59)
Age		49.3 ± 12.9	56.5 ± 12.7
Females, n (%)		22 (78.6%)	52 (88.1%)
Disease duration, yrs		-	12.5 ± 10.4
Subsets, n (%)	definite SSC		6 (10.1%)
	lcSSC		43 (72.8%)
	dcSSC		10 (17.1%)
Autoantibody, n (%)	ANA		56 (94.9%)
	ACA		23 (39%)
	Topo I		23 (39%)
FVC, % predicted		-	100.1 ± 21.1
DLco, % predicted		-	73.1 ± 21.8
ILD, n (%)		-	18 (30.5%)
History of DU, n (%)		-	28 (47.4%)
History of arthritis, n (%)		-	9 (15.3%)
GERD, n (%)		-	36 (61%)
Intestinal involvement, n (%)		-	33 (55.9%)
Immunosuppressants, n (%)			12 (20.3%)
	MMF		3 (5.1%)
	AZA		7 (11.8%)
	CYC		2 (3.4%)
	Biologicals, n (%)		9 (15.2%)
	Abatacept		3 (5.1%)
	Tocilizumab		5 (8.5%)
	Anti-TNF		1 (1.7%)

3.3 Altered metabolites in urine

Preliminary results based on PCA show a slight separation between SSC patients and control samples (**Figure 1a**).Our LC-MS data contain signals from several hundreds of metabolites. The number of candidate variable is typically relatively large compared with the sample size. Because many of the candidate are irrelevant to the study, variable selection based on univariate statistical test was a crucial step for metabolomics data analysis and modelling.

Overall 136 urine metabolites were selected via univariate analysis; these metabolites were used to build multivariate PLS-DA models. When PCA was applied to PLS-DA models two main components could be sorted out, (PC1, 20.9 %, PC2, 9.6 %). The scores plot of the model is represented in **Figure 2a**. Extensive cross-validation of the univariate selection procedure and multivariate PLS-DA modelling yielded an AUROC value of 0.807 ± 0.031 (mean ± standard deviation) with a permutation p-value less than 0.001. (**Figure 2b**).

Of the 136 peak metabolites choosen via univariate selection, a reduced set of 45 potential biomarkers metabolites with a variable importance in projection (VIP) score higher than 1.0 were selected for identification. **Table 2** shows the annotated metabolites along with their masses, retention times and molecular formulas, univariate t-test p-values, scores, errors, FC, VIP values, individual AUROC values and the characteristic MS/MS fragments. Some of the compounds were tentatively identified by their molecular formula obtained from their exact masses and isotopic distributions.

Unknown compounds that could not be tentatively annotated are reported in **Table 1S**. Of all the metabolites characterized, N1-methyl-4-pyridine-3-carboxamide, N1-methyl-2-pyridine-5-carboxamide, D-sorbitol and dimethylheptonoylcarnitine were the metabolites showing the highest probability of differentiation between cases and controls. **Figure 3a** represents the ROC curves of these metabolites.

Table 2. Molecular and statistical details of annotated urinary metabolites that presented significant differences between SSC patients and healthy volunteers. (FC<0, metabolites overexpressed in SSC).

RT (min)	Mass (Da)	p-value	FC	VIP-value	AUC	Molecular Formula	Score	Error (ppm)	Metabolite	MS/MS fragments
0.9	182.0788	9.02 E-4	-5.48	1.813	0.800	$C_6H_{16}O_6$	92.4	1.58	D-Sorbitol	69.0441/83.0598/183.0886
0.9	136.0617	8.37 E-3	1.52	1.148	0.700	$C_7H_8N_2O$	99.0	1.13	N-Methylnicotinamide	92.0456/94.0567/120.0550/137.0718
1.0	143.0946	7.89 E-3	2.42	1.607	0.728	$C_7H_{13}NO_2$	97.8	-2.80	Proline Betaine	42.0335/58.0652/84.0810
1.0	113.0588	7.89 E-3	1.44	1.007	0.728	$C_4H_7N_3O$	97.6	-4.09	Creatinine	43.0288/44.0495/86.0715
1.4	127.0997	0.016	-4.91	1.406	0.716	$C_7H_{13}NO$	90.1	2.06	N-cyclohexylformamide	Annotated by Formula
2.5	299.1477	0.026	-4.73	1.346	0.677	$C_{13}H_{21}N_3O_5$	98.9	3.20	Ser Pro Pro	Annotated by Formula
2.9	143.0585	0.025	2.50	1.055	0.628	$C_6H_9NO_3$	99.8	0.09	Vinylacetylglycine	41.0374/58.0283/69.0336/98.0607/144.0699
3.1	152.0583	3.75 E-4	1.56	1.227	0.812	$C_7H_8N_2O_2$	95.8	-0.44	N1-methyl-4pyridine-3-carboxamide	84.0440/108.0442/136.0389/153.0654
3.5	152.0583	1.19 E-3	1.68	1.367	0.770	$C_7H_8N_2O_2$	95.9	-2.55	N1-methyl-2-pyridine-5-carboxamide	42.0333/53.0385/78.0333/108.0441/110.0603/153.0602
3.7	325.0789	0.022	-1.75	1.095	0.661	$C_{14}H_{15}NO_8$	95.3	2.60	Dihydroxy-1H-indole glucuronide	150.055/326.085
6.6	230.1263	0.079	1.56	1.097	0.739	$C_{10}H_{18}N_2O_4$	93.5		Hydroxyprolyl-Valine	Annotated by Formula
10.2	246.1206	0.030	2.08	1.003	0.653	$C_{10}H_{18}N_2O_5$	96.4	4.68	L-beta-aspartyl-L-Leucine	74.0246/86.0966/132.1013/201.1236/247.1296
11.6	246.1372	0.011	1.99	1.745	0.678	$C_{14}H_{18}N_2O_2$	92.5	-1.62	Hypaphorine	Annotated by Formula
14.8	129.0429	0.0079	-1.52	1.038	0.695	$C_5H_7NO_3$	98.8	-1.42	Pyroglutamic acid	45.0337/58.0367/84.0447/130.0501
14.8	264.1150	0.014	-1.46	1.001	0.677	$C_{13}H_{16}N_2O_4$	92.4	-4.41	Alpha-N-Phenylacetyl – L glutamine	91.0536/101.0715/129.0662/130.0495/147.077
20.7	265.0951	0.0079	-1.62	1.326	0.710	$C_{13}H_{15}NO_5$	78.4	4.04	2-(2-Phenylacetoxy)propinylglycine	57.0339/119.0489/266.1007
34.3	285.1935	0.038	1.78	1.426	0.656	$C_{15}H_{27}NO_4$	96.8	-1.95	2-octenoyl-carnitine	55.0539/85.0284
33.8	285.1943	0.038	1.45	1.068	0.646		95.1	-0.63		
36.6	309.1921	0.027	1.41	1.036	0.669	$C_{17}H_{27}NO_4$	99.8	0.55	Decatrienoylcarnitine	85.0288/251.1297/310.2026
37.1	309.1922	0.038	1.44	1.021	0.623		86.8	5.00		
36.8	299.2078	6.29 E-3	1.70	1.292	0.739	$C_{16}H_{29}NO_4$	96.6	5.00	2-Nonenoylcarnitine	85.0218/300.2181
37.8	301.2235	9.02 E-4	1.91	1.468	0.776	$C_{16}H_{31}NO_4$	94.6	0.90	2,6-Dimethylheptanoyl carnitine	60.0804/85.0282/302.228
38.2	313.2245	0.010	1.62	1.061	0.701	$C_{17}H_{31}NO_4$	96.4	1.53	9-Decenoylcarnitine	60.0804/85.0282/157.0501/255.159
38.3	313.2218	7.89 E-3	1.73	1.160	0.701		96.6	1.27		
38.6	357.2498	0.010	1.87	1.173	0.673	$C_{19}H_{35}NO_5$	96.6	-0.67	Hydroxydodecenoylcarnitine	60.081/81.0698/85.0282/95.0856/137.1333/155.1437
39.0	327.2389	7.89 E-3	2.44	1.431	0.706	$C_{18}H_{33}NO_4$	96.1	-1.86	Undecenoyl carnitine	85.0283/328.2484

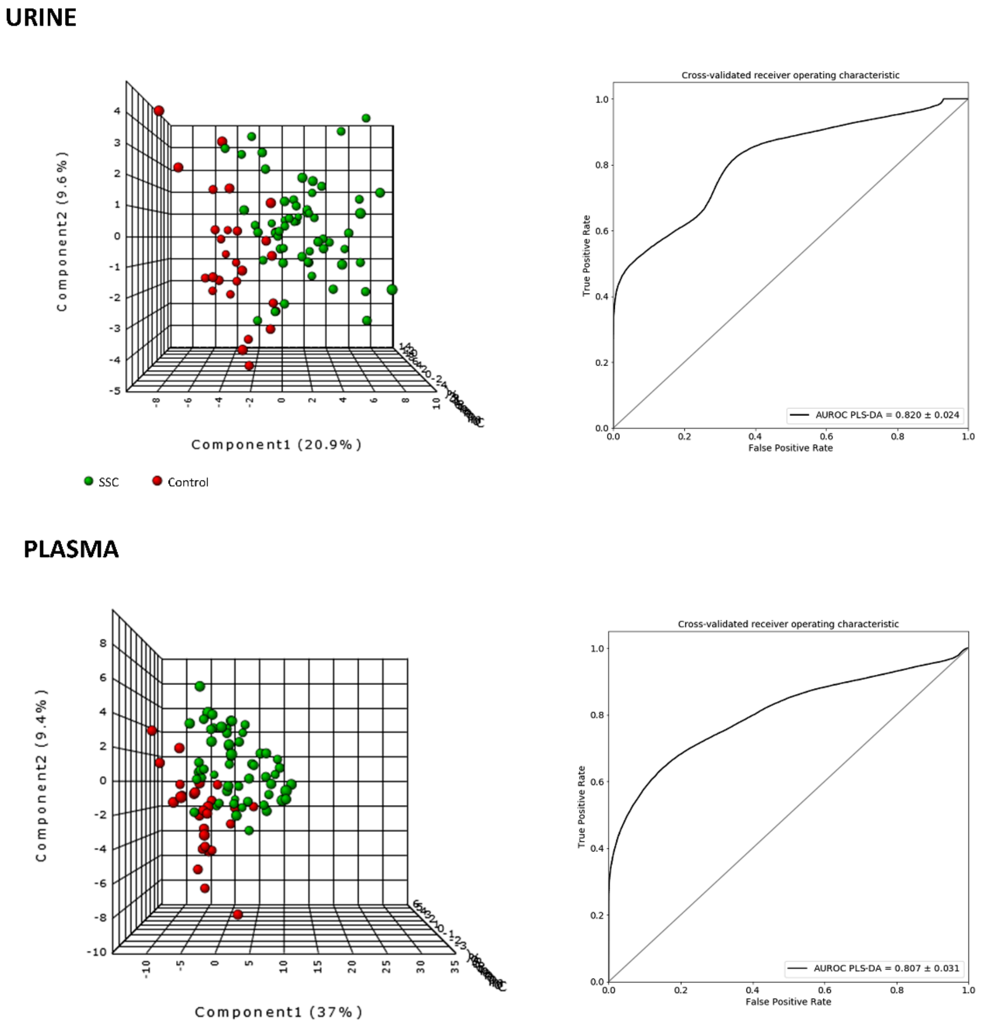


Figure 2. A supervised Partial Least Squares Discriminant Analysis (PLS-DA) from urine and plasma samples. **Urine**: PLS-DA scores plot (**2a),** ROC curve for PLS-DA model validation (**2b**) **Plasma**: PLS-DA scores plot (**2c),** ROC curve for PLS-DA model validation (**2d**). (SSC, green plots; HCs, red plots)

3.4 Altered metabolites in plasma

Plasma samples were analysed analogously to urine samples. As the result of univariate filtering, a total of 46 features were selected. Using these features, a PLS-DA model was created with 2 principal components (PC1, 37.0 %, PC2, 9.4 %). This feature selection procedure was also validated with an AUROC value of 0.820 ± 0.024. **Figure 2** shows the scores plot (**Figure 2c**) and the ROC curve (**Figure 2d**). The permutation p-value was lower than 0.001. A total of 12 features were chosen (VIP values > 1.0) and tentatively identified with the methodology described above.

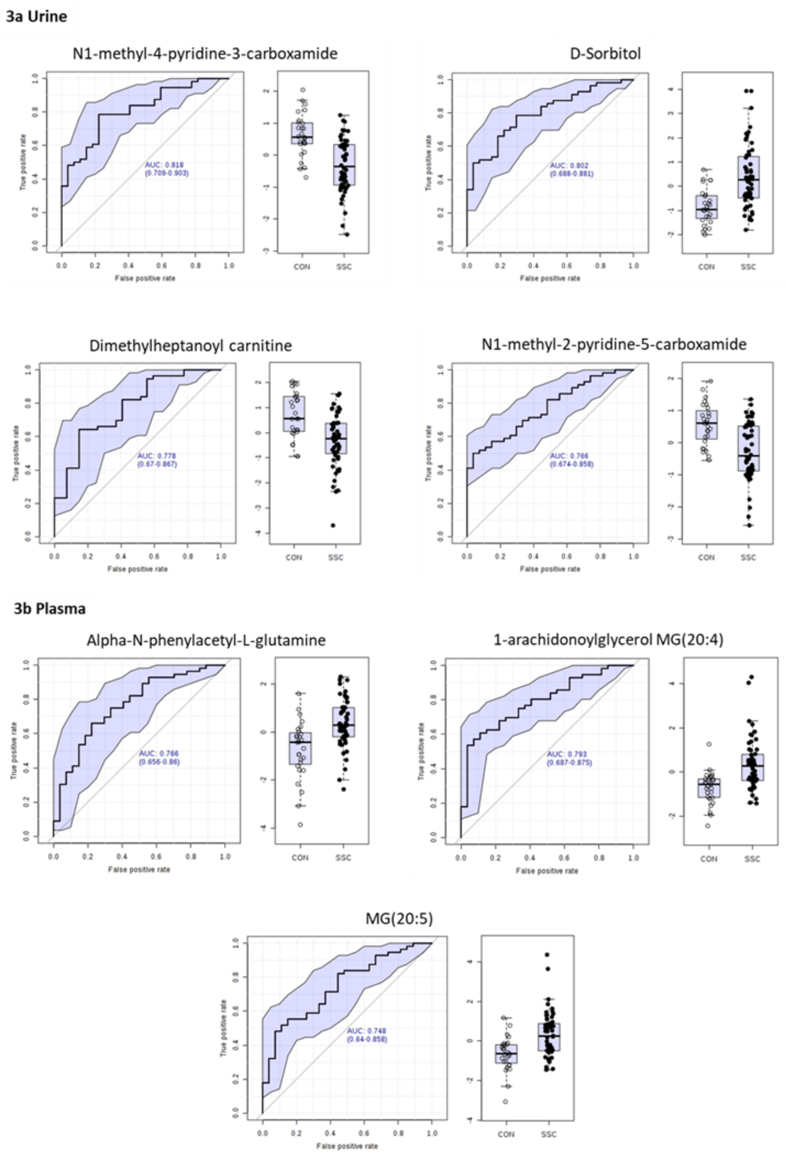


Figure 3. ROC curves of metabolites that present AUROC values higher than 0.74. Urine (**3a**): N1-methyl-4pyridine-3-carboxamide, D-sorbitol, dimethylheptanoyl carnitine and N1-methyl-2-pyridine-5-carboxamide. Plasma (**3b**): MG(20:4), alpha-N-phenylacetyl-L-glutamine and MG(20:5). The box plots show the median, the quartiles and the whole range of concentration of these metabolites measured in the samples.

Table 3. Molecular and statistical details of annotated plasma metabolites that presented significant differences between SSC patients and healthy volunteers. (FC<0, metabolites overexpressed in SSC).

	RT (min)	Mass (Da)	p-value	FC	VIP-value	AUC	Molecular Formula	Score	Error (ppm)	Metabolite	MS/MS fragments
Plasma	5.6	231.1482	0.035	-1.58	1.098	0.689	$C_{11}H_{21}NO_4$	93.8	3.42	Butyrilcarnitine	85.0288/113.9698/173.0816/232.1544
	10.4	245.1633	0.026	-1.35	1.005	0.624	$C_{12}H_{23}NO_4$	91.9	-.0.77	Valerylcarnitine	60.0814/85.0285/144.1017/187.0949/246.1695
	11.3	264.1132	0.0047	-2.23	1.745	0.765	$C_{13}H_{16}N_2O_4$	92.7	-4.92	Alpha-N-phenylacetyl-L-glutamine	84.0436/91.0533/129.0654/130.0490/136.0752/147.0757
	12.4	231.9800	0.037	-1.82	1.168	0.684	$C_6H_4N_2O_7S$	96.8	-0.54	2-4-dinitrobenzenesulfonic acid	Annotated by Formula
	22.9	376.2628	0.022	-2.95	1.559	0.747	$C_{23}H_{36}O_4$	84.6	4.98	MG (20:5)	57.0685/93.0679/377.2627
	24.3	378.2720	0.021	-3.38	1.668	0.792	$C_{23}H_{38}O_4$	90.3	4.92	1-arachidonoylglycerol MG(20:4)	67.0533/81.0685/95.0839/305.2095/379.2770
	26.0	282.2539	0.037	-1.57	1.068	0.659	$C_{18}H_{34}O_2$	86.8	-5.86	Oleic acid	55.0524/69.0683/67.0526/83.0834/265.2472/283.2586

Table 3 shows the significant annotated metabolites from plasma samples; whereas unknown metabolites are reported in **Table 1S**.

Univariate AUROC values were lower than those obtained in urine samples. Nevertheless, N-phenylacetyl-L-glutamine, 1 arachidonoylglycerol MG(20:4) and MG(20:5) obtained AUROC values close to 0.80 (**Figure 3b**).

3.5 Specific metabolites in relation to disease subsets and lung involvement

Metabolic characteristics of dcSSC and lcSSC subjects were investigated both in plasma and urine as described above. The only remarkable differences were found after univariate analysis in urine samples, while no difference could be found from plasma metabolites. The significant urinary metabolites found in the dcSSC vs lcSSC analysis were 3-methylglutarylcarnitine, 5-hydroxyindoleacetic acid, indospicine, L-arogenate and N(5-amino-2hydroxybenzoyl)glycine (**Table 2S**). No multivariate model could discriminate between the two major SSC subsets.

Finally, metabolites associated with lung involvement were explored. Urine PLS-DA models showed that scleroderma patients with lung involvement are better classified with respect to healthy controls (AUROC=0.922 ± 0.038, **Figure S1**) than SSC subjects without lung involvement (AUROC = 0.795 ± 0.034, **Figure S2**). Results from univariate analyses are detailed in **Tables 3S-4S**.

When comparing SSC-ILD and SSC without ILD, univariate analysis found several significant differences, specifically valyl-valine, kynurenic acid, L-proline, proline-histidine, quinolinic acid and β-D-glucopyrapyranosil anthranilate in urine and three fructosamines (1-amino-1-deoxy-D-fructose) derived from leucine, isoleucine and valine in plasma. (**Table 5S**). Multivariate models showed no good classification between SSC-ILD and SSC without ILD showing an AUROC value of 0.639 ± 0.062.

4. DISCUSSION

To our knowledge, our work represents the largest and most comprehensive attempt to analyse metabolite intermediates in SSC so far. Urine and plasma samples were analysed to sort out differences against age- and sex-matched healthy controls or between SSC subsets and our findings point out to a number of alterations that may reflect several pathophysiological processes of SSC. Overall we were capable of confirming and expanding previous preliminary observations made in a smaller case-series of SSC samples [6], of providing consistent novel results about specific metabolic alterations as well as describing exploratory findings that may warrant further investigations. These results have thoroughly been described through the paper and its related supplemental materials, however, for sake of brevity, hereafter we will solely comment on the most consistent and interesting findings.

Firstly, several metabolites derived from amino acids were found altered in urine from SSC patients compared to healthy controls. These metabolites include proline-betaine, hydroxyprolyl-valine and a tripeptide formed by serine and two units of proline (Ser-Pro-Pro). These results suggest that SSC patients present deregulation in amino acid pathways, specifically proline metabolism. Alterations in proline metabolism can be related to the perturbed collagen turnover that characterizes SSC and indeed collagen is a microenviromental reservoir of proline [17,18]. Of interest, urinary L-proline and proline-histidine were also increased in patients with lung fibrosis as compared to SSC subjects without interstitial lung involvement, further strengthening the role of L-proline derivates as a biomarker of fibrosis. Nonetheless, we were not capable of finding any difference between lcSSC and dcSSC subjects as far as L-proline or derivates are concerned. This negative result could be due to the lack of power in the lcSSC/dcSSC comparison or to the long disease duration in our dcSSC samples (mean = 13.5 years). Collagen turnover and histopathology may indeed change in late skin lesion of dcSSC patients [19] and it would be of interest in further studies to analyse and focus on the metabolic differences between late/early dcSSC subjects.

Amino acid metabolism appears to be disturbed in SSC patients also when other metabolites are concerned. This is the case for instance, of alpha-N-phenylacetyl-L-glutamine that was found to be consistently up-regulated both in plasma and urine samples from SSC patients. This metabolite is formed by the conjugation of glutamine and phenylacetate and the latter may accumulate in SSC as a consequence of intestinal dysbiosis. Previous studies have found in SSC a reduced number of commensal bacteria [20], which possess the capability to catabolize phenylacetate [21]. Dysbiosis and the reduction of this bacteria would thus promote an excess of substrate favouring phenylacetylglutamine formation.

The finding that amino acid metabolism is altered in SSC is not completely new, although it has not been described before. Only in [6] pyroglutamic acid was preliminarily found to be altered in SSC patients. Our results support and confirm these findings, although their significance is largely speculative. A defective glutathione metabolism would lead to the accumulation of pyroglutamic acid and urine excretion. Indeed, previous studies have observed an association between the null allele of the glutathione S-transferase that result in a loss of enzymatic activity and cardiovascular complications of SSC [22].

N1-methyl-4-pyridine-3-carboxamide and its isomer N1-methyl-2-pyridine-5-carboxamide could be related to renal failure of the patients. These metabolites have been described as end products of nicotinamide-adenine-dinucleotide (NAD) degradation being uremic toxins. They have also been found altered in patients with chronic renal failure [23]. Both metabolites are involved in nicotinate and nicotinamide metabolism and they come from N-methylnicotinamide by action of aldehyde oxidase. This metabolite was also found altered in urine samples.

In those urine samples we also found an overall deregulation of acylcarnitines, including dimethylheptanoylcarnitine, undecenoylcarnitine, 9-hydroxydecenoylcarnitine, 9-decenoylcarnitine, nonenoylcarnitine and 2-octenoylcarnitine, while butyrylcarnitine was found in plasma. Acylcarnitines are involved in the metabolism of fatty acids, being responsible for the transport of fatty acids in the mitochondria, as well as amino acid oxidation. The alterations in acylcarnitines together with the oleic acid up-regulation in SSC plasma samples indicate that fatty acid β-oxidation metabolism is somehow affected in SSC patients. This result agrees with those reported in sera from SLE, SjS and SSC [6]. Acylcarnitines, as other lipid intermediates, may have effect on the membrane function that harbours the insulin receptor and their accumulation has been associated with insulin resistance [24]. The overall reduction of acylcarnitines we found in SSC patients would thus reflect and could help explain the reduced risk of diabetes observed in SSC patients [25]. Another family of metabolites also involved in fatty acid beta-oxidation metabolism are acylglycines. Two metabolites of this family (vinylacetylglycine and 2-2-phenylacetoxypropinylglycine), which are minor metabolites of FAs, have been found altered in urine samples.

Another interesting finding in the present study was the detection of D-sorbitol as potential biomarker in urine. This metabolite is involved in the polyol pathway also called sorbitol-aldose reductase pathway which transforms glucose to fructose. The high concentration of this compound in urine could be related to its malabsorption due to the gastrointestinal dysfunction which is a frequent affection in SSC patients [26] rather than a possible failure in glucose metabolism. In addition, D-sorbitol is associated to fructose because they compete with the same transporter GLUT5. Previous studies have shown fructose malabsorption in systemic sclerosis [27]. Other metabolites suggestive of alterations in fructose metabolism were the three fructosamines (1-amino-1-deoxy-D-fructose) found significantly different between patients with and without lung fibrosis. Fructosamines, also called amadori products, are metabolites derived from the early stage of the maillard reaction. These metabolites are directly related to the advanced glycation end products (AGE), which have been previously studied showing differences in pulmonary fibrosis [28].

Finally, two monoacylglycerol compounds (MG(20:4) and MG(20:5)), were significantly up regulated in plasma samples of SSC patients. 2-arachidonoylglycerol (MG(20:4)) has been described as the second endogenous cannabinoid ligand which could have implications for SSC pathogenesis [29]. The endocannabinoid system exerts a role in fibroblast activation, vasculature activation and the immune response. A previous study has shown the role of the cannabinoid pathways in the induction and propagation of systemic sclerosis and autoimmunity using a mouse model [30].

CONCLUSIONS

According to our results, the main deregulated compounds were alpha-N-phenylacetyl-L-glutamine, acylcarnitines, acylglycines, monoacylglycerols and metabolites derived from aminoacids. Thus, the metabolic mechanisms affected in scleroderma patients are fatty acid beta oxidation and aminoacid, mainly proline, histidine and glutamine, pathways. In this way, SSC patients show several alterations in metabolic pathways, mainly, but not exclusively, involving amino acids, most likely as a result of deranged collagen metabolism and turnover. It has also been confirmed the deregulations in endocannabinoid system by the alteration of 2-arachidonoylglycerol in plasma. Nevertheless, further studies should be carried out to confirm these results with a higher number of samples.

Conflict of interest.

All authors declare that they have no conflict of interest.

Acknowledgements.

The work described has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement No. 115565, the resources for which are composed of a financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and the European Federation of Pharmaceutical Industries and Associations (EFPIA) companies' in-kind contribution. The author IBL gratefully acknowledges the Spanish Ministry of Economy and Competitiveness

(MINECO) in association with the European Social Fund (FSE) for the contract PTQ-13-06429, the author AFO received support from the Spanish Ministry of Education, Culture and Sports (FPU grant 14/03992).

Author contributions

A.F did the sample treatments, realized their analysis, performed the data treatment and wrote the manuscript in collaboration with L.B and R.Q. I.B and R.Q contributed with the analysis and results discussion. PRECISESADS Clinical Consortium (**Table 6S**) included and collected clinical data from patients in this study. D.G participated in data processing and metabolite identification. L.B contributed with statistical analysis and results discussion. A.S and M.E.A designed and supervised the study.

Bibliography.

[1] A.C. Desbois, P. Cacoub, Systemic sclerosis: An update in 2016, Autoimmun. Rev. 15 (2016) 417–426. doi:10.1016/j.autrev.2016.01.007.

[2] M. Nikpour, M. Baron, Mortality in systemic sclerosis, Curr. Opin. Rheumatol. 26 (2014) 131–137. doi:10.1097/BOR.0000000000000027.

[3] J. Varga, M. Trojanowska, M. Kuwana, Pathogenesis of systemic sclerosis: recent insights of molecular and cellular mechanisms and therapeutic opportunities, J. Scleroderma Relat. Disord. 2 (2017) 137–152. doi:10.5301/jsrd.5000249.

[4] J. Kang, L. Zhu, J. Lu, X. Zhang, Application of metabolomics in autoimmune diseases: Insight into biomarkers and pathology, J. Neuroimmunol. 279 (2015) 25–32. doi:10.1016/j.jneuroim.2015.01.001.

[5] M. Manetti, Emerging biomarkers in systemic sclerosis, Curr. Opin. Rheumatol. 28 (2016). doi:10.1097/BOR.0000000000000324.

[6] A.A. Bengtsson, J. Trygg, D.M. Wuttge, G. Sturfelt, E. Theander, M. Donten, T.

Moritz, C.J. Sennbro, F. Torell, C. Lood, I. Surowiec, S. R??nnar, T. Lundstedt, Metabolic profiling of systemic lupus erythematosus and comparison with primary Sjögren's syndrome and systemic sclerosis, PLoS One. 11 (2016). doi:10.1371/journal.pone.0159384.

[7] A. Agin, D. Heintz, E. Ruhland, J.M. Chao de la Barca, J. Zumsteg, V. Moal, A.S. Gauchez, I.J. Namer, Metabolomics - an overview. From basic principles to potential biomarkers (part 1), Med. Nucl. 40 (2016) 4–10. doi:10.1016/j.mednuc.2015.12.006.

[8] G.A.N. Gowda, S. Zhang, H. Gu, V. Asiago, N. Shanaiah, D. Raftery, Metabolomics-based methods for early disease diagnostics., Expert Rev. Mol. Diagn. 8 (2008) 617–33. doi:10.1586/14737159.8.5.617.

[9] F. Van Den Hoogen, D. Khanna, J. Fransen, S.R. Johnson, M. Baron, A. Tyndall, M. Matucci-Cerinic, R.P. Naden, T.A. Medsger Jr., P.E. Carreira, G. Riemekasten, P.J. Clements, C.P. Denton, O. Distler, Y. Allanore, D.E. Furst, A. Gabrielli, M.D. Mayes, J.M. Van Laar, J.R. Seibold, L. Czirjak, V.D. Steen, M. Inanc, O. Kowal-Bielecka, U. Müller-Ladner, G. Valentini, D.J. Veale, M.C. Vonk, U.A. Walker, L. Chung, D.H. Collier, M.E. Csuka, B.J. Fessler, S. Guiducci, A. Herrick, V.M. Hsu, S. Jimenez, B. Kahaleh, P.A. Merkel, S. Sierakowski, R.M. Silver, R.W. Simms, J. Varga, J.E. Pope, 2013 classification criteria for systemic sclerosis: An american college of rheumatology/European league against rheumatism collaborative initiative, Arthritis Rheum. 65 (2013). doi:10.1002/art.38098.

[10] B. Vigone, A. Santaniello, M. Marchini, G. Montanelli, M. Caronni, A. Severino, L. Beretta, Role of class II human leucocyte antigens in the progression from early to definite systemic sclerosis, Rheumatol. (United Kingdom). 54 (2014) 707–711. doi:10.1093/rheumatology/keu381.

[11] A.J. Chetwynd, A. Abdul-Sada, S.G. Holt, E.M. Hill, Use of a pre-analysis osmolality normalisation method to correct for variable urine concentrations and for improved metabolomic analyses, J. Chromatogr. A. 1431 (2016) 103–110. doi:10.1016/j.chroma.2015.12.056.

[12] A.M. De Livera, M. Olshansky, T.P. Speed, Chapter 20 Statistical Analysis of Metabolomics Data, Methods Mol. Biol. 1055 (n.d.). doi:10.1007/978-1-62703-577-4_20.

[13] P.H. Westfall, S.S. Young, Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment, Technometrics. 35 (1993) 450. doi:10.2307/1270279.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res. 12 (2012) 2825–2830. doi:10.1007/s13398-014-0173-7.2.

[15] J. Xia, D.S. Wishart, Using MetaboAnalyst 3.0 for Comprehensive Metabolomics Data Analysis., Curr. Protoc. Bioinformatics. 55 (2016) 14.10.1-14.10.91. doi:10.1002/cpbi.11.

[16] H. Mizuno, K. Ueda, Y. Kobayashi, N. Tsuyama, K. Todoroki, J.Z. Min, T. Toyo'oka, The great importance of normalization of LC-MS data for highly-accurate non-targeted metabolomics, Biomed. Chromatogr. 31 (2017) 1–7. doi:10.1002/bmc.3864.

[17] A. Barbul, Proline precursors to sustain Mammalian collagen synthesis., J. Nutr. 138 (2008) 2021S–2024S.

[18] M. Phang, James, W. Liu, C.N. Hancock, J.W. Fischer, Proline metabolism and cancer: emerging links to glutamine and collagen, Curr Opin Clin Nutr Metab Care. 18 (2015) 71–77. doi:10.2741/4022.

[19] T. Krieg, K. Takehara, Skin disease: a cardinal feature of systemic sclerosis, Rheumatology. 48 (2006) iii14-iii18. doi:10.1093/rheumatology/kep108.

[20] E.R. Volkmann, A.-M. Hoffmann-Vold, Y.-L. Chang, J.P. Jacobs, K. Tillisch, E.A. Mayer, P.J. Clements, J.R. Hov, M. Kummen, Ø. Midtvedt, V. Lagishetty, L. Chang, J.S. Labus, Ø. Molberg, J. Braun, Systemic sclerosis is associated with

specific alterations in gastrointestinal microbiota in two independent cohorts, BMJ Open Gastroenterol. 4 (2017) e000134. doi:10.1136/bmjgast-2017-000134.

[21] R. Teufel, V. Mascaraque, W. Ismail, M. Voss, J. Perera, W. Eisenreich, W. Haehnel, G. Fuchs, Bacterial phenylalanine and phenylacetate catabolic pathway revealed, Proc. Natl. Acad. Sci. 107 (2010) 14390–14395. doi:10.1073/pnas.1005399107.

[22] C.N.A. Palmer, V. Young, M. Ho, A. Doney, J.J.F. Belch, Association of common variation in glutathione S-transferase genes with premature development of cardiovascular disease in patients with systemic sclerosis, Arthritis Rheum. 48 (2003) 854–855. doi:10.1002/art.10955.

[23] A. Lenglet, S. Liabeuf, S. Bodeau, L. Louvet, A. Mary, A. Boullier, A.S. Lemaire-Hurtel, A. Jonet, P. Sonnet, S. Kamel, Z.A. Massy, N-methyl-2-pyridone-5-carboxamide (2PY) — Major metabolite of nicotinamide: An update on an old uremic Toxin, Toxins (Basel). 8 (2016). doi:10.3390/toxins8110339.

[24] M.G. Schooneman, F.M. Vaz, S.M. Houten, M.R. Soeters, Acylcarnitines: Reflecting or inflicting insulin resistance?, Diabetes. 62 (2013) 1–8. doi:10.2337/db12-0466.

[25] C.-C. Tseng, S.-J. Chang, W.-C. Tsai, T.-T. Ou, C.-C. Wu, W.-Y. Sung, M.-C. Hsieh, J.-H. Yen, Reduced incidence of Type 1 diabetes and Type 2 diabetes in systemic sclerosis: A nationwide cohort study, Jt. Bone Spine. 83 (2016) 307–313. doi:10.1016/j.jbspin.2015.06.017.

[26] M.. A. Recasens, C. Puig, V. Ortiz-Santamaria, Nutrition in Systemic Sclerosis, Reumatol. Clínica (English Ed. 8 (2012) 135–140. doi:10.1016/j.reumae.2011.09.003.

[27] I. Marie, A.-M. Leroi, G. Gourcerol, H. Levesque, J.-F. Ménard, P. Ducrotte, Fructose Malabsorption in Systemic Sclerosis, Medicine (Baltimore). 94 (2015) e1601. doi:10.1097/MD.0000000000001601.

[28] T. Matsuse, E. Ohga, S. Teramoto, M. Fukayama, R. Nagai, S. Horiuchi, Y. Ouchi,

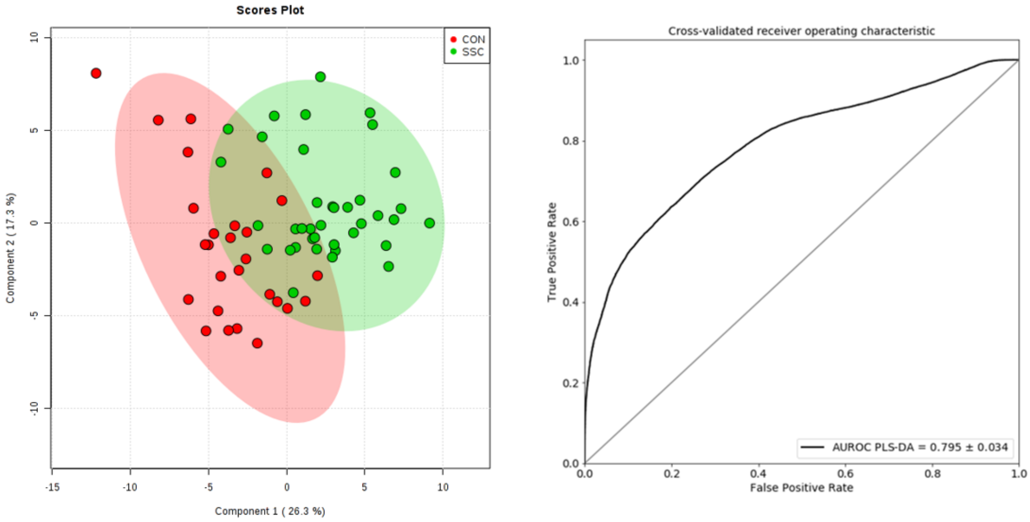
Immunohistochemical localisation of advanced glycation end products in pulmonary fibrosis, J Clin Pathol. 51 (1998) 515–519. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=9797728.

[29] D. Pattanaik, M. Brown, B.C. Postlethwaite, A.E. Postlethwaite, Pathogenesis of systemic sclerosis, Front. Immunol. 6 (2015). doi:10.3389/fimmu.2015.00272.

[30] A. Servettaz, N. Kavian, C. Nicco, V. Deveaux, C. Chéreau, A. Wang, A. Zimmer, S. Lotersztajn, B. Weill, F. Batteux, Targeting the cannabinoid pathway limits the development of fibrosis and autoimmunity in a mouse model of systemic sclerosis., Am. J. Pathol. 177 (2010) 187–96. doi:10.2353/ajpath.2010.090763.

Supplementary Material

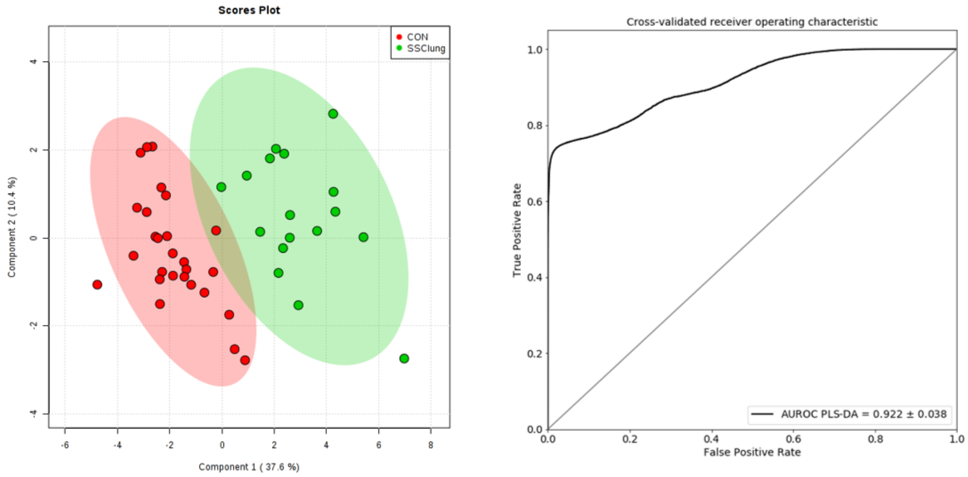
Healthy Controls vs SSc without lung fibrosis



AUROC: 0.795 ± 0.034

Figure 1S. A supervised Partial Least Squares Discriminant Analysis (PLS-DA) from urine samples (SSC with lung fibrosis vs Healthy controls). PLS-DA scores plot (**1Sa),** ROC curve for PLS-DA model validation (**1Sb**). (SSC without lung fibrosis, green plots; HCs, red plots)

Healthy Controls vs SSc with lung fibrosis



AUROC: 0.922 ± 0.038

Figure 2S. A supervised Partial Least Squares Discriminant Analysis (PLS-DA) from urine samples (SSC without lung fibrosis vs Healthy controls). PLS-DA scores plot (**2Sa),** ROC curve for PLS-DA model validation (**2Sb**). (SSC with lung fibrosis, green plots; HCs, red plots)

Table 1S. Molecular and statistical details of unknown compounds that presented significant differences between SSC patients and healthy volunteers. (FC<0, metabolites overexpressed in SSC).

		RT (min)	Mass (Da)	p-value	Fold Change	VIP-value	AUC
Unknowns compounds	Urine	0.84	225.9398	0.0336	-1.55	1.029	0.691
		0.94	150.1349	0.014	-1.85	1.039	0.661
		0.96	226.1170	0.0079	1.64	1.119	0.720
		1.00	258.1934	8.29 E-3	1.60	1.346	0.707
		1.02	236.1671	0.0079	1.91	1.333	0.704
		1.03	218.1647	0.030	2.14	1.051	0.659
		1.04	345.2367	0.022	1.37	1.211	0.674
		1.04	161.1031	0.0045	2.27	1.542	0.728
		1.06	252.1316	0.013	1.67	1.038	0.684
		1.10	364.0932	7.89 E-3	-2.88	1.290	0.717
		1.57	227.0797	0.028	-1.52	1.042	0.653
		1.89	145.0751	0.014	2.46	1.214	0.668
		2.00	230.1376	0.010	1.54	1.013	0.698
		2.45	299.1477	0.027	-1.61	1.344	0.673
		3.66	167.0594	8.58 E-3	-1.92	1.201	0.697
		3.70	149.0461	0.010	-1.94	1.254	0.692
		5.89	177.0818	0.023	1.33	1.037	0.672
		8.06	273.1105	0.018	-1.60	1.018	0.668
		14.77	415.1399	7.89 E-3	-1.68	1.391	0.695
	Plasma	0.95	192.1143	0.039	-1.44	1.405	0.629
		2.27	126.0307	0.038	-1.58	1.004	0.694
		18.1	290.1143	0.036	-1.44	1.544	0.638
		25.36	646.4097	0.046	-1.97	1.335	0.625
		25.98	698.4485	0.039	-1.84	1.328	0.654

Table 2S. Molecular and statistical details of identified urinary metabolites that presented significant differences the between subclasses of SSC disease (dcSSC vs lcSSC; fibrosis vs non-fibrosis). (FC<0, metabolites overexpressed in lcSSC).

	RT (min)	Mass (Da)	p-value	Fold Change	AUC	Formula	Score	Error (ppm)	Metabolite	MS/MS fragments
URINE	1.57	227.080	0.040	-2.124	0.741	$C_{10}H_{13}NO_5$	82.9	-2.98	L-Arogenate	Annotated by Formula
	2.92	173.117	0.016	-1.835	0.695	$C_7H_{15}N_3O_2$	82.9	4.13	Indospicine	Annotated by Formula
	4.26	191.058	0.013	1.678	0.755	$C_{10}H_9NO_3$	90.1	3.74	5-hydroxyindleacetic acid	Annotated by Formula
	6.58	210.064	0.041	-2.766	0.732	$C_9H_{10}N_2O_4$	86.3	1.13	N-(5-amino-2hydroxybenzoyl)glycine	42.0339/80.0493/108.0436/109.0522
	9.51	289.151	0.012	2.243	0.768	$C_{13}H_{23}NO_6$	86.6	2.92	3-methylglutarylcarnitine	60.0808/73.0290/85.0292/101.024

Table 3S. Metabolite differences between healthy controls and systemic sclerosis patients with lung fibrosis. (FC<0, metabolites overexpressed in SSC)

	RT (min)	Mass (Da)	adjusted p-value	Fold Change	AUC	Formula	Score	Error (ppm)	Metabolite	MS/MS fragments
Urine	0.91	182.0788	0.028	-5.99	0.824	$C_6H_{16}O_6$	92.4	1.58	D-Sorbitol	69.0441/83.0598/183.0886
	0.96	113.0588	0.047	1.554	0.776	$C_4H_7N_3O$	97.6	-4.09	Creatinine	43.0288/44.0495/86.0715
	1.37	127.0997	0.019	-5.78	0.878	$C_7H_{13}NO$	90.1	2.06	N-cyclohexylformamide	Annotated by Formula
	1.41	228.1478	0.032	1.339	0.800	$C_{11}H_{20}N_2O_3$	95.4	1.03	L-Leucyl-L-Proline	58.0649/60.0803/70.0648/96.0805
	3.12	135.0323	0.019	1.782	0.837	$C_7H_8N_2O_2$	95.8	-0.44	N1-Methyl-4-pyridine-3-carboxamide	84.0440/108.0442/136.0389/153.0654
	3.53	152.0583	0.044	1.593	0.776	$C_7H_8N_2O_2$	95.9	-2.55	N1-Methyl-2-pyridine-5-carboxamide	42.0333/53.0385/78.0333/108.0441/110.0603/153.0602
	10.68	243.1471	0.042	1.704	0.784	$C_{12}H_{21}NO_4$	96.5	-0.37	Tiglylcarnitine	83.0502/85.0288/185.0810/186.0366
	16.16	189.0423	0.019	1.994	0.847	$C_{10}H_7NO_3$	84.1	3.64	Kynurenic acid	89.0389/116.0489/144.0418
	37.81	301.2235	0.046	2.046	0.776	$C_{16}H_{31}NO_4$	94.6	0.90	Dimethylheptanoyl carnitine	60.0804/85.0282/302.2280
Plasma	1.09	204.0738	0.047	-1.21	0.782	$C_9H_{17}NO_4$	90.4	0.9	O-AcetylCarnitine	Annotated by Formula
	1.10	159.1233	0.027	-1.51	0.824	$C_8H_{17}NO_2$	84.4	4.39	DL-2-Aminooctanoic acid	Annotated by Formula
	22.85	376.2628	0.019	-2.24	0.822	$C_{23}H_{36}O_4$	84.6	4.98	MG(20:5)	57.0685/93.0679/377.2627
	24.34	378.272	0.019	-2.44	0.854	$C_{23}H_{38}O_4$	90.3	6.92	1-Arachinodylglycerol	67.0533/81.0685/95.0839/305.2095/379.2770
	36.48	759.5789	0.019	-1.30	0.793	$C_{42}H_{82}NO_8P$	92.6	0.3	PC(34:1)	60.0805/86.0690/184.0730
	40.27	616.4997	0.038	-1.57	0.797	$C_{39}H_{68}O_5$	89.7	-6.39	DG (18:1/18:3)	Annotated by Formula
	40.88	642.5145	0.021	-1.62	0.824	$C_{41}H_{70}O_5$	68.5	-7.02	DG (38:5)	Annotated by Formula

Table 4S. Metabolite differences between healthy controls and systemic sclerosis patients without lung fibrosis. (FC<0, metabolites overexpressed in SSC)

	RT (min)	Mass (Da)	p-value	FC	AUC	Formula	Score	Error (ppm)	Metabolite	MS/MS fragments
Urine	0.91	182.0788	0.015	-5.88	0.797	$C_6H_{16}O_6$	92.4	1.58	Sorbitol	69.0441/83.0598/183.0886
	1.04	143.0946	0.036	3.25	0.753	$C_7H_{13}NO_2$	97.8	-2.80	Proline Betaine	42.0335/58.0652/84.0810
	3.12	152.0583	0.015	1.65	0.796	$C_7H_8N_2O_2$	95.8	-0.44	N1-Methyl-4-pyridine-3-carboxamide	84.0440/108.0442/136.0389/153.0654
	3.53	152.0583	0.033	1.54	0.762	$C_7H_8N_2O_2$	95.9	-2.55	N1-Methyl-2-pyridine-5-carboxamide	42.0333/53.0385/78.0333/108.0441/110.0603/153.0602
	36.8	299.2078	0.043	2.15	0.753	$C_{16}H_{29}NO_4$	96.6	5.90	Nonenoylcarnitine	85.0218/300.2181
	37.81	301.2235	0.020	2.20	0.782	$C_{16}H_{31}NO_4$	94.6	0.90	Dimethylheptonoyl carnitine	60.0804/85.0282/302.228
Plasma	11.29	264.1132	0.049	-2.22	0.751	$C_{13}H_{16}N_2O_4$	92.7	-4.92	alpha-N-Phenyl-L-Glutamine	84.0436/91.0533/129.0654/130.0490/136.0752/147.0757

Table 5S. Metabolite differences between systemic sclerosis patients with and without lung fibrosis. (FC>0, metabolites overexpressed in SSC without lung involvement)

	RT (min)	Mass (Da)	p-value	Fold Change	AUC	Formula	Score	Error (ppm)	Metabolite	MS/MS fragments
Urine	1.20	216.1466	0.028	2.24	0.732	$C_{10}H_{20}N_2O_3$	86.5	-0.33	Valyl Valine	Annotated by Formula
	1.66	167.1053	0.037	3.20	0.681	$C_7H_5NO_4$	85.1	-1.9	Quinolinic acid	Annotated by Formula
	1.89	115.0636	0.037	2.50	0.675	$C_5H_9NO_2$	86.6	2.93	L-Proline	Annotated by Formula
	2.97	299.0997	0.041	1.89	0.652	$C_{13}H_{17}NO_7$	83.3	0.59	B-D-glucopyrapyranosil anthranilate	Annotated by Formula
	13.26	252.1205	0.037	3.42	0.689	$C_{11}H_{16}N_4O_3$	84.6	4.59	Pro-His	Annotated by Formula
	16.16	189.0423	0.036	1.53	0.696	$C_{10}H_7NO_3$	84.1	3.64	Kynurenic acid	89.0389/116.0489/144.0418
Plasma	1.34	279.1325	0.049	1.52	0.669	$C_{11}H_{21}NO_7$	99.5	0.8	N-(1-Deoxy-1-fructosyl)-Valine	72.0807/216.1234/244.1188/262.1292
	2.51	293.1488	0.033	1.48	0.679	$C_{12}H_{23}NO_7$	83.9	-1.43	N-(1-Deoxy-1-fructosyl)-leucine	86.0967/144.1020/230.1400/258.1349/276.1452
	2.70	293.1492	0.048	1.52	0.670	$C_{12}H_{23}NO_7$	88.7	-2.95	N-(1-Deoxy-1-fructosyl)-Isoleucine	86.0967/144.1020/230.1400/276.1452

Table 6S. Members from the Precisesads Clinical Consortium.

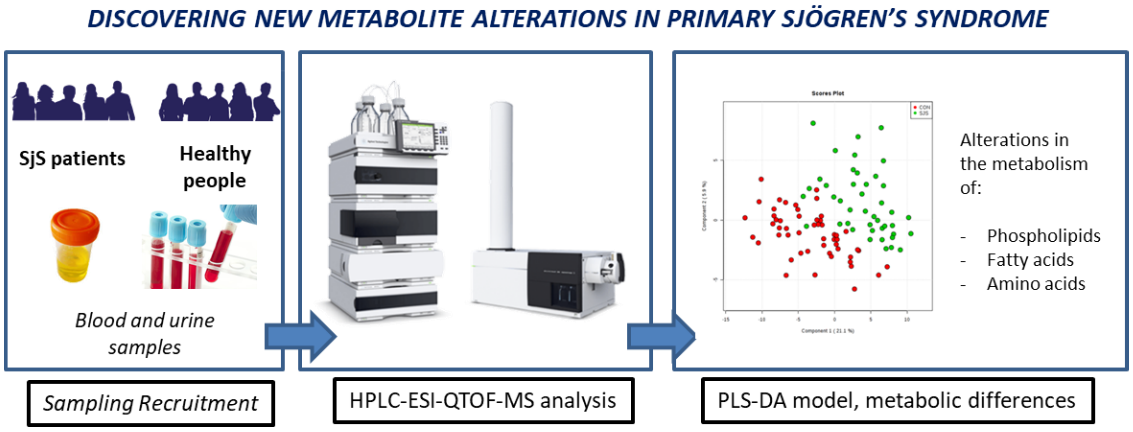
Clinical center	Principal investigator	Clinicians
Hospital Regional de Málaga, Servicio Andaluz de Salud , Málaga (Spain)	Enrique de Ramón Garrido	

Hospital Universitario San Cecilio, Servicio Andaluz de Salud, Granada (Spain)	Norberto Ortego; Enrique Raya	**María Concepción Fernández Roldán;
		José Luis Callejas Rubio;
		Raquel Ríos Fernández;
		Inmaculada Jiménez Moleón;**

Hospital Universitario Reina Sofía Andaluz de Salud, Córdoba (Spain)	Eduardo Collantes	**Rafaela Ortega-Castro Mª Angeles Aguirre-Zamorano Alejandro Escudero- Contreras Mª Carmen Castro-Villegas**
Centre Hospitalier Universitaire de Brest, Hospital de la Cavale Blanche, Brest, (France)	Jacques-Olivier Pers	**Alain Saraux Valérie Devauchelle-Pensec Divi Cornec Sandrine Jousse-Joulin**
Fondazione IRCCS Ca Granda Ospedale Maggiore Policlinico, Milano (Italy)	Lorenzo Beretta	
Deutsches Rheuma-Forschungszentrum Berlin, (Germany)	Falk Hiepe	
Hospitaux Universitaires de Genève (Switzerland)	Carlo Chizzolini	
Centro Hospitalar do Porto (Portugal)	Carlos Vasconcelos	**Ana Campar António Marinho Fátima Farinha Isabel Almeida Mariana Brandão Raquel Faria**
Medizinische Hochschule Hannover (Germany)	Torsten Witte	**Niklas Baerlecken**
Katholieke Universiteit Leuven (Belgium)	Rik Lories	**Ellen De Langhe**
Université catholique de Louvain (Belgium)	Bernard Lauwerys	
Università degli studi di Milano, (Italy)	Pier Luigi Meroni	
Klinikum Der Universitaet Zu Koeln, (Germany)	Nicolas Hunzelmann	**Doreen Belz**
Medizinische Universitat Wien, (Austria)	Georg Stummvoll	**Michael Zauner, Michaela Lehner**
University of Szeged, (Hungary)	Laszló Kovács	**Attila Balog, Magdolna Deák, Márta Bocskai, Sonja Dulic, Gabriella Kádár**
Hospital Clinic I Provicia, Institut d'Investigacions Biomèdiques August Pi i Sunyer, Barcelona (Spain)	Ricard Cervera	**Ignasi Rodríguez-Pintó, Gerard Espinosa**
Hospital Universitario Marqués de Valdecilla, Servicio Cántabro de Salud, Santander (Spain)	Miguel A. González-Gay	**Ricardo Blanco Alonso Alfonso Corrales Martínez**
Andalusian Public Health System Biobank	Blanca Miranda	**Rocío Aguilar Quesada**
Project Office - Recruitment and data follow up	**Jacqueline Marovac & Tania Gomes Anjos**	

Capítulo 5

Estudio del síndrome de Sjögren en muestras de orina y plasma mediante una estrategia metabolómica no dirigida basada en HPLC-ESI-QTOF-MS


DISCOVERING NEW METABOLITE ALTERATIONS IN PRIMARY SJÖGREN'S SYNDROME
SjS patients
Healthy people
Blood and urine samples
Sampling Recruitment
HPLC-ESI-QTOF-MS analysis
Scores Plot
CON
SJS
Alterations in the metabolism of:
- Phospholipids
- Fatty acids
- Amino acids
PLS-DA model, metabolic differences

Álvaro Fernández-Ochoa, Isabel Borrás-Linares, Rosa Quirantes-Piné, PRECISESADS Clinical Consortium, Marta E. Alarcón Riquelme, Lorenzo Beretta, Antonio Segura-Carretero

Journal of Pharmaceutical and Biomedical Analysis

Discovering new metabolite alterations in primary Sjögren's Syndrome in urinary and plasma samples using an HPLC-ESI-QTOF-MS methodology

ABSTRACT

Sjögren's Syndrome (SjS) is a complex autoimmune disease characterized by the affection of the exocrine glands and the involvement of multiple organs. Although a greater number of biomarker studies have been carried out in recent years, the origin and pathogenesis are not yet well known and therefore there is a need to continue studying this pathology. This work aims to find metabolic changes in biological samples (plasma and urine), which could help identify the metabolic pathways affected by the SjS pathogenesis. The samples collected from SjS patients and healthy volunteers were analyzed by a fingerprinting metabolomic approach based on HPLC-ESI-QTOF-MS methodology. After feature pre-selection by univariate statistical tests, an integrated PLS-DA model using data from urine and plasma was constructed obtaining a good classification between cases and controls (AUROC = 0.839 ± 0.021). 31 and 38 metabolites in plasma and urine, respectively, showed significant differences between healthy volunteers and SjS patients and were proposed for their identification. From them, 12 plasma and 24 urinary metabolites could be annotated. In general, the main metabolic pathways altered in SjS patients were related to the metabolism of phospholipids, fatty acids, and amino acids, specially tryptophan, proline and phenylalanine.

Keywords. Metabolomics, HPLC-ESI-QTOF-MS, Sjögren's Syndrome, biomarkers, unsaturated fatty acids, phosphatidylinositols,

1. Introduction

Sjögren's syndrome (SjS) is a complex systemic autoimmune disease mainly characterized by lymphocytic infiltration of exocrine tissues, leading to salivary and lacrimal hypofunction. In addition, because of a widespread activation of the immune system, SjS can also affect several organs and tissues, including skin, lung, heart, joints, nervous system or kidney, among others [1]. SjS may present alone (primary SjS, pSjS) or in association with other autoimmune diseases, such as systemic lupus erythematosus (SLE), rheumatoid arthritis (RA) or systemic sclerosis (SSC) (secondary SjS, sSjS).

Major advances in understanding SjS pathophysiology have been achieved during the last years thanks to the application and advances in the field of the "omic" sciences [2].

Metabolomics, which is focused on the study of molecules with low molecular weight present in biological systems, is the last link of the 'omics' cascade. This tool is able to find alterations in metabolic pathways that allow to improve the knowledge about the origin and pathogenesis of diseases [3]. In SjS, fingerprinting metabolomic studies have previously been performed using blood [4,5] and salivary samples [6,7]. In these works, the samples were analyzed by gas-chromatography coupled to mass spectrometry (GC-MS) [6,8] proton magnetic resonance spectroscopy (1H-MRS) [7] and, more recently by liquid chromatography coupled to mass spectrometry (LC-MS) [5]. NMR has better characteristics for quantification and sample treatment whereas the advantages of MS are related to a higher selectivity, sensitivity as well as the possibility of being coupled to chromatographic techniques allowing to detect a higher number of metabolites [9].

It is interesting to observe that no study has been conducted so far on urine samples from SjS subjects despite the benefit of analyzing this biofluid. This easily accessible biological fluid may be of high diagnostic value and reveal important information related to renal involvement [10]. In this way, interesting biomarkers have been found in urine in other immune-mediated inflammatory diseases which may be related to pSjS such as SLE or RA, among others [11].

In order to improve the knowledge of pSjS, this work focused, for the first time, on the metabolomic analysis of urine and plasma samples using the powerful analytical platform, HPLC-ESI-QTOF-MS.

2. Materials and methods

2.1. Study population and design.

Forty-three patients with a diagnosis of pSjS according to the American-European Consensus Group (AECC) criteria for the classification of pSjS recruited from European centers participating in the PRECISESADS project (www.precisesads.eu; clinicaltrials.gov registration number NCT0289012) were included. 52 age- and sex-matched healthy volunteers (Healthy controls, HC) participated in the study. All the participants signed a written informed consent for the study that was approved by local ethic committees.

Plasma samples were collected after centrifugation (1500 g, 10 min, 4ºC) of blood samples obtained in tubes with dipotassium ethylenediaminetetraacetic (K_2EDTA). On the other hand, single urine-plot samples were collected and centrifuged (2500 g, 10 min, 4 ºC). These biological samples were frozen at -80 ºC until sample treatment and analysis.

2.2. Sample treatment.

Frozen plasma and urinary samples were thawed on ice. Quality control samples (QC samples) for each matrix were constituted by combining equal volumes (20µl) of each sample. These QC samples were treated as the rest of the samples as detailed below.

Urine samples were diluted with Milli-Q H_2O in order to obtain an osmolality value of 100 mOsm/kg. This measurement of osmolality was performed by an OSMOMAT 3000 osmometer (Gonotec, Berlin, Germany). In this way, the different participants' hydration states were corrected [12].

Regarding plasma samples, the protein content in 100 µl aliquots was precipitated by adding 200 µl of a mixture of methanol and ethanol (50:50, v/v). In order to guarantee an effective precipitation, the mixture was kept in cold (-20 ºC) during 20 min. The supernatant was evaporated to dryness under vacuum. The solid residue was reconstituted with 0.1% aqueous formic acid:methanol (95:5, v/v).

The last step of the sample treatment for both matrices was a centrifugation for 10 min at 25830 g and 4º C. The supernatants were transferred to vials to be analysed by HPLC-ESI-QTOF-MS.

2.3. HPLC-ESI-QTOF-MS analysis.

The instrument used for the analysis of both biological samples consisted of an Agilent 1260 HPLC system, a Jet Stream dual ESI interface and an Agilent 6540 Ultra High Definition (UHD) Accurate Mass Q-TOF spectrometer.

For the chromatographic separation, a C18 analytical column (Zorbax Eclipse Plus, 2.1×150 mm, 3.5 µm) was used operating at 25 ºC. The mobile phases consisted of water containing 0.1% of formic acid (A) and methanol (B) which were delivered at 0.4 ml/min. Specific chromatographic conditions for urine and plasma, such as injection volumes and elution gradients are detailed in the Supplementary Material (**Table 1S)**.

The MS scanned from 50 to 1700 m/z and the masses were calibrated by means of continuous infusion of purine (m/z 121.050873) and HP-921 (m/z 922.009798). Ultrahigh pure nitrogen was used as drying and nebulizer gas at temperatures of 200 and 350 ºC and flows of 10 and 12 L/min, respectively.

Due to the large number of samples, these were analysed in three different analytical batches. A QC sample was injected each five randomized biological samples analysed throughout sequences in order to check the instrumental reproducibility. The QC sample was also analysed by tandem mass spectrometry (MS/MS) to facilitate the identification of potential biomarkers. For these analyses, nitrogen was used as collision gas with energy values of 10 eV, 20 eV and 40 eV.

2.4. Data processing.

Firstly, the chemical features found in the acquired data were extracted, aligned and integrated using MassHunter Profinder Software (B.06.00, Agilent Technologies), which provides untargeted molecular feature extraction for batches of LC-MS data. For feature alignment, a retention time window of 0.25 min and a mass range of 20 ppm ± 2 mDa were assigned. For peak noise removal, features with intensity lower than 1000 counts were filtered. Meanwhile $[M+H]^+$, $[M+Na]^+$ and $[M+H-H_2O]^+$ were the possible adducts searched for. Due to the large number of sample files and their size, the chosen strategy was to perform the molecular feature extraction in the quality control files acquired along the three analysis batches. Those molecular features found in QC samples were used to perform a targeted search in the tested samples of all participants. For peak processing, Agile2 was the method of integration although the obtained areas were manually supervised to correct failures in the automatic integration step. In order to avoid features coming from diet or drugs, the features present in less than 25 % of the studied samples were removed. Furthermore, an additional filtration step was performed according to the variability throughout the sequence. Thus, features were also filtered if the relative standard deviation (RSD) was higher than 25 % in the QC samples after normalization.

2.5. Statistical analysis.

Firstly, Principal Component Analysis (PCA) was performed in MetaboAnalyst 3.0 software [13] in order to verify the analytical reproducibility and detect possible outliers. Due to the analytical drifts observed in plasma sample analysis, a normalization strategy was applied using the batch effect correction tool from MetaboAnalyst 3.0 software [13]. This procedure is based on empirical Bayes method. In order to correct the within-batch effect, each analysis was normalised by MS total useful signal (MSTUS). Urine data was also normalized by MSTUS.

Before multivariate data analysis, univariate analysis based on t-test and fold change (FC) was used to eliminate the non-significant features. The univariate t-test was conducted on cube-root transformed data and p-values were corrected with the False Discovery Rate (FDR) method with a cut-off of 0.05 and the FC threshold was of 1.2.

In order to get a better explanation and to detect possible correlations, features of urine and plasma from the same person were grouped as a single file. To make features comparable to each other data were transformed by cube root transformation and Pareto scaling. Supervised Partial Least Squares Discriminant Analysis (PLS-DA) models were created with the integrated data of plasma and urine samples. The discriminatory ability of PLS-DA models after feature selection was evaluated by means of the Area Under Receiver Operating Characteristic curve (AUROC) after extensive internal cross-validation (CV); to this end 1000 runs of 10-fold CV were used.

Univariate and PLS-DA analysis were performed using the scikit-learn algorithm (http://scikit-learn.org/stable/index.html) along with custom python codes developed by LB.

2.6. Metabolite identification and metabolite pathway analysis.

The identification was carried out through the comparison of the accurate mass, isotopic distribution and fragmentation patterns obtained in MS/MS analysis with the online available metabolomic databases. Databases were searched by CEU Mass Mediator tool. This tool allowed the simultaneous metabolite search in several databases such as LipidMaps, KEGG as well as Human Metabolome Database. The MS/MS patterns were also compared with MS/MS fragmentation resources, concretely

MetFrag. For a better biological interpretation, the Pathway Analysis module in MetaboAnalyst 3.0 [13] was used with the metabolites that were annotated.

3. Results and discussion.

As mentioned before, the main objective of the present research was to look for differentiated biomarkers in plasma of urine collected from pSjS patients compared to healthy controls. To our knowledge, this is the first study that delves into the Sjögren's syndrome providing altered metabolites from urinary samples.

Urine and plasma data were aggregated to create a unique PLS-DA model capable of discriminating pathological cases with respect to HC. The metabolites responsible for this differentiation point out to a number of alterations in metabolic pathways of SjS patients that could reflect several of the pathophysiological processes. In the following subsections, the results obtained are presented together with the discussion of the most relevant metabolites found affected in this pathology.

3.1. Data quality assessment.

The Molecular Feature Extraction process (MFE) exposed 554 and 571 molecular features extracted from the QCs of plasma and urine samples, respectively. An initial overview of the run quality was obtained by PCA of the whole data set including all the QC injections. As it is shown in PCA scores (**Fig. 1a, Fig. 1b**), QC samples were not well grouped. This behaviour was much more significant in plasma samples. The observed non-grouping effect was caused by instrumental variations between the three analytical batches and within each batch due to the injection order. In this sense, non-biological experimental variations or batch effects are commonly observed across multiple batches of LC-MS analyses. The sum of these influences make samples from different batches not directly comparable. In our case, both effects were clearly affecting QCs and study samples **(Fig. 1Sa)** and a data normalization procedure had to be implemented to make statistics meaningful and obtain biological information. The applied normalization step based on Bayes method enabled the equalisation of the intensity between batches. However, the drift due to injection order effect remained, as it is depicted in **Fig. 1Sb**. In order to correct the within-batch effect, each analysis was also normalised by the MS total useful signal (MSTUS). After applying both normalisation steps, PCA showed good clustering of QCs (**Fig. 1c**).

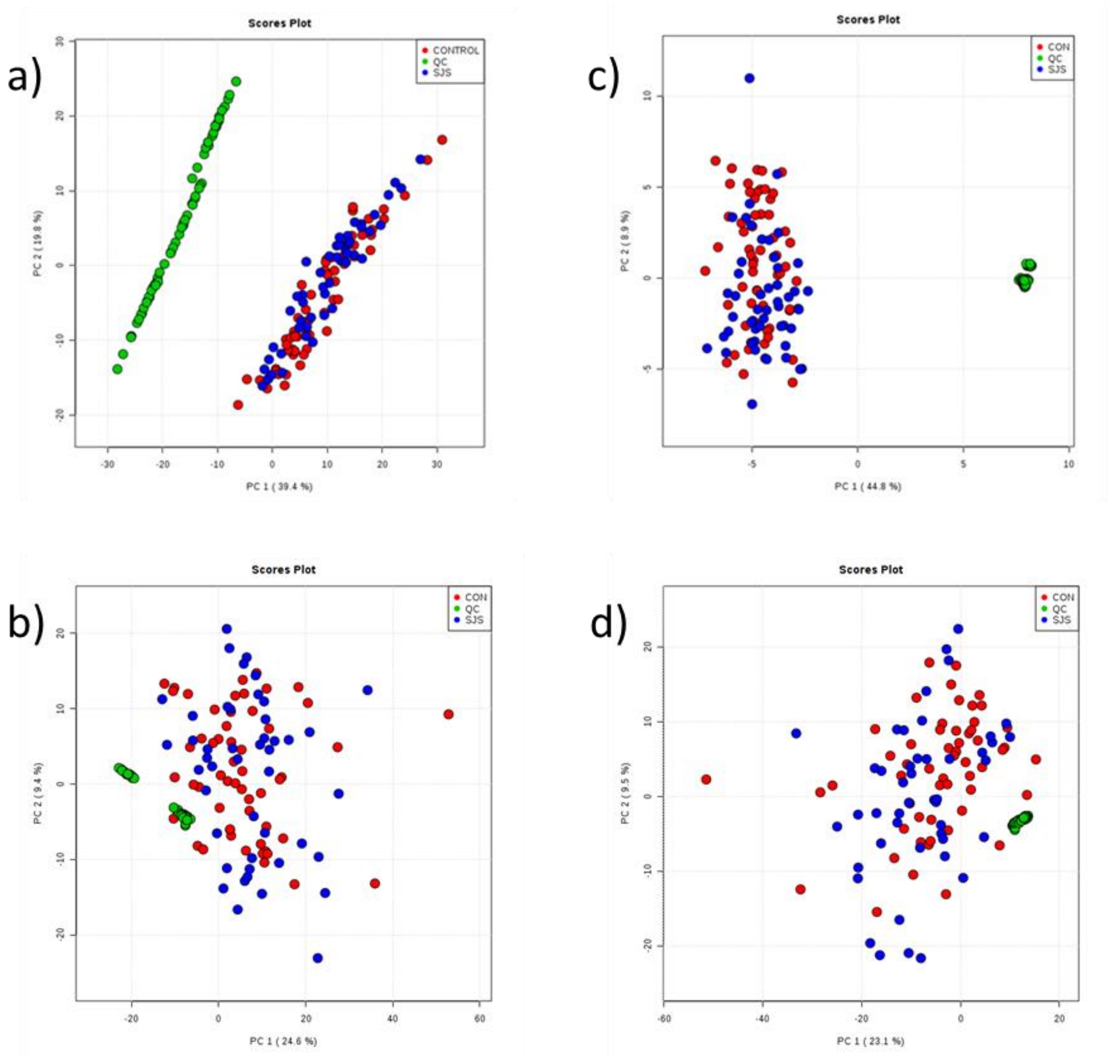


Fig. 1. PCA scores plot from data before batch normalization (1a: Plasma, 1b: Urine) and after batch normalization (1c: Plasma, 1d: Urine). (Healthy controls, red dots; pSjS, blue dots; QCs, green dots)

On the contrary, urinary data did not show those large drifts due to between-batch effect observed in plasma samples. Nevertheless, only one normalization step based on MSTUS was performed in this data set in order to correct the small differences due to the injection order (**Fig. 1d**).

3.2. Feature selection in urine and plasma.

Because many variables are irrelevant in response to the study question, variable selection by univariate analysis was performed before multivariate modelling. Thus, 60 and 108 molecular features were obtained with FDR-adjusted p-values less than 0.05 and FC higher than 1.2 in plasma and urine, respectively. The peak area of these features in both types of samples were integrated and used for the PLS-DA analysis. Two principal components were selected for the model, which explained 21.1 % and 5.9 % of data variability. **Fig. 2a** shows the scores plot of the two principal components where a separation between HC and pSjS samples is clearly produced. For the PLS-DA assessment, values of 0.835, 0.659 and 0.430 were obtained for accuracy, R^2 and Q^2, respectively. Generally, a Q^2 value higher than 0.4 and differences between R^2 and Q^2 less than 0.3 are considered acceptable in biological studies [14].

An AUROC value of 0.839 ± 0.021 (mean ± standard deviation) (**Fig. 2b**) with a permutation p-value less than 5 E-4 (**Fig. 2c**) were obtained after the extensive cross-validation of the model created with the integrated data from urine and plasma. The same methodology was used for the sets of data separately. AUROC values of 0.723 ± 0.030 and 0.771 ± 0.025 (**Fig. 2S**) for urine and plasma dataset, respectively, were obtained. Therefore, the integration of urine and plasma data for the creation of a single PLS-DA model improved the results obtaining a better classification of the samples compared to the single plasma and urine datasets, separately.

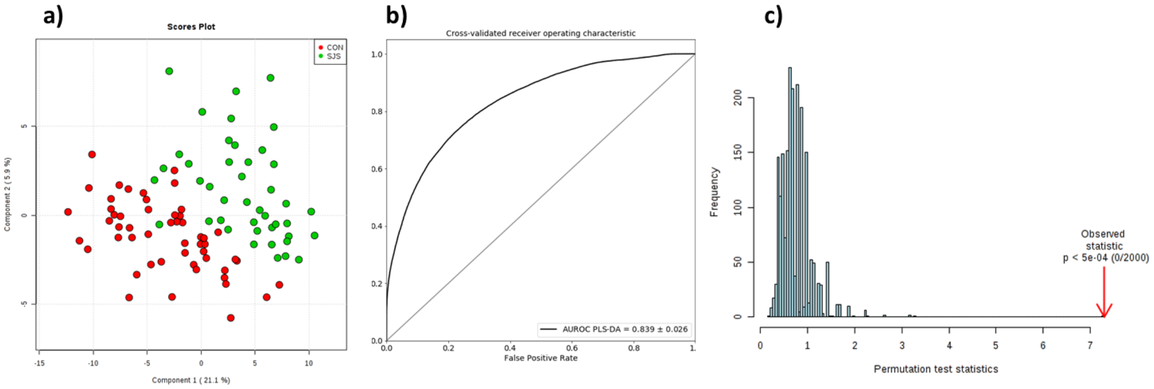


Fig. 2. PLS-DA model from the integration of urine and plasma data. PLS-DA Scores plot (2a) (HC, red dots; pSjS, green dots). ROC Curve (2b). Permutation test results (2c)

The molecular features with a VIP-value higher than 1.0 in the PLS-DA model were proposed for identification. According to the Metabolomics Standards Initiative (MSI), the proposed metabolites were putatively annotated. **Tables 1-2** list these annotated metabolites with their chemical characteristics and their statistical results. Other characterization information (MS/MS fragments, scores and errors) are shown in **Tables 2S-3S**. As metabolite identification remains a bottleneck in metabolomic studies based on mass spectrometry, several features could not be identified and their results are detailed in **Tables 4S-5S**.

Table 1. Retention times (RT), masses and statistical results from significant metabolites present in plasma samples. (FC < 0, metabolites overexpressed in pSjS). The superscripts in the metabolite column indicate the level of identification of the metabolites.

RT (min)	Mass (Da)	p-value	FDR	Fold Change	VIP-value	AUC	Molecular Formula	Metabolite
1.48	149.0484	6.3 E-4	2.6 E-3	-1.27	1.22	0.700	$C_5H_{11}NO_2S$	*L-Methionine*
4.34	208.0857	2.5 E-3	9.5 E-3	1.44	1.32	0.636	$C_{10}H_{12}N_2O_3$	L-kynurenine
9.05	184.1221	4.9 E-3	0.013	1.29	1.14	0.678	$C_9H_{16}N_2O_2$	*N-(3-aminopropyl)pyrrolidin-2-one*
21.13	519.3225	4.7 E-3	0.015	-1.27	1.02	0.677	$C_{26}H_{50}NO_7P$	LysoPC(18:2)
21.18	543.3324	6.0 E-5	1.4 E-3	1.53	1.42	0.739	$C_{28}H_{50}NO_7P$	LysoPC(20:4)
24.24	328.2354	8.8 E-4	0.030	1.78	1.19	0.693	$C_{22}H_{32}O_2$	Docosahexaenoic acid (DHA)
24.52	302.2129	3.6 E-3	0.012	1.65	1.05	0.701	$C_{20}H_{30}O_2$	Eicosapentaenoic acid (EPA)
24.52	928.623	1.6 E-3	0.011	1.98	1.13	0.697	$C_{49}H_{95}O_{12}P$	PI(P20:0/20:0)
25.59	278.2184	6.6 E-4	0.024	1.75	1.22	0.752	$C_{18}H_{30}O_2$	Linolenic acid
26.20	934.6784	1.7 E-4	0.019	1.89	1.34	0.769	$C_{51}H_{99}O_{12}P$	PI(P20:0/22:O)
26.20	282.2518	4.0 E-5	1.9 E-3	1.60	1.44	0.748	$C_{18}H_{34}O_2$	Oleic acid
30.84	672.5185	7.3 E-4	0.010	1.43	1.21	0.720	$C_{37}H_{73}N_2O_6P$	PE-Cer(d14:2/21:0)

Table 2. Retention times (RT), masses and statistical results from significant metabolites present in urinary samples. (FC < 0, metabolites overexpressed in pSjS). The superscripts in the metabolite column indicate the level of identification of the metabolites.

RT (min)	Mass (Da)	p-value	FDR	Fold Change	VIP-value	AUC	Molecular Formula	Metabolite
0.95	113.0612	3.0 E-3	9.5 E-3	-1.28	1.10	0.677	$C_4H_7N_3O$	Creatinine
0.96	131.0702	3.7 E-3	8.7 E-3	-1.23	1.05	0.674	$C_4H_9N_3O_2$	Creatine
1.32	268.1164	7.7 E-4	2.6 E-3	-1.53	1.20	0.713	$C_{11}H_{16}N_4O_4$	Histidinyl-Hydroxyproline
1.51	138.0449	2.7 E-4	1.4 E-3	-1.35	1.30	0.723	$C_6H_6N_2O_2$	Urocanic acid
1.53	187.0960	2.9 E-3	7.8 E-3	-1.44	1.34	0.687	$C_7H_{13}N_3O_3$	5-guanidino-3methyl-2-oxopentanoate
1.62	167.0219	3.5 E-3	0.026	1.26	1.38	0.600	$C_7H_5NO_4$	Quinolinic acid
1.96	214.1316	8.2 E-3	0.049	-1.27	1.15	0.656	$C_{10}H_{18}N_2O_3$	Valyl-Proline
2.86	228.1475	1.2 E-3	8.5 E-3	-1.31	1.17	0.695	$C_{11}H_{20}N_2O_3$	Leucyl-Proline
3.37	261.1570	3.8 E-4	4.2 E-3	-1.33	1.27	0.703	$C_{12}H_{23}NO_5$	Hydroxyisovaleryoyl carnitine
4.85	166.0488	1.6 E-3	0.039	-1.69	1.13	0.687	$C_6H_6N_4O_2$	MethylXanthine
4.93	246.1211	1.6 E-3	0.012	-1.27	1.13	0.679	$C_{10}H_{18}N_2O_5$	L-gamma-glutamyl-L-valine
4.96	165.0795	3.8 E-3	0.026	-1.25	1.04	0.697	$C_9H_{11}NO_2$	L-phenylalanine
5.65	231.1468	3.0 E-3	0.012	-1.27	1.06	0.692	$C_{11}H_{21}NO_4$	Butyrylcarnitine
6.56	210.0639	9.5 E-4	0.012	1.86	2.10	0.680	$C_9H_{10}N_2O_4$	N-(5-amino)-2-hydroxybenoylglycine
7.05	297.1070	9.8 E-3	0.016	-1.20	1.16	0.653	$C_{11}H_{15}N_5O_5$	Methylguanosine
8.43	285.0960	5.2 E-3	0.042	-1.35	1.01	0.650	$C_{11}H_{15}N_3O_6$	N4-Acetlycytidine
8.60	157.0726	1.9 E-3	8.3 E-3	-1.56	1.12	0.680	$C_7H_{11}NO_3$	Tiglylglycine
9.31	159.0894	3.2 E-4	1.4 E-3	-1.44	1.28	0.730	$C_7H_{13}NO_3$	Isovaleroylglycine
10.55	187.0635	2.4 E-3	0.010	-1.31	1.09	0.689	$C_{11}H_9NO_2$	Indoleacrylic acid
10.68	243.147	2.6 E-3	9.5 E-3	-1.33	1.08	0.677	$C_{12}H_{21}NO_4$	Tiglylcarnitine
11.3	260.1377	8.5 E-3	0.049	-1.20	1.16	0.658	$C_{11}H_{20}N_2O_5$	Gamma-glutamylisoleucine
16.17	189.0423	3.4 E-3	0.026	-1.20	1.10	0.636	$C_{10}H_7NO_3$	Kynurenic acid
36.17	329.2200	8.6 E-3	0.049	-1.45	1.06	0.644	$C_{17}H_{31}NO_5$	6-keto-decenoylcarnitine
40.36	488.237	6.6 E-4	1.0 E-4	-2.32	1.71	0.783	$C_{23}H_{32}N_6O_6$	Peptide

3.3. *Analysis of the metabolic pathways.*

The pathway analysis was performed with the annotated metabolites detailed in the previous section. **Fig 3.** shows the main metabolic pathways affected by the Sjögren's Syndrome according to the p-value and impact values. The results highlighted the most affected pathways were the following: phenylalanine metabolism, tryptophan metabolism, histidine metabolism, alpha-linolenic acid metabolism, arginine and proline metabolism, and cysteine and methionine metabolism.

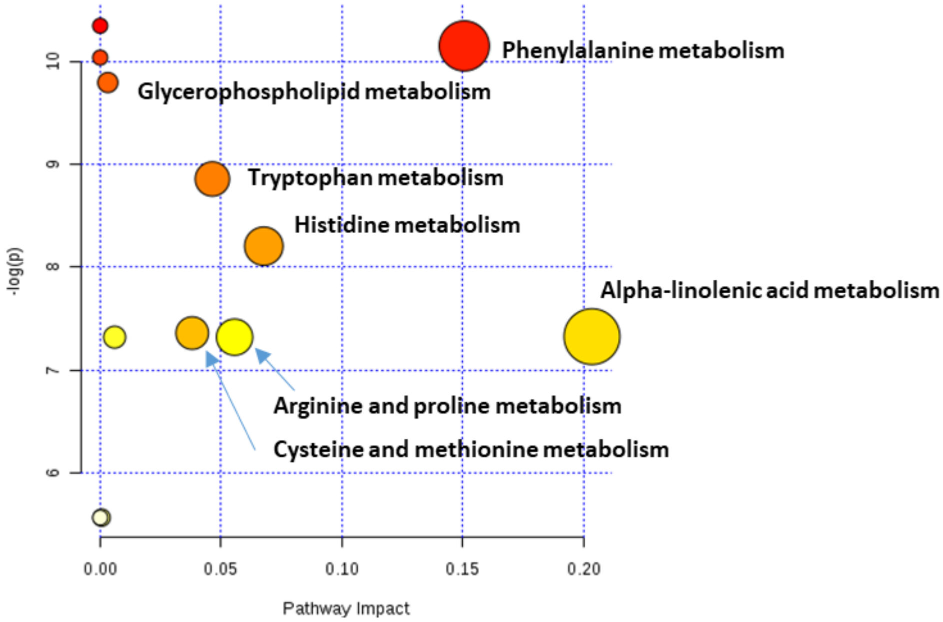


Fig. 3. The pathway impact analysis of the significant metabolites using Metaboanalyst 3.0.

One of the most outstanding results has been the number of metabolites derived from amino acids found deregulated in urine samples, much of them dipeptides. These metabolites include histidinyl-hydroxyproline, leucyl-proline, valyl-proline, phenylalanine, gamma-glutamylvaline and gamma-glutamylisoleucine. This is a clear evidence of amino acid metabolism alteration in Sjögren's Syndrome. Mainly, the main amino acid involved in these derivatives was proline. According to this finding, phenylalanine, proline and glycine were also reported as altered in salivary samples of SjS patients in previous studies [7]. In addition, another recent research reported a higher concentration of L-proline in SjS patients than in the healthy population indicating the possible role that this amino acid could have in this disease [5].

Moreover, metabolites involved in tryptophan metabolism were also found to be significant. Specifically, these altered metabolites were quinolinic and kynurenic acids in urine and L-kynurenine in plasma. They are related to kynurenine pathway, which is responsible for the synthesis of nicotinamide adenine dinucleotide from tryptophan. The kynurenine pathway has a large impact in recent years due to its relation with the immune system, inflammation and neurological processes [15]. In fact, our findings are in concordance with previous results which have suggested that tryptophan metabolism regulate the immune response in primary Sjögren’s syndrome [16]. As these previous studies showed, these results indicate the increased activity of the enzyme indoleamine-pyrrole 2,3-dioxygenase, which is involved in tryptophan degradation. These metabolites derived from the tryptophan metabolism seem to be highly related to the gut microbiota [17]. In this way, recent studies have suggested the role that exists between the pathogenesis of autoimmune diseases and the microbiome [18]

The results of differential metabolite identification suggest that disorders of the metabolism of unsaturated fatty acids (UFAs) were also involved in SjS. Docosahexaenoic (DHA), eicosapentaenoic (EPA), linolenic and oleic acids appeared up-regulated in plasma. Although the high level of free fatty acids in blood, especially palmitic acid, have shown to be involved in the progression of SjS, the effect of UFAS seems to be the opposite [19]. Previous work has demonstrated that omega-3 fatty acids are beneficial in several human inflammatory and autoimmune diseases [20]. The significance of increased UFAS levels in SjS is unclear, but it could be related to a change in their metabolism, namely in beta oxidation. This is supported by the presence of other families of metabolites involved in this pathway that have also been detected altered in plasma and urine, as the case of acylcarnitines.

In this sense, acylcarnitines are involved in the metabolism of fatty acids as well as amino acid oxidation. The alterations of 6-keto-decenoylcarnitine, tiglylcarnitine, hydroxyisovaleroyl carnitine and butyrylcarnitine in urinary samples indicate that fatty acid β-oxidation pathway was affected. Nevertheless, these discrepancies suggest that unsaturated fatty acids metabolism imbalance in SjS patients deserves further attention.

Another family of metabolites also involved in fatty acid beta-oxidation metabolism are acylglycines. Metabolites belonging to this family were found altered in urine samples of SjS patients. Specifically, these metabolites were N-(5-amino)-2-hydroxybenoylglycine, tiglylglycine and isovalerylglycine. This family of compounds, which are frequently produced by the enzyme glycine N-acyltransferase, has been used in the detection of inborn errors of metabolism where amino acids and organic acids are involved. In this way, the deregulation of the excretion of these metabolites has shown to be highly related with perturbations in the mitochondrial fatty acid beta-oxidation [21]. So, the results may be indicating deregulations in these pathways in SjS patients that should be further explored.

Two lysophosphatidylcholines (LPCs) were found disturbed in plasma samples of SjS patients (LysoPC(18:2), LysoPC(20:4). LPCs are bioactive phospholipids which are originated by the hydrolysis of phospathidylcholines (PCs) mediated by the phospholipase A_2 in living cells. Thus, this result evidenced that phospholipid metabolism is affected in SjS patients. In fact, the relationship of LPCs with inflammatory processes and the modulation of the immune response is well-known [22]. Previous studies have reported the implication of LPCs in the pathogenesis of atherosclerosis [23] and SLE [24]. Thus, the deregulation of LPCs concentration in plasma has been reported in the literature in numerous studies mainly focused on cancer [25], obesity, or type 2 diabetes, among others [26]. These studies reported that lower concentrations of LPCs are related to a higher risk of suffering from these diseases and our findings may justify the increased risk of diabetes mellitus observed in SjS subjects [27]. Related to these compounds, other phospholipid metabolites identified as differential metabolites of SjS in plasma samples were the phosphatidylinositols PI(20:0/20:0) and PI(P20:0/22:O). These phospholipids are important precursors for many biologically active mediators of metabolism including eicosanoids, diacylglycerol and platelet-activating factor. The abnormal phospholipid metabolic pathway may not only result in the abnormal physiology and metabolism via a variety of pathways, but also promote the systemic inflammatory state [28]. These results are supported by the phospholipid disturbance previously described in diseases such as SLE and RA [29,30].

4. Conclusions

The integrated PLS-DA model using aggregated HPLC-MS data from urine and plasma analysis of HC and pSjS patients showed better results than models built considering data from each single biofluid separately. Our optimized methodology put into light that the most deregulated metabolites in pSjS patients were unsaturated fatty acids, phosphatidylinositols, acylglycines, lysophosphatidylcholines, acylcarnitines and metabolites related to amino acid pathways specially tryptophan, proline and phenylalanine metabolism. These findings may have implications in the pathogenesis or in the progression of the disease and reflect the systemic affectation. Nevertheless, larger studies should be carried out to confirm these results and to better highlight their functional implications.

Compliance with Ethical Standards

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study

Conflict of interest

All authors declare that they have no conflict of interest.

Author contributions

A.F did the sample treatments, realized their analysis, performed the data treatment and wrote the manuscript in collaboration with R.Q. PRECISESADS Clinical Consortium (**Table 6S**) included and collected clinical data from patients in this study. L.B contributed with statistical analysis and results discussion. A.S and M.E.A designed and supervised the study.

Acknowledgements

The work described has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement No. 115565, the resources for which are composed of a financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and the European Federation of Pharmaceutical Industries and Associations (EFPIA) companies' in-kind contribution. We thank all the members of PRECISESADS Clinical Consortium (**Table 6S**) for their effort in the sample recruitment, distribution and assessment of the samples used in this study. The author AFO gratefully acknowledges the Spanish Ministry of Education, Culture and Sports for the FPU grant (FPU14/03992).

Bibliography

[1] G.E. Katsifis, N.M. Moutsopoulos, S.M. Wahl, T Lymphocytes in Sjögren's Syndrome: Contributors to and Regulators of Pathophysiology, Clin. Rev. Allergy Immunol. 32 (2007) 252–264. doi:10.1007/s12016-007-8011-8.

[2] J. Ai, S. Feng, K. Misuno, S. Hu, Integrated Omics Analysis of Sjogren's Syndrome, J. Integr. OMICS. 2 (2012) 6–10. doi:10.5584/jiomics.v2i2.97.

[3] A. Agin, D. Heintz, E. Ruhland, J.M. Chao de la Barca, J. Zumsteg, V. Moal, A.S. Gauchez, I.J. Namer, Metabolomics - an overview. From basic principles to potential biomarkers (part 1), Med. Nucl. 40 (2016) 4–10. doi:10.1016/j.mednuc.2015.12.006.

[4] A.A. Bengtsson, J. Trygg, D.M. Wuttge, G. Sturfelt, E. Theander, M. Donten, T. Moritz, C.-J. Sennbro, F. Torell, C. Lood, I. Surowiec, S. Rännar, T. Lundstedt, Metabolic profiling of systemic lupus erythematosus and comparison with primary Sjögren's syndrome and systemic sclerosis, PLoS One. 11 (2016). doi:10.1371/journal.pone.0159384.

[5] J. Li, N. Che, L. Xu, Q. Zhang, Q. Wang, W. Tan, M. Zhang, LC-MS-based serum metabolomics reveals a distinctive signature in patients with rheumatoid arthritis, Clin. Rheumatol. 37 (2018) 1493–1502. doi:10.1007/s10067-018-4021-6.

[6] G. Kageyama, J. Saegusa, Y. Irino, S. Tanaka, K. Tsuda, S. Takahashi, S. Sendo, A. Morinobu, Metabolomics analysis of saliva from patients with primary Sjögren's syndrome, Clin. Exp. Immunol. 182 (2015) 149–153. doi:10.1111/cei.12683.

[7] J.J. Mikkonen, Metabolic Profiling of Saliva in Patients with Primary Sj?gren?s syndrome, J. Postgenomics Drug Biomark. Dev. 03 (2013). doi:10.4172/2153-0769.1000128.

[8] A.A. Bengtsson, J. Trygg, D.M. Wuttge, G. Sturfelt, E. Theander, M. Donten, T. Moritz, C.J. Sennbro, F. Torell, C. Lood, I. Surowiec, S. R??nnar, T. Lundstedt, Metabolic profiling of systemic lupus erythematosus and comparison with primary Sjögren's syndrome and systemic sclerosis, PLoS One. 11 (2016) 1–15. doi:10.1371/journal.pone.0159384.

[9] A.H.M. Emwas, The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research, Methods Mol. Biol. 1277 (2015) 161–193. doi:10.1007/978-1-4939-2377-9_13.

[10] S. Bouatra, F. Aziat, R. Mandal, A.C. Guo, M.R. Wilson, C. Knox, T.C. Bjorndahl, R. Krishnamurthy, F. Saleem, P. Liu, Z.T. Dame, J. Poelzer, J. Huynh, F.S. Yallou, N. Psychogios, E. Dong, R. Bogumil, C. Roehring, D.S. Wishart, The Human Urine Metabolome, PLoS One. 8 (2013) e73076. doi:10.1371/journal.pone.0073076.

[11] A. Alonso, A. Julià, M. Vinaixa, E. Domènech, A. Fernández-Nebro, J.D. Cañete, C. Ferrándiz, J. Tornero, J.P. Gisbert, P. Nos, A.G. Casbas, L. Puig, I. González-Álvaro, J.A. Pinto-Tasende, R. Blanco, M.A. Rodríguez, A. Beltran, X. Correig, S. Marsal, Urine metabolome profiling of immune-mediated inflammatory diseases, BMC Med. 14 (2016) 133. doi:10.1186/s12916-016-0681-8.

[12] A.J. Chetwynd, A. Abdul-Sada, S.G. Holt, E.M. Hill, Use of a pre-analysis

osmolality normalisation method to correct for variable urine concentrations and for improved metabolomic analyses, J. Chromatogr. A. 1431 (2016) 103–110. doi:10.1016/j.chroma.2015.12.056.

[13] J. Xia, D.S. Wishart, Using MetaboAnalyst 3.0 for Comprehensive Metabolomics Data Analysis., Curr. Protoc. Bioinformatics. 55 (2016) 14.10.1-14.10.91. doi:10.1002/cpbi.11.

[14] J.A. Westerhuis, H.C.J. Hoefsloot, S. Smit, D.J. Vis, A.K. Smilde, E.J.J. Velzen, J.P.M. Duijnhoven, F.A. Dorsten, Assessment of PLSDA cross validation, Metabolomics. 4 (2008) 81–89. doi:10.1007/s11306-007-0099-6.

[15] I. Davis, A. Liu, What is the tryptophan kynurenine pathway and why is it important to neurotherapeutics?, Expert Rev. Neurother. 15 (2015) 719–721. doi:10.1586/14737175.2015.1049999.

[16] N.I. Maria, C.G. van Helden-Meeuwsen, Z. Brkic, S.M.J. Paulissen, E.C. Steenwijk, V.A. Dalm, P.L. van Daele, P. Martin van Hagen, F.G.M. Kroese, J.A.G. van Roon, A. Harkin, W.A. Dik, H.A. Drexhage, E. Lubberts, M.A. Versnel, Association of Increased Treg Cell Levels With Elevated Indoleamine 2,3-Dioxygenase Activity and an Imbalanced Kynurenine Pathway in Interferon-Positive Primary Sjögren’s Syndrome, Arthritis Rheumatol. 68 (2016) 1688–1699. doi:10.1002/art.39629.

[17] A. Agus, J. Planchais, H. Sokol, Gut Microbiota Regulation of Tryptophan Metabolism in Health and Disease, Cell Host Microbe. 23 (2018) 716–724. doi:10.1016/J.CHOM.2018.05.003.

[18] F. De Luca, Y. Shoenfeld, The microbiome in autoimmune diseases, Clin. Exp. Immunol. 195 (2019) 74–85. doi:10.1111/cei.13158.

[19] Y. Shikama, Y. Kudo, N. Ishimaru, M. Funaki, Potential Role of Free Fatty Acids in the Pathogenesis of Periodontitis and Primary Sjögren’s Syndrome, Int. J. Mol. Sci. 18 (2017) 836. doi:10.3390/ijms18040836.

[20] A.P. Simopoulos, Omega-3 fatty acids in inflammation and autoimmune diseases., J. Am. Coll. Nutr. 21 (2002) 495–505.

http://www.ncbi.nlm.nih.gov/pubmed/12480795 (accessed February 20, 2019).

[21] C.G. Costa, W.S. Guérand, E.A. Struys, U. Holwerda, H.J. ten Brink, I. Tavares de Almeida, M. Duran, C. Jakobs, Quantitative analysis of urinary acylglycines for the diagnosis of β-oxidation defects using GC-NCI-MS, J. Pharm. Biomed. Anal. 21 (2000) 1215–1224. doi:10.1016/S0731-7085(99)00235-6.

[22] J.H.S. Kabarowski, Y. Xu, O.N. Witte, Lysophosphatidylcholine as a ligand for immunoregulation, Biochem. Pharmacol. 64 (2002) 161–167. doi:10.1016/S0006-2952(02)01179-6.

[23] A.J. Lusis, Atherosclerosis., Nature. 407 (2000) 233–241. doi:10.1038/35025203.

[24] R. Wu, E. Svenungsson, I. Gunnarsson, B. Andersson, I. Lundberg, L. Schäfer Elinder, J. Frostegård, Antibodies against lysophosphatidylcholine and oxidized LDL in patients with SLE, Lupus. 8 (1999) 142–150. doi:10.1191/096120399678847434.

[25] T. Kühn, A. Floegel, D. Sookthai, T. Johnson, U. Rolle-Kampczyk, W. Otto, M. von Bergen, H. Boeing, R. Kaaks, Higher plasma levels of lysophosphatidylcholine 18:0 are related to a lower risk of common cancers in a prospective metabolomics study, BMC Med. 14 (2016) 13. doi:10.1186/s12916-016-0552-3.

[26] M.N. Barber, S. Risis, C. Yang, P.J. Meikle, M. Staples, M.A. Febbraio, C.R. Bruce, Plasma lysophosphatidylcholine levels are reduced in obesity and type 2 diabetes, PLoS One. 7 (2012) e41456. doi:10.1371/journal.pone.0041456.

[27] M. Ramos-Casals, P. Brito-Zerón, A. Sisó, A. Vargas, E. Ros, A. Bove, R. Belenguer, J. Plaza, J. Benavent, J. Font, High prevalence of serum metabolic alterations in primary Sjögren’s syndrome: Influence on clinical and immunological expression, J. Rheumatol. (2007).

[28] S. Manzi, M.C.M. Wasko, Inflammation-mediated rheumatic diseases and atherosclerosis, Ann. Rheum. Dis. (2000) 321–325. doi:10.1136/ard.59.5.321.

[29] Y. Gu, C. Lu, Q. Zha, H. Kong, X. Lu, A. Lu, G. Xu, Plasma metabonomics study of

rheumatoid arthritis and its Chinese medicine subtypes by using liquid chromatography and gas chromatography coupled with mass spectrometry, Mol. Biosyst. 8 (2012) 1535. doi:10.1039/c2mb25022e.

[30] X. Ding, J. Hu, C. Wen, Z. Ding, L. Yao, Y. Fan, Rapid resolution liquid chromatography coupled with quadrupole time-of-flight mass spectrometry-based metabolomics approach to study the effects of jieduquyuziyin prescription on systemic lupus erythematosus, PLoS One. 9 (2014) 1–11. doi:10.1371/journal.pone.0088223.

SUPPLEMENTARY MATERIAL

Table 1S. Elution gradients and injection volumes of chromatographic methods for plasma and urine samples.

	PLASMA			URINE		
	Time (min)	% A	% B	Time (min)	% A	% B
Elution Gradients	0	95	5	0	95	5
	5	90	10	30	70	30
	15	15	85	40	0	100
	32-40	0	100	50-60	95	5
	45-50	95	5			
Injection Volume (µL)	3.00			5.00		

Table 2S. MS/MS fragments, formula scores and error (ppm) of the annotated metabolites in plasma samples.

RT (min)	Mass (Da)	Molecular Formula	Metabolite	Score	Error (ppm)	MS/MS fragments
1.48	149.0484	$C_5H_{11}NO_2S$	L-Methionine	92.7	-5.00	56.0482/104.0512/132.0638
4.34	208.0857	$C_{10}H_{12}N_2O_3$	L-kynurenine	96.4	-1.89	74.0240/120.0449/146.0605/192.0661
9.05	184.1221	$C_9H_{16}N_2O_2$	N-*(3-aminopropyl)*pyrrolidin-2-one	86.5	0.46	126.0903/143.1153185.1261
21.13	519.3225	$C_{26}H_{50}NO_7P$	LysoPC(18:2)	91.7	-1.91	86.0956/104.1064/184.0717/483.2466/484.2496
21.18	543.3324	$C_{28}H_{50}NO_7P$	LysoPC(20:4)	88.2	-2.54	86.0955/104.1056/146.9805/184.0725/507.246
24.24	328.2354	$C_{22}H_{32}O_2$	Docosahexaenoic acid (DHA)	80.8	-8.26	66.9845/80.9998/91.0519/95.0156
24.52	302.2129	$C_{20}H_{30}O_2$	Eicosapentaenoic acid (EPA)	89.7	-2.63	67.0525/81.000/81.0686/95.0858/107.0849/207.1202
24.52	928.623	$C_{49}H_{95}O_{12}P$	PI(P20:0/20:0)	80.5	-5.41	Annotated by Formula
25.59	278.2184	$C_{18}H_{30}O_2$	Linolenic acid	88.6	-1.81	91.0518/95.0827/107.0834/123.1134/137.1296/159.1145
26.2	282.2518	$C_{18}H_{34}O_2$	Oleic acid	88.8	-5.86	55.0524/69.0683/67.0526/83.0834/265.2472/283.2586
26.2	934.6784	$C_{51}H_{99}O_{12}P$	PI(P20:0/22:O)	90.2	4.95	Annotated by Formula
30.84	672.5185	$C_{37}H_{73}N_2O_6P$	PE-Cer(d14:2/21:0)	86.6	4.10	55.0544/146.9819/512.4438/513.4469/636.4320

Table 3S. MS/MS fragments, formula scores and error (ppm) of the annotated metabolites in urine samples.

RT (min)	Mass (Da)	Molecular Formula	Metabolite	Score	Error (ppm)	MS/MS Fragments
0.95	113.0612	$C_4H_7N_3O$	Creatinine	97.2	-4.86	43.0287/44.0492/58.0651/86.0810
0.96	131.0702	$C_4H_9N_3O_2$	Creatine	98.9	-1.84	43.0288/44.0494/114.0666
1.32	268.1164	$C_{11}H_{16}N_4O_4$	Histidinyl-Hydroxyproline	98.8	0.51	72.0447/110.0718/111.0743/114.055/156.0764
1.51	138.0449	$C_6H_6N_2O_2$	Urocanic acid	94.5	-3.62	66.0353/68.0474/93.0446/121.0387
1.53	187.0960	$C_7H_{13}N_3O_3$	5-guanidino-3methyl-2-oxopentanoate	85.2	-1.45	41.0383/43.0287/55.0541/57.0335
1.62	167.0219	$C_7H_5NO_4$	Quinolinic acid	86.2	-0.46	84.043/86.0952/130.0472/132.1005
1.96	214.1316	$C_{10}H_{18}N_2O_3$	Valyl-Proline	86.8	0.26	43.0173/70.0647/72.0805/113.0708/114.0549/159.0772
2.86	228.1475	$C_{11}H_{20}N_2O_3$	Leucyl-Proline	94.0	-3.99	58.065/59.0723/60.0804/70.0651
3.37	261.1570	$C_{12}H_{23}NO_5$	Hydroxyisovaleroyl carnitine	98.3	-0.98	85.0286/262.1658
4.85	166.0488	$C_6H_6N_4O_2$	MethylXanthine	95.9	4.85	42.0334/67.0289/69.0446/124.0507
4.93	246.1211	$C_{10}H_{18}N_2O_5$	L-gamma-glutamyl-L-valine	98.2	-2.40	72.0800/84.0446/102.0550/118.0863/130.0498/184.0962/230.1028
4.96	165.0795	$C_9H_{11}NO_2$	L-phenylalanine	92.5	-2.76	42.0331/77.0382/103.0541/120.0804/124.0502
5.65	231.1468	$C_{11}H_{21}NO_4$	Butyrylcarnitine	89.2	-6.74	85.0290/232.1556
6.56	210.0639	$C_9H_{10}N_2O_4$	N-(5-amino)-2-hydroxybenoylglycine	84.2	-0.97	42.0337/53.0382/80.0492/108.0446/109.0528
7.05	297.1070	$C_{11}H_{15}N_5O_5$	Methylguanosine	97.5	0.35	109.0506/110.0350/135.0300/149.0461/166.0724/167.0750
8.43	285.0960	$C_{11}H_{15}N_3O_6$	N4-Acetlycytidine	70.3	7.69	112.0493/154.0591
8.60	157.0726	$C_7H_{11}NO_3$	Tiglylglycine	85.9	0.42	55.0534/69.0682
9.31	159.0894	$C_7H_{13}NO_3$	Isovaleroylglycine	82.0	-0.86	57.0693/85.0641/146.0824
10.55	187.0635	$C_{11}H_9NO_2$	Indoleacrylic acid	98.7	0.71	65.0379/91.0532/115.0529/142.0638
10.68	243.147	$C_{12}H_{21}NO_4$	Tiglylcarnitine	85.1	2.17	85.0280/185.0787
11.30	260.1377	$C_{11}H_{20}N_2O_5$	Gamma-glutamylisoleucine	86.2	-1.25	86.0959/130.0505/132.1019/244.117
16.17	189.0423	$C_{10}H_7NO_3$	Kynurenic acid	97.8	-1.72	89.0388/116.0494/144.0443
36.17	329.2200	$C_{17}H_{31}NO_5$	6-keto-decenoylcarnitine	84.7	-1.57	60.0798/52.0281/109.1005/127.1106
40.36	488.237	$C_{23}H_{32}N_6O_6$	Peptide	79.4	-6.64	Annotated by formula

Table 4S. Retention times, masses, and statistical results from unknown metabolites present in plasma samples.

RT (min)	Mass (Da)	p-value	Fold Change	VIP-value	AUC
1.47	103.0446	3.0 E-5	-1.33	1.47	0.742
1.47	132.0218	5.0 E-4	-1.25	1.24	0.712
11.64	947.5146	3.0 E-3	1.56	1.15	0.676
14.32	600.2563	9.3 E-3	-1.37	1.39	0.662
16.05	926.6471	7.0 E-5	2.82	1.40	0.790
21.61	1128.6258	5.5 E-3	-1.20	1.06	0.674
21.87	585.2987	1.7 E-4	1.42	1.34	0.728
23.16	322.1777	2.6 E-3	2.05	1.08	0.720
23.16	278.2150	1.5 E-3	1.96	1.14	0.731
24.35	326.2135	6.3 E-3	1.36	1.03	0.668
24.52	359.2042	2.9 E-3	1.38	1.07	0.694
25.96	562.4048	3.8 E-3	-1.40	1.04	0.628
26.59	468.3662	5.2 E-3	-1.35	1.20	0.627
26.61	468.3486	1.1 E-2	-1.30	1.10	0.618
27.29	596.4205	1.1 E-2	-1.39	1.23	0.618
27.91	616.4585	3.2 E-3	-1.47	1.34	0.634
28.36	396.2115	2.5 E-4	1.22	1.30	0.710
28.60	616.4648	6.3 E-4	-1.89	1.47	0.677
33.69	825.5238	2.2 E-2	1.20	1.10	0.642

Table 5S. Retention times, masses, and statistical results from unknown metabolites present in urine samples.

RT (min)	Mass (Da)	p-value	Fold Change	VIP-value	AUC
0.90	129.0427	2.2 E-3	-1.30	1.10	0.687
0.94	117.0540	1.8 E-3	-1.47	1.12	0.684
1.04	156.0542	4.1 E-3	-1.28	1.07	0.671
1.04	174.0640	2.9 E-3	-1.23	1.07	0.681
1.05	161.0449	7.5 E-3	-1.32	1.20	0.698
1.06	380.0710	2.3 E-2	1.97	1.83	0.594
1.06	364.0959	2.2 E-2	1.66	1.62	0.602
1.13	103.0632	4.8 E-3	-1.59	1.02	0.672
1.32	277.1163	1.1 E-4	-1.51	1.37	0.747
1.47	225.0741	4.7 E-2	2.14	1.43	0.651
2.19	172.0472	1.6 E-2	1.23	1.70	0.634
3.36	240.1466	9.7 E-3	-1.27	1.27	0.658
4.82	253.1537	4.0 E-3	-1.43	1.04	0.677
7.15	601.1855	3.1 E-2	-1.28	1.23	0.630

Table 6S. Members from the Precisesads Clinical Consortium.

Clinical center	Principal investigator	Clinicians

Hospital Universitario San Cecilio, Servicio Andaluz de Salud, Granada (Spain)	Norberto Ortego; Enrique Raya	**María Concepción Fernández Roldán;
		José Luis Callejas Rubio;
		Raquel Ríos Fernández;
		Inmaculada Jiménez Moleón;**

Hospital Universitario Reina Sofía Andaluz de Salud, Córdoba (Spain)	Eduardo Collantes	**Rafaela Ortega-Castro Mª Angeles Aguirre-Zamorano Alejandro Escudero- Contreras Mª Carmen Castro-Villegas**
Centre Hospitalier Universitaire de Brest, Hospital de la Cavale Blanche, Brest, (France)	Jacques-Olivier Pers	**Alain Saraux Valérie Devauchelle-Pensec Divi Cornec Sandrine Jousse-Joulin**
Fondazione IRCCS Ca Granda Ospedale Maggiore Policlinico, Milano (Italy)	Lorenzo Beretta	
Deutsches Rheuma-Forschungszentrum Berlin, (Germany)	Falk Hiepe	
Hospitaux Universitaires de Genève (Switzerland)	Carlo Chizzolini	
Katholieke Universiteit Leuven (Belgium)	Rik Lories	**Ellen De Langhe**
Université catholique de Louvain (Belgium)	Bernard Lauwerys	
Hospital Clinic I Provicia, Institut d'Investigacions Biomèdiques August Pi i Sunyer, Barcelona (Spain)	Ricard Cervera	**Ignasi Rodríguez-Pintó, Gerard Espinosa**
Andalusian Public Health System Biobank	Rocío Aguilar Quesada	**Rocío Aguilar Quesada**
Project Office - Recruitment and data follow up	**Jacqueline Marovac & Tania Gomes Anjos**	

Figure 1S. PCA scores plot from plasma data before normalization (1Sa) and after the between-batches normalization step (1Sb). (Batch1, red plots; Batch2, green plots; Batch3, blue plots)

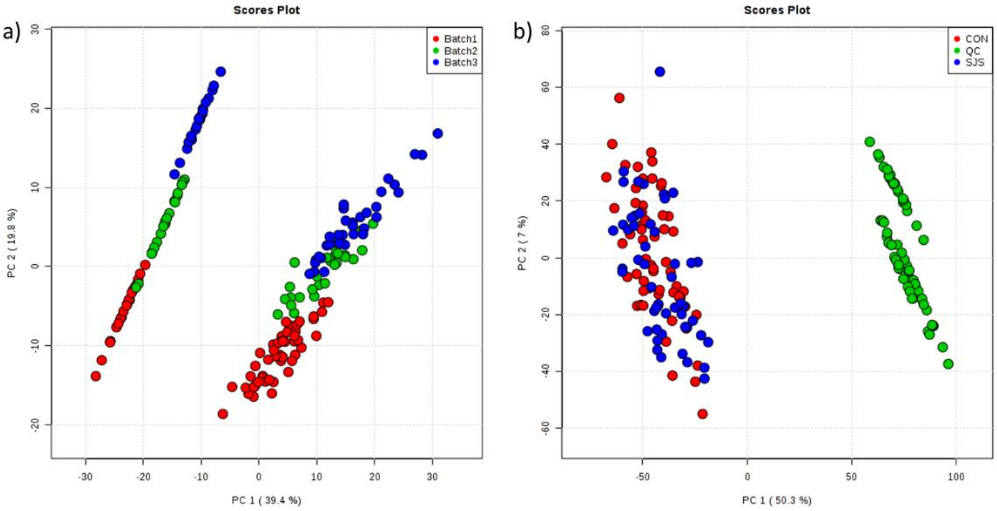
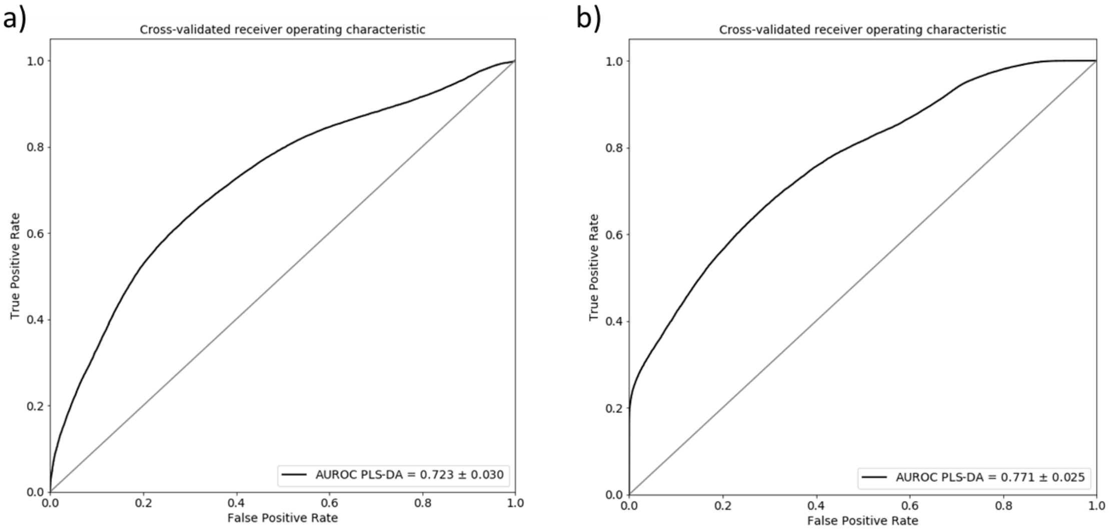
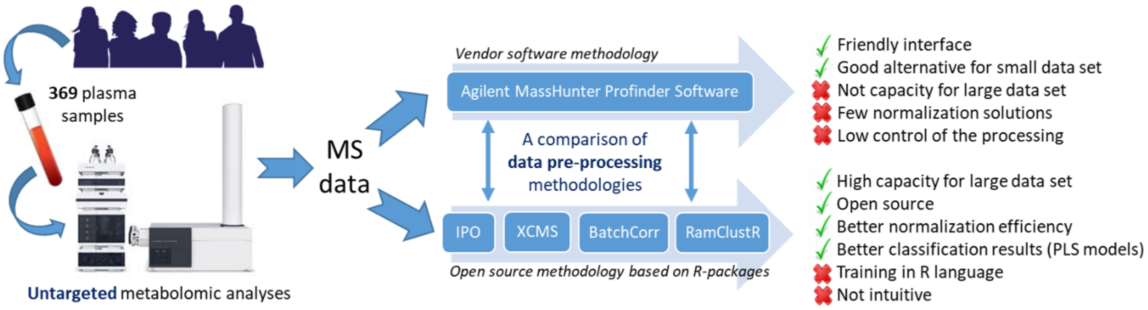


Figure 2S. ROC Curves for PLS-DA models from plasma data (2Sa) and urine data (2Sb)



Capítulo 6

Análisis comparativo de las herramientas de pre-procesamiento de datos metabólomicos basadas en software comerciales y programas de acceso libre


369 plasma samples
Untargeted metabolomic analyses
MS data
Vendor software methodology
Agilent MassHunter Profinder Software
A comparison of data pre-processing methodologies
IPO
XCMS
BatchCorr
RamClustR
Open source methodology based on R-packages
✓ Friendly interface
✓ Good alternative for small data set
✗ Not capacity for large data set
✗ Few normalization solutions
✗ Low control of the processing
✓ High capacity for large data set
✓ Open source
✓ Better normalization efficiency
✓ Better classification results (PLS models)
✗ Training in R language
✗ Not intuitive

Álvaro Fernández-Ochoa, Rosa Quirantes-Piné, Isabel Borrás-Linares, PRECISESADS Clinical Consortium, Marta E. Alarcón Riquelme, Carl Brunius, Antonio Segura-Carretero.

Costs and benefits of switching from vendor-based to open source pipelines for untargeted LC-MS metabolomics

ABSTRACT

Data pre-processing of the LC-MS data is a critical step in untargeted metabolomics studies in order to achieve correct biological interpretations related to the question of the study. Several tools have been developed for pre-processing, and these can be classified into either commercial or open source software. This study aims to compare two of these different methodologies (vendor vs open-source) for untargeted metabolomic with a large number of samples. Specifically, 369 plasma samples (306 case samples and 63 Quality Control) were analyzed by HPLC-ESI-QTOF-MS. The collected data were pre-processed by both methodologies and later evaluated by several parameters (number of peaks, degree of missingness, quality of the peaks, degree of misalignments, and robustness in multivariate models). The vendor software was characterized by ease of use, friendly interface and good quality of the graphs. The open source methodology could more effectively correct the drifts due to between and within batch effects. In addition, the evaluated statistical methods achieved better classification results with higher parsimony for the open source methodology, indicating higher data quality. However, the open source methodology also required a higher degree of computational experience. Although both methodologies have strengths and weaknesses, the open source methodology seems to be more appropriate for studies with a large number of samples due mainly to its higher

capacity and versatility that allows combining different packages, functions and methods in a single environment.

Keywords. Metabolomics; Data pre-processing; Mass spectrometry; Liquid chromatography; R packages; Vendor software

1. Introduction.

Metabolomics is defined as the complete characterization of low molecular weight molecules (metabolites) present in a biological system, such as cells, tissues, biofluids or organisms [1]. Untargeted metabolomics is frequently used to compare metabolic profiles between subjects to identify differences associated with the underlying study question (e.g. disease, diet, etc.). [2]

Untargeted metabolomics studies are carried out through a series of the following steps: (i) study design and sample recruitment, (ii) sample preparation, (iii) instrumental analysis, (iv) data pre-processing and statistical analysis, (v) compound identification and (vi) biological interpretation [3,4]. It is necessary to carry these steps out thoroughly with high precision and accuracy to maintain data quality throughout the pipeline to be able to correctly interpret the results and address the underlying biological question of the study [5]. The analytical techniques most frequently used in this type of studies are proton Nuclear Magnetic Resonance Spectroscopy (^{1}H-NMR) and Mass Spectrometry (MS). The main advantages of NMR are the high reproducibility/repeatability and accurate quantification as well as capacity of structure elucidation. MS, on the other hand, is able to detect a much higher number of metabolites due to its higher sensitivity [6]. In addition, MS is usually coupled to various separation techniques at the front end, such as Liquid Chromatography (LC-MS). This hyphenation is able to separate the analytes prior to MS detection in order to achieve better MS performance. In metabolomics, LC-MS is the most employed analytical technique [7]. Among the different steps involved in untargeted metabolomic workflows using LC-MS, this work is primarily focused on data pre-processing.

Data pre-processing of the LC-MS data is a critical step whose purpose is to reduce the complexity of the raw data, extract the main features and transform them in order to subsequently perform adequate statistical tests [8]. This process involves a series of steps, such as baseline correction, noise filtering, peak detection, peak alignment, normalization, missing data imputation and annotation [8–10]. Different vendor and open source software have been developed to perform these functions [11,12]. In this sense, the main commercial platforms at present correspond to the major instrument vendors: Mass Profinder/Profiler from Agilent Technologies [13,14], Progenesis QI from Waters Corporation, Compound Discover from Thermo Scientific, MetaboScape from Brucker and SIEVE from Thermo Scientific. On the other hand, open source software has gained in popularity in recent years [7]. Some highly popular softwares are MZmine [15], Workflow4Metabolomics [16], MetAlign [17], OpenMS [18], and XCMS [19]. Several, if not most, software modules are based on the programming language R [7], with a recent survey showing that the most used tool to pre-process LC-MS data is XCMS [20].

Ideally, the perfect platform to data processing in metabolomics should be intuitive with a user-friendly interface, open-source and offer a comprehensive coverage of all steps (or at least with easy integration to other steps) of the pipeline [7]. While commercial software stands out for being intuitive and user-friendly, open source solutions are free to use and provide more versatility to the needs of the users. However, in general they are also less intuitive and have steeper learning curves [7]. Moreover, it is also quite common that different tools show great effectiveness in some of the data processing steps but not in others. Users therefore have to stitch together different tools to carry out the entire pre-processing pipeline which often demands more advanced bioinformatics and/or programming skills [8]. As there are a lot of tools for metabolomics data processing, there is a need to compare these methodologies and put into light the pros or cons of them. [7,21].

In the present work, two methodologies, vendor vs open-source, for data pre-processing in untargeted metabolomics studies were compared. We highlight differences in these two pipelines using 369 plasma samples analyzed by LC-MS, from a dataset aimed to investigate the metabolism of Systemic Autoimmune Diseases within the PRECISESADS project (http://www.precisesads.eu/). Specifically, we have processed the data using Agilent software, representing a vendor pipeline. On the other hand, an open source methodology based on R was represented by the combination of several packages, namely IPO [22], XCMS [19], batchCorr [23] and RAMClustR [24]. Our aim was to provide insights into benefits and disadvantages of using these two methodologies, thereby aiding metabolomics researchers in their choice of data pre-processing strategies. Although tutorials may exist for individual pre-processing modules, tutorials on how to stitch together modules into entire pipelines are lacking. In this way, a detailed tutorial on how to start using an R-based methodology is provided while offers users with outlines from which to build their own custom pre-processing pipelines.

2. Materials and methods.

2.1 Dataset.

Metabolomic data were obtained from samples of the PRECISESADS project (www.precisesads.eu). The aim of this project is to find clinically useful biomarkers in order to obtain a new reclassification of 7 systemic autoimmune diseases (systemic lupus erythematous, rheumatoid arthritis, systemic sclerosis, mixed connective tissue disease, antiphospholipid syndrome, Sjögren's syndrome and undifferentiated connective tissue disease). This metabolomic analysis is ancillary to the written informed consent obtained from each participant of the study, which was registered on clinicaltrials.gov with the code NCT02890121.

Plasma samples from 247 patients with the above diseases and 59 healthy volunteers were analyzed. Subjects were recruited from different study centers across Europe. Biological samples were obtained and stored at -80 ºC until analysis. A Quality Control (QC) sample was obtained by mixing 20 µl of each study sample including both controls and case samples. After thawing on ice, a protein precipitation step was carried out with a mix of methanol-ethanol (1:1; v/v). Samples were analyzed using an Agilent 1260 HPLC instrument coupled to an Agilent 6540 Ultra High Definition (UHD) Accurate Mass Q-TOF equipped with a Jet Stream dual ESI interface. Metabolites were separated using a reversed-phase C18 analytical column (Agilent Zorbax Eclipse Plus, 3.5 µm, 2.1×150 mm) and detected in positive-ion mode over a range from 50 to 1700 m/z. The analytical methodology is described in detail elsewhere [25].

The QC sample was injected five times at the beginning of each sequence in order to stabilize the equipment and also continuously throughout the analytical sequence (each five study samples) to monitor system performance and perform feature intensity drift correction. Due to the large number of samples, instrumental analysis was performed in three batches. In addition, MS/MS analysis of the QC sample was performed in order to obtain a representative fragmentation pattern of the main metabolites present in the majority of the samples. This analysis was carried out using nitrogen as the collision gas with 10 eV, 20 eV and 40 eV as collision energies.

2.2. Data Pre-processing.

Fig. 1 shows schematically a summary of both methodologies carried out for pre-processing of the data. Both methodologies are described in detail in the following subsections.

2.2.1. Vendor software approach.

Data was processed using *Agilent MassHunter Profinder B.06.00* software using Automatic peak finding by the two-step method. This software was installed on a Windows 7 computer with 3.20 GHz Intel Core i7 and 32 Gb of RAM memory.

First, a batch recursive feature extraction was performed using data from QC samples as a representative sample in which all endogenous metabolites should be present. Due to the large number of sample files and their size, molecular feature extraction of the QC files was performed in the first place. Second, the molecular features found in the QC samples were then used to guide feature selection in the case and control study samples. In this step, peaks with intensity lower than 1000 counts were filtered out. Isotopes and adducts were grouped into a molecular feature with a maximum charge of 2. Feature alignment was performed with 20 ppm ± 2 mDa mass and 0.25 minutes retention time windows. Peaks were manually inspected and corrected before integration. Both Percentile Shift and Quantile normalization methods were tested using Mass Profiler Professional software (Agilent Technologies). Due to the large between-batch and within-batch effects, the data were normalized using two methods consecutively. Firstly, the Bayes method from MetaboAnalyst 4.0 [26–28], and secondly, the Mass Total Useful Signal (MSTUS) method. Finally, the molecular features with high variability in QC samples (relative standard deviation, RSD, higher than 30%) were removed.

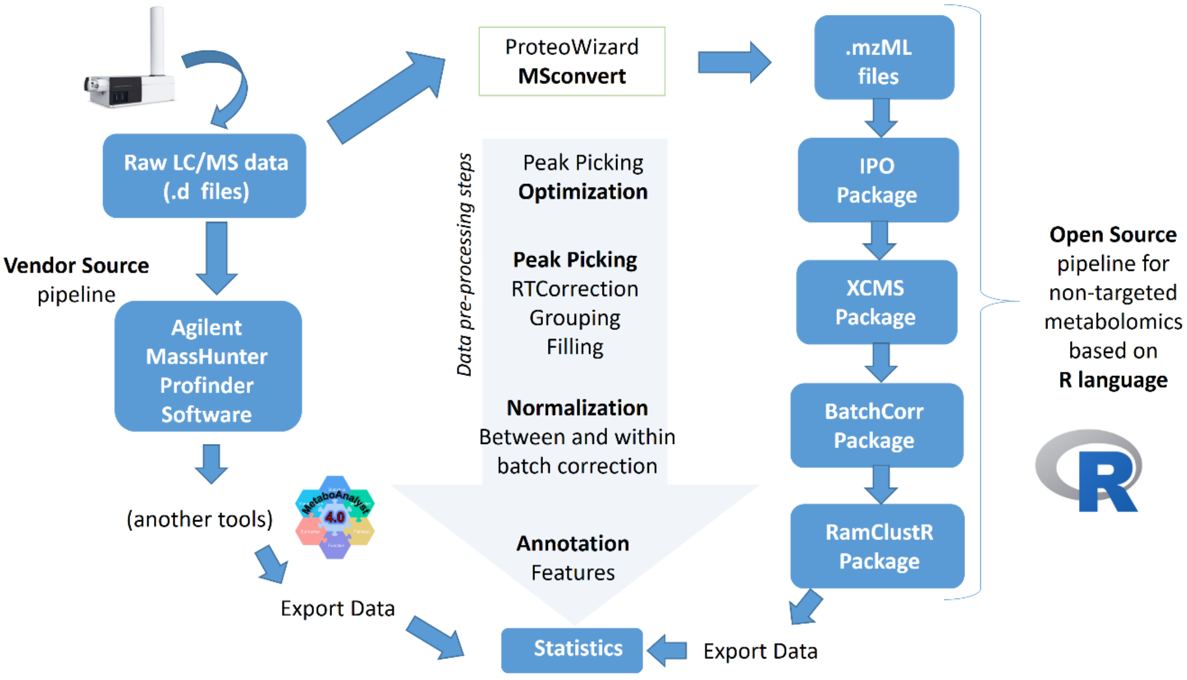


Fig. 1. A summary scheme of both methodologies (Vendor Software vs Open Source) used in the comparative study.

2.2.2. R-based approach.

First, Agilent .d files were converted to .mzML file format using the *MSConvertGUI* software [29] to be able to import them into the R open source environment (version 3.5.1) [30]. The R scripts, packages and commands were applied in RStudio environment (version 1.1.456) [31] to facilitate use and visualization of the results.

The *XCMS* package was used for peak picking retention time alignment, grouping and filling of missing features [19]. XCMS parameters were optimized using a combination of the *IPO* package [22] and manual optimization. For IPO optimization, 6 QC files spanning the multi-batch injections sequence were selected. The final optimized parameters for peak picking using the "centwave" method were the following: peakwidth = c(12.45, 35), mzdiff = 0.00175, prefilter = c(3, 1000). Retention time adjustment was performed with the "obiwarp" method using the following optimized values: profStep = 0.3, response = 13.84, gapInit = 0.352, gapExtend = 2.436. Finally, feature correspondence was achieved with the "density" method using the following optimized parameters: bw = 5.0 and mzwid = 0.047.

Imputation of values still missing after XCMS peak filling was performed using an in-house script based on RandomForest (https://gitlab.com/CarlBrunius/StatTools; mvImpWrap() function). The obtained data was corrected for within- and between-batch intensity drift using the *batchCorr* package [23]. Moreover, the features with high variability after normalization (RSD > 30 %) were filtered out.

Finally, grouping of features (isotopes, adducts and fragments) corresponding to the same metabolites was achieved using the RAMClustR package [24]. RAMClust grouping is based on similarity between features in retention time and intensity correlations between samples. The similarity parameters (σ_t, σ_r) were optimized using an in-house procedure and were set at values of 1.33 and 0.3, respectively.

All R scripts used in this work are available in full detail with comments as a tutorial in the Supplementary Material.

2.3. Statistics and metabolite annotation

In order to compare both methodologies, different statistical tests were performed. The Pearson correlation test was used to study the similarity of metabolite features obtained with both methodologies. Moreover, a subset of samples (53 healthy control and 45 patients with systemic sclerosis) was chosen for multivariate data analysis. PLS-DA models were performed using MetaboAnalyst 4.0 [32] and the R MUVR package [33]. Permutation tests were performed in both models for validation [34] .

To provide biological meaning to the results, metabolites of interest were annotated according to Metabolomics Standard Initiative (MSI) guidelines [35]. Annotation was performed by comparing MS and MS/MS spectra with information from metabolomics databases (LipidMaps, KEGG Human Metabolome Database and METLIN) as well as MS/MS fragmentation resources such as MetFrag and Sirius [36,37].

3. Results and discussion

Starting from the idea that data pre-processing is a critical step to decrease the risk of chance findings and misinterpretation and achieve correct biological interpretations, we have compared two pre-processing pipelines, representing vendor and open source software, respectively.

Due to the large number of samples, the methodology based on vendor software was not able to perform the data processing of all samples in a single step due to the capacity of the computer. Consequently, batch recursive feature extraction had to be performed separately for five different subsets of the entire data set. In contrast, the open source methodology allowed pre-processing of all samples at once, and depending on the number of available computer cores, the pre-processing would be more or less fast.

In the next sections, we present and discuss several results obtained from the two methodologies, i.e. number of peaks, degree of missingness, quality of the peaks, degree of misalignments, and robustness in multivariate models.

3.1. Peak Peaking

After grouping of features likely arising from the same metabolite (merging of isotopes, adducts and fragments, both methodologies obtained similar number of peaks (Vendor methodology: 548, R-based methodology: 531) and degrees of missing data (Vendor methodology: 8.91 %; R-based methodology: 9.59 %). The peaks were cross-checked by retention time (RT) and m/z, and, in total, 445 were picked by both methodologies.

RT drift was well aligned by XCMS, which modifies the RT for the samples to achieve superpositioning of the chromatograms (**Fig. 1S**, Supplementary Material). In contrast, vendor methodology does not modify the RT of the chromatograms, but instead, this tries to find the features in the samples within a RT range. With the high number of injections, RT drifts were pronounced, resulting in poor peak matching for several features. Those failures need to be corrected by the operator one by one, being a very time-consuming step. An example of this type of failure is shown in **Fig. 2S**

(Supplementary Material), where it can be observed the comparison of the result obtained by XCMS. Since manual supervision and correction of the results is highly time consuming, an advantage of the commercial software is the ease of visualizing the molecular features. Manual inspections and corrections, are, however, much more tedious in the R-based approach. XCMS integrations was therefore indirectly assessed by Pearson correlation with peak areas obtained from the Agilent workflow after manual inspection and correction, (**Fig. 3S, Supplementary Material**), which showed overall very high accordance. Interestingly, peak area correlations decreased somewhat when comparing XCMS peaks to those obtained from vendor software prior to manual correction (**Fig. 4S**, Supplementary Material), suggesting better results obtained using R-based pipeline in terms of alignment and integration. We hypothesize that this could be highly related to the greater number of parameters that can be modified and optimized. In contrast, the vendor software does not allow adjusting so many parameters and there is no automatic optimization process.

3.2. Normalization results

The metabolomic data from the three batches was collected in different months and each batch lasted for about a week. These facts produced large between-batch and within-batch effects. The magnitude of these drifts was detected by the distribution of the QC samples in the PCA score plots [4] from raw data obtained with both vendor software (**Fig. 2a**) and open source (**Fig. 3a**) methodologies. These effects are quite common in large-scale LC-MS studies due to different reasons such as matrix effects, variations in chromatographic conditions, loss of mass ionization efficiency or variability in MS sensibility [38]. Consequently, normalization is one of the most critical step in any pre-processing pipelines, to ensure that the data is comparable, without losing valuable biological information [39]. The number of normalization methods in vendor software is in general limited. Specifically, in Agilent Mass Profiler, the offered methods are by internal standards, quantile and percentile shift. Normalization by internal standards is widely considered not fit for purpose in untargeted metabolomics [40]. The others techniques did not provided satisfactory normalization, in that study samples were visibly separated by batch (**Fig. 2**). These normalization methods are based on the signal intensity distributions [41] and do not consider possible feature drift patterns [23]. In order to improve the obtained results by vendor software, data was also normalized by the open access platform MetaboAnalyst 4.0, which showed improved efficacy (**Fig. 5S**, Supplementary Material). Furthermore, MetaboAnalyst has the advantages that it is both free to use and has a friendly, intuitive web-based interface (https://www.metaboanalyst.ca/). However, it is also important to note that this tool offered is mainly oriented to statistical analysis and not pre-processing.

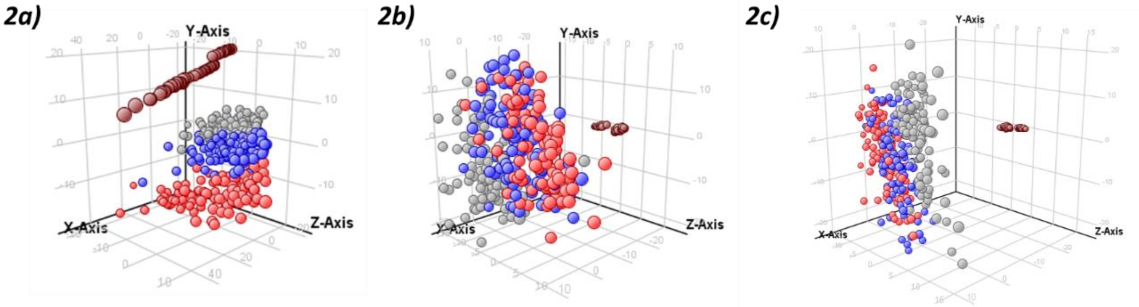


Fig. 2. PCA scores plot from data obtained by methodology based on vendor software. 1a) Raw Data; 1b) data normalized by the quantile method (Mass Profiler Professional, MPP); 1c) data normalized by the percentile shift (75.0) method (MPP). Batch 1 in red, batch 2 in blue, batch 3 in grey and QCs in brown.

Unlike commercial software, there are several open source programs based on R to carry out the normalization step in large untargeted LC-MS metabolomics studies, such as MetNormalizer [40], BatchCorr [23], MixNorm [42], Normalyzer [43] or NormalizeMets [44], among others. Most of them are based on QC samples taking into consideration the possible feature drift patterns. In this way, the open source package (bathCorr) applied on our data showed good results getting a well-behaved grouping of the QC samples and allowed to correct in a higher degree the batch effects (**Fig. 3**). The main advantage of batchCorr is that it takes into account different possible drift trends along the sequence [23] and examples of some of these different patterns are shown in **Fig. 6S** (Supplementary Material). Therefore, different correction functions are used depending on the detected drifts. However, as an example of the less thought through user experience in most R packages, the native PCA plots provided by the batchCorr package were very rudimentary (**Fig. 7S**, Supplementary Material). The data were therefore imported in MetaboAnalyst to obtain more visually pleasing figures (**Fig. 3**).

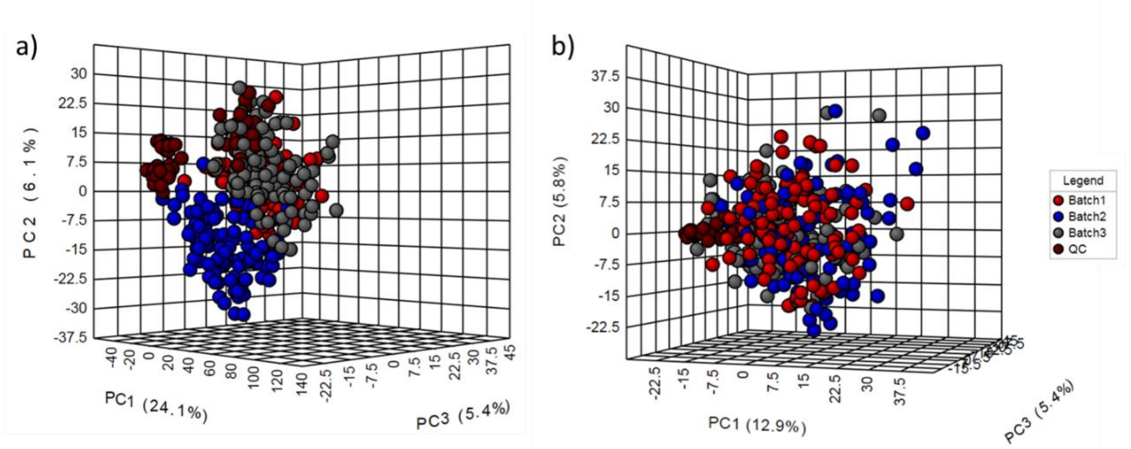


Fig. 3. PCA scores plot from data obtained by open source methodology. 1a) Raw Data; 1b) data normalized by the batchCorr package (R environment); Batch 1 in red, batch 2 in blue, batch 3 in grey and QCs in brown.

3.3. Multivariate models

A subset of samples (systemic sclerosis patients and healthy controls) was selected for multivariate modelling. The same statistical tests were performed for the data using both methodologies. First, PLS-DA models performed by MetaboAnalyst 4.0 showed slightly higher classification accuracy and predictive power using data obtained from the R pipeline (**Fig. 4**). More details information on the top-ranked metabolite features (**Fig. 4g-h**) are available in the Supplementary Material (**Tables 1S** and **2S**). Six metabolites (L-kynurenine, PS(18:0), Pipecolic acid, Theophylline and two unknowns) were found among the 15 most important in both PLS-DA models.

Table 1. Main results (number of variables (nVar), classification rate (class, %), Area Under the Curve (AUC), number of components (nComp) and p-value of permutations test) obtained for the PLS models using MUVR package.

PLS-MUVR models	nVar	class (%)	AUC	nComp	p-value (Permutation)
R Data	15	86.8	0.931	2	1.38 E-6
Vendor software Data	67	81.7	0.893	3	2.80 E-5

PLS models were also performed in R using the MUVR package, which employs a more prudent cross-validation scheme (repeated double cross-validation) and also performs unbiased variable selection [33]. Analogously to the PLS analyses performed using MetaboAnalyst, slightly better classifying results were found with the data obtained in R (**Table 1**). Overall, better modelling results were obtained for the R data, including parsimony, represented by a lower number of selected variables. Misclassifications and the confusion matrices are shown in **Fig. 5**, and complete lists with annotated metabolites are provided in the Supplemental **Tables 3S** and **4S**. The higher number of components and variables in the model with data from vendor software may make the biological interpretation of the results more difficult [45]. In addition, the ideal model would be the one that achieves better classifying results with a smaller number of variables. Therefore, the better results obtained with R data indicate higher data quality compared to the commercial pre-processing pipeline.

In view of the annotated metabolites, L-kynurenine and phosphatidylserine (PS) were found among the most significant variables in all four multivariate models. The kynurenine pathway has shown a large impact in recent years due to its relation with the immune system, inflammation and neurological processes (Davis and Liu 2015).

Furthermore, the dysregulation of the kynurenine pathway are in agreement with results from other autoimmune diseases such as systemic lupus erythematosus (SLE) [46].

Other differential metabolites in the majority of the models were acylcarnitines, unsaturated fatty acids (UFAs), and phospholipids. The dyregulation of these metabolites, mainly acylcarnitines and UFAs, are in line with previous work on a smaller number of volunteers [25], which gives consistency to the data obtained by both methodologies in this subset of samples.

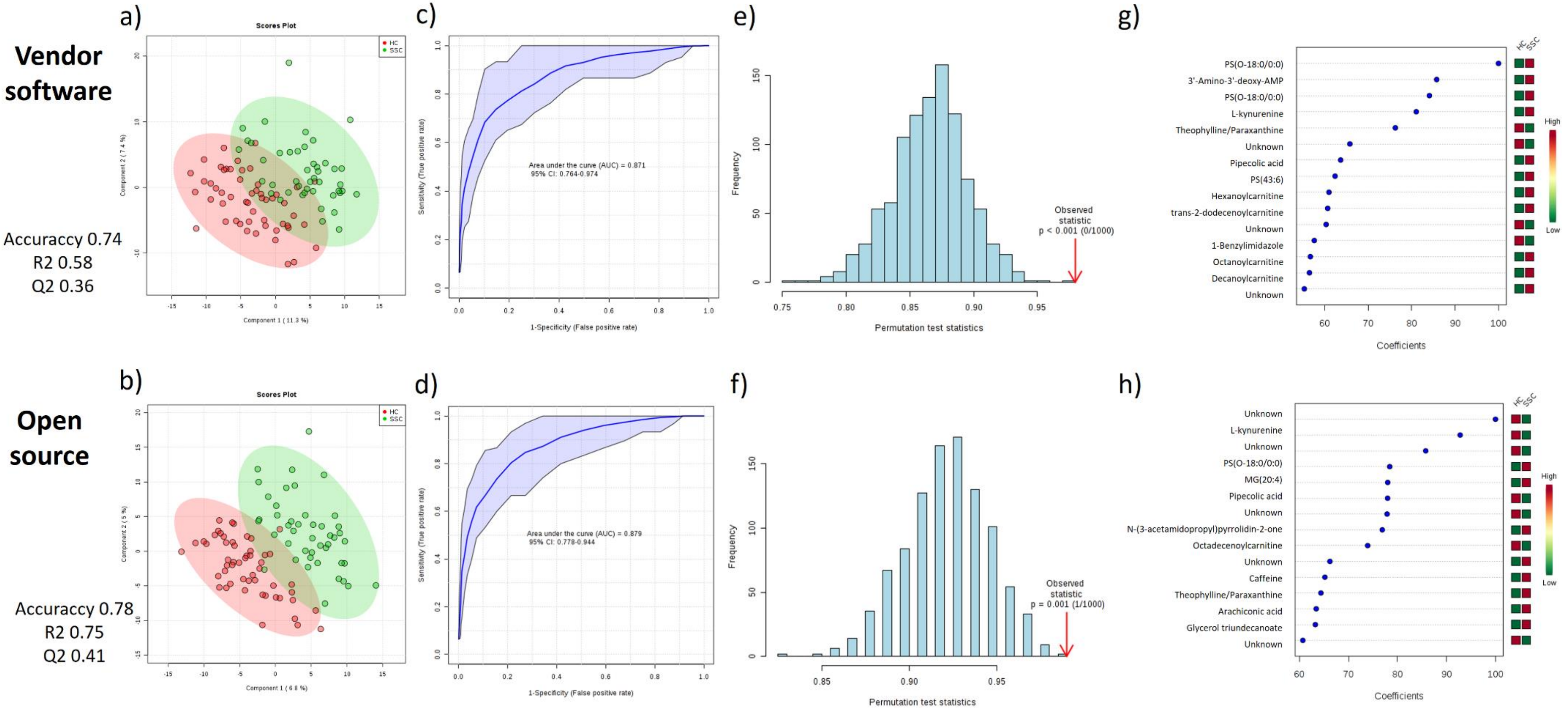
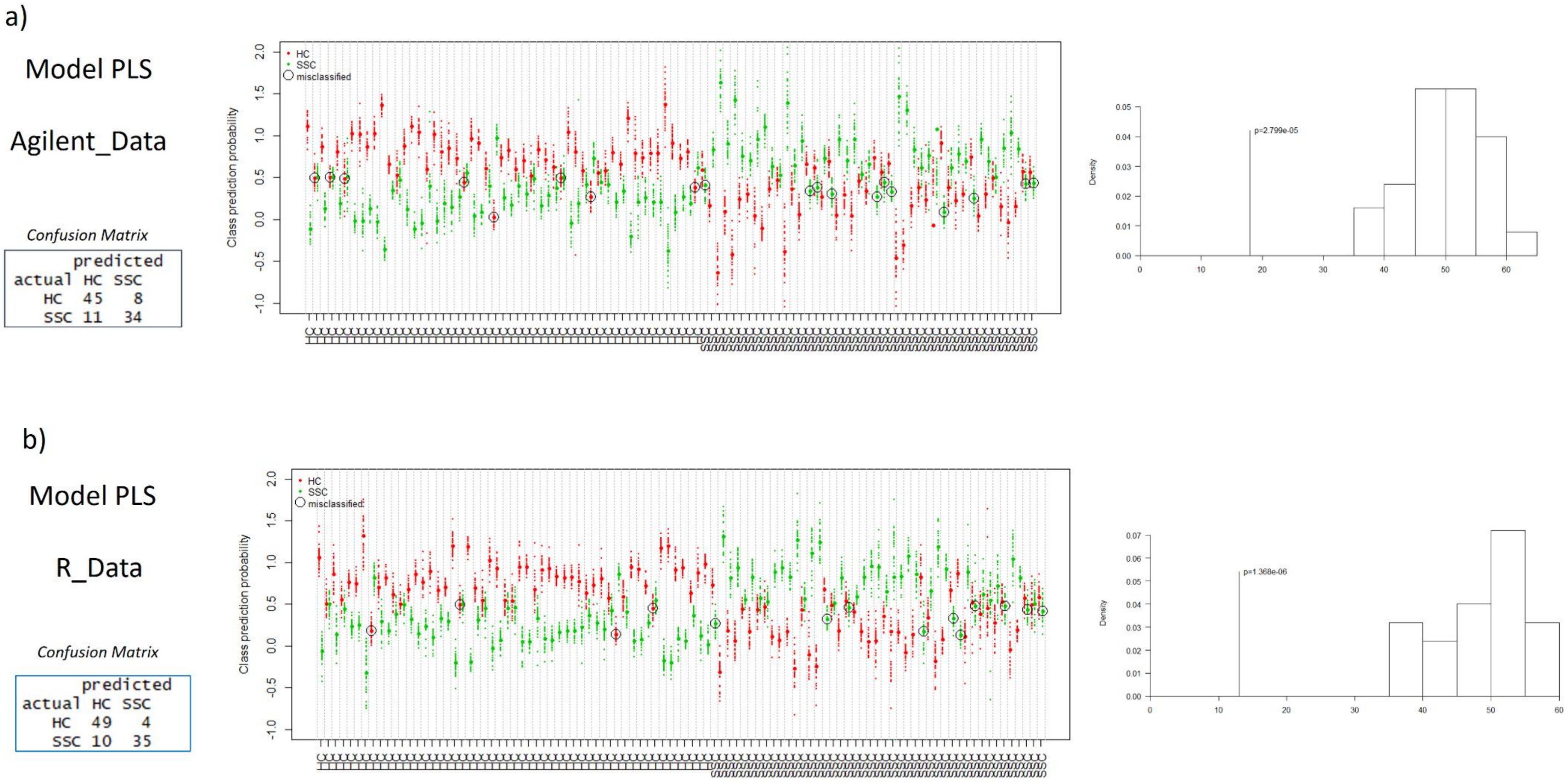


Fig. 4. A supervised Partial Least Squares Discriminant Analysis (PLS-DA) performed in MetaboAnalyst 4.0 software. PLS-DA scores plot (4a,Vendor Software Data (V), 4b, open source data (O)), ROC curve for PLS-DA model validation (4c, V; 4d, O), Permutation test result (4e, V; 4f, O), 15 most significant features (4g, V; 4h, O).



```
        predicted
actual  HC  SSC
   HC   45    8
   SSC  11   34
```

```
        predicted
actual  HC  SSC
   HC   49    4
   SSC  10   35
```

Fig. 5. Confusion matrices, permutation test results, and predictive classification of individuals according to PLS results obtained using MUVR package. a) vendor software data, b) open source data.

3.4. Global comparison of both methodologies

Based on the results obtained in the previous sections, the main advantages and disadvantages of each methodology are highlighted in **Table 2**. The vendor software methodology is characterised by its ease of use, a high level of dedicated support and good integration with annotation modules. In view of the results obtained, commercial software seems to be appropriate for studies of metabolomics with a smaller number of samples, where there is little drift in m/z, RT or signal intensity over time. However, for metabolomics studies with larger number of samples, as in the case of the example shown, commercial software have limitations mainly in capacity to process a high number of samples as well as in correcting for signal drifts. In addition, the high occurrence of incorrect peak integration requires extensive efforts by researchers for manual correction. Fortunately, these disadvantages can be addressed using open source methodology, e.g. in R, although this environment is not as user friendly or intuitive as commercial software. Furthermore, if the user has never worked with R-based methodologies, the initial learning curve is very steep. To compensate for this difficulty and to aid R beginners in setting up a data pre-preprocessing and analytical pipeline, a tutorial is provided in the online supplementary material.

Candidate biomarkers discovery should ideally be independent on the methodology used for data processing [11]. However, we have shown differences in the selection of candidate metabolites obtained by the two different methodologies in the presented example related to Systemic Sclerosis. In fact, multivariate models had higher classification rate and were more parsimonious using data obtained by the open source R methodology. These observed differences are likely related to the quality of the data used to create such models. In view of the results, the differences in data quality can be highly influenced predominantly by the normalization step, which has been showed as the main weakness of the vendor software methodology.

Table 2. Main advantages and disadvantages of the use of vendor software and R packages for pre-processing of metabolomics data obtained by HPLC-ESI-QTOF-MS.

Vendor Software methodology		***R-based methodology***	
✓	✗	✓	✗
Easy to use, User- friendly interface	Licence fee	Open source	Steep learning curve
High quality of the plots	Limited capacity to process a high number of samples	Greater number of packages, functions and methods (e.g. normalization)	Low plot quality
No need to transform the format of the data	Few normalization techniques. Difficulties to normalize large between-batch effects	High capacity for faster processing of a high number of samples	Data format transformation
Easy to inspect features, integration results and MS spectra. Easy to predict molecular formula.	Errors in peak integration	Possibility of carrying out all the steps of pre-processing and statistical analysis in the same environment	More cumbersome to show integration results, MS spectra, and to predict molecular formula.
Easy to manually correct areas	Low control of the processing (only some parameters can be modified)	Flexibility and versatility	Some level of coding skills is required

4. Conclusions

Both vendor and open source methodologies have strength and weaknesses. However, we have shown that the open source methodology is the most suitable option for metabolomic studies with larger number of samples in multiple batches. First, this methodology is to a much higher degre able to correct the large between- and within-batch effects. In addition, it stands out for being free and open source, having a greater capacity and versatility to use a large number of packages, functions and methods in a single environment. Nevertheless, this environment is also less intuitive, frequently with lower quality graphical output and with a distinctly steeper learning curve. We provide a detailed tutorial to help users of commercial software to start processing data through R-based methodology.

Acknowledgements

The work described has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement No. 115565, the resources for which are composed of a financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and the European Federation of Pharmaceutical Industries and Associations (EFPIA) companies' in-kind contribution. The author AFO received support from the Spanish Ministry of Education, Culture and Sports (FPU grant 14/03992). We thank all the members of PRECISESADS Clinical Consortium (Table 5S) for their effort in the sample recruitment, distribution and assessment of the samples used in this study.

Bibliography

[1] A. Agin, D. Heintz, E. Ruhland, J.M. Chao de la Barca, J. Zumsteg, V. Moal, A.S. Gauchez, I.J. Namer, Metabolomics - an overview. From basic principles to potential biomarkers (part 1), Med. Nucl. 40 (2016) 4–10. doi:10.1016/j.mednuc.2015.12.006.

[2] E. Parfieniuk, M. Zbucka-Kretowska, M. Ciborowski, A. Kretowski, C. Barbas, Untargeted metabolomics: an overview of its usefulness and future potential in prenatal diagnosis, Expert Rev. Proteomics. 15 (2018) 809–816. doi:10.1080/14789450.2018.1526678.

[3] A. Alonso, S. Marsal, A. Julià, Analytical methods in untargeted metabolomics: state of the art in 2015., Front. Bioeng. Biotechnol. 3 (2015) 23. doi:10.3389/fbioe.2015.00023.

[4] M.M. Ulaszewska, C.H. Weinert, A. Trimigno, R. Portmann, C. Andres Lacueva, R. Badertscher, L. Brennan, C. Brunius, A. Bub, F. Capozzi, M. Cialiè Rosso, C.E. Cordero, H. Daniel, S. Durand, B. Egert, P.G. Ferrario, E.J.M. Feskens, P. Franceschi, M. Garcia-Aloy, F. Giacomoni, P. Giesbertz, R. González-Domínguez, K. Hanhineva, L.Y. Hemeryck, J. Kopka, S.E. Kulling, R. Llorach, C. Manach, F. Mattivi, C. Migné, L.H. Münger, B. Ott, G. Picone, G. Pimentel, E. Pujos-Guillot, S. Riccadonna, M.J. Rist, C. Rombouts, J. Rubert, T. Skurk, P.S.C. Sri Harsha, L. Van Meulebroek, L. Vanhaecke, R. Vázquez-Fresno, D. Wishart, G. Vergères, Nutrimetabolomics: An Integrative Action for Metabolomic Analyses in Human Nutritional Studies, Mol. Nutr. Food Res. 63 (2019) 1800384. doi:10.1002/mnfr.201800384.

[5] Y. Wu, L. Li, Sample normalization methods in quantitative metabolomics, J. Chromatogr. A. 1430 (2016) 80–95. doi:10.1016/J.CHROMA.2015.12.007.

[6] A.H.M. Emwas, The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research, Methods Mol. Biol. 1277 (2015) 161–193. doi:10.1007/978-1-4939-2377-9_13.

[7] R. Spicer, R.M. Salek, P. Moreno, D. Cañueto, C. Steinbeck, Navigating freely-available software tools for metabolomics analysis., Metabolomics. 13 (2017) 106. doi:10.1007/s11306-017-1242-7.

[8] S. Castillo, P. Gopalacharyulu, L. Yetukuri, M. Orešič, Algorithms and tools for the preprocessing of LC–MS metabolomics data, Chemom. Intell. Lab. Syst. 108 (2011) 23–32. doi:10.1016/J.CHEMOLAB.2011.03.010.

[9] M. Sugimoto, M. Kawakami, M. Robert, T. Soga, M. Tomita, Bioinformatics Tools for Mass Spectroscopy-Based Metabolomic Data Processing and Analysis, 2012. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3299976/pdf/CBIO-7-96.pdf (accessed February 19, 2019).

[10] M. Katajamaa, M. Orešič, Data processing for mass spectrometry-based metabolomics, J. Chromatogr. A. 1158 (2007) 318–328. doi:10.1016/j.chroma.2007.04.021.

[11] L. Hao, J. Wang, D. Page, S. Asthana, H. Zetterberg, C. Carlsson, O.C. Okonkwo, L. Li, Comparative Evaluation of MS-based Metabolomics Software and Its Application to Preclinical Alzheimer’s Disease, Sci. Rep. 8 (2018) 9291. doi:10.1038/s41598-018-27031-x.

[12] R. Vettukattil, Preprocessing of Raw Metabonomic Data, in: Humana Press, New York, NY, 2015: pp. 123–136. doi:10.1007/978-1-4939-2377-9_10.

[13] L. Vaclavik, O. Lacina, J. Hajslova, J. Zweigenbaum, The use of high performance liquid chromatography-quadrupole time-of-flight mass spectrometry coupled to advanced data mining and chemometric tools for discrimination and classification of red wines according to their variety, Anal. Chim. Acta. 685 (2011) 45–51. doi:10.1016/j.aca.2010.11.018.

[14] V. Sánchez De Medina, F. Priego-Capote, M.D. Luque De Castro, Characterization of refined edible oils enriched with phenolic extracts from olive leaves and pomace, J. Agric. Food Chem. 60 (2012) 5866–5873. doi:10.1021/jf301161v.

[15] T. Pluskal, S. Castillo, A. Villar-Briones, M. Orešič, MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data, BMC Bioinformatics. 11 (2010) 395. doi:10.1186/1471-2105-11-395.

[16] F. Giacomoni, G. Le Corguille, M. Monsoor, M. Landi, P. Pericard, M. Petera, C. Duperier, M. Tremblay-Franco, J.-F. Martin, D. Jacob, S. Goulitquer, E.A. Thevenot, C. Caron, Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics, Bioinformatics. 31 (2015) 1493–1495. doi:10.1093/bioinformatics/btu813.

[17] A. Lommen, MetAlign: Interface-Driven, Versatile Metabolomics Tool for Hyphenated Full-Scan Mass Spectrometry Data Preprocessing, Anal. Chem. 81 (2009) 3079–3086. doi:10.1021/ac900036d.

[18] A. Bertsch, C. Gröpl, K. Reinert, O. Kohlbacher, OpenMS and TOPP: Open Source Software for LC-MS Data Analysis, in: Methods Mol. Biol., 2011: pp. 353–367. doi:10.1007/978-1-60761-987-1_23.

[19] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification, Anal. Chem. 78 (2006) 779–787. doi:10.1021/ac051437y.

[20] R.J.M. Weber, T.N. Lawson, R.M. Salek, T.M.D. Ebbels, R.C. Glen, R. Goodacre, J.L. Griffin, K. Haug, A. Koulman, P. Moreno, M. Ralser, C. Steinbeck, W.B. Dunn, M.R. Viant, Computational tools and workflows in metabolomics: An international survey highlights the opportunity for harmonisation through Galaxy, Metabolomics. 13 (2017) 12. doi:10.1007/s11306-016-1147-x.

[21] Z. Li, Y. Lu, Y. Guo, H. Cao, Q. Wang, W. Shui, Comprehensive evaluation of untargeted metabolomics data processing software in feature detection, quantification and discriminating marker selection, Anal. Chim. Acta. 1029 (2018) 50–57. doi:10.1016/J.ACA.2018.05.001.

[22] G. Libiseller, M. Dvorzak, U. Kleb, E. Gander, T. Eisenberg, F. Madeo, S. Neumann, G. Trausinger, F. Sinner, T. Pieber, C. Magnes, IPO: a tool for automated optimization of XCMS parameters, BMC Bioinformatics. 16 (2015) 118. doi:10.1186/s12859-015-0562-8.

[23] C. Brunius, L. Shi, R. Landberg, Large-scale untargeted LC-MS metabolomics data correction using between-batch feature alignment and cluster-based within-batch

signal intensity drift correction, Metabolomics. 12 (2016) 1–13. doi:10.1007/s11306-016-1124-4.

[24] C.D. Broeckling, F.A. Afsar, S. Neumann, A. Ben-Hur, J.E. Prenni, RAMClust: A novel feature clustering method enables spectral-matching-based annotation for metabolomics data, Anal. Chem. 86 (2014) 6812–6817. doi:10.1021/ac501530d.

[25] Á. Fernández-Ochoa, R. Quirantes-Piné, I. Borrás-Linares, D. Gemperline, M.E. Alarcón Riquelme, L. Beretta, A. Segura-Carretero, Urinary and plasma metabolite differences detected by HPLC-ESI-QTOF-MS in systemic sclerosis patients, J. Pharm. Biomed. Anal. 162 (2019) 82–90. doi:10.1016/j.jpba.2018.09.021.

[26] J. Xia, D.S. Wishart, Using MetaboAnalyst 3.0 for Comprehensive Metabolomics Data Analysis., Curr. Protoc. Bioinformatics. 55 (2016) 14.10.1-14.10.91. doi:10.1002/cpbi.11.

[27] J. Xia, I. V Sinelnikov, B. Han, D.S. Wishart, MetaboAnalyst 3.0—making metabolomics more meaningful, Nucleic Acids Res. 43 (2015) W251–W257. doi:10.1093/nar/gkv380.

[28] W.E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods, Biostatistics. 8 (2007) 118–127. doi:10.1093/biostatistics/kxj037.

[29] R. Adusumilli, P. Mallick, Data Conversion with ProteoWizard msConvert, in: Humana Press, New York, NY, 2017: pp. 339–368. doi:10.1007/978-1-4939-6747-6_23.

[30] R.D.C. Team, R. R Development Core Team, R: A Language and Environment for Statistical Computing, R Found. Stat. Comput. (2016). doi:10.1007/978-3-540-74686-7.

[31] RStudio, RStudio: Integrated development for R, [Online] RStudio, Inc., Boston, MA URL Http//Www. Rstudio. Com. (2017). doi:10.1007/978-81-322-2340-5.

[32] J. Chong, O. Soufan, C. Li, I. Caraus, S. Li, G. Bourque, D.S. Wishart, J. Xia, MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis, Nucleic Acids Res. 46 (2018) W486–W494. doi:10.1093/nar/gky310.

[33] L. Shi, J.A. Westerhuis, J. Rosén, R. Landberg, C. Brunius, Variable selection and validation in multivari- ate modelling, Bioinformatics. (2018) 1–9. doi:10.1093/bioinformatics/bty710.

[34] F. Lindgren, B. Hansen, W. Karcher, M. Sjöström, L. Eriksson, Model validation by permutation tests: Applications to variable selection, J. Chemom. 10 (1996) 521–532. doi:10.1002/(SICI)1099-128X(199609)10:5/6<521::AID-CEM448>3.0.CO;2-J.

[35] L.W. Sumner, A. Amberg, D. Barrett, M.H. Beale, R. Beger, C.A. Daykin, T.W.M. Fan, O. Fiehn, R. Goodacre, J.L. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A.N. Lane, J.C. Lindon, P. Marriott, A.W. Nicholls, M.D. Reily, J.J. Thaden, M.R. Viant, Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI), Metabolomics. 3(3) (2007) 211–221. doi:10.1007/s11306-007-0082-2.

[36] K. Dührkop, M. Fleischauer, M. Ludwig, A.A. Aksenov, A. V. Melnik, M. Meusel, P.C. Dorrestein, J. Rousu, S. Böcker, SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information, Nat. Methods. (2019). doi:10.1038/s41592-019-0344-8.

[37] A. Gil de la Fuente, E. Grace Armitage, A. Otero, C. Barbas, J. Godzien, Differentiating signals to make biological sense – A guide through databases for MS-based non-targeted metabolomics, Electrophoresis. 38 (2017) 2242–2256. doi:10.1002/elps.201700070.

[38] B.A. Ejigu, D. Valkenborg, G. Baggerman, M. Vanaerschot, E. Witters, J.-C. Dujardin, T. Burzykowski, M. Berg, Evaluation of normalization methods to pave the way towards large-scale LC-MS-based metabolomics profiling experiments., OMICS. 17 (2013) 473–85. doi:10.1089/omi.2013.0010.

[39] H. Mizuno, K. Ueda, Y. Kobayashi, N. Tsuyama, K. Todoroki, J.Z. Min, T. Toyo'oka, The great importance of normalization of LC-MS data for highly-accurate non-targeted metabolomics, Biomed. Chromatogr. 31 (2017) 1–7. doi:10.1002/bmc.3864.

[40] X. Shen, X. Gong, Y. Cai, Y. Guo, J. Tu, H. Li, T. Zhang, J. Wang, F. Xue, Z.-J. Zhu, Normalization and integration of large-scale metabolomics data using support vector regression, Metabolomics. 12 (2016) 89. doi:10.1007/s11306-016-1026-5.

[41] J. Lee, J. Park, M. Lim, S.J. Seong, J.J. Seo, S.M. Park, H.W. Lee, Y.-R. Yoon, Quantile normalization approach for liquid chromatography-mass spectrometry-based metabolomic data from healthy human volunteers., Anal. Sci. 28 (2012) 801–5. http://www.ncbi.nlm.nih.gov/pubmed/22878636 (accessed March 15, 2019).

[42] M. Nodzenski, M.J. Muehlbauer, J.R. Bain, A.C. Reisetter, W.L. Lowe, D.M. Scholtens, Metabomxtr: an R package for mixture-model analysis of non-targeted metabolomics data, Bioinformatics. 30 (2014) 3287–3288. doi:10.1093/bioinformatics/btu509.

[43] A. Chawade, E. Alexandersson, F. Levander, Normalyzer: A Tool for Rapid Evaluation of Normalization Methods for Omics Data Sets, J. Proteome Res. 13 (2014) 3114–3120. doi:10.1021/pr401264n.

[44] A.M. De Livera, G. Olshansky, J.A. Simpson, D.J. Creek, NormalizeMets: assessing, selecting and implementing statistical methods for normalizing metabolomics data, Metabolomics. 14 (2018) 54. doi:10.1007/s11306-018-1347-7.

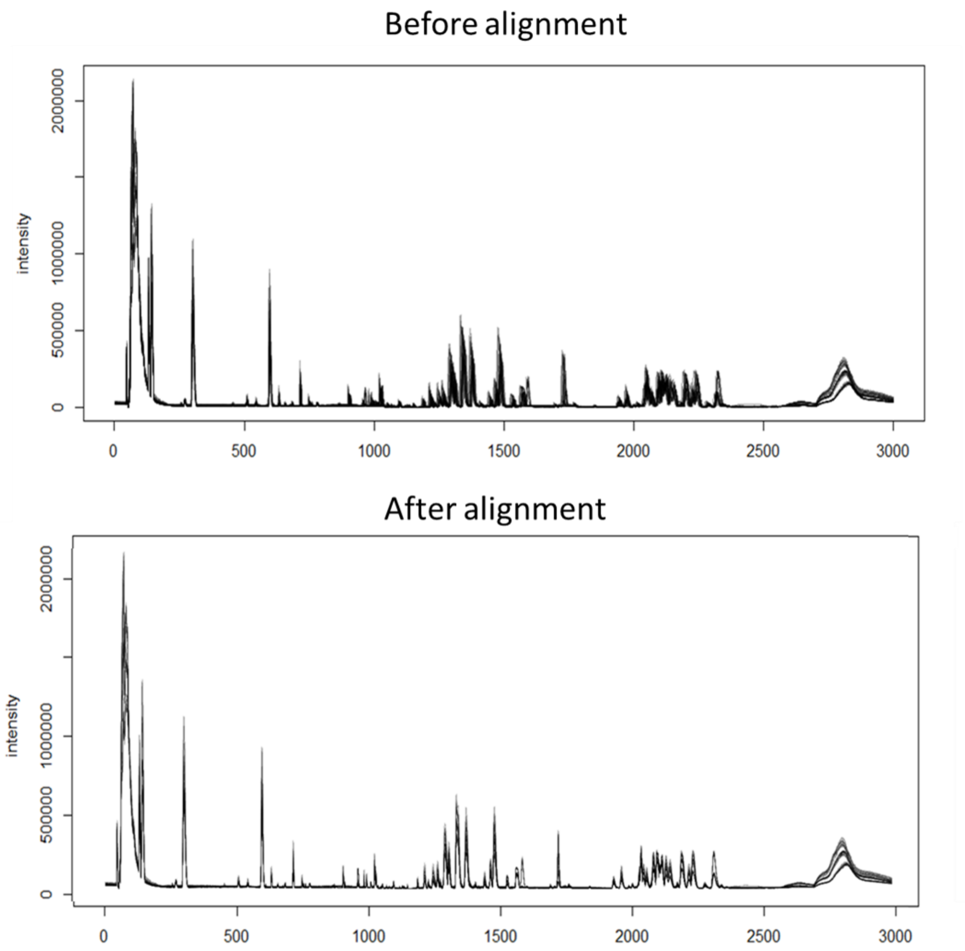
[45] E. Szymańska, E. Saccenti, A.K. Smilde, J.A. Westerhuis, Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies, Metabolomics. 8 (2012) 3–16. doi:10.1007/s11306-011-0330-3.

[46] A.A. Bengtsson, J. Trygg, D.M. Wuttge, G. Sturfelt, E. Theander, M. Donten, T. Moritz, C.J. Sennbro, F. Torell, C. Lood, I. Surowiec, S. R??nnar, T. Lundstedt, Metabolic profiling of systemic lupus erythematosus and comparison with primary Sjögren’s syndrome and systemic sclerosis, PLoS One. 11 (2016) 1–15. doi:10.1371/journal.pone.0159384.

Supplementary material

Costs and benefits of switching from vendor-based to open source pipelines for untargeted LC-MS metabolomics.

Fig. 1S. Chromatograms before and after the alignment step performed by XCMS using the parameters optimized by IPO. In each figure, there are 21 superposed chromatograms from QC samples analysed throughout the three batches.



Before alignment

After alignment

Fig. 2S. a) An example of the failure in the automatic area integration step produced using the methodology based on vendor software. It shows how retention times do not change after molecular feature extraction and consequently some peaks are not well integrated. b) The same ion extracted by XCMS c) The result of the example after the alignment of the retention times by XCMS (It shows how retention times change after this step.) c) The result of the example highlighting the section integrated by XCMS. The integration error observed in figure 2a is not observed in this case.

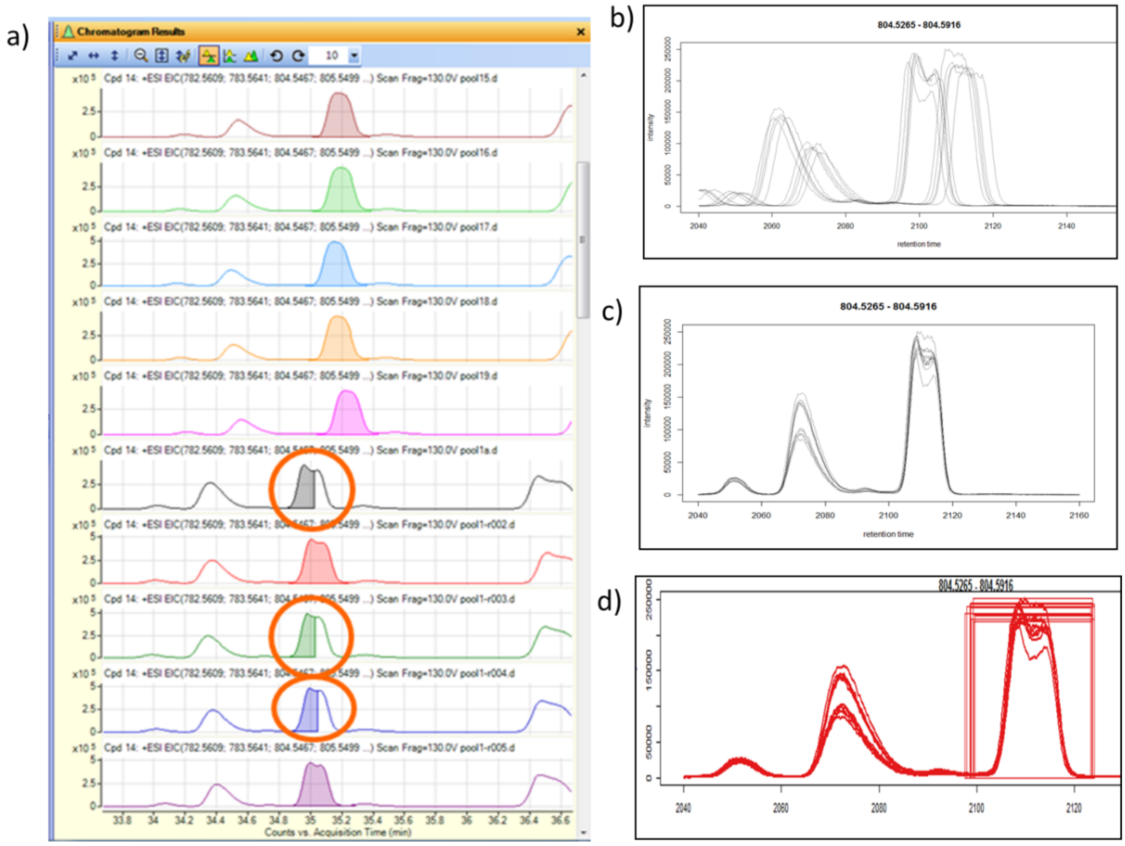


Fig. 3S. a) Histogram of the results of the correlations between the peaks areas which were found by both methodologies (*Peak Areas obtained by R versus Peak Areas obtained by Agilent Mass Profinder after manually correction*). **b)** Scatter plots that compares the integrated areas obtained by R and Agilent (after manually correction) of 12 random peaks.

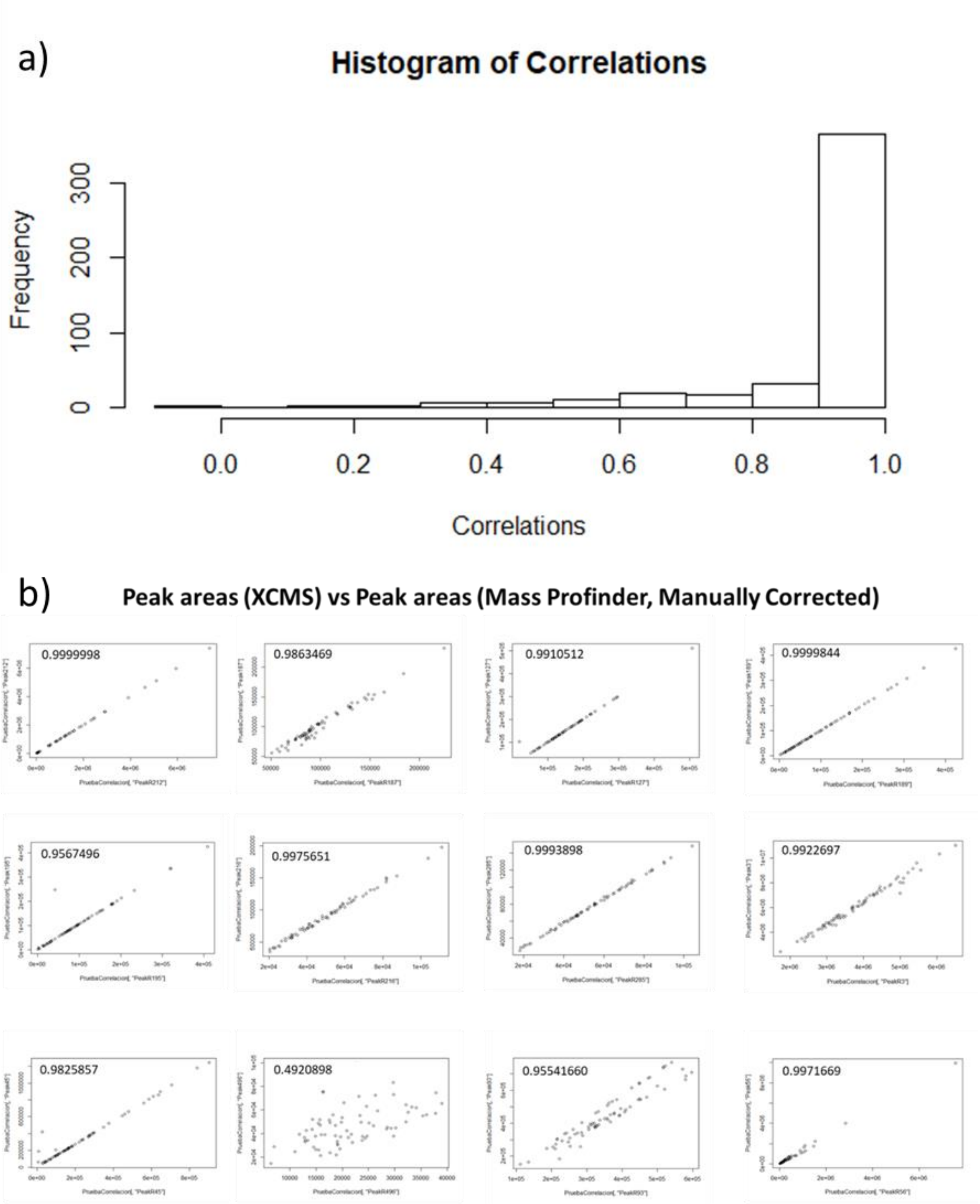


Fig. 4S. Histograms of the results of the correlations between the peaks areas which were found by both methodologies (*Peak Areas obtained by R versus Peak Areas obtained by Agilent Mass Profinder before (a) or after (b) manually correction*).

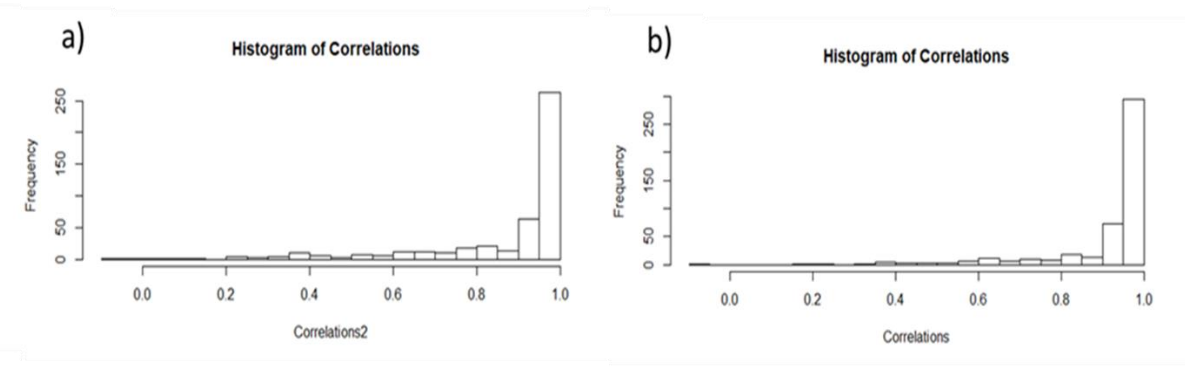


Fig. 5S. PCA scores plot from data obtained by methodology based on vendor software. 1a) Raw Data; 1b) data obtained after Bayes method (Metaboanalyst 3.0); 1c) data obtained after bayes and MSTUS methods (Metaboanalyst 3.0) (batch 1, red plots; batch 2, green plots; batch3, blue plots; QCs, inside the circle).

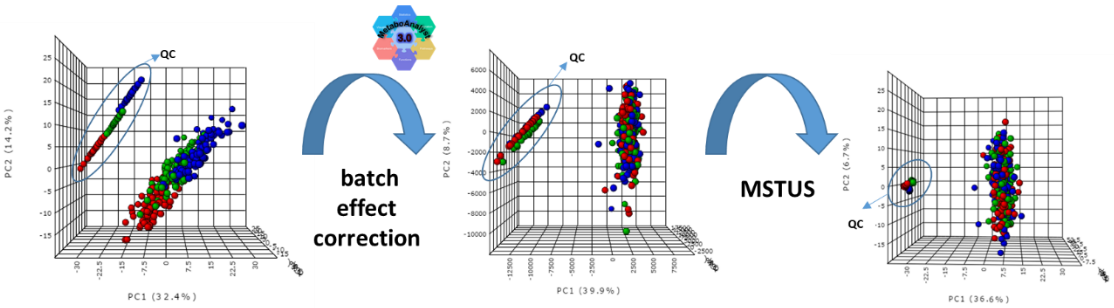


Fig. 6S. Examples of different drift trends observed throughout the analytical sequence. These different trends are taken into account in batchCorr normalization.

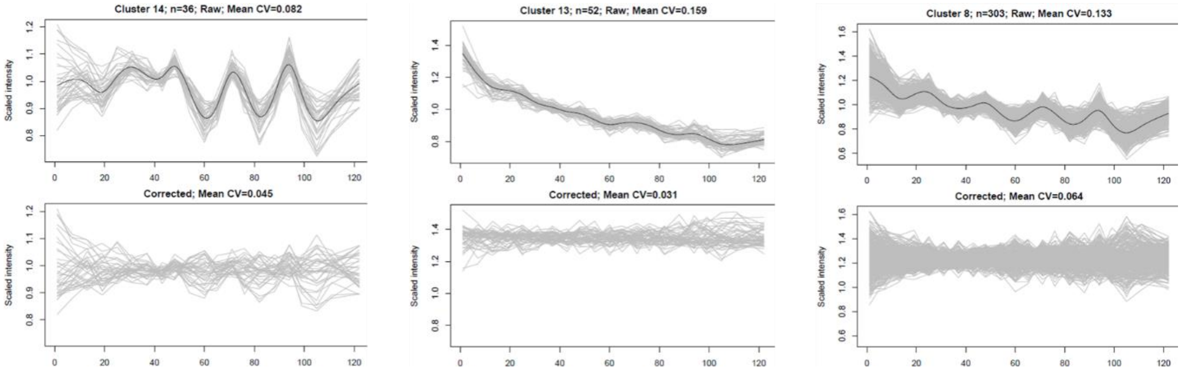


Fig. 7S. PCA score plots obtained with data obtained by R methodology. a) Raw data, b) after within batch normalization (batchCorr) c) after between-batch normalization (batchCorr)

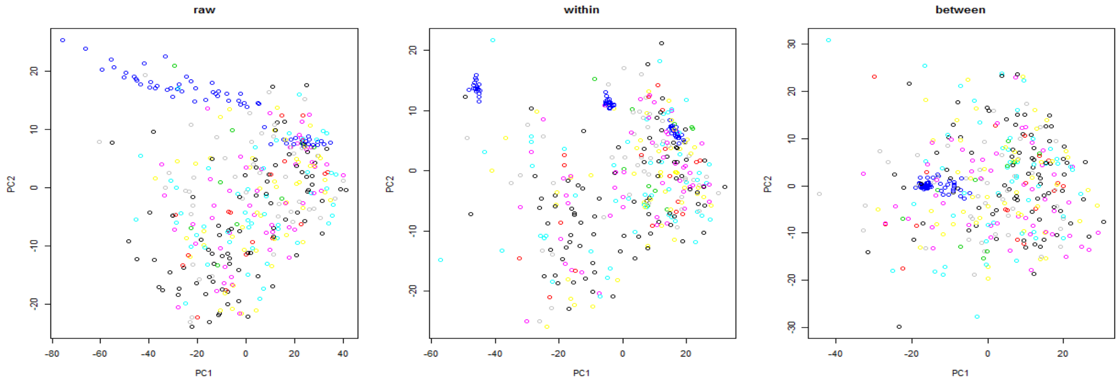


Table 1S. Detailed analytical information (name, mass, retention time (RT), score, molecular formula and fragments) of the 15 most important peaks in the PLS-DA (Metaboanalyst 3.0) model developed with the data obtained from commercial software.

Peak Nº (Agilent)	Overall Score (%)	Mass	RT (min)	Molecular Formula	Score (%)	Compound name	MS/MS Fragments	Identification Database
507	100.0	511.3117	18.20	$C_{24}H_{50}NO_8P$	86.32	PS(O-18:0/0:0) (iso1)	351.2465/475.2374/534.3116	LMGP03060002
401	85.77	346.0787	13.83	$C_{10}H_{15}N_6O_6P$	77.49	3'-Amino-3'-deoxy-AMP	120.0773/194.0561/222.0507/347.0816	C07026
122	84.10	511.3218	18.40	$C_{24}H_{50}NO_8P$	89.31	PS(O-18:0/0:0) (iso2)	351.2465/475.2374/534.3116	LMGP03060002
127	81.07	208.0857	4.34	$C_{10}H_{12}N_2O_3$	85.66	L-kynurenine	74.0216/94.0627/120.0413/146.0561	HMDB00684
61	76.26	180.0650	10.58	$C_7H_8N_4O_2$	97.25	Theophylline/Paraxanthine	55.0283/96.0454/124.0490/181.0805	HMDB01889/ HMDB01860
496	65.80	919.5195	34.18	--	--	Unknown	--	--
151	63.67	129.0729	1.32	$C_6H_{11}NO_2$	85.83	Pipecolic acid	56.0479/69.0569/84.0804	HMDB01860
305	62.39	899.5472	36.26	$C_{49}H_{84}NO_{10}P$	87.42	PS(43:6)	146.9787/649.507/731.4886/773.4973/832.57007	LMGP03010004
511	60.98	259.1783	12.78	$C_{13}H_{25}NO_4$	86.61	Hexanoylcarnitine	60.0804/71.0856/85.0287	HMDB00705
494	60.69	341.2469	17.14	$C_{19}H_{35}NO_4$	96.48	trans-2-dodecenoylcarnitine	85.0263/163.0386	HMDB13326
251	60.33	616.4635	28.76	--	--	Unknown	--	--
373	57.60	158.0837	10.01	$C_{10}H_{10}N_2$	86.30	1-Benzylimidazole	159.0837	MetlinID63058
168	56.70	287.2069	15.28	$C_{15}H_{29}NO_4$	96.21	Octanoylcarnitine	60.0785/85.0262/288.2109	HMDB00834
91	56.49	315.2366	16.64	$C_{17}H_{33}NO_4$	94.56	Decanoylcarnitine	60.0788/85.0262/316.2425	HMDB62631
485	55.31	430.2593	19.05	--	--	Unknown	--	--

Table 2S. Detailed analytical information (name, mass, retention time (RT), score, molecular formula and fragments) of the 15 most important peaks in the PLS-DA (Metaboanalyst 3.0) model developed with the data obtained from open source.

Peak Nº (R)	Overall Score (%)	Mass	RT (min)	Molecular Formula	Score (%)	Compound name	MS/MS Fragments	Identification Database
R_355	100	412.1294	17.75	--	--	Unknown	--	--
R_18	98.58	208.0857	4.33	$C_{10}H_{12}N_2O_3$	85.66	L-kynurenine	74.0216/94.0627/120.0413/146.0561	HMDB00684
R_248	95.04	174.0342	10.64	--	--	Unknown	--	--
R_75	89.44	511.3259	18.39	$C_{24}H_{50}NO_8P$	89.31	PS(O-18:0/0:0)	351.2465/475.2374/534.3116	LMGP03060002
R_52	87.61	378.5454	24.64	$C_{23}H_{38}O_4$	85.09	MG(20:4)	67.0533/81.0685/95.0839	HMDB11578
R_223	84.36	129.0792	1.21	$C_6H_{11}NO_2$	85.83	Pipecolic acid	56.0479/69.0569/84.0804/130.0865	HMDB01860
R_266	82.29	216.1011	7.92	--	--	Unknown	--	--
R-17	80.40	184.1211	9.01	$C_9H_{16}N_2O_2$	86.53	N-(3-acetamidopropyl)pyrrolidin-2-one	56.9408/86.9471/98.0591/126.0905/185.1262	HMDB61384
R_361	80.05	425.3468	19.30	$C_{25}H_{47}NO_4$	72.95	Octadecenoylcarnitine	85.0274/426.3541	HMDB94687
R_417	77.91	619.48	29.40	--	--	Unknown	--	
R_13	71.55	194.0808	11.98	$C_8H_{10}N_4O_2$	94.86	Caffeine	56.0483/69.0433/83.0593/110.0687/138.064	HMDB01847
R_11	71.34	180.0650	10.61	$C_7H_8N_4O_2$	97.25	Theophylline/Paraxanthine	55.0283/96.0454/124.0490/181.0805	HMDB01889/ HMDB01860
R_40	70.51	304.2390	26.30	$C_{20}H_{32}O_2$	87.11	Arachidonic acid	55.0530/57.0685/67.0526/71.0842/ 107.0811/121.0995/123.0117/161.112/177.3601	HMDB01043
R_414	70.39	596.4926	28.60	$C_{36}H_{68}O_6$	86.75	Glycerol triundecanoate	337.2304/339.2447/361.2273/	HMDB31089
R_486	66.54	919.5195	34.18	--	--	Unknown	--	--

Table 3S. Detailed analytical information (name, mass, retention time (RT), score, molecular formula and fragments) of the important peaks obtained in the MUVR (PLS) model with the data from vendor software.

Peak Nº (Agilent)	Rank	Mass	RT (min)	Molecular Formula	Score (%)	Compound Name	MS/MS fragments	Identification Database
PEAK507	3.17	511.3117	18.20	$C_{24}H_{50}NO_8P$	86.32	PS(O-18:0/0:0)	351.2465/475.2374/534.3116	LMGP03060002
PEAK520	5.52	397.3082	18.53	$C_{23}H_{43}NO_4$	77.79	Hexadecanoyl carnitine	85.0286/398.3256	HMDB13207
PEAK367	6.08	934.6784	26.2	$C_{51}H_{99}O_{12}P$	80.66	PI(P-42:0)	Level 3	LMGP06020067
PEAK288	9.03	676.4521	26.4	$C_{36}H_{69}O_9P$	88.88	PG(P-30:1)	Level 3	LMGP04030004
PEAK305	9.35	877.5695	36.26	$C_{49}H_{84}NO_{10}P$	87.42	PS(43:6)	146.9787/649.507/731.4886/773.4973/832.57007	LMGP03010004
PEAK253	9.75	282.2481	26.36	$C_{18}H_{34}O_2$	81.47	Oleic acid	57.0684/69.0686/83.0832/97.0984/135.1141/149.1283	HMDB00207
PEAK79	12.72	630.4439	26.36	$C_{35}H_{67}O_7P$	84.47	PA(P-32:1)	57.0685/283.2572	LMGP10030031
PEAK127	13.21	208.0857	4.34	$C_{10}H_{12}N_2O_3$	85.66	L-kynurenine	74.0216/94.0627/120.0413/146.0561	HMDB00684
PEAK227	16.07	361.2158	26.33	$C_{17}H_{34}N_2O_4P$	84.31	Unknown	Level 4	--
PEAK122	17.78	511.3218	18.40	$C_{24}H_{50}NO_8P$	89.31	PS(O-18:0/0:0)	351.2465/475.2374/534.3116	LMGP03060002
PEAK29	18.16	326.2206	26.18	$C_{20}H_{32}O_2$	87.11	Arachidonic acid	55.0530/57.0685/67.0526/71.0842/107.0811/121.0995/123.0117/161.112/177.3601	HMDB01043
PEAK414	28.86	436.1598	25.72	$C_{16}H_{28}N_4O_8S$	64.49	Thr Trp Met	Level 3	Metlin16635
PEAK131	29.41	507.3473	24.45	$C_{26}H_{54}NO_6P$	94.21	PC(P-18:0/0:0)	88.1102/508.3473	HMDB13122
PEAK165	32.15	394.1974	26.2	$C_{16}H_{26}N_8O_4$	97.84	Unknown	Level 4	--
PEAK226	34.93	335.2003	25.73	$C_{20}H_{32}O_2P$	84.08	Unknown	Level 4	--
PEAK160	37.45	368.1765	25.73	$C_{14}H_{24}N_8O_4$	97.26	Glycylhystidilargenine	295.1890/369.1924	Metlin21026
PEAK52	39.30	278.2184	25.59	$C_{18}H_{30}O_2$	87.00	Linolenic Acid	69.0660/95.0833/123.1134/279.2177	HMDB01388
PEAK200	48.99	672.5185	30.84	$C_{37}H_{73}N_2O_6P$	71.00	SM(d18:2/14:0)	86.0948/146.9801/512.4402	LMSP03010034
PEAK209	49.16	926.6471	16.05	$C_{45}H_{76}N_{21}O$	96.24	Unknown	Level 4	--
PEAK525	58.03	598.4061	24.66	$C_{26}H_{58}N_6O_9$	97.85	Unknown	Level 4	--
PEAK190	59.13	328.2419	24.37	$C_{22}H_{32}O_2$	83.41	DHA	107.0814/145.0975/173.1304/161.1285	HMDB02183
PEAK312	61.25	733.5548	36.29	$C_{40}H_{80}NO_8P$	81.94	PC(32:0)	Level 3	HMDB07871
PEAK186	64.95	781.5627	35.22	$C_{44}H_{80}NO_8P$	96.43	PC(36:4)	86.0956/146.9812/184.0724/621.4832	LMGP01010927
PEAK30	68.77	302.2129	24.52	$C_{20}H_{30}O_2$	80.91	Eicosapentaenoic acid	81.0686/91.0535/119.0165/303.2146/325.2115	HMDB01999
PEAK264	74.40	549.3545	25.44	$C_{28}H_{56}NO_7P$	94.91	LysoPC(20:1)	86.0939/104.1048/146.9795/367.3151/	HMDB11148
PEAK75	76.36	164.0457	2.39	$C_9H_8O_3$	95.15	m-Coumaric acid	91.0523/119.0471/136.0733/147.0409	HMDB01713

Table 3S(Cont)

Peak Nº (Agilent)	Rank	Mass	RT (min)	Molecular Formula	Score (%)	Compound Name	MS/MS fragments	Identification Database
PEAK494	76.53	341.2469	17.14	$C_{19}H_{35}NO_4$	96.48	Trans-2-dodecenoylcarnitine	85.0263/163.0386	HMDB13326
PEAK276	78.42	1537.5468	35.47	--	--	Unknown	--	--
PEAK405	78.55	694.3850	24.65	$C_{26}H_{50}N_{18}O_5$	98.98	Unknown	--	--
PEAK386	80.64	396.2115	28.36	$C_{16}H_{28}N_8O_4$	99.50	Unknown	--	--
PEAK384	82.07	525.2812	21.69	$C_{27}H_{44}NO_7P$	81.45	LysoPE(22:6)	Level 3	HMDB11526
PEAK389	82.42	326.2135	24.35	--	--	Unknown	--	--
PEAK444	83.89	359.2042	24.52	--	--	Unknown	--	--
PEAK347	85.93	1653.0686	35.24	--	--	Unknown	--	--
PEAK401	85.95	346.0787	13.58	$C_{10}H_{15}N_6O_6P$	77.49	3'-Amino-3'-deoxy-AMP	120.0773/194.0561/222.0507/347.0816	C07026
PEAK470	95.41	928.623	24.52	$C_{51}H_{93}O_{12}P$	75.32	PI(O-42:4)	Level 3	LMGP06020069
PEAK224	103.32	369.2780	17.92	$C_{21}H_{39}NO_4$	94.08	cis-5-Tetradecenoylcarnitine	85.0270/370.2902	HMDB02014
PEAK410	103.56	350.2162	24.37	$C_{19}H_{30}N_2O_4$	80.55	Perindopril lactam	351.2251	MetlinID1799
PEAK61	106.74	180.065	10.58	$C_7H_8N_4O_2$	97.25	*Theophylline/Paraxanthine*	55.0283/96.0454/124.0490/181.0805	HMDB01889/ HMDB01860
PEAK511	111.06	259.1783	12.78	$C_{13}H_{25}NO_4$	86.61	Hexanoylcarnitine	60.0804/71.0856/85.0287	HMDB00705
PEAK103	132.69	153.0789	2.39	C_8H_9NO	97.59	Dopamine	65.0370/91.0523/119.0471/136.0731	MetlinID64
PEAK4	136.48	519.3335	21.63	$C_{26}H_{50}NO_7P$	87.88	LysoPC(18:2)	104.1062/184.072	HMDB10386
PEAK166	137.88	515.3059	20.99	$C_{26}H_{46}NO_7P$	81.33	LysoPC(18:3)	104.1060/146.9804/184.0720	HMDB10389
PEAK175	140.23	541.3168	21.26	$C_{28}H_{48}NO_7P$	79.65	LysoPC(20:5)	86.0956/146.9805/184.0718	
PEAK267	142.65	103.0446	1.47	C_4H_9NS	98.39	(±)-2-Methylthiazolidine	56.0481/104.0530	HMDB31682
PEAK144	145.04	517.3666	21.87	$C_{26}H_{48}NO_7P$	60.99	LysoPC(18:3)	96.0954/104.1058/184.0715	HMDB10387
PEAK141	149.88	184.1221	9.05	$C_9H_{16}N2O_2$	86.53	*N-(3-*aminopropyl)*pyrrolidin-2-one*	56.9408/86.9471/98.0591/126.0905/185.1262	HMDB61384
PEAK47	161.45	767.5627	36.48	$C_{44}H_{82}NO_7P$	95.32	PC(36:3)	86.0942/146.9794/184.0702	HMDB11246
PEAK373	162.25	158.0837	10.00	$C_{10}H_{10}N_2$	86.30	1-Benzylimidazole	159.0837	MetlinID63058
PEAK379	165.79	775.5092	33.59	$C_{44}H_{74}NO_8P$	77.61	PC(36:7)	86.095/593.4508/717.4421	HMDB08723
PEAK251	166.08	616.4648	28.60	--	--	Unknown	--	--

Table 3S(Cont)

Peak Nº (Agilent)	Rank	Mass	RT (min)	Molecular Formula	Score (%)	Compound Name	MS/MS fragments	Identification Database
PEAK298	166.44	468.3006	26.80	$C_{30}H_{44}O_4$	77.96	Unknown	--	--
PEAK169	178.04	700.5492	32.85	$C_{39}H_{77}N_2O_6P$	74.75	SM(d18:1/16:1)	86.0950/184.0715	LMSP03010041
PEAK177	178.38	795.5868	38.85	$C_{46}H_{86}NO_7P$	96.07	PC(37:4)	86.0971/146.9827	HMDB11252
PEAK151	179.63	129.0729	1.32	$C_6H_{11}NO_2$	85.83	Pipecolic acid	56.0479/69.0569/84.0804/130.0865	HMDB01860
PEAK336	180.93	780.5849	38.70	$C_{45}H_{85}N_2O_6P$	97.37	C22:3 Sphingomyelin	Level 3	HMDB13468
PEAK349	183.71	743.5593	36.67	$C_{42}H_{82}NO_7P$	67.14	PC(34:1)	Level 3	HMDB11240
PEAK358	184.03	1273.2775	17.56	--	--	Unknown	--	--
PEAK91	185.86	315.2366	16.64	$C_{17}H_{33}NO_4$	94.56	Decanoylcarnitine	60.0788/85.0262/316.2425	HMDB62631
PEAK168	189.69	287.2069	15.28	$C_{15}H_{29}NO_4$	96.21	Octanoylcarnitine	60.0785/85.0262/288.2109	HMDB00834
PEAK23	191.26	117.0749	1.27	$C_5H_{11}NO_2$	97.78	L-Valine	55.0528/72.0792/117.0749	
PEAK331	192.73	1585.5867	35.25	--	--	Unknown	--	--
PEAK332	194.29	1128.6258	21.61	--	--	Unknown	--	--
PEAK496	194.69	919.5195	34.18	--	--	Unknown	--	--
PEAK528	194.76	253.1263	15.07	$C_{13}H_{19}NO_4$	62.1	3-Indolecarboxylic acid	254.1301	MetlinID6660
PEAK365	197.61	322.1777	23.16	$C_{12}H_{26}N_4O_6$	83.21	Perindoprilat lactam A	323.1926	MetlinID1800
PEAK25	205.07	519.3335	21.63	$C_{26}H_{50}NO_7P$	87.88	LysoPC(18:2)	104.1062/184.072	HMDB10386

Table 4S. Details (name, mass, retention time (RT), score, molecular formula and fragments) of the important peaks obtained in the MUVR (PLS) model with the data from R.

Peak Nº (R)	Rank	Mass	RT (min)	Molecular Formula	Score (%)	Compound Name	MS/MS Fragments	Identification Database
R_355	1.52	412.1294	17.76	--	--	Unknown	--	--
R_99	4.50	676.4521	26.35	$C_{51}H_{99}O_{12}P$	80.66	PI(P-42:0)	Level 3	LMGP06020067
R_18	4.79	208.0857	4.33	$C_{10}H_{12}N_2O_3$	85.66	L-kynurenine	74.0216/94.0627/120.0413/146.0561	HMDB00684
R_40	6.69	304.2390	26.35	$C_{20}H_{32}O_2$	87.11	Arachidonic acid	55.0530/57.0685/67.0526/71.0842/ 107.0811/121.0995/123.0117/161.112/177.3601	HMDB01043
R_361	8.25	425.3468	19.32	$C_{21}H_{39}NO_4$	72.95	Octadecenoylcarnitine	85.0274/426.3541	HMDB94687
R_248	9.35	174.0342	10.64	--	--	Unknown	--	--
R_75	14.16	511.3259	18.40	$C_{24}H_{50}NO_8P$	89.31	PS(O-18:0/0:0)	351.2465/475.2374/534.3116	LMGP03060002
R_33	14.56	278.2194	25.72	$C_{18}H_{30}O_2$	87.00	Linolenic Acid	69.0660/95.0833/123.1134/279.2177	HMDB01388
R_36	23.21	672.5185	24.65	$C_{37}H_{73}N_2O_6P$	71.00	SM(d18:2/14:0)	86.0948/146.9801/512.4402	LMSP03010034
R_32	50.79	276.2032	23.72	$C_{18}H_{28}O_2$	60.57	Stearidonic acid	55.0521/69.0687/95.0825/161.9743/	HMDB06547
R_48	52.06	328.2368	24.35	$C_{22}H_{32}O_2$	83.41	Docosahexanoic acid	107.0814/145.0975/173.1304/161.1285	HMDB02183
R_39	57.54	302.2195	24.65	$C_{20}H_{30}O_2$	80.91	Eicosapentaenoic acid	81.0686/91.0535/119.0165/303.2146/325.2115	HMDB01999
R_46	94.75	348.2002	24.49	--	--	Unknown	--	--
R_41	103.28	517.3666	21.87	$C_{26}H_{48}NO_7P$	60.99	LysoPC(18:3)	96.0954/104.1058/184.0715	HMDB10387
R_17	103.43	184.1211	9.04	$C_9H_{16}N_2O_2$	86.53	N-(3-acetamidopropyl)pyrrolidin-2-one	56.9408/86.9471/98.0591/126.0905/185.1262	HMDB61384

Table 5S. Members from the Precisesads Clinical Consortium:

Researchers and clinicians	***Clinical center***
Lorenzo Beretta, Barbara Vigone	Referral Center for Systemic Autoimmune Diseases, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico di Milano, Italy.
Jacques-Olivier Pers, Alain Saraux, Valérie Devauchelle-Pensec, Divi Cornec, Sandrine Jousse-Joulin	Centre Hospitalier Universitaire de Brest, Hospital de la Cavale Blanche, Brest, France.
Bernard Lauwerys, Julie Ducreux, Anne-Lise Maudoux	Pôle de pathologies rhumatismales systémiques et inflammatoires, Institut de Recherche Expérimentale et Clinique, Université catholique de Louvain, Brussels, Belgium.
Carlos Vasconcelos, Ana Tavares, Esmeralda Neves, Raquel Faria, Mariana Brandão	Centro Hospitalar do Porto, Portugal.
Ana Campar, António Marinho, Fátima Farinha, Isabel Almeida	Servicio Cantabro de Salud, Hospital Universitario Marqués de Valdecilla, Santander, Spain.
Miguel Angel Gonzalez-Gay Mantecón, Ricardo Blanco Alonso, Alfonso Corrales Martínez	Servicio Cantabro de Salud, Hospital Universitario Marqués de Valdecilla, Santander, Spain.
Ricard Cervera, Ignasi Rodríguez-Pintó, Gerard Espinosa	Hospital Clinic I Provicia, Institut d'Investigacions Biomèdiques August Pi i Sunyer, Barcelona, Spain.
Rik Lories, Ellen De Langhe	Katholieke Universiteit Leuven, Belgium.
Nicolas Hunzelmann, Doreen Belz	Klinikum der Universitaet zu Koeln, Cologne, Germany.
Torsten Witte, Niklas Baerlecken	Medizinische Hochschule Hannover, Germany.
Georg Stummvoll, Michael Zauner, Michaela Lehner	Medical University Vienna, Vienna, Austria.
Eduardo Collantes, Rafaela Ortega-Castro, Mª Angeles Aguirre-Zamorano, Alejandro Escudero-Contreras, Mª Carmen Castro-Villegas	Servicio Andaluz de Salud, Hospital Universitario Reina Sofía Córdoba, Spain.

Table 5S (Cont).

Researchers and clinicians	Clinical center
Norberto Ortego, María Concepción Fernández Roldán	Servicio Andaluz de Salud, Complejo hospitalario Universitario de Granada (Hospital Universitario San Cecilio), Spain.
Enrique Raya, Inmaculada Jiménez Moleón	Servicio Andaluz de Salud, Complejo hospitalario Universitario de Granada (Hospital Virgen de las Nieves), Spain.
Enrique de Ramon, Isabel Díaz Quintero	Servicio Andaluz de Salud, Hospital Regional Universitario de Málaga, Spain
Pier Luigi Meroni, Maria Gerosa, Tommaso Schioppo, Carolina Artusi,	Università degli studi di Milano, Milan, Italy.
Carlo Chizzolini, Aleksandra Zuber, Donatienne Wynar,	Hospitaux Universitaires de Genève, Switzerland.
Laszló Kovács, Attila Balog, Magdolna Deák, Márta Bocskai, Sonja Dulic, Gabriella Kádár	University of Szeged, Szeged, Hungary.
Falk Hiepe, Velia Gerl, Silvia Thiel	Charite, Berlin, Germany.
Manuel Rodriguez Maresca, Antonio López-Berrio, Rocío Aguilar-Quesada, Héctor Navarro-Linares	Andalusian Public Health System Biobank, Granada, Spain
Yiannis Ioannou, Chris Chamberlain, Jacqueline Marovac.	UCB Pharma, Slough, United Kingdom (PRECISESADS Project office)
Marta Alarcón Riquelme, Tania Gomes Anjos.	Department of Medical Genomics, Center for Genomics and Oncological Research (GENYO), Granada, Spain (PRECISESADS Project Office)

A brief Tutorial on *R-based approach*

Costs and benefits of switching from vendor-based to open source pipelines for untargeted LC-MS metabolomics.

This tutorial aims to detail step-by-step how the R approach has been developed. We pretend to show the way we have applied the different R packages in our data. In this sense, our main objective is to help beginners in R language to be able to use the different packages and scripts used in the manuscript. Although tutorials may exist for individual pre-processing modules, in this document is showed how to stitch together these modules into entire pipelines. Nevertheless, we recommend you reading the different manuscripts and tutorials of each R package available in the following links:

IPO (Libiseller et al. 2015):
https://bioconductor.org/packages/release/bioc/vignettes/IPO/inst/doc/IPO.html
XCMS (Smith et al. 2006):
https://bioconductor.org/packages/release/bioc/vignettes/xcms/inst/doc/xcms.html
https://jotsetung.github.io/metabolomics2018/xcms-preprocessing.html
batchCorr (Brunius et al. 2016):
https://gitlab.com/CarlBrunius/batchCorr/tree/master/Tutorial
RAMClustR (Broeckling et al. 2014):
http://pubs.acs.org/doi/abs/10.1021/ac501530d
MUVR (Shi et al. 2018):
https://github.com/CarlBrunius/MUVR/blob/master/README.md

1) Installation.

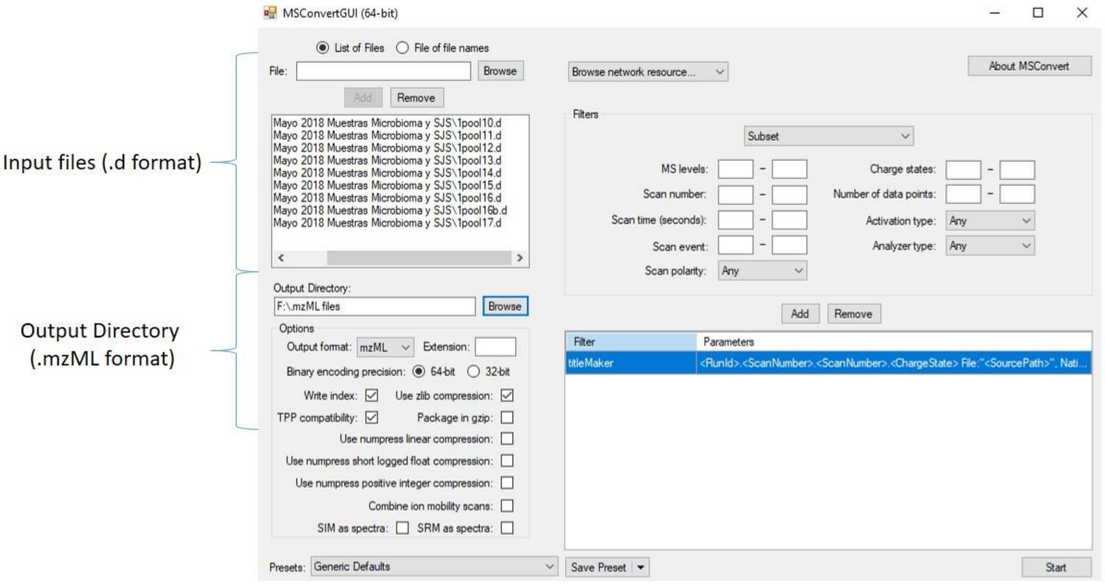
First, R and RStudio were downloaded from the websites (https://www.r-project.org/, https://www.rstudio.com/) and installed on a Windows 7 computer. Second, the different packages used were installed in the R environment with the following codes:

```
install.packages("BiocManager", repos="http://cran.us.r-project.org",
dependencies=TRUE)
install.packages("devtools", repos="http://cran.us.r-project.org",
dependencies=TRUE)
```

```
library(devtools)
library(BiocManager)

BiocManager::install("IPO", version = "3.8")
BiocManager::install("xcms", version = "3.8")
install_github("cbroeckl/RAMClustR", build_vignettes = TRUE, dependenc
ies = TRUE)
devtools::install_git("https://gitlab.com/CarlBrunius/batchCorr.git")
```

```
library(IPO)
library(xcms)
library(RAMClustR)
library(batchCorr)
```

In order to be able to read the data in the R environment, it is necessary to transform the data format into an adequate format (mzML, mzXML, mzData, NetCDF). In our particular case, data were collected by an Agilent instrument in a .d format. We transformed this format into .mzML using *MSConvertGUI* software. This tool from proteowizard can be downloaded from the website http://proteowizard.sourceforge.net/download.html. This software is easy to use as it is shown schematically in the following figure.



● List of Files ○ File of file names

File:

Mayo 2018 Muestras Microbioma y SJS\1pool10.d
Mayo 2018 Muestras Microbioma y SJS\1pool11.d
Mayo 2018 Muestras Microbioma y SJS\1pool12.d
Mayo 2018 Muestras Microbioma y SJS\1pool13.d
Mayo 2018 Muestras Microbioma y SJS\1pool14.d
Mayo 2018 Muestras Microbioma y SJS\1pool15.d
Mayo 2018 Muestras Microbioma y SJS\1pool16.d
Mayo 2018 Muestras Microbioma y SJS\1pool16b.d
Mayo 2018 Muestras Microbioma y SJS\1pool17.d

Output Directory: F:\.mzML files

Options

Output format: mzML Extension:

Binary encoding precision: ● 64-bit ○ 32-bit

Write index: ☑ Use zlib compression: ☑

TPP compatibility: ☑ Package in gzip: □

Use numpress linear compression: □

Use numpress short logged float compression: □

Use numpress positive integer compression: □

Combine ion mobility scans: □

SIM as spectra: □ SRM as spectra: □

Presets: Generic Defaults

Filters

Subset

MS levels: – Charge states: –

Scan number: – Number of data points: –

Scan time (seconds): – Activation type: Any

Scan event: – Analyzer type: Any

Scan polarity: Any

Filter	Parameters
titleMaker	<RunId>.<ScanNumber>.<ScanNumber>.<ChargeState> File:"<SourcePath>", Nati...

2) IPO.

The aim of this package is to optimize the xcms parameters. For this purpose, 6 QC files well distributed throughout the sequence were selected.

```
library(xcms)
library(RColorBrewer)
library(pander)
library(magrittr)
library("IPO", lib.loc="~/R/win-library/3.5")
library("msdata", lib.loc="~/R/win-library/3.5")
```

The selected files must be saved in a folder within the working directory. If you do not know what this directory is, use the command “>getwd()”. In our case, the 6 QC files were located in a folder called “QC” and we used the following command to import them into the R environment.

```
#Open files
datafiles <- list.files("QC", recursive = TRUE, full.names = TRUE, pat
tern=".mzML")
peakpickingParameters <- getDefaultXcmsSetStartingParams('centWave')
```

To work more quickly in R, it is very useful to work in parallel using the different computer cores. For this, we use the “doParallel” package. In order to continue working with the computer while the R commands are executing, a good number of cores is the numbers of available cores minus one.

```
library(doParallel)
nCore=detectCores()-1
```

Next, the different parameters for peak peaking were selected. As our data were obtained by a high resolution mass spectrometer, the peak picking method was ‘CentWave’. Some parameters were optimized by IPO package selecting a reasonable range according to our experiments. In this way, *min_peakwidth*, *max_peakwidth* and *ppm* were the optimized parameters. However, the *noise* and *prefilter* parameters were set at a single value. These parameters were selected after observing the current noise level in the chromatograms. The selection of the different parameters was performed with the following code:

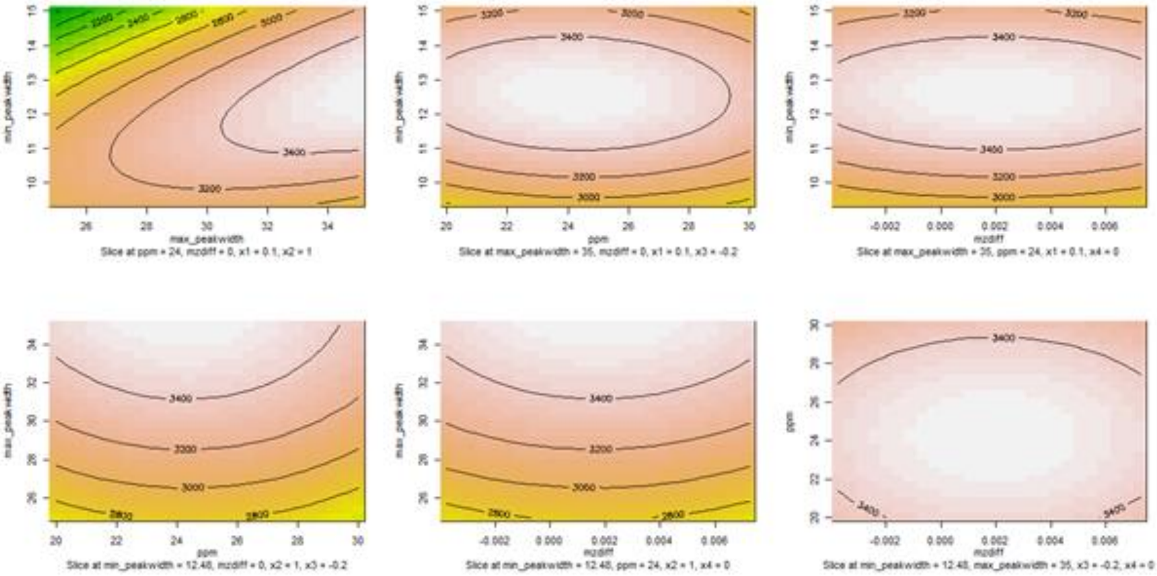
```
#PeakPickingParameters
peakpickingParameters <- getDefaultXcmsSetStartingParams('centWave')
peakpickingParameters$noise=1000
peakpickingParameters$value_of_prefilter=(3,800)
peakpickingParameters$min_peakwidth<- c(8,15)
peakpickingParameters$max_peakwidth<- c(30,40)
peakpickingParameters$ppm<- c(25,35)
param=SnowParam(workers = 5)
```

Once the parameters were selected, the optimization of the parameters set in a range was performed by IPO.

```
resultPeakpicking <-
  optimizeXcmsSet(files = datafiles,
                  params = peakpickingParameters,
                  BPPARAM = param,
                  nSlaves = 1,
                  subdir = NULL,
                  plot = TRUE)
```

The IPO optimization is based on Design of Experiments. The results were shown both numerically and graphically in contour graphs.

```
#best parameter settings:
#  min_peakwidth: 12.48
#max_peakwidth: 35
#ppm: 24
#mzdiff: 0.00175
#snthresh: 10
#noise: 1000
#prefilter: 3
#value_of_prefilter: 800
#mzCenterFun: wMean
#integrate: 1
#fitgauss: FALSE
#verbose.columns: FALSE
```



Analogously to the peak picking optimization, the retention time alignment and grouping parameters were also optimized in order to be applying in XCMS package. The optimized parameters for retention time alignment using the “obiwarp” method were *profStep*, *response*, *gapInit* and *gapExtend*, and for correspondence using “density” method were *bw* and *mzwid*. In our study, the parameter *minfrac* was fixed at value of 0.5.

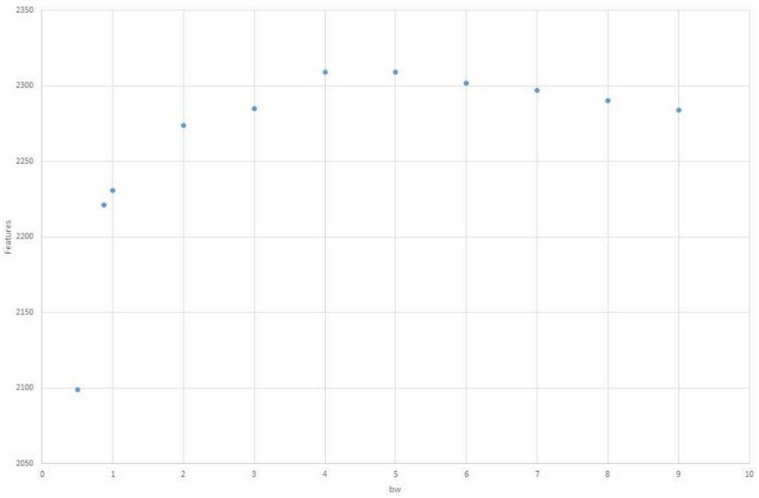
```
optimizedXcmsSetObject <- resultPeakpicking$best_settings$xset

retcorGroupParameters <- getDefaultRetGroupStartingParams()
retcorGroupParameters$profStep <- c(0.33,1)
retcorGroupParameters$gapExtend <- c(2.1,2.9)
retcorGroupParameters$minfrac=0.5
retcorGroupParameters$response=c(9.6,17.6)
retcorGroupParameters$gapInit=c(0.14, 0.54)
retcorGroupParameters$mzwid=c(0.023, 0.047)

BiocParallel::register(BiocParallel::SerialParam())

resultRetcorGroup <-
  optimizeRetGroup(xset = optimizedXcmsSetObject,
                   params = retcorGroupParameters,
                   nSlaves = 1,
                   subdir = NULL,
                   plot = TRUE)
```

The results were obtained by the following code and these are shown below:

```
writeRScript(resultPeakpicking$best_settings$parameters, resultRetcorG
roup$best_settings)
```

```
xset <- retcor(
xset,
method         = "obiwarp",
plottype       = "none",
distFunc       = "cor_opt",
profStep       = 0.3,
center         = 6,
response       = 13.84,
gapInit        = 0.352,
gapExtend      = 2.436,
factorDiag     = 2,
factorGap      = 1,
localAlignment = 0)
xset <- group(
xset,
method   = "density",
bw       = 0.879999999999999,
mzwid    = 0.047,
minfrac  = 0.5,
minsamp  = 1,
max      = 50)
```

Despite the optimization made by IPO, the obtained value for *bw* was considered very low. For this reason, this parameter was optimized manually. Different tests were performed with different values of bw (from 0.5 to 10). The optimal parameter of bw was chosen when the greatest number of features was obtained. In our case, it was obtained at 5.0 as it is showed in the next figure.



3) XCMS

Before using XCMS, we imported all data in R analogously to the importation performed previously with the 6 QC samples for the IPO optimization. Thus, we pasted the .mzML files of all the samples in a folder called “PlasmaData” within the workspace and run the following code.

```
datafiles <- list.files("PlasmaData", recursive = TRUE, full.names = T
RUE, pattern=".mzML")
```

The samples were divided in two subfolders in order to create a phenodata data.frame depending on if they are cases or QC samples.

```
# Create a phenodata data.frame
pd <- data.frame(sample_name = sub(basename(datafiles), pattern = ".mz
ML",
                                   replacement = "", fixed = TRUE),
                 sample_group = c(rep("Case", 306), rep("QC", 63)),
                 stringsAsFactors = FALSE)
```

The peak picking, retention time correction, and feature correspondence were performed using XCMS with the parameters optimized in the previous step with IPO.

```
raw_data <- readMSData(files = datafiles, pdata = new("NAnnotatedDataF
rame", pd), mode = "onDisk")
```

#PeakPicking Step

```
cwp  <-  CentWaveParam(peakwidth  =  c(12.48,  35),  ppm  =  24,  mzdiff  =
0.00175, snthresh = 10, noise = 1000, prefilter = c(3,800))
xdataL1 <- findChromPeaks(raw_data, param = cwp)

[...Detecting chromatographic peaks in 2519 regions of interest ... OK
: 1755 found.
Detecting mass traces at 24 ppm ... OK
Detecting chromatographic peaks in 2574 regions of interest ... OK: 18
25 found.
Detecting mass traces at 24 ppm ... OK
Detecting chromatographic peaks in 2634 regions of interest ... OK: 18
49 found. ...]
```

#Grouping and RT Correction

```
#RT Correction

BiocParallel::register(BiocParallel::SerialParam())
xdataL2 <- adjustRtime(xdataL1, param = ObiwarpParam(gapInit = 0.352,
gapExtend = 2.436, response = 13.84, binSize = 0.3))

[...Sample number 185 used as center sample.
Aligning 0601429.mzML against 3920435.mzML ... OK
Aligning 0610035.mzML against 3920435.mzML ... ]
```

```
#Correspondence
pdp    <-    PeakDensityParam(sampleGroups    =    xdataL2$sample_group,
minFraction = 0.5, bw = 5, binSize = 0.047)

xdataL3 <- groupChromPeaks(xdataL2, param = pdp)
Processing 67195 mz slices ... OK
```

The next step is to look for the peaks which were assigned as missing values (NA) in some samples because the peak detection algorithm was not able to find them. Therefore, this step of "filling peaks" is performed with the following code.

```
xdataL4 <- fillChromPeaks(xdataL3)
[Defining peak areas for filling-in .... OK
Start integrating peak areas from original files
Requesting 488 missing peaks from 0601429.mzML ... got 394.
Requesting 622 missing peaks from 0610035.mzML ... got 510.
Requesting 553 missing peaks from 0611111.mzML ... got 393.
...]
```

Despite this new step, there are still missing values in the samples for different reasons (e.g. not present in the biological samples, very low concentrations below the detection limits, algorithm failures). It is shown below the total number of missing values of all samples before and after the filling step.

```
#Number of missing values before filling step
xdataL3b = featureValues(xdataL3, value = "into")
sum(is.na(xdataL3b))
[1] 136846
#Number of missing values after filling step
xdataL4b = featureValues(xdataL4, value = "into")
sum(is.na(xdataL4b))
[1] 43476
```

Once these steps have been performed, we proceed to extract the integrated areas and the information of the peaks (mass and retention times) with the following codes. In this step, the features are named according to their mass and retention times using the "featureDefinitions" function as follows.

```
xdataL5 = featureValues(xdataL4, value = "into")
xdataL5= t(xdataL5)
featData=featureDefinitions(xdataL4)
featData=featData@listData
featNames=paste0(featData$mzmed,"_",featData$rtmed)
colnames(xdataL5)=featNames
```

For example, the first feature (mz: 100.075826364437, rt: 273.94580078125 s) was called "100.075826364437_273.94580078125".

Missing value imputation

Imputation of values still missing after XCMS peak filling was performed using an in-house script based on RandomForest methodology (https://gitlab.com/CarlBrunius/StatTools; mvImpWrap() function). For it, it is necessary to install firstly the *StatTools* package.

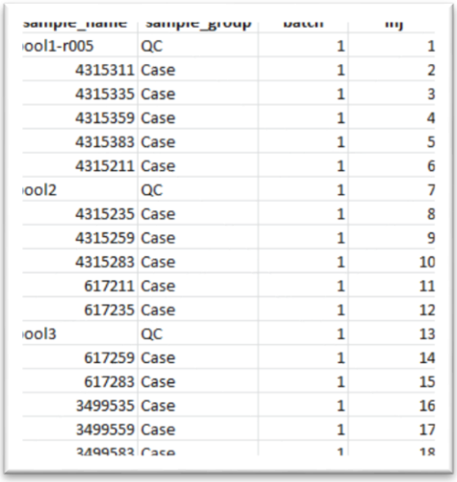
```
#Installation of package StatTools
devtools::install_git("https://gitlab.com/CarlBrunius/StatTools.git")

#Missing Value Imputation
nCore=detectCores()-1
cl=makeCluster(nCore)
registerDoParallel(cl)
Imp <- StatTools::mvImpWrap(MAT = xdataL5, method = "RF")
xdataL5 <- Imp
stopCluster(cl)

#Number of missing values after missing value imputation
sum(is.na(xdataL5))
[1] 0
```

4) BatchCorr

The BatchCorr package aims to normalize the data due to between-batch and within-batch drifts produced in LC-MS analysis. This normalization strategy is based on the QC samples, the batches, and the injection order. For these reasons, it is necessary to introduce this information (injection, batch, and QC/Case) in the R environment.

Previously, a dataframe called “pd” was created with the information of the kind of sample (QC or Case sample). Therefore, we needed to add in this dataframe two columns to indicate the injection position and the batch of each sample.

First of all, it is necessary to read a .csv file with the information of injection order and batch of each sample. This file (in our case is named “pdinjbatch.csv”) have to be located in the workspace.



sample_name	sample_group	batch	inj
ool1-r005	QC	1	1
4315311	Case	1	2
4315335	Case	1	3
4315359	Case	1	4
4315383	Case	1	5
4315211	Case	1	6
ool2	QC	1	7
4315235	Case	1	8
4315259	Case	1	9
4315283	Case	1	10
617211	Case	1	11
617235	Case	1	12
ool3	QC	1	13
617259	Case	1	14
617283	Case	1	15
3499535	Case	1	16
3499559	Case	1	17
3499583	Case	1	18

```
pd_injbatch = read.csv(file='pd_injbatch.csv', head = TRUE, sep = ',')
```

Once this file was loaded in the R environment, the columns “inj” and “batch” were added to the dataframe *pd*. For this step, it is needed to make sure that the samples are sorted in both files (pd and pd_injbatch) in the same way. Below are the instructions to carry out this stage:

```
# Check the identical order of the samples.
rownames(pd) = pd$sample_name
rownames(pd_injbatch) = pd_injbatch$sample_name
identical(rownames(pd), rownames(pd_injbatch))
[1] TRUE

# Add the “inj” and “batch” columns to the dataframe pd.
pd<- data.frame(pd, inj= pd_injbatch$inj)
pd<- data.frame(pd, batch= pd_injbatch$batch)
```

To perform the *batchCorr* package correctly is required that the samples are sorted by their injection order. If they are not ordered in the previous files in this way, it is mandatory to apply the following commands.

```
xdataL5Sort= xdataL5[order(pd$inj),]
pdSort = pd[order(pd$inj),]
```

The normalization by BatchCorr is performed by means of three steps: 1: between-batch correspondence/alignment. 2: within-batch intensity drift correction and 3: between-batch normalization. In the following lines are detailed how we used this package in our data. To better understanding the scripts and functions, we strongly recommend that you read the tutorial of this package (https://gitlab.com/CarlBrunius/batchCorr/tree/master/Tutorial)

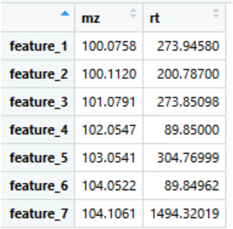
In order to carry out these 3 steps, we needed these 3 objects: a) dataframe (pd) with information about batches, sample groups (QC/case) and injection orders; b) peak table without missing values (xdataL5Sort); c) peak table with missing values obtained before filling step. In order to obtain this last peak table (xdataL3bSort), we applied the following code:

```
xdataL3b = featureValues(xdataL3, value = "into")
xdataL3b= t(xdataL3b)
featData=featureDefinitions(xdataL4)
featData=featData@listData
featNames=paste0(featData$mzmed,"_",featData$rtmed)
colnames(xdataL3b)=featNames
xdataL3bSort=xdataL3b[order(pd$inj),]
```

Step 1. Between-batch correspondence/alignment

The aim of the first step is to align the features misaligned between batches. Firstly, it is fundamental to extract the retention times and masses from each feature using the function “peakInfo”. As our features were called, “mass_rt”, we applied the function as follows:

```
peakIn <- peakInfo(PT = xdataL3bSort, sep = '_', start = 1)
```



	mz	rt
feature_1	100.0758	273.94580
feature_2	100.1120	200.78700
feature_3	101.0791	273.85098
feature_4	102.0547	89.85000
feature_5	103.0541	304.76999
feature_6	104.0522	89.84962
feature_7	104.1061	1494.32019

Secondly, we used the "alignBatches" function to carry out the purpose of this first step.

```
alignBat <- alignBatches(peakInfo = peakIn, PeakTabNoFill = xdataL3bSo
rt, PeakTabFilled = xdataL5Sort, batches = pdSort$batch, sampleGroups
= pdSort$grp, selectGroup = 'QC')

IMPORTANT NOTIFICATION

There are no alignment candidates. Therefore,
  between-batch alignment is not possible.
```

```
Consider expanding mzdiff and/or rtdiff
  of that correspondence is accurate between batches.
Returning NULL.
```

This output indicates that no alignment candidates were found and, consequently, that no between-batch alignment was possible. This fact reflects that this stage was carried out correctly in the previous step with XCMS package.

Step 2. Within-batch intensity drift correction

The second step aims to normalize and correct the drifts produced throughout each batch. In addition, the features with a RSD higher than 30% in QC samples after the normalization were filtered.

We used the functions "getbatch" and "correctDrift" to extract each batch, and carry out the normalization, respectively.

```
batchB_F <- getBatch(peakTable = xdataL5Sort, meta = pdSort, batch = p
dSort$batch, select = '1')

batch1 <- correctDrift(peakTable = batchB_F$peakTable, injections = ba
tchB_F$meta$inj, sampleGroups = batchB_F$meta$sample_group, QCID = 'QC
', G = seq(5,35,by=3), modelNames = c('VVE', 'VEE'))#

    __  ___________    __  _____________
   /  |/  / ____/ /   / / / / ___/_  __/
  / /|_/ / /   / /   / / / /\__ \ / /
 / /  / / /___/ /___/ /_/ /___/ // /
/_/  /_/\____/_____/\____//____//_/    version 5.4.2
Type 'citation("mclust")' for citing this R package in publications.

Mclust fitting ...
  |===================================================================
======================================================================
=======| 100%

MClust final model with 14 clusters and VEE geometry.
BIC performed in 14.54 seconds and clustering in 0.91 seconds.

Calculation of QC drift profiles performed.

Drift correction of 12 out of 14 clusters using QC samples only.
Corrected peak table in $TestFeatsCorr

Filtering by QC CV < 0.3 -> 1205 features out of 1312 kept in the peak
table.
Peak table in $TestFeatsFinal, final variables in $finalVars and clust
er info in $actionInfo.

# Batch 2
batchA_F <- getBatch(peakTable = xdataL5Sort, meta = pdSort, batch = p
dSort$batch, select = '2')
```

```
batch2 <- correctDrift(peakTable = batchA_F$peakTable, injections = ba
tchA_F$meta$inj, sampleGroups = batchA_F$meta$sample_group, QCID = 'QC
', G = seq(5,35,by=3), modelNames = c('VVE', 'VEE'))#

Mclust fitting ...
  |===================================================================
======================================================================
=======| 100%

MClust final model with 14 clusters and VEE geometry.
BIC performed in 24.26 seconds and clustering in 1.25 seconds.

Calculation of QC drift profiles performed.

Drift correction of 13 out of 14 clusters using QC samples only.
Corrected peak table in $TestFeatsCorr

Filtering by QC CV < 0.3 -> 1110 features out of 1312 kept in the peak
table.
Peak table in $TestFeatsFinal, final variables in $finalVars and clust
er info in $actionInfo.

# Batch 3
batchH_F <- getBatch(peakTable = xdataL5Sort, meta = pdSort, batch = p
dSort$batch, select = '3')

batch3 <- correctDrift(peakTable = batchH_F$peakTable, injections = ba
tchH_F$meta$inj, sampleGroups = batchH_F$meta$sample_group, QCID = 'QC
', G = seq(5,35,by=3),modelNames = c('VVE', 'VEE'))


Mclust fitting ...
  |===================================================================
======================================================================
=======| 100%

MClust final model with 8 clusters and VEE geometry.
BIC performed in 10.59 seconds and clustering in 0.22 seconds.

Calculation of QC drift profiles performed.

Drift correction of 4 out of 8 clusters using QC samples only.
Corrected peak table in $TestFeatsCorr

Filtering by QC CV < 0.3 -> 701 features out of 1312 kept in the peak
table.
Peak table in $TestFeatsFinal, final variables in $finalVars and clust
er info in $actionInfo.
```

Step 3. Between-batch normalization

The third step aims to merge the drift-corrected data (mergeBatches() function) and normalize the effects between batches (normalizeBatches() function).

```
mergedData <- mergeBatches(list(batch1, batch2, batch3))
```

```
normData <- normalizeBatches(peakTable = mergedData$peakTable, batches
= pdSort$batch, sampleGroup = pdSort$sample_group, population = 'all')
```

Once these functions were carried out, we extracted the peakTable (features and intensities) and exported it in a .csv file.

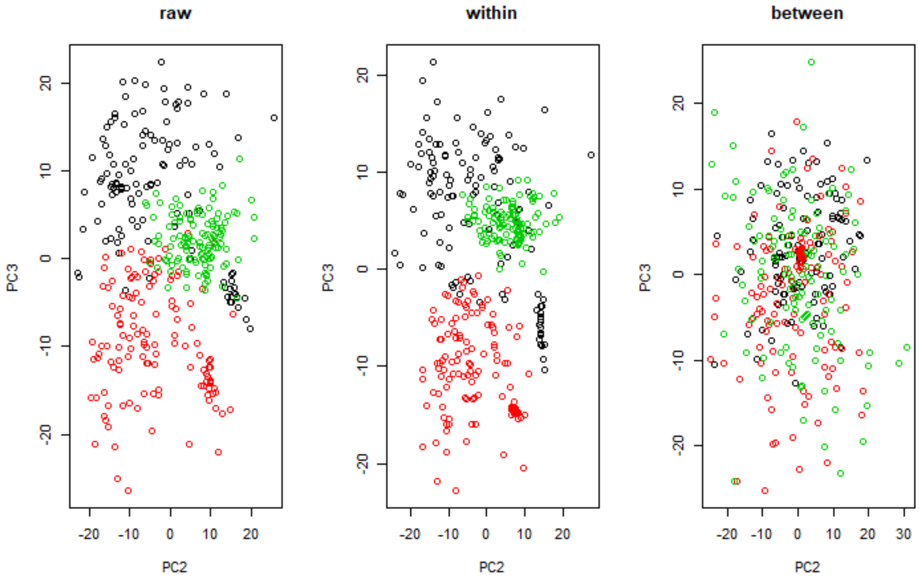
```
PTnorm <- normData$peakTable
write.csv(PTnorm, file="PTnorm_F.csv")
```

A good way to observe how the data has been transformed with this package is using Principal Component Analysis (PCA). In the following lines, it is shown the code for PCA analysis used for the three data sets (raw data, data after within-batch normalization step and data after between-batch normalization step).

```
pca1 <- prcomp(x = xdataL5Sort, center = T, scale. = T)
pca2 <- prcomp(x = mergedData$peakTable, center = T, scale. = T)
pca3 <- prcomp(x = PTnorm, center = T, scale. = T)
```

It is shown below the code to create the graphs that show the scores plots of each PCA model.

```
par(mfrow=c(1,3))
plot(pca1$x[,2:3], col=pdSort$batch, main = 'raw')
plot(pca2$x[,2:3], col=pdSort$batch, main = 'within')
plot(pca3$x[,2:3], col=pdSort$batch, main = 'between')
```



5) RAMClust

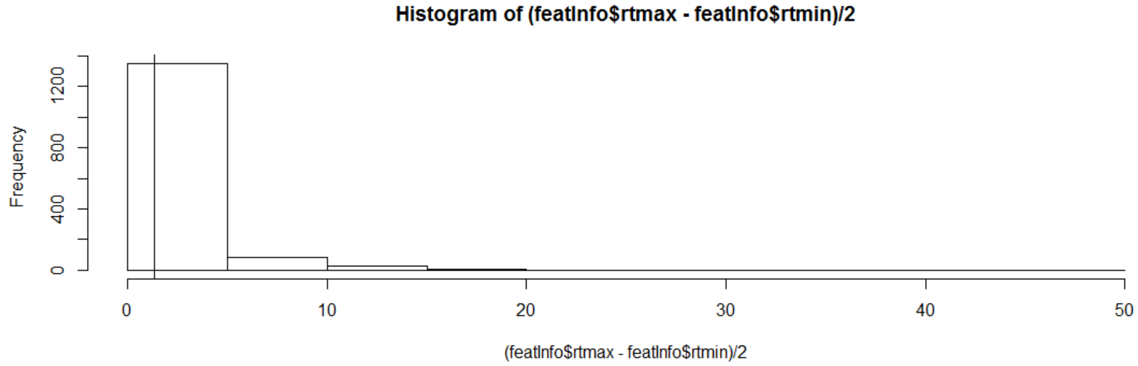
The aim of this package is to perform the annotation step. Briefly, it consists of grouping all MS signals related to a single compound (isotopes, adducts, isomers, fragments products, etc.) into a single file called "feature". To perform this step, the package RAMClust is mainly based on two parameters: (i) similarity in their retention times (st) and (ii) correlation in the abundances across different samples (sr).

```
library(RAMClustR)
```

Unlike the IPO package that allows optimizing the parameters for XCMS, there is still no package to optimize the mentioned RAMClust parameters, st and sr. So, we created our own methods and functions to optimize these values and achieve a better result of the annotation. In this way, the functions *getRamSt* and *plotClust* were created (see **Annex**).

The *getRamSt* function aims to obtain the optimal value of the **st** parameter.

```
getRamSt(xdataL4)
[1] 1.33
```



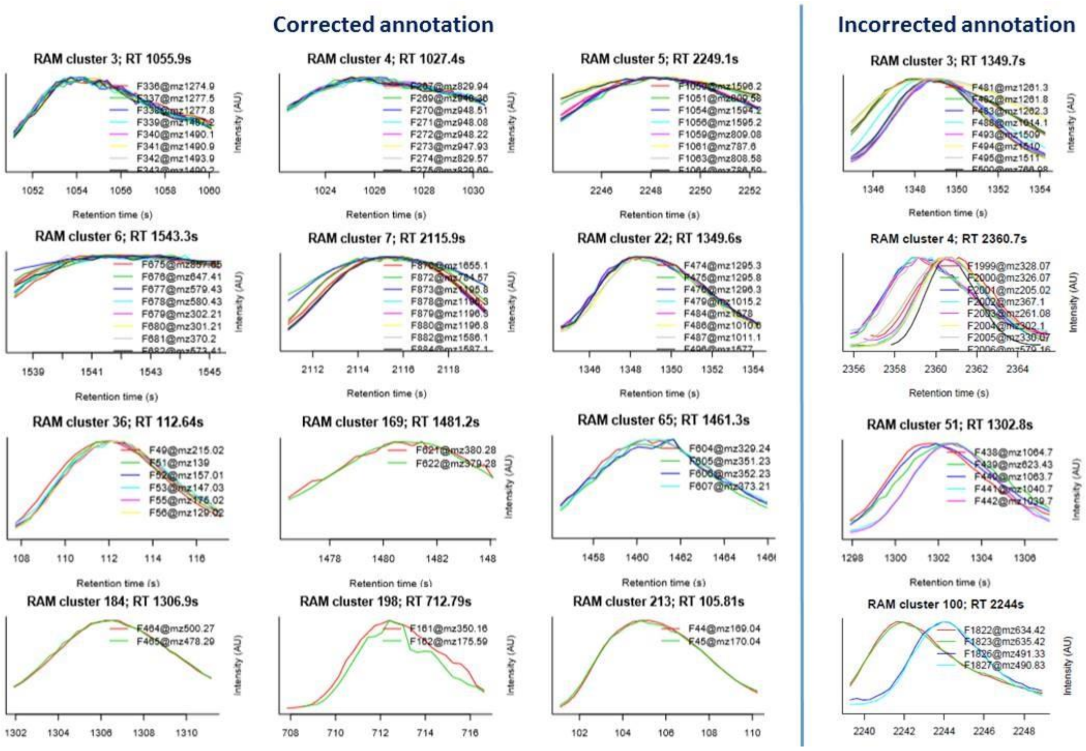
Once the st parameter was optimized, we applied the `ramclustR` function with this st value and a range of sr parameter (sr = 0.3, 0.4, 0.5). In this script, we used the plotClust function to obtain the plots of 20 features in order to observe the groupings of the signals and obtain the optimal sr value.

```
expDes=defineExperiment(force.skip = T)
sr=c(.3,.4,.5)
st=1.33
```

```
maxt=c(5)
par=expand.grid(st=st,sr=sr,maxt=maxt)
str(par)
'data.frame':   3 obs. of  3 variables:
 $ st  : num  1.33 1.33 1.33
 $ sr  : num  0.3 0.4 0.5
 $ maxt: num  5 5 5
 - attr(*, "out.attrs")=List of 2
  ..$ dim     : Named int  1 3 1
  .. ..- attr(*, "names")= chr  "st" "sr" "maxt"
  ..$ dimnames:List of 3
  .. ..$ st  : chr "st=1.33"
  .. ..$ sr  : chr  "sr=0.3" "sr=0.4" "sr=0.5"
  .. ..$ maxt: chr "maxt=5"

nClust=nSing=sizeMax=sizeMed=sizeMean=numeric(nrow(par))
nFeat=list()
samps=sample(1:205,20)
register(bpstart(SnowParam(7))) # Choose as many cores as you can/want
for (i in 1:nrow(par)) {
  RRP=ramclustR(ms='PTnorm_F.csv', st = par$st[i], sr=par$sr[i], maxt
= par$maxt, timepos = 2, sampNameCol = 1, featdelim = '_', ExpDes = ex
pDes)
  nClust[i]=length(RRP$cmpd)
  nSing[i]=RRP$nsing
  sizeMax[i]=max(RRP$nfeat)
  sizeMed[i]=median(RRP$nfeat)
  sizeMean[i]=mean(RRP$nfeat)
  nFeat[[i]]=RRP$nfeat
  pdf(file=paste0('clusts_par',i,'.pdf'),width=15,height=8)
  par(mfrow=c(4,5),mar=c(4,4,2,0)+.5)
  clusts=round(c(2:6,seq(7,max(RRP$featclus),length.out = 15)))
  for (c in clusts) {
    plotClust(ram = RRP,clustnr = c,xcmsData = xdataL4,samps = samps)
  }
  dev.off()
}
```

Then, the results were obtained for each combination of values. To choose the optimal sr value, we visually inspected the generated files. In the following figure, examples of correct or incorrect annotation cases are shown. In this way, the optimized sr value was chosen where the greatest number of features with a correct annotation was obtained, (a value of 0.3 was obtained in our case).

```
cbind(par,nClust,nSing,sizeMax,sizeMean,sizeMed)
    st  sr maxt nClust nSing sizeMax sizeMean sizeMed
1 1.33 0.3    5    213   306      16 3.778302       3
2 1.33 0.4    5    213   257      17 4.009434       3
3 1.33 0.5    5    204   217      19 4.384236       3
```



Finally, we applied the ramclustR function with the optimal values of st and sr.

```
RC1 <- ramclustR(ms='PTnorm_F.csv',
                 featdelim = "_",
                 st = 1.33,
                 sr = 0.3,
                 ExpDes=expDes,
                 sampNameCol = 1)
  organizing dataset
  normalizing dataset
calculating ramclustR similarity: nblocks =  1
1  RAMClust feature similarity matrix calculated and stored:
RAMClust distances converted to distance object
fastcluster based clustering complete
dynamicTreeCut based pruning complete
RAMClust has condensed 1109 features into 213 spectra
collapsing feature into spectral signal intensities
writing msp formatted spectra
msp file complete


RC1 <- do.findmain(RC1, mode = "positive", mzabs.error = 0.02, ppm.err
or = 10)
10 of 213
20 of 213
30 of 213
40 of 213
50 of 213
60 of 213
70 of 213
80 of 213
90 of 213
100 of 213
```

```
110 of 213
120 of 213
130 of 213
140 of 213
150 of 213
160 of 213
170 of 213
180 of 213
190 of 213
200 of 213
210 of 213
plotting findmain annotation results
finished
```

6) Export Data.

The last step of pre-processing corresponds to the export of data to use them in statistical analyses. We exported both the features annotated by RAMClustR and the signals that were not assigned to any cluster (singletons). In the following lines it is showed how the exportation was performed. The tidyverse package is necessary.

```
library(tidyverse)
```

#Look for the ion with the highest intensity in each group obtained by RAMClustR

```
Max_int<- lapply(RC1$M.ann, function(x) x[which.max(x$int), ])
Molecular_ions <- bind_rows(Max_int)
Molecular_ions$name <- paste(round(Molecular_ions$mz,4), round(RC1$clr
t,2), sep = "_")
PTnorm_F2_mz <- colnames(PTnorm_F2)[-1] %>% str_split(.,"\\_") %>% lap
ply(.,function(x) x[1]) %>% unlist() %>%as.numeric()
```

#check if it works for your data with the following table. The number of TRUES should be the same to the cluster number.

```
PTnorm_F2_mz %in% Molecular_ions$mz %>% table()
FALSE  TRUE
  896   213

PTnorm_F2_molecularIons_rawInt<- PTnorm_F2[,-1][,PTnorm_F2_mz %in% Mol
ecular_ions$mz]
```

#Look for the singletons

```
clustered_mz<- lapply(RC1$M.ann, function (x) x$mz) %>% unlist() %>% a
s.numeric()
singletons <- PTnorm_F2[,-1][,!PTnorm_F2_mz%in%clustered_mz]
```

#Combine singletons and molecular ions

```
PTnorm_F2_ramclust_annotedPT <- cbind.data.frame(PTnorm_F2_molecularIons_rawInt, singletons)

rownames(PTnorm_F2_ramclust_annotedPT) <- paste0("PTnorm_F2",rownames(PTnorm_F2_ramclust_annotedPT))

colnames(PTnorm_F2_ramclust_annotedPT) <- paste0("PTnorm_F2",colnames(PTnorm_F2_ramclust_annotedPT))

write.csv(PTnorm_F2_ramclust_annotedPT, file = "PTnorm_F2_ramclust_annotedPT0920.csv")
```

Finally, a .csv file (samples vs intensities of each feature) was obtained in the working directory. This can be imported into different statistical software to perform the corresponding tests that are desired.

We hope this guide allows you to use R in the same way that we have used it in our work.

Good luck and thanks for your interest in this tutorial.

Álvaro Fernández and Carl Brunius

PS. Please don’t hesitate to contact us if you have problems with the codes and functions or any suggestions or questions.

alvaroferochoa@ugr.es
carl.brunius@chalmers.se

Annex

getRamSt:

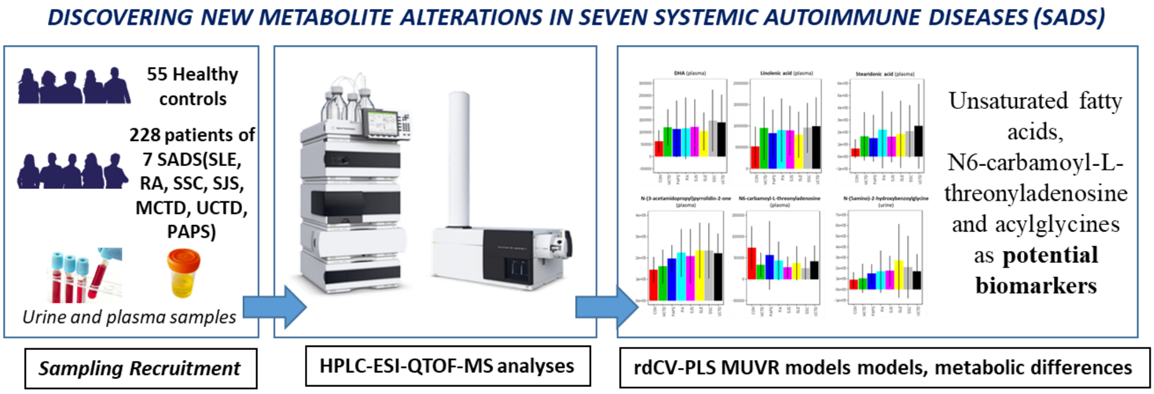
```
getRamSt <- function(XObj) {
  featInfo <- featureDefinitions(XObj)
  hist((featInfo$rtmax-featInfo$rtmin)/2)
  st <- round(median(featInfo$rtmax-featInfo$rtmin)/2, digits = 2)
  abline(v=st)
  return(st)
}
```

plotClust:

```
plotClust=function(ram,clustnr,xcmsData,samps,dtime=5,dmz=.05) {
  if(missing(samps)) {
    nSamp=nrow(ram$SpecAbund)
    samps=1:nSamp
  } else nSamp=length(samps)
  whichFeats=which(ram$featclus==clustnr)
  peakMeta=cbind(ram$fmz,ram$frt)
  pkMetaGrp=peakMeta[whichFeats,]
  rtr=ram$clrt[clustnr]+c(-dtime,dtime)
  rtr[rtr<0]=0
  mzr=cbind(ram$fmz[whichFeats]-dmz,ram$fmz[whichFeats]+dmz)
  chr <- chromatogram(xcmsData, mz = mzr, rt = rtr)
  plot(0:1,0:1,type='n',axes=F,xlab='Retention time (s)', ylab='Intens
ity (AU)',main=paste0('RAM cluster ',clustnr,'; RT ',signif(ram$clrt[c
lustnr],5),'s'))
  box(bty='l')
  for (pk in 1:length(whichFeats)) {
    rts=ints=list()
    for (samp in 1:nSamp) {
      rts[[samp]]=chr[pk,samps[samp]]@rtime
      ints[[samp]]=chr[pk,samps[samp]]@intensity
    }
    nrts=min(sapply(rts,length))
    rts=sapply(rts,function(x) x[1:nrts])
    rts=rowMeans(rts)
    ints=sapply(ints,function(x) x[1:nrts])
    ints=rowMeans(ints,na.rm=T)
    par(new=T)
    plot(rts,ints,type='l',col=pk+1,ylim=c(0,max(ints,na.rm=T)),axes=F
,xlab='',ylab='')
  }
  axis(1)
  legend('topright',legend = paste0('F',whichFeats,'@mz',signif(pkMeta
Grp[,1],5)), lty=1,col=(1:length(whichFeats))+1,bty='n')
```

Capítulo 7

Detección de alteraciones metabólicas en muestras de orina y plasma de pacientes de siete enfermedades autoinmunes sistémicas mediante HPLC-ESI-QTOF-MS



DISCOVERING NEW METABOLITE ALTERATIONS IN SEVEN SYSTEMIC AUTOIMMUNE DISEASES (SADS)

55 Healthy controls

228 patients of 7 SADS(SLE, RA, SSC, SJS, MCTD, UCTD, PAPS)

Urine and plasma samples

Unsaturated fatty acids, N6-carbamoyl-L-threonyladenosine and acylglycines as **potential biomarkers**

Sampling Recruitment

HPLC-ESI-QTOF-MS analyses

rdCV-PLS MUVR models models, metabolic differences

DHA (plasma)
Linolenic acid (plasma)
Stearidonic acid (plasma)
N-(3-acetamidopropyl)pyrrolidin-2-one (plasma)
N6-carbamoyl-L-threonyladenosine (plasma)
N-(5amino)-2-hydroxybenzoylglycine (urine)
CON
MCTD
PAPS
RA
SJS
SLE
SSC
UCTD

Álvaro Fernández-Ochoa, Isabel Borrás-Linares, Rosa Quirantes-Piné, PRECISESADS Clinical Consortium, Marta E. Alarcón Riquelme, Carl Brunius, Antonio Segura-Carretero

Metabolic disturbances in urinary and plasma samples from seven different systemic autoimmune diseases detected by HPLC-ESI-QTOF-MS

ABSTRACT

Systemic autoimmune diseases (SADs) are characterized by failures of the immune system causing multiple damages in several tissues, and organs of the organism. Several of these pathologies are systemic lupus erythematosus, systemic sclerosis, Sjögren`s syndrome, rheumatoid arthritis, antiphospholipid syndrome, mixed connective tissue disease and undifferentiated connective tissue disease (UCTD). Currently, there are great difficulties in their diagnosis due to the combination of similar symptoms and signs. Hence, the aim of this work has been improve the knowledge of these seven diseases through the search for differentiating metabolites in biological samples. For this purpose, a preliminary fingerprinting-based metabolomic study was carried out by LC-MS analyses of plasma and urinary samples from 228 SADs patients and 55 healthy volunteers. Afterwards, multivariate statistical models were applied in order to find metabolic differences between different diseases and healthy controls. In addition, this kind of statistical model was also performed to compare UCTD against its related diseases, due to the great difficulties to define this particular disease. The results showed acceptable classifications (AUC>0.7) in the models that compared the different diseases against healthy controls. The families of metabolites that showed a greater contribution in these classificatory models were unsaturated fatty acids, acylglycines, acylcarnitines and amino acids. However, the specific models focused on UCTD showed limitations to differentiate this disease from its related pathologies. These results are in accordance with the current difficulties in defining this type of disease. Therefore, further studies integrating different "omics" approaches are needed for a better understanding of the complex processes involved in the systemic autoimmune diseases.

Keywords. metabolomics, untargeted, mass spectrometry, biomarkers, systemic autoimmune diseases, multivariate models

1. Introduction

Autoimmune diseases are characterized by the attack of the immune system to the healthy cells and tissues of the own organism by mistake [1]. The origin and causes of these diseases are not fully known in most cases, although they appear to be related to genetic, environmental, hormonal or infectious factors [2]. This kind of diseases are mainly classified into the following two categories: organ-specific and systemic [3,4]. The organ-specific autoimmune diseases are characterized by the attack of the immune system focused on the antigens expressed in a single organ. However, systemic autoimmune diseases (SADs) are distinguished by failures of the immune system which turns against several tissues and organs of the organism with no apparent relationship between them [5,6].

SADs affects approximately 1% of the world population and, in general, there are few treatment options for them as well as great difficulties in their diagnosis due to the combination of similar symptoms and signs [5]. These diseases are mainly represented by systemic lupus erythematosus (SLE), systemic sclerosis (SSC) and rheumatoid arthritis (RA). Other SADs have extensive clinical overlap with these diseases causing major complications in their diagnosis, such as Sjögren's syndrome (SjS), mixed connective tissue disease (MCTD), primary antiphospholipid syndrome (PAPS) or undifferentiated connective tissue disease (UCTD) [7–9]. MCTD and UCTD represent the clearest examples of the diagnostic difficulties of this set of diseases. MCTD shows symptoms and signs present in other SADs diseases, such as SSC, polymyositis and SLE, although is defined by having a clinically distinct entity [10,11]. However, UCTD encompasses cases that present symptoms and signs characteristic of others SADs, such as SLE, SjS, RA or SSC, but do not completely satisfy the condition of one of them. According to the UCTD evolution over time, one third of the cases are finally diagnosed as a specific SADs [12,13].

In the last decade, omic sciences, including genomics, transcriptomics, proteomics, metabolomics or microbiomics, are allowing to know better the biological mechanisms and the pathophysiology of the SADs. In this sense, these technologies have great potential to identify biomarkers in order to improve their diagnosis, prognosis, treatments and even reclassify this set of diseases. [14]

Specifically, metabolomics is the end point of the omics' cascade showing the greatest relationship with the phenotypic response. Metabolomics is focused on studying all low molecular weight molecules present in biological systems in order to find alterations and interactions in the organism due to different conditions or causes [15]. During the last decade, metabolomic studies have been increased thanks to the development of analytical platforms, in particular Nuclear Magnetic Resonance spectroscopy (NMR) and Mass Spectrometry (MS), which is usually used coupled to liquid or gas chromatography [16]. In consequence, several research have recently focused on the study of SADs based on metabolomic strategies. These have mainly studied the most common and predominant SADs, such as SLE, RA, SSC or SjS [4,17–19]. In contrast, there is a lack of metabolomic studies of SADs with more difficulties for diagnosis, such as MCTD or UCTD.

In general, metabolomic studies of SADs have mainly analyzed serum or plasma samples in order to find biomarkers and to increase the knowledge of the pathogenesis. In contrast, studies using urine or feces have not practically been carried out in spite of the advantages of this type of biological samples, such as the large amount of represented metabolites, as well as their non-invasive sample collection [20,21]. In addition, several autoimmune connective tissue diseases, such as SLE, SjS, SSC or RA, show kidneys damage or renal manifestations [22]. Therefore, urine samples present a great potential to offer valuable information about these pathological processes.

The aim of this research is to identify metabolic disturbances in urine and plasma samples of seven SADs (SSC, SjS, SLE, RA, PAPS, MCTD and UCTD) in order to increase the knowledge of their pathogenesis and the biochemical mechanisms involved. For this purpose, an untargeted metabolomic strategy was carried out based on LC-MS. Multivariate statistical models were created with the objective of detecting metabolic differences between the different SADs and the healthy controls (HC) as well as between UCTD and its related SADs.

2. Material and Methods

2.1. *Biological samples*

Biological samples were collected from volunteers participating in the PRECISESADS project (www.precisesads.eu). Specifically, plasma and urinary samples from 228 patients with SADs (47 RA, 46 SLE, 46 SjS, 43 SSC, 22 UCTD, 14 MCTD, 10 PAPS) and 55 healthy people (HC) were collected. The main information about the volunteers (age, gender or duration of the diseases) are collected in the **Table 1S (Supplementary Material).** Plasma samples were obtained after treatment of blood with K_2EDTA and centrifugation (1500g, 10 min). Single urine samples were treated with HCl (1M) and centrifuged. Both kind of samples were stored at -80 ºC until treatment. This metabolomic analysis is ancillary to the written informed consent that was obtained from each participant of the study, which was registered on clinicaltrials.gov with the code NCT02890121.

2.2. *Sample treatment*

100 µl of plasma sample was treated with organic solvents (ethanol:methanol, 50:50, v/v) in a 1:2 ratio in order to precipitate the proteins [23]. After that, the mixture was kept at -20 ºC (30 min) and centrifuged (14800, 4ºC, 10 min). The supernatant was evaporated and reconstituted with 80 µl of 0.1% water (0.1% formic acid) and methanol (95:5, v/v) and, an aliquot was transferred into vials and kept at -80 ºC until the HPLC-MS analysis. On the other hand, urine samples were diluted with H_2O- Milli-Q in order to obtain the same osmolarity value of 100 mOsm/Kg in all samples in order to correct the hydration differences among volunteers [24]. After that, the sample was centrifuged (14800, 4ºC, 10 min) and, an aliquot was transferred into vials and kept at -

80 ºC until analysis. A quality control (QC) sample was made up of equal amounts of all the study samples, for urine and plasma separately. QC samples were treated as detailed above.

2.3. HPLC-ESI-QTOF-MS analysis

The LC-MS methodology was optimized and used by our research group in a previous work carried out with a smaller number of samples of SSC (Fernández-Ochoa et al., 2019). The metabolites were separated using a 1260 HPLC instrument (Agilent Technologies, Palo Alto, CA, USA), in reverse phase mode using a C18 column. Detection by MS was carried out in positive-ion mode with a 6540 UHD Accurate Mass Q-TOF (Agilent Technologies, Palo Alto, CA, USA) equipped with a Jet Stream dual ESI interface. Briefly, the main parameters of the analytical methodology are listed in **Table 1**. Due to the large number of samples, analyses were performed on three and six different batches for plasma and urine samples, respectively. QC samples were analysed every 6 samples of the study in order to control the analytical reproducibility. In addition, this sample was analysed five times at the beginning of the batches in order to stabilize the chromatographic and detection conditions. Finally, MS/MS analyses were performed in QC samples to obtain the characteristic fragments that help in the task of identifying the metabolites.

Table 1. Experimental parameters of the liquid chromatographic and mass spectrometry conditions for the untargeted metabolomics analysis of plasma and urine samples.

Parameters		Plasma	Urine
HPLC conditions	Column	Agilent Zorbax Eclipse Plus, 3.5 µm, 2.1×150 mm	
	Mobile Phase A	H_2O- Milli-Q containing 0.1% of formic acid	
	Mobile Phase B	Methanol	
	HPLC gradient	0 min 95 % (A) 5 min 90 % (A) 15 min 15 % (A) 32-40 min 0 % (A) 45-50 min 95 % (A)	0 min 95 % (A) 30 min 70 % (A) 40 min 0 % (A) 50-60min 95 % (A)
	Column T	25 ºC	25 ºC
	Autosampler T	4 ºC	4 ºC
	Flow rate	0.4 mL/min	0.4 mL/min
	Injection Volume	5 µl	3 µl
MS conditions	Reference masses	Purine (m/z 121.050873)	
		hexakis(1H, 1H, 3H-tetrafluoropropoxy) phosphazine or HP-921 (m/z 922.009798)	
	Mass range	100 – 1700 m/z	50 – 1700 m/z
	Drying gas	Ultrahigh pure N_2 (200 ºC, 10 L/min)	
	Nebulizer gas	Ultrahigh pure N_2 (350 ºC, 12 L/min)	
	Collision energies (MS/MS)	10, 20 and 40 eV	

2.4. *Data processing*

Data were preprocessed using an open-source methodology by the combination of different packages developed in R (version 3.5.1) [25]. First, data were converted to .mzML file format using *MSConvertGUI* tool [26]. *IPO* [27] was the first package of the pipeline in order to optimize the parameters to carry out the steps of peak picking, alignment and grouping with the *XCMS* package [28]. The optimized parameters are shown in the **Table 2S** (**Supplemental Material**). A *filling peak* step was also carried out with XCMS to find missing values (NA) obtained in previous steps. In this steps, signals with a percentage of NA among all study samples (SADs and HC) higher than 30 %, were filtered in order to discard exogenous metabolites related to treatments. After that, the remaining missing values were imputed using an in-house script based on

RamdomForest (https://gitlab.com/CarlBrunius/StatTools; mvImpWrap() function). The next package used was *batchCorr* [29] in order to normalize the between-batch and within-batch effects, as well as to filter the features with a relative standard deviation (RSD) higher than 30% in QC samples. The last package used for data processing was *RamClustR* [30], with the aim of grouping the corresponding features to the same metabolite. This package is based on the two following parameters: similarity in retention time (σ_t) and similarity in intensity (σ_r) between samples. The similarity parameters (σ_t, σ_r) were set at (1.33, 0.3) and (2.50, 0.3) for plasma and urine, respectively.

2.5. Statistical analysis

First, principal component analyses (PCA) were performed in order to ensure the quality of the data. For a better visualization of the PCA scores plots, data were imported and analysed in MetaboAnalyst 4.0 (https://www.metaboanalyst.ca/) [31]. Supervised multivariate analysis were performed using *MUVR* package [32] within the R environment. This algorithm performed a minimum variable selection through Partial least squares regression (PLS) models by a repeated double cross-validation procedure. PLS models were created with the seven SADs and controls. Models were also created using two groups (healthy controls vs particular diseases), to obtain a better vision of the metabolic alterations in the particular diseases. In addition, MUVR-PLS models that compare UCTD with its related ones (SLE, RA, SSC or SjS) were also created. For PLS models, variables were scaled to unit variance. As the number of patients in each group were different, the MUVR models took into account a number of samples compensated among all the groups to avoid overfitting. These compensated rdCV-PLS models were repeated ten times with different samples. The significant variables present in all repeated models were selected for identification. To ensure the validity of the models as well as the degree of overfitting, permutation analysis were performed.

2.6. Metabolite identification

The identification was carried out through the comparison of the accurate mass, isotopic distribution and fragmentation patterns obtained in MS/MS analysis with the online available metabolomic databases (METLIN, LipidMaps, KEGG, HMDB). CEU Mass Mediator tool was used to facilitate the search of the metabolites in the database allowing the simultaneous search in several databases [33]. The MS/MS patterns were also compared with MS/MS fragmentation resources, concretely Sirius [34] and MetFrag (http://msbi.ipb-halle.de/MetFrag/).

3. Results and discussion

As mentioned before, the main aim of the present work was to look for metabolic alterations in urine and plasma samples of seven SADs. To our knowledge, this is the first investigation that deepens the metabolic alterations in urine and/or plasma of seven autoimmune disease in a single study. Metabolomic data were used to create PLS models in order to discriminate pathological cases with respect to healthy control. In the same way, UCTD cases were compared with its related pathologies, such as SLE, SjS, SSC and RA. In the following subsections, the results of the PLS models are detailed together with the discussion of the most relevant metabolites responsible for the discrimination of these models.

3.1. Data quality assessment

After all the data processing steps, a total of 531 and 825 features were obtained for plasma and urine samples, respectively. Before statistical analysis, the quality of the data derived from the metabolomic analysis was evaluated based on the distribution of the QC samples in the PCA analyses. **Figure 1**(**a-b**) shows the results of the PCA scores obtained with the raw data from urine and plasma samples. As shown in the figure, large batch effects are clearly observed affecting sample distributions in both kind of biological samples. These batch effects were corrected after normalization using batchCorr package, which takes into account different possible drift trends for each metabolite along the analytical sequences [29]. The PCA scores plots obtained with the normalized data showed the correction of the batch effects (**Figure 1c-d**) demonstrating, therefore, a good efficiency of the normalization process as well as a good quality of the data used in the statistical analyses. Finally, PCA were also performed with the final data without the QC samples in which no outliers were detected. However, the clusters between the different categories of the samples were not observed in the PCA scores plots (**Figure 1S**). The biological variability among individuals together with the fact that SADs are closely related produce extremely complex metabolomic matrices. Therefore, unsupervised methods based on PCA were not enough to separate samples according to the different diseases.

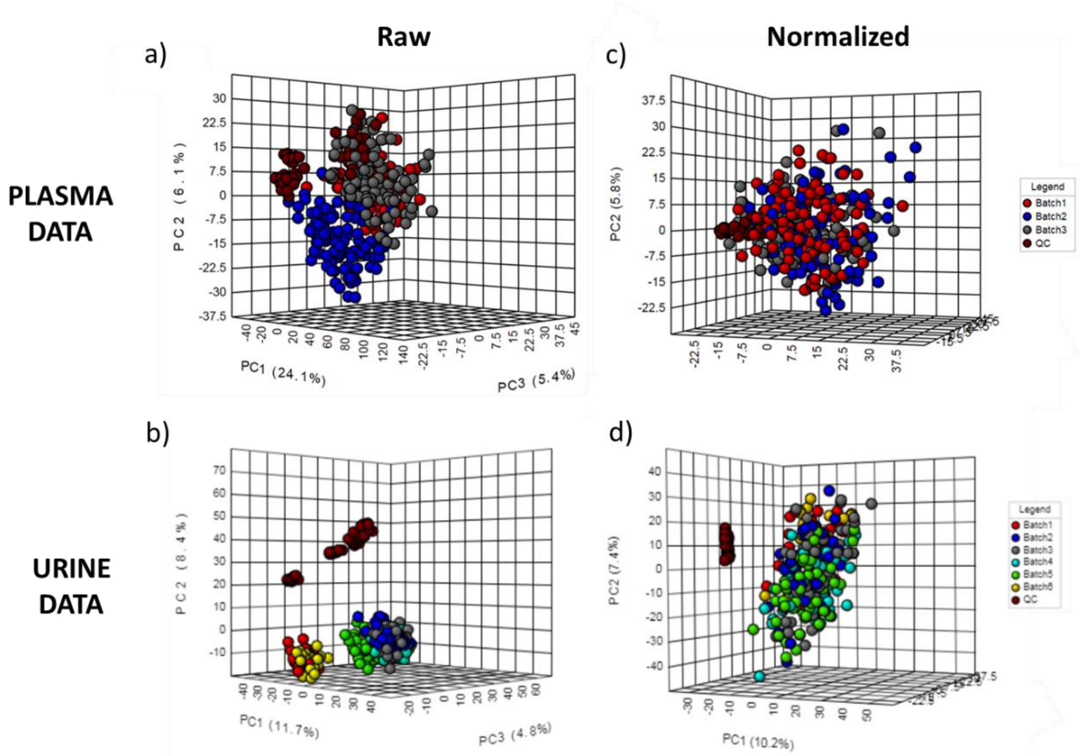
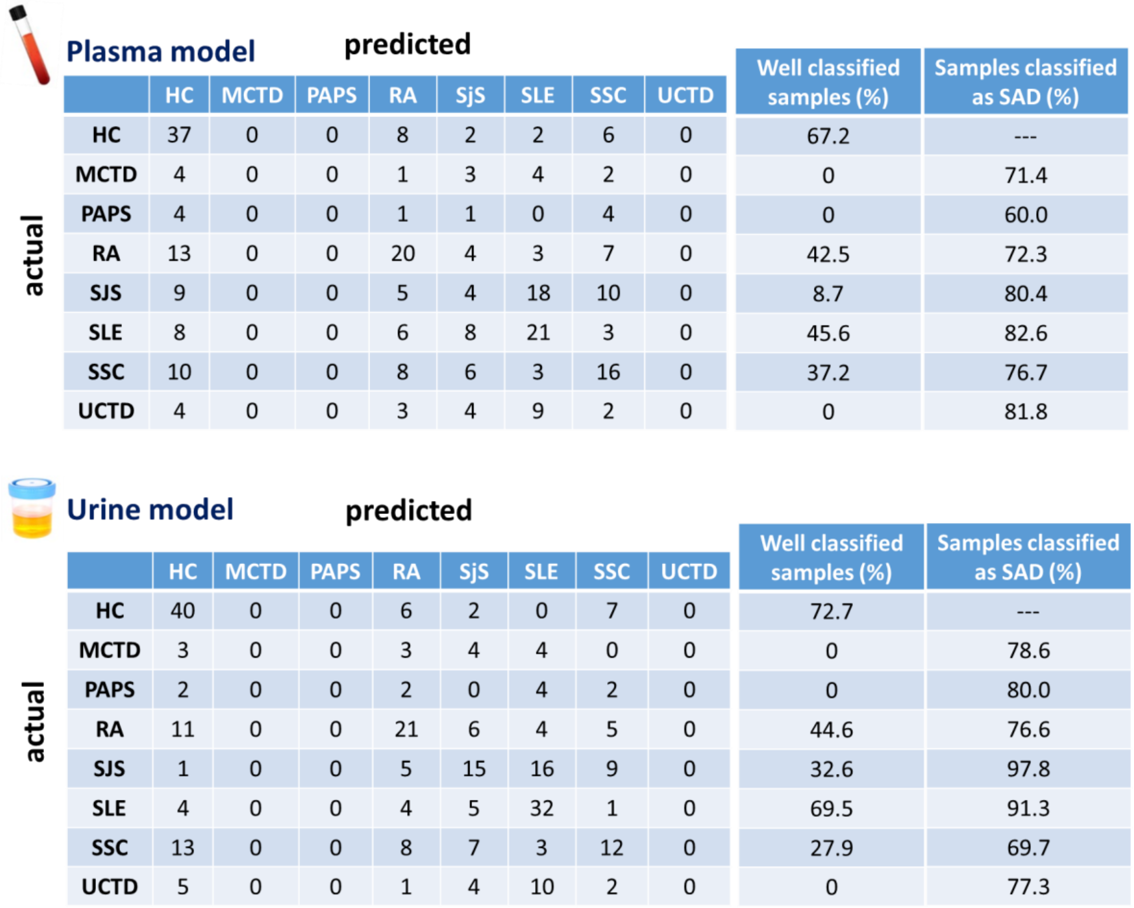


Figure 1. PCA scores plots from data before (1a. Plasma; 1b, Urine) and after (1c: Plasma; 1d: Urine) batchCorr normalization. (color code by analytical batches).

3.2. *Multivariate statistical analyses*

As a first approximation, two rdCV-PLS models were created with all data obtained from plasma and urine samples, respectively. Therefore, in these global models the dependent variables were the different categories of the samples (HC, SLE, SSC, SjS, RA, PAPS, MCTD and UCTD). The numbers of misclassification were used to evaluate the results and the quality of these multivariate models. **Figure 2** shows the confusion matrixes obtained in both multivariate models. In general, the rdCV-PLS model based on urine data got slightly better results compared to that of plasma.

These results show that the predictive capabilities of the rdCV-PLS models in order to classify all cases correctly using metabolic data were limited. Nevertheless, most of the cases belonging to SADs were classified in any of these diseases, in a percentage range between 60 and 97.8%, depending on the particular disease. Among all these results, it also stands out as a high percentage of the HC were correctly classified. Regarding SADs cases, SLE and RA were the diseases that obtained the best classification results. The results show that no sample belonging to the PAPS, MCTD and UCTD pathologies was well classified. These results may be highly related to the fact that these experimental groups were represented by a smaller number of subjects. In this way, the definition of these groups in the rdCV-PLS models could be more difficult compared to the other pathologies. Nevertheless, around 80% of the UCTD samples were classified in its related pathologies in both models. Comparing these classifications, it stands out that the majority of UCTD cases were classified as SLE.



Plasma model predicted

actual

	HC	MCTD	PAPS	RA	SjS	SLE	SSC	UCTD	Well classified samples (%)	Samples classified as SAD (%)
HC	37	0	0	8	2	2	6	0	67.2	---
MCTD	4	0	0	1	3	4	2	0	0	71.4
PAPS	4	0	0	1	1	0	4	0	0	60.0
RA	13	0	0	20	4	3	7	0	42.5	72.3
SJS	9	0	0	5	4	18	10	0	8.7	80.4
SLE	8	0	0	6	8	21	3	0	45.6	82.6
SSC	10	0	0	8	6	3	16	0	37.2	76.7
UCTD	4	0	0	3	4	9	2	0	0	81.8

Urine model predicted

actual

	HC	MCTD	PAPS	RA	SjS	SLE	SSC	UCTD	Well classified samples (%)	Samples classified as SAD (%)
HC	40	0	0	6	2	0	7	0	72.7	---
MCTD	3	0	0	3	4	4	0	0	0	78.6
PAPS	2	0	0	2	0	4	2	0	0	80.0
RA	11	0	0	21	6	4	5	0	44.6	76.6
SJS	1	0	0	5	15	16	9	0	32.6	97.8
SLE	4	0	0	4	5	32	1	0	69.5	91.3
SSC	13	0	0	8	7	3	12	0	27.9	69.7
UCTD	5	0	0	1	4	10	2	0	0	77.3

Figure 2. Confusion matrices, percentages of well classified samples and percentages of samples classified as SAD obtained in the rdCV-PLS models created with the metabolomic data of the 7 SADs and HC (1a: Plasma Data, 1b: Urine data).

Table 2. Results of the rdCV-PLS models created with individual SAD and healthy controls. (nVar: average number of variables; class (%): average percentage of samples correctly classified; AUC: area under the roc curve; nComp: number of principal components; selected variables: number of variables selected in the ten rdCV-PLS models; p-value: results of the permutation test)

rdCV-PLS MUVR models (HC-SADS)		nVar	class (%)	AUC	nComp	Selected Variables	p-value
PLASMA	**SSC** (n=43) vs HC (n=43)	31	82.6	0.931	2	12	$1.87\ 10^{-5}$
	SjS (n=46) vs HC (n=46)	62	83.7	0.888	3	32	$6.42\ 10^{-7}$
	SLE (n=46) vs HC (n=46)	73	83.6	0.922	2	24	$6.89\ 10^{-7}$
	RA (n=47) vs HC (n=47)	44	77.6	0.818	3	24	$6.63\ 10^{-7}$
	PAPS (n=10) vs HC (n=10)	8	60.0	0.550	1	0	$0.90\ 10^{-3}$
	MCTD (n=14) vs HC (n=14)	18	57.1	0.536	2	2	$0.34\ 10^{-3}$
	UCTD (n=22) vs HC(n=22)	26	77.2	0.736	2	2	$0.10\ 10^{-4}$
URINE	**SSC** (n=43) vs HC (n=43)	75	71.7	0.814	2	16	$7.90\ 10^{-5}$
	SjS (n=46) vs HC (n=46)	75	87.2	0.836	2	28	$3.88\ 10^{-3}$
	SLE (n=46) vs HC (n=46)	64	81.9	0.866	3	19	$6.23\ 10^{-7}$
	RA (n=47) vs HC (n=47)	65	77.4	0.811	2	16	$4.96\ 10^{-5}$
	PAPS (n=10) vs HC (n=10)	27	59.5	0.670	1	0	>0.05
	MCTD (n=14) vs HC (n=14)	53	61.4	0.668	1	0	>0.05
	UCTD (n=22) vs HC(n=22)	45	73.9	0.736	2	0	$1.12\ 10^{-3}$

Table 3. Annotated metabolites selected from the rdCV-MUVR-PLS models created with plasma data from individual SADs and healthy controls.

Mass (Da)	RT (min)	Diseases (Rank)	Molecular Formula	Score (%)	Compound Name	MS/MS Fragments	Identification Database
117.0780	1.3	SJS (45.5), RA(105.6), SLE(71.2)	$C_5H_{11}NO_2$	86.81	L-Valine	55.0543/57.0572/72.0807	HMDB00043
149.0503	1.5	SJS (24.6), RA(25.6)	$C_{10}H_{12}N_2O_3$	96.40	L-Methionine	56.0482/104.0512/132.0638	HMDB00696
176.0947	1.8	SJS (25.0), RA(29.9), SLE(4.2)	$C_{10}H_{12}N_2O$	86.82	Serotonin	56.0481/102.0524/135.0415/146.0416/160.0586	HMDB00259
131.0751	2.4	SJS (23.4), RA(47.1), SLE(71.9)	$C_6H_{13}NO_2$	90.53	L-Leucine	69.0686/86.0944	HMDB00687
208.0848	4.3	SJS (26.1), SSC(6.2), MCTD(44.3)	$C_{10}H_{12}N_2O_3$	85.66	L-kynurenine	74.0216/94.0627/120.0413/146.0561	HMDB00684
231.1471	5.7	RA(148.2)	$C_{11}H_{21}NO_4$	93.8	Butyrilcarnitine	85.0288/113.9698/173.0816/232.1544	HMDB62510
184.1212	9.0	SJS (108.1), SSC(53.9), UCTD(13.8), SLE(6.6), RA(11.6)	$C_9H_{16}N_2O_2$	86.53	N-(3-acetamidopropyl)pyrrolidin-2-one	56.9408/86.9471/98.0591/126.0905/185.1262	HMDB61384
159.1260	9.9	RA(25.6),	$C_8H_{17}NO_2$	87.4	DL-2-amino-octanoic acid	97.0991/142.1201	HMDB00991
253.0828	11.3	RA(207.3)	$C_9H_{11}N_5O_4$	81.6	L-Threoneopterin	52.9993/140.06454	HMDB00727
259.1783	12.8	SJS (127.5),	$C_{13}H_{25}NO_4$	83.65	Hexanoylcarnitine	85.0287	HMDB00705
518.1392	15.8	RA(18.8), SLE(93.1)	$C_{25}H_{26}O_{12}$	85.69	Medicarpin 3-O-(6'-malonylglucoside)	229.0421/343.1077	HMDB38776
412.1325	17.8	SJS (1.3), SSC(1.3), RA(5.5), SLE(9.9)	$C_{15}H_{20}N_6O_8$	84.47	N6-Carbamoyl-L-threonyladenosine	136.0602/162.0392/281.09852/413.1398	Metlin95993
511.3225	18.4	SSC(32.2),	$C_{24}H_{50}NO_8P$	89.31	PS(O-18:0/0:0)	351.2465/475.2374/534.3116	LMGP03060002
423.3305	18.8	SLE(98.4)	$C_{25}H_{45}NO_4$	82.15	Linoleoyl carnitine	85.0271/144.0998	HMDB13212
399.3308	19.1	RA(19.6)	$C_{23}H_{45}NO_4$	93.3	Palmitoylcarnitine	85.0271/144.0984	HMDB00846
425.3461	19.3	SSC(13.6), RA(42.8),	$C_{21}H_{39}NO_4$	72.95	Octadecenoylcarnitine	85.0274/426.3541	HMDB94687
521.3462	23.1	RA(90.3),	$C_{26}H_{52}NO_7P$	94.67	LysoPC(18:1)	86.08950/104.1056/146.9803	HMDB02815
276.2031	23.7	SJS (61.1), SSC(43.5), RA(63.2), SLE(135.2),	$C_{18}H_{28}O_2$	60.57	Stearidonic acid	55.0521/69.0687/95.0825/161.9743/	HMDB06547
328.2363	24.4	SJS (30.0), MCTD(26.0), UCTD(8.7), SLE(91.4), SSC(23.1)	$C_{22}H_{32}O_2$	83.41	Docosahexanoic acid	107.0814/145.0975/173.1304/161.1285	HMDB02183
507.3634	24.4	SLE(37.3)	$C_{26}H_{54}NO_6P$	88.9	LysoPC(18:0)	86.0942/104.1048/146.9798	HMDB13122
302.2194	24.7	SJS (89.7), SLE(58.3)	$C_{20}H_{30}O_2$	80.91	Eicosapentaenoic acid	81.0686/91.0535/119.0165/303.2146/325.2115	HMDB01999
278.2193	25.7	SJS (13.1), SSC(14.7), RA(160.1), SLE(80.4)	$C_{18}H_{30}O_2$	87.00	Linolenic Acid	69.0660/95.0833/123.1134/279.2177	HMDB01388
304.2351	26.4	SJS (5.0), SSC(7.9), SLE(28.4)	$C_{20}H_{32}O_2$	87.11	Arachidonic acid	55.0530/57.0685/67.0526/71.0842/ 107.0811/121.0995/123.0117/161.112/177.3601	HMDB01043
676.4568	26.4	SJS (8.9), SLE(35.4),SSC(4.8)	$C_{36}H_{69}O_9P$	97.2	PG(P-30:1)	699.4461 (Level 3)	LMGP04030004
306.2498	28.5	SJS (23.6), SSC(84.3)	$C_{20}H_{34}O_2$	83.9	Sciadonic acid	55.0524/95.0155	HMDB31058
823.5258	34.1	SLE(13.4)	$C_{45}H_{78}NO_{10}P$	85.1	PS(39:5)	146.979	LMGP03010503
755.5397	34.2	SLE(26.2)	$C_{42}H_{78}NO_8P$	96.3	PC(34:3)	86.0949/184.0713	HMDB07975

Table 3 (Cont).

Mass (Da)	RT (min)	Diseases (Rank)	Molecular Formula	Score (%)	Compound Name	MS/MS Fragments	Identification Database
902.4951	34.2	SJS (125.7),	$C_{46}H_{80}O_3P_2$	77.2	PGP(40:6)	86.0947/595.4633/669.4813/719.4583/720.4605	HMDB13516
767.5726	38.3	SLE(164.9)	$C_{44}H_{82}NO_7P$	77.1	PC(36:3)	86.0938/184.0695/709.4987	HMDB11246
780.6045	38.7	RA(123.8), SLE(125.2)	$C_{45}H_{85}N_2O_6P$	94.43	C22:3 Sphingomyelin	86.0939/184.0699/598.5429	HMDB13468
810.6537	39.6	SLE(122.3)	$C_{52}H_{90}O_6$	80.5	TG(49:4)	Level 3	HMDB42842

Table 4. Annotated metabolites selected from the rdCV-MUVR-PLS models created with urine data from individual SADs and healthy controls.

Mass	RT (min)	Diseases (Rank)	Molecular Formula	Score (%)	Compound Name	MS/MS Fragments	Identification Database
169.0855	1.0	SJS(266.6)	$C_7H_{11}N_3O_2$	95.67	Methyl Histidine	42.0331/81.0449/95.0602/96.0683	HMDB04935
113.0596	1.0	SJS(49.7)	$C_4H_7N_3O$	97.2	Creatinine	43.0287/44.0492/58.0651/86.0810	HMDB00562
277.1162	1.3	SJS(29.9), RA(2.88)	$C_{11}H_{19}NO_7$	84.57	N-(1-Deoxy-1-fructosyl)proline	100.0755/130.0494/214.1060/232.1174	HMDB38493
268.1174	1.3	SJS(53.7), SLE(32.4), SSC(91.1)	$C_{11}H_{16}N_4O_4$	98.8	Histidinyl-Hydroxyproline	72.0447/110.0718/111.0743/114.055/156.0764	HMDB28886
112.0277	1.5	RA(68.3)	$C_4H_4N_2O_2$	86.3	Uracil	113.0350 (level 3)	HMDB00300
227.0797	1.6	RA(2.7)	$C_{10}H_{13}NO_5$	83.76	L-Arogenate	228.0870 (level3)	C00826
214.1323	2.0	SJS(116.9)	$C_{10}H_{18}N_2O_3$	86.8	Valyl-Proline	43.0173/70.0647/72.0805/113.0708/114.0549/159.0772	HMDB29135
203.1270	2.2	SSC(11.3)	$C_8H_{17}N_3O_3$	99.15	Lysyl-Glycine	57.0442/158.1282	HMDB28951
228.1478	2.9	SJS(69.0)	$C_{11}H_{20}N_2O_3$	97.89	L-isoleucyl-L-proline	60.0806/70.0653/114.0552	HMDB11174
216.1112	3.3	SLE(86.4)	$C_9H_{16}N_2O_4$	87.1	Threoninyl-Proline	70.0652/116.0714	HMDB29069
261.1577	3.4	RA(94.5)	$C_{12}H_{23}NO_5$	99.3	hydroxyisovaleroyl carnitine	85.0288	HMDB62555
221.0723	3.7	SJS(220.9)	$C_8H_{15}NO_4S$	98.29	N-lactoyl-Methionine	57.0336/77.0385/85.0285/91.0541	HMDB62182
271.1644	4.9	SLE(59.7)	$C_{11}H_{21}N_5O_3$	96.04	Arginylproline	100.0757/167.0565	HMDB28717
246.1217	5.0	SJS(201.4)	$C_{10}H_{18}N_2O_5$	97.94	Aspartyl-Leucine	42.0332/56.9421/71.0228	HMDB28757
366.1427	5.1	SJS(229.9)	$C_{17}H_{22}N_2O_7$	99.45	N-(1-Deoxy-1-fructosyl)tryptophan	229.0968/230.0807/247.1076/350.1252	Metlin92655
231.1475	5.7	RA(7.4)	$C_{11}H_{21}NO_4$	98.63	Butyrylcarnitine	60.0807/85.0288/173.0810	HMDB62510
137.0476	6.2	SJS(107.3)	$C_7H_7NO_2$	72.1	Trigonelline	51.0239/39.0224/65.0398/79.0428	C01004
210.0642	6.6	SJS(5.8, RA(291.5), SSC(51.9), SLE(102.9)	$C_9H_{10}N_2O_4$	83.17	N-(5amino)-2-hydroxybenzoylglycine	108.0449/109.0526/150.0418/165.0648	HMDB61683
99.0683	6.6	SSC(171.7)	C_5H_9NO	87.19	2-Piperidinone	44.0127/56.0487	HMDB11749
297.1074	7.1	SJS(209.5)	$C_{11}H_{15}N_5O_5$	97.5	Methylguanosine	109.0506/110.0350/135.0300/149.0461/166.0724/167.0750	HMDB01563
157.0740	8.6	SSC(29.8)	$C_7H_{11}NO_3$	81.90	Tiglylglycine	55.0516/80.0493/108.6441	HMDB00959
303.1567	8.8	SJS(193.2)	$C_{16}H_{21}N_3O_3$	84.35	Tryptophyl-Valine	124.0497/149.0440/179.0789/194.1010	HMDB29096

Table 4 (Cont).

Mass	RT (min)	Diseases (Rank)	Molecular Formula	Score (%)	Compound Name	MS/MS Fragments	Identification Database
159.0890	9.3	SSC(12.7)	$C_7H_{13}NO_3$	81.96	Isovalerylglycine	41.0384/43.0171/57.0700/160.0963	HMDB00678
243.1473	10.7	SJS(85.3), RA(332.3)	$C_{12}H_{21}NO_4$	85.1	Tiglylcarnitine	85.0280/185.0787	HMDB02366
311.1233	11.0	SJS(256.9)	$C_{12}H_{17}N_5O_5$	95.32	1,7-Dimethylguanosine	180.0760	HMDB01961
260.1373	11.3	SJS(161.3)	$C_{11}H_{20}N_2O_5$	86.2	Gamma-glutamylisoleucine	86.0959/130.0505/132.1019/244.117	HMDB11171
383.1078	12.7	SJS(329.9)	$C_{14}H_{17}N_5O_8$	98.76	Succinyladenosine	136.0611/192.0512/234.0620/252.0756	HMDB00912
189.0428	16.2	RA(276.5)	$C_{10}H_7NO_3$	97.8	Kynurenic acid	89.0388/116.0494/144.0443	HMDB00715
412.1344	22.1	SJS(184.2)	$C_{15}H_{20}N_6O_8$	96.42	N6-Carbamoyl-L-threonyladenosine	136.0602/162.0392/281.09852	HMDB41623
287.2098	36.0	SJS(13.8), RA(21.5)	$C_{15}H_{29}NO_4$	77.4	L-octanoylcarnitine	60.0802/85.0278	HMDB00834
357.2515	38.6	SSC(146.5)	$C_{19}H_{35}NO_5$	95.57	9-hydroxydecenoylcarnitine	60.0799/85.0291	LMFA07070024
303.3891	41.4	SLE(227.1)	$C_{22}H_{44}$	85.97	1-Docosene	41.0380/43.0537/46.0645/57.0698/186.2210	HMDB62602
255.2569	43.6	SJS(26.9)	$C_{16}H_{33}NO$	85.9	Palmitic amide	278.2442 (level 3)	HMDB12273
392.2938	43.8	SSC(42.3)	$C_{34}H_{40}O_4$	96.43	Allodeoxycholic acid	57.0694/71.0848/113.1321	HMDB00478

3.2.1. *SAD(s) vs HC models*

Table 2 shows the results of the rdCV-PLS models created with individual SADs versus healthy control. In general, acceptable results (percentages of correctly classified samples around 70-90 %; and AUC around 0.7-0.93) [35] were obtained for most diseases with the exception of PAPS and MCTD. Comparing the results of the individual rdCV-PLS models created with the plasma and urine samples, similar classification results were obtained for both types of biological samples. The biggest difference between them corresponds to the average number of variables used to create the models. This fact is reasonable given the high number of variables detected in urine samples compared to plasma samples.

As mentioned before, PAPS and MCTD were represented by a very small number of samples. Therefore, this low representativeness of these experimental groups may be responsible for the unsatisfactory results achieved for these in multivariate models.

The molecular features selected in the ten corresponding rdCV-PLS models were chosen for identification. According to the criteria of the Metabolomics Standards Initiative (MSI) [36], the selected metabolites were putatively annotated. **Tables 3** and **4** list the annotated metabolites selected in the plasma and urine models, respectively. These tables contain information about the analytical (retention times, masses, retention times) and identification (scores, metabolite dabase, MS/MS fragments) results as well as the particular rdCV-PLS models where these variables were selected with their corresponding ranks. The values of these ranks are based on the variable importance of projection (VIP) for the PLS models. Low values of these ranks indicate a greater weight of the metabolite in the multivariate model [32]. As metabolite identification step remains the main limitation of the metabolomic studies based on

MS, several selected variables in the models were assigned as unknown identities (**Table 3S**).

Many of the metabolites identified in the models based on plasma samples were selected in several of the diseases, mainly SSC, SLE, SSC, and RA. These metabolites may have high potential as indicators of the development of a systemic autoimmune disease process. The main families of selected metabolites in plasma models were unsaturated fatty acids, acylcarnitines, lysophosphatidylcholines and amino acids. However, in urine models, metabolites were selected in particular models in most cases.

In these models, a remarkable result is the number of acylglycines selected in the different pathologies. **Figure 3** shows the trends of the six selected metabolites (docosahexanoic acid, linolenic acid, stearidonic acid, N-(3-acetamidopropyl)pyrrolidin-2-one, N6-carbamoyl-L-threonyladenosine, 2-hydroxybenzoylglycine) in a larger number of rdCV-PLS models. It is clearly observable the different levels of concentration of these metabolites in SADS groups compared to healthy controls. Among these metabolites, N6-carbamoyl-L-threonyladenosine is the one with the greatest potential as biomarker of these pathologies due to its high level of concentration in control samples compared to pathological cases. This compound is a member of the family of metabolites known as purine nucleosides, which is formed by a universal transformation of RNA where different enzymes and bacteria, such as *Escherichia coli.,* have a role in its synthesis [37]. The deregulation of this metabolite in SADs may be closely related to renal involvement, which is quite frequent in these types of diseases [22,38] due to the fact that previous studies have demonstrated the potential as a biomarker of this metabolite for renal dysfunction in cases of diabetes

[39,40]. On the other hand, N-(3-acetamidopropyl)pyrrolidin-2-one was another metabolite that presented high differences in plasma levels between HC and SADs cases. This analyte is a catabolic product of spermidine, whose intake has been related to protective effects against different diseases such as cancer, cardiovascular or neurodegenerative diseases [41].

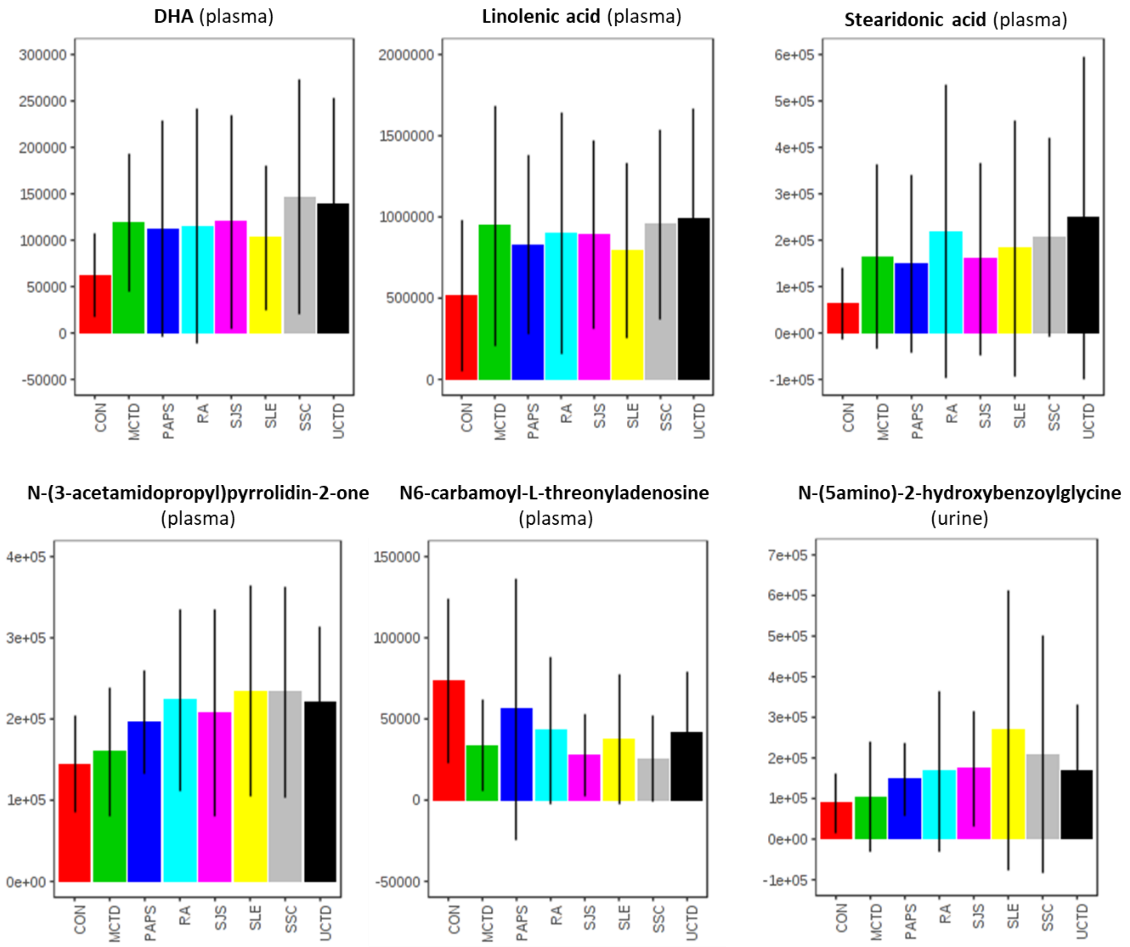


Figure 3. Relative concentration levels of the six metabolites selected in a larger number of rdCV-PLS multivariate models.

As shown in **Figure 3**, unsaturated fatty acids (UFAs) were one of the main families of deregulated metabolites in SADS, suggesting that most of the studied SADs involve serious disorders of the metabolism of UFAs. Concretely, several omega-3 and omega-6 fatty acids, such as docosahexanoic (DHA), eicosapentaenoic acid (EPA), stearidonic, linolenic or arachidonic acids, appeared up-regulated in plasma samples of different SADs. These results agree with those reported by Bengtsson et al. [42] in sera from SLE, SJS and SSC as well as with the elevated plasma levels of free fatty acids in RA patients relative to controls described by Gu *et al*. [43]. These results can be a general tendency of autoimmune diseases. However, these up regulations could be a cause of a diet enriched in UFAs by SADs patients. Previous studies have shown that omega-3 fatty acids have benefits in several inflammatory and autoimmune diseases in humans [44,45]. Due to these evidences, the up-regulation of the levels of the UFAs in plasma samples of SADs patients could be highly related to the increased intake of these types of compounds in the diet.

Additionally, acylcarnitines, involved in fatty acids transportation, were generally increased in several SADs in plasma and urine samples. It has been described that patients with a disturbed fatty acid oxidation or branched chain amino acid catabolism usually accumulate abnormal acyl-CoA species, eventually leading to the accumulation of related unusual acylcarnitines [46]. Similar results obtained in RA plasma were attributed to a drastic acceleration of lipids mobilization [43]. On the contrary, Wu *et al.* [47] reported a significant reduction of long chain FA, including DHA, EPA and linolenic acid as well as acylcarnitines in SLE whereas levels of medium chain FA and lipid oxidation products were elevated. In this case, authors hypothesise that the reduction in long chain FA may be due to a low availability of acyl coA and co-factors, dietary differences [48] as well as an acceleration in metabolic consumption or conversion, either through lipid peroxidation or heightened leukotriene synthesis. The discrepancies in data suggest that the fatty acids metabolism imbalance in SADs patients deserves further attention.

Another family of metabolites also involved in fatty acid beta-oxidation metabolism are acylglycines. Metabolites belonging to this family were found altered in urine samples in several SADs, such as SSC, SJS, RA or SLE. Specifically, these metabolites were N-(5-amino)-2-hydroxybenoylglycine, isovalerylglycine, tiglylglycine and lysyl-glycine. This family of compounds, which are frequently produced by the enzyme glycine N-acyltransferase, has been used in the detection of inborn errors of metabolism where amino acids and organic acids are involved. In this way, the deregulation of the excretion of these metabolites has shown to be highly related with perturbations in the mitochondrial fatty acid beta-oxidation [49].

Furthermore, different metabolites related to the amino acids metabolism were selected in several multivariate models. For example, essential amino acids, such as valine, methionine or leucine; metabolites derived from essential amino acids, as the case of serotonin, kynurenine, N-lactoyl-Methionine, as well as dipeptides such as Tryptophyl-Valine, Aspartyl-Leucine, Arginylproline or Valyl-Proline.

The alterations in proline metabolism are in concordance with previous results found in a previous study of systemic sclerosis [17]. These alterations seem to be generalizable to other pathologies, such as SjS, RA or SLE, given the selected metabolites in the urine multivariate models of these diseases. On the other hand, metabolites related to tryptophan metabolism were also seleced in the models, such as L-kynurenine, kynurenic acid, N-(1-Deoxy-1-fructosyl)tryptophan, tryptophyl-valine or serotonin. Deregulation of tryptophan-related metabolic pathways has been shown to be related to the immune system, inflammation and neurological processes [50,51].

3.2.2. *UCTD vs SAD(s)*

Table 5 shows the results of the rdCV-PLS models created with cases of UCTD versus its related diseases (SSC, SjS, SLE, RA and MCTD). These results (percentages of correctly classified samples around 47-65 %; and AUC around 0.46-0.72) indicate the great difficulties in discriminating UCTD from the rest of SADs bases on metabolomics data. In fact, in some models the permutation tests were not satisfactory (p-value > 0.05) indicating that no differences at the metabolic level were detected between the diseases studied in the rdCV-PLS models. These results could be highly related to the fact that patients with UCTD eventually end up being diagnosed in another of the defined SADs, such as SLE, SSC, RA, MCTD, systemic vasculitis or polydermatomyositis [52]. Among all the models, the ones that showed the best classification results were those that compared UCTD with SSC.

Different variables were selected after the 10-fold cross validation of the MUVR models. These variables from plasma and urine are listed in **Table 6** with their analytical parameters. The MUVR models based on urine data showed slightly better results that revealed greater power to classify samples through a number of selected variables. Among the identified variables, it stands out the selected variables in the urine model that compare UCTD versus SJS. Most of these variables are common with the selected in multivariate models that compare SJS with healthy controls, such as N6-carbamoyl-L-threonyladenosine, methylguanosine, tryptophyl-valine, succinyladenosine, or 1,7-dimethylguanosine. Therefore, this group of metabolites could have a high predictive potential to classify Sjögren's syndrome with respect to healthy controls and UCTD. Most of them are related to the nucleosides of guanosine and adenosine. These results indicate the significance of the adenosinergic pathway in this kind of diseases, which has been described in previous reports [53]. In view of these results and given the difficulties in differentiating diseases from each other from metabolic data, it is necessary to continue studying this type of pathologies through the integration of data obtained from different omic techniques.

Conclusions

The rdCV-PLS models using HPLC-MS data from urine and plasma analysis of HC and individual SAD patients showed better results than multivariate models that compared UCTD with its related diseases. The most significant families of metabolites between HC and SADs were unsaturated fatty acids, acylcarnitines, acylglycines and amino acids. It is also remarkable the higher level of N6-carbamoyl-L-threonyladenosine in HC controls compared to SADs cases, which would be highly related to the renal involvement present in these diseases. The low classificatory efficiency of the MUVR models that compared UCTD versus different SADs indicated the great difficult of diagnose these diseases and the need to continue studying them by means of data integration of different omic methodologies.

Table 5. Results of the rdCV-PLS models created with UCTD versus its related diseases (SSC, SjS, SLE, RA, and MCTD). (nVar: average number of variables; class (%): average percentage of samples correctly classified; AUC: area under the roc curve; nComp: number of principal components; selected variables: number of variables selected in the ten rdCV-PLS models; p-value: results of the permutation test)

rdCV-PLS MUVR models (HC-SADS)		**nVar**	**class (%)**	**AUC**	**nComp**	**Selected Variables**	**p-value**
PLASMA	**SSC** (n=43) vs **UCTD** (n=22)	26	60.7	0.620	2	3	0.011
	SjS (n=46) vs **UCTD** (n=22)	18	47.5	0.535	2	0	>0.05
	SLE (n=46) vs **UCTD** (n=22)	37	50.0	0.468	2	0	>0.05
	RA (n=47) vs **UCTD** (n=22)	22	57.5	0.590	2	3	0.013
	MCTD (n=14) vs **UCTD** (n=14)	26	63.2	0.566	1	5	0.047
URINE	**SSC** (n=22) vs **UCTD** (n=22)	32	60.7	0.628	2	5	0.024
	SjS (n=22) vs **UCTD** (n=22)	32	54.5	0.551	2	12	0.049
	SLE (n=22) vs **UCTD** (n=22)	45	54.7	0.521	2	6	0.047
	RA (n=22) vs **UCTD** (n=22)	38	64.5	0.519	2	0	>0.05
	MCTD (n=14) vs **UCTD** (n=14)	15	50.4	0.525	1	0	>0.05

Table 6. Annotated metabolites selected from the rdCV-MUVR-PLS models created with plasma data from individual SADs versus UCTD.

	Mass (Da)	RT (min)	Diseases (Rank)	Molecular Formula	Score (%)	Compound Name	MS/MS Fragments	Identification Database
PLASMA	159.0527	1.1	MCTD(8.1)	$C_6H_9NO_4$	96.77	N-Methyl-2-oxoglutaramate	56.0480/74.0584/86.0586/114.0527	C03623
	112.0271	2.1	MCTD(18.6)	----	----	Unknown	-----	-----
	164.0498	2.4	RA(20.6), MCTD(71.9)	$C_9H_8O_3$	98.94	m-Coumaric acid	77.0370/91.0522/119.0469/147.0410	HMDB62774
	208.0848	4.3	SSC(49.3)	$C_{10}H_{12}N_2O_3$	85.66	L-kynurenine	74.0216/94.0627/120.0413/146.0561	HMDB00684
	156.0763	8.1	SSC(11.9)	----	----	Unknown	-----	-----
	287.1204	15.2	RA(100.2)	$C_{15}H_{29}NO_4$	93.89	L-Octanoylcarnitine	60.0785/85.0262/310.1934	HMDB00834
	495.3308	21.9	RA(80.8)	$C_{24}H_{50}NO_7P$	90.62	LysoPC(16:0)	86.0953/104.1060/146.9804/459.2470	HMDB10382
	569.3426	22.3	MCTD(32.2)	$C_{30}H_{52}NO_7P$	68.90	LysoPC(22:5)	86.0954/104.1055/146.9802/483.2450/533.2598	HMDB10403
	481.3109	24.9	MCTD(24.7)	$C_{23}H_{48}NO_7P$	71.25	LysoPC(15:0)	482.3182 (level3)	HMDB10381
	616.4617	28.7	SSC(31.6)	----	----	Unknown	-----	-----
URINE	113.0596	1.0	SJS(260.5)	$C_4H_7N_3O$	97.2	Creatinine	43.0287/44.0492/58.0651/86.0810	HMDB00562
	112.0274	1.2	SLE(44.5), SSC(33.7)	$C_4H_4N_2O_2$	86.3	Uracil	113.0350 (level 3)	HMDB00300
	225.0753	1.5	SLE(147.7)	----	----	Unknown	-----	-----
	240.1480	1.5	SJS(226.9)	----	----	Unknown	-----	-----
	244.0669	1.5	SJS(344.2)	$C_9H_{12}N_2O_6$	93.7	Uridine	82.0291/125.0348/139.0503/155.0450/209.0552	HMDB00296
	136.0338	1.9	SJS(89.2), SSC(56.6)	$C_5H_4N_4O$	95.33	Hypoxanthine	55.0285/67.0289/82.0395/94.0398/110.0351/119.0353	HMDB00157
	304.0905	2.4	SSC(66.9), SJS(92.1)	$C_{11}H_{16}N_2O_8$	84.21	N-Acetylaspartylglutamic acid	305.0978(level3)	HMDB01067
	228.1478	2.9	SJS(351.9)	$C_{11}H_{20}N_2O_3$	97.89	L-isoleucyl-L-proline	60.0806/70.0653/114.0552	HMDB11174
	198.0706	5.5	SLE(242.0)	----	----	Unknown	-----	-----
	99.0683	6.6	SLE(443.6)	C_5H_9NO	87.19	2-Piperidinone	44.0127/56.0487	HMDB11749
	297.1074	7.1	SJS(182.4)	$C_{11}H_{15}N_5O_5$	97.5	Methylguanosine	109.0506/110.0350/135.0300/149.0461/166.0724/167.0750	HMDB01563
	150.0543	7.3	SJS(305.2)	$C_6H_6N_4O$	96.87	1-Methylhypoxanthine	42.0332/55.0283/82.0397/94.0394/110.0347	HMDB13141
	303.1567	8.8	SJS(77.7)	$C_{16}H_{21}N_3O_3$	84.35	Tryptophyl-Valine	124.0497/149.0440/179.0789/194.1010	HMDB29096
	311.1233	11.0	SJS(87.8)	$C_{12}H_{17}N_5O_5$	95.32	1,7-Dimethylguanosine	180.0760	HMDB01961
	383.1078	12.7	SJS(248.6)	$C_{14}H_{17}N_5O_8$	98.76	Succinyladenosine	136.0611/192.0512/234.0620/252.0756	HMDB00912

Table 6 (Cont)

	Mass (Da)	RT (min)	Diseases (Rank)	Molecular Formula	Score (%)	Compound Name	MS/MS Fragments	Identification Database
URINE	512.2145	16.4	SSC(95.5)	----	----	Unknown	-----	-----
	349.1562	19.9	SSC(115.9)	----	----	Unknown	-----	-----
	412.1344	22.1	SJS(114.1)	$C_{15}H_{20}N_6O_8$	96.42	N6-Carbamoyl-L-threonyladenosine	136.0602/162.0392/281.09852	HMDB41623
	116.5186	40.5	SLE(33.7)	----	----	Unknown	-----	-----
	317.2333	41.9	SLE(492.46)	----	----	Unknown	-----	-----

Acknowledgements

The work described has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement No. 115565, the resources for which are composed of a financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and the European Federation of Pharmaceutical Industries and Associations (EFPIA) companies' in-kind contribution. The author AFO gratefully acknowledges the Spanish Ministry of Education, Culture and Sports for the received grant FPU14/03992.

Bibliography

[1] Schmitt J. Recombinant autoantigens for diagnosis and therapy of autoimmune diseases. Biomed Pharmacother 2003;57:261–8.

[2] Singh S, Wal P, Wal A, Srivastava V, Tiwari R, Dutt R. UNDERSTANDING AUTOIMMUNE DISEASE: AN UPDATE REVIEW. vol. 3. 2016.

[3] Druet P. Diagnosis of autoimmune diseases. J Immunol Methods 1992;150:177–84. doi:10.1016/0022-1759(92)90076-6.

[4] Ferreira HB, Pereira AM, Melo T, Paiva A, Domingues MR. Lipidomics in autoimmune diseases with main focus on systemic lupus erythematosus. J Pharm Biomed Anal 2019;174:386–95. doi:10.1016/J.JPBA.2019.06.005.

[5] Shi G, Zhang J, Zhang ZJ, Zhang X. Systemic autoimmune diseases. Clin Dev Immunol 2013;2013:728574. doi:10.1155/2013/728574.

[6] Raval P. Systemic (non-Organ Specific) Autoimmune Disorders. XPharm Compr Pharmacol Ref 2007:1–6. doi:10.1016/B978-008055232-3.60773-1.

[7] Jury EC, D'Cruz D, Morrow WJW. Autoantibodies and overlap syndromes in autoimmune rheumatic disease. J Clin Pathol 2001;54:340–7. doi:10.1136/JCP.54.5.340.

[8] Shah S, Chengappa K, Negi V. Systemic lupus erythematosus and overlap: A clinician perspective. Clin Dermatology Rev 2019;3:12. doi:10.4103/CDR.CDR_44_18.

[9] Ramos-Casals M, Brito-Zerón P, Font J. The Overlap of Sjögren's Syndrome with Other Systemic Autoimmune Diseases. Semin Arthritis Rheum 2007;36:246–55. doi:10.1016/j.semarthrit.2006.08.007.

[10] Venables PJW. Mixed connective tissue disease. Lupus 2006;15:132–7.

[11] Hoffman RW, Greidinger EL. Mixed Connective Tissue Disease 2002:347–57. doi:10.1007/978-1-59259-239-5_23.

[12] Antunes M, Scirè CA, Talarico R, Alexander T, Avcin T, Belocchi C, et al. Undifferentiated connective tissue disease: state of the art on clinical practice guidelines. RMD Open 2019;4:e000786. doi:10.1136/rmdopen-2018-000786.

[13] Spinillo A, Beneventi F, Caporali R, Ramoni V, Montecucco C. Undifferentiated connective tissue diseases and adverse pregnancy outcomes. An undervalued association? Am J Reprod Immunol 2017;78:e12762. doi:10.1111/aji.12762.

[14] Teruel M, Chamberlain C, Alarcón-Riquelme ME. Omics studies: their use in diagnosis and reclassification of SLE and other systemic autoimmune diseases. Rheumatology 2016;56:kew339. doi:10.1093/rheumatology/kew339.

[15] Agin A, Heintz D, Ruhland E, Chao de la Barca JM, Zumsteg J, Moal V, et al. Metabolomics - an overview. From basic principles to potential biomarkers (part 1). Med Nucl 2016;40:4–10. doi:10.1016/j.mednuc.2015.12.006.

[16] Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, Anderson N, et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nat Protoc 2011;6:1060–83. doi:10.1038/nprot.2011.335.

[17] Fernández-Ochoa Á, Quirantes-Piné R, Borrás-Linares I, Gemperline D, Alarcón Riquelme ME, Beretta L, et al. Urinary and plasma metabolite differences detected by HPLC-ESI-QTOF-MS in systemic sclerosis patients. J Pharm Biomed Anal 2019;162:82–90. doi:10.1016/j.jpba.2018.09.021.

[18] Bengtsson AA, Trygg J, Wuttge DM, Sturfelt G, Theander E, Donten M, et al. Metabolic profiling of systemic lupus erythematosus and comparison with primary Sjögren's syndrome and systemic sclerosis. PLoS One 2016;11. doi:10.1371/journal.pone.0159384.

[19] Li J, Che N, Xu L, Zhang Q, Wang Q, Tan W, et al. LC-MS-based serum metabolomics reveals a distinctive signature in patients with rheumatoid arthritis. Clin Rheumatol 2018;37:1493–502. doi:10.1007/s10067-018-4021-6.

[20] Bouatra S, Aziat F, Mandal R, Guo AC, Wilson MR, Knox C, et al. The Human Urine Metabolome. PLoS One 2013;8:e73076. doi:10.1371/journal.pone.0073076.

[21] Karu N, Deng L, Slae M, Guo AC, Sajed T, Huynh H, et al. A review on human fecal metabolomics: Methods, applications and the human fecal metabolome database. Anal Chim Acta 2018;1030:1–24. doi:10.1016/j.aca.2018.05.031.

[22] Kronbichler A, Mayer G. Renal involvement in autoimmune connective tissue diseases. BMC Med 2013;11:95. doi:10.1186/1741-7015-11-95.

[23] Bruce SJ, Tavazzi I, Parisod V, Rezzi S, Kochhar S, Guy P a. Investigation of human blood plasma sample preparation for performing metabolomics using ultrahigh performance liquid chromatography/mass spectrometry. Anal Chem 2009;81:3285–96. doi:10.1021/ac8024569.

[24] Chetwynd AJ, Abdul-Sada A, Holt SG, Hill EM. Use of a pre-analysis osmolality normalisation method to correct for variable urine concentrations and for improved metabolomic analyses. J Chromatogr A 2016;1431:103–10. doi:10.1016/j.chroma.2015.12.056.

[25] Team RDC, R Development Core Team R. R: A Language and Environment for Statistical Computing. R Found Stat Comput 2016. doi:10.1007/978-3-540-74686-7.

[26] Adusumilli R, Mallick P. Data Conversion with ProteoWizard msConvert, Humana Press, New York, NY; 2017, p. 339–68. doi:10.1007/978-1-4939-6747-6_23.

[27] Libiseller G, Dvorzak M, Kleb U, Gander E, Eisenberg T, Madeo F, et al. IPO: a tool for automated optimization of XCMS parameters. BMC Bioinformatics 2015;16:118. doi:10.1186/s12859-015-0562-8.

[28] Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem 2006;78:779–87. doi:10.1021/ac051437y.

[29] Brunius C, Shi L, Landberg R. Large-scale untargeted LC-MS metabolomics data correction using between-batch feature alignment and cluster-based within-batch signal intensity drift correction. Metabolomics 2016;12:1–13. doi:10.1007/s11306-016-1124-4.

[30] Broeckling CD, Afsar FA, Neumann S, Ben-Hur A, Prenni JE. RAMClust: A novel feature clustering method enables spectral-matching-based annotation for metabolomics data. Anal Chem 2014;86:6812–7. doi:10.1021/ac501530d.

[31] Chong J, Soufan O, Li C, Caraus I, Li S, Bourque G, et al. MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. Nucleic Acids Res 2018;46:W486–94. doi:10.1093/nar/gky310.

[32] Shi L, Westerhuis JA, Rosén J, Landberg R, Brunius C. Variable selection and validation in multivari- ate modelling. Bioinformatics 2018:1–9. doi:10.1093/bioinformatics/bty710.

[33] Gil de la Fuente A, Grace Armitage E, Otero A, Barbas C, Godzien J. Differentiating signals to make biological sense – A guide through databases for MS-based non-targeted metabolomics. Electrophoresis 2017;38:2242–56. doi:10.1002/elps.201700070.

[34] Dührkop K, Fleischauer M, Ludwig M, Aksenov AA, Melnik A V., Meusel M, et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat Methods 2019. doi:10.1038/s41592-019-0344-8.

[35] Westerhuis JA, Hoefsloot HCJ, Smit S, Vis DJ, Smilde AK, Velzen EJJ, et al. Assessment of PLSDA cross validation. Metabolomics 2008;4:81–9. doi:10.1007/s11306-007-0099-6.

[36] Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, et al. Proposed minimum reporting standards for chemical analysis: Chemical Analysis

Working Group (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics 2007;3(3):211–21. doi:10.1007/s11306-007-0082-2.

[37] Thiaville PC, Iwata-Reuyl D, de Crécy-Lagard V. Diversity of the biosynthesis pathway for threonylcarbamoyladenosine (t(6)A), a universal modification of tRNA. RNA Biol 2014;11:1529–39. doi:10.4161/15476286.2014.992277.

[38] Steen VD. Kidney involvement in systemic sclerosis. Press Medicale 2014;43. doi:10.1016/j.lpm.2014.02.031.

[39] Niewczas MA, Mathew A V., Croall S, Byun J, Major M, Sabisetti VS, et al. Circulating Modified Metabolites and a Risk of ESRD in Patients With Type 1 Diabetes and Chronic Kidney Disease. Diabetes Care 2017;40:383–90. doi:10.2337/dc16-0173.

[40] Mathew A V, Jaiswal M, Ang L, Michailidis G, Pennathur S, Pop-Busui R. Impaired Amino Acid and TCA Metabolism and Cardiovascular Autonomic Neuropathy Progression in Type 1 Diabetes. Diabetes 2019;68:2035–44. doi:10.2337/db19-0145.

[41] Madeo F, Eisenberg T, Pietrocola F, Kroemer G. Spermidine in health and disease. Science (80-) 2018;359:eaan2788. doi:10.1126/science.aan2788.

[42] Bengtsson AA, Trygg J, Wuttge DM, Sturfelt G, Theander E, Donten M, et al. Metabolic profiling of systemic lupus erythematosus and comparison with primary Sjögren's syndrome and systemic sclerosis. PLoS One 2016;11. doi:10.1371/journal.pone.0159384.

[43] Gu Y, Lu C, Zha Q, Kong H, Lu X, Lu A, et al. Plasma metabonomics study of rheumatoid arthritis and its Chinese medicine subtypes by using liquid chromatography and gas chromatography coupled with mass spectrometry. Mol Biosyst 2012;8:1535. doi:10.1039/c2mb25022e.

[44] Simopoulos AP. Omega-3 fatty acids in inflammation and autoimmune diseases. J Am Coll Nutr 2002;21:495–505.

[45] Akbar U, Yang M, Kurian D, Mohan C. Omega-3 Fatty Acids in Rheumatic Diseases. JCR J Clin Rheumatol 2017;23:330–9. doi:10.1097/RHU.0000000000000563.

[46] Costa CG, Struys EA, Bootsma A, Ten Brink HJ, Dorland L, Tavares De Almeida I, et al. Quantitative analysis of plasma acylcarnitines using gas chromatography chemical ionization mass fragmentography. J Lipid Res 1997;38.

[47] Wu T, Xie C, Han J, Ye Y, Weiel J, Li Q, et al. Metabolic disturbances associated with systemic lupus erythematosus. PLoS One 2012;7:1–9. doi:10.1371/journal.pone.0037210.

[48] Simopoulos AP. Omega-3 Fatty Acids in Inflammation and Autoimmune Diseases n.d.

[49] Costa CG, Guérand WS, Struys EA, Holwerda U, ten Brink HJ, Tavares de Almeida I, et al. Quantitative analysis of urinary acylglycines for the diagnosis of β-oxidation defects using GC-NCI-MS. J Pharm Biomed Anal 2000;21:1215–24. doi:10.1016/S0731-7085(99)00235-6.

[50] Mondanelli G, Iacono A, Carvalho A, Orabona C, Volpi C, Pallotta MT, et al. Amino acid metabolism as drug target in autoimmune diseases. Autoimmun Rev 2019;18:334–48. doi:10.1016/J.AUTREV.2019.02.004.

[51] Opitz CA, Wick W, Steinman L, Platten M. Tryptophan degradation in autoimmune diseases. Cell Mol Life Sci 2007;64:2542–63. doi:10.1007/s00018-007-7140-9.

[52] Mosca M, Tani C, Talarico R, Bombardieri S. Undifferentiated connective tissue diseases (UCTD): simplified systemic autoimmune diseases. Autoimmun Rev 2011;10:256–8. doi:10.1016/j.autrev.2010.09.013.

[53] Dong K, Gao Z-W, Zhang H-Z. The role of adenosinergic pathway in human autoimmune diseases. Immunol Res 2016;64:1133–41. doi:10.1007/s12026-016-8870

Supplementary Material

Material and Methods

Chemicals

All chemical were of analytical reagent grade and used as received. Formic acid and LC-MS grade methanol for analytical chromatography were purchased from Fluka, Sigma-Aldrich (Steinheim, Germany) and Fisher Scientific (Madrid, Spain), respectively. Water was purified by a Milli-Q system from Millipore (Bedford, MA, USA). For plasma treatment, ethanol and methanol (Fisher Scientific Madrid, Spain) were used.

Table 1S. Characteristics of the volunteers participating in the study (age, gender, duration of the diseases).

	HC	**RA**	**SLE**	**SjS**	**SSC**	**UCTD**	**MCTD**	**PAPS**
Subjects	55	47	46	46	43	22	14	10
Age	44.4 ± 12.0	59.2 ± 11.5	48.0 ± 11.4	57.1 ± 12.3	61.1 ± 10.5	52.5 ± 11.3	51.7 ± 16.6	43.4 ± 14.2
Females, n (%)	44 (80.0)	35 (74.4)	44 (95.6)	44 (95.6)	36 (83.7)	21 (95.5)	13 (92.8)	9 (90.0)
Duration (year)	--	13.6 ± 8.4	13.9 ± 9.4	9.2 ± 5.5	12.2 ± 10.2	9.4 ± 10.9	9.6 ± 7.2	8.8 ± 7.3

Table 2S. Parameters optimized by IPO that were used for peak peaking, alignment and grouping carried out with the XCMS package. *These parameters were fixed and therefore, not optimized by IPO.

Parameters		Urine	Plasma
Peak Picking	method	*centwave*	*centwave*
	min_peakwidth	11.2	12.45
	max_peakwidth	48.0	35.0
	ppm	29.8	24.0
	mzdiff	0.00725	0.00175
	noise*	1000	1000
	prefilter*	(3,800)	(3,800)
Retention time alignment	method	*obiwarp*	*obiwarp*
	profStep	1.00	0.30
	response	4.42	13.84
	gapInit	0.768	0.352
	gapExtend	2.98	2.44
Grouping	method	*density*	*density*
	bw	5.48	5.00
	minfrac*	0.5	0.5
	mzwid	0.033	0.047

Figure 1S. PCA scores plots of the normalized data of plasma (a) and urine data (b). (Color code by sample categories)

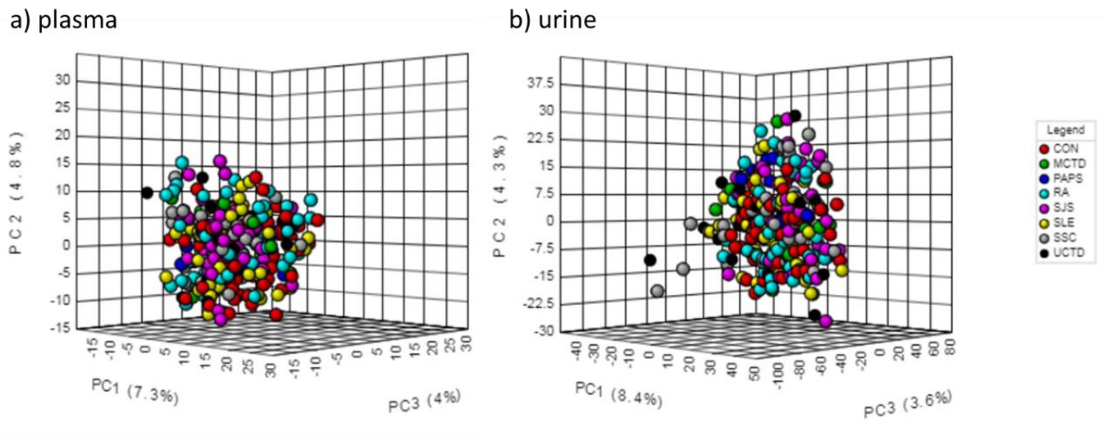


Table 3S. List of the parameters (m/z, RT, and ranks) of unknown compounds selected from the rdCV-MUVR-PLS models created with individual SADs and healthy controls.

Unknowns (Plasma samples)			Unknowns (Urine samples)		
m/z	**RT (min)**	**Diseases (Rank)**	**m/z**	**RT (min)**	**Diseases (Rank)**
175.0341	10.64	SJS (109.8)	157.0609	1.04	SSC(71.6), RA(258.9)
216.0664	12.67	RA(144.1)	187.0585	1.10	RA(40.1)
199.1309	13.04	RA(74.6)	278.1235	1.32	SSC(9.3), SLE(36.5), SJS(26.9)
601.2651	14.39	SJS (139.3)	142.0862	1.37	RA(187.4)
171.1488	14.73	SJS (103.4), SLE(151.2)	254.0885	1.38	RA(25.9), SSC(28.9), SJS(33.5)
268.0597	15.2	RA(127.5), SLE(114.5)	226.0826	1.48	SLE(5.7), SJS(16.1)
374.1953	15.20	RA(118.4)	241.1553	1.52	SJS(204.7)
475.3245	16.08	RA(129.4)	150.0186	1.63	SSC(184.5)
276.9894	16.17	SJS (82.3), RA(79.5)	305.0978	2.35	SSC(38.1)
398.2388	17.27	SLE(139.9)	238.0934	2.54	RA(38.5)
415.2777	20.47	SJS (31.1)	273.1268	6.95	SLE(52.9)
526.2871	21.69	SJS (155.2)	159.5907	13.89	SLE(30.6)
619.4329	22.80	SJS (137.8)	146.5826	14.02	SLE(79.7)
680.3053	23.07	SJS (162.0)	512.2218	16.62	SLE(85.8)
303.2235	24.64	SLE(73.5)	322.1315	18.35	SLE(29.7)
319.1910	24.65	SJS (50.8), SSC(29.2), SLE(27.4)	276.1262	19.90	SLE(42.5)
461.1812	24.66	SJS (42.4)	175.5557	19.91	SLE(39.4)
396.2137	26.35	SJS (198.1)	366.1577	25.55	SLE(98.3)
364.2483	28.49	SJS (8.8)	484.1482	27.44	SLE(81.0)
617.4690	28.74	SJS (141.1)	350.1628	30.21	SLE(10.7)
769.5841	36.50	SJS (124.2)	332.2433	34.61	SSC(43.5)
806.0530	35.59	SJS (112.8)	277.1412	39.36	SJS(166.3), RA(44.9)
643.5204	41.24	SJS (176.5)	250.1186	40.15	SLE(204.5)
			219.1744	40.56	SLE(217.7)
			535.2877	41.00	RA(11.2), SJS(38.0)
			318.2406	41.85	RA(31.9), UCTD()
			269.2094	42.02	SSC(91.1)
			640.5906	43.38	SJS(29.5)
			522.5991	43.49	RA(92.3)
			282.2800	43.99	SSC(52.9), SJS(36.0)

CONCLUSIONES

CONCLUSIONS

CONCLUSIONES

Durante la presente tesis doctoral, se han desarrollado, optimizado y aplicado diferentes estrategias metabolómicas no dirigidas en el ámbito de los compuestos bioactivos para el estudio de su absorción y metabolismo así como del efecto de su ingesta prolongada en el metabolismo endógeno; y por otra parte, en el ámbito de las enfermedades autoinmunes sistémicas, demostrando su idoneidad y potencialidad en la clasificación de diferentes estados fisiopatológicos. A continuación, se detallan las conclusiones específicas de los diferentes capítulos abordados:

1. Mediante los análisis por HPLC-ESI-QTOF-MS de las muestras biológicas obtenidas en el estudio de perfusión intestinal llevado a cabo en ratas Wistar, se ha profundizado en el conocimiento acerca de la absorción y el metabolismo de los compuestos bioactivos procedentes del romero demostrando su biodisponibilidad y poniendo de manifiesto el mecanismo de metabolización de varios de ellos. De esta forma, se ha dejado patente que los compuestos que presentan mayor biodisponibilidad son los pertenecientes a la familia de diterpenos fenólicos de tipo abietano, a los cuales se le han atribuido, entre otras, capacidad antioxidante y anticancerígena. Además, la detección en plasma de los derivados glucuronizados procedentes de estos diterpenos, ha permitido comprobar que la principal ruta de metabolización de dichos compuestos está basada en reacciones de glucuronización.
2. La aplicación de una estrategia metabolómica no dirigida basada en análisis mediante HPLC-ESI-QTOF-MS en un estudio de intervención nutricional llevado a cabo en ratas Wistar diabéticas a las que se les administró un extracto

bioactivo de mango, permitió identificar en muestras de plasma e hígado varios biomarcadores de los trastornos diabéticos relacionados con el metabolismo de ácidos grasos, aminoácidos, ácidos biliares, nucleótidos y pantotenato. Además, se identificaron dos metabolitos biodisponibles en hígado (euxantona y glutatión) vinculados con la mangiferina, la cual ha sido descrita como uno de los compuestos bioactivos más importantes del mango. Los resultados obtenidos sugieren la mejora del estado antioxidante del hígado de las ratas diabéticas debido al consumo de una dieta rica en compuestos fenólicos procedentes del mango cv. Ataulfo.

3. A través del desarrollo y aplicación de estrategias metabolómicas no dirigidas en un ensayo de intervención nutricional en el que se administró un suplemento alimentario basado en un extracto bioactivo de ajo a voluntarios sanos, se ha comprobado que la ingesta prolongada de dicho suplemento afecta al metabolismo de los fosfolípidos. Entre los principales compuestos alterados destaca un aumento de la concentración de acilcarnitinas, lisofosfatidilcolinas y lisofospatidiletanolaminas, junto con una disminución de fructosaminas después de la ingesta del suplemento. Estas alteraciones detectadas podrían estar altamente asociadas con las propiedades antioxidantes y de antiglicación descritas previamente para los extractos de ajo.
4. El estudio de la esclerosis sistémica mediante un enfoque metabolómico no dirigido en muestras de orina y plasma sanguíneo de voluntarios sanos y enfermos, permitió detectar diferentes metabolitos desregulados en dicha patología, como es el caso de alfa-N-fenilacetil-L-glutamina, acilcarnitinas, acilglicinas, monoacilgliceroles y derivados de aminoácidos. Por lo tanto, los

mecanismos metabólicos afectados en pacientes con esclerodermia se pueden relacionar con reacciones de β-oxidación de ácidos grasos así como con el metabolismo de aminoácidos (prolina, histidina y glutamina). Además, la alteración del compuesto 2-araquidonilglicerol en las muestras de plasma sugiere la desregulación del sistema endocannabinoide en los pacientes de esclerodermia.

5. La integración de los datos adquiridos por HPLC-ESI-QTOF-MS en los análisis de las muestras de orina y plasma sanguíneo de pacientes con síndrome de Sjögren primario e individuos sanos mediante un modelo PLS-DA mostró una mejor clasificación entre ambos grupos que los modelos multivariantes de cada matriz biológica por separado. Además, la metodología optimizada permitió detectar varios metabolitos alterados en estas muestras biológicas, como el caso de los ácidos grasos insaturados, fosfatidilinositoles, acilglicinas, lisofosfatidilcolinas, acilcarnitinas y derivados de varios aminoácidos como el triptófano, prolina y fenilalanina.
6. Tras la optimización y comparación de herramientas de pre-procesamiento de datos basadas en un software comercial y en programas de acceso libre basados en paquetes desarrollados en lenguaje R, se ha demostrado que esta última metodología es la opción más adecuada para estudios de metabolómica no dirigida que involucran un número elevado de muestras. De hecho, esta metodología tiene una mayor capacidad para normalizar las derivas analíticas producidas generalmente entre y a lo largo de las diferentes secuencias analíticas de HPLC-ESI-QTOF-MS. Sin embargo, este entorno es menos intuitivo y requiere un cierto nivel de experiencia por los usuarios. Debido a estas

limitaciones y para facilitar su aplicación a usuarios principiantes se ha desarrollado una guía detallada de los pasos a seguir para la utilización de este entorno.

7. La clasificación entre individuos sanos y pacientes de siete enfermedades autoinmunes sistémicas mediante modelos multivariantes utilizando datos adquiridos por HPLC-ESI-QTOF-MS de muestras de orina y plasma, mostraron buenos resultados clasificatorios, siendo los metabolitos más influyentes los ácidos grasos insaturados, acilcarnitinas, acilglicinas y varios aminoácidos. Además, el metabolito N6-carbamoil-L-treoniladenosina se ha propuesto como un potencial biomarcador de este conjunto de enfermedades, al detectarse niveles superiores en las muestras de los sujetos sanos respecto a las de los individuos con las diferentes patologías. Sin embargo, los modelos multivariantes mostraron ciertas limitaciones para diferenciar la enfermedad del tejido conectivo indiferenciado de sus patologías relacionadas (SLE, RA, SJS, SSC, MCTD) a partir de los datos metabolómicos adquiridos. Por ello, sería necesario continuar estudiando e investigando este tipo de patologías mediante enfoques que integren datos obtenidos por diferentes metodologías ómicas.

CONCLUSIONS

During the present PhD thesis, different untargeted metabolomic strategies have been developed, optimized and applied in two different fields of knowledge. On the one hand, the absorption and metabolism of bioactive compounds as well as the effect of their prolonged intake on endogenous metabolism have been studied. On the other hand, untargeted methodologies have been performed in the field of systemic autoimmune diseses, demonstrating their suitability and potential for the classification of different pathophysiological states. The specific conclusions of the different chapters are detailed below:

1. The knowledge about the absorption and metabolism of bioactive compounds from rosemary has been increased through HPLC-ESI-QTOF-MS analysis of the biological samples obtained in the intestinal perfusion study conducted in Wistar rats. In fact, the biovailabily and metabolism mechanisms of several compounds have been demonstrated. In this way, abietene-type diterpenes, whose antioxidant and anticancer properties have been demonstrated, are the compounds that presented the greatest bioavailability. In addition, their glucuronized derivatives were detected in plasma samples verifying the glucuronization reactions as the main route of metabolization of these compounds.
2. The application of an untargeted metabolomic strategy based on HPLC-ESI-QTOF-MS analysis in a nutritional intervention study carried out in diabetic Wistar rats that ingested a bioactive extract of mango, allowed to identify in liver and plasma samples several biomarkers of diabetic disorders. These

metabolites were related to the metabolism of fatty acids, amino acids, bile acids, nucleotides and pantothenate. In addition, two metabolites, euxanthone and glutathione, were identified bioavailable in liver. These metabolites are linked to mangiferin, which has been descrided as one of the most important bioactive compounds in mango. The obtained results suggest the improvement of the antioxidant state of the liver of diabetic rats due to the consumption of a diet rich in phenolic compounds from mango cv. Ataulfo.

3. The development and application of an untargeted metabolomic strategy in a nutritional intervention assay, in which a food supplement based on a bioactive garlic extract was administered to healthy volunteers, has shown that prolonged intake of the supplement affects the phospholipid metabolism. Among the main altered metabolites, an increase in the concentration of acylcarnitines, lysophosphatidylcholines and lysophospatidylethanolamines was detected togheter with a decrease in frucosamines concentrations after the intake of the supplement. These alterations could be highly associated with the antioxidant and antiglication properties, which have been previously described for garlic extracts.
4. The study of systemic sclerosis by means of an untargeted metabolomic approach in urine and plasma samples from patients and healthy volunteers, allowed the detection of different deregulated metabolites in this disease. The main altered compounds were alpha-N-phenylacetyl-L-glutamine, acylcarnitines, acylglycines, monoacylglycerols and metabolites derived from amino acids. Therefore, the metabolic pathways affected in scleroderma patients seemed to be fatty acid beta-oxidation and aminoacid (proline,

histidine and glutamine) pathways. In addition, it has also been confirmed the deregulatons in endocannabinoid system by the alteration of 2-arachidonylglycerol in plasma samples.

5. The integrated PLS-DA model using aggregated HPLC-ESI-QTOF-MS data from urine and plasma analysis of healthy controls and Sjögren’s syndrome patients showed better results than multivariate models built considering data from each single biofluid separately. In addition, the optimized methodology put into light that the most deregulated metabolites in these biological samples were unsaturated fatty acids, phosphatidylinositols, acylglycines, lysophosphatidylcholines, acylcarnitines and metabolites related to amino acid pathways, specially tryptophan, proline and phenylalanine metabolism.
6. After the optimization and comparison of data pre-processing tools based on commercial software and open source based on packages developed in R language, it has been highlighted that open source methodology is the most suitable option for untargeted metabolomic studies that involve a large number of samples. In fact, the open source methodology is, to a much higher degree able to correct the large between- and within-batch effects produced in HPLC-ESI-QTOF-MS analyses. On the contrary, the R environment is less intuitive, frequently with a distinctly steeper learning curve. For these reasons, a detailed tutorial was provided to help users of commercial software to start processing data through R-based methodology.
7. The classification between healthy individuals and patients of seven systemic autoimmune diseases by multivariate models using data acquired by HPLC-ESI-QTOF-MS from urine and plasma samples, showed good classification results.

The most important metabolites found in the models were unsaturated fatty acids, acylcarnitines, acylglycins and several amino acids. In addition, the metabolite N6-carbamoyl-L-threoniladenosine has been proposed as a potential biomarker of this set of diseases since higher levels were detected in healthy subjects compared to those of individuals with different pathologies. However, multivariate models showed certain limitations to differentiate undifferentiated connective tissue disease from its related pathologies (SLE, RA, SJS, SSC, MCTD) using the acquired metabolomic data. Therefore, it would be necessary to continue studying and investigating this type of pathologies through approaches that integrate data obtained by different omic methodologies.