Programa de Doctorado en Fisiología

Departament de Fisiologia – Facultat de Medicina i Odontologia

ESTRATEGIAS PARA EL ANÁLISIS DE DATOS METABOLÓMICOS DIRIGIDOS AL DIAGNÓSTICO CLÍNICO

TESIS DOCTORAL presentada por:

Clara Pérez Rambla

Dirigida por:

Dr. Antonio Pineda Lucena

Dra. Leonor Puchades Carrasco

Valencia, febrero 2018

Antonio Pineda Lucena, Doctor en Ciencias Químicas, Investigador Principal, Unidad Descubrimiento de Fármacos, Instituto de Investigación Sanitaria La Fe.

Leonor Puchades Carrasco, Doctora en Bioquímica, Investigadora, Programa Metabolómica, Centro investigación Príncipe Felipe.

CERTIFICA/N:

Que la presente memoria, titulada "ESTRATEGIAS PARA EL ANÁLISIS DE DATOS METABOLÓMICOS DIRIGIDOS AL DIAGNÓSTICO CLÍNICO", corresponde al trabajo realizado bajo su dirección por D/Dña. **Clara Pérez Rambla**, para su presentación como Tesis Doctoral en el Programa de Doctorado en Fisiología de la Universitat de València.

Y para que conste firma/n el presente certificado en Valencia, 20 de Diciembre de 2017

Dr. Antonio Pineda Lucena, DNI: 30517484A

Dra. Leonor Puchades Carrasco, DNI: 24373634M

Dr. Federico Vte. Pallardó Calatayud, tutor de la tesis, da el visto bueno.

AGRADECIMIENTOS

Me gustaría dar las gracias, en primer lugar, a Antonio Pineda, quien en el 2013 2013 me dio la oportunidad de unirme a su grupo de investigación y empezar esta interesante aventura en el mundo de la metabolómica. Me gustaría mostrar mi gratitud hacia él en el campo científico, profesional y personal. En segundo lugar, me gustaría agradecer a Leonor Puchades todo el apoyo y confianza que me ha dado en estos años, todo lo que me ha enseñado y la paciencia que ha tenido en esta recta final. Ha sido un placer contar con tu experiencia y tus conocimientos.

Seguidamente, he de destacar el apoyo y soporte del resto de personas del equipo del CIPF como Martina, Leti, Sara y Paco. Muchas gracias por vuestra confianza. También me gustaría dar las gracias al equipo de la Red Valenciana del Biobanco: Jacobo, Carol, Andrea y Lidia. Con ellos empezó mi primer proyecto y ha sido un placer colaborar con gente tan predispuesta a ayudar y agradable.

Agradecer también al grupo del Dr. Jose Antonio López del IVO y al grupo del Dr. Carlos Camps y la Dra. Eloisa Jantus del Hospital General de Valencia con los que he colaborado en diversos proyectos que forman parte de esta tesis.

A mis compañeros del Laboratorio del Hospital Dr. Peset de Valencia, donde me forme como residente y me apoyaron en mis inicios como investigadora.

Finalmente y no por ello menos importante, me gustaría agradecer el cariño y la comprensión recibida por mi familia. Desde el primer momento me han dado soporte para llegar a donde he llegado. También agradecer a Carlos su apoyo y su positivismo en todo, a pesar de estar alejado del mundo de la ciencia, siempre ha sabido cómo ayudarme a avanzar. Y también a nuestro hijo Adrián que desde pequeño ha sabido estar a la altura y respetar mis tiempos de concentración.

Finalmente dedicar unas palabras a todas las persona que de una manera u otra se ven afectadas por el Cáncer, esta tesis supone un pequeño granito para avanzar en la lucha contra esta enfermedad.

ABREVIATURAS

ABREVIATURAS

1D	Experimento monodimensional
^{1}H-RMN	Resonancia magnética nuclear de protón
2D	Experimento bidimensional
AACR	Aminoácidos de cadena ramificada
AUC	*Area under the curve* Área bajo la curva
BC	Bronquitis crónica
CaP	Cáncer de próstata
CoeffCS	*Regression coefficient centered and scaled* Coeficiente de regresión centrado y escalado
CoeffCScvSE	*Regression coefficient centered and scaled cross validated standard error* Error estándar del coeficiente de regresión tras la validación cruzada
CP	Cáncer de pulmón
CPMG	Carr Purcell Meiboom y Gill
CPNM	Cáncer de pulmón no microcítico
CV	Coeficiente de variación
Da	Dalton
EDTA	Ácido etilendiaminotetraacético
EM	Espectrometría de masas
EPOC	Enfermedad pulmonar obstructiva crónica
FID	*Free induction decay* Decaimiento de inducción libre
GNMT	Glicina-N-metiltransferasa
HBP	Hiperplasia benigna de próstata
HMDB	*Human metabolome database* Base de datos del metaboloma humano
Hz	Hercios
LDL	*Low density lipoproteins*

	Lipoproteínas de baja densidad
OR	*Odds ratio* Razón de probabilidades
NOESY	*Nuclear Overhauser effect spectroscopy* Espectroscopia de efecto nuclear Overhauser
OPLS-DA	*Orthogonal partial least square-discriminant analysis* Análisis discriminante de mínimos cuadrados con corrección ortogonal
PCA	*Principal component analysis* Análisis de componentes principales
PCs	*Principal components* Componentes principales
PLS-DA	*Partial least square-discriminant analysis* Análisis discriminante por mínimos cuadrados
PPM	Partes por millón
PQN	*Probabilistic quotient normalization* Normalización con cocientes probabilísticos
PSA	*Prostatic specific antigen* Antígeno prostático específico
RMN	Resonancia magnética nuclear
ROC	*Receiver operating curve* Característica operativa del receptor
SARDH	Sarcosina deshidrogenasa
S.E.M	*Standard error of the mean* Error estándar de la media
SUS-plot	*Shared and unique structures plot* Gráfico de características compartidas y únicas
SVA	*Surrogate variable analysis* Análisis de variable subrogada
TB	Tuberculosis
TF	Transformada de Fournier

TSP	3-trimetilsilil propionato
UV	*Unit variance* Varianza unitaria
UPLC-ToFMS	*Ultra performance liquid chromatography coupled to time-of-flight mass spectrometry* Cromatografía líquida de ultra-alta resolución acoplada a espectrometría de masas
VIP	*Variable importance in the projection* Importancia de la variable en la proyección
VLDL	*Very low-density lipoprotein* Lipoproteína de muy baja densidad

I. INTRODUCCIÓN

1. BIOLOGÍA DE SISTEMAS

Los organismos vivos, incluso los más sencillos, son sistemas de gran complejidad cuando se analizan sus componentes. Todos los procesos celulares, a pesar de estar muy regulados, están influenciados por el ambiente y por las reacciones internas que ocurren en él. Para entender mejor esta complejidad se hace necesario aplicar un enfoque multidisciplinar (Lindon *et al*, 2011).

La Biología de Sistemas permite analizar de forma holística el funcionamiento de todos los componentes celulares (proteínas, genes, metabolitos, etc.) y profundizar en el conocimiento de cómo sus interacciones internas y con otros sistemas conducen a la aparición de nuevas propiedades (Joyce & Palsson, 2006).

Entre las disciplinas que contribuyen a la Biología de Sistemas destacan la genómica, la transcriptómica, la proteómica y la metabolómica. Todas ellas aportan una gran cantidad de datos, proporcionando un conocimiento más amplio de las funciones biológicas desde distintos niveles de organización biológica (**Figura 1**).

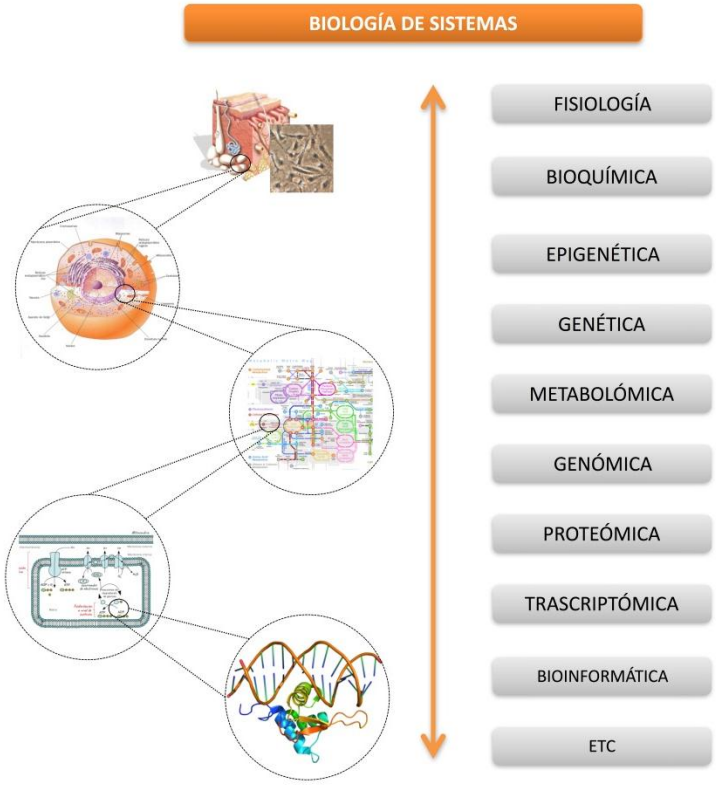


Figura 1: Diagrama de las disciplinas que contribuyen a la Biología de Sistemas.

1.1 Metabolómica

En el contexto de la Biología de Sistemas, la metabolómica es considerada, a nivel bioquímico, como el final de la cascada "ómica" (genómica → transcriptómica → proteómica → metabolómica). La metabolómica mide la respuesta metabólica dinámica y global de los sistemas biológicos frente a estímulos biológicos o manipulaciones genéticas, indica lo que realmente ha pasado y proporciona información del fenotipo de los seres vivos. El análisis metabolómico consiste en la identificación y cuantificación sistemática de los metabolitos en células, tejidos o fluidos.

Los metabolitos que se producen en el curso de las reacciones metabólicas que tienen lugar en los seres vivos son moléculas de bajo y medio peso molecular (<1.500 Da). Los metabolitos sirven como indicadores directos de la actividad bioquímica celular. Son el producto final de procesos de regulación celular y sus niveles pueden ser considerados como la última respuesta de un sistema biológico a cambios genéticos o ambientales, de manera que permiten una interpretación global (Lindon *et al*, 2011).

Por extensión, el metaboloma se define como el conjunto dinámico de pequeñas moléculas presentes en un organismo vivo, sean sintetizados *de novo* por el propio organismo o incorporados desde el exterior. Así pues, no es de extrañar que el estudio del metaboloma se esté utilizando cada vez más en el diagnóstico clínico, ya que el perfil metabolómico ofrece una oportunidad para interrogar el funcionamiento celular y los mecanismos bioquímicos que ocurren en la célula, para así relacionarlos con el fenotipo observado (Fiehn, 2002; Spratlin *et al*, 2009; Weckwerth, 2003).

1.2 Aplicaciones metabolómicas

Durante las últimas dos décadas las aplicaciones de la metabolómica se han centrado en diferentes áreas (Puchades-Carrasco & Pineda-Lucena, 2015): la biotecnología de plantas (Gomez-Casati *et al*, 2013), la microbiología (Aldridge & Rhee, 2014), la nutrición (Astarita & Langridge, 2013) y el campo preclínico (Anderson & Kodukula, 2014; Lindon *et al*, 2007) y clínico (Armitage & Barbas, 2014), entre otras.

En el ámbito de la biomedicina, el objetivo de los estudios de metabolómica es analizar las fluctuaciones en los niveles de los metabolitos presentes en una muestra biológica y buscar correlaciones sobre la existencia o no de una patología, la respuesta a un tratamiento farmacológico, etc. La metabolómica permite caracterizar los procesos bioquímicos que ocurren en la enfermedad, pudiendo aportar información

de gran utilidad en el diagnóstico de la enfermedad, la monitorización de los pacientes, o la evaluación de la respuesta al tratamiento en diferentes patologías.

La metabolómica hace uso de diferentes estrategias de análisis, dependiendo del objetivo particular del estudio:

-Análisis dirigido: se aplica cuando existe un conocimiento previo de los metabolitos implicados en el proceso biológico de estudio. Es un enfoque selectivo, condicionado por el objeto de estudio, y que generalmente implica una ruta bioquímica concreta (Dudley *et al,* 2010). Cuando se utiliza el enfoque dirigido todas las etapas de experimentación y análisis de los datos se optimizan para la detección y cuantificación de dichos compuestos.

-Análisis no dirigido: se emplea cuando se quiere medir y comparar de forma simultánea el mayor número de metabolitos posibles, sin ser necesario un conocimiento previo sobre los metabolitos presentes en la muestra biológica. En este tipo de estudios se caracteriza el perfil metabolómico de las muestras analizadas. En este enfoque, los datos obtenidos son más complejos y extensos, obteniendo una visión global del metaboloma, incluidos aquellos procesos que son desconocidos o están poco caracterizados (Patti *et al*, 2012).

1.3 Metabolómica y cáncer

El cáncer se describe como una enfermedad compleja donde una célula sana se transforma en una célula tumoral. La enfermedad se caracteriza, entre otras modificaciones, por la alteración del metabolismo celular y del microambiente tumoral (Wishart *et al*, 2016). A principios del siglo XX ya se demostró que las células cancerosas presentan un fenotipo metabólico distinto, consumiendo más cantidades de glucosa que las células sanas. El aumento de la glicolisis aerobia, conocido como efecto Warburg (Warburg *et al*, 1927) se ha identificado y estudiado en muchos tipos de células tumorales.

En los últimos años, se han publicado numerosos estudios en los que se ha demostrado la utilidad de la metabolómica en la investigación del fenotipo metabólico del cáncer. Se ha demostrado cómo la célula tumoral utiliza en mayor medida metabolitos como la glucosa y el glutamato para producir energía y sintetizar carbohidratos, ácidos grasos, aminoácidos y nucleótidos que son necesarios para la síntesis de proteínas y para su proliferación celular (Jimenez *et al*, 2013; Jobard *et al*, 2014; Kim & Dang, 2006; Tan *et al*, 2013; Vander Heiden, 2011).

Estos estudios han permitido la identificación de nuevos biomarcadores relacionados con distintas patologías oncológicas y el descubrimiento de los

denominados "oncometabolitos" (Wishart, 2015). Los oncometabolitos son metabolitos endógenos que se producen como consecuencia del desarrollo de un tumor y del proceso de metástasis. El primer oncometabolito identificado fue el 2-hidroxiglutarato, encontrado en concentraciones altas en pacientes con gliomas (Ward *et al*, 2010).

A partir de este descubrimiento, se identificaron otros oncometabolitos como el fumarato (carcinoma de células renales), el succinato (paraganglioma), la sarcosina (cáncer próstata), la asparagina (leucemia), la colina (cáncer de próstata, cerebro y mama) y las poliaminas (muchos tipos de cánceres). Todos estos metabolitos participan en rutas metabólicas claves en el cáncer, incluyendo la glicolisis aerobia, glutaminólisis y el metabolismo de un carbono (Wishart, 2015).

Los avances en metabolómica permiten un conocimiento más completo de los aspectos relacionados con el metabolismo de la célula tumoral y una mejor comprensión de los mecanismos claves en el desarrollo del cáncer. Por tanto, la investigación en este campo podría contribuir al desarrollo de nuevas estrategias terapéuticas contra el cáncer basadas en un conocimiento más profundo, con resultados más efectivos y menos efectos adversos (Beger, 2013).

1.4 Plataformas analíticas

La realización de estudios de metabolómica requiere del uso de plataformas analíticas que posean características tales como la reproducibilidad, la sensibilidad, la precisión y la sencillez (Lu *et al*, 2008). Sin embargo, es complicado encontrar una técnica que reúna todas estas características. Las principales técnicas analíticas empleadas en metabolómica están basadas en la espectroscopia de Resonancia Magnética Nuclear (RMN) y en la Espectrometría de Masas (EM). Las dos técnicas poseen diferentes ventajas y limitaciones desde el punto de vista técnico, ofreciendo información complementaria (Lenz & Wilson, 2007).

La EM es una técnica con gran sensibilidad y selectividad, que permite la identificación y cuantificación de moléculas, basada en su relación masa/carga (m/z), en diferentes matrices (líquidas, sólidas y gaseosas). Para conseguir la separación de las moléculas, previa al análisis, es necesario acoplar esta técnica a otras técnicas cromatográficas como la cromatografía líquida-espectrometría de masas (CL-EM), la cromatografía de gases-espectrometría de masas (CG-EM) o, en menor medida, la electroforesis capilar-espectrometría de masas (EC-EM). La EM permite analizar una gran cantidad de metabolitos con sensibilidades del orden picomolar y además, eligiendo el método de separación adecuado, se consigue una gran selectividad. Sin embargo, existen ciertas limitaciones como son la falta de estandarización de los

métodos, la baja reproducibilidad, un tratamiento de muestra tedioso y la necesidad de usar patrones internos para la identificación inequívoca de compuestos.

La RMN es una técnica cuantitativa que utiliza las propiedades magnéticas de los átomos para dilucidar la estructura química de los compuestos. Como inconveniente, destaca su baja sensibilidad, del orden micromolar, y el solapamiento de señales que dificulta la identificación de metabolitos. A pesar de eso, es considerada una potente herramienta analítica que permite la detección simultánea de un amplio rango de metabolitos estructuralmente diversos de forma fiable y repetitiva. La preparación de la muestra es mínima y, empleando espectrómetros de alto campo con un alto nivel de automatización (p.ej., 600 MHz), se requieren tiempos de medida que en la mayoría de los casos no superan los 4-5 minutos.

1.4.1 Principios básicos de RMN

La RMN es un tipo de espectroscopia basada en la absorción de radiación electromagnética. Su uso en metabolómica se fundamenta en la propiedad que presentan algunas moléculas de poseer núcleos atómicos magnéticamente activos. En este sentido, en metabolómica por RMN, el núcleo más utilizado es el protón ya que es un átomo magnéticamente activo y de gran abundancia natural (~99.9%).

En presencia de un campo magnético externo, los núcleos atómicos experimentan un desdoblamiento de niveles de energía, dando lugar a estados de baja y alta energía. En estas condiciones, la aplicación de un pulso de radio frecuencia, perpendicular a la dirección del campo magnético, provoca transiciones energéticas. La energía de la radiación necesaria para producir este salto de nivel dependerá del tipo de núcleo, del entorno químico de éste, del tipo de núcleos presentes en sus cercanías y del campo externo aplicado.

Cuando el pulso de radiofrecuencia finaliza, los núcleos excitados liberan el exceso de energía regresando al estado de equilibrio, fenómeno conocido como relajación. La vuelta al estado de equilibrio se caracteriza por la desaparición de la magnetización, dando lugar a lo que se conoce como *Free Induction Decay* (FID), la cual proporciona información sobre la muestra irradiada. La FID es una función dependiente del tiempo que puede transformarse en un espectro de señales dependiente de las frecuencias utilizando una función matemática, la transformada de Fournier (TF) (**Figura 2**).

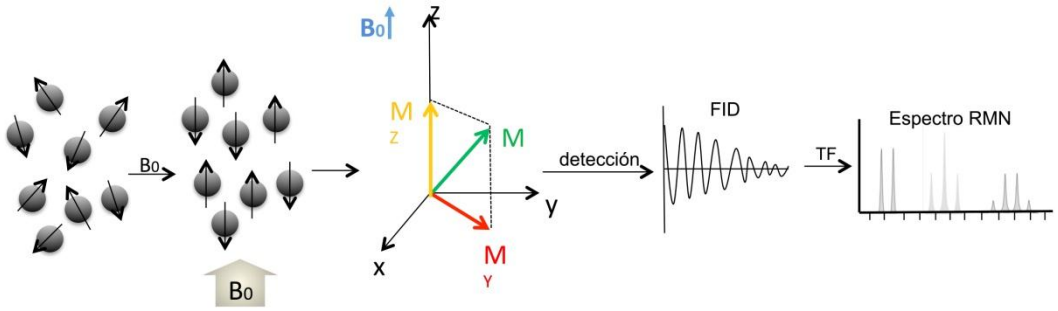


Figura 2: Esquema de las distintas fases en la adquisición de un espectro de RMN.

1.4.2 Características de los espectros de RMN

Un espectro de RMN consiste en una representación gráfica donde la posición en la que aparece la señal permite identificar un átomo o grupo de átomos en la molécula y el área bajo la señal es proporcional a la concentración de la misma.

Una de las características más importantes de los espectros de RMN es el desplazamiento químico (δ). Las unidades del desplazamiento químico se expresan en partes por millón (ppm), es una escala universal que permite comparar espectros registrados en equipos de diferente intensidad de campo magnético.

El desplazamiento químico varía en función de la densidad electrónica que rodea al núcleo. Los protones que forman parte de moléculas orgánicas no se encuentran aislados sino que están influenciados por su entorno químico. En situaciones donde la densidad electrónica en el entorno de un núcleo es alta, éste aparece en el espectro de RMN a desplazamientos químicos pequeños. En cambio, cuando la densidad electrónica alrededor del núcleo es pequeña, éste aparece a desplazamientos químicos más altos. El desplazamiento químico se utiliza para la identificación de grupos funcionales.

Otras características del espectro de RMN que aportan información sobre la muestra analizada son la integral o área bajo la señal y la multiplicidad de la misma. El área bajo la señal proporciona información sobre el número de protones que integran la señal y se utiliza con fines cuantitativos. La multiplicidad corresponde al número de picos de una misma señal y ayuda a conocer qué grupos funcionales están unidos entre sí. Esto es debido a que el campo magnético que influye sobre un núcleo concreto está afectado también por el campo que crean los núcleos vecinos. De esta manera, la señal correspondiente a un núcleo concreto se desdobla en (n+1) picos, donde “n” es el número de núcleos equivalentes vecinos que dan lugar al desdoblamiento de una señal (**Figura 3**).

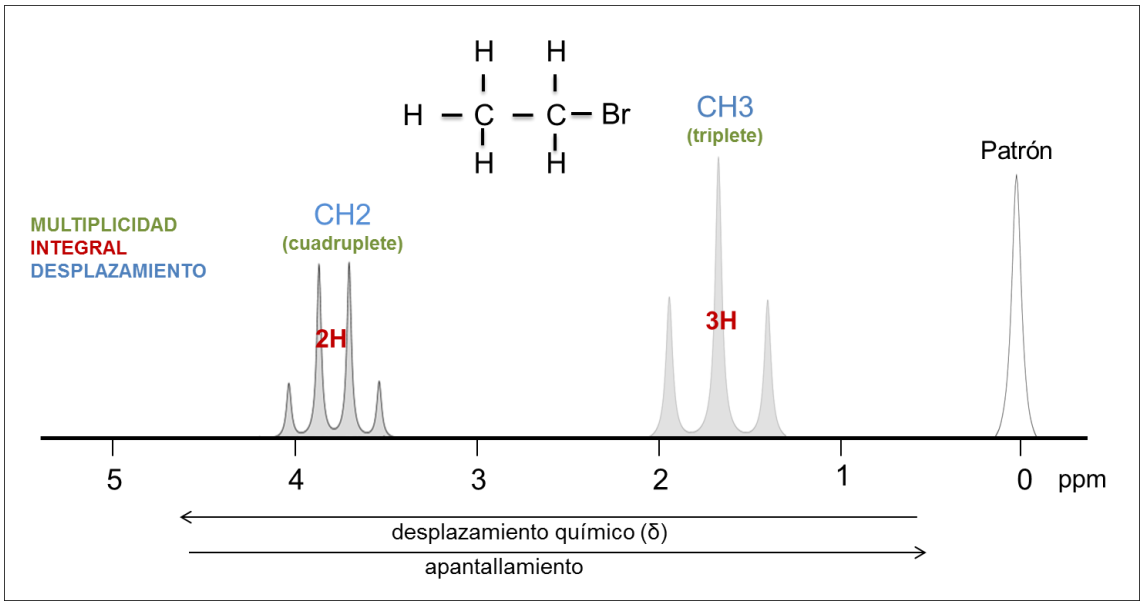


Figura 3 Espectro del etilo de bromidio. El desplazamiento químico a 3,75 ppm del grupo -CH_2 es consecuencia del apantallamiento del bromo y aparece como cuadruplete por la presencia del grupo metilo adyacente. El grupo metilo aparece a 1,5 ppm y como un triplete debido al metileno adyacente. El área bajo la señal o integral corresponde al número de protones que integral la señal.

2. FASES DE UN ESTUDIO METABOLÓMICO

Las distintas etapas que se utilizan generalmente cuando se llevan a cabo estudios metabolómicos tienen una gran importancia para lograr resultados relevantes y de buena calidad. El diagrama de flujo que se aplica comúnmente en metabolómica por RMN engloba cuatro pasos: diseño del estudio y preparación de la muestra, adquisición de espectros, análisis de los datos e interpretación biológica (**Figura 4**).

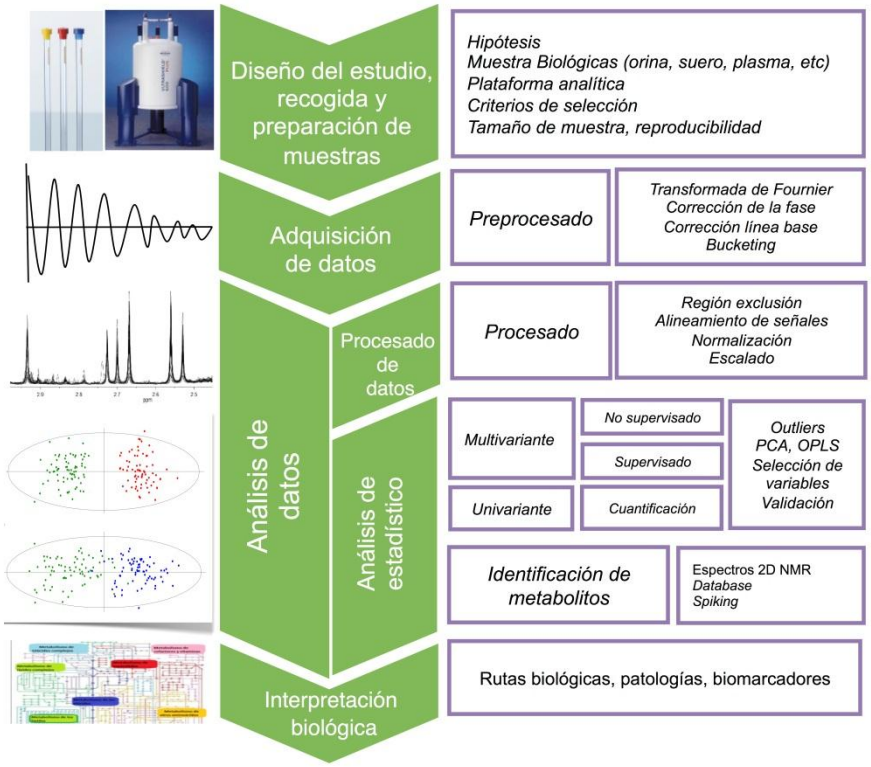


Figura 4: Esquema completo del flujo de trabajo de un estudio metabolómico por RMN.

2.1 Diseño del estudio

En los estudios de metabolómica clínica es esencial contar con un buen diseño del estudio, adaptado a la situación clínica a estudiar y acorde con los objetivos planteados. Es necesario plantear una serie de cuestiones científicas a las cuales se pretende dar respuesta a lo largo del proyecto. Por ejemplo, en el diseño de un estudio para el descubrimiento de biomarcadores, es necesario conocer qué tipo de patología vamos a estudiar, qué posibles rutas metabólicas están implicadas, qué tipo de muestra es la más adecuada para el análisis, el número de pacientes y/o de controles necesarios, las características de los individuos y/o pacientes (edad, sexo o raza), etc.

Las variables más importantes que hay que tener en cuenta a la hora del diseño del estudio son:

-Grupos de estudio: en metabolómica predominan los estudios caso-control, donde se comparan pacientes con una condición médica específica con individuos control que no tienen esa condición. También se realizan estudios con cohortes longitudinales donde el paciente es su propio control; en este caso, se consigue reducir la variabilidad interindividual (Baumgartner *et al*, 2011).

Los criterios de inclusión/exclusión definidos para el estudio deben ser rigurosos y precisos. Así, se consigue una población lo más homogénea posible evitando la

variabilidad en los grupos de muestras debida a factores como el estilo de vida, la dieta, el hábito tabáquico, el consumo de medicamentos, la presencia de otras enfermedades crónicas, el ejercicio, etc.

-Tipo de muestra: la selección del tipo de muestra es clave para obtener resultados óptimos. La muestra elegida debe reflejar la condición que se pretende estudiar. En los estudios de metabolómica clínica se emplean diferentes tipos de muestras biológicas: la orina, el plasma, el suero, la saliva, el fluido seminal, el líquido cefalorraquídeo, el tejido, etc (Claus & Swann, 2013; Lindon *et al*, 2000).

La orina y la sangre (suero y plasma) son las muestras que con más frecuencia se utilizan en los estudios de metabolómica clínica, ya que su obtención es sencilla y poco invasiva para el paciente. En la búsqueda de nuevos biomarcadores estas características son importantes para su posterior traslación a la práctica clínica.

La sangre es una fuente de metabolitos que refleja lo que ocurre en los diferentes procesos bioquímicos, es una ventana metabólica que informa sobre el estado fisiológico del organismo en el momento en que se obtiene la muestra. Las alteraciones del perfil metabólico del plasma o del suero surgen cuando la función homeostática de la sangre es alterada o dañada (Claus & Swann, 2013). Tanto el plasma como el suero derivan de la sangre total, proporcionando información de gran utilidad clínica sobre el estado de un organismo.

La orina es un biofluido que contiene productos de desecho del organismo, cuya composición metabólica es el resultado global de las actividades que ocurren en el mismo. La orina no tiene un mecanismo de regulación homeostática, lo cual la convierte en un medio con un elevado número de biomarcadores con respecto a otros biofluidos (Gao, 2013). Es producida por el riñón y recoge los productos de desecho del cuerpo humano desde la sangre. Contiene información no solo de los riñones y del tracto urinario, sino también de órganos distantes que, vía plasmática, son recogidos a través de la filtración glomerular. Es un biofluido muy utilizado en estudios de metabolómica (Carrola *et al*, 2011; Salek *et al*, 2007; Shi *et al*, 2016).

El uso de la orina como biofluido en los estudios de metabolómica por RMN presenta algunas dificultades en comparación con la sangre. Este hecho hace necesario el uso de técnicas específicas de pre-tratamiento de los datos (Čuperlović-Culf, 2012). Por un lado, la concentración de proteínas y metabolitos en la orina varía a lo largo del día en función de la ingesta de líquidos y de la dieta. Y por otro lado, pueden existir fuentes de variabilidad que dificultan la interpretación de las señales de los espectros, como son los cambios de pH y la abundancia de compuestos que

pueden encontrarse en este biofluido. Esto dificulta la identificación y cuantificación de los metabolitos, bien por solapamientos de las señales o por los desplazamientos de las señales en el espectro.

-Tamaño muestral: el número de individuos incluidos en el estudio influye en la potencia estadística del mismo. La prevalencia de la enfermedad y la heterogeneidad dentro de los diferentes grupos de muestras incluidos en el estudio pueden suponer una limitación importante a la hora de escoger el tamaño muestral en los estudios de metabolómica. Las características individuales de los pacientes (edad, sexo, estilo de vida, ingesta de fármacos, subtipo de enfermedad, grado de evolución de la enfermedad, etc) juegan un papel importante a la hora de poder seleccionar y recoger muestras representativas de cada grupo, influyendo así de forma significativa en los resultados del estudio (Lawton *et al*, 2008). Una estrategia para compensar la posible heterogeneidad dentro de los grupos de estudio es aumentar el tamaño muestral.

2.2 Recogida de muestras

La recogida de muestras es una etapa que debe estar estandarizada para garantizar la reproducibilidad, disminuir la variabilidad no relacionada con la variable de estudio y reducir el riesgo de contaminación. Para cada tipo de muestra, debido a sus características físicas, el proceso de recogida necesita unas condiciones preanalíticas determinadas para mantener su estabilidad.

En el caso de la sangre, la recogida de muestra es diferente dependiendo de si la muestra que se quiere obtener es suero o plasma. El suero se obtiene tras la coagulación de la sangre. Los coágulos de fibrina formados durante la coagulación, junto con las células sanguíneas y los factores de coagulación, se separan del suero por centrifugación. El plasma es un compuesto líquido y acelular de la sangre similar al suero pero, a diferencia de éste, contiene los factores de coagulación. La elección del tubo de extracción se hace, por tanto, dependiendo del tipo de biofluido que se desa analizar. Si se quiere obtener plasma hay que utilizar tubos con anticoagulante, EDTA o heparina.

Entre los factores preanalíticos que influyen en el proceso de recogida y de extracción de las muestras sangre para estudios de metabolómica, el tubo de recogida es fundamental. Los anticoagulantes son sustancias químicas que impiden o retrasan la coagulación de la sangre, de manera que facilitan la manipulación, el fraccionamiento de la sangre y el análisis de la muestra. Los anticoagulantes más utilizados para la toma de muestras de sangre son EDTA, heparina o citrato. El tubo

escogido para la recogida de la muestra está condicionado por la técnica utilizada para el análisis y en función de si se quiere analizar suero o plasma.

La recogida de muestra de orina es la menos invasiva. Cuando se trabaja con muestras de orina es importante mantener su estabilidad y evitar su contaminación. A veces, es necesario añadir algún buffer de estabilización o algún agente antimicrobiano.

Hay otros factores preanalíticos, como son el procesado de la muestra o el almacenamiento de las mismas, que también deben ser controlados para garantizar unos resultados reproducibles. La falta de estandarización en estos procedimientos puede introducir una variabilidad significativa en la composición molecular de las muestras biológicas y, en consecuencia, interferir en el resultado experimental o afectar a su reproducibilidad (Lippi *et al*, 2006).

2.3 Preparación de muestras

La preparación de las muestras consiste en acondicionar el biofluido para facilitar su análisis mediante la técnica seleccionada. En el caso de la RMN, el procedimiento de preparación de la muestra debe reunir los siguientes criterios: reproducibilidad, facilidad, mínima variabilidad entre muestras, calidad espectral (resolución y sensibilidad), ausencia de artefactos analíticamente detectables, etc.

Generalmente, muestras como el plasma, suero u orina no requieren de procedimientos sofisticados en la preparación para su análisis por RMN. La adición de un buffer es suficiente, en la mayoría de los casos, para reducir posibles variaciones debidas a diferencias en el pH o a la viscosidad de las muestras (Beckonert *et al,* 2007). En cambio, cuando se trabaja con tejidos o células, la preparación de este tipo de muestras precisa del uso de disolventes orgánicos y agua para obtener dos fases: la fase acuosa, que contiene los metabolitos polares y la fase orgánica con los metabolitos apolares (p.ej., lípidos).

2.4 Adquisición de datos por RMN

El primer paso para la adquisición de los espectros de RMN es la calibración de los parámetros del equipo, así como la adecuación de las secuencias de pulsos a las características particulares de las muestras del estudio. De esta manera, es posible mantener constantes las condiciones de adquisición de los espectros, garantizando la reproducibilidad de los análisis realizados dentro de un mismo estudio.

Debido a las características que presentan las muestras de metabolómica y los compuestos que se pretenden medir, se han desarrollado distintos experimentos de

RMN con algunas modificaciones en sus secuencias de pulsos que permiten optimizar la obtención de espectros de calidad en estas condiciones. Por un lado, el agua presente en los biofluidos interfiere en el análisis del resto de señales del espectro por lo que se hace necesario optimizar su eliminación. Para minimizar este efecto de la señal del disolvente (agua) se irradia de forma continua la señal del agua para que sus ^{1}H no contribuyan a la señal observada. Por otro lado, las muestras utilizadas en metabolómica presentan una mezcla diversa de compuestos, dando lugar a muchas señales de baja intensidad en el espectro. Estas señales, además, pueden solaparse dificultando el análisis de las señales que pueden ser más interesantes para el estudio.

El principal experimento de RMN que se emplea en metabolómica es el experimento monodimensional (1D) del hidrógeno. Las secuencias más utilizadas son:

- **1D NOESY**. Esta secuencia contiene un pequeño período de mezcla que mejora la supresión de la señal del agua, la fase y la línea base. Este experimento proporciona un método simple, altamente reproducible y robusto para la adquisición de espectros de RMN de alta calidad en soluciones acuosas. Esta técnica es, sin duda, la más empleada en lo que se refiere a la supresión de agua (Hoult, 1976) y se aplica en una secuencia de pulsos de RMN conocida como *noesy-presat* (Kumar *et al*, 1980), donde la señal de agua es saturada a través de una irradiación selectiva.

- **CPMG** (*Carr Purcell Meiboom and Gill*). Esta secuencia están basadas en las propiedades de relajación de los núcleos y son muy útiles en metabolómica (Meiboom & Gill, 1958). Es la más utilizada para el análisis de muestras de plasma y suero, ya que contienen un alto nivel de proteínas. Las lipoproteínas y proteínas son moléculas de alto peso molecular que experimentan una relajación transversal muy rápida, dando lugar a señales anchas en el espectro. Esta secuencia permite la eliminación de estas señales, mejorando así la resolución de señales de bajo peso molecular.

Adicionalmente, se utilizan experimentos bidimensionales (2D) que complementan la información obtenida en los 1D y permiten obtener información más detallada. Los experimentos 2D se realizan sobre muestras representativas de cada grupo de estudio y permiten solucionar las dificultades asociadas al solapamiento de señales en los 1D, ya que los desplazamientos químicos se resuelven en dos dimensiones. Los 2D informan acerca de la relación existente entre las distintas señales del espectro, lo cual es muy útil si se quiere asignar a qué molécula corresponde cada señal presente en el espectro. Dentro de los estudios 2D destacan la secuencia **COSY** (COrrelations SpectroscopY), **TOCSY** (TOtal Correlation SpectroscopY) y **NOESY** (Nuclear

Overhauser Effect SpectroscopY), que son experimentos de gran utilidad para el análisis de las relaciones homonucleares, a través de enlace químico y del espacio (Beckonert *et al*, 2007). Existen otros experimentos que aportan información complementaria sobre relaciones heteronucleares, como es el **HSQC** (Heteronuclear Single Quantum Correlation).

2.5 Análisis de los datos

Una vez adquiridos los espectros de RMN, el paso siguiente es el análisis de la información obtenida. Este proceso se subdivide en:

2.5.1 Pre-procesado

Las variaciones en las características físicas y composición química de las muestras, debidas a diferencias de pH, temperatura, contenido iónico y concentración de metabolitos, hacen necesaria la aplicación de diferentes pasos previos al procesado de los datos. En este sentido, las técnicas de pre-procesado pueden transformar los datos y obtener resultados más robustos.

Entre las técnicas de pre-procesado de datos básicas destacan la transformada de Fourier (TF), la corrección de la fase y la línea base del espectro. La TF transforma la señal FID en un espectro de señales distribuidas en función de su frecuencia. La corrección de la línea base permite eliminar las posibles distorsiones que puedan afectar a las señales del espectro. Una inadecuada corrección puede hacer difícil el proceso de identificación de señales, así como introducir errores significativos en la cuantificación de los metabolitos. El ajuste de la fase es importante para conseguir que todas las señales tengan el mismo signo, su ajuste está condicionado por la concentración de la muestra.

Otro paso importante dentro del pre-procesado es el conocido como *binning o bucketing*. El *binning* consiste en la subdivisión e integración del espectro en pequeñas regiones que se denominan *bins* o *buckets*, permitiendo una reducción significativa de la complejidad de los datos. En la mayoría de los casos se emplea un tamaño de *bucket* constante, de 0.04 ppm, aunque dependiendo del tipo de muestra el tamaño de *bucket* puede variar. Los espectros de orina presentan muchas señales estrechas y muy próximas entre ellas, por lo que es frecuente usar tamaños de *bucket* más reducidos (0.01-0.001 ppm), que permitan evaluar diferencias en regiones muy próximas en el espectro.

Estos pasos previos al análisis estadístico se realizan de manera semi-automatica, utilizando los paquetes informáticos que se emplean para la adquisición de los espectros. Existen también diferentes plataformas y/o programas que aportan

herramientas complementarias y específicas para el pre-procesado de los datos (Puchades-Carrasco *et al*, 2015).

2.5.2 Procesado

Una vez realizado este proceso se aplican distintos procedimientos que permiten obtener información comparable entre los distintos espectros.

-Regiones de exclusión: es necesario excluir del análisis las señales espectrales no informativas o con información no reproducible. En este sentido, se incluye en el análisis únicamente la región espectral que contiene las señales de interés: generalmente entre 0.2 y 10 ppm. Una región presente en todos los espectros de metabolómica, y que es necesario excluir, es la señal residual del agua (4.6-5.4 ppm) que es altamente variable. Dependiendo del biofluido con el que se esté trabajando pueden existir otras señales que necesiten ser excluidas, como por ejemplo la señal de la urea (5.4-6.0 ppm). La urea es un metabolito que no puede ser cuantificarse por RMN con exactitud al encontrarse sus protones en continuo intercambio con el agua. También es necesario excluir otras señales que puedan interferir en el análisis, como las debidas a aditivos, fármacos, contaminantes, etc. Antes de excluir estas señales hay que realizar un estudio exploratorio para valorar su exclusión, ya que pueden ser señales que estén presentes en todos los espectros o sólo en algunos y pueden o no tener interés clínico.

-Alineamiento de señales: consiste en la corrección de pequeñas variaciones en el desplazamiento químico de las señales que puedan existir entre los espectros que forman parte de un mismo estudio. El alineamiento incorrecto de picos supone una dificultad a la hora de analizar los espectros de RMN. Este fenómeno deriva de la gran heterogeneidad en la composición de los biofluidos. Existe una gran cantidad de variables que juegan un papel importante en estos cambios, como son diferencias en el pH, la fuerza iónica, la temperatura, etc. Esta situación hace que las mismas señales aparezcan a desplazamientos químicos distintos en los espectros, lo que complica su comparación.

Para corregir este fenómeno se realiza una calibración de los valores de desplazamiento químico de todas las señales del espectro respecto a una señal de referencia (p.ej., 3-trimetilsilil propionato, TSP) o en relación con un metabolito interno (p.ej., glucosa, alanina) cuyo desplazamiento químico es insensible a estos cambios. La alanina se utiliza cuando se trabaja con muestras de plasma o de suero ya que es una señal presente en todas las muestras, aparece bien definida y su desplazamiento es insensible a las características específicas de la muestra. Adicionalmente, puede

ocurrir que no todas las señales del espectro se vean afectadas por estas variaciones de la misma forma, ni incluso las que pertenecen a un mismo metabolito. La orina es uno de los biofluidos en los que este fenómeno tiene mayor relevancia. Por este motivo, cuando se trabaja con espectros de orina, es necesario aplicar algoritmos específicos que faciliten el alineamiento de las señales en estas condiciones, como MetaboLab, Automics, Icoshift o "speaq" (Vu *et al*, 2011). Algunos de estos algoritmos realizan un alineamiento que afecta a todas las señales del espectro, para ello, seleccionan un espectro de referencia a partir del cual se alinean el resto de los espectros (Vu & Laukens, 2013). Otros, permiten realizar el alineamiento de regiones aisladas del espectro tomando como referencia un espectro representativo. El alineamiento de las regiones seleccionadas se lleva a cabo sin modificar la posición del resto de las señales. En función de las variaciones encontradas en las muestras del estudio, será necesario valorar el método de corrección más adecuado.

-Normalización: se realiza para minimizar el efecto de las variaciones relacionadas con la distinta concentración de las muestras o el peso de los metabolitos presentes dentro de una misma muestra. La normalización puede ser necesaria también para corregir diferencias técnicas entre los distintos espectros incluidos en un análisis (Torgrip, 2008). Si los espectros se adquieren empleando un número diferente de scans o empleando diferentes equipos, los valores absolutos de los espectros pueden ser diferentes, de manera que se hace necesario un proceso de normalización antes de poder analizarlos de manera conjunta.

Existen diferentes tipos de normalización, uno de los más utilizados es la normalización respecto al área total del espectro, en la cual la intensidad de cada *bucket* es dividida por la integral del espectro completo (**Figura 5**). Sin embargo, dependiendo del tipo de biofluido puede ser conveniente aplicar métodos más específicos, como es la normalización por cociente probabilístico (PQN) (Dieterle *et al*, 2006). Este método se emplea en biofluidos como la orina, donde la intensidad de las señales de los metabolitos analizados dentro de una misma muestra es muy distinta. Se trata de un método más exacto y robusto que la normalización por área total. Considera que los cambios en las concentraciones individuales de un metabolito concreto sólo influyen en parte del espectro, mientras que las variaciones en la concentración global de las muestras afectan al espectro completo. Esta normalización emplea un factor de dilución específico para cada espectro que se calcula dividiendo el valor de cada una de las regiones del espectro por la intensidad de esa misma región en el espectro de referencia (Dieterle *et al*, 2006).

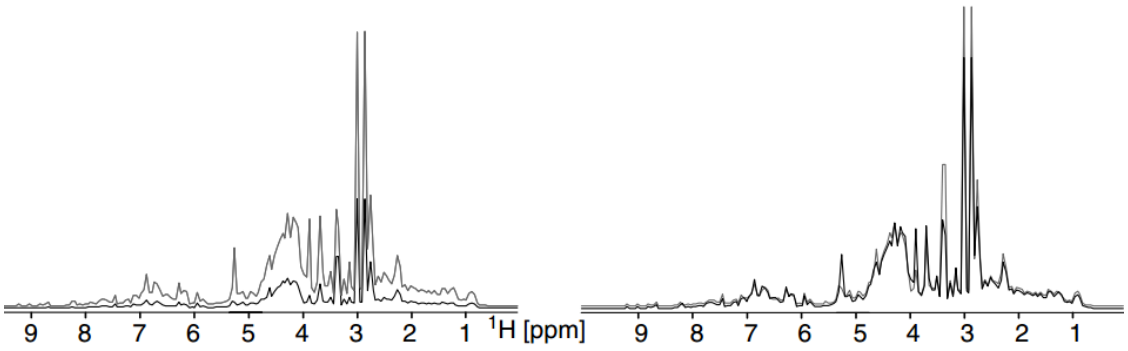


Figura 5: Efecto de la normalización a área total en el procesado de espectros de muestras con diferencias en su concentración: superposición de los espectros antes (izquierda) y después de la normalización (derecha).

En otros casos, también se utiliza, aunque en menor medida, la normalización en función de la concentración de creatinina. Este método se utiliza tradicionalmente en la analítica clínica, cuando se trabaja con muestras de orina. Se asume que la excreción de creatinina es constante, sirviendo como indicador de la concentración de la muestra. La creatinina se puede cuantificar mediante RMN integrando en el espectro las señales de 3.05 y 4.05 ppm. En la práctica, la aplicación de este método puede estar contraindicada por problemas técnicos. Por un lado, su desplazamiento químico pude variar al estar influenciado por variaciones del pH entre muestras y, por otro lado, debido al solapamiento de picos, bastante frecuente en los espectros de orina.

Los métodos de procesado anteriormente citados, van dirigidos a corregir la variabilidad existente entre las distintas muestras del estudio que podría interferir en el análisis posterior. Otro aspecto a tener en cuenta es que grandes diferencias en la magnitud de las variaciones de las distintas señales del espectro en la comparación entre los grupos de estudio pueden dificultar de forma considerable el análisis (Craig *et al*, 2006). Este efecto se corrige con un procedimiento conocido como **escalado.** El escalado se realiza sobre los datos una vez normalizados. Los tres métodos de escalado más utilizados en los estudios de metabolómica son:

-Centrado: Resta el valor medio para todas las muestras a los datos originales de ese *bucket* en cada muestra. Este método permite reducir el ruido de fondo, centrando el análisis únicamente en la parte fluctuante de los datos (Craig *et al*, 2006).

-Varianza Unitaria (UV): Divide los datos originales por la desviación estándar de cada variable. De este modo, todas las variables tienen el mismo peso en el análisis, independientemente de cual sea la magnitud de su diferencia entre los grupos de estudio. El inconveniente de este método es que puede aumentar el peso de señales debidas al ruido de fondo o la variabilidad derivada de errores en la instrumentación,

dificultando así el análisis de las variables de interés en el estudio (Eriksson *et al*, 2013).

-Pareto: Divide cada variable por la raíz cuadrada de la desviación estándar de esa variable entre las muestras del estudio. Este método es ampliamente utilizado par ael tratamiento de datos espectroscópicos, ya que continúa manteniendo una distribución en el peso de las variables similar a la de los datos originales (Eriksson *et al*, 2013).

2.5.3 Análisis estadístico

Tras el procesado de los datos se obtiene una matriz de datos metabolómicos que contiene toda la información bioquímica de las muestras. En los estudios de metabolómica, el número de variables generalmente es muy superior al número de observaciones (muestras). Las técnicas estadísticas utilizadas en este tipo de estudios permiten reducir la dimensionalidad de los datos y extraer la información más relevante. Los métodos de análisis estadístico multivariante se utilizan para extraer la información sobre qué variables son relevantes en el estudio. Posteriormente, la magnitud de las diferencias observadas para esas variables, así como su relevancia estadística se evalúa de manera univariante.

a) Análisis estadístico multivariante

Los métodos de análisis multivariante estudian el comportamiento de tres o más variables al mismo tiempo. Su objetivo es simplificar el modelo estadístico, donde el número de variables puede ser un problema y, así, comprender mejor la relación entre varios grupos de variables. Dentro del análisis multivariante destacan dos métodos. El primero se basa en la aplicación de métodos no supervisados, donde no se tiene en cuenta la información acerca de la estructura de los datos. El segundo es el método supervisado, donde la clasificación de las muestras se fundamenta en el conocimiento previo del sistema.

El método no supervisado se utiliza para realizar una primera exploración de los datos. Uno de los modelos más utilizados es el análisis de componentes principales (PCA). El PCA tiene el objetivo de reducir las variables perdiendo la mínima cantidad de información posible. De esta manera, las múltiples variaciones quedan agrupadas de acuerdo a su importancia en la distribución de las muestras observadas en lo que se conoce como componentes principales (PC). Los PC que se obtienen son una combinación lineal de las variables originales, y son independientes entre sí. Geométricamente, el PCA es una representación de las muestras en un nuevo sistema de coordenadas construido con un número de variables inferior al utilizado inicialmente.

El primer componente principal es una combinación lineal de las variables (*k)* que explican la máxima variabilidad de las muestras (*n*). El segundo componente se escoge de forma que sea ortogonal al primero y que explique la máxima variabilidad de las muestras una vez restada la explicada por el primer PC y así, sucesivamente. Las nuevas variables son combinaciones lineales de las anteriores y se van construyendo según el orden de importancia en cuanto a la variabilidad total que recogen de las muestras. Esta información se representa en forma de matriz. Cada elemento de ésta representa las correlaciones entre las variables y los PCs. El PCA resulta útil para identificar patrones o tendencias dentro de los grupos de muestras analizados.

Dos representaciones gráficas de gran utilidad en el análisis de un modelo PCA son el ***score plot*** *y* el ***loading plot.*** El *score plot* es un gráfico de puntuaciones de las muestras en función de los componentes principales. Este gráfico representa la localización de las muestras (*n*) y la relación entre ellas, revelando sus agrupaciones y tendencias. El *loading plot* es un gráfico que complementa la información de los *score plot*. Son gráficos de puntuaciones de las variables en función de los PCs, y definen la relación entre las variables (*k)* que integran la matriz de datos original (X) con las mismas direcciones que las del *score plot*. Este gráfico informa de la existencia de correlaciones positivas o negativas entre las variables del modelo y, además, muestra cuáles son las variables que más influyen en los valores de cada componente principal (**Figura 6**).

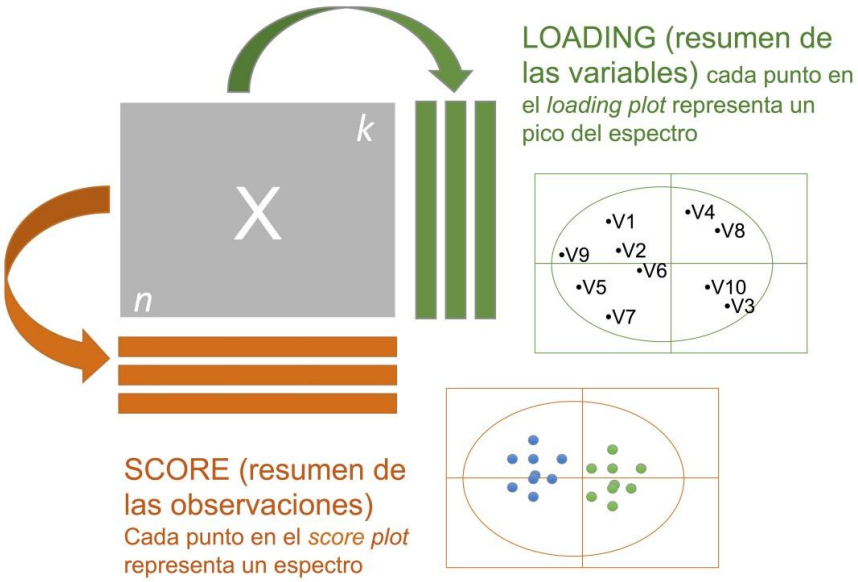


Figura 6: Esquema de los gráficos *score plot* (observaciones) y *loading plot* (variables) obtenidos a partir de la matriz de datos.

Otro gráfico que aporta información para identificar posibles *outliers* es el gráfico ***T^2 Hotelling***. Un *outlier* es aquella muestra que presenta un comportamiento diferente dentro de su grupo. El *T^2 Hotelling* mide la variación de cada muestra dentro del modelo PCA permitiendo detectar si la variación incluida en los componentes principales es más grande que la que le correspondería si sólo influyeran variaciones aleatorias. Los gráficos *T^2 Hotelling* permiten identificar todas las muestras que se encuentran fuera del intervalo de confianza del 95% en el modelo no supervisado.

En conjunto, todos estos gráficos ayudan a estudiar la homogeneidad de los datos, evaluar la calidad de los mismos, caracterizar la presencia de *outliers* o identificar la existencia de sesgos. La presencia de *outliers* en el estudio puede interferir negativamente en los resultados por lo que identificar estas muestras y razonar su exclusión o inclusión del estudio es fundamental.

El otro método de análisis utilizado para diseñar modelos de discriminación entre grupos de estudio es el análisis supervisado. En este tipo de análisis se incluye información sobre la clase a la que pertenece cada muestra. Los métodos más utilizados en metabolómica para el análisis supervisado son el análisis discriminante por mínimos cuadrados (PLS-DA) y su modificación con corrección ortogonal (OPLS-DA) (Gu *et al*, 2011; Pan *et al*, 2007).

El método PLS-DA es un método de regresión lineal supervisada basado en la combinación de una matriz de observaciones (datos espectrales) y una matriz de valores cualitativos (variables informativas). El método OPLS-DA representa una modificación del método PLS basado en la separación sistemática de los datos metabolómicos estructurada en dos niveles: la variación que está correlacionada con la pertenencia a una u otra clase en el modelo, y aquella que no está relacionada (ortogonal) (Trygg & Wold, 2002). Los modelos derivados de la aplicación de este método se pueden representar con gráficos (**Figura 7**) de manera similar al PCA.

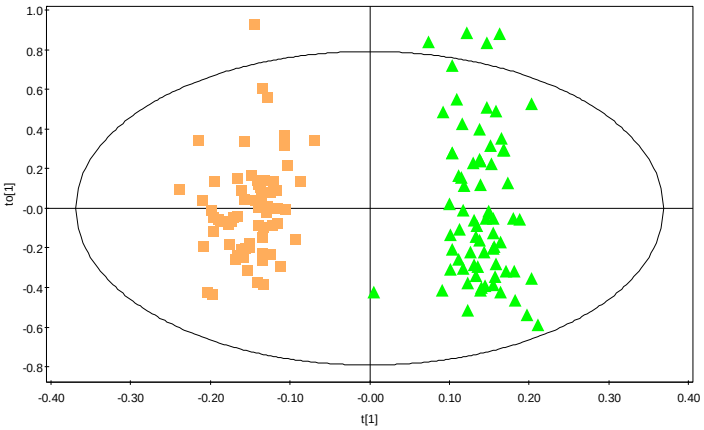
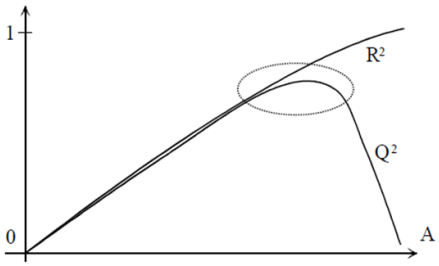


Figura 7: Representación de un modelo OPLS-DA. El primer componente (eje abscisas) contiene la información predictiva o directamente relacionada con la variable respuesta (p.ej., caso-control) y la información no relacionada con la respuesta aparece en el componente ortogonal (eje ordenadas).

Los parámetros que se emplean para evaluar estos modelos son la bondad de ajuste (R^2) y la capacidad predictiva (Q^2). Un valor elevado de Q^2 muestra una buena capacidad predictiva, es decir, que el modelo es capaz de predecir muestras no incluidas en él. El ajuste y la capacidad predictiva del modelo aumentan a medida que se añaden componentes, hasta un punto en el cual la capacidad predictiva empieza a disminuir, al incluir en los nuevos componentes información no relevante para el modelo. El número de componentes del modelo se elige teniendo en cuenta estos valores y es necesario que exista un compromiso entre R^2 y Q^2. Los valores de estos parámetros oscilan entre 0 y 1. En los estudios de metabolómica se considera que un modelo discriminante es de buena calidad cuando $Q^2 > 0.5$ y la diferencia entre ambos (R^2, Q^2) parámetros no es superior a 0.2-0.3 (Eriksson *et al*, 2013).

Lo ideal es que el número de componentes del modelo sea tal que los valores de R^2 y Q^2 sean lo más altos posibles (Eriksson *et al*, 2013) (**Figura 8**). Cuando la relación entre R^2 y Q^2 no se mantiene el modelo está sobreajustado. Los modelos con sobreajuste se caracterizan por una baja capacidad de predicción cuando se aplican a la clasificación de un conjunto de muestras externo, es decir, el modelo se ajusta muy bien a los datos existentes pero tiene un pobre rendimiento para predecir nuevos resultados.



R2 estima la bondad de ajuste
Q2 estima capacidad predictiva

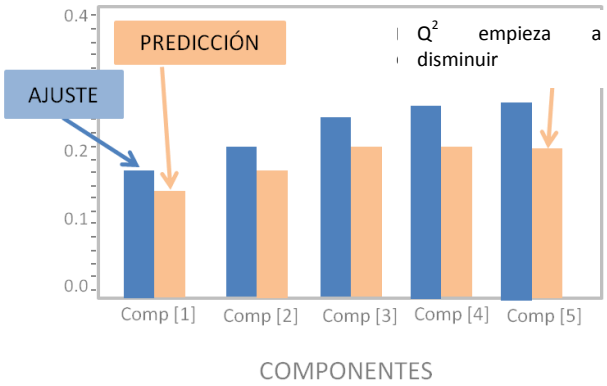


Figura 8: Relación entre los valores de R^2 y Q^2 en función de los componentes utilizados para crear un modelo PLS-DA. En ambos gráficos el eje de abscisas corresponde al número de componentes que forman el modelo y el eje de ordenadas al valor de R^2 ó Q^2.

La selección del número de componentes se realizará siempre buscando obtener la máxima capacidad de predicción (Q^2). Una de las opciones para mejorar el valor de Q^2 en los modelos discriminantes es aplicar métodos de selección de variables. La selección de variables permite extraer sólo la información más relevante y que contribuye de forma significativa a la discriminación entre los grupos, excluyendo las variables redundantes o no informativas antes de realizar el análisis multivariante (Andersen & Bro, 2010).

La selección de variables, además de eliminar la información redundante, facilita la visualización y la comprensión de los datos, reduce la proporción entre variables y muestras, incrementa la capacidad de predicción de los modelos y mejora su interpretación (Di Anibal *et al*, 2011). Las estrategias para realizar la selección de variables son diversas. Algunos métodos se basan en la eliminación de las variables de baja intensidad o el ruido de fondo con el objetivo de reducir el número de datos. Otros métodos aplican procedimientos supervisados para identificar las variables que más predominan en la discriminación para lograr la mejor separación cuando se trabaja con diferentes grupos de muestras.

Entre los métodos más utilizados destaca el método de los coeficientes de regresión (CoeffCS) y sus correspondientes errores (CoeffCScvSE). Los coeficientes se utilizan para interpretar la influencia de la variable X sobre la Y, y su error o desviación interesa que sea lo más pequeña posible. El coeficiente de regresión mide la intensidad de la relación lineal entre dos variables, de tal manera que cuánto más cercano a uno sea el valor del coeficiente, más robusto será la asociación lineal entre las dos variables.

Otro método para la selección de variables es el conocido del inglés como *interval Partial Least Squares method* (iPLS) (Nørgaard et al, 2000), que consiste en estudiar

las regiones del espectro con mayor influencia en la discriminación de los grupos, y calcular nuevos modelos PLS-DA. También se emplean con frecuencia los algoritmos genéticos (Hibbert, 1993; Leardi, 2001), que permiten encontrar las variables óptimas a partir de un subconjunto inicial aleatorio de variables mediante un proceso iterativo.

El análisis de los modelos discriminantes genera una gran cantidad de información que puede ser analizada a través de gráficos que facilitan el análisis e interpretación de los datos. Estas herramientas nos permiten extraer la información sobre cuáles son las variables más relevantes en la discriminación de los grupos. Entre ellas destacan:

-VIP list (*Variable Importance in the Projection*): proporciona información sobre las variables (*buckets*) que más importancia tienen en el modelo, asignando a cada *bucket* un valor de VIP. Los valores VIP reflejan la correlación entre las condiciones de estudio y las variables implicadas.

-**Contribution plot**: representa mediante un diagrama de barras la relevancia de cada variable incluida en el modelo según su contribución a la discriminación entre las clases analizadas en el modelo.

-SUS-plot (*Shared and Unique Structures*): permite el análisis de las diferencias entre dos modelos supervisados. Este gráfico ayuda a la identificación de biomarcadores. Detecta cuáles son las diferencias entre las distintas clases de cada modelo que son únicas y cuáles están presentes en ambos modelos. Representa la covarianza y la correlación del modelo en un diagrama de dispersión.

Una aproximación complementaria y de gran utilidad para el análisis de los resultados obtenidos de los modelos de discriminación son los modelos estadísticos basados en la regresión logística. El objetivo de estos modelos es conocer la relación entre una variable dependiente, cualitativa o dicotómica, con una o más variables explicativas independientes o covariables. Una de sus aplicaciones es el cálculo de ecuaciones de probabilidad. Para ello, la regresión logística utiliza la siguiente fórmula:

$$\text{Log}\left(\frac{p}{1-p}\right) = b_0 + b_1x_1 + \ldots + b_nx_n$$

Donde "*p"* es la probabilidad de que ocurra el evento de interés, "*x"* representa las variables independientes y "*b"* los coeficientes de regresión asociados a cada variable.

Se emplean con frecuencia en el campo de la Biomedicina para la identificación de resultados en base a lo que se conoce como OR (*odds ratio*). A partir de los coeficientes de regresión (*b*) de las variables independientes implicadas en el modelo se puede obtener directamente la de OR cada una de ellas. Así, la OR se utiliza como medida de la relación entre la variable de estudio y la intensidad de la variable explicativa.

b) Validación interna de modelos

En los modelos supervisados (PLS, OPLS), previo a su análisis, es importante evaluar su calidad y su fiabilidad. Con ese fin, es necesario realizar pruebas de validación interna (Westerhuis *et al*, 2008):

- **Validación cruzada (*cross-validation*):** consiste en dividir aleatoriamente el conjunto de muestras en un grupo de validación y un grupo de entrenamiento. Se construye un modelo con las muestras del grupo de entrenamiento y se utiliza ese modelo para predecir la clasificación de las muestras del grupo de validación, permitiendo así calcular el error de predicción del modelo. Durante la validación cruzada, se construye el modelo excluyendo en cada iteración un grupo de muestras distinto, obteniendo como resultado final un valor de Q^2 que refleja la capacidad predictiva global del modelo.
- **Prueba de permutación:** se utiliza para evaluar la fiabilidad y solidez de los valores de R^2 y Q^2 de un modelo determinado. El test de permutación consiste en construir un número determinado de modelos (n permutaciones), en los que el sistema asigna la clasificación de las muestras de manera aleatoria. La comparación de los valores de R^2 y Q^2 del modelo original, frente a los valores obtenidos en las distintas permutaciones, proporciona información sobre la validez del modelo. El modelo se considera correctamente validado cuando los valores obtenidos en las permutaciones son muy inferiores a los del modelo real. La valoración de los resultados del test de permutación se realiza calculando las rectas de regresión para los valores de R^2 y Q^2, respectivamente, de todas las permutaciones y el modelo original (**Figura 9**). La ordenada en el origen para la recta de regresión entre los distintos valores de R^2 debe situarse entre 0.3-0.4, y el de la recta de regresión de Q^2 debe ser inferior a 0.05

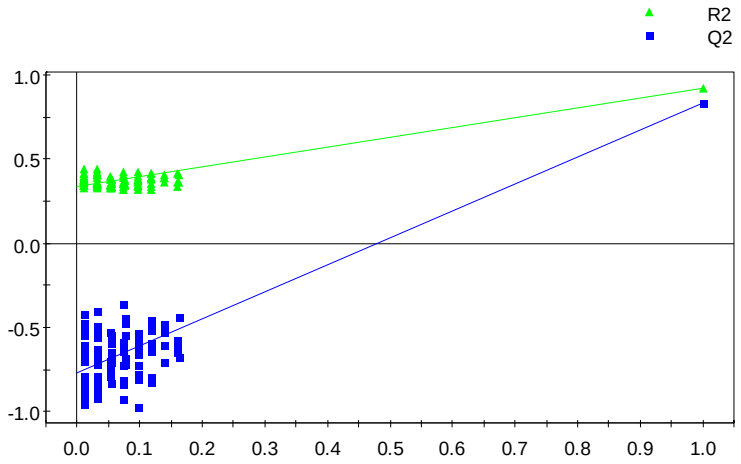


Figura 9: Resultados del test de permutación de un modelo PLS-DA: el eje de ordenadas representa los valores obtenidos para R^2 (Y) (verde) y Q^2 (Y) (azul), el eje de abscisas representa el valor del coeficiente de correlación entre el valor del modelo original y los valores de cada permutación.

c) Análisis univariante

Como resultado del análisis multivariante se obtiene un conjunto de variables significativas o relevantes para la discriminación entre clases. Estas señales son integradas y cuantificadas en el espectro de RMN. El análisis cuantitativo tiene como objetivo estudiar la relevancia estadística de las diferencias en los niveles de concentración de los metabolitos entre los diferentes grupos de estudio.

A partir de las intensidades obtenidas para cada metabolito de interés, se pueden aplicar diferentes test estadísticos univariantes, en función de los grupos que se comparan. Entre ellos, los test estadísticos como el test de Mann-Whitney, t-student, ANOVA, etc. Las características de los datos y el objetivo del análisis determinarán la elección de un test estadístico distinto.

2.6 Identificación de metabolitos

Una de las fases más importantes de los estudios de metabolómica es la asignación e identificación de los compuestos presentes en las diferentes regiones del espectro. La mayoría de las señales son identificadas en los espectros 1D y son asignadas en función de su desplazamiento químico. En este tipo de espectro, la presencia de múltiples especies moleculares en la misma muestra provoca solapamiento de señales, dificultando su identificación. En estos casos, los espectros 2D son de gran utilidad para identificar algunas de las señales del espectro.

La identificación y asignación de los metabolitos se realiza comparando los desplazamientos químicos obtenidos para cada señal con los desplazamientos químicos descritos en la literatura (Bouatra *et al*, 2013; Nicholson *et al*, 1995;

Psychogios *et al*, 2011; Salek *et al*, 2007) y en las bases de datos disponibles como HMDB: *Human Metabolome Data Base* (Wishart *et al*, 2009), BMRB: *Biological Magnetic Resonance Data Bank* ("BMRB - Biological Magnetic Resonance Bank") *, etc.*

El uso de algoritmos específicos facilita la identificación automática de metabolitos. Existen diferentes programas como Chenomix (Weljie *et al*, 2006), o paquetes informáticos como FOCUS (Alonso *et al*, 2014), metaboMiner (Xia *et al*, 2008), rNMR (Lewis *et al*, 2009), MetaboHunter (Tulpan *et al*, 2011), SpinAssign (Chikayama *et al*, 2010), MetaboID (MacKinnon *et al*, 2013) (**Tabla1**) diseñados para este fin.

Tabla 1: Resumen de algunas bases de datos utilizadas en los estudios de metabolómica (Puchades-Carrasco *et al*, 2015).

Nombre	Disponibilidad	Observaciones	Acceso
Human Metabolome Database	libre	3.000 metabolitos, información biológica de 40,000 metabolitos	http://www.hmdb.ca/
Biological Magnetic Resonance Data Bank	libre	1.000 compuestos, no aporta información biológica	http://www.bmrb.wisc.edu/metabolomics/
Spectral Database for Organic Compounds	libre	15.000 compuestos	http://sdbs.db.aist.go.jp/
Madison-Qingdao Metabolomics Consortium Database	libre	794 compuestos	http://www.mmcd.nmrfam.wisc.edu/
Platform for RIKEN Metabolomics	libre	80 compuestos	http://prime.psc.riken.jp/
Birminghan Metabolite Library	libre	208 compuestos a pHs diferentes	http://www.bml-nmr.org/
BBIOREFCODE	Requiere licencia	600 metabolitos	https://www.bruker.com/
Chenomx	Requiere licencia	500 metabolitos	http://www.chenomx.com/

Cuando la asignación de una señal es incierta porque su desplazamiento químico varía o está próximo a la señal de otro metabolito, es útil aplicar la técnica de *spiking* (**Figura 10**). Los experimentos de *spiking* se realizan añadiendo a la muestra de estudio una determinada concentración del metabolito puro a identificar. El nuevo espectro se compara con el espectro inicial con el objetivo de observar qué señal aumenta debido a la adición de ese metabolito, confirmando la asignación del mismo. (Psychogios *et al*, 2011). En la **Figura** 1**0** se comparan dos espectros de RMN obtenidos siguiendo esta aproximación experimental. El espectro inferior presenta una

señal poco definida, cercana a 3.35 ppm, de difícil asignación. El espectro superior, corresponde al análisis de esa misma muestra después de haber añadido una cantidad definida del metabolito X, confirmando así que el metabolito X presenta una señal de RMN en esa región del espectro.

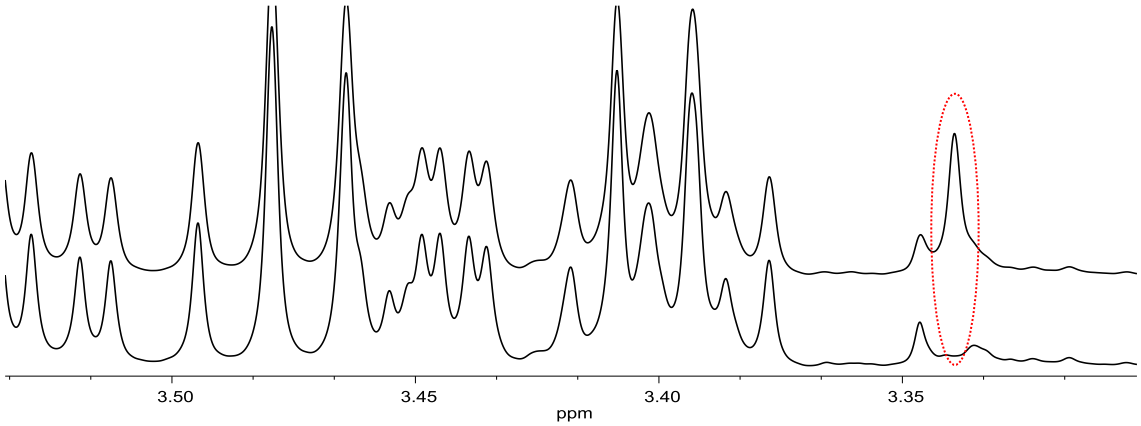


Figura 10: Experimento de *spiking* utilizado para la confirmación de la asignación de un metabolito.

2.7 Interpretación biológica

El objetivo de esta etapa del análisis es relacionar los resultados encontrados con la fisiopatología de la enfermedad o del problema de estudio. La naturaleza del metabolismo celular, en la que un mismo metabolito está implicado en varias rutas metabólicas, hace que la interpretación sea una tarea complicada (Aggio *et al*, 2010).

Para la interpretación biológica de los cambios metabólicos observados es frecuente recurrir al uso de bases de datos. La HMDB es una base de datos que proporciona información química y biológica sobre los compuestos. Hay otras bases de datos más específicas sobre la implicación de metabolitos en rutas metabólicas como KEGG *pathway analysis database* (Kanehisa, 2002) y *ConsensusPathDB-human* (Kamburov *et al*, 2012). Las bases de datos y los estudios previos publicados en relación al estudio ayudan a relacionar las rutas metabólicas implicadas con la patología de estudio y a entender el papel que las variaciones encontradas en los niveles de los metabolitos juegan en relación el problema de estudio.

La dificultad en esta etapa consiste en interpretar los resultados como un todo y no como una suma de resultados individuales. En este punto, la enfermedad o patología de estudio debe describirse como la acción correlativa de muchos factores moleculares, un fenómeno multifactorial. El análisis de las rutas metabólicas trata de establecer las conexiones que hay entre los metabolitos detectados para poder formar un mapa completo de las distintas rutas biológicas alteradas (Weckwerth & Morgenthal, 2005).

3. BÚSQUEDA DE BIOMARCADORES DE INTERÉS CLÍNICO

Los metabolitos son los productos finales de todos los procesos que se producen en las células. Las fluctuaciones en las concentraciones de los metabolitos reflejan los cambios bioquímicos que se producen en los distintos estados fisiológicos, incluidos los patológicos. Actualmente, existen identificados más de 3.000 metabolitos diferentes, endógenos y exógenos, resultantes de la actividad enzimática celular y la respuesta a estímulos ambientales, como la dieta (Wishart *et al*, 2007).

En el año 2001, el *National Institutes of Health* estadounidense estandarizó la definición de biomarcador como "una característica que se puede medir y evaluar objetivamente como indicador de procesos biológicos normales, procesos patogénicos o respuestas farmacológicas a una intervención terapéutica".

Los biomarcadores se han consolidado como recursos esenciales para el diagnóstico de enfermedades en estadios tempranos (Manna *et al*, 2011), para la identificación de nuevas dianas terapéuticas (Nicholson *et al*, 2011), o para el pronóstico de enfermedades de forma más precisa (Puchades-Carrasco *et al*, 2013).

En los últimos años, el desarrollo y búsqueda de nuevos biomarcadores ha despertado gran interés, principalmente en el área de la oncología debido a la naturaleza heterogénea de la enfermedad. El desarrollo de las tecnologías ómicas (genómica, proteómica, transcriptómica, metabolómica) ha incrementado potencialmente la investigación en biomarcadores basados en ADN, ARN, proteínas o metabolitos. La idea del análisis de una única molécula como biomarcador está siendo reemplazada por el análisis multiparamétrico de genes, proteínas y metabolitos, centrado en la identificación de lo que se denomina como "firma o huella" del tumor. En los últimos años se están desarrollando distintas estrategias para el descubrimiento de biomarcadores en cáncer utilizando estas aproximaciones (**Figura 11**).

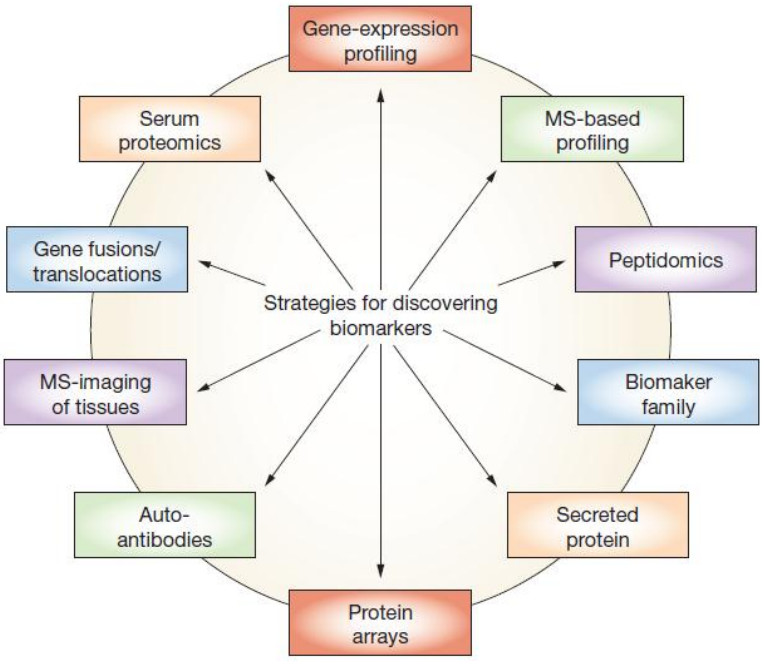


Figura 11: Esquema de las estrategias para el descubrimiento de biomarcadores a través de aproximaciones ómicas (Kulasingam & Diamandis, 2008).

Un ejemplo de estos desarrollos, basado en perfiles de expresión génica es el caso del comercializado panel de *MammaPrint*® (Van De Vijver *et al*, 2002). En este panel se analizan 70 genes de una muestra tumoral y se evalua la información para determinar si el cáncer presenta un bajo o alto riesgo de recurrencia en el transcurso de los diez años posteriores al diagnóstico. Este test fue aprobado en Febrero del 2007, convirtiéndose en el primer panel multigen aprobado por la FDA (*Food and drug administration)* con valor pronóstico en cáncer de mama (Kulasingam & Diamandis, 2008).

Otro ejemplo del potencial que presenta en este tipo de análisis el perfil proteómico es su aplicación para el diagnóstico de cáncer de ovario. Petricoin y colaboradores (Petricoin *et al*, 2002) desarrollaron una herramienta bioinformática que permitía identificar patrones proteómicos en suero, distinguiendo las enfermedades neoplásica de no neoplásica con una sensibilidad y especificidad del 100% y 95%, respectivamente.

En el área de la metabolómica, se han desarrollado estudios dirigidos a la identificación de biomarcadores no sólo asociados a la aparición de la enfermedad, sino también a la evaluación de la eficacia y seguridad de los tratamientos administrados o la monitorización de pacientes. En el estudio publicado por Puchades-Carrasco y colaboradores (Puchades-Carrasco *et al*, 2013) realizado en pacientes diagnosticados con mieloma múltiple, se identificaron cambios metabólicos específicos

asociados, a la progresión de la enfermedad y a la respuesta al tratamiento, respectivamente (Puchades-Carrasco *et al*, 2013).

3.1 Características de los biomarcadores

La naturaleza de los biomarcadores es muy variable ya que puede incluir proteínas, ácidos nucleicos, anticuerpos, metabolitos, genes, etc. El material biológico sobre el que se determina la presencia de biomarcadores también puede ser de naturaleza muy distinta (orina, plasma, suero, líquido cefalorraquídeo, tejido, saliva, etc).

Entre las características esenciales de un biomarcador, destacan la especificidad, la sensibilidad, la estabilidad, la accesibilidad y que sea fácilmente cuantificable (Lescuyer *et al*, 2007).

El biomarcador ideal debería reunir todas esas características y, además, demostrar tener utilidad clínica sobre otros marcadores que ya se encuentren implantados en la práctica clínica diaria y en la investigación. Durante las dos últimas décadas, poco más de 12 biomarcadores han sido aprobados por la FDA para la monitorización de respuesta, seguimiento o recurrencia del cáncer (Anderson & Anderson, 2002). La falta de sensibilidad y especificidad de los métodos de ensayo utilizados para su determinación hace que muchos biomarcadores estudiados en la investigación básica no lleguen a trasladarse a la práctica clínica (Srivastava *et al*, 2001).

La sensibilidad es la capacidad de clasificar correctamente a un individuo enfermo, es decir, la probabilidad de que para un sujeto enfermo la prueba logre un resultado positivo. La especificidad es la capacidad de clasificar correctamente a un individuo sano, es decir, la probabilidad de que para un sujeto sano la prueba logre un resultado negativo (Wagner *et al*, 2004). La sensibilidad y especificidad varían según el punto de corte que se utilice, ya que una depende de la otra. En función del punto de corte, la sensibilidad y especificidad variarán, aumentando o disminuyendo a su vez el número de falsos positivos y falso negativos. (**Figura 12**). El punto de corte que se define para la sensibilidad y especificidad varía en función de la utilidad final de la prueba (*screening*, pronóstico, etc). Cuando el objetivo es detectar la enfermedad en su estadio inicial, el test de detección de la enfermedad será más sensible a expensas de una menor especificidad. Así, todos los pacientes enfermos serán detectados, evitando su evolución a fases avanzadas de la enfermedad donde las posibilidades de cura son menores.

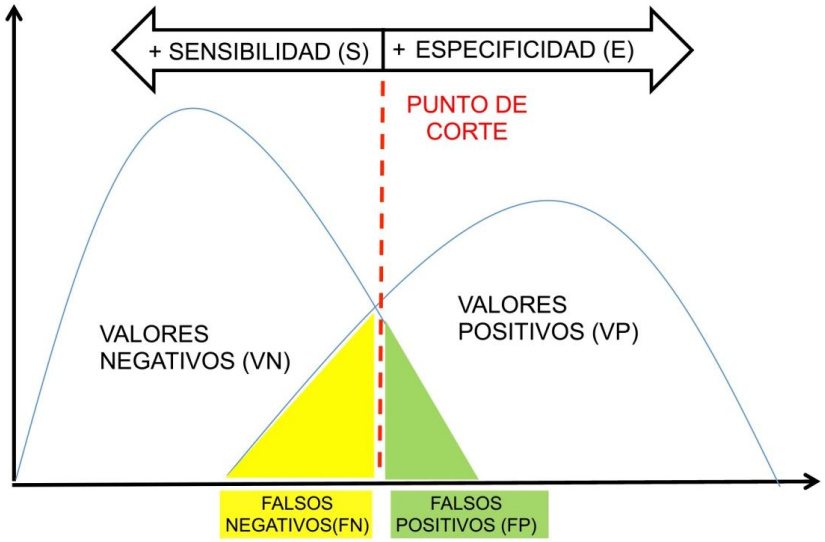


Figura 12: Representación gráfica de la relación entre la sensibilidad y especificidad.

Las curvas ROC son una representación gráfica de la relación existente entre sensibilidad y especificidad para cada punto de corte posible (**Figura 13**). El biomarcador ideal será aquel para el que el área bajo la curva en la representación de la curva ROC tenga un valor máximo, su curva en la gráfica coincidiría con los lados izquierdo y superior. Una prueba que fuera ineficaz su curva tendería a formar una línea recta entre las esquinas inferior-izquierda y superior-derecha del gráfico. En la práctica, las curvas se situarán en una posición intermedia entre esas dos opciones.

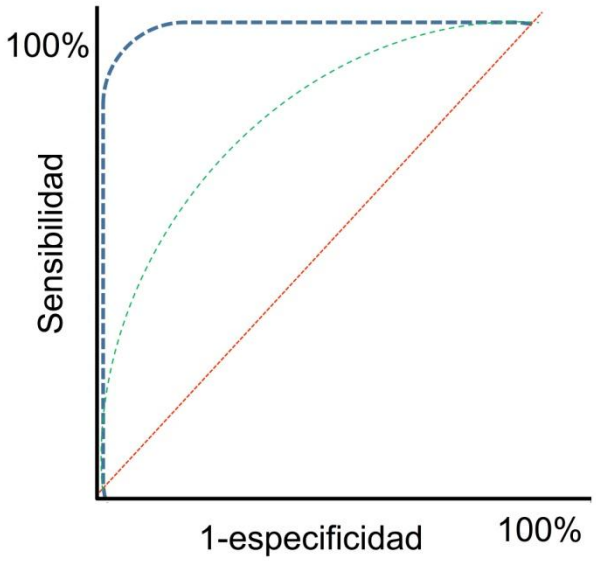


Figura 13: Curva ROC de distitntos biomarcadores. Se representan los valores de sensibilidad y especificidad para tres biomarcadores con distinta sensibilidad: alta (azul), intermedia (verde) y baja (rojo).

Ningún biomarcador clínico utilizado en la actualidad presenta una especificidad y sensibilidad del 100%. Por ejemplo, el antígeno prostático específico (PSA), presenta

una alta sensibilidad (más del 90%) pero una baja especificidad (aproximadamente del 25%), lo que se traduce en que a muchos hombres se les realiza una biopsia sin tener cáncer de próstata (Bangma *et al*, 1995; Gann *et al*, 1995). El marcador tumoral en suero para el cáncer de mama, CA 15.3, tiene solo un 23% de sensibilidad y un 60% de especificidad, por lo que sólo tiene utilidad para monitorizar la terapia o evaluar la recurrencia de la enfermedad. La alfa fetoproteína es un marcador de suero empleado en el diagnóstico de cáncer hepatocelular pero tiene una sensibilidad (39-64%) y especificidad (76-91%), muy bajas para usarlo como marcador de detección temprana (Marrero *et al*, 2003).

En la actualidad, con las nuevas tecnologías, se recurre al uso de paneles donde se combinan varios biomarcadores o al uso de tecnologías ómicas para obtener patrones que incluyan diferentes proteínas, genes o metabolitos, con el objetivo de aumentar la sensibilidad y especificidad.

3.2 Fases del desarrollo

En metabolómica, el proceso de identificación y validación de biomarcadores consta de distintas etapas:

-Selección de biomarcador/es: implica la identificación de un conjunto de metabolitos con características óptimas que permitan una máxima discriminación entre los grupos de muestras estudiados.

-Evaluación o análisis: estudia paneles de biomarcadores, analizando su relación con la patología de estudio, y valorando la presencia de los biomarcadores encontrados en un número más amplio de muestras.

-Creación del modelo: evalúa la utilización de una ecuación matemática o algoritmos que permitan combinar el panel de metabolitos seleccionado en una sola prueba cuyo resultado pueda relacionarse con un determinado resultado clínico.

-Validación de modelos: analizan la especificidad y la selectividad los biomarcadores candidatos en la población general, realizando una validación externa del modelo. Posteriormente, se estudian nuevas estrategias para la aplicación clínica del biomarcador, entre ellas validar el modelo utilizando otras técnicas analíticas más sencillas o más accesibles en el entorno clínico.

-Aplicación clínica: utiliza estrategias para aplicar el biomarcador seleccionado a la práctica clínica, buscando técnicas accesibles, evaluando la sensibilidad y especificidad del test en estudios prospectivos, testando el biomarcador frente a muestras recogidas longitudinalmente en estudios de cohorte de más población, etc.

3.3 Tipos

Los biomarcadores han adquirido gran importancia por su papel fundamental en el diagnóstico precoz, prevención y tratamiento de enfermedades como el cáncer. Distintos biomarcadores han demostrado ser de gran utilidad en diagnóstico temprano de la enfermedad y en el tratamiento de este tipo de pacientes. De acuerdo a sus características y utilidad, los biomarcadores se pueden clasificar en las siguientes categorías (Manne *et al*, 2005; Srivastava *et al*, 2001):

- Detección temprana: se aplican para el *screening* o cribado de pacientes con enfermedad subclínica.
- Diagnóstico: se emplean para evaluar presencia o ausencia de enfermedad.
- Pronóstico: se utilizan para evaluar la supervivencia del paciente, detectar un fenotipo agresivo de la enfermedad o la progresión de la misma. Deben ser capaces de diferenciar pacientes con pronóstico favorable frente a los de peor pronóstico.
- Predictivos: se destinan para identificar subpoblaciones de pacientes con mayor probabilidad de responder favorablemente a una terapia determinada.
- Diana: son biomarcadores relacionados con alguna potencial diana terapéutica contra la que sea posible desarrollar o administrar algún tratamiento farmacológico.

4. APLICACIONES CLÍNICAS DE LA METABOLÓMICA

La metabolómica permite la realización de un amplio abanico de estudios en función del objetivo deseado. Por un lado, la metabolómica permite identificar biomarcadores relacionados con la calidad de las muestras biológicas, lo que puede facilitar la evaluación de la estabilidad y calidad de los biofluidos destinados a estudios de metabolómica clínica. Por otro lado, pueden emplearse en la realización de estudios dirigidos a la identificación de biomarcadores en patologías de difícil diagnóstico debido a la elevada variabilidad entre pacientes, o estudios de validación de biomarcadores con utilidad clínica.

4.1 Impacto preanalítico

La fase preanalítica es un punto crítico en el diseño de un estudio metabolómico. Esta fase incluye diferentes etapas: procesado, almacenamiento, transporte, extracción, etc. Estas etapas han estado, hasta el momento, poco estandarizadas, representando una fuente de variabilidad importante en los resultados.

En el diagnóstico clínico de rutina, las variaciones preanalíticas pueden explicar entre el 60% y el 80% de los errores en las pruebas de laboratorio (Carraro *et al*, 2012; Szecsi & Odum, 2009). La información suministrada por los laboratorios clínicos influye hasta en un 60-70% de las decisiones clínicas, lo que da una idea de la envergadura que pueden adquirir estos errores en la atención de los pacientes. Por tanto, uno de los retos actuales de los laboratorios clínicos consiste en la mejora de los procesos preanalíticos.

Además de las distintas fuentes de variabilidad que deben tenerse en consideración en el diseño de cualquier proyecto de investigación clínica, en metabolómica existen otras fuentes de variabilidad que puede condicionar la reproducibilidad y calidad de los estudios. La principal fuente de variabilidad en el análisis metabolómico afecta a la recogida de la muestra. Existe una variabilidad asociada al muestreo y operaciones posteriores, tales como los ciclos de congelación/descongelación o las condiciones de almacenamiento inadecuadas (Lauridsen *et al*, 2007). En los estudios de metabolómica es muy frecuente la utilización de muestras seleccionadas de manera retrospectiva, una vez finalizada la fase de recogida y procesado de las muestras. Este hecho supone una limitación a la hora de poder determinar la idoneidad de las muestras para el estudio.

Un correcto diseño del estudio y el cumplimiento de unos protocolos normalizados de trabajo adecuados son cruciales para minimizar la variabilidad existente en esta fase inicial del estudio. La importancia de estos factores ha sido descrita en varios estudios (Gika *et al*, 2007; Kamlage *et al*, 2014; Lauridsen *et al*, 2007; Lehmann, 2015; Saigusa *et al*, 2016; Shurubor *et al*, 2007; Teahan *et al*, 2006).

En los últimos años, la fase preanalítica se ha intentado estandarizar y protocolizar a través de los biobancos. El objetivo de los biobancos es poder proporcionar un número óptimo de muestras biológicas para estudios de investigación, garantizando la calidad de las mismas. Entre las medidas implantadas en los biobancos, la aplicación del código SPREC (*Standard Preanalytical Coding for Biospecimens*) está recomendada para garantizar el registro sistemático y la preparación de procedimientos de trabajo normalizados en la recogida y extracción de muestras de sangre (Betsou *et al*, 2010).

La existencia de biobancos especializados resulta de gran utilidad a la hora de diseñar un estudio donde se requiere un conjunto de muestras homogéneo y una cantidad razonable de muestras que garantice la validez del estudio. Sin embargo, el almacenamiento masivo de muestras biológicas plantea cuestiones técnicas complejas

que afectan desde la propia recogida de la muestra, su transporte, su identificación, la trazabilidad, su conservación a diferentes temperaturas, la recuperación de la muestra guardada, el tratamiento informático de los datos, etc. En este contexto, es importante identificar metabolitos que sirvan como indicadores de calidad de las muestras y garanticen que las etapas previas a su análisis han sido adecuadas, garantizando la estabilidad de las mismas.

4.2 Variabilidad biológica

Las variaciones biológicas, como la dieta, el ejercicio físico, la regulación homeostática, etc. pueden tener un reflejo en la composición metabólica de las muestras biológicas. Este tipo de variaciones afectan especialmente a biofluidos como la orina, cuya composición se ve muy influenciada por la dieta y el estilo de vida del paciente. Sin embargo, la orina es un biofluido muy utilizado en estudios de metabolómica debido a la diversidad de metabolitos que pueden aparecer en ella y a su sencilla obtención (Bouatra *et al*, 2013). En patologías como el cáncer de próstata (CaP), la orina puede resultar un medio idóneo para la búsqueda de biormarcadores. Actualmente, esta patología no dispone de un método de diagnóstico ideal. La falta de sensibilidad y especificidad da lugar a falsos positivos o falsos negativos dificultando un adecuado diagnóstico de esta patología.

El CaP es el cáncer más frecuente en hombres, con una incidencia cercana al 22% del total de nuevos casos diagnosticados en 2014 y una prevalencia a 5 años del 31,4% (SEOM, 2014). En la actualidad, el cribado para detección del CaP incluye la determinación de los niveles de PSA en sangre y el examen del tacto rectal (DRE, *digital rectal exam*). Sin embargo, los estudios realizados en relación a la detección precoz de CaP no han demostrado una disminución en la mortalidad por este tumor (Ilic *et al*, 2013). El test del PSA presenta una baja especificidad, dando lugar a una tasa alta de falsos positivos que se traduce en biopsias prostáticas innecesarias y en un sobretratamiento de tumores con bajo potencial maligno (Draisma *et al*, 2003). Además, la biopsia prostática es la prueba imprescindible para la confirmación histológica del CaP, pero presenta una alta tasa de falsos negativos (alrededor 30%) (Rabbani *et al*, 1998; Schoenfield *et al*, 2007).

En este contexto, la metabolómica representa una oportunidad para la búsqueda de nuevos biomarcadores y la identificación de rutas biológicas implicadas en la aparición y progresión del CaP. Hasta la fecha se han realizado diversos estudios metabólicos en diferentes biofluidos (suero, orina, tejido, fluido prostático,…) con este objetivo (Kumar *et al*, 2015; Stabler *et al*, 2011; Zhang *et al*, 2013a). A pesar de que la

búsqueda de biomarcadores en orina mediante la metabolómica por RMN representa una oportunidad en la profundización de los mecanismos moleculares del CaP. El análisis de este biofluido es extremadamente complejo debido al elevado número de metabolitos que contiene y la variabilidad interindividual reflejada en este tipo de muestras. En este sentido, se requiere la aplicación de diferentes estrategias en las fases del estudio metabolómico y la aplicación de métodos estadísticos robustos que permitan discriminar los cambios asociados al problema biológico de interés (CaP) de las variaciones aleatorias (variaciones en la dieta, estilo de vida, edad, etc.).

4.3 Validación de biomarcadores

Un biomarcador antes de llegar a la clínica debe ser validado utilizando un número considerable de muestras y demostrar su reproducibilidad, sensibilidad y especificidad (Mamas *et al*, 2011). Actualmente, existen pocas publicaciones donde los modelos metabolómicos hayan sido validados estadísticamente (Beckonert *et al*, 2003; Jonsson *et al*, 2005). Uno de los retos actuales en metabolómica es, por tanot, la validación de los estudios metabólicos y de los biomarcadores derivados de los mismos. La falta de validación de los estudios de metabolómica supone una limitación en la traslación de nuevos biomarcadores al ámbito clínico. Una de las dificultades en la validación de este tipo de estudios es la necesidad de obtener un número elevado de muestras representativas.

En los últimos años, se han publicado un gran número de estudios con resultados prometedores en la caracterización del perfil metabolómico de pacientes con cáncer de pulmón no microcítico (CPNM) (Carrola *et al*, 2011; Deja *et al*, 2014; Jordan *et al*, 2010; Nobakht M. Gh *et al*, 2015; Rocha *et al*, 2011). Sin embargo, no se ha llevado a cabo la validación externa de estos estudios para la detección temprana de esta enfermedad, ya que requiere de análisis con un mayor número de muestras que permitan la identificación y confirmación de los biomarcadores identificados.

El cáncer de pulmón (CP) es la causa más común de muerte por cáncer en el mundo, representa aproximadamente el 12% de todos los casos de cáncer, con una incidencia de casi dos millones de nuevos casos anuales en el mundo (Ferlay *et al*, 2015). En este contexto, la validación un nuevo biomarcador, o biomarcadores, con utilidad en la detección temprana de esta enfermedad, podría representar un gran avance en el contexto de la enfermedad.

II. OBJETIVOS Y ESTRUCTURA

El objetivo de la presente Tesis Doctoral es la evaluación del potencial de la metabolómica en el ámbito clínico. Para ello se plantean diversos estudios que abarcan aspectos claves en el descubrimiento de biomarcadores de utilidad clínica, como son:

-La evaluación del impacto de diferentes variaciones preanalíticas sobre el perfil metabólico de muestras de plasma y suero.

-El desarrollo de una estrategia de análisis dirigido a la caracterización de los cambios metabólicos en la orina de pacientes con CaP asociados a la enfermedad, y su discriminación frente a los cambios metabolómicos derivados de la variabilidad biológica entre muestras.

-La aplicación de una estrategia específica para la validación de un conjunto de metabolitos que puedan emplearse como biomarcadores de diagnóstico precoz de CPNM.

III. METODOLOGÍA

1. DISEÑO EXPERIMENTAL DE LOS ESTUDIOS Y RECOGIDA DE MUESTRAS

La recogida de muestras de los estudios se realizó tras la obtención del consentimiento informado de los pacientes, de conformidad con los principios fundamentales establecidos en la Declaración de Helsinki y con la aprobación de los Comités Éticos de Investigación de los centros que participaron en los estudios.

Las muestras de sangre y orina se procesaron y almacenaron siguiendo los protocolos previamente descritos para estudios de metabolómica por RMN (Beckonert *et al*, 2007).

1.1 Impacto preanalítico

El estudio se realizó con muestras de sangre periférica procedente de 40 voluntarios sanos del Hospital Universitario Dr. Peset de Valencia y del Centro Superior de Investigación en Salud Pública (CSISP) recogidas de manera prospectiva.

A cada paciente incluido en el estudio se le extrajo un volumen máximo de 40 ml de sangre. Las muestras de sangre se centrifugaron después de la extracción, antes de 30 minutos a 4ºC, a una velocidad de 1600g durante 15 minutos. La sangre extraída a cada voluntario se dividió en alícuotas (300µL) y se almacenaron 5 réplicas para cada condición de estudio a -80ºC.

En la **Tabla 2** se describen los distintos factores de variabilidad preanalítica que se evaluaron en este estudio, que incluyen:

a) La presencia de distintos aditivos en los tubos de extracción. Para ello, se analizó sangre recogida en 5 tubos de extracción: EDTA, secos sin partículas (SST), con gel separador (*clot activator*), con heparina y con citrato. Para cada tipo de tubo se recogió también un blanco (50 g/L albúmina humana en suero salino).

b) La temperatura y tiempo al que se conservaron las muestras desde la extracción hasta su centrifugación y almacenamiento.

c) El grado de hemólisis en las muestras. Se evaluó el efecto que tienen distintos grados de hemólisis en muestras de plasma. La hemólisis se provocó de forma mecánica, haciendo pasar la sangre por una aguja (0,9x25mm) y posteriormente se midió el índice de hemólisis con un analizador ARCHITECT con el fin de cuantificar los diferentes grados de hemólisis (control, hemólisis moderada y hemólisis intensa). Las muestras control fueron muestras obtenidas del mismo paciente, sin hemolizar.

d) Los ciclos de congelación-descongelación. Las alícuotas de suero y plasma se sometieron a distintos ciclos de congelación y descongelación. El proceso de descongelación se realizó dejando las muestras a temperatura ambiente durante dos horas, posteriormente homogeneizando en el vórtex y volviéndose a congelar a -80ºC.

e) El tiempo de almacenamiento. Las muestras se almacenaron a -80ºC durante 6 meses, 1 año o 2 años.

Tabla 2: Resumen de las condiciones preanalíticas evaluadas en el estudio.

a) Distintos aditivos en los tubos de extracción	EDTA	SST (gelosa)	Clot activator (seco)	HEPARINA	CITRATO
b) Temperatura y tiempo	30 min.	1 hora	2 horas	6 horas	24 horas
	4ºC y Temperatura ambiente				
c) Grado de hemólisis	Control	Moderada	Intensa		
d) Ciclos de congelación-descongelación	1-5 ciclos				
e) Tiempo de almacenamiento	Control	6 meses	1 año	2 años	

1.2 Variabilidad biológica

El estudio incluyó un total de 140 muestras de orina recogidas en el Servicio de Urología y el Biobanco del Instituto Valenciano de Oncología (IVO), distribuidos de la siguiente manera:

- 71 pacientes con CaP: 46 pacientes de bajo riesgo y 25 pacientes de alto riesgo.
- 69 muestras de pacientes con Hiperplasia benigna de próstata (HBP).

La clasificación de los pacientes se realizó teniendo en cuenta los niveles de PSA, tacto rectal, el resultado de la biopsia y la puntuación del *Gleason* (**Tabla 3**). La biopsia se realizó analizando al menos seis cilindros y la clasificación de los individuos incluidos en el estudio se realizó de acuerdo con la guía de cáncer de próstata EAU-ESTRO-SIOG (Mottet *et al*, 2016). El grupo control incluía hombres sin CaP, diagnosticados de HBP atendiendo a sus niveles de PSA y los resultados negativos del tacto rectal y de la biopsia.

Tabla 3: Características clínicas de los pacientes incluidos en el estudio.

	HBP (media, rango)	**CaP (media, rango)**
Edad (años)	61.7 (41.4-74.5)	64.9 (50.0-86.3)
BMI (kg m^{-2})	27 (22.8-34)	27.5 (23-33)
Volumen prostático (ml)	52 (24-171)	40.5 (2-134)
PSA (ng/ml)	4.97 (1.02-11.29)	7.10 (0.85-71.41)
Núbero de cilindros	13.45 (10-19)	16.06 (6-54)
Cilindros positivos	No aplica	10.86% (0.05-100)
Carga tumoral	No aplica	8.94% (0.14-67.74)
Gleason score	No aplica	6.47 (5-9)

Las muestras de orina (10 ml) se recogieron en tubos sin aditivos (BD, Vacutainer, sin aditivos) a los que se añadió NaN_3 (ázida sódica) al 0.05%. Las muestras se almacenaron a -80ºC hasta su análisis por RMN.

1.3 Validación de biomarcadores

Este estudio es la continuación de un estudio previo en el que se analizaron mediante resonancia magnética nuclear de protón (^{1}H-RMN) los perfiles metabólicos de 216 muestras: 74 individuos sanos y 142 pacientes con cáncer de pulmón no microcítico (CPNM). Las muestras de los pacientes con CPNM se clasificaron en estadios tempranos (n=72) y estadios avanzados (n=70) (Greene FL, 2003). Este es el grupo de muestras que constituye el set de entrenamiento en este estudio de validación.

En el set de validación se incluyeron 80 muestras recogidas de manera externa al set de entrenamiento (**Tabla 4**): 13 muestras pertenecientes a individuos sanos, 20 pacientes con CPNM en estadios tempranos y 20 en estadios avanzados. En el grupo control se incluyeron 27 muestras de suero procedentes de pacientes diagnosticados con enfermedad pulmonar benigna (EPB). Se recogieron de 1-2 ml sangre en un tubo libre de anticoagulante o de cualquier otro aditivo (BD Vacutainer, sin aditivos).

Tabla 4: Características clínicas y demográficas de las muestras incluidas en el set de validación y en el set de entrenamiento.

	Set de entrenamiento			Set de validación			
	Control	**CPNM-ET**	**CPNM-EA**	**Control**	**EPB**	**CPNM-ET**	**CPNM-EA**
Número total	74	72	70	13	27	20	20
Sexo							
Femenino	13	8	12	2	13	5	7
Masculino	61	64	58	11	14	15	13
Edad	56 ± 1.55	63 ± 1.17	63 ± 1.29	47 ± 1.78	52 ± 2.88	68 ± 1.48	61 ± 2.28
Hábito tabáquico							
Ex-fumador	22	25	21	1	8	9	10
Fumador	31	42	20	8	7	8	4
No fumador	21	5	5	3	11	3	6
Desconocido			24	1	1		
Histología							
Adenocarcinoma		24	32			9	16
Carcinoma de células grandes		2	4			1	1
Carcinoma epidermoide		38	27			9	
Otras		8	7			1	3
Estadios							
IA		10				5	
IB		27				4	
IIA		1				5	
IIB		17				4	
IIIA		17				2	
IIIB			18				
IV			52				20
Otra patología							
EPOC					9		
TB					3		
Neumonía					4		
BC					2		
Otras					9		

2. PREPARACIÓN DE LAS MUESTRAS

Todas las muestras permanecieron congeladas a -80ºC hasta su análisis por RMN. De acuerdo con los protocolos desarrollados para este tipo de estudios (Dona *et al*, 2014) las muestras se prepararon de la siguiente manera para su análisis por RMN:

2.1 Orina

La muestras de orina se descongelaron en hielo y se centrifugaron a una velocidad de 6000g durante 5 minutos a temperatura ambiente. Se añadió 60 µL de solución tampón 100% D_2O (1.5 M KH_2PO4, 0.1% TSP, 0.05% NaN3, pH 7.4) a 540 µL del

sobrenadante de la muestra de orina. Finalmente, 500 µL de la mezcla se transfirieron a un tubo de 5-mm de diámetro de RMN para su análisis.

2.2 Suero y plasma

Las muestras de suero y de plasma se descongelaron en hielo. Posteriormente, se utilizó 300µL de la muestra y se les añadió un volumen equivalente de una solución tampón (140 mM Na_2HPO_4, 5 mM TSP, 0.04% NaN_3, pH 7.4). La solución tampón se preparó con un 100% de D_2O para el estudio de variabilidad preanalítica y con un 10% de D_2O para las muestras del estudio de validación de biomarcadores. De la mezcla resultante, se transfirieron 550 µL a un tubo de RMN de 5mm de diámetro para su análisis.

3. ADQUISICIÓN DE LOS ESPECTROS DE RMN

Siguiendo los protocolos recomendados por Bruker (Bruker, Biospin) para estudios de metabolómica con muestras biológicas, y antes de proceder a la adquisición de cada conjunto de experimentos, se procedió a la optimización y calibración de todos los parámetros críticos en este tipo de estudios. Primero se calibró la temperatura con una muestra de metanol deuterado al 99.8%. Posteriormente, se optimizaron los *shims* y la frecuencia de la señal del agua (O1), por un lado, con una muestra patrón de sacarosa (2mM sacarosa en 90% H_2O + 10% D_2O) y, por otro, se ajustó para una muestra representativa del estudio (50% suero/plasma + 50% solución tampón y 90% orina + 10% solución tampón, respectivamente) hasta conseguir la calidad de los espectros recomendada para estudios de metabolómica: línea base plana, buena corrección de fase, señal simétrica de TSP (anchura <1Hz) y señal simétrica del agua (anchura <100 Hz).

La adquisición de los experimentos de ^{1}H-RMN para cada muestra incluida en el estudio se realizó a 310K (suero/plasma) o a 300K (orina). Se empleó un espectrómetro Bruker AVII 500 MHz para el análisis del estudio de la variabilidad preanalitica y para el estudio de variabilidad biológica. Para el análisis de validación de biomarcadores se utilizó un espectrómetro Bruker Avance II 600 MHz equipado con una criosonda TCI de 5 mm. Tanto la adquisición como el procesado de estos experimentos se realizaron con el programa TOPSPIN 3.0 (Bruker, Biospin).

Para cada muestra incluida en el estudio se obtuvo un experimento ^{1}H-RMN CPMG (Meiboom & Gill, 1958). La secuencia de comandos que se utilizó incluía:

-Ajuste de la sintonía de la sonda.

-Estabilización y homogenización del campo magnético.

-Cálculo automático del pulso de 90º (P1) para cada muestra.

-TF.

-Pre-procesado automático de los espectros.

La adquisición de los espectros de RMN en cada grupo de muestras analizadas se realizó de manera conjunta y alternando de manera aleatoria el orden de las muestras en función de los distintos grupos incluidos en cada estudio.

Las condiciones de adquisición se adaptaron en función del biofluido analizado (**Tabla 5**):

Tabla 5: Parámetros de adquisición de los espectros ^{1}H-RMN CPMG.

	Impacto preanalítico	**Variabilidad biológica**	**Validación de biomarcadores**
Secuencia de pulso	cpmgpr1d.comp		
Temperatura	310 K	300 K	310K
TD	61440	66560	73728
NS	128	32	64
DS	8	4	8
SW (ppm)	20.0482	19.9947	20.0276
D1	4 s	4 s	4 s
RG	362	90.5	80.6

(K: grados Kelvin; TD: dominio tiempo; NS: número de scans; DS: dummy scans; SW: anchura espectral; D1: periodo de repetición del experimento; P1: pulso de 90º; RG: ganancia del receptor).

Adicionalmente, para algunas muestras de cada uno de los estudios se obtuvieron experimentos 2D (TOCSY) para facilitar la asignación y la identificación de metabolitos. La secuencia de pulso que se utilizó fue dipsi2gpph19.

4. ANALISIS DE LOS DATOS

4.1 Pre-procesado y procesado de datos de los espectros de RMN

La corrección de la fase y la línea base de todos los experimentos monodimensionales se realizó de manera automática utilizando el comando “apk0.noe” del paquete informático TopSpin 3.0 (Bruker, BioSpin).

Los experimentos de RMN utilizados en los tres estudios para construir los modelos estadísticos fueron ^{1}H-RMN CPMG. La integración de los espectros se realizó utilizando el programa AMIX (Analysis of MIXtures, Bruker BioSpin). Los parámetros utilizados en cada estudio se describen en la **Tabla 6**.

Tabla 6: Parámetros empleados en la integración de los espectros en los diferentes estudios.

Impacto preanalítico	
Método de integración	Suma de intensidades absolutas
Señal de referencia	Grupo metilo alanina (1.47 ppm)
Región de espectro	8.61-0.26 ppm
Regiones excluidas	5.14-4.19 ppm (agua):plasma y suero 5.79-5.65 ppm (urea): plasma y suero 3.63-3.54 ppm (EDTA): plasma 3.28-3.04 ppm (EDTA): plasma 2.78-2.65 ppm (EDTA): plasma 2.58-2.51 ppm (EDTA): plasma
Tamaño de *buckets*	0.04 ppm
Normalización	Área total
Escalado	Pareto
Variabilidad biológica	
Método de integración	Suma de intensidades absolutas
Señal de referencia	TSP (0.00 ppm) / speaq
Región de espectro	9.50-0.15 ppm
Regiones excluidas	4.55-5.09 ppm (agua) 6.10-5.52 ppm (urea)
Tamaño de *buckets*	0.001 ppm
Normalización	PQN
Escalado	Varianza unitaria
Validación de biomarcadores	
Método de integración	Suma de intensidades absolutas
Señal de referencia	Grupo metilo alanina (1.47 ppm)
Región de espectro	9.02-0.14 ppm.
Regiones excluidas	5.06-4.30 ppm (agua) 5.85-5.60 ppm (urea)
Tamaño de *buckets*	0.01 ppm
Normalización	Área total
Escalado	Pareto

El tipo de escalado que se utilizó fue distinto en cada estudio con el objetivo de conseguir modelos estadísticos más óptimos según el tipo de muestras analizadas. Para el estudio de la impacto preanalítico y el de validación de biomarcadores se utilizó el método Pareto y para el estudio de variabilidad biológica se utilizó el método de escalado de varianza unitaria. El escalado se aplicó tras la normalización de los datos.

En el estudio de la variabilidad biológica se aplicó sobre los datos de la integración de las muestras de orina una corrección para conseguir una correcta alineación de los espectros. Para ello, se utilizó “speaq” (Vu *et al*, 2011) de R (http://www.R-project.org), esta herramienta incluye un algoritmo de alineamiento llamado CluPa (*Hierarchical Cluster-based Peak Alignment*), que permite definir un valor para el ruido e identificar como señales los picos con intensidades por encima de este valor. Posteriormente se

escoge un espectro de referencia que es utilizado como patrón a partir del cual se alinean el resto de espectros. El algoritmo selecciona aquel espectro que tiene mayor número de picos comunes al resto de espectros. Finalmente, siguiendo el criterio de máxima verosimilitud se alinean las señales de todos los espectros.

En el estudio de validación de biomarcadores, sobre los datos obtenidos tras la integración de los espectros, se realizó un paso adicional de normalización para facilitar la combinación de los dos conjuntos de datos incluidos en el estudios (Stein *et al*, 2015). Para ello, se empleó la función ComBat incluida en el paquete "sva" de R. Este paquete bioinformático permite eliminar la variabilidad derivada de la utilización de distintos lotes de muestras, medidos en diferentes tiempos, y evitar así la variabilidad metodológica. La función SVA (*Surrogate Variable Analysis*) corrige los efectos en el lote de entrenamiento, creando un algoritmo que utiliza como clasificador, basado en este conjunto de datos de entrenamiento limpio o filtrado. Posteriormente, utiliza la probabilidad y los coeficientes estimados de la base de datos del lote de entrenamiento para eliminar los efectos de lote de las nuevas muestras (Parker *et al*, 2014).

4.2 Análisis estadístico de los datos

4.2.1 Análisis estadístico multivariante

El análisis estadístico multivariante se realizó empleando SIMCA-P 12.0 (Umetrics AB, Suecia). En los tres proyectos se realizó un análisis no supervisado mediante modelos PCA para el estudio global de las muestras. Se evaluó la homogeneidad dentro de los grupos y se identificaron posibles *outliers* utilizando los gráficos T^2 de *Hotelling*, *score plot* y la inspección visual.

En el estudio de variabilidad biológica con muestras de orina fue necesario aplicar una selección de variables previa a la construcción de los modelos OPLS-DA, como estrategia para facilitar la interpretación y el estudio de los modelos. La selección de variables que se utilizó fue la de los coeficientes de regresión (CoeffCS) y sus correspondientes errores estándar obtenidos tras la validación cruzada (CoeffCScvSE). Se mantuvieron sólo aquellas variables cuya relación CoeffCS/CoeffCScvSE fue superior a uno (Diaz *et al*, 2013).

Posteriormente, se realizó un análisis supervisado para identificar posibles diferencias entre los grupos analizados mediante el diseño de modelos OPLS-DA. En determinados casos, donde la variable dependiente analizada era continua, se utilizó una variante del modelo OPLS-DA, llamada Y-OPLS. De los modelos OPLS-DA generados se obtuvo un listado con las K-variables empleadas en el modelo y su

correspondiente valor de VIP. Se consideró que los descriptores con valores de VIP superiores a uno eran los que contribuían en mayor medida a las discriminación entre los grupos de estudio. Posteriormente, se identificaron y se asignaron los metabolitos correspondientes a estas señales en el espectro de RMN.

En todos los estudios se realizó una validación interna utilizando la validación cruzada y el test de permutación, a partir de modelos equivalentes de PLS-DA con el mismo número de componentes que el modelo OPLS-DA original.

4.2.2 Cuantificación de los metabolitos relevantes

Una vez identificadas las señales más relevantes en la discriminación entre grupos y los metabolitos responsables de las diferencias, se realizó de nuevo una integración de los espectros de RMN, esta vez dirigida específicamente a las señales identificadas. Esta nueva integración proporcionó una tabla de tamaño variable (*Variable Size bucketing*) en la cual cada región integrada contenía una señal de un metabolito. Las señales del espectro se inspeccionaron visualmente para identificar las regiones que pertenecían a un mismo metabolito y se realizó la integración de forma manual o semi-automática.

La cuantificación de los metabolitos de interés se realizó utilizando el programa AMIX (Bruker BioSpin). En el estudio de caracterización de biomarcadores, debido a la complejidad de los espectros de orina, se utilizó también la herramienta de integración disponible en MNova (MestReNova v8.1.2) para la integración de los metabolitos.

4.2.3 Análisis estadístico univariante

El análisis estadístico univariante se realizó con el paquete estadístico “stats” de R y el paquete IBM SPSS Statistics 19, considerando un valor de $p<0.05$ como estadísticamente significativo. También se empleó el paquete “Limma” de R para el análisis estadístico del estudio variabilidad preanalitica.

5. ASIGNACIÓN DE LOS ESPECTROS

La asignación de las señales de los espectros se basó en la información disponible en bases de datos especializadas (HMBD, BMRB, PRIMe, etc.) combinada con la base de datos comercial BBIOREFCODE (Bruker BioSpin). Esta información fue complementada con la información disponible en la literatura sobre la asignación de metabolitos en muestras de suero (Bouatra *et al*, 2013; MacIntyre *et al*, 2010; Nicholson *et al*, 1995; Psychogios *et al*, 2011; Yang *et al*, 2008) y en orina (Bouatra *et al*, 2013; Diaz *et al*, 2013; Jacobs *et al*, 2012; Yang *et al*, 2008). De forma complementaria, se adquirieron espectros de RMN de compuestos patrón que

facilitaron la confirmación de algunas asignaciones, así como el estudio de espectros 2D para facilitar la asignación de señales. También se utilizó la técnica de *spiking* para resolver la identificación de determinadas señales.

IV. RESULTADOS Y DESARROLLO ARGUMENTAL

1. IMPACTO PREANALÍTICO

Una adecuada fase preanalítica es muy importante para minimizar cualquier variabilidad que pueda influir en la calidad de los resultados obtenidos en los estudios de metabolómica. La aplicación de protocolos de trabajo estandarizados y la identificación de biomarcadores que sirvan como indicadores adecuados para el control de calidad de las muestras utilizadas en investigación resulta esencial. El objetivo del presente estudio fue evaluar el impacto de distintos factores que pueden afectar durante la fase preanalítica a la calidad de las muestras de suero y plasma destinadas a estudios de metabolómica por RMN.

1.1 Presencia de aditivos

En primer lugar se analizaron los espectros ^{1}H-RMN CPMG correspondientes a los blancos de los cinco tubos de extracción de sangre elegidos para el estudio. Tres de ellos son utilizados habitualmente en el ámbito sanitario para la recogida de muestras de plasma (citrato, heparina y EDTA) y dos para la recogida de muestras de suero (*clot activator* y SST). En la **Figura 14** se muestran los espectros de las distintas muestras analizadas. Como se puede observar, tanto el tubo que contiene citrato, como el que contiene EDTA muestran señales en el espectro que podrían interferir con la cuantificación de determinadas regiones en el espectro de las muestras de plasma (citrato: 2,54 y 2,66 ppm; EDTA: 3,23 y 3,62 ppm). Estas señales son debidas a los protones de la estructura química del anticoagulante. Las muestras obtenidas para los otros tres tubos: heparina para la recogida de plasma y el tubo seco (*clot activator)* o con gel separador (SST) para la recogida de suero, no presentan señales que interfieran en el análisis del perfil metabolómico por ^{1}H-RMN.

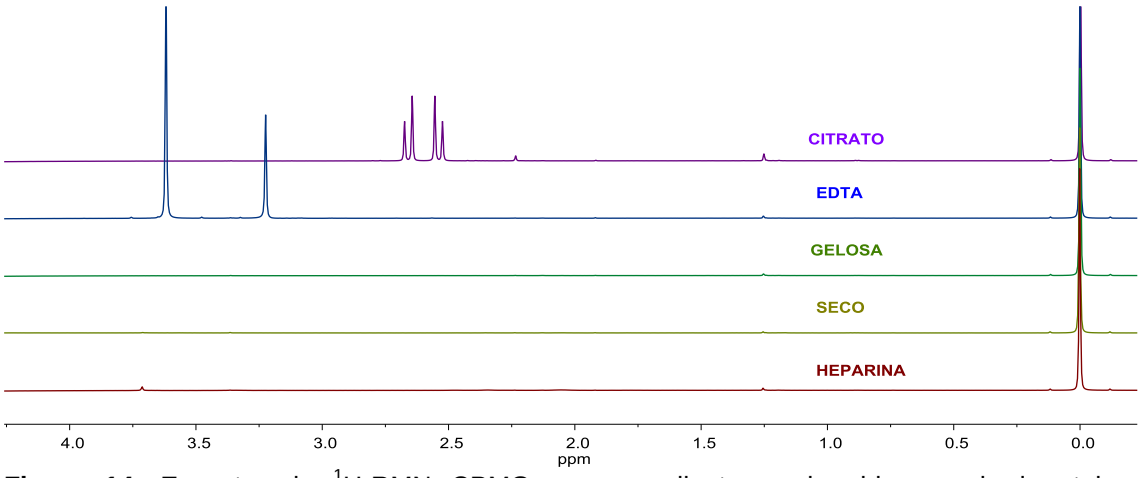


Figura 14: Espectro de ^{1}H-RMN CPMG correspondientes a los blancos de los tubos analizados.

Posteriormente, se analizaron los espectros ^{1}H-RMN CPMG obtenidos para muestras recogidas utilizando cada uno de los tubos incluidos en el estudio. El objetivo de este análisis fue evaluar si existían interacciones entre el anticoagulante del tubo y el biofluido que pudieran dar lugar a la presencia de otras señales en el espectro.

En este análisis, sólo se observaron nuevas señales en el espectro correspondiente a la muestra de plasma recogida con el tubo de EDTA (**Figura 15**). Al superponer el espectro de plasma recogida en el tubo con EDTA y el espectro de plasma recogida con el tubo de heparina se observó que, aparte de las señales propias de la molécula libre de EDTA (3.24 y 3.63 ppm), aparecían otras señales. Estas nuevas señales se observaron a: 2.58, 2.72 y 3.14 ppm, y corresponden a complejos formados entre la molécula de EDTA y los iones presentes en la sangre (Ca^{2+} y Mg^{2+}) (**Figura 15A**). Estas nuevas señales aparecen en el espectro de RMN en unos desplazamientos químicos en los que también lo hacen otras señales correspondientes a metabolitos de interés presentes en las muestras de plasma, como son colina, dimetilamina, glicerol, glucosa, citrato, etc. La presencia de estas señales dificultaría en gran medida la cuantificación de estos metabolitos en las muestras de plasma recogidas en este tipo de tubos. Estos resultados concuerdan con los encontrados por el grupo de Pinto y colaboradores (Pinto *et al*, 2014).

En resumen, el tipo de tubo elegido para la extracción de las muestras de sangre puede suponer una fuente importante de interferencias exógenas, no relacionadas con el biofluido analizado. En los estudios de metabolómica por RMN, el uso de tubos con EDTA o citrato para la extracción de plasma no está recomendado, mientras que la recogida de muestras de plasma con tubos de heparina no presenta problemas. En cambio, este tipo de tubos es el más adecuado para los estudios de metabolómica por EM. En EM, los tubos con heparina de litio o el uso de anticoagulantes con otros cationes no son recomendables ya que interfieren con los metabolitos en el proceso de ionización (Yin *et al*, 2013).

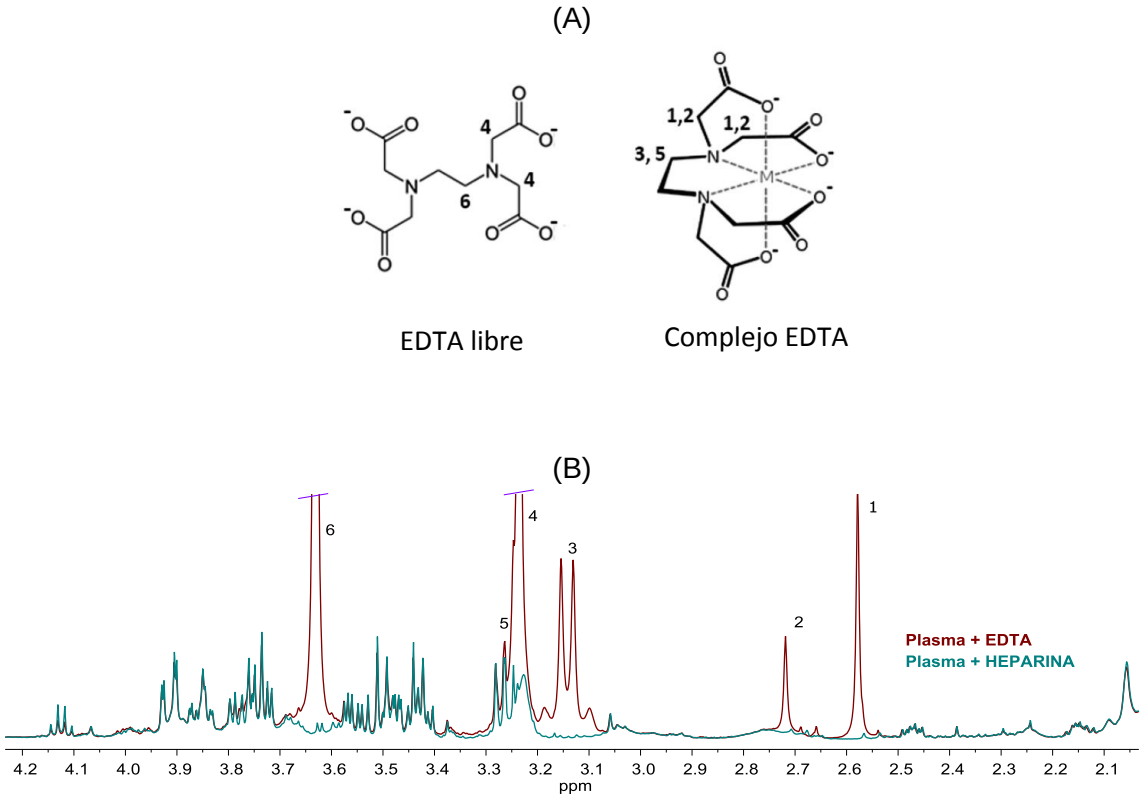


Figura 15: (A) Estructura química de la molécula de EDTA libre y cuando forma complejo con metales. (B) Comparación de los espectros ^{1}H-RMN de dos muestras de plasma recogidas en tubos con anticoagulantes distintos (granate: EDTA; verde: heparina). La numeración de las señales en el espectro (B) se corresponde con la numeración de los protones en la molécula de EDTA libre y cuando forma el complejo (A).

1.2 Temperatura y tiempo de procesado

El tiempo y la temperatura a la que permanecen las muestras desde su extracción hasta que son almacenadas pueden influir en la estabilidad de las mismas. Para estudiar el posible impacto de estas variables en la calidad de las muestras, se analizaron los espectros de muestras de plasma y suero procesadas en distintas condiciones. Unas muestras se procesaron a temperatura ambiente (TA) y otras a 4ºC, en diferentes tiempos: 30 minutos, 1, 2, 6 y 24 horas.

Inicialmente, se llevó a cabo el análisis no supervisado (PCA) con la finalidad de explorar el comportamiento de las muestras incluidas en esta condición del estudio. En la **Figura 16** se observa cómo, dentro de las muestras correspondientes a cada voluntario incluido en el estudio, existe una agrupación de las muestras en función del tiempo que han tardado en ser procesadas, siendo ésta más acusada en las muestras conservadas a TA que en las conservadas a 4ºC. Este fenómeno se observó tanto en las muestras de plasma como en las muestras de suero (**Figura 16**). Las muestras de los dos voluntarios incluidos en esta condición presentaron el mismo comportamiento.

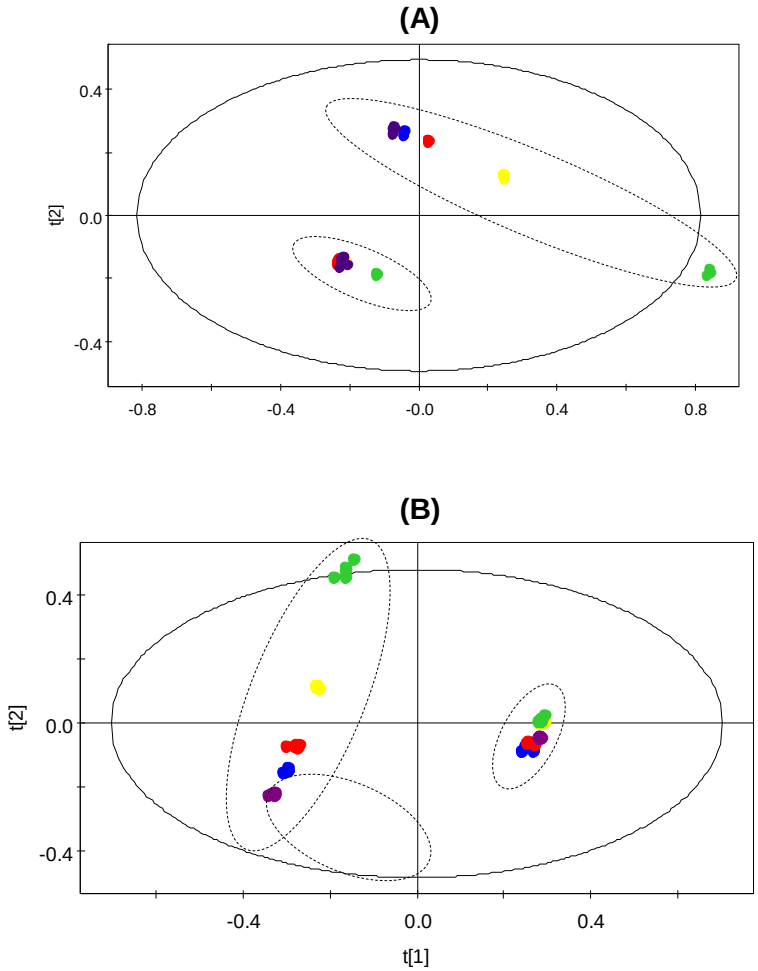


Figura 16: Análisis de componentes principales (PCA) de los perfiles metabolómicos correspondiente a muestras de plasma (A) y suero (B) procesadas a TA y 4ºC, respectivamente, en distintos tiempos (● 30 minutos, ● 1 hora, ● 2 horas, ● 6 horas, ● 24 horas). Las elipses circunscritas de mayor y menor tamaño dentro de cada gráfico corresponden a TA y 4ºC, respectivamente.

En la **Figura 17** se muestra, el *loading plot* obtenido para el análisis realizado sobre las muestras procesadas a TA. En el gráfico se representa la contribución de cada variable en el espectro a la separación entre las muestras procesadas a TA. Las variables que presentan mayor variación entre las muestras procesadas a distintos tiempos coinciden en ambos biofluidos. Este tipo de gráficos proporciona información útil para identificar cuáles son las variables responsables de las tendencias observadas en la distribución de las distintas muestras. En el gráfico de la **Figura 17** se observa cómo la señal marcada a 1.32 y a 4.12 ppm, correspondientes al lactato, están relacionadas con el componente principal que define la distribución de las muestras en función del tiempo de procesado en el modelo, siendo el lactato uno de los metabolitos con alteración significativa en función del tiempo de procesado.

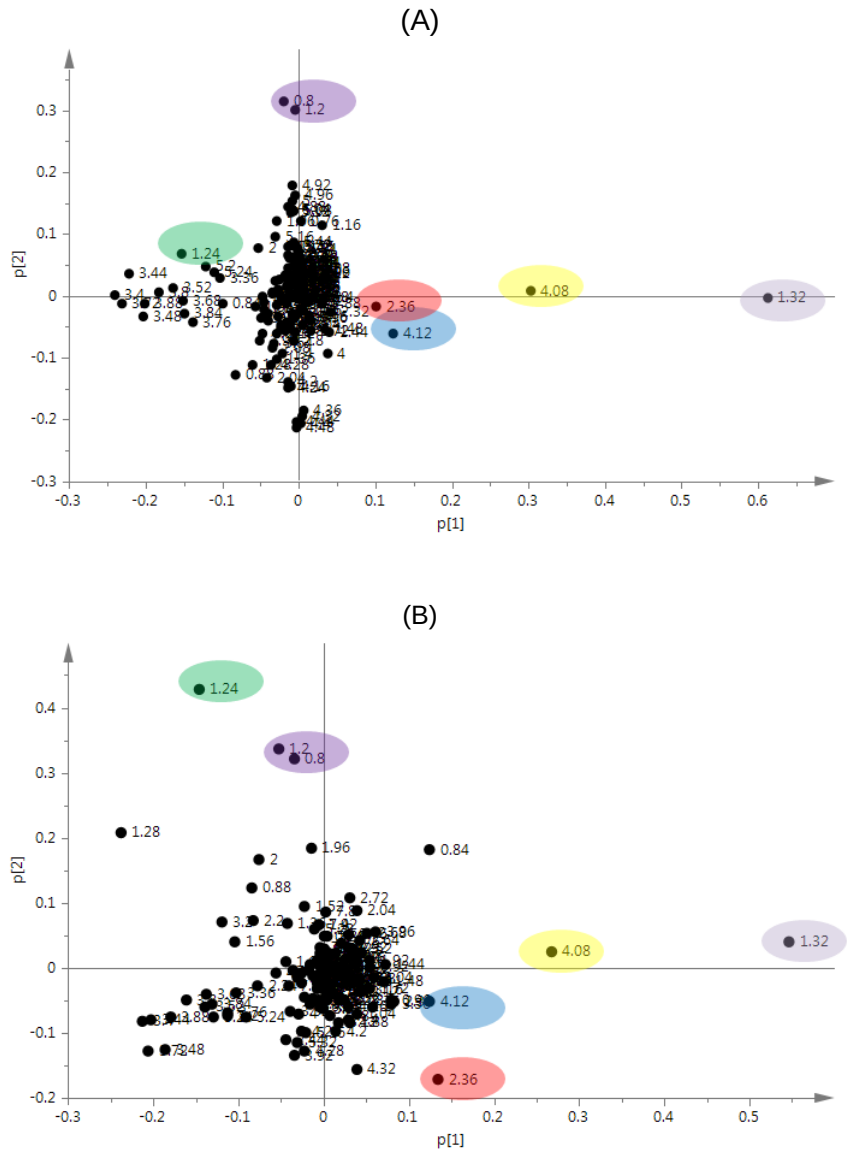


Figura 17: Representación de los *loading plot* correspondientes a las muestras de plasma (A) y suero (B), respectivamente, procesadas a TA. Las variables que coinciden en ambos biofluidos se muestran coloreadas.

El análisis supervisado de los espectros ^{1}H-RMN CPMG se realizó tras el análisis exploratorio. Con el fin de analizar las diferencias en la estabilidad de las muestras en función del tiempo que permanecían a TA o a 4ºC se generaron modelos OPLS-Y, utilizando como variable para la clasificación el tiempo de procesado (**Tabla 7**). Se construyó un modelo para cada temperatura de estudio.

Tabla 7: Valores de R^2 (Y) y Q^2 (Y) para los modelos OPLS-Y en los que las muestras se clasifican en función del tiempo de procesado.

BIOFLUIDO	OPLS	R^2(Y)	Q^2(Y)
PLASMA	**TA**	0.988	0.932
	4ºC	0.993	0.841
SUERO	**TA**	0.996	0.989
	4ºC	0.875	0.834

Dentro de cada modelo Y-OPLS se identificaron cuáles eran las variables que mejor caracterizaban a cada uno de los grupos, y se analizaron cuáles eran las regiones que contribuían mayoritariamente a las diferencias observadas en función del tiempo de procesado. Posteriormente, se llevó a cabo el análisis univariante de cada una de estas variables, con la finalidad de confirmar la significación estadística de los cambios observados. Para cada biofluido se generó un patrón de integración común para la integración de las señales de los metabolitos identificados como relevantes en cualquiera de los modelos analizados. Se realizó un procedimiento equivalente para la integración de las muestras de suero y plasma, por separado.

En la **Tabla 8** se muestra el porcentaje de variación en las intensidades de las señales de los metabolitos que se identificaron como relevantes en los OPLS-Y correspondientes entre las muestras procesadas 24 h después de la recogida y los niveles encontrados en la condición de control (30 minutos), a las distintas temperaturas estudiadas en las muestras de plasma.

Tabla 8: Coeficiente de variación de los metabolitos más relevantes en la comparación entre la condición control (30 minutos) y las muestras procesadas tras 24 horas a TA y 4ºC, respectivamente, en muestras de plasma.

		TA		**4ºC**	
METABOLITO	**ppm**	**Variación**	**p-valor[a]**	**Variación**	**p-valor[a]**
Lactato	1.32-1.30	527.53 ↑	0.00*	55.61 ↑	0.00*
Glucosa	5.23-5.20	-61.21 ↓	0.00*	-6.76 ↓	0.00*
VLDL/HDL	0.90-0.76	-2.70 ↓	0.00*	-1.74 ↓	0.00*
VLDL/LDL	1.30-1.19	-9.30 ↓	0.00*	-0.55 ↓	0.00*
Glutamato	2.35-2.28	37.56 ↑	0.00*	-0.28 ↓	0.01*
Piruvato	2.36-2.35	176.36 ↑	0.00*	-14.32 ↓	0.00*
Ácido adípico	1.61-1.50	-15.77 ↓	0.00*	-0.10 ↓	0.13
Grupos CH_2-CO lípidos	2.24-2.18	-8.43 ↓	0.00*	1.12 ↑	0.13
Grupos CH=CH lípidos	5.36-5.23	-3.96 ↓	0.00*	-1.73 ↓	0.00*

[a] p valor calculado mediante la prueba t de *Student*
* Diferencias estadísticamente significativas ($p<0.05$)

Un análisis equivalente, en las muestras de suero (**Tabla 9**), permitió identificar señales en el espectro que presentaban variaciones estadísticamente significativas entre las distintas condiciones evaluadas y que no habían sido identificadas en el análisis en las muestras de plasma debido a que, en este tipo de muestras, existen regiones del espectro que no pueden ser analizadas por RMN al encontrarse solapadas con las señales de los anticoagulantes. Además de los cambios observados en plasma, se observaron alteraciones en otros metabolitos como: creatina, colina, glicerol, distintos aminoácidos, N-óxido trimetilamina, N-acetíl-cisteína y algunas señales correspondientes a grupos lipídicos. Las señales de algunos de estos metabolitos (colina, glicerol, y algún aminoácido) solapan en el espectro con las señales de EDTA, lo que impide su correcta integración y análisis cuando se trabaja con este tipo de muestras (tubos con EDTA).

Tabla 9: Coeficiente de variación de los metabolitos más relevantes en la comparación entre la condición control (30 minutos) y las muestras procesadas tras 24 horas a TA y 4ºC, respectivamente, en muestras de suero.

		4ºC		TA	
Metabolito	**ppm**	**Variación**	**p-valor [a]**	**Variación**	**p-valor [a]**
Ácido adípico	1.61-1.50	3,89 ↑	0.002*	-34.39↓	0.000*
Alanina	1.48-1.45	2,13 ↑	0.055	29.91↑	0.000*
Colina	3.21-3.16	2.12 ↑	0.055	-8.98↓	0.000*
Creatina	4.04-4.02	13.10 ↑	0.000*	13.62 ↑	0.000*
Glicerol	3.57-3.55	2.06↑	0.059	7.94 ↑	0.000*
Glucosa	5.23-5.20	-12.40 ↓	0.000*	-36.94 ↓	0.000*
Glutamato	2.35-2.28	8.64 ↑	0.000*	43.55 ↑	0.000*
H α/β aminoácidos	3.97-3.95	1.12 ↑	0.295	5.11 ↑	0.000*
Isoleucina	0.93-0.90	6.21 ↑	0.000*	5.82 ↑	0.000*
Lactato	4.12-4.06	4.56 ↑	0.000*	102.51 ↑	0.000*
Leucina	0.96-0.94	12.34 ↑	0.000*	26.60 ↑	0.000*
VLDL/HDL	0.90-0.76	8.98 ↑	0.000*	11.06 ↑	0.000*
VLDL/LDL	1.30-1.19	2.97 ↑	0.011*	-28.47 ↓	0.000*
Grupos CH_2-CH=CH lípidos	2.01-1.93	5.09 ↑	0.000*	-8.15 ↓	0.000*
Grupos CH_2-CO lípidos	2.24-2.18	3.06 ↑	0.009*	-37.97 ↓	0.000*
N-acetíl-cisteína	2.05-2.01	3.95 ↑	0.001*	5.14 ↑	0.000*
N-óxido trimetilamina	3.24-3.24	-0.84 ↓	0.426	7.21 ↑	0.000*
Piruvato	2.36-2.35	-33.57 ↓	0.000*	25.43 ↑	0.000*
Treonina	4.28-4.21	6.06 ↑	0.000*	-4.17 ↓	0.000*
Valina	0.97-0.96	8.08 ↑	0.000*	18.76 ↑	0.000*

[a] p valor calculado mediante la prueba t de *Student*
* Diferencias estadísticamente significativas (p<0.05)

Como era de esperar, tanto en las muestras de plasma como en las de suero la variación en los niveles de los metabolitos identificados fueron mayores en las

muestras procesadas a TA frente a las que se mantuvieron a 4ºC. La alteración de los lípidos fue más rápida y con una variación más acusada a TA. Además se observaron más tipos de lípidos alterados y con una variación más acusada respecto a la condición control en TA.

En general, los metabolitos que más afectados se vieron por el tiempo de procesado fueron el lactato, glucosa, glutamato y piruvato, probablemente siendo reflejo de cambios en una misma vía metabólica; la de la glucólisis y la fermentación láctica (Baynes & Dominiczak, 2010). Se observó un aumento progresivo en los niveles de lactato y una disminución en los niveles de glucosa en función del tiempo transcurrido desde que se recoge la muestra hasta su procesado. Estos cambios podrían ser debidos al prolongado contacto de la sangre con los eritrocitos, los cuales al carecer de mitocondrias obtienen la energía a través de la fermentación láctica, mecanismo por el cual, se consume glucosa para la generar lactato (Trezzi *et al*, 2016).

Por otro lado, los niveles de piruvato en los distintos tiempos analizados, presentan variaciones en distinto sentido en función de la temperatura de procesado de las muestras, ya que a temperatura ambiente su concentración aumenta, mientras que a 4ºC disminuye. Estos resultados coinciden con los reportados por Bernini en un estudio previo (Bernini *et al*, 2011) y hasta la fecha no se ha encontrado una razón justificada.

En la **Figura 18** se muestra cómo los metabolitos analizados en la muestra de plasma en función del tiempo de procesado varían en la misma dirección en ambas temperaturas, siendo el cambio más acusado a TA. Se aprecia también cómo la dirección de la variación en los niveles de piruvato no coincide en ambas temperaturas.

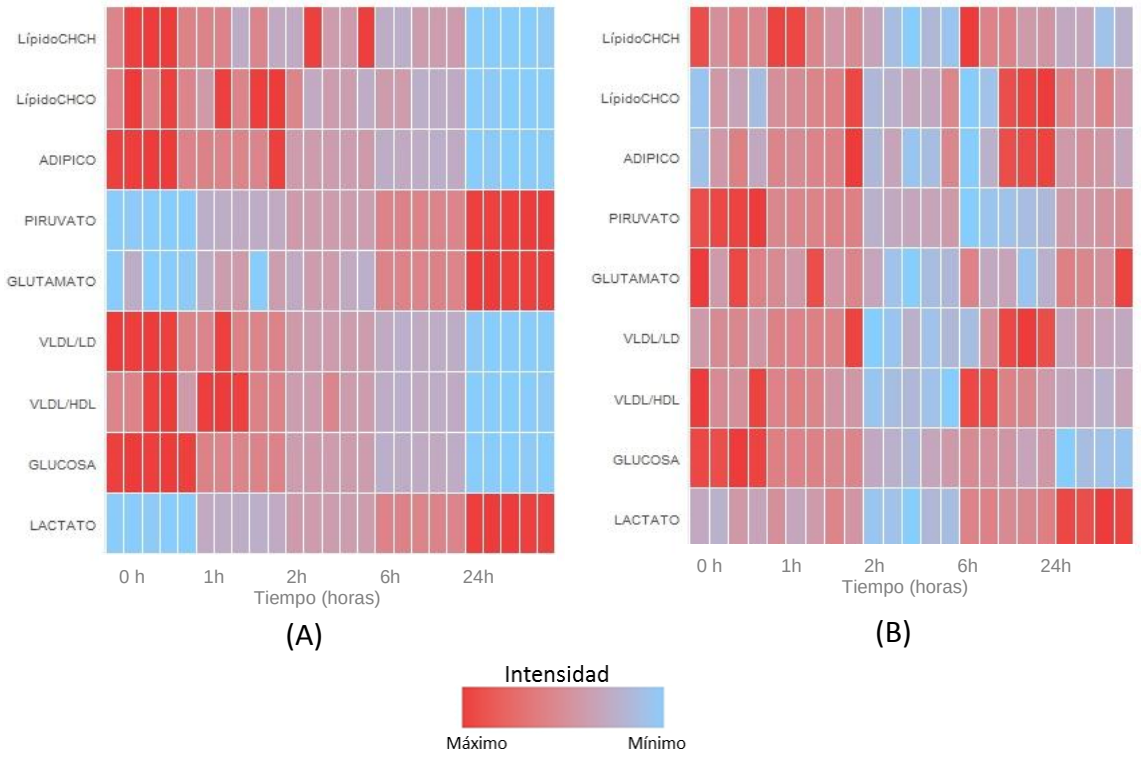


Figura 18: Heatmap correspondiente a los metabolitos alterados en las muestras de plasma a TA (A) y a 4ºC (B) en función del tiempo transcurrido hasta el procesado de las muestras.

De manera global, se observó que las muestras presentaban más alteración o degradación cuando se conservan a TA hasta su procesado, siendo esta alteración proporcional al tiempo de retraso en el procesado. Este efecto es más acusado en las muestras de plasma que en las muestras de suero. A partir de una hora de espera entre la recogida de la muestra y su procesado, a TA, se ven afectados aproximadamente el 50% de los metabolitos identificados en el estudio, mientras que a 4ºC el perfil metabolómico permanece estable durante más tiempo, hasta las 6 horas.

Jain y colaboradores (Jain *et al*, 2017) realizaron un estudio donde analizaron el efecto del tiempo que transcurre desde la extracción al procesado de las muestras de plasma y observaron que el 18,5% de los metabolitos analizados experimentaba un cambio significativo consecuencia del tiempo de retraso en el procesado (0h-20h). Los metabolitos alterados estaban relacionados con el metabolismo de los eritrocitos, ya que estos representan el 40-45% del volumen total de la sangre. Estos metabolitos podrían ser usados como biomarcadores para la detección de variaciones preanalíticas en los estudios de metabolómica. Entre los candidatos propuestos en este estudio como posibles biomarcadores destacan la 5-oxoprolina, el lactato, y la proporción de ornitina/arginina (Jain *et al*, 2017).

Otro estudio realizado previamente por Shurubor y colaboradores (Shurubor *et al*, 2007) concluyó que al menos entre 41 y 67% de los metabolitos analizados permanecieron estables tras dejar la sangre 48 horas a TA hasta su procesado. Estos resultados no concuerdan con los nuestros, ni con otras publicaciones, como las de Yin y colaboradores (Yin *et al*, 2013). Este último, en su estudio mediante CL-EM, concluye que a las dos horas de estar la sangre a TA ya se observan alteraciones en metabolitos como la hipoxantina o la carnitina (Yin *et al*, 2013). Otro estudio realizado por RMN concluye que, al igual que se observa en nuestro análisis, al permanecer las muestras 24 horas a TA hasta su procesado, la concentración de metabolitos como la glucosa, el lactato o el piruvato no se mantiene estable (Bernini *et al*, 2011).

1.3 Efecto de la hemólisis

La hemólisis se produce por una rotura de la membrana de los hematíes, liberando su contenido al exterior al medio. Las condiciones en la que se conservan las muestras durante la fase preanalítica pueden provocar la hemólisis y alterar la concentración de los metabolitos presentes en la muestras. En esta parte del estudio se extrajo sangre a dos voluntarios sanos y se provocó la hemolisis de las muestras *in vitro,* obteniéndose 5 alícuotas sin presencia de hemólisis, 5 con hemólisis moderada y 5 con hemólisis intensa.

El análisis no supervisado se realizó para explorar las muestras incluidas en esta condición del estudio. A través del análisis del PCA (**Figura 19)** se exploró el comportamiento de las muestras pertenecientes a los individuos incluidos en el estudio y se analizó el efecto de los distintos grados de hemólisis en ellos.

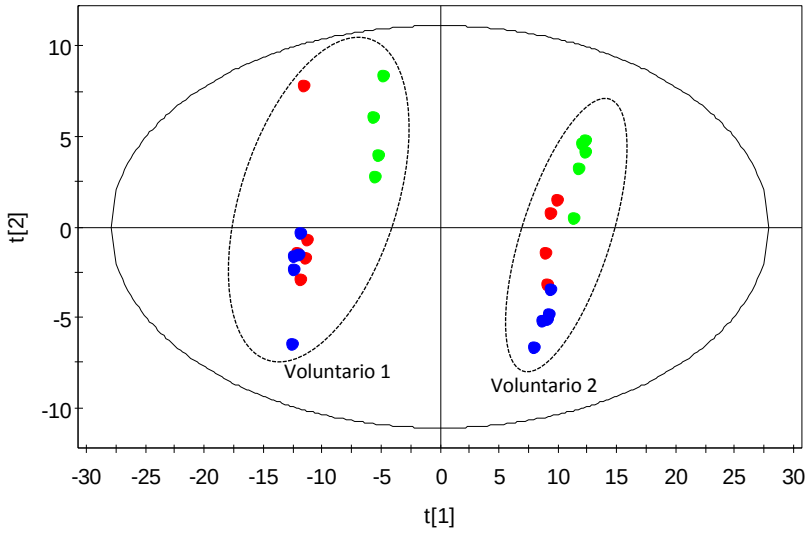


Figura 19: *Score plot* del análisis de componentes principales (PCA) de los perfiles metabolómicos de las muestras de plasma incluidas en el estudio de hemólisis (●: sin hemólisis;●: hemólisis moderada; ●: hemolisis intensa).

Por un lado, se observó que el grado de hemólisis de las muestras tenía un efecto gradual en el perfil metabolómico de las muestras. Por otro lado, la afectación de las muestras fue diferente en ambos voluntarios debido a la variación intraindividual observada en el perfil metabolómico propio de cada voluntario. En el PCA se observó como la hemólisis intensa afecta de forma más acusada en uno de los individuos. En relación a lo observado en el PCA, al analizar visualmente las señales del espectro a las que se debían las diferencias observadas entre las distintas condiciones analizadas, y los distintos pacientes, los cambios más relevantes que se observaron fueron (**Figura 20**): **i)** las diferencias entre el perfil metabolómico de ambos voluntarios se debían principalmente a señales correspondiente a lípidos (2.00 y 2.22 ppm) y, **ii)** los espectros de las muestras con hemólisis intensa, en referencia al espectro de la muestras sin hemólisis para cada voluntario incluido en el estudio, presenta una disminución de los niveles de los metabolitos que resuenan en la zona de 2.30-2.35 y 2.41-2.45 ppm (glutamato, glutamina).

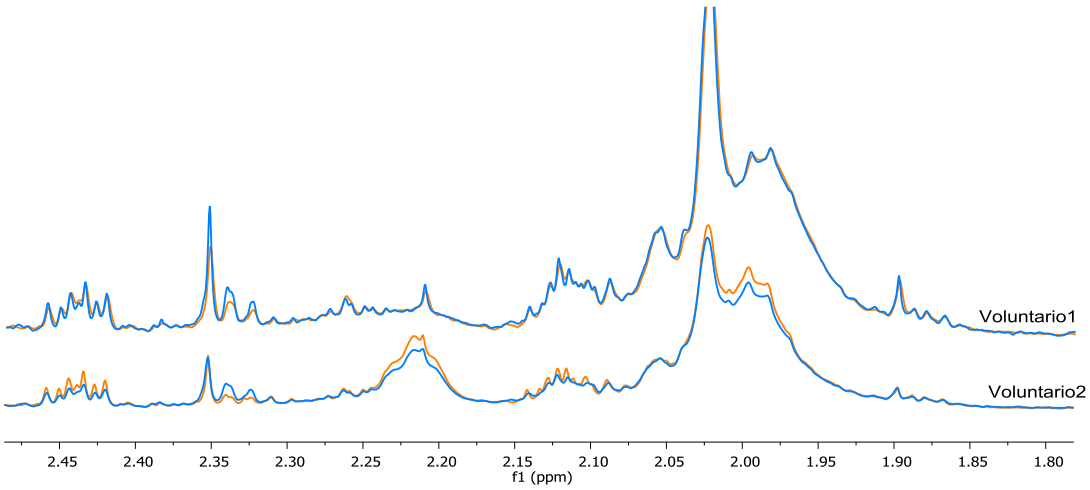


Figura 20: Perfil metabolómico de las muestras de los dos voluntarios incluidos en el estudio de hemólisis: hemólisis intensa (naranja); y sin hemólisis (azul).

En el análisis supervisado se construyó un modelo Y-OPLS-DA utilizando el grado de hemólisis como variable continua con el fin de estudiar las diferencias existentes entre los perfiles metabolómicos de las muestras. Los valores obtenidos para el ajuste y la capacidad de predicción del modelo OPLS-DA fueron $R^2(Y)$= 0,991 y $Q^2(Y)$= 0,954. Estos valores indican que existe una buena separación entre los tres grupos de muestras estudiados, es decir, que existen diferencias significativas en los perfiles metabolómicos de las muestras en función de su grado de hemólisis (**Figura 21**).

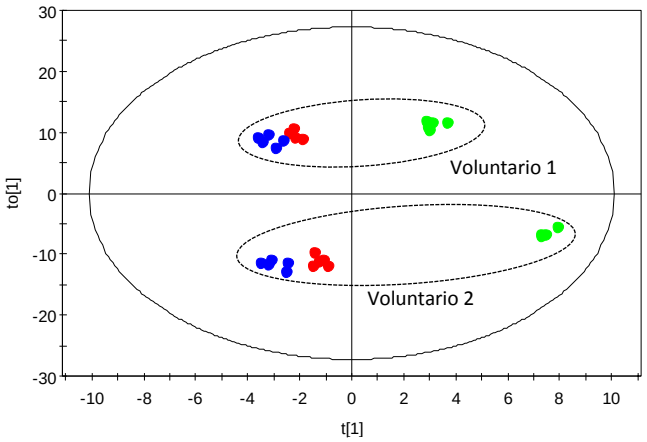


Figura 21: *Score plot* correspondiente al modelo OPLS-DA resultante de la comparación de los perfiles metabolómicos de los espectros de la condición de hemólisis (●: sin hemólisis; ●: hemólisis moderada; ●: hemolisis intensa).

A partir del modelo OPLS-DA se identificaron cuáles eran las variables que mejor caracterizaban a cada uno de los grupos estudiados y se integraron en los espectros las regiones que más contribuían a las diferencias observadas según el grado de hemólisis. En el análisis se incluyeron tanto las variables significativas comunes en ambos individuos, como las variables significativas individuales. Una vez identificadas y asignadas las variables con mayor variación se realizó el análisis estadístico univariante.

En el análisis estadístico univariante se estudiaron los cambios en los niveles de concentración de un total de 43 metabolitos. Entre ellos, se observaron cambios en los niveles de creatinina, distintos aminoácidos (tirosina, histidina, valina, glutamato, glutamina, etc) y señales correspondientes a algunos lípidos. El porcentaje de metabolitos alterados fue mayor en el grupo de muestras con hemólisis intensa que en las que presentaban un grado moderado de hemólisis respecto al control, como era de esperar (**Figura 22**). Las muestras con hemólisis moderada presentaban más similitud con las muestra control que con las muestras con hemólisis intensa. En ambos pacientes, se observó la misma tendencia en los cambios de los metabolitos por lo que en el análisis estadístico univariante se analizaron todas las muestras de manera conjunta atendiendo a su clasificación según el grado de hemólisis, e independientemente del voluntario del que provenían las muestras.

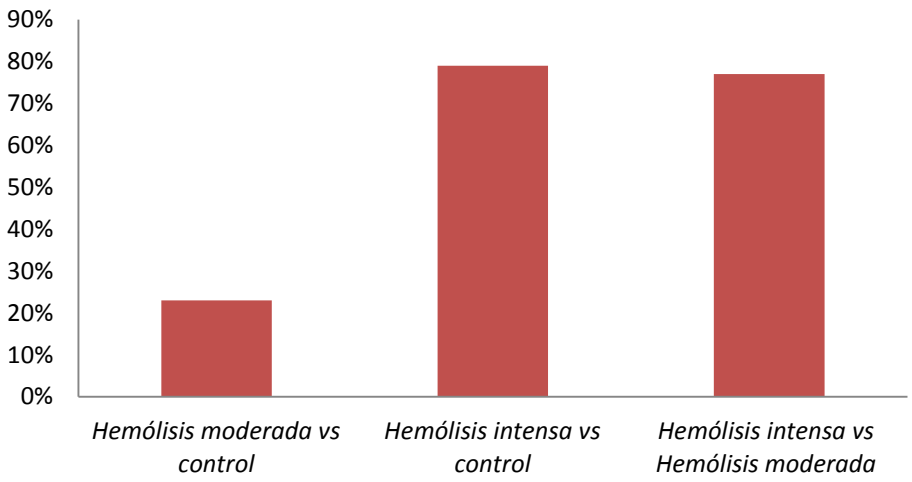


Figura 22: Porcentaje de metabolitos alterados de forma significativa (p valor < 0.05, test t de *Student*) en función del grado de hemólisis.

De manera global, la composición metabólica de las muestras con hemólisis intensa resultó muy alterada con respecto a las muestras control. Un porcentaje elevado de metabolitos presentaron niveles distintos a los controles (**Figura 23**).

La variación en la concentración de los metabolitos en las muestras con hemólisis podría ser debido al efecto de dilución que sufren los metabolitos presentes en el plasma tras la hemólisis y/o a la actividad de las enzimas glicolíticas de los eritrocitos con los metabolitos presentes en el plasma. Por otro lado, en el caso de metabolitos que se encuentran en mayor concentración en el interior del eritrocito, cuando éste se rompe y libera su contenido al exterior, la concentración de esos metabolitos se verá aumentada.

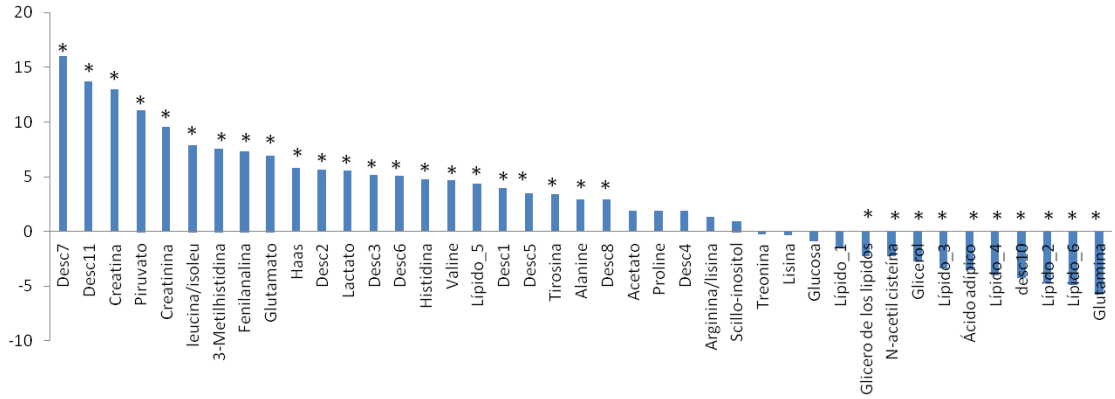


Figura 23: Metabolitos alterados en muestras de sangre con hemólisis intensa con respecto a las muestras control. (*p-valor < 0.05, *t Student* test).

Yin y colaboradores (Yin *et al*, 2013) empleando CL-EM, observaron alteración en la intensidad de los metabolitos debido a la hemólisis, un total de 69 metabolitos presentaron cambios significativos por el efecto de la hemólisis (Yin *et al*, 2013). Kamlage y colaboradores (Kamlage *et al*, 2014) también observaron alteraciones

significativas en un 18% y 30% de los metabolitos en muestras con hemólisis moderada e intensa, respectivamente, en un estudio realizado por EM (Kamlage *et al*, 2014).

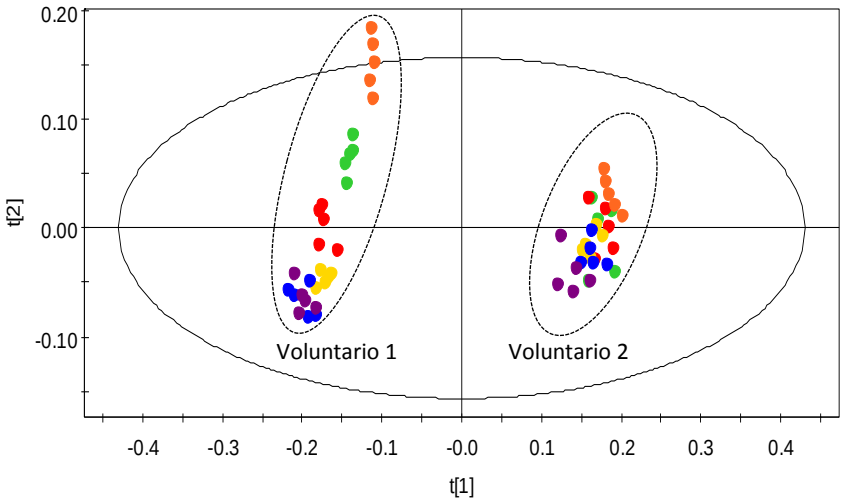
La hemólisis es una interferencia frecuente que afecta a los resultados de los laboratorios clínicos y su prevalencia puede llegar a ser del 3,3% en el total de muestras que se analizan de rutina en un laboratorio, siendo esta interferencia más común que otras interferencias como son muestra insuficiente, muestra coagulada, error de identificación, etc (Lippi *et al*, 2008). Tal y como se ha mostrado, la presencia de hemólisis en las muestras destinadas a estudios de metabolómica por RMN, tiene un gran impacto sobre el perfil metabólico de las muestras alterando al 70% de los metabolitos (**Figura 22**).

1.4 Ciclos de congelación y descongelación

Las muestras que se utilizan para investigación suelen permanecer almacenadas a -80ºC. Generalmente, en función de las necesidades o de las características del estudio, es necesario congelarlas y descongelarlas en distintos momentos para realizar diferentes análisis o experimentos sobre la misma muestra. En este estudio se evaluó la posible degradación de las muestras en función de los ciclos de congelación y descongelación a los que habían sido sometidas.

En el análisis exploratorio se estudió el comportamiento de las muestras a partir de modelos de PCA (**Figura 24**). En este análisis se apreció una distribución de las muestras en función del número de ciclos de congelación y descongelación a los que habían sido sometidas. Este comportamiento se observó en los dos biofluidos estudiados: plasma y suero. Además, como se observa en la **Figura 24**, se observó que la afectación de las muestras fue diferente en ambos voluntarios debido a la variación interindividual observada en el perfil metabolómico propio de cada voluntario.

(A)



(B)

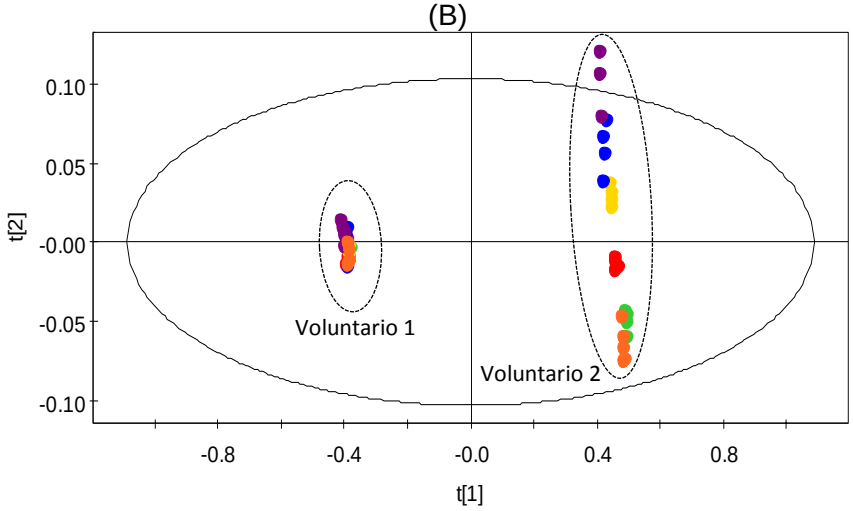


Figura 24: *Score plot* correspondiente al modelo PCA de los perfiles metabolómicos de los espectros de las muestras de plasma (A) y suero (B) incluidos en el estudio del efecto de los ciclos de congelación-descongelación (●control, ●1ciclo, ●2 ciclos, ●3 ciclos, ●4 ciclos, ● 5 ciclos).

En el análisis supervisado, tanto para plasma como para suero, se construyó un modelo Y-OPLS-DA. El número de ciclos de congelación y descongelación se utilizó como variable continua, con el fin de estudiar las diferencias existentes entre los perfiles metabólicos de las muestras según los ciclos a los que habían sido sometidas (plasma: R^2 (Y) =0.924, Q^2 (Y) =0.896, suero: R^2 (Y) =0.981, Q^2 (Y) =0.916) (**Figura 25**).

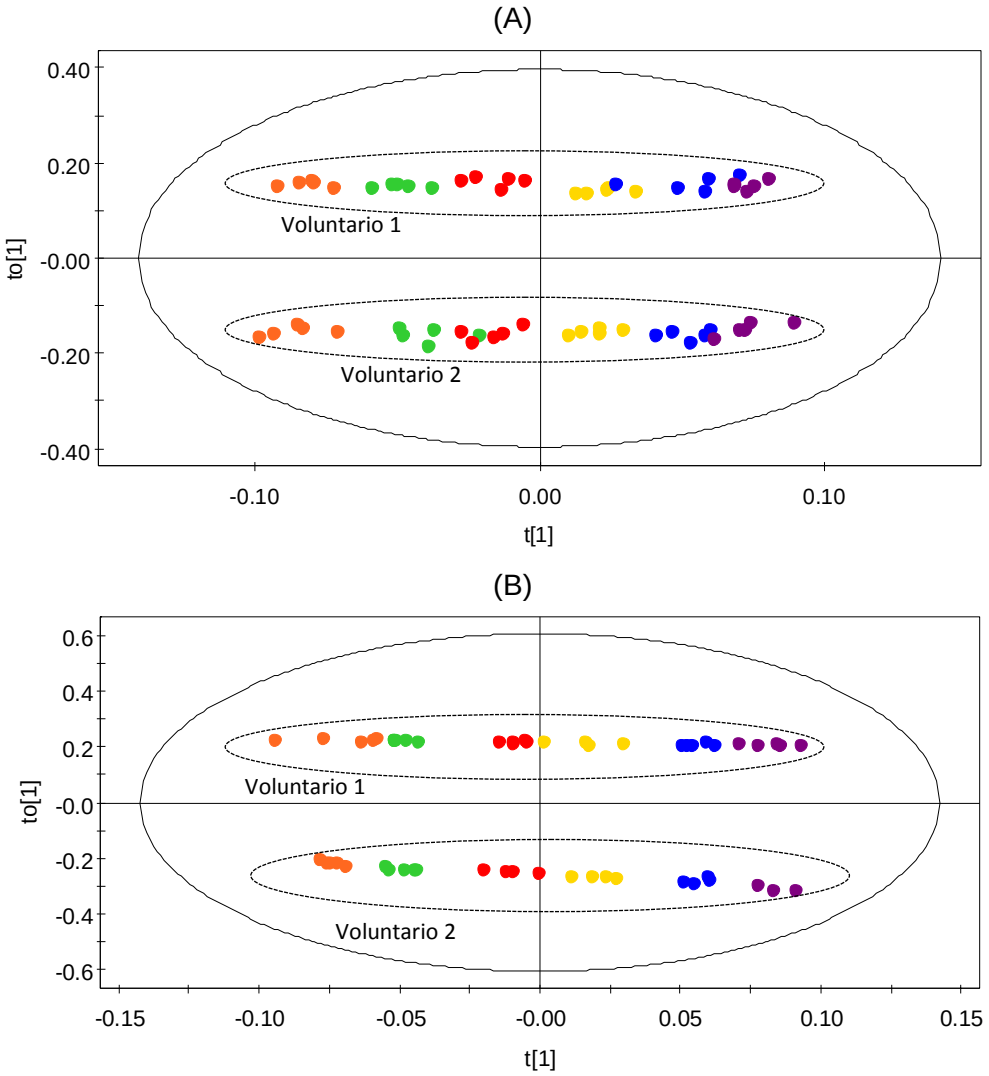


Figura 25: *Score plot* correspondiente al modelo Y-OPLS-DA resultante de la comparación de los perfiles metabolómicos de los espectros de las muestras de plasma (A) y suero (B) en función del número de ciclos de congelación-descongelación (●control, ●1ciclo, ●2 ciclos, ●3 ciclos, ●4 ciclos, ● 5 ciclos).

A partir de los modelos OPLS se analizaron e identificaron aquellas regiones del espectro que más contribuían a las diferencias observadas entre los distintos grupos. En base a estos resultados, y teniendo en cuenta aquellas señales que presentaban un valor de VIP superior a uno, se procedió a la integración de las señales correspondientes en los espectros ^{1}H-RMN CPMG de las muestras incluidas del estudio. Debido a la variabilidad interindividual observada en los modelos, y con la finalidad de analizar cualquier variación relacionada con el número de ciclos de congelación-descongelación, independientemente del origen de la muestra, la selección de las variables con VIP > 1 se realizó tanto de los modelos OPLS de cada individuo como del OPLS que incluía las muestras de ambos individuos. Se identificaron un total de 22 metabolitos en plasma y 35 metabolitos en suero. La

diferencia en el número de metabolitos identificados en plasma y en suero es debida a las señales de EDTA presentes en los perfiles metabolómicos de las muestras de plasma, que se encuentran solapadas con las señales de otros metabolitos. Entre los metabolitos que no se pudieron identificar en las muestras de plasma se encuentran la colina, el piruvato, el citrato y el glicerol. En la **Figura 26** se observa cómo, a pesar de que el plasma se degrada en mayor porcentaje en el ciclo 1, el suero se degrada de manera más intensa en los primeros ciclos. En ambos biofluidos a partir del cuarto ciclo de congelación-descongelación la degradación es más acusada, alterándose más del 50% de los metabolitos estudiados.

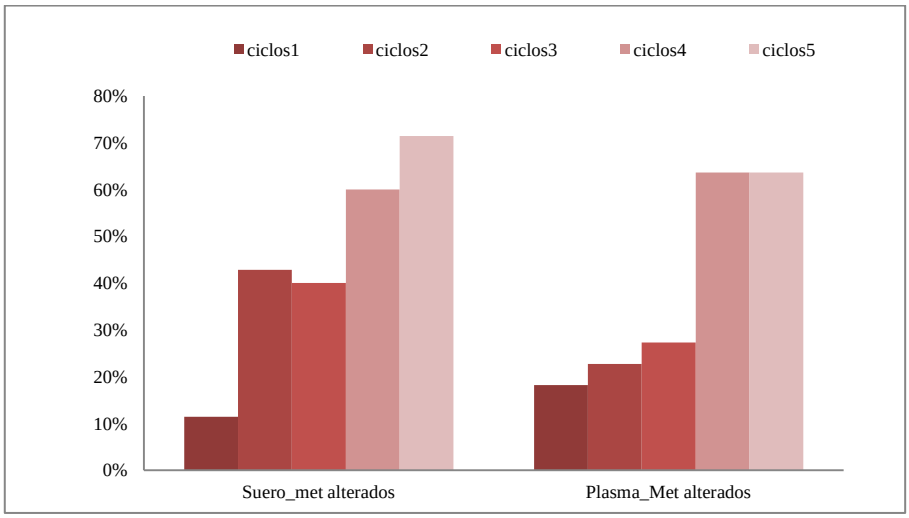


Figura 26: Porcentaje de metabolitos alterados de forma significativa (p valor < 0.05, test t de *Student*) tras los diferentes ciclos de congelación-descongelación.

Los metabolitos más alterados en ambos fluidos fueron los lípidos, aminoácidos, glucosa y creatina (**Tabla 10**). Estas alteraciones concuerdan con los resultados encontrados en otros estudios en los que se estudiaron los cambios en la composición del plasma tras varios ciclos de congelación-descongelación. Pinto y colaboradores (Pinto *et al*, 2014) observaron cambios en la composición de las muestras de plasma, en las que se vieron alterados los niveles de lípidos, colina, valina, alanina y lactato (Pinto *et al*, 2014). En otro estudio realizado en suero se observaron alteraciones en determinados metabolitos cuando se realizaban varios ciclos de congelación-descongelación, disminuyendo metabolitos como la prolina, el glicerol, el metanol o la colina (Fliniaux *et al*, 2011).

Tabla 10: Metabolitos alterados de forma significativa a partir de 4º ciclo de congelación-descongelación en comparación con el 1º ciclo (p-valor < 0.05, *t Student* test) en plasma y suero.

Biofluido	Metabolito	Variación
PLASMA	Ácido adípico	-6.29 ↓
	Lípidos CH_2CO	-6.51 ↓
	VLDL-LDL	-2.95 ↓
	Glutamato	4.47 ↑
	Valina	3.61 ↑
	Leucina	3.36 ↑
	Isoleucina	4.50 ↑
	Lípidos CH=CH	3.65 ↑
	Glutamina	3.37 ↑
	Glucosa	3.02 ↑
	Creatina	3.78 ↑
	Lisina	3.50 ↑
SUERO	Ácido adípico	-6.60 ↓
	Citrato	-5.45 ↓
	Colina	-2.97 ↓
	Desconocido 12	-2.73 ↓
	Glicerol	-5.65 ↓
	Glutamina	-4.23 ↓
	VLDL-HDL	-5.68 ↓
	VLDL-LDL	-7.08 ↓
	Lípidos CH_2-C=C	-3.74 ↓
	Lípidos CH_2CO	-6.38 ↓
	N-acetíl cisteína	-4.38 ↓
	Piruvato	6.01 ↑
	Lactato	4.37 ↑
	Acetato	2.33 ↑
	Creatina	2.91 ↑
	Glicina	3.43 ↑
	Glutamato	3.33 ↑
	Isoleucina	2.67 ↑
	Lisina	2.28 ↑
	Treonina	2.91 ↑
	Valina	3.44 ↑
	Glucosa	5.50 ↑
	Leucina	9.63 ↑
	N-óxido de trimetilamina	5.17 ↑

1.5 Tiempo de almacenamiento

Las muestras que se utilizan para investigación suelen permanecer almacenadas a -80ºC. El tiempo que permanecen almacenadas varía en función del tipo de estudio. La estabilidad de las muestras puede perderse a lo largo del tiempo. En este sentido, en este estudio se compara la composición de las muestras que han permanecido almacenadas a -80ºC después de 1 mes, 6 meses, 1 año y 2 años tras su obtención.

El análisis no supervisado se realizó para explorar el comportamiento de las muestras incluidas en esta condición del estudio. A partir de los modelos de PCA generados se observó si existía una agrupación de las muestras en función del tiempo durante el que han permanecido almacenadas (**Figura 27**). Este análisis se realizó tanto en las muestras de plasma como en las de suero. En ambos biofluidos se observó un cambio importante a los dos años de almacenamiento. En las muestras de suero del voluntario 2, las muestras de un año de almacenamiento muestran una variabilidad no explicada por el experimento ni por cambios relacionados con el tiempo de almacenamiento, seguramente debido a alguna modificación en la recogida y procesado de esas muestras.

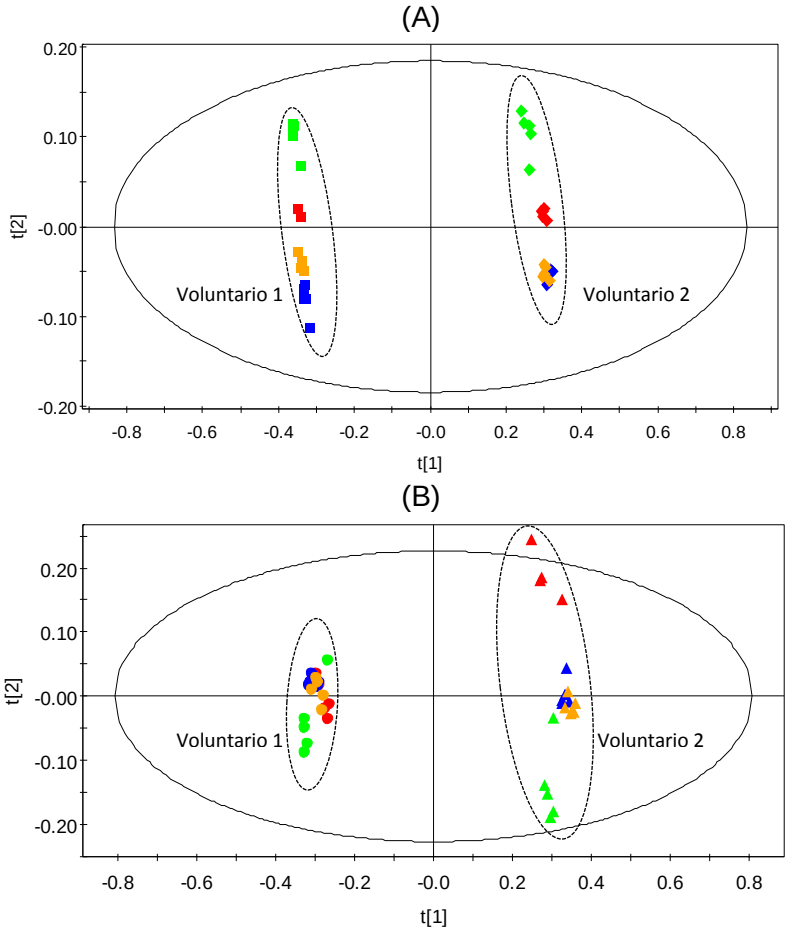


Figura 27: *Score plot* correspondiente al modelo PCA de los perfiles metabolómicos de los espectros le las muestras de plasma (A) y suero (B) de las muestras incluidas en el estudio del efecto del tiempo de almacenamiento a -80ºC. (●control, ● 6 meses, ● 1 año, ● 2 años).

Tras el análisis exploratorio en ambos biofluidos, se realizó el análisis supervisado. Se construyeron modelos OPLS-Y (Plasma: $R^2(Y)$ =0,982 y $Q^2(Y)$ =0,95; Suero: $R^2(Y)$ =0,887 y $Q^2(Y)$ =0,844) utilizando como variable continua el tiempo de almacenamiento y se comparó la estabilidad de esas muestras en función del tiempo almacenado (**Figura 28**). A partir del modelo OPLS, se analizaron e identificaron aquellas regiones de espectro que más contribuían a las diferencias observadas entre los diferentes

grupos. En base a estos resultados, y teniendo en cuenta aquellas señales que presentaban un valor de VIP superior a uno, se obtuvo un listado de las variables más representativas y se procedió a la integración de esas señales en los espectros 1D-CPMG de las muestras incluidas del estudio. Tal y como se ha explicado en análisis anteriores, se tuvo en cuenta para el análisis tanto los modelo OPLS de los voluntarios de forma individual como el del modelo OPLS de ambos voluntarios.

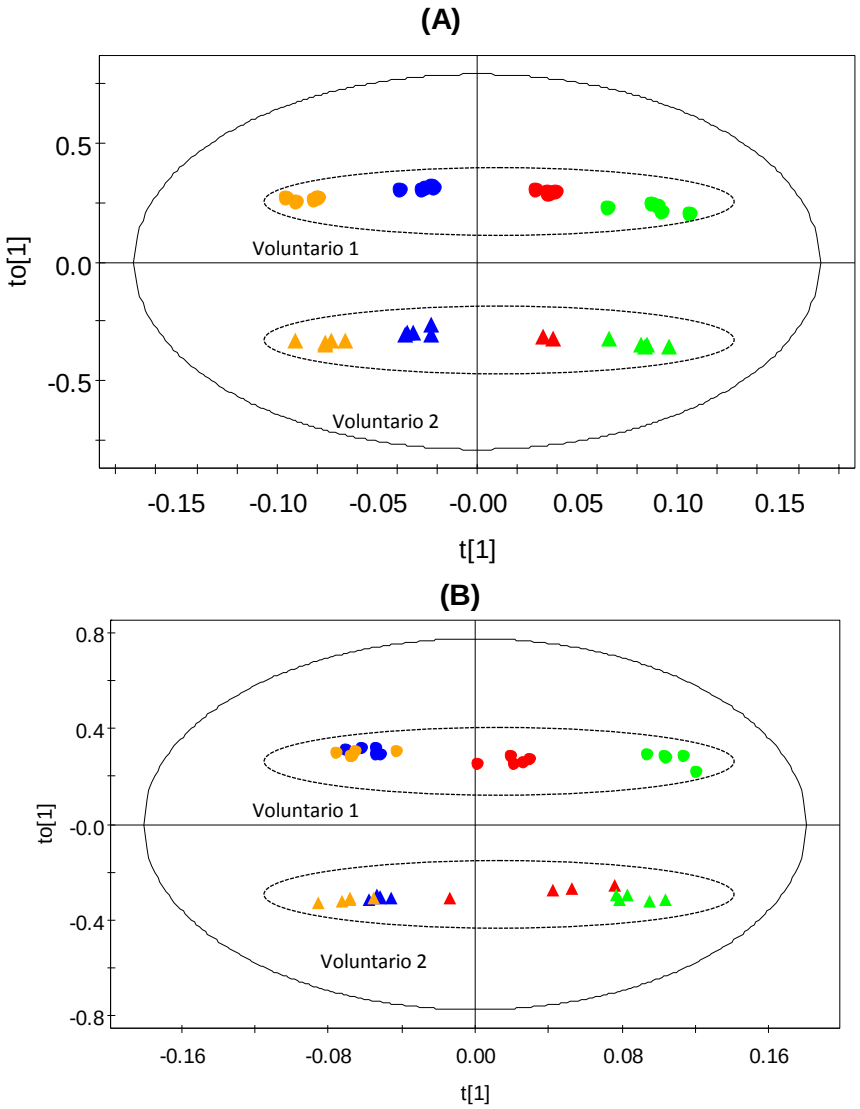


Figura 28: *Score plot* correspondiente al modelo OPLS-Y-DA resultante de la comparación de los perfiles metabolómicos de los espectros de las muestras de plasma (A) y suero (B) de la condición de almacenamiento (●control, ● 6 meses, ● 1 año, ● 2 años).

El análisis univariante reveló que en los primeros 6 meses de almacenamiento apenas se producía alteración en la composición metabólica de las muestras con respecto al control. En cambio, a los dos años se observó la concentración alterada en casi el 50% los metabolitos analizados debido a la pérdida de la estabilidad en la composición metabólica de las muestras (**Figura 29**). Se observa cómo la degradación

en la composición metabólica de las muestras tras dos años de almacenamiento afecta de manera similar al suero y al plasma.

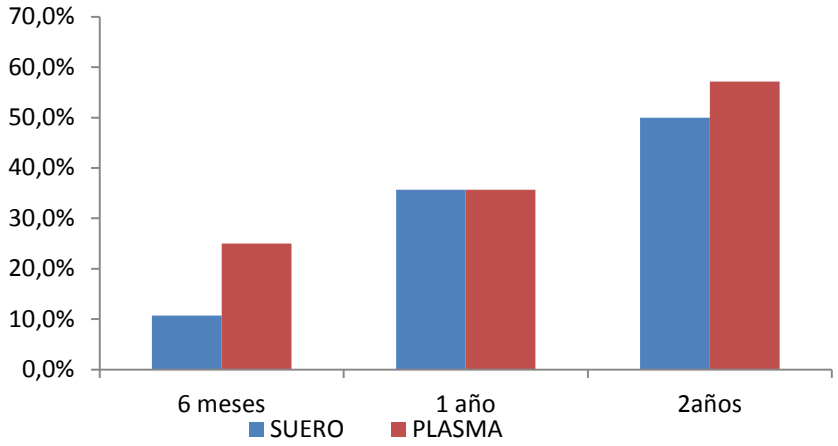


Figura 29: Porcentaje de metabolitos alterados de manera significativa (p valor < 0.05, test t de *Student*) en muestras de plasma y suero tras permanecer 6 meses, 1 años o 2 años almacenados hasta su análisis.

Los resultados muestran cómo, a pesar de que la magnitud de los cambios es similar en ambos biofluidos, los metabolitos afectados de forma significativa tras permanecer las muestras dos años almacenadas hasta su análisis por RMN fueron ligeramente distintos en los dos biofluidos. En las muestras de plasma los metabolitos más afectados fueron los lípidos, la creatinina y aminoácidos como treonina, tirosina e histidina. Mientras que en suero, aparte de algunos lípidos, los metabolitos más afectados fueron el citrato y aminoácidos como glutamato, valina o lisina (**Tabla 11)**. Estos resultados concuerdan con los encontrados por Pinto y colaboradores (Pinto *et al*, 2014) en un estudio de metabolómica por RMN, donde tras los 20-30 meses de almacenamiento a -80ºC observaron cambios en el colesterol, N-acetil-glicoproteínas y creatinina (Pinto *et al*, 2014).

Tabla 11: Metabolitos alterados en muestras de suero y plasma tras permanecer dos años almacenadas hasta su análisis.

Control *vs*. 2 años (SUERO)				Control *vs*. 2 años (PLASMA)			
ID	**ppm**	**Variación**	**p-valor** [a]	**ID**	**ppm**	**Variación**	**p-valor** [a]
Ácido adípico	1.60-1.52	-1,84 ↓	0.130	**Ácido adípico**	1.52-1.60	-7.72 ↓	0.000*
Acetato	1.91-1.89	6,54 ↑	0.000*	**Acetato**	1.88-1.91	5.11 ↑	0.000*
Citrato	2.54-2.52	5,93 ↑	0.000*	**Creatinina**	4.03-4.04	-14.11 ↓	0.000*
Colina	3.21-3.16	0,08 ↑	0.934	**Desconocido 1**	3.29-3.32	3.70 ↑	0.002*
Creatinina	4.04-4.03	-2,07 ↓	0.086	**Desconocido 2**	3.32-3.34	2.61 ↑	0.025*
Desconocido 3	3.29-3.32	5,57 ↑	0.000*	**Glicerol**	3.62-3.65	1.88 ↑	0.106
Desconocido 4	3.37-3.36	2,53 ↑	0.032*	**Glucosa**	3.47-3.48	3.91 ↑	0.001*
Desconocido 2	2.92-2.86	-1,44 ↓	0.606	**Glutamato**	2.05-2.08	-1.51 ↓	0.173
Desconocido 5	7.68-7.61	-0,62 ↓	0.000*	**Glutamina**	2.08-2.15	0.39 ↑	0.698
Glicerol	3.65-3.62	-7,10 ↑	0.017*	**H α/β de aa**	3.94-4.00	-2.88 ↓	0.016*
Glucosa	3.48-3.47	1.61 ↑	0.179	**Hidroxibutirato**	1.14-1.18	-3.86 ↓	0.001*
Glutamato	2.35-2.30	4,44↑	0.000*	**Histidina**	7.74-7.79	-6.23 ↓	0.000*
Glutamina	2.15-2.08	1,64↑	0.179	**Lactato**	1.30-1.32	-0.63 ↓	0.566
H α/β de aa	4.00-3.94	4,97↑	0.000*	**LDL.VLDL**	1.18-1.30	-2.68↓	0.024*
Histidina	7.79-7.74	-5,77↓	0.000*	**Lípido CH_2CC**	1.91-2.02	-1,77↓	0.127
Lactato	1.32-1.30	0.71 ↑	0.558	**Lípido CH_2CO**	2.25-2.17	-8,95↓	0.000*
LDL.VLDL	1.30-1.18	-3,54↓	0.003*	**Lípido CH-CH**	5.36-5.23	-15,86↓	0.000*
Lípido CH_2CO	2.25-2.17	-1,27↓	0.295	**Glicerol**	5.20-5.13	-7,31↓	0.000*
Lípido CHCH	5.36-5.23	-5,74↓	0.000*	**Lisina**	1.62-1.75	1,19↑	0.281
Lisina	1.75-1.62	3,77↑	0.002*	**Metanol**	3.35-3.34	0,61↑	0.566
Tirosina	6.90-6.87	-0,51↓	0.660	**N-Acetil**	2.01-2.04	-1,16↓	0.286
Treonina	4.34-4.20	1,24↑	0.296	**Tirosina**	6.87-6.90	-2,66↓	0.024*
Valina	1.04-1.01	3,93↑	0.001*	**Treonina**	4.34-4.20	-4,75↓	0.000*
VLDL.HDL	0.90-0.76	-0,87↓	0.476	**VLDL.HDL**	0.76-0.90	-1,94↓	0.099

[a] p valor calculado mediante la prueba t de *Student*
* Diferencias estadísticamente significativas (p<0.05)

En otro estudio realizado por Hebels y colaboradores (Hebels *et al*, 2013) se evaluó, utilizando diferentes técnicas ómicas, cómo afectaba el tiempo en el que permanecían las muestras de plasma en biobancos durante casi dos décadas. A diferencia de nuestro estudio y de otros estudios, en éste no se detectaron diferencias significativas tras las evaluación de las muestras después del periodo de almacenamiento (Hebels *et al*, 2013). Esto podría deberse a que el método de análisis utilizado es distinto, ya que ellos lo analizan por UPLC-ToFMS y en nuestro estudio se analizar por RMN.

1.6 Impacto analítico

Los resultados presentados en este estudio muestran cómo distintas variaciones en los procesos incluidos en la fase preanalítica (tubos de recogida, temperatura de procesado, tiempo trascurrido hasta el procesamientos, temperatura y tiempo de almacenamiento) pueden afectar de manera muy significativa al perfil metabolómico de las muestras. Así pues, variaciones en esta fase, durante la recogida, procesado y almacenamiento de las muestras biológicas, podrían introducir desviaciones muy significativas en los datos obtenidos y, como consecuencia, en los resultados y conclusiones alcanzadas en los estudios a los que se destinan esas muestras. Generalmente, en estos procesos preanalíticos interviene personal muy diverso y existen protocolos distintos en función del laboratorio o centro de investigación en el que se obtengan las muestras. Esta situación crea la necesidad de implantar procedimientos normalizados de trabajo que garanticen la calidad y estandarización de la fase preanalítica.

Finalmente, a partir del análisis de los resultados obtenidos en los distintos apartados de este estudio, podríamos destacar algunas recomendaciones a tener en cuenta en la recogida, procesado y almacenamiento de muestras de suero y plasma para su uso en estudios de metabolómica. Entre las recomendaciones se incluiría:

-Evitar la hemólisis en la medida de lo posible y rechazar aquellas muestras que lleguen al laboratorio con hemólisis intensa.

-Procesar y almacenar las muestras en el menor tiempo posible (30 minutos), manteniendo una temperatura constante de 4ºC desde la extracción de la muestra hasta su almacenamiento. En cualquier caso, nunca sobrepasar 1 hora entre la extracción y el procesado en el caso en que las muestras se mantengan a TA y no sobrepasar las 6 horas si se conservan a 4°C durante este tiempo.

-Intentar no someter las muestras a ciclos de congelación y descongelación. Sería recomendable almacenar las muestras en distintas alícuotas, de manera que sea posible descongelar únicamente el volumen necesario de muestra en cada momento. En el caso de ser necesario congelar y descongelar la muestra se recomienda nunca superar los 2 ciclos ya que, a partir del tercer ciclo, la degradación de la muestra afecta a más del 50% de los metabolitos analizados.

-La estabilidad de las muestras tras su almacenamiento sería de un año máximo, y si se desea almacenar las muestras más tiempo por requerimientos del estudio, hay que tener en cuenta que la estabilidad de metabolitos como los lípidos se puede ver alterada.

2. VARIABILIDAD BIOLÓGICA

El diagnóstico actual del CaP consiste en una primera parte de cribado, basada en el análisis de los niveles de PSA en suero y el tacto rectal, y una segunda etapa de diagnóstico donde la prueba de referencia es la biopsia prostática. Las limitaciones en el diagnóstico de esta enfermedad son, por un lado, la baja especificidad del PSA como biomarcador de la enfermedad y, por otro lado, la alta tasa de resultados falsos negativos en la biopsia prostática, siendo además una prueba invasiva.

Uno de los principales retos en el estudio del CaP es el desarrollo de métodos de diagnóstico más específicos y no invasivos. Sin embargo, ninguno de los biomarcadores propuestos hasta el momento ha permitido un avance significativo en este sentido. Este trabajo, donde se comparan pacientes con CaP frente a individuos con HBP, constituye el primer estudio dirigido a identificar posibles biomarcadores de utilidad en el diagnóstico clínico de esta patología a través de RMN y utilizando un biofluido no invasivo como la orina.

2.1 Análisis y asignación de los espectros

Las muestras de orina de los individuos con HBP y de los pacientes con CaP presentaron espectros de RMN de buena calidad. En la **Figura 30** se muestra un espectro típico de orina ^{1}H-RMN-CPMG de un paciente con CaP, indicando la asignación de las señales más representativas presentes en la muestra.

En general, los espectros de muestras de orina se caracterizan por presentar señales de metabolitos de diversa clase química, incluyendo ácidos orgánicos, aminoácidos, azúcares simples y polisacáridos. En particular, el espectro de orina está dominado por señales como la urea, la creatinina, el trimetilamina-N-óxido, la dimetilamina, el ácido hipúrico o el ácido cítrico, entre otras señales.

Una de las características de la orina es que las señales que aparecen en el espectro de RMN pueden presentar intensidades muy distintas entre ellas. El perfil metabolómico de la orina presenta picos de alta intensidad, correspondientes a metabolitos como la creatinina, y picos de muy baja intensidad, correspondientes a metabolitos como la alanina o el fumarato.

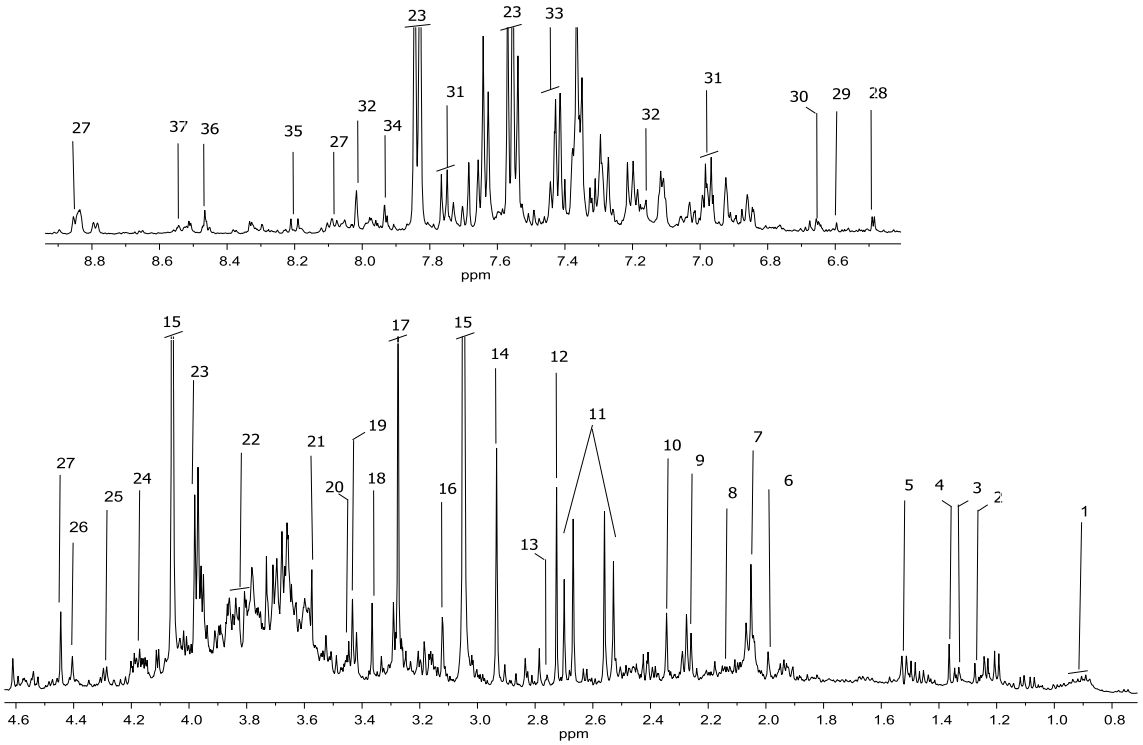


Figura 30: 1, AACR (aminoácidos de cadena ramificada); 2, 3-hidroxisovalerato; 3, Lactato; 4, 2-hidroxisobutirato; 5, Alanina; 6, Acetato; 7, grupos N-acetil; 8, Glutamato; 9, 2-hidroxi-glutarato; 10, Piruvato; 11, ácido cítrico; 12, Dimetilamina; 13, Sarcosina; 14, Dimetilglicina; 15, Creatinina; 16, Ácido cis-aconítico; 17, Trimetilamina-N-oxido; 18, Metanol; 19, Ácido trans-aconítico; 20, Taurina; 21, Glicina; 22, Serina; 23, Ácido hipúrico; 24, Pseudouridina; 25, L-treonina; 26, dihidroxiacetona; 27,Trigonelina; 28, Desconocido; 29, Fumarato; 30; 2-Furoylglicina; 31, 4-hidroxibenzoato; 32, 3-metilhistidine; 33, Fenilalanina; 34, Histidina; 35, Hipoxantina; 36, Formato; 37, 4-imidazolacetato.

La orina es un biofluido que proporciona información sobre el estado metabólico de los pacientes y tiene la ventaja de que puede obtenerse de manera no invasiva. Sin embargo, se considera un medio complejo para los estudios de metabolómica debido a la variabilidad interindividual que presenta.

Por un lado, la orina presenta cambios significativos en su pH entre individuos/días. La variación en el pH puede hacer que las señales de determinados metabolitos en los espectros tengan un desplazamiento químico que varía entre muestras de distintos pacientes/días. Por otro lado, la abundancia de metabolitos que se encuentran en la orina hace que las señales del espectro de RMN se solapen, dificultando la cuantificación de los mismos.

Además, la orina es un fluido que presenta alta variabilidad en su composición metabólica ya que depende de la dieta, del ejercicio, de la ingesta de líquidos, etc. La intensidad de las señales también presenta variabilidad entre las muestras debido a que la concentración de los metabolitos puede experimentar variaciones en función de lo diluida o no que esté la muestra.

Estos factores hacen que el tratamiento de los datos que se obtienen del análisis de los espectros del perfil metabolómico de este tipo de muestras deba ser riguroso y específico.

2.2 Procesado de los datos

En el procesado de los datos se emplearon diferentes herramientas para lograr que la información contenida en las señales del espectro pudiera analizarse de manera reproducible. Las señales de los espectros de orina presentaron variaciones significativas en sus desplazamientos químicos, dificultando la correcta integración del espectro. Un alineamiento adecuado de las señales de los espectros es necesario para que en el proceso de *bucketing* una misma señal no se integre de forma distinta en los espectros que se están comparando, dando lugar a cambios importantes en el análisis posterior de los datos. Distintos métodos de alineamiento fueron evaluados para solucionar estas variaciones.

En primer lugar, se referenciaron los espectros utilizando la señal del TSP (0 ppm). Debido al distinto desplazamiento químico en algunas señales de algunos metabolitos como consecuencia de variaciones en el pH de las muestras, no se consiguió un correcto alineamiento global de los picos. Como alternativa, se aplicó la herramienta "speaq" (Vu *et al*, 2011) con la que se obtuvo un alineamiento de los espectros que permitió la correcta integración de las señales y comparación de los espectros. La herramienta "speaq" selecciona un espectro que toma como referencia, posteriormente se alinean el resto de los espectros permitiendo alinear de manera aislada las distintas regiones del espectro, no sólo en referencia a una única señal como ocurre cuando se toma como única referencia la señal del TSP. En la **Figura 31** se puede observar cómo las señales correspondientes a los dobletes de la molécula del ácido cítrico (2.55 y 2.70 ppm), no se encuentran correctamente alineadas en los espectros originales debido a las variaciones de pH entre muestras. Este efecto fue corregido tras la aplicación del algoritmo "speaq" sobre los datos crudos obtenidos de los espectros de RMN originales. En la **Figura 31** se observa también cómo estas diferencias en el pH de las muestras tienen distinto efecto sobre las diferentes señales que aparecen en los espectros.

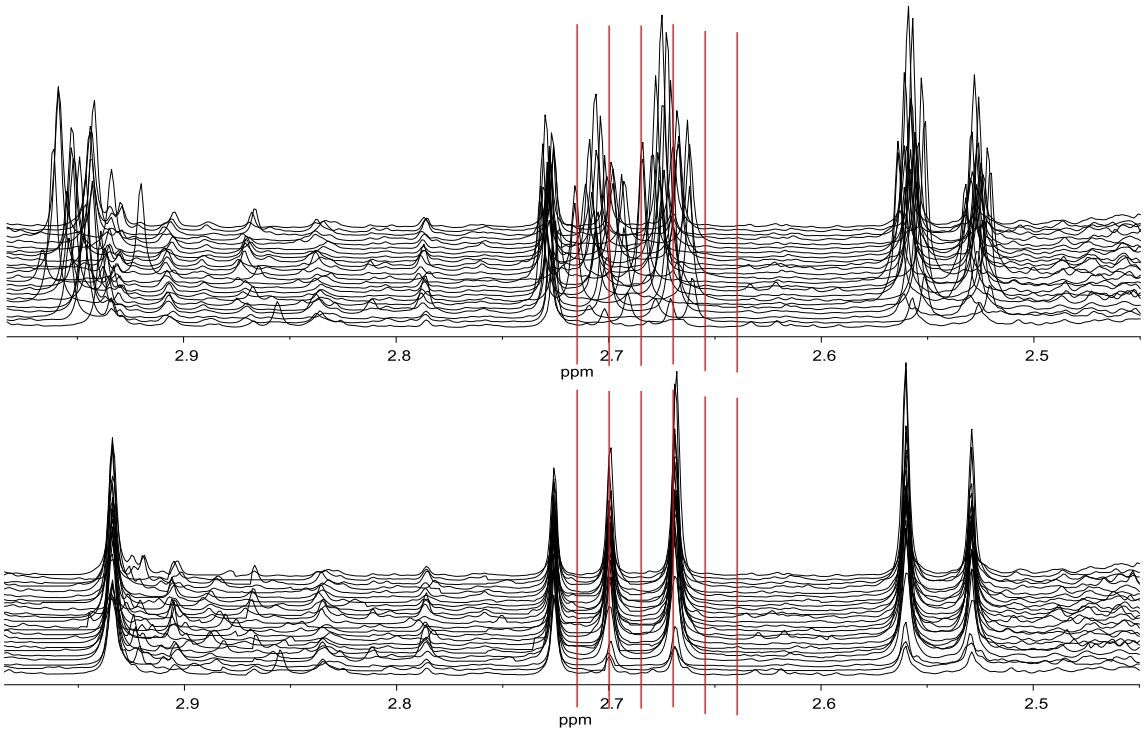


Figura 31: Alineamiento de espectros de RMN. La imagen superior representa un conjunto de espectros ^{1}H-RMN-CPMG correspondientes a muestras de orina originales. La imagen inferior representa en mismo conjunto de muestras después de aplicar "speaq" como herramienta de corrección del alineamiento.

Una vez las señales de los espectros fueron alineadas correctamente, se evaluaron distintos métodos de normalización sobre los datos. Como se ha comentado anteriormente, en las muestras de orina de pacientes, existe una variabilidad en la intensidad de las señales entre las muestras debido a las variaciones en el volumen o dilución de las mismas. El objetivo de la normalización fue conseguir datos comparables y corregir esta variabilidad. La normalización respecto al área total, muy utilizada en muestras de suero o plasma, no logró corregir completamente estas variaciones, debido a las diferencias en las intensidades de las señales que presentaban los espectros de orina. Por esta razón, se optó por normalizar los espectros utilizando el método PQN (Dieterle et al, 2006).

La normalización mediante PQN proporciona buenos resultados en aquellas muestras donde las intensidades de los metabolitos varían mucho dentro de la misma muestra y, a su vez, se comparan espectros con distintas intensidades entre ellos. Se basa en el cálculo de un factor de dilución para cada muestra, obtenido a partir del cálculo de los cocientes de las intensidades de cada señal respecto a las intensidades de esas señales en un espectro de referencia. Este cálculo permite corregir la intensidad de cada *bucket*, teniendo en cuenta no sólo la intensidad global del propio espectro sino también la intensidad de esa señal en el resto de espectros. Una vez normalizados, los espectros procesados fueron transformados en una matriz con una

anchura de *bucket* de 0.01 ppm, con la que se llevó a cabo el análisis estadístico de los datos.

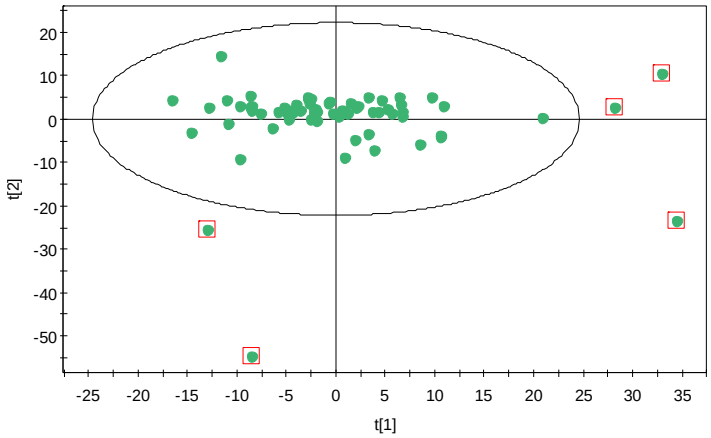
2.3 Análisis no supervisado

Tras el procesado de los datos se llevó a cabo el análisis no supervisado. En éste análisis se busca, en primer lugar, poder identificar y estudiar cualquier muestra que presente un comportamiento anómalo en relación al de su grupo (posibles *outliers*). Una vez identificadas estas muestras y, tras decidir de manera justificada su inclusión/exclusión en el estudio, se estudia la existencia de posibles tendencias o agrupaciones que puedan existir entre las muestras, debidas a variables distintas a la variable del estudio.

La identificación y análisis de este tipo de muestras y/o tendencias es muy importante en los estudios de metabolómica ya que la presencia de fuentes de heterogeneidad no identificadas, podrían interferir en el análisis y la correcta interpretación de los resultados. Además, como se ha mencionado anteriormente, la orina se caracteriza por presentar un elevado número de metabolitos y una gran variabilidad interindividual. Por esta razón, una vez realizado el procesado de los datos se evaluó detalladamente la homogeneidad de las muestras.

Para ello, se realizó un modelo de PCA de todo el conjunto de muestras y uno para cada grupo de muestras incluidas en el estudio. A través del estudio de los PCA se identificaron las muestras que estaban fuera del intervalo de confianza del 95% en el gráfico T^2 de *Hotelling* y que mostraban un comportamiento estadísticamente distinto al del resto de su grupo. La **Figura 32** representa el gráfico T^2 *Hotelling* y el *score plot* que se utilizaron para la identificación de las muestras *outliers* en el grupo de pacientes con HBP.

(A)



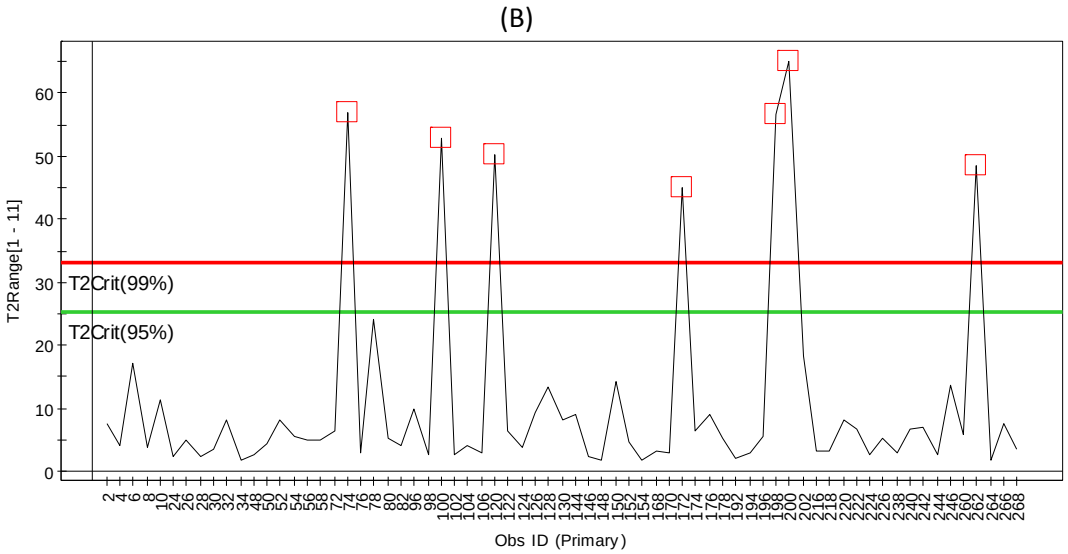


Figura 32: *Score plot* (a) y gráfico T^2 de *Hotelling* (b) derivados del análisis PCA de los espectros ^{1}H-RMN CPMG correspondientes a las muestras de orina del grupo de HBP. Las muestras marcadas con el cuadrado rojo se identificaron como *outliers* dentro de su grupo.

Por un lado, se identificaron once muestras en el grupo de HBP y cinco muestras de pacientes con CaP que presentaban señales muy intensas en el espectro de H^1-RMN que no estaban presentes, o lo estaban en menor intensidad, en el resto de las muestras incluidas en el estudio. Estas señales, entre las que se incluía la glucosa, el ácido hipúrico, la creatinina o el TMAO (N-óxido de Trimetilamina) (**Figura 33**), fueron asociadas a diferencias en la dieta o a distintas condiciones fisiológicas de los pacientes. La presencia de glucosa en orina se conoce como glucosuria, y puede estar asociada a la diabetes o bien a un defecto hereditario en su reabsorción a nivel del túbulo renal. El TMAO deriva del metabolismo de la colina y está regulado por la microflora intestinal, su incremento en orina puede asociase a dietas con alto contenido en carne roja, pescado o consumo de bebidas energéticas o suplementos dietéticos, aunque también se puede asociar a fallo renal o problemas cardiovasculares. El ácido hipúrico deriva del metabolismo del benzoato de sodio, compuesto que se utiliza como aditivo alimentario. Su incremento en orina se ha observado en individuos que siguen dietas ricas en polifenoles, sobre todo derivados del té. La concentración de creatinina en orina puede variar en función del momento de recogida de la muestra, de la concentración de la misma, del ejercicio físico o de patologías asociadas a fallo renal (Griffin *et al*, 2015; Slupsky *et al*, 2007; Walsh *et al*, 2006).

Otras dos muestras del grupo de pacientes con CaP y cuatro en el grupo de HBP presentaron un espectro de baja calidad probablemente debido a la degradación de la muestras durante el proceso de almacenamiento (**Figura 33**). El proceso de almacenamiento y procesado de las muestras es muy importante ya que los

metabolitos pueden degradarse si las condiciones preanalíticas no son adecuadas, alterando la calidad de las mismas, como se ha indicado anteriormente (Yin *et al*, 2013).

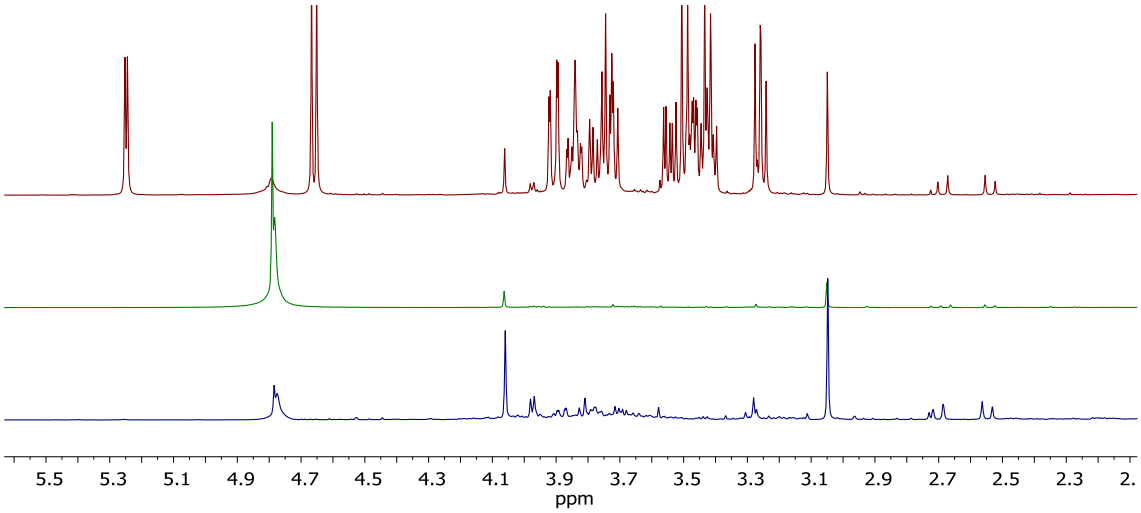


Figura 33: Comparación de un espectro correspondiente a una muestra clasificada como *outlier* por presencia de glucosa (granate) y otra clasificada como *outlier* por baja calidad (verde) con un espectro correspondiente a una muestra del mismo grupo con buena calidad (azul).

Por otro lado, dos muestras del grupo de HBP presentaron señales en el espectro correspondientes a distintos contaminantes presentes durante el procesado y preparación de las muestras (manitol y etanol). Finalmente, una muestra del grupo de HBP presentó niveles altos de paracetamol.

En total, el análisis del *score plot* y del gráfico T^2 *Hotelling,* junto con el análisis visual de los espectros, llevó a la exclusión justificada de 25 muestras del análisis, debido a su condición de *outliers*. Una vez excluidas las muestras se realizaron nuevos modelos PCA para observar el comportamiento del resto de las muestras tras la exclusión de la variabilidad derivada de estas muestras. Generalmente, los *outliers* se caracterizan por presentar señales muy diferentes al resto de muestras (p.ej., medicamentos, alimentos,…) y conducen a modelos erróneos. Por esta razón, es necesario crear nuevos modelos tras su exclusión, para evaluar el comportamiento del resto de muestras sin la presencia de ese tipo de muestras.

Una vez estudiada y justificada la exclusión del estudio de las 25 muestras identificadas como *outliers* se prosiguió con el análisis no supervisado del resto de muestras de forma conjunta (**Tabla 12**).

Tabla 12: Número de muestras incluidas en el estudio antes y después de la identificación de *outliers*.

	HBP	CaP
n inicial	69	71
n final	51	64

Tras la exclusión de los *outliers*, el análisis PCA de las muestras incluidas en el estudio se empleó para evaluar si existía alguna tendencia debida a otras variables: edad, niveles de PSA, carga tumoral, etc (**Figura 34**). Como se puede apreciar, no se observa ningún patrón o agrupación en función de estos descriptores que pudiera interferir en el posterior análisis supervisado entre pacientes con HBP *vs*. pacientes con CaP. Además, se construyó un modelo PCA para evaluar si existía, de manera no supervisada, alguna agrupación entre los dos grupos de estudio, no evidenciándose ninguna agrupación entre ellos.

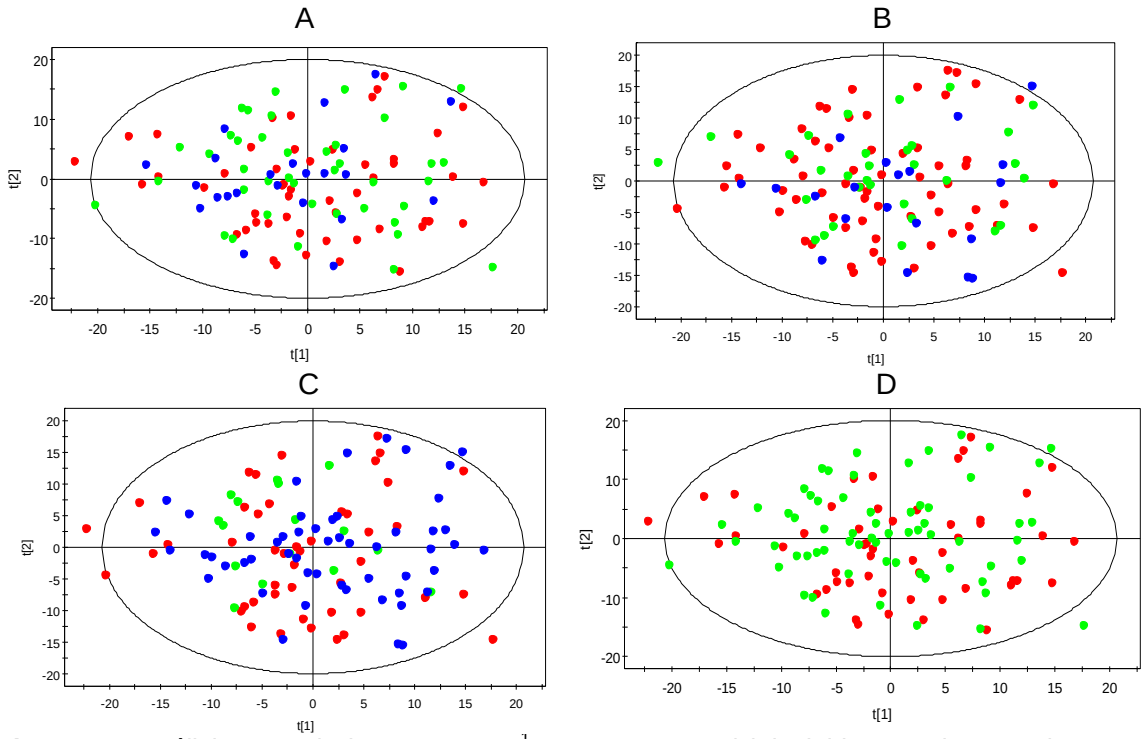


Figura 34: Análisis PCA de los espectros ^{1}H-RMN CPMG: (a) incluidos en el grupo de CaP según el índice de *Gleason* (●: *Gleason* 0; ●:*Gleason* 5-6; ●: *Gleason* >7), según la edad (b) (●: < 60 años; ●: 60-70; ●: >70), según el valor del PSA (c) (●: < 3; ●: 3-5; ●: >5) y según el grupo al que pertenecen HBP (●) o pacientes con CaP (●).

2.4 Análisis supervisado

Una vez estudiadas las posibles variables que podían interferir en el análisis de los resultados se procedió al análisis supervisado de los resultados de RMN. Este análisis emplea el conocimiento previo que se tiene de las muestras, es decir, la pertenencia a un determinado grupo del estudio. El análisis de los modelos estadísticos construidos (OPLS-DA) permite identificar las diferencias entre los grupos de muestras incluidos en el modelo.

En el modelo OPLS-DA obtenido para la comparación entre las muestras de pacientes con HBP y pacientes con CaP, representado en la **Figura 35,** se obtuvo un valor de ajuste de los datos ($R^2(Y)$=0.586) adecuado para un estudio de metabolómica. En cambio, como resultado de la validación cruzada, el modelo presentaba una capacidad de predicción ($Q^2(Y)$=-0.230) baja, indicando la incapacidad del modelo para predecir la clasificación de nuevas muestras y el posible sobreajuste del modelo.

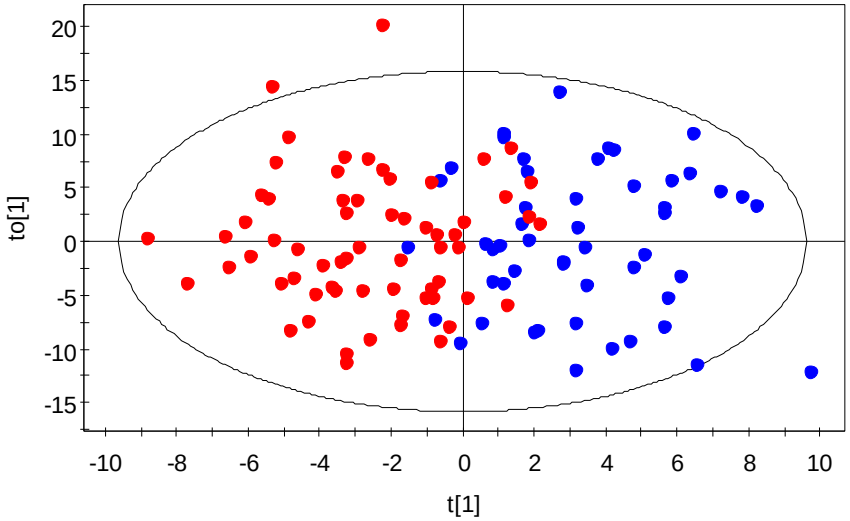


Figura 35: *Score plot* correspondiente al modelo OPLS-DA resultante de la comparación entre los perfiles metabólicos de los espectros ^{1}H-RMN CPMG de los pacientes con CaP (●) y el grupo de muestras con HBP (●).

En los modelos supervisados, es importante evaluar la calidad de los análisis realizados y la fiabilidad de los modelos obtenidos. En este tipo de modelos existe el riesgo de que, debido a un sobreajuste del modelo, el análisis conduzca a la identificación de variables no informativas, no relacionadas con la discriminación entre los grupos de muestras del estudio. La validación interna del modelo estadístico multivariante (OPLS-DA) se realizó mediante el test de permutación (n=100) para el modelo PLS-DA equivalente. Al comparar los valores de R^2 (Y) y Q^2 (Y) permutados con respecto a los valores del modelo original, se observó cómo los valores obtenidos de R^2 (Y) y de Q^2 (Y) en el modelo original no eran significativamente superiores a los

valores obtenidos en las permutaciones (**Figura 36**). Como se ha explicado anteriormente, para considerar válido el modelo OPLS-DA en metabolómica, los valores permutados de bondad de ajuste y de capacidad de predicción de los modelos generados de forma aleatoria deben ser significativamente inferiores a los del modelo original.

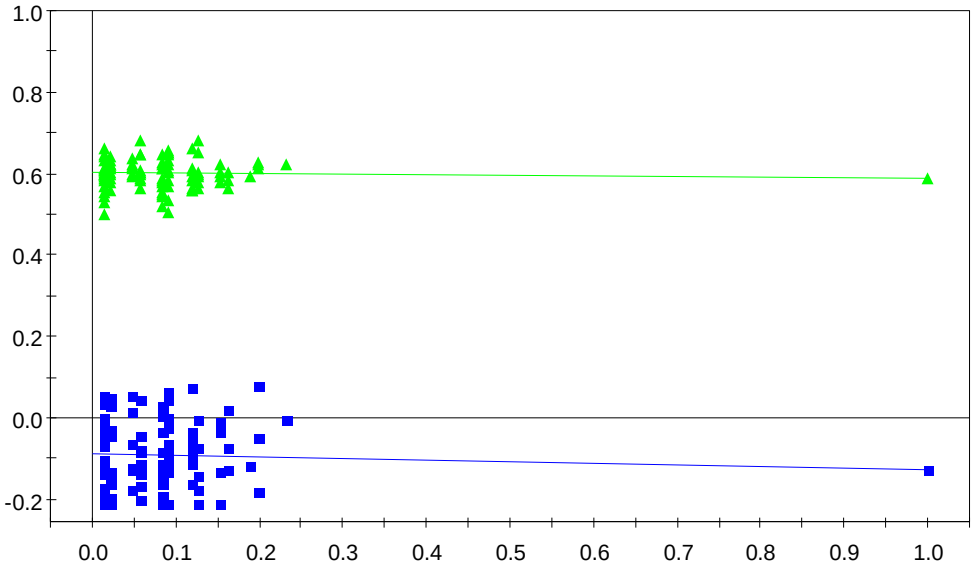


Figura 36: Resultado del test de permutación obtenido para el modelo PLS-DA de pacientes con CaP e individuos con HBP. En el eje de abscisas se representa el valor del coeficiente de correlación entre el valor del modelo original y los valores de cada permutación. El eje de ordenadas representa los valores obtenidos para R^2 (Y) (verde) y Q^2 (Y) (azul).

Por otro lado, el test de permutación permitió, además, identificar que el modelo estaba sobreajustado. En los modelos con sobreajuste el sistema encuentra una combinación de variables capaces de discriminar los grupos de estudio pero tiene un rendimiento pobre para predecir nuevos resultados en un conjunto de muestras externas (Mehmood *et al*, 2012; Saccenti *et al*, 2014). Los factores que influyeron en el sobreajuste observado en nuestro modelo fueron, por un lado, el bajo número de muestras incluidas en el estudio con respecto al número de variables obtenidas para cada muestra. A esta situación, que se da en la mayoría de los modelos de metabolómica, se suma el hecho de que cuando se trabaja con muestras de orina la abundancia de señales hace necesario dividir el espectro en *buckets* más pequeños, aumentando aún más el número de variables. Por otro lado, las muestras de orina, como se ha comentado anteriormente, presentan una elevada variabilidad entre muestras, enmascarando las diferencias entre los grupos de estudio.

2.4.1 Selección de variables

Una aproximación para mejorar la capacidad predictiva del modelo y poder identificar aquellas variables que están relacionadas con la patología de estudio

consiste en aplicar un método de selección de variables. La selección de variables se fundamenta en mantener las regiones del espectro que contribuyen de forma significativa a la discriminación entre los grupos y excluir las variables redundantes y no informativas (Quintás *et al*, 2012). El método de selección de variables empleado se basó en la relación entre los coeficientes de regresión y sus correspondientes errores (CoeffCS/CoeffCScvSE) (Diaz *et al*, 2013). Se mantuvieron sólo aquellas variables cuya relación CoeffCS/CoeffCScvSE fuese superior a uno.

La selección de variables se aplicó a la matriz de datos, reduciendo el número de variables de 823 a 108. La selección de variables no garantiza que el nuevo modelo no presente sobreajuste y, por tanto, no mejore la capacidad de predicción del mismo. En nuestro caso, en el nuevo modelo OPLS-DA mejoró significativamente el valor predictivo respecto al del modelo anterior ($Q^2(Y)$= 0.416). (**Figura 37**), mientras que el valor de $R^2(Y)$ apenas varió con respecto al del modelo original ($R^2(Y)$= 0.600). Este dato confirmó que las variables excluidas en el nuevo modelo no estaban relacionadas con la discriminación entre los grupos, por lo que su exclusión no provocó la pérdida de información relevante para el análisis. A su vez, estos resultados confirman que, las variables que se mantuvieron en el modelo, estaban directamente relacionadas con la discriminación entre el perfil metabolómico de los pacientes con CaP y los individuos diagnosticados con HBP.

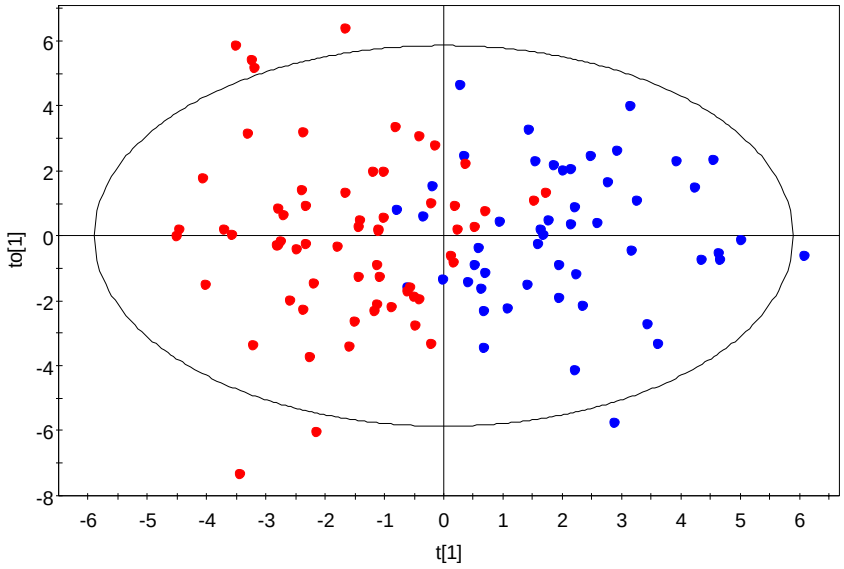


Figura 37: *Score plot* correspondiente al modelo OPLS-DA, tras la selección de variables, resultante de la comparación entre los perfiles metabólicos de los espectros ^{1}H-RMN CPMG de los pacientes con CaP (●) y el grupo de muestras con HBP (●).

La validación interna del modelo estadístico multivariante (OPLS-DA) obtenido tras la selección de variables se realizó de nuevo mediante el test de permutación (n=100) para el modelo PLS-DA equivalente. En este caso, los resultados obtenidos en el test

de permutación mostraron un valor de R^2 (Y) y de Q^2 (Y) en el modelo original significativamente superiores a los valores obtenidos en las permutaciones, confirmando así la validez del modelo estadístico tras la selección de variables (intersección recta R^2 (Y) = 0,346 y recta Q^2 (Y) = -0,239, respectivamente) (**Figura 38**).

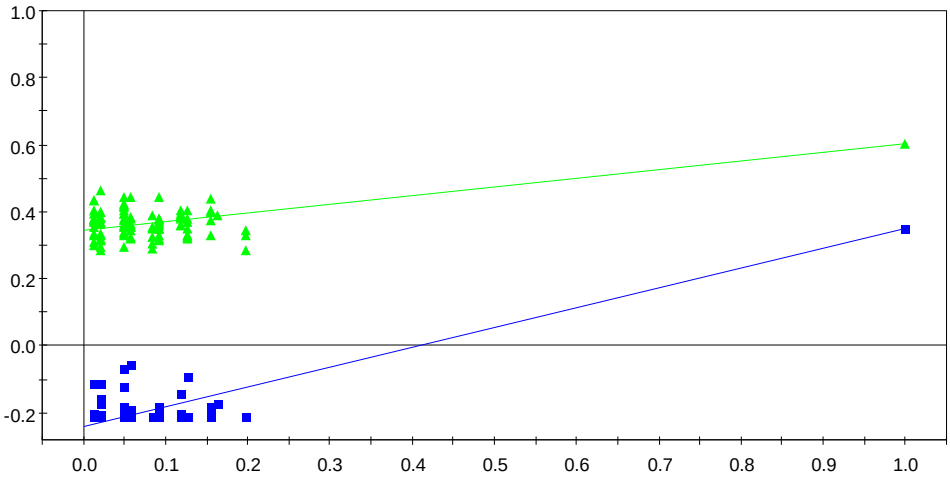


Figura 38: Resultado del test de permutación obtenido para el modelo PLS-DA de pacientes con CaP y de individuos diagnosticados con HBP empleando 100 permutaciones aleatorias después de la selección de variables. El eje de ordenadas representa los valores de R^2 (Y) (verde) y Q^2 (Y) (azul) para cada permutación con respecto a los valores del modelo original, mientras que el eje abscisas muestra los coeficientes de correlación entre los modelos permutados y el modelo original.

2.5 Identificación y cuantificación de metabolitos

La identificación de los metabolitos se realizó a partir del modelo OPLS-DA construido con las 108 variables. Se identificaron las regiones del espectro ^{1}H-RMN que más contribuían a la discriminación entre los distintos grupos de muestras incluidos en el estudio. La lista de los valores del VIP del modelo OPLS-DA tras la selección de variables se utilizó para identificar las regiones más relevantes. De las 108 variables obtenidas en la selección de variables, 40 presentaron un valor de VIP superior a uno y se incluyeron como regiones relevantes en la discriminación entre los pacientes con CaP y los individuos con HBP.

El análisis de los valores del coeficiente de regresión de las regiones de espectro con valor de VIP superior a uno permitió estudiar la magnitud y la dirección de los cambios observados (**Figura 39)**.

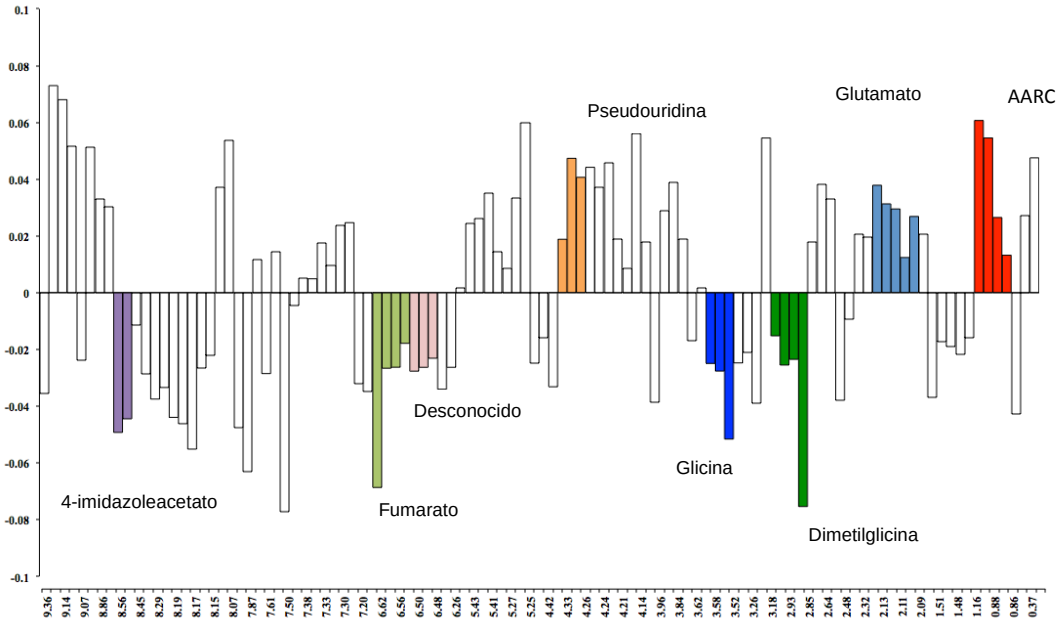


Figura 39: Representación gráfica de los coeficientes de regresión del modelo OPLS-DA para la comparación del perfil metabolómico de la orina de los pacientes con CaP frente a la de individuos diagnosticados con HBP. Las regiones coloreadas en el gráfico corresponden a los metabolitos que presentaron diferencias estadísticamente significativas (p valor < 0.05, U Mann Whitney) entre ambos grupos de muestras. El eje ordenadas representa el desplazamiento químico de las señales y, el eje abscisas representa valor del coeficiente de regresión.

A partir de las 40 regiones incluidas en el análisis, basándose en los desplazamientos químicos de las distintas señales presentes en los espectros de RMN, se procedió a la integración de las señales correspondientes. Finalmente, se identificaron 26 señales. Se calcularon las medias y valores de s.e.m para cada grupo, así como el coeficiente de variación entre los niveles encontrados en ambos grupos de muestras. La significación estadística de las diferencias de intensidad de los metabolitos identificados entre ambos grupos fueron analizadas de forma univariante, utilizando la prueba U de Mann Whitney. Ocho metabolitos presentaron diferencias en sus concentraciones estadísticamente significativas cuando se comparó el perfil metabolómico de la orina de pacientes con CaP frente al de individuos diagnosticados con HBP (**Tabla 13**).

La orina de pacientes con CaP presentó niveles altos de AACR, glutamato y pseudouridina, y niveles bajos de glicina, dimetilglicina, fumarato, 4-imidazolacetato y un metabolito desconocido cuando se comparó con los niveles encontrados en la orina de individuos diagnosticados con HBP.

Tabla 13: Intensidades medias y s.e.m de los metabolitos que presentan variaciones estadísticamente significativas en la comparación entre individuos con HBP y pacientes con CaP.

Metabolito	Ppm	HBP	CaP	p-valor	% Variación
		Media ± s.e.m	Media ± s.e.m		
AACR	0.93-0.84	13.10 ± 0.26	14.33 ± 0.37	0.011*	9.37
Glutamato	2.11-2.08	11.72 ± 0.18	12.42 ± 0.21	0.012*	5.98
Glicina	3.58-3.56	23.29 ± 1.01	20.49 ± 0.94	0.015*	-12.00
Desconocido 1	6.49-6.47	0.69 ± 0.09	0.45 ± 0.04	0.027*	-34.50
Fumarato	6.55-6.49	0.99 ± 0.05	0.87 ± 0.03	0.021*	-12.54
4-imidazolacetato	8.56-8.51	1.77 ± 0.11	1.37 ± 0.62	0.006*	-22.52
Pseudouridina	4.30-4.27	6.88 ± 0.13	7.68 ± 0.41	0.049*	11.65
Dimetilglicina	2.94-2.92	20.56 ± 0.91	17.17 ± 1.11	0.034*	-16.48

*p-valor < 0.05; s.e.m: Error estándar de la media.

2.6 Interpretación biológica

Los resultados de este estudio revelaron la existencia de diferencias estadísticamente significativas entre el perfil metabolómico en orina de pacientes con CaP e individuos diagnosticados con HBP. Estas diferencias se basan en un conjunto de metabolitos específicos, cuya información podría ser de utilidad en el diagnóstico precoz del CaP. El análisis de las alteraciones en estos metabolitos reveló que el CaP podría estar asociado con cambios importantes relacionados con el metabolismo energético.

En nuestro estudio, se observó un descenso en los niveles de glicina y dimetilglicina en el perfil metabolómico de la orina de pacientes con CaP en relación a los niveles observados en los individuos con HBP. Estos resultados coinciden con otros estudios publicados donde analizaron muestras de suero (Kumar *et al*, 2015) y orina (Struck-Lewicka *et al*, 2015) de pacientes con CaP e individuos sanos. Kumar y colaboradores encontraron niveles elevados de sarcosina y niveles de disminuidos de glicina en muestras de suero de pacientes con CaP cuando las comparaban con muestras de individuos sanos. Además, el estudio realizado por CL-EM y CG-EM realizado por Struck-lewicka y colaboradores reveló niveles bajos de glicina cuando comparaba los perfiles metabolómicos de la orina de pacientes con CaP e individuos con HBP.

La glicina es convertida en sarcosina por el enzima glicina-N-metiltransferasa (GNMT). La sarcosina es un derivado metabólico del aminoácido N-metilglicina que ha sido asociado a CaP en estudios previos (Sreekumar *et al*, 2009). Sus niveles son regulados por la sarcosina deshidrogenasa (SARDH), enzima capaz de transformar la sarcosina a glicina, y por la enzima dimetilglicina deshidrogenasa (DMGDH), la cual

genera dimetilglicina a partir de la sarcosina (Sreekumar *et al*, 2009). El papel que juega la sarcosina en el CaP ha sido sujeto de muchos estudios (Bianchi *et al*, 2011; Issaq & Veenstra, 2011; Jentzmik *et al*, 2010; Khan *et al*, 2013; Kumar *et al*, 2015; Lucarelli *et al*, 2012; Miyake *et al*, 2012; Sreekumar *et al*, 2009), generando controversia y no dejando clara la utilidad de este biomarcador (Ploussard & De La Taille, 2010). En nuestro estudio, se encontraron niveles elevados de este metabolito en pacientes con CaP, aunque la variación no fue estadísticamente significativa. Globalmente, nuestros resultados se ciernen en una interconversión entre glicina/dimetilglicina y sarcosina mediante la activación de la DMGH y la GNMT, y regulado por la SARDH. También existen otros mecanismos que podrían contribuir a la disminución en los niveles de glicina. Un estudio reciente de CaP por metabolómica mostró que el consumo de glicina está asociado con una supresión de la proliferación celular mediante su participación en el metabolismo del intercambio de carbonos (Zhang *et al*, 2012b). Esta ruta metabólica ha sido tradicionalmente considerada como un proceso de limpieza ("*housekeeping process*"), e implica una red metabólica compleja basada en las reacciones químicas relacionadas con los compuestos de folato. Estudios recientes sugieren también que la hiperactivación de esta vía podría ser la causante de la oncogénesis y de la persistencia de tumores (Locasale, 2013). En este contexto, se ha observado que el metabolismo de la glicina está implicado en la transformación celular y en el proceso tumoral. Este proceso podría estar mediado por la activación de la glicina deshidrogenasa descarboxilante (GLDC) que cataliza la descarboxilación de la glicina. Además, la pérdida de control que se produce durante la proliferación celular requiere de un exceso de energía para poder llevarse a cabo (Zhang *et al*, 2012a). Por lo tanto, el piruvato, los ácidos grasos y, en especial los aminoácidos, pueden suministrar sustratos al ciclo de Krebs para mantener la producción mitocondrial en la células cancerosas (Chen & Russo, 2012).

Uno de los factores que contribuye a la disponibilidad de aminoácidos es el síndrome metabólico que se da aproximadamente en el 60% de los pacientes con CaP, conocido como caquexia (Utech *et al*, 2012). Este proceso produce un incremento en el catabolismo proteico mediante la activación de la proteólisis y tiene un impacto importante en los niveles de AACR (O'Connell, 2013). En condiciones normales, la oxidación de los AACR en el musculo esquelético proporciona el 6-7% de la energía necesaria, pero en circunstancias donde existe un incremento catabólico como la caquexia, la contribución puede llegar a ser del 20% (Lam & Poon, 2008). En esas condiciones, se podría esperar un aumento de los niveles de AACR circulantes, como ocurre en nuestro estudio y en otros estudios realizados en tejido prostático

(Giskeødegård *et al*, 2013; McDunn *et al*, 2013) y suero (Giskeødegård *et al,* 2015) de pacientes con CaP. Esto podría explicar también los resultados obtenidos en estudios previos, donde se muestran niveles significativamente elevados de AACR en ciertos procesos neoplásicos, por ejemplo, en cáncer gástrico o esofágico (Fan *et al*, 2012; Zhang *et al*, 2013b).

Los AACR se pueden convertir en acetíl-CoA y otras moléculas orgánicas implicadas en el ciclo de Krebs. La flexibilidad metabólica que se da por los múltiples precursores del ciclo de Krebs permite a las células cancerosas responder correctamente a las necesidades metabólicas que se dan en la evolución del tumor (Boroughs & DeBerardinis, 2015). Del mismo modo, el catabolismo de los AACR proporciona una fuente importante para la síntesis de aminoácidos, especialmente en la síntesis de glutamina y glicina. Diferentes estudios en cáncer han demostrado alteraciones en los niveles de glutamina que podrían estar asociados con el aumento en la actividad metabólica derivada de las condiciones de hipoxia y del hipermetabolismo que ocurre en el entorno del tumor (Eigenbrodt *et al*, 1998).

Durante la proliferación, las células consumen glutamina y la convierten en glutamato a través de un conjunto de reacciones de desamidación y transamidación, participando la glutaminasa amidohidrolasa mitocondrial (Hensley *et al*, 2013). En este contexto, el aumento de glutamato en la orina de pacientes con CaP podría explicarse como consecuencia de la hidrólisis de la glutamina para la generación de amonio y glutamato con el objetivo de intentar equilibrar el pH de las células tumorales, contrarrestando el exceso de lactato producido por el efecto Warburg. Estos resultados concuerdan con estudios previos realizados en suero (Giskeødegård *et al*, 2015) y tejido prostático (McDunn *et al*, 2013) en pacientes con CaP. El glutamato es transformado posteriormente en α-cetoglutarato, a través de una serie de reacciones bioquímicas llamadas glutaminolisis, que contribuyen a reponer los metabolitos intermediarios del ciclo de krebs que se van consumiendo (DeBerardinis *et al*, 2008).

También en relación con el metabolismo de los aminoácidos, se observó un descenso significativo en la orina de los pacientes con CaP del metabolito 4-imidazolacetato, compuesto asociado al metabolismo de la histidina. Este metabolito fue identificado en un estudio previo llevado a cabo en muestras de suero de pacientes con CaP, recogidas 20 años antes de ser diagnosticados de CaP (Mondul *et al*, 2015). Mondul y colaboradores encontraron una relación entre los niveles de este metabolito y el riesgo de sufrir CaP, así como con la agresividad del tumor. En otros estudios previos también se ha observado niveles elevados de histidina en suero (Giskeødegård *et al*, 2015) y tejido (McDunn *et al*, 2013) de pacientes con CaP.

Nuestros hallazgos podrían ser reflejo de la capacidad limitada de las células con CaP para procesar este aminoácido. Tanto las alteraciones en el metabolismo de la histidina como en el metabolismo de los AACR (valina, leucina e isoleucina) han sido observadas también en otros tipos de canceres (cáncer de ovario, cáncer de mama, etc) (Ke *et al*, 2015; Schramm *et al*, 2010).

Los niveles urinarios de pseudouridina se encontraron aumentados en el perfil metabolómico de pacientes con CaP en comparación con individuos diagnosticados con HBP. La pseudouridina es un isómero del nucleósido uridina en el que el radical uracilo está unido por un enlace C-glicosídico, en lugar de un enlace N-glucosídico. El aumento en los niveles de uracilo o de otros metabolitos derivados del uracilo, como la 2´-pseudouridina (Mondul *et al*, 2015), también ha sido descrito en estudios previos de CaP (Jiang *et al*, 2010; McDunn *et al*, 2013; Spur *et al*, 2013; Sreekumar *et al*, 2009), sugiriendo la implicación de este metabolito en el desarrollo de la enfermedad. Alteraciones en los niveles de pseudouridina han sido observados en otros procesos patológicos (Masaki *et al*, 2006; Rasmuson & Bjork, 1995; Vicente-Munoz *et al*, 2015) y se han relacionado con la progresión de la enfermedad, la carga tumoral y el estadio clínico (Tamura *et al*, 1987).

Finalmente, se encontraron variaciones estadísticamente significativas entre los dos grupos incluidos en el estudio en los niveles de fumarato, metabolito clave del ciclo de Krebs. Dentro del ciclo de Krebs, el complejo succinato deshidrogenasa convierte al succinato en fumarato, que posteriormente es transformado en malato a través de la fumarato hidratasa (FH). Las mutaciones en este enzima (FH) han sido previamente relacionadas con carcinoma de células renales, uterinos o cáncer de piel (Tomlinson *et al*, 2002). El descenso en los niveles de otros intermediarios del ciclo de Krebs (isocitrato, aconitato y succinato) en la orina de pacientes con CaP y, como consecuencia, la alteración del metabolismo energético, han sido descritos también en estudios previos (Struck-Lewicka *et al*, 2015). Además, es interesante destacar la relación entre el descenso de los niveles de fumarato en la orina de pacientes con CaP, cuando se compara con individuos diagnosticados con HBP, y su correlación positiva con estudios previos donde muestran acumulación de este metabolito en pacientes con CaP con metástasis ósea (Thapar & Titus, 2014) y en el tejido prostático (McDunn *et al*, 2013). En este sentido, destaca cómo el succinato, otro metabolito que experimenta una disminución de sus niveles en la orina de pacientes con CaP, también tiende a acumularse en las células cancerosas. Ambos metabolitos pertenecen a la familia de compuestos conocidos como oncometabolitos, de los cuales se sabe que se acumulan en células cancerosas y facilitan la progresión del cáncer

(Yang *et al*, 2013). Estos dos oncometabolitos están relacionados con una incorrecta estabilización de HIF-1α (subunidad alfa del factor 1 inducible por hipoxia) (Semenza, 2010), proteína clave en el cáncer, que se encuentra sobreexpresada en células de CaP (Thomas & Kim, 2008)

2.7 Relevancia de los resultados

En este estudio dirigido al análisis del perfil metabolómico en orina de pacientes con CaP e individuos diagnosticados con HBP se han combinado una serie de herramientas que han permitido un correcto tratamiento de los datos, facilitando así el análisis e interpretación de los resultados. La estrategia utilizada en el tratamiento de los datos ha consistido en el alineamiento por regiones de los espectros, normalización mediante PQN y la selección de variables. La combinación de estas herramientas ha permitido afrontar en este estudio la variabilidad biológica propia de las muestras de orina, condiciendo a la obtención de resultados de gran relevancia en CaP.

En la actualidad, la investigación en la búsqueda de biomarcadores no invasivos en CaP que puedan utilizarse para el cribado, diagnóstico, predicción o monitorización de la enfermedad con alta especificidad y sensibilidad sigue siendo fundamental (Thapar & Titus, 2014). En este contexto, hasta la fecha, sólo se ha publicado un estudio preliminar donde se analiza el perfil metabolómico de la orina de pacientes con CaP usando ^{1}H-RMN (Zaragoza *et al*, 2014). Sin embargo, nuestro trabajo es el primer estudio donde se consigue identificar, en orina, metabolitos capaces de discriminar entre pacientes con CaP e individuos diagnosticados con HBP utilizando la ^{1}H-RMN.

Existen publicados otros estudios metabolómicos realizados por RMN o EM con resultados también prometedores utilizando otros biofluidos (Kumar *et al*, 2015; Serkova *et al*, 2008; Stabler *et al*, 2011; Struck-Lewicka *et al*, 2015). Los resultados obtenidos en este estudio revelan la posibilidad de caracterizar, usando una herramienta no invasiva, el perfil metabólico urinario de pacientes con CaP para la identificación de posibles biomarcadores y así entender mejor las alteraciones metabólicas que se dan en este proceso neoplásico. La confirmación de las diferencias observadas en un grupo de muestras independiente podría ser clave en la validación de los resultados y en el avance del diagnóstico de esta enfermedad.

3. VALIDACIÓN DE BIOMARCADORES

3.1 Antecedentes

En un estudio previo realizado por nuestro grupo de investigación (Puchades-Carrasco, 2013) se analizó el perfil metabolómico de muestras de suero de pacientes diagnosticados con CPNM (estadios avanzados y estadios tempranos) e individuos sanos. El objetivo era caracterizar los perfiles metabolómicos de estos pacientes e identificar biomarcadores de utilidad para el diagnóstico temprano del CPNM. Los resultados alcanzados en este estudio permitieron la obtención de dos modelos de predicción, uno basado en un análisis multivariante y otro en una ecuación de regresión logística. Tras el análisis estadístico de los datos se construyó un modelo multivariante OPLS-DA (R^2 (Y)=0.931 y Q^2 (Y)=0.873) y, a partir de las variables relevantes de este modelo, se realizó un análisis de regresión logística que dio como resultado una ecuación probabilística que incluía cinco metabolitos: treonina, glutamina, lactato, colina y metanol (**Figura 40**).

$$\text{Log}\left[\frac{p}{1-p}\right] = \frac{\text{EXP}(17.70-(1.82\cdot\text{treonina})-(1.70\cdot\text{glutamina})-(0.41\cdot\text{colina})+(4.60\cdot\text{metanol})+(0.34\cdot\text{lactato}))}{(1+\text{EXP}(17.70-(1.82\cdot\text{treonina})-(1.70\cdot\text{glutamina})-(0.41\cdot\text{colina})+(4.60\cdot\text{metanol})+(0.34\cdot\text{lactato})))}$$

Figura 40: Ecuación de regresión logística del set de entrenamiento.

La diferencia fundamental entre el modelo de discriminación OPLS-DA y la ecuación de probabilidad obtenida a través de la regresión logística es el número de variables que se utilizan para la predicción de la clasificación de nuevas muestras. Por un lado, el modelo OPLS-DA se construye a partir de todas las regiones del espectro, es decir, contiene información de todas las variables que pueden estar relacionadas con la pertenencia a un grupo u otro. En cambio, en la ecuación de regresión se eliminan las variables no informativas o que aportan información redundante en el modelo, y utiliza para el cálculo de la probabilidad los niveles de únicamente cinco metabolitos directamente relacionados con la patología. Así pues, la ventaja de la ecuación de regresión con respecto al modelo OPLS-DA es la posibilidad de clasificar muestras de nuevos pacientes midiendo los niveles de un número de metabolitos muy reducido, lo que facilita la traslación a la práctica clínica mediante la cuantificación de esos metabolitos con otros métodos analíticos.

Como se ha comentado con anterioridad, la validación es una de las fases más importantes en los estudios de metabólica basados en la búsqueda e identificación de nuevos biomarcadores para el diagnóstico de enfermedades. La validación permite

determinar la sensibilidad y especificidad de los nuevos biomarcadores y su utilidad en la práctica clínica.

La validación de un modelo estadístico discriminante comprende una primera etapa de validación interna, en la que se busca confirmar la capacidad predictiva dentro de ese mismo conjunto de muestras y, posteriormente, una etapa de validación externa utilizando muestras independientes, en la que se busca confirmar la utilidad del modelo al extrapolarlo a otro conjunto de muestras. Tanto el modelo OPLS-DA (Q^2=0.873) como la curva ROC (AUC=0.985) proporcionaron buenos resultados en la validación interna. En la **Figura 41** se representa la validación interna de la ecuación de regresión mediante la curva ROC.

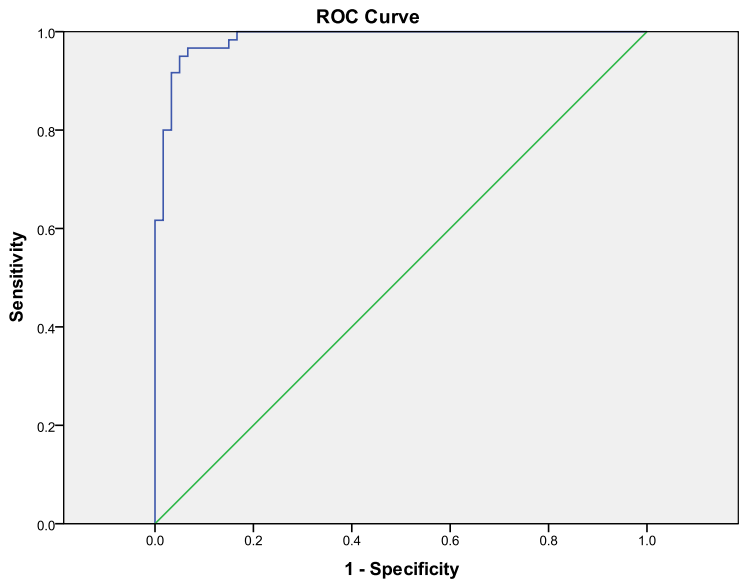


Figura 41: Curva ROC de la ecuación de regresión.

En la segunda parte del estudio, los dos modelos generados fueron evaluados con un conjunto de muestras externo. Generalmente, los estudios de metabolómica suelen incluir un número relativamente pequeño de muestras que se analizan al mismo tiempo. Para realizar estudios de validación es necesario utilizar un mayor número de muestras, por lo que normalmente son analizadas posteriormente en condiciones diferentes al conjunto de muestras inicial. Esto implica que, durante el proceso de adquisición de los espectros, la calibración del equipo de medida, el tratamiento de los datos, etc, pueden afectar variaciones técnicas entre los dos grupos de muestras independientes, dificultando así la correcta validación de los resultados.

En este sentido, resulta fundamental solventar estos problemas y poder utilizar datos procedentes de muestras de distintos lotes para validar los resultados, ampliar el tamaño muestral o poder comparar muestras analizadas en diferentes laboratorios. Entre las estrategias que se han desarrollado en los últimos años para superar esta

limitación destaca la aplicación de herramientas informáticas que eliminan esa variabilidad.

3.2 Set de validación

El set de validación en este estudio incluyó 40 muestras de pacientes con CPNM, 13 muestras de voluntarios sanos y 27 muestras de individuos con EPB. El grupo de EPB se incluyó como grupo control para analizar cómo se predecían este tipo de muestras en el modelo y estudiar si las diferencias identificadas eran específicas de CPNM o si eran similares para cualquier tipo de patología pulmonar.

En este sentido, los espectros de RMN de las muestras de suero del set de validación se compararon con los espectros de RMN de las muestras de suero del set de entrenamiento. Los dos conjuntos de muestras se estudiaron mediante análisis no supervisado. El análisis exploratorio demostró que existía una variabilidad entre ambos sets debida al efecto de lote (efecto *batch*), como se muestra en el gráfico del PCA correspondiente (**Figura 42A**).

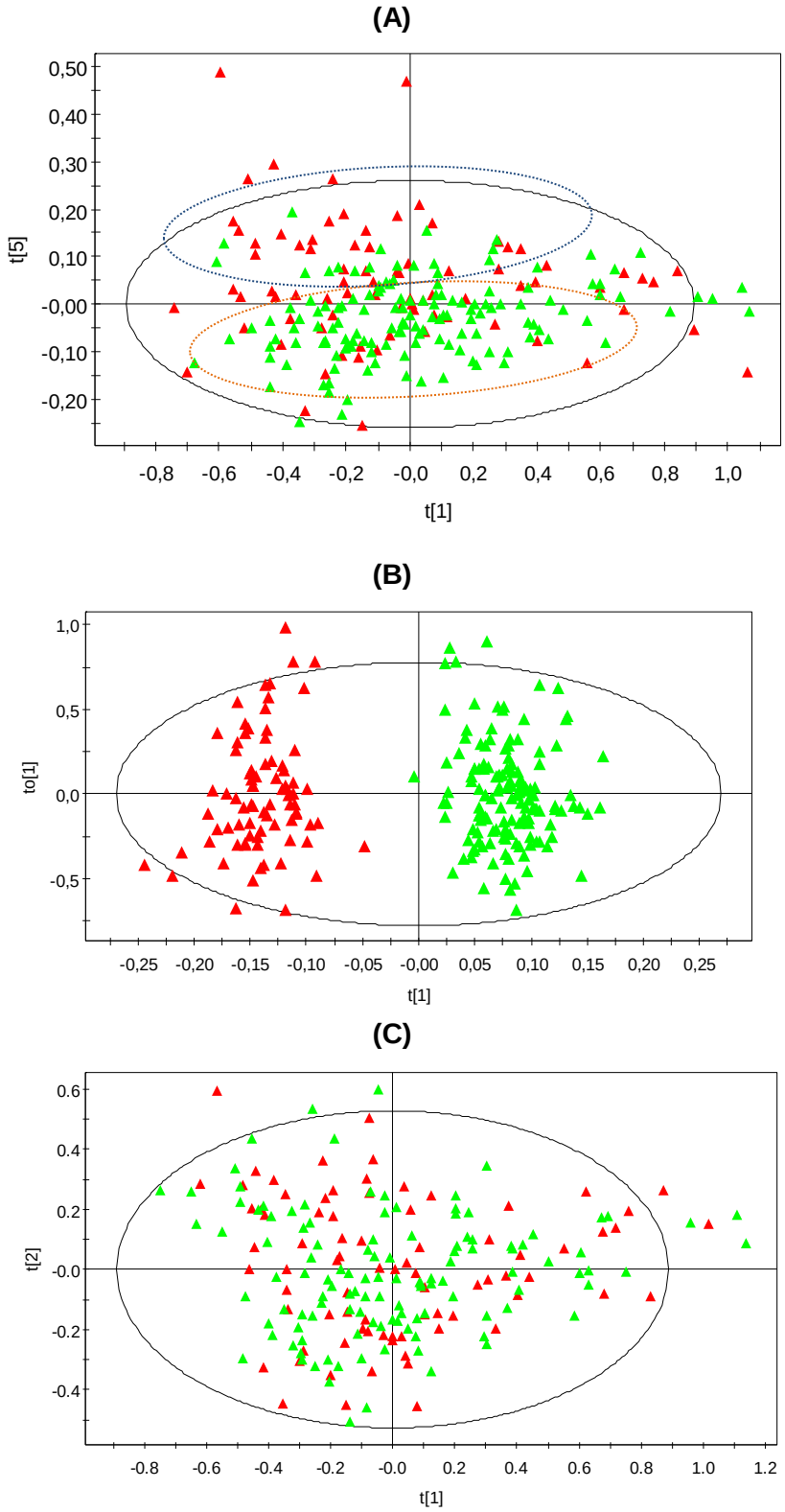


Figura 42: (A) PCA del set de entrenamiento (verde) y del set de validación (rojo) antes de usar la función ComBat. (B) *Score plot* del modelo OPLS-DA entre el set de entrenamiento (verde) y el set de validación (rojo) antes de usar la función ComBat. (C) PCA del set de entrenamiento (verde) y del set de validación (rojo) después de usar la función ComBat.

Con el objetivo de estudiar a qué variables se debían las diferencias características entre ambos conjuntos de muestras se realizó un modelo OPLS-DA (**Figura 42B**). El

análisis de los espectros ^{1}H-RMN se observó que las variaciones eran debidas a diferencias en regiones no informativas del espectro, relacionadas con la supresión del agua, la línea base, etc. A pesar de haber mantenido los mismos parámetros para la adquisición de los espectros y haber seguido los mismos protocolos para la adquisición, se observaron pequeñas variaciones en las condiciones de medición.

Con el fin de tener datos comparables entre ambos sets de muestras, se aplicó una herramienta para filtrar los datos y evitar el sesgo producido por la variabilidad entre las muestras. La función ComBat incluida en el paquete "sva" de R se utilizó en esta parte del análisis para integrar los datos analíticos obtenidos por RMN para ambos set de muestras. Cuando se aplica esta herramienta es importante que los grupos de muestras estén equilibrados, ya que la finalidad del algoritmo empleado en el análisis es eliminar la variabilidad existente entre lotes sin modificar la variabilidad existente entre los grupos de estudio. Por este motivo, se seleccionó un grupo representativo de muestras, equivalente en ambos sets (**Tabla 14**). Una descompensación en el número de muestras entre los sets o entre los distintos grupos del estudio dentro de los sets de muestras podría interferir en el análisis, dando como lugar la eliminación de otra variabilidad, no relacionada con la adquisición de los espectros.

Tabla 14: Número de muestras incluidas en el set de entrenamiento y en el set de validación al aplicar la función ComBat.

	Pacientes CPNM	Individuos Control
Set de entrenamiento	60	60
Set de validación	40	40

Tras el tratamiento de los datos, la homogeneidad de los sets fue verificada realizando un nuevo modelo de PCA (**Figura 42C**) que confirmó que la agrupación en función del set ya no estaba presente. El análisis estadístico no fue capaz de generar un modelo OPLS-DA, lo que confirmó la homogeneidad de las muestras.

La aplicación de la función ComBat permitió la utilización de forma conjunta de los datos de ambos sets de muestras para su posterior análisis y eliminó la variabilidad metodológica entre ambos sets. En la **Figura 43** se representan las diferencias entre las intensidades medias de las señales de los espectros en ambos sets de muestras antes y después de aplicar la aplicar la función ComBat, respectivamente. Se observa cómo la diferencia entre los dos sets de muestras se redujo drásticamente, facilitando el análisis posterior, en función del grupo de estudio (CPNM, EPB y voluntarios sanos).

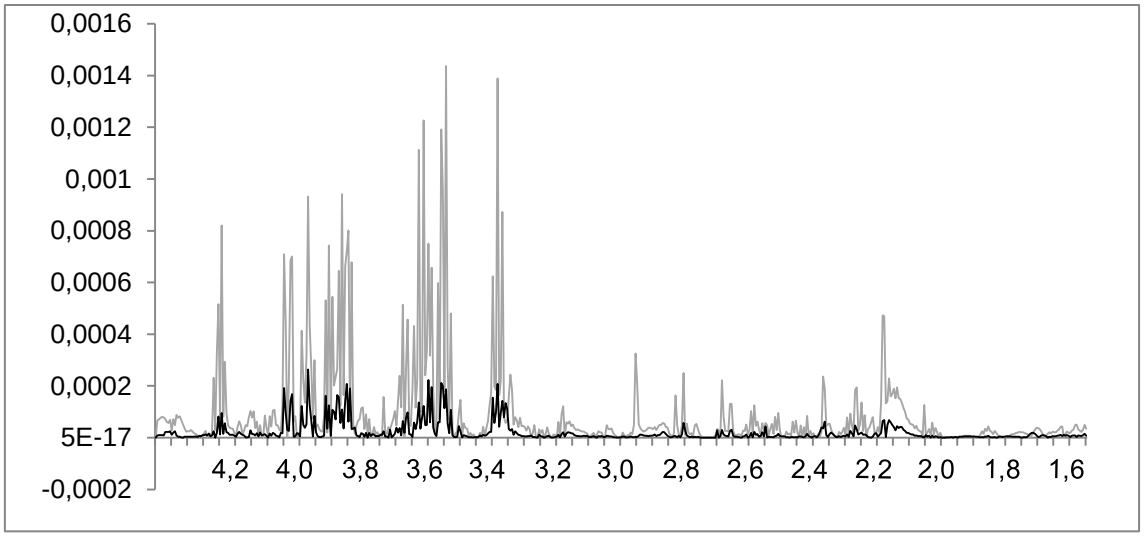


Figura 43: Variación en las intensidades medias de los espectros de RMN entre los dos lotes de muestras, antes (gris) y después (negro) de aplicar la función ComBat.

Antes de proseguir con el análisis, se calculó el modelo OPLS-DA, incluyendo únicamente las muestras del set de entrenamiento utilizado en el primer estudio, para la comparación entre las muestras de los pacientes con CPNM y las muestras de los individuos sanos tras aplicar ComBat. La finalidad de este nuevo modelo fue confirmar que el procesado de los datos para eliminar el efecto *batch* entre los sets de muestras, no había afectado a las diferencias identificadas como relevantes en la comparación entre los grupos del estudio en el modelo inicial.

3.3 Validación de los modelos estadísticos

Una vez corregido el efecto lote se evaluó la capacidad predictiva de los dos modelos generados, el OPLS-DA y la ecuación de regresión logística, con el set de muestras externo.

3.3.1 Capacidad predictiva del modelo OPLS-DA

Los modelos OPLS-DA están basados en la regresión por mínimos cuadrados parciales. Estos métodos permiten analizar muchas variables correlacionadas o no al mismo tiempo y obtener modelos con buen poder de predicción. Se basan en eliminar la información no relacionada con la variable de interés y maximizar aquella que evidencia una relación entre variables independientes y una o más variables dependientes.

En la validación externa del modelo OPLS-DA, el 95% de los pacientes diagnosticados con CPNM del set de validación fueron correctamente clasificados y, todos menos uno de los individuos sanos incluidos en el grupo de validación se clasificaron correctamente en su grupo. El 85,2% (23 muestras) del grupo de EPB se clasificaron como individuos sanos. Como se muestra en la **Figura 44** y en la **Tabla 15**

el modelo estadístico multivariante obtenido para la discriminación entre pacientes con CPNM e individuos sanos presentó una sensibilidad del 95% y una especificidad del 92,30% (87,50% para todas las muestra en ausencia de cáncer).

Tabla 15: Resultados de la predicción derivada del modelo OPLS-DA correspondiente a la comparación entre individuos sanos y pacientes con CPNM.

Set de validación	Clasificado como CPNM	Clasificado como sano	Correctamente clasificado	
Pacientes CPNM (40)	38	2	95%	
Individuos sanos (13)	1	12	92.31%	87,5%
Grupo EPB (27)	4	23	85,2%	

En el CPNM uno de los principales retos es el desarrollo de métodos de diagnóstico capaces de detectar la presencia de enfermedad en estadios tempranos. En los casos donde es necesario diagnosticar la enfermedad en estadios tempranos es muy importante tener valores altos de sensibilidad. Esto permitiría iniciar el tratamiento antes, evitando que la enfermedad evolucione a estadios más avanzados donde el tratamiento es más complicado y las tasas de supervivencia se ven reducidas (Jantus-Lewintre *et al*, 2012). Al mismo tiempo, la especificidad también es importante si el objetivo es trasladar los resultados a un método de cribado con un número de falsos positivos bajo. En nuestro caso, los resultados de sensibilidad y la especificidad fueron elevados.

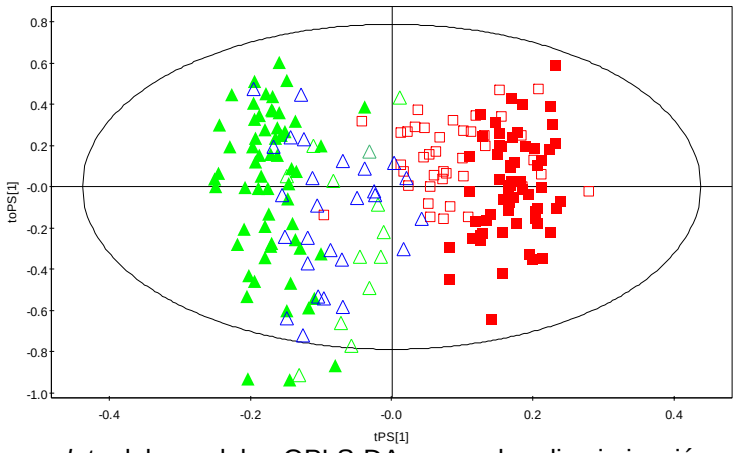


Figura 44: *Score plot* del modelo OPLS-DA para la discriminación entre pacientes diagnosticados con CPNM e individuos sanos, con la predicción de las muestras incluidas en el set de validación (Δ: individuos sanos-set de validación-;□: pacientes con CPNM-set de validación-; Δ:pacientes con EPB-set de validación-;▲: individuos sanos-set de entrenamiento-; ■: pacientes con CPNM-set de entrenamiento-).

3.3.2 Capacidad predictiva de la ecuación de regresión logística

El análisis de regresión logística es un método estadístico de gran utilidad en la generación de modelos predictivos en los que una o varias variables independientes

pueden explicar matemáticamente la probabilidad de que ocurra un evento (p.ej., sano/enfermo). La capacidad predictiva de la ecuación de regresión logística fue evaluada con el set de muestras externo. Los niveles de los cinco metabolitos de las muestras del set de validación fueron cuantificados y se aplicó la ecuación de regresión logística. Los valores superiores a 0.5 se clasificaron como pacientes con CP y los valores inferiores a 0.5 se clasificaron como individuos sanos. Los resultados obtenidos fueron de un 77,3% de las muestras clasificadas correctamente en el set de validación. El 77,5% y el 76,9% de las muestras de los pacientes con CPNM y de individuos sanos del set de validación, respectivamente, se clasificaron correctamente (**Tabla 16**). La especificidad y sensibilidad del modelo de regresión logística fue más baja que la del modelo OPLS-DA como era de esperar. Como se ha comentado anteriormente, el OPLS-DA incluye información sobre todas las regiones del espectro, mientras que la ecuación de probabilidad sólo incluye información de los niveles de cinco metabolitos.

Tabla 16: Resultados de la predicción de la ecuación de regresión logística correspondiente a la comparación entre individuos sanos y pacientes con CPNM.

Set de validación	Clasificado como CPNM	Clasificado como sano	Correctamente clasificado	
Pacientes CPNM (40)	**31**	9	77.50%	
Individuos sanos (13)	3	**10**	76.90%	70,00%
Grupo EPB (27)	9	**18**	66,66%	

Además, el 66,6% de las muestras diagnosticadas con EPB fueron clasificadas como individuos sanos. La disminución en el porcentaje de muestras con EPB clasificadas como individuos sanos (66,6% regresión logística) frente al 83,2% obtenido utilizando el modelo OPLS-DA se investigó evaluando las variaciones en los niveles de los cinco metabolitos incluidos en la ecuación (lactato, metanol, glutamina, colina y treonina) en los distintos grupos de muestras (**Figura 45**). Los resultados del análisis revelaron la existencia de diferencias entre los pacientes con CPNM e individuos sanos que también se observaban en la comparación entre los pacientes diagnosticados con EPB y los individuos sanos, como se observó en los niveles metanol y de lactato.

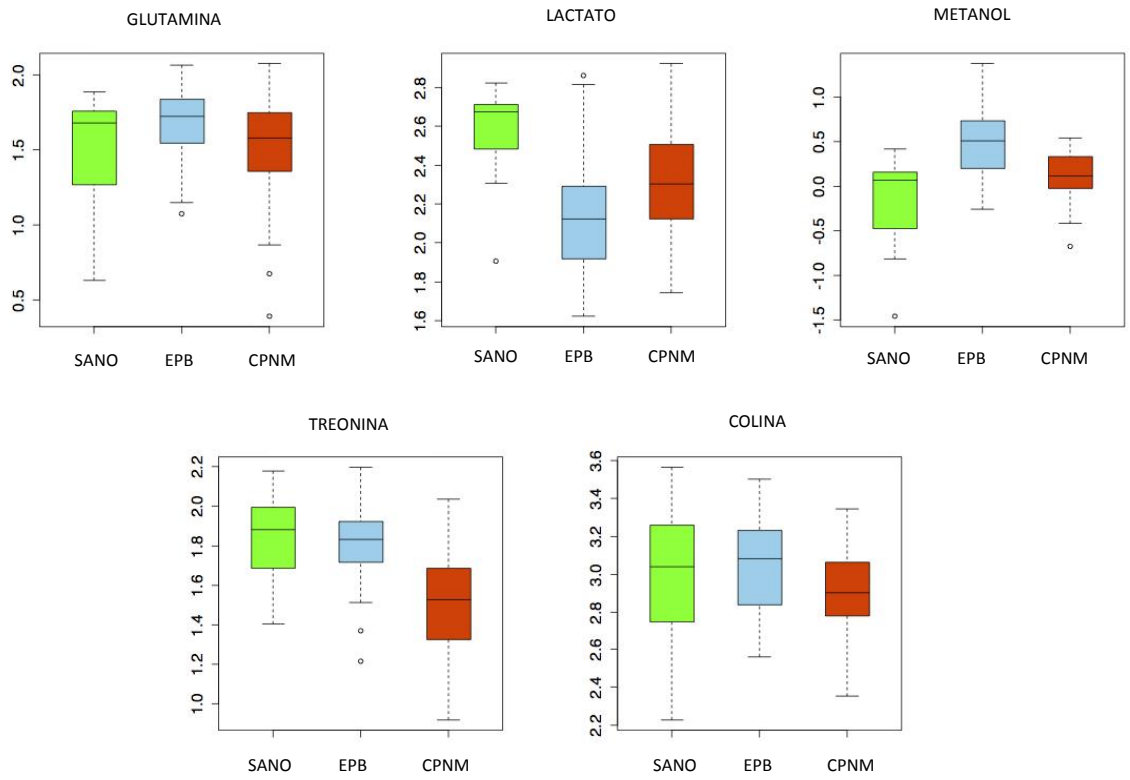


Figura 45: Diagrama de cajas representando la intensidad de los metabolitos incluidos en la ecuación de regresión logística para los diferentes grupos del estudio. Para cada caja, la línea central corresponde a la mediana, los extremos superior e inferior de la caja a los cuartiles y los bigotes o extremos más allá de la caja corresponden a ±1.5 el rango intercuartílico; los *outliers* se muestran como un punto.

3.4 Caracterización del perfil metabólico de pacientes con EPB

Con la finalidad de caracterizar las diferencias específicas entre los pacientes diagnosticados con EPB frente a los individuos sanos y frente a los pacientes con CPNM se analizaron los modelos OPLS-DA para la comparación del perfil metabolómico del suero de los pacientes con EPB frente a los distintos grupos. Este análisis aportó información complementaria que podría mejorar la interpretación de los modelos y conocer qué cambios son específicos de los pacientes con CPNM y cuáles son comunes en las enfermedades pulmonares.

Se generaron modelos estadísticos OPLS-DA para la comparación de individuos con EPB frente a individuos sanos ($R^2(Y)=0.963$; $Q^2(Y)=0.782$) y frente a pacientes con CPNM ($R^2(Y)=0.972$; $Q^2(Y)=0.856$) (**Figura 46**).

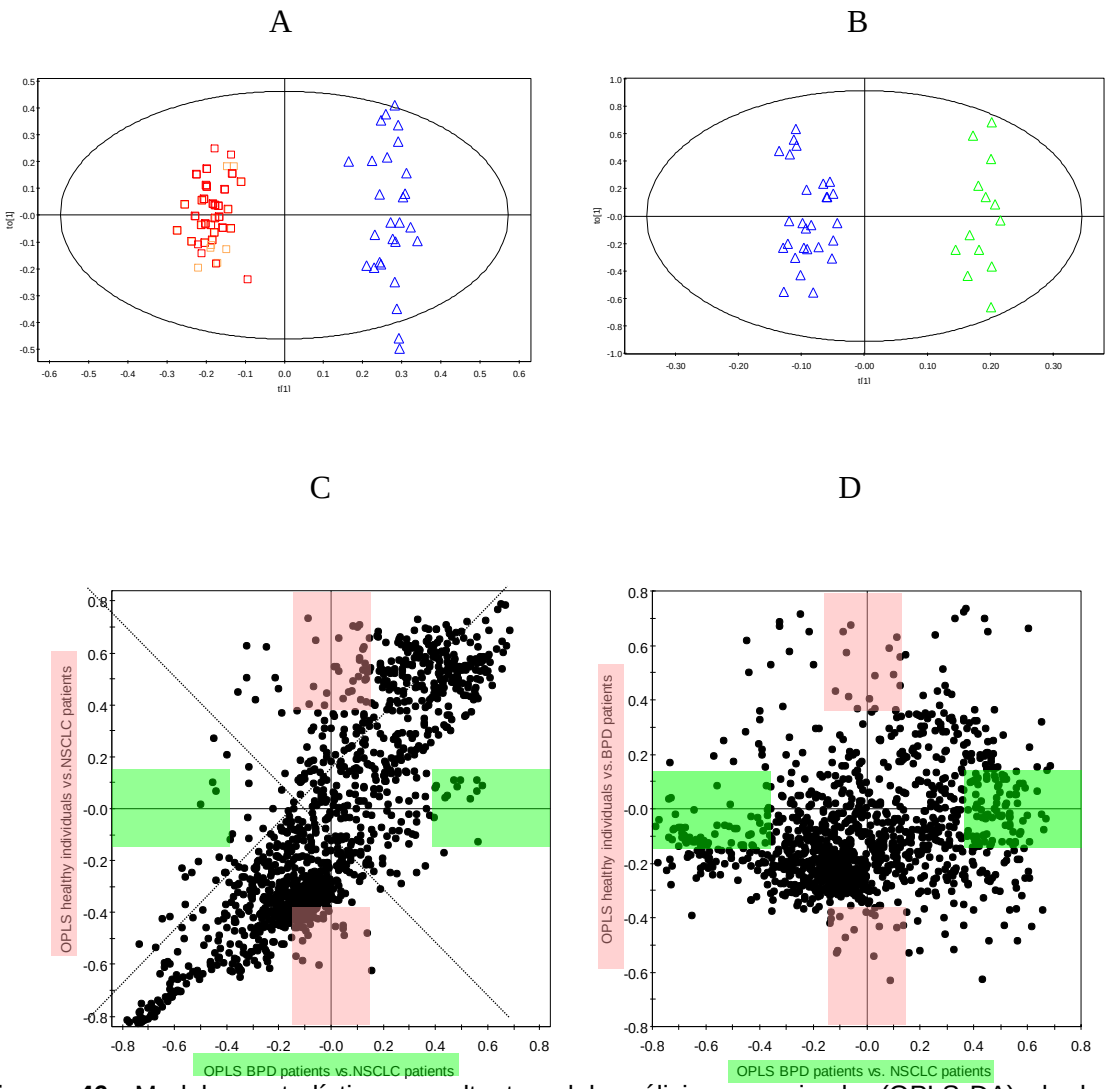


Figura 46: Modelos estadísticos resultantes del análisis supervisado (OPLS-DA) de los espectros ^{1}H-RMN CPMG obtenidos (A) para muestras de suero de pacientes diagnosticados con EPB (Δ) frente a pacientes con CPNM (□) y (B) para muestras de suero de pacientes diagnosticados con EPB frente a individuos sanos (Δ). *SUS-plots* correspondientes a la comparación de los modelos OPLS-DA entre (C) pacientes con CPNM y pacientes con EPB frente a individuos sanos y pacientes con CPNM y entre (D) pacientes con EPB y pacientes con CPNM frente a individuos sanos y pacientes con EPB. Los cuadrados de color rosa y/o verde engloban las señales específicas del modelo marcado en ese mismo color.

A partir de estos modelos se generaron los correspondientes *SUS-plot.* Mediante el análisis de los *SUS-plot* es posible identificar cuáles son las diferencias, únicas y/o comunes, en la comparación entre dos modelos OPLS-DA. En la gráfica, el eje de ordenadas y de abscisas representa el coeficiente de correlación de cada variable entre los modelos que se comparan. Los extremos de la diagonal muestran las variables comunes a ambos modelos, que pueden variar en la misma dirección o en dirección opuesta, y las variables que se encuentran en el eje vertical y horizontal son variables específicas de cada modelo, respectivamente. El análisis de estos modelos permitió identificar las regiones del espectro que contribuían de forma significativa a la

discriminación de los distintos grupos, así como aquellas que eran comunes o específicas de los distintos modelos.

De manera general, la mayoría de las regiones que son relevantes en los modelos en los que se comparan los perfiles metabolómicos de individuos sanos frente a pacientes con CPNM, lo son también en la comparación entre los perfiles de los individuos con EPB y los pacientes con CPNM (**Figura 46C**). Cuando se compara el modelo en el que se representa los perfiles de los individuos con EPB frente a pacientes con CPNM con el modelo donde se representa los perfiles metabólicos de individuos sanos frente a los perfiles de los individuos con EPB, se observa un mayor número de regiones que únicamente son relevantes para cada modelo (**Figura 46D**). Estos resultados revelan que el perfil metabolómico de las muestras de los pacientes con EPB es más similar al perfil metabolómico de los individuos sanos que al de los pacientes con CPNM.

En base a estos resultados, se procedió a la identificación de los metabolitos más significativos que contribuyen a la discriminación entre individuos con EPB frente a individuos sanos y frente a pacientes con CPNM. Este análisis reveló que el perfil metabolómico del suero de estos pacientes se caracterizaba por niveles elevados de metanol y niveles bajos de lactato, estadísticamente significativos, cuando se comparaba con el perfil metabolómico del suero de los individuos sanos. Además, estos pacientes presentaban niveles más elevados de metanol, colina y LDL/VLDL, y niveles bajos de lactato y glucosa, estadísticamente significativos, cuando se comparaban con los niveles de estos metabolitos en pacientes con CPNM (**Tabla 17**).

Tabla 17: Variación de la media de las intensidades de los metabolitos más relevantes, implicados en la discriminación entre pacientes con EPB y pacientes con CPNM o individuos sanos.

Metabolitos	**δ ^{1}H (ppm)[a]**	**Individuos sanos *vs*. EPB**		**CPNM *vs*. EPB**	
		% variación	***p*-valor[b]**	**% variación**	***p*-valor[b]**
HDL (CH_3)	0.85-0.79	-6.55	0.3602	-9.37	0.1073
VLDL (CH_3)	0.91-0.85	-0.19	0.7539	-16.47	0.0002*
Leucina/Isoleucina	0.97-0.91	2.52	0.9319	8.54	0.0360*
3-hidroxibutirato	1.20-1.18	-2.98	0.5490	20.64	0.0735
LDL/VLDL $(CH_2)_n$	1.31-1.20	1.61	0.9319	-23.28	0.0000*
Desconocido1	1.40-1.36	34.21	0.1424	-7.80	0.0676
Ácido Adípico	1.60-1.52	4.30	0.8867	-28.57	0.0001*
Acetato	1.91-1.90	-15.53	0.7979	6.96	0.0078*
Lípidos (CH_2-C=C)	2.02-1.93	-6.28	0.1424	-15.00	0.0000*
N-Acetil-cisteína	2.05-2.02	-8.02	0.0392*	-6.80	0.0152*
Lípidos (CH_2-CO)	2.26-2.19	4.63	0.8867	-12.33	0.1674
Glutamato	2.36-2.33	-7.84	0.2175	43.35	0.0000*
Glutamina	2.47-2.41	-15.34	0.1065	-14.08	0.0136*
Lípidos (CH=CH-CH_2-CH=CH)	2.79-2.68	-0.06	0.6077	-15.40	0.0000*
Colina –$N(CH_3)_3^+$	3.21-3.18	-3.46	0.8197	-14.45	0.0461*
Prolina	3.30-3.34	-27.10	0.0130*	-15.46	0.0434*
Metanol	3.36-3.35	-45.45	0.0010*	-34.12	0.0002*
Desconocido 2	3.55-3.54	-10.03	0.3166	17.93	0.0091*
Desconocido 3	3.58-3.56	-13.86	0.2765	9.10	0.1714
Valina	3.61-3.59	-11.64	0.2068	9.14	0.2150
Glicerol	3.80-3.78	-7.20	0.4406	17.94	0.0027*
Creatina	3.92-3.91	-12.39	0.2175	22.52	0.0005*
Creatinina	4.04-4.03	-0.24	0.6898	-24.18	0.0000*
Myo-inositol	4.07-4.05	10.28	0.4578	-33.58	0.0000*
Lactato	4.13-4.08	49.44	0.0002*	16.92	0.0136*
Treonina	4.30-4.21	4.85	0.5878	-26.09	0.0000*
Glucosa	5.24-5.21	-5.59	0.2639	42.78	0.0000*
Lípidos (CH=CH)	5.37-5.24	2.23	0.3602	-21.01	0.0000*

[a] Desplazamiento químico usado para la cuantificación; [b]*p*-valor calculado usando el test de la U Mann–Whitney; *Estadísticamente significativo (P<0.05)

3.5 Interpretación biológica

Los resultados de este estudio muestran cómo una combinación especifica de cinco metabolitos, obtenidos mediante el análisis por regresión logística, es capaz de predecir la clasificación de muestras de suero de individuos sanos y de pacientes con CPNM, procedentes de un set de muestras independiente, con una sensibilidad del 77,5% y una especificidad del 76,9%. Estos resultados revelan la existencia de un perfil metabolómico característico para pacientes con CPNM, individuos sanos y pacientes con EPB.

Los resultados del estudio previo realizado por nuestro grupo de investigación revelaron que el perfil metabólico en suero de pacientes con CPNM está caracterizado por un conjunto de metabolitos que varían de forma muy específica. El perfil metabólico de pacientes diagnosticados con CPNM, comparado con el de individuos sanos, se caracteriza por la presencia de niveles reducidos de treonina, glutamina y colina, y por niveles elevados de metanol y lactato.

En esta segunda fase del estudio se incluyó, dentro del grupo control, un grupo de muestras de individuos con otras patologías pulmonares lo que ha permitido conocer el perfil metabolómico de este grupo de individuos. Esto ha resultado fundamental para demostrar la existencia de diferencias específicas en los niveles de metabolitos para los pacientes con CPNM y diferencias específicas en los niveles de metabolitos para los individuos con EPB. Nuestro análisis reveló diferencias estadísticamente significativas entre el perfil metabolómico de individuos diagnosticados con EPB e individuos sanos ($R^2(Y) = 0.963$; $Q^2(Y) = 0.782$) y entre pacientes con CPNM ($R^2(Y) = 0.972$; $Q^2(Y) = 0.856$).

Nuestros resultados concuerdan con un estudio reciente de metabolómica por RMN, donde se comparan muestras de suero de pacientes con CPNM y pacientes con enfermedad pulmonar obstructiva crónica (EPOC) (Deja *et al*, 2014). Deja y colaboradores observaron que el perfil metabólico del suero de los pacientes con CPNM, comparado con pacientes diagnosticados de EPOC, estaba caracterizado por niveles altos de lactato y niveles bajos de metanol**.** Nuestros resultados coinciden con estos al comparar el perfil metabolómico del suero de individuos con EPB frente a pacientes con CPNM.

Por el contrario, ellos observaron niveles elevados de colina en el suero de pacientes con CPNM, mientras que nosotros observamos niveles reducidos de este metabolito en los pacientes con CPNM. Los resultados de nuestro estudio, respecto a

la colina, concuerdan con lo encontrado en otros análisis de muestras de tejido procedente de tumores con cáncer de pulmón (Rocha *et al*, 2010).

La colina es un metabolito que tiene un papel como precursor de fosfolípidos de membrana. Los niveles de colina encontrados en el suero de los individuos con EPB varían en la misma dirección que en el suero de los individuos sanos. Ambos grupos presentan niveles superiores a los pacientes con CPNM. Los niveles bajos de colina encontrados en el suero de los pacientes con CPNM podrían estar asociados con un incremento en la demanda de este metabolito por la células tumorales debido a su alta proliferación (Banez-Coronel et al, 2008; Gallego-Ortega et al, 2009). Estos resultados, coinciden con los publicados por Rocha y colaboradores, que observaron una mayor presencia, en relación a los niveles encontrados en individuos sanos, de lípidos, así como de colina, glicerofosfocolina, y fosfocolina en tejido tumoral de pacientes con CPNM (Rocha et al, 2010). Por otro lado, las variaciones en los niveles de lípidos en pacientes oncológicos han sido previamente descritas asociándose a un incremento del consumo, por parte de la células tumorales, de colesterol, componente esencial de las membranas celulares (Puchades-Carrasco *et al*, 2013).

En relación a la composición aminoacídica, nuestros resultados coinciden con los obtenidos por Deja y colaboradores en un estudio en el que observaron que los pacientes con EPOC presentaban niveles inferiores de valina, leucina e isoleucina y niveles más elevados de glutamina, cuando se comparaban con el suero de los pacientes con CPNM (Deja *et al*, 2014). Además, en nuestro estudio también se observaron niveles más elevados de treonina en el suero de los individuos con EPB en comparación con el suero de los pacientes con CPNM. Tanto los individuos sanos como los individuos con EPB, presentaron niveles más elevados de este metabolito en comparación con los pacientes con CPNM.

Las alteraciones en la composición aminoacídica en el suero de pacientes con cáncer ha sido observada en otros estudios previos, probablemente reflejo del estado hipermetabólico y del incremento de la demanda de aminoácidos durante el desarrollo del tumor (Lai *et al*, 2005; Maeda *et al*, 2010; Pisters & Pearlstone, 1993). El descenso de los niveles en suero de treonina e histidina observados en los pacientes con CPNM cuando se comparan con individuos sanos, podría ser consecuencia de la sobreregulación del sistema glicina/serina/treonina y de la vía metabólica de las pirimidinas, respectivamente. Estas características han sido previamente descritas como características metabólicas específicas en células iniciadoras de tumores en CPNM (Zhang *et al*, 2012b).

El descenso observado en los niveles de glutamina en los pacientes con CPNM, que coincide con lo observado en otros pacientes oncológicos (Gao et al, 2008; Urayama et al, 2010; Zira et al, 2010), se ha asociado al aumento de la actividad metabólica (glucolisis y glutaminolisis) derivada de las condiciones de hipoxia e hipermetabolismo observadas en el entorno tumoral (Eigenbrodt et al, 1998). Un estudio reciente de proteómica, realizado con células derivadas de adenocarcinoma pancreático, puso de manifiesto el papel fundamental que juega la glutamina, bajo esas condiciones, como fuente de nitrógeno para la síntesis de nucleótidos y aminoácidos (Zhou et al, 2012). En estos estudios también se observó una disminución en los niveles de la alanina aminotransferasa, la enzima que cataliza la producción de alanina a partir de glutamato. Este hecho, unido a la posible hidrólisis de la glutamina para la generación de glutamato y amonio, que contribuiría a equilibrar el pH en las células tumorales contrarrestando el exceso de lactato producido como consecuencia del efecto Warburg, podría explicar los niveles de glutamato elevados que se observan en el grupo de pacientes diagnosticados con CPNM incluidos en nuestro estudio.

3.6 Relevancia de los resultados

Este estudio representa un gran avance en la identificación de biomarcadores de utilidad clínica en el diagnóstico temprano de CPNM. Estos resultados proporcionan un mejor conocimiento sobre la fisiopatología de esta enfermedad y la caracterización de perfiles metabólicos capaces de discriminar entre paciente e individuos sanos. Nuestro trabajo representa el primer estudio en este área que engloba un número de muestras significativo y que es validado con un set de muestras independiente.

La validación de los estudios de metabolómica es un avance importante en el contexto de esta tecnología, que permite dar solidez y validez a los estudios. En este contexto, las herramientas bioinformáticas aplicadas en la clínica se han convertido en un elemento esencial en la investigación traslacional (Baumgartner et al, 2011). En este trabajo se empleó por primera vez la herramienta ComBat para integrar el análisis de datos metabolómicos procedentes de muestras medidas en distintos lotes. Esta herramienta ha demostrado ser una estrategia muy útil para la corrección del efecto *batch* que se da entre los distintos sets de medida en la validación externa de estudios metabolómicos ampliando el tamaño muestral con muestras de diferentes lotes.

Existen múltiples métodos estadísticos utilizados en estudios ómicos que permiten generar modelos predictivos, con el fin de predecir la probabilidad de que ocurra un evento esperado o no (p.ej., sano/enfermo) (Bahado-Singh et al, 2015; Li et al, 2016).

Algunos ejemplos son los algoritmos de máquinas de soporte vectorial (Support Vector Machines), regresión de mínimos cuadrados parciales (PLS), la regresión logística, etc. (Xi et al, 2014). En nuestro estudio se ha validado la utilidad clínica de una ecuación matemática, usando la regresión logística, que permite predecir si una muestra sin diagnosticar pertenece al grupo de enfermos o sanos, con una especificidad y sensibilidad elevada. La ecuación de regresión logística es un método sencillo que permite, incluyendo un número reducido de metabolitos, adaptar el análisis a un método analítico alternativo a la metabolómica, abordable en la práctica clínica habitual.

V. CONCLUSIONES

I. La metabolómica por RMN es una técnica accesible y no invasiva que permite caracterizar el perfil metabolómico de muestras biológicas y la identificación y cuantificación de biomarcadores.

II. Las variaciones en el procesado y/o almacenamiento de las muestras biológicas en fase preanalítica se traduce en alteraciones en los niveles de algunos metabolitos que pueden condicionar los resultados de los estudios de investigación. La fase preanalítica es una etapa crucial que debe estar controlada para poder obtener muestras de calidad en los estudios de investigación.

III. El análisis de los espectros de orina de pacientes con CaP revela, cuando se compara con muestras de orina de individuos diagnosticados con HBP, un perfil metabolómico caracterizado por un aumento en los niveles de aminoácidos de cadena ramificada, glutamato y pseudouridina, y niveles disminuidos de glicina, dimetilglicina, fumarato, 4-imidazolacetato y un metabolito desconocido.

IV. La metabolómica es una técnica de gran utilidad para discriminar los perfiles metabolómicos específicos de pacientes con EPB, pacientes con CPNM e individuos sanos. Los pacientes con EPB presentan niveles más elevados de metanol y más reducidos de lactato, en comparación con los otros dos grupos. La composición aminoacídica y lipídica de estos pacientes es similar a la de los individuos sanos.

V. La validación externa de una ecuación de probabilidad basada en los niveles de 5 metabolitos (lactato, treonina, glutamina, colina y metanol) permite clasificar, con una especificidad del 77,5% y una sensibilidad del 76,9%, muestras pertenecientes a individuos sanos, pacientes con EPB y pacientes con CPNM.

VI. BIBLIOGRAFIA

Aggio RB, Ruggiero K, Villas-Bôas SG (2010) Pathway Activity Profiling (PAPi): from the metabolite profile to the metabolic pathway activity. Bioinformatics 26(23): 2969-2976

Aldridge BB, Rhee KY (2014) Microbial metabolomics: innovation, application, insight. Current opinion in microbiology 19: 90-96

Alonso A, Rodriguez MA, Vinaixa M, Tortosa R, Correig X, Julia A, Marsal S (2014) Focus: a robust workflow for one-dimensional NMR spectral analysis. Analytical chemistry 86(2): 1160-9

Andersen CM, Bro R (2010) Variable selection in regression—a tutorial. Journal of chemometrics 24(11- 12): 728-737

Anderson D, Kodukula K (2014) Biomarkers in pharmacology and drug discovery. Biochemical pharmacology 87(1): 172-188

Anderson NL, Anderson NG (2002) The human plasma proteome: history, character, and diagnostic prospects. Molecular & cellular proteomics : MCP 1(11): 845-67

Armitage EG, Barbas C (2014) Metabolomics in cancer biomarker discovery: current trends and future perspectives. Journal of pharmaceutical and biomedical analysis 87: 1-11

Astarita G, Langridge J (2013) An emerging role for metabolomics in nutrition science. Journal of nutrigenetics and nutrigenomics 6(4-5): 181-200

Bahado-Singh RO, Syngelaki A, Akolekar R, Mandal R, Bjondahl TC, Han B, Dong E, Bauer S, Alpay-Savasan Z, Graham S, Turkoglu O, Wishart DS, Nicolaides KH (2015) Validation of metabolomic models for prediction of early-onset preeclampsia. American journal of obstetrics and gynecology 213(4): 530 e1-530 e10

Banez-Coronel M, Ramirez de Molina A, Rodriguez-Gonzalez A, Sarmentero J, Ramos MA, Garcia-Cabezas MA, Garcia-Oroz L, Lacal JC (2008) Choline kinase alpha depletion selectively kills tumoral cells. Current cancer drug targets 8(8): 709-19

Bangma CH, Grobbee D, Schröder F (1995) Volume adjustment for intermediate prostate-specific antigen values in a screening population. European Journal of Cancer 31(1): 12-14

Baumgartner C, Osl M, Netzer M, Baumgartner D (2011) Bioinformatic-driven search for metabolic biomarkers in disease. Journal of clinical bioinformatics 1(1): 1

Baynes J, Dominiczak M (2010) Medical Biochemistry, 4th Edition. Elsevier, Amsterdam

Beckonert O, Bollard ME, Ebbels TM, Keun HC, Antti H, Holmes E, Lindon JC, Nicholson JK (2003) NMR-based metabonomic toxicity classification: hierarchical cluster analysis and k-nearest-neighbour approaches. Analytica chimica acta 490(1): 3-15

Beckonert O, Keun HC, Ebbels TM, Bundy J, Holmes E, Lindon JC, Nicholson JK (2007) Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. Nature protocols 2(11): 2692-703

Beger RD (2013) A review of applications of metabolomics in cancer. Metabolites 3(3): 552-574

Bernini P, Bertini I, Luchinat C, Nincheri P, Staderini S, Turano P (2011) Standard operating procedures for pre-analytical handling of blood and urine for metabolomic studies and biobanks. Journal of biomolecular NMR 49(3-4): 231-43

Betsou F, Lehmann S, Ashton G, Barnes M, Benson EE, Coppola D, DeSouza Y, Eliason J, Glazer B, Guadagni F, Harding K, Horsfall DJ, Kleeberger C, Nanni U, Prasad A, Shea K, Skubitz A, Somiari S, Gunter E (2010) Standard preanalytical coding for biospecimens: defining the sample PREanalytical code. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 19(4): 1004-11

Bianchi F, Dugheri S, Musci M, Bonacchi A, Salvadori E, Arcangeli G, Cupelli V, Lanciotti M, Masieri L, Serni S, Carini M, Careri M, Mangia A (2011) Fully automated solid-phase microextraction-fast gas chromatography-mass spectrometry method using a new ionic liquid column for high-throughput analysis of sarcosine and N-ethylglycine in human urine and urinary sediments. Analytica chimica acta 707(1-2): 197-203

BMRB - Biological Magnetic Resonance Bank. www.bmrb.wisc.edu

Boroughs LK, DeBerardinis RJ (2015) Metabolic pathways promoting cancer cell survival and growth. Nature cell biology 17(4): 351-359

Bouatra S, Aziat F, Mandal R, Guo AC, Wilson MR, Knox C, Bjorndahl TC, Krishnamurthy R, Saleem F, Liu P, Dame ZT, Poelzer J, Huynh J, Yallou FS, Psychogios N, Dong E, Bogumil R, Roehring C, Wishart DS (2013) The human urine metabolome. PloS one 8(9): e73076

Carraro P, Zago T, Plebani M (2012) Exploring the initial steps of the testing process: frequency and nature of pre-preanalytic errors. Clinical chemistry 58(3): 638-42

Carrola J, Rocha CM, Barros AS, Gil AM, Goodfellow BJ, Carreira IM, Bernardo J, Gomes A, Sousa V, Carvalho L, Duarte IF (2011) Metabolic signatures of lung cancer in biofluids: NMR-based metabonomics of urine. Journal of proteome research 10(1): 221-230

Chen JQ, Russo J (2012) Dysregulation of glucose transport, glycolysis, TCA cycle and glutaminolysis by oncogenes and tumor suppressors in cancer cells. Biochimica et biophysica acta 1826(2): 370-84

Chikayama E, Sekiyama Y, Okamoto M, Nakanishi Y, Tsuboi Y, Akiyama K, Saito K, Shinozaki K, Kikuchi J (2010) Statistical indices for simultaneous large-scale metabolite detections for a single NMR spectrum. Analytical chemistry 82(5): 1653-8

Claus SP, Swann JR (2013) Nutrimetabonomics:applications for nutritional sciences, with specific reference to gut microbial interactions. Annual review of food science and technology 4: 381-99

Craig A, Cloarec O, Holmes E, Nicholson JK, Lindon JC (2006) Scaling and normalization effects in NMR spectroscopic metabonomic data sets. Analytical chemistry 78(7): 2262-7

Čuperlović-Culf M (2012) NMR metabolomics in cancer research: Elsevier

DeBerardinis RJ, Lum JJ, Hatzivassiliou G, Thompson CB (2008) The biology of cancer: metabolic reprogramming fuels cell growth and proliferation. Cell metabolism 7(1): 11-20

Deja S, Porebska I, Kowal A, Zabek A, Barg W, Pawelczyk K, Stanimirova I, Daszykowski M, Korzeniewska A, Jankowska R, Mlynarz P (2014) Metabolomics provide new insights on lung cancer staging and discrimination from chronic obstructive pulmonary disease. Journal of pharmaceutical and biomedical analysis 100: 369-380

Di Anibal CV, Callao MP, Ruisánchez I (2011) 1 H NMR variable selection approaches for classification. A case study: The determination of adulterated foodstuffs. Talanta 86: 316-323

Diaz SO, Barros AS, Goodfellow BJ, Duarte IF, Galhano E, Pita C, Almeida Mdo C, Carreira IM, Gil AM (2013) Second trimester maternal urine for the diagnosis of trisomy 21 and prediction of poor pregnancy outcomes. Journal of proteome research 12(6): 2946-57

Dieterle F, Ross A, Schlotterbeck G, Senn H (2006) Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. Analytical chemistry 78(13): 4281-90

Dona AC, Jimenez B, Schafer H, Humpfer E, Spraul M, Lewis MR, Pearce JT, Holmes E, Lindon JC, Nicholson JK (2014) Precision high-throughput proton NMR spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping. Analytical chemistry 86(19): 9887-94

Draisma G, Boer R, Otto SJ, van der Cruijsen IW, Damhuis RA, Schroder FH, de Koning HJ (2003) Lead times and overdetection due to prostate-specific antigen screening: estimates from the European Randomized Study of Screening for Prostate Cancer. Journal of the National Cancer Institute 95(12): 868-78

Dudley E, Yousef M, Wang Y, Griffiths WJ (2010) Targeted metabolomics and mass spectrometry. Advances in protein chemistry and structural biology 80: 45-83

Eigenbrodt E, Kallinowski F, Ott M, Mazurek S, Vaupel P (1998) Pyruvate kinase and the interaction of amino acid and carbohydrate metabolism in solid tumors. Anticancer research 18(5A): 3267-74

Eriksson L, Byrne T, Johansson E, Trygg J, Vikström C (2013) Multi-and megavariate data analysis basic principles and applications: Umetrics Academy

Fan J, Hong J, Hu J-D, Chen J-L (2012) Ion chromatography based urine amino acid profiling applied for diagnosis of gastric cancer. Gastroenterology research and practice 2012

Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F (2015) Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. International journal of cancer 136(5): E359-E386

Fiehn O (2002) Metabolomics–the link between genotypes and phenotypes. Plant molecular biology 48(1-2): 155-171

Fliniaux O, Gaillard G, Lion A, Cailleu D, Mesnard F, Betsou F (2011) Influence of common preanalytical variations on the metabolic profile of serum samples in biobanks. Journal of biomolecular NMR 51(4): 457-65

Fraser C (2001) Biological variation: from principles to practice: AACCPress

Gallego-Ortega D, Ramirez de Molina A, Ramos MA, Valdes-Mora F, Barderas MG, Sarmentero-Estrada J, Lacal JC (2009) Differential role of human choline kinase alpha and beta enzymes in lipid metabolism: implications in cancer onset and treatment. PloS one 4(11): e7819

Gann PH, Hennekens CH, Stampfer MJ (1995) A prospective evaluation of plasma prostate-specific antigen for detection of prostatic cancer. Jama 273(4): 289-294

Gao H, Dong B, Liu X, Xuan H, Huang Y, Lin D (2008) Metabonomic profiling of renal cell carcinoma: high-resolution proton nuclear magnetic resonance spectroscopy of human serum with multivariate data analysis. Analytica chimica acta 624(2): 269-77

Gao Y (2013) Urine-an untapped goldmine for biomarker discovery? Science China Life sciences 56(12): 1145-6

Gika HG, Theodoridis GA, Wingate JE, Wilson ID (2007) Within-day reproducibility of an HPLC-MS-based method for metabonomic analysis: application to human urine. Journal of proteome research 6(8): 3291-303

Giskeødegård GF, Bertilsson H, Selnæs KM, Wright AJ, Bathen TF, Viset T, Halgunset J, Angelsen A, Gribbestad IS, Tessem M-B (2013) Spermine and citrate as metabolic biomarkers for assessing prostate cancer aggressiveness. PloS one 8(4): e62375

Giskeødegård GF, Hansen AF, Bertilsson H, Gonzalez SV, Kristiansen KA, Bruheim P, Mjos SA, Angelsen A, Bathen TF, Tessem MB (2015) Metabolic markers in blood can separate prostate cancer from benign prostatic hyperplasia. British journal of cancer 113(12): 1712-9

Gomez-Casati DF, Zanor MI, Busi MV (2013) Metabolomics in plants and humans: applications in the prevention and diagnosis of diseases. BioMed research international 2013

Greene FL PDL, Fleming I.D, Fritz A.G, Balch C.M, Haller D.G, Morrow, M. (2003) AJCC Cancer Staging Manual,6th edition, . Philadelphia: Lippincott Raven Publishers

Griffin JL, Wang X, Stanley E (2015) Does Our Gut Microbiome Predict Cardiovascular Risk? Circulation: Cardiovascular Genetics 8(1): 187-191

Gu H, Pan Z, Xi B, Asiago V, Musselman B, Raftery D (2011) Principal component directed partial least squares analysis for combining nuclear magnetic resonance and mass spectrometry data in metabolomics: application to the detection of breast cancer. Analytica chimica acta 686(1-2): 57-63

Hebels DG, Georgiadis P, Keun HC, Athersuch TJ, Vineis P, Vermeulen R, Portengen L, Bergdahl IA, Hallmans G, Palli D (2013) Performance in omics analyses of blood samples in long-term storage: opportunities for the exploitation of existing biobanks in environmental health research. Environmental Health Perspectives (Online) 121(4): 480

Hensley CT, Wasti AT, DeBerardinis RJ (2013) Glutamine and cancer: cell biology, physiology, and clinical opportunities. The Journal of clinical investigation 123(9): 3678-3684

Hibbert DB (1993) Genetic algorithms in chemistry. Chemometrics and Intelligent Laboratory Systems 19(3): 277-293

Hoult D (1976) Solvent peak saturation with single phase and quadrature Fourier transformation. Journal of Magnetic Resonance (1969) 21(2): 337-347

Ilic D, Neuberger MM, Djulbegovic M, Dahm P (2013) Screening for prostate cancer. The Cochrane database of systematic reviews 1: CD004720

Issaq HJ, Veenstra TD (2011) Is sarcosine a biomarker for prostate cancer? Journal of separation science 34(24): 3619-21

Jacobs DM, Spiesser L, Garnier M, de Roo N, van Dorsten F, Hollebrands B, van Velzen E, Draijer R, van Duynhoven J (2012) SPE-NMR metabolite sub-profiling of urine. Analytical and bioanalytical chemistry 404(8): 2349-61

Jain M, Kennedy AD, Elsea SH, Miller MJ (2017) Analytes related to erythrocyte metabolism are reliable biomarkers for preanalytical error due to delayed plasma processing in metabolomics studies. Clinica chimica acta; international journal of clinical chemistry 466: 105-111

Jantus-Lewintre E, Usó M, Sanmartín E, Camps C (2012) Update on biomarkers for the detection of lung cancer. Lung Cancer: Targets and Therapy 3: 21-29

Jentzmik F, Stephan C, Miller K, Schrader M, Erbersdobler A, Kristiansen G, Lein M, Jung K (2010) Sarcosine in urine after digital rectal examination fails as a marker in prostate cancer detection and identification of aggressive tumours. European urology 58(1): 12-8; discussion 20-1

Jiang Y, Cheng X, Wang C, Ma Y (2010) Quantitative determination of sarcosine and related compounds in urinary samples by liquid chromatography with tandem mass spectrometry. Analytical chemistry 82(21): 9022-9027

Jimenez B, Mirnezami R, Kinross J, Cloarec O, Keun HC, Holmes E, Goldin RD, Ziprin P, Darzi A, Nicholson JK (2013) 1H HR-MAS NMR spectroscopy of tumor-induced local metabolic "field-effects" enables colorectal cancer staging and prognostication. Journal of proteome research 12(2): 959-68

Jobard E, Pontoizeau C, Blaise BJ, Bachelot T, Elena-Herrmann B, Tredan O (2014) A serum nuclear magnetic resonance-based metabolomic signature of advanced metastatic human breast cancer. Cancer letters 343(1): 33-41

Jonsson P, Bruce SJ, Moritz T, Trygg J, Sjostrom M, Plumb R, Granger J, Maibaum E, Nicholson JK, Holmes E, Antti H (2005) Extraction, interpretation and validation of information for comparing samples in metabolic LC/MS data sets. The Analyst 130(5): 701-7

Jordan K, Adkins C, Su L, Halpern E, Mark E, Christiani D, Cheng L (2010) Comparison of squamous cell carcinoma and adenocarcinoma of the lung by metabolomic analysis of tissue–serum pairs. Lung Cancer 68(1): 44-50

Joyce AR, Palsson BO (2006) The model organism as a system: integrating 'omics' data sets. Nature reviews Molecular cell biology 7(3): 198-210

Kamburov A, Stelzl U, Lehrach H, Herwig R (2012) The ConsensusPathDB interaction database: 2013 update. Nucleic acids research: gks1055

Kamlage B, Maldonado SG, Bethan B, Peter E, Schmitz O, Liebenberg V, Schatz P (2014) Quality markers addressing preanalytical variations of blood and plasma processing identified by broad and targeted metabolite profiling. Clinical chemistry 60(2): 399-412

Kanehisa M (2002) The KEGG database. Silico simulation of biological processes 247: 91-103

Ke C, Hou Y, Zhang H, Fan L, Ge T, Guo B, Zhang F, Yang K, Wang J, Lou G, Li K (2015) Large-scale profiling of metabolic dysregulation in ovarian cancer. International journal of cancer Journal international du cancer 136(3): 516-26

Khan AP, Rajendiran TM, Ateeq B, Asangani IA, Athanikar JN, Yocum AK, Mehra R, Siddiqui J, Palapattu G, Wei JT, Michailidis G, Sreekumar A, Chinnaiyan AM (2013) The role of sarcosine metabolism in prostate cancer progression. Neoplasia 15(5): 491-501

Kim J-w, Dang CV (2006) Cancer's molecular sweet tooth and the Warburg effect. Cancer research 66(18): 8927-8930

Kulasingam V, Diamandis EP (2008) Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. Nature clinical practice Oncology 5(10): 588-599

Kumar A, Ernst RR, Wuthrich K (1980) A two-dimensional nuclear Overhauser enhancement (2D NOE) experiment for the elucidation of complete proton-proton cross-relaxation networks in biological macromolecules. Biochemical and biophysical research communications 95(1): 1-6

Kumar D, Gupta A, Mandhani A, Sankhwar SN (2015) Metabolomics-derived prostate cancer biomarkers: fact or fiction? Journal of proteome research 14(3): 1455-64

Lai HS, Lee JC, Lee PH, Wang ST, Chen WJ (2005) Plasma free amino acid profile in cancer patients. Seminars in cancer biology 15(4): 267-76

Lam VW, Poon RT (2008) Role of branched- chain amino acids in management of cirrhosis and hepatocellular carcinoma. Hepatology Research 38(s1)

Lauridsen M, Hansen SH, Jaroszewski JW, Cornett C (2007) Human urine as test material in 1H NMR-based metabonomics: recommendations for sample preparation and storage. Analytical chemistry 79(3): 1181-6

Lawton KA, Berger A, Mitchell M, Milgram KE, Evans AM, Guo L, Hanson RW, Kalhan SC, Ryals JA, Milburn MV (2008) Analysis of the adult human plasma metabolome. Pharmacogenomics 9(4): 383-97

Leardi R (2001) Genetic algorithms in chemometrics and chemistry: a review. Journal of chemometrics 15(7): 559-569

Lehmann R (2015) Preanalytics: what can metabolomics learn from clinical chemistry? Bioanalysis 7(8): 927-30

Lenz EM, Wilson ID (2007) Analytical strategies in metabonomics. Journal of proteome research 6(2): 443-58

Lescuyer P, Hochstrasser D, Rabilloud T (2007) How shall we use the proteomics toolbox for biomarker discovery? Journal of proteome research 6(9): 3371-6

Lewis IA, Schommer SC, Markley JL (2009) rNMR: open source software for identifying and quantifying metabolites in NMR spectra. Magnetic resonance in chemistry : MRC 47 Suppl 1: S123-6

Li Y, Wang L, Ju L, Deng H, Zhang Z, Hou Z, Xie J, Wang Y, Zhang Y (2016) A Systematic Strategy for Screening and Application of Specific Biomarkers in Hepatotoxicity Using Metabolomics Combined With ROC Curves and SVMs. Toxicological sciences : an official journal of the Society of Toxicology 150(2): 390-9

Lindon JC, Holmes E, Nicholson JK (2007) Metabonomics in pharmaceutical R & D. Febs Journal 274(5): 1140-1151

Lindon JC, Nicholson JK, Holmes E (2011) The handbook of metabonomics and metabolomics: Elsevier

Lindon JC, Nicholson JK, Holmes E, Everett JR (2000) Metabonomics: metabolic processes studied by NMR spectroscopy of biofluids. Concepts in Magnetic Resonance 12(5): 289-320

Lindon JC, Nicholson JK, Holmes E, Keun HC, Craig A, Pearce JT, Bruce SJ, Hardy N, Sansone S-A, Antti H (2005) Summary recommendations for standardization and reporting of metabolic analyses. Nature biotechnology 23(7): 833-839

Lippi G, Blanckaert N, Bonini P, Green S, Kitchen S, Palicka V, Vassault AJ, Plebani M (2008) Haemolysis: an overview of the leading cause of unsuitable specimens in clinical laboratories. Clinical chemistry and laboratory medicine 46(6): 764-72

Lippi G, Guidi GC, Mattiuzzi C, Plebani M (2006) Preanalytical variability: the dark side of the moon in laboratory testing. Clinical chemistry and laboratory medicine 44(4): 358-65

Locasale JW (2013) Serine, glycine and one-carbon units: cancer metabolism in full circle. Nature reviews Cancer 13(8): 572-83

Lu X, Zhao X, Bai C, Zhao C, Lu G, Xu G (2008) LC-MS-based metabonomics analysis. Journal of chromatography B, Analytical technologies in the biomedical and life sciences 866(1-2): 64-76

Lucarelli G, Fanelli M, Larocca AM, Germinario CA, Rutigliano M, Vavallo A, Selvaggi FP, Bettocchi C, Battaglia M, Ditonno P (2012) Serum sarcosine increases the accuracy of prostate cancer detection in patients with total serum PSA less than 4.0 ng/ml. The Prostate 72(15): 1611-21

MacIntyre DA, Jimenez B, Lewintre EJ, Martin CR, Schafer H, Ballesteros CG, Mayans JR, Spraul M, Garcia-Conde J, Pineda-Lucena A (2010) Serum metabolome analysis by 1H-NMR reveals differences between chronic lymphocytic leukaemia molecular subgroups. Leukemia 24(4): 788-97

MacKinnon N, Somashekar BS, Tripathi P, Ge W, Rajendiran TM, Chinnaiyan AM, Ramamoorthy A (2013) MetaboID: a graphical user interface package for assignment of 1H NMR spectra of bodyfluids and tissues. J Magn Reson 226: 93-9

Maeda J, Higashiyama M, Imaizumi A, Nakayama T, Yamamoto H, Daimon T, Yamakado M, Imamura F, Kodama K (2010) Possibility of multivariate function composed of plasma amino acid profiles as a novel screening index for non-small cell lung cancer: a case control study. BMC cancer 10: 690

Mamas M, Dunn WB, Neyses L, Goodacre R (2011) The role of metabolites and metabolomics in clinically applicable biomarkers of disease. Archives of toxicology 85(1): 5-17

Manna SK, Patterson AD, Yang Q, Krausz KW, Idle JR, Fornace AJ, Gonzalez FJ (2011) UPLC-MS-based urine metabolomics reveals indole-3-lactic acid and phenyllactic acid as conserved biomarkers for alcohol-induced liver disease in the Ppara-null mouse model. Journal of proteome research 10(9): 4120-33

Manne U, Srivastava RG, Srivastava S (2005) Recent advances in biomarkers for cancer diagnosis and treatment. Drug discovery today 10(14): 965-76

Marrero JA, Su GL, Wei W, Emick D, Conjeevaram HS, Fontana RJ, Lok AS (2003) Des- gamma carboxyprothrombin can differentiate hepatocellular carcinoma from nonmalignant chronic liver disease in american patients. Hepatology 37(5): 1114-1121

Masaki Y, Itoh K, Sawaki T, Karasawa H, Kawanami T, Fukushima T, Kawabata H, Wano Y, Hirose Y, Suzuki T, Sugai S, Umehara H (2006) Urinary pseudouridine in patients with lymphoma: comparison with other clinical parameters. Clinica chimica acta; international journal of clinical chemistry 371(1-2): 148-51

McDunn JE, Li Z, Adam KP, Neri BP, Wolfert RL, Milburn MV, Lotan Y, Wheeler TM (2013) Metabolomic signatures of aggressive prostate cancer. The Prostate 73(14): 1547-60

Mehmood T, Liland KH, Snipen L, Sæbø S (2012) A review of variable selection methods in partial least squares regression. Chemometrics and Intelligent Laboratory Systems 118: 62-69

Meiboom S, Gill D (1958) Modified spin- echo method for measuring nuclear relaxation times. Review of scientific instruments 29(8): 688-691

Miyake M, Gomes Giacoia E, Aguilar Palacios D, Rosser CJ (2012) Sarcosine, a biomarker for prostate cancer: ready for prime time? Biomarkers in medicine 6(4): 513-4

Mondul AM, Moore SC, Weinstein SJ, Karoly ED, Sampson JN, Albanes D (2015) Metabolomic analysis of prostate cancer risk in a prospective cohort: The alpha-tocolpherol, beta-carotene cancer prevention (ATBC) study. International journal of cancer Journal international du cancer 137(9): 2124-32

Mottet N, Bellmunt J, Briers E, Bolla M, Cornford P, De Santis M, Henry A, Joniau S, Lam T, Mason M (2016) EAU-ESTRO-SIOG.

Nicholson JK, Foxall PJ, Spraul M, Farrant RD, Lindon JC (1995) 750 MHz 1H and 1H-13C NMR spectroscopy of human blood plasma. Analytical chemistry 67(5): 793-811

Nicholson JK, Wilson ID, Lindon JC (2011) Pharmacometabonomics as an effector for personalized medicine. Pharmacogenomics 12(1): 103-11

Nobakht M. Gh BF, Aliannejad R, Rezaei-Tavirani M, Taheri S, Oskouie AA (2015) The metabolomics of airway diseases, including COPD, asthma and cystic fibrosis. Biomarkers 20(1): 5-16

Nørgaard L, Saudland A, Wagner J, Nielsen J, Munck L, Engelsen S (2000) Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. Applied Spectroscopy 54(3): 413-419

O'Connell TM (2013) The complex role of branched chain amino acids in diabetes and cancer. Metabolites 3(4): 931-945

Pan Z, Gu H, Talaty N, Chen H, Shanaiah N, Hainline BE, Cooks RG, Raftery D (2007) Principal component analysis of urine metabolites detected by NMR and DESI-MS in patients with inborn errors of metabolism. Analytical and bioanalytical chemistry 387(2): 539-49

Parker HS, Bravo HC, Leek JT (2014) Removing batch effects for prediction problems with frozen surrogate variable analysis. PeerJ 2: e561

Patti G, Yanes O, Siuzdak G (2012) Metabolomics: the apogee of the omic triology. Nature reviews Molecular cell biology 13(4): 263-269

Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC (2002) Use of proteomic patterns in serum to identify ovarian cancer. The lancet 359(9306): 572-577

Pinto J, Domingues MR, Galhano E, Pita C, Almeida Mdo C, Carreira IM, Gil AM (2014) Human plasma stability during handling and storage: impact on NMR metabolomics. The Analyst 139(5): 1168-77

Pisters PW, Pearlstone DB (1993) Protein and amino acid metabolism in cancer cachexia: investigative techniques and therapeutic interventions. Critical reviews in clinical laboratory sciences 30(3): 223-72

Ploussard G, De La Taille A (2010) Urine biomarkers in prostate cancer. Nature Reviews Urology 7(2): 101-109

Psychogios N, Hau DD, Peng J, Guo AC, Mandal R, Bouatra S, Sinelnikov I, Krishnamurthy R, Eisner R, Gautam B (2011) The human serum metabolome. PloS one 6(2): e16957

Puchades-Carrasco L. Aplicaciones de la RMN a la identificación de nuevos biomarcadores de utilidad clínica en oncología. Tesis Doctoral 2013

Puchades-Carrasco L, Lecumberri R, Martinez-Lopez J, Lahuerta JJ, Mateos MV, Prosper F, San-Miguel JF, Pineda-Lucena A (2013) Multiple myeloma patients have a specific serum metabolomic profile that changes after achieving complete remission. Clinical cancer research : an official journal of the American Association for Cancer Research 19(17): 4770-4779

Puchades-Carrasco L, Palomino-Schatzlein M, Perez-Rambla C, Pineda-Lucena A (2015) Bioinformatics tools for the analysis of NMR metabolomics studies focused on the identification of clinically relevant biomarkers. Briefings in bioinformatics

Puchades-Carrasco L, Pineda-Lucena A (2015) Metabolomics in pharmaceutical research and development. Current opinion in biotechnology 35: 73-77

Quintás G, Portillo N, García-Cañaveras JC, Castell JV, Ferrer A, Lahoz A (2012) Chemometric approaches to improve PLSDA model outcome for predicting human non-alcoholic fatty liver disease using UPLC-MS as a metabolic profiling tool. Metabolomics 8(1): 86-98

Rabbani F, Stroumbakis N, Kava BR, Cookson MS, Fair WR (1998) Incidence and clinical significance of false-negative sextant prostate biopsies. The Journal of urology 159(4): 1247-50

Rasmuson T, Bjork GR (1995) Urinary excretion of pseudouridine and prognosis of patients with malignant lymphoma. Acta Oncol 34(1): 61-7

Rocha CM, Barros AS, Gil AM, Goodfellow BJ, Humpfer E, Spraul M, Carreira IM, Melo JB, Bernardo J, Gomes A, Sousa V, Carvalho L, Duarte IF (2010) Metabolic profiling of human lung cancer tissue by 1H high resolution magic angle spinning (HRMAS) NMR spectroscopy. Journal of proteome research 9(1): 319-32

Rocha CM, Carrola J, Barros AS, Gil AM, Goodfellow BJ, Carreira IM, Bernardo J, Gomes A, Sousa V, Carvalho L (2011) Metabolic signatures of lung cancer in biofluids: NMR-based metabonomics of blood plasma. Journal of proteome research 10(9): 4314-4324

Ross A, Schlotterbeck G, Dieterle F, Senn H (2007) NMR spectroscopy techniques for application to metabonomics: Elsevier, Amsterdam

Saccenti E, Hoefsloot HC, Smilde AK, Westerhuis JA, Hendriks MM (2014) Reflections on univariate and multivariate analysis of metabolomics data. Metabolomics 10(3): 0

Saigusa D, Okamura Y, Motoike IN, Katoh Y, Kurosawa Y, Saijyo R, Koshiba S, Yasuda J, Motohashi H, Sugawara J, Tanabe O, Kinoshita K, Yamamoto M (2016) Establishment of Protocols for Global Metabolomics by LC-MS for Biomarker Discovery. PloS one 11(8): e0160555

Salek RM, Maguire ML, Bentley E, Rubtsov DV, Hough T, Cheeseman M, Nunez D, Sweatman BC, Haselden JN, Cox R (2007) A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat, and human. Physiological genomics 29(2): 99-108

Schoenfield L, Jones JS, Zippe CD, Reuther AM, Klein E, Zhou M, Magi-Galluzzi C (2007) The incidence of high-grade prostatic intraepithelial neoplasia and atypical glands suspicious for carcinoma on first-time saturation needle biopsy, and the subsequent risk of cancer. BJU international 99(4): 770-4

Schramm G, Surmann EM, Wiesberg S, Oswald M, Reinelt G, Eils R, Konig R (2010) Analyzing the regulation of metabolic pathways in human breast cancer. BMC medical genomics 3: 39

Semenza GL (2010) HIF-1: upstream and downstream of cancer metabolism. Current opinion in genetics & development 20(1): 51-56

SEOM (2014) Las Cifras del Cáncer en España 2014 - SEOM.

Serkova NJ, Gamito EJ, Jones RH, O'Donnell C, Brown JL, Green S, Sullivan H, Hedlund T, Crawford ED (2008) The metabolites citrate, myo-inositol, and spermine are potential age-independent markers of prostate cancer in human expressed prostatic secretions. The Prostate 68(6): 620-8

Shi H, Li X, Zhang Q, Yang H, Zhang X (2016) Discovery of urine biomarkers for bladder cancer via global metabolomics. Biomarkers 21(7): 578-88

Shurubor YI, Matson WR, Willett WC, Hankinson SE, Kristal BS (2007) Biological variability dominates and influences analytical variance in HPLC-ECD studies of the human plasma metabolome. BMC clinical pathology 7: 9

Slupsky CM, Rankin KN, Wagner J, Fu H, Chang D, Weljie AM, Saude EJ, Lix B, Adamko DJ, Shah S (2007) Investigations of the effects of gender, diurnal variation, and age in human urinary metabolomic profiles. Analytical chemistry 79(18): 6995-7004

Spratlin JL, Serkova NJ, Eckhardt SG (2009) Clinical applications of metabolomics in oncology: a review. Clinical Cancer Research 15(2): 431-440

Spur EM, Decelle EA, Cheng LL (2013) Metabolomic imaging of prostate cancer with magnetic resonance spectroscopy and mass spectrometry. European journal of nuclear medicine and molecular imaging 40 Suppl 1: S60-71

Sreekumar A, Poisson LM, Rajendiran TM, Khan AP, Cao Q, Yu J, Laxman B, Mehra R, Lonigro RJ, Li Y, Nyati MK, Ahsan A, Kalyana-Sundaram S, Han B, Cao X, Byun J, Omenn GS, Ghosh D, Pennathur S, Alexander DC, Berger A, Shuster JR, Wei JT, Varambally S, Beecher C, Chinnaiyan AM (2009) Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. Nature 457(7231): 910-4

Srivastava S, Verma M, Henson DE (2001) Biomarkers for early detection of colon cancer. Clinical cancer research : an official journal of the American Association for Cancer Research 7(5): 1118-26

Stabler S, Koyama T, Zhao Z, Martinez-Ferrer M, Allen RH, Luka Z, Loukachevitch LV, Clark PE, Wagner C, Bhowmick NA (2011) Serum methionine metabolites are risk factors for metastatic prostate cancer progression. PloS one 6(8): e22486

Stein CK, Qu P, Epstein J, Buros A, Rosenthal A, Crowley J, Morgan G, Barlogie B (2015) Removing batch effects from purified plasma cell gene expression microarrays with modified ComBat. BMC bioinformatics 16: 63

Struck-Lewicka W, Kordalewska M, Bujak R, Yumba Mpanga A, Markuszewski M, Jacyna J, Matuszewski M, Kaliszan R, Markuszewski MJ (2015) Urine metabolic fingerprinting using LC-MS and GC-MS reveals metabolite changes in prostate cancer: A pilot study. Journal of pharmaceutical and biomedical analysis 111: 351-61

Szecsi PB, Odum L (2009) Error tracking in a clinical biochemistry laboratory. Clinical chemistry and laboratory medicine 47(10): 1253-7

Tamura S, Fujioka H, Nakano T, Hada T, Higashino K (1987) Serum pseudouridine as a biochemical marker in small cell lung cancer. Cancer research 47(22): 6138-41

Tan B, Qiu Y, Zou X, Chen T, Xie G, Cheng Y, Dong T, Zhao L, Feng B, Hu X, Xu LX, Zhao A, Zhang M, Cai G, Cai S, Zhou Z, Zheng M, Zhang Y, Jia W (2013) Metabonomics identifies serum metabolite markers of colorectal cancer. Journal of proteome research 12(6): 3000-9

Teahan O, Gamble S, Holmes E, Waxman J, Nicholson JK, Bevan C, Keun HC (2006) Impact of analytical bias in metabonomic studies of human blood serum and plasma. Analytical chemistry 78(13): 4307-18

Thapar R, Titus MA (2014) Recent Advances in Metabolic Profiling And Imaging of Prostate Cancer. Current Metabolomics 2(1): 53-69

Thomas R, Kim MH (2008) HIF-1 alpha: a key survival factor for serum-deprived prostate cancer cells. The Prostate 68(13): 1405-15

Tomlinson IP, Alam NA, Rowan AJ, Barclay E, Jaeger EE, Kelsell D, Leigh I, Gorman P, Lamlum H, Rahman S, Roylance RR, Olpin S, Bevan S, Barker K, Hearle N, Houlston RS, Kiuru M, Lehtonen R, Karhu A, Vilkki S, Laiho P, Eklund C, Vierimaa O, Aittomaki K, Hietala M, Sistonen P, Paetau A, Salovaara R, Herva R, Launonen V, Aaltonen LA (2002) Germline mutations in FH predispose to dominantly inherited uterine fibroids, skin leiomyomata and papillary renal cell cancer. Nature genetics 30(4): 406-10

Torgrip RÅ, KM.; Alm, E.; Schuppe-Koistinen I.; Lindberg, J. (2008) A note on normalization of biofluid 1D 1H-NMR data. Metabolomics 4(2): 114-212

Trezzi JP, Bulla A, Bellora C, Rose M, Lescuyer P, Kiehntopf M, Hiller K, Betsou F (2016) LacaScore: a novel plasma sample quality control tool based on ascorbic acid and lactic acid levels. Metabolomics 12: 96

Trygg J, Wold S (2002) Orthogonal projections to latent structures (O- PLS). Journal of chemometrics 16(3): 119-128

Tulpan D, Leger S, Belliveau L, Culf A, Cuperlovic-Culf M (2011) MetaboHunter: an automatic approach for identification of metabolites from 1H-NMR spectra of complex mixtures. BMC bioinformatics 12: 400

Urayama S, Zou W, Brooks K, Tolstikov V (2010) Comprehensive mass spectrometry based metabolic profiling of blood plasma reveals potent discriminatory classifiers of pancreatic cancer. Rapid communications in mass spectrometry : RCM 24(5): 613-20

Utech AE, Tadros EM, Hayes TG, Garcia JM (2012) Predicting survival in cancer patients: the role of cachexia and hormonal, nutritional and inflammatory markers. Journal of cachexia, sarcopenia and muscle 3(4): 245-251

Van De Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ (2002) A gene-expression signature as a predictor of survival in breast cancer. New England Journal of Medicine 347(25): 1999-2009

Vander Heiden MG (2011) Targeting cancer metabolism: a therapeutic window opens. Nature reviews Drug discovery 10(9): 671-684

Vicente-Munoz S, Morcillo I, Puchades-Carrasco L, Paya V, Pellicer A, Pineda-Lucena A (2015) Nuclear magnetic resonance metabolomic profiling of urine provides a noninvasive alternative to the identification of biomarkers associated with endometriosis. Fertility and sterility 104(5): 1202-9

Vu TN, Laukens K (2013) Getting your peaks in line: a review of alignment methods for NMR spectral data. Metabolites 3(2): 259-76

Vu TN, Valkenborg D, Smets K, Verwaest KA, Dommisse R, Lemiere F, Verschoren A, Goethals B, Laukens K (2011) An integrated workflow for robust alignment and simplified quantitative analysis of NMR spectrometry data. BMC bioinformatics 12: 405

Wagner PD, Verma M, Srivastava S (2004) Challenges for biomarkers in cancer detection. Annals of the New York Academy of Sciences 1022(1): 9-16

Walsh MC, Brennan L, Malthouse JPG, Roche HM, Gibney MJ (2006) Effect of acute dietary standardization on the urinary, plasma, and salivary metabolomic profiles of healthy humans. The American journal of clinical nutrition 84(3): 531-539

Warburg O (1956) On the origin of cancer cells. Science 123(3191): 309-14

Warburg O, Wind F, Negelein E (1927) The metabolism of tumors in the body. The Journal of general physiology 8(6): 519

Ward PS, Patel J, Wise DR, Abdel-Wahab O, Bennett BD, Coller HA, Cross JR, Fantin VR, Hedvat CV, Perl AE (2010) The common feature of leukemia-associated IDH1 and IDH2 mutations is a neomorphic enzyme activity converting α-ketoglutarate to 2-hydroxyglutarate. Cancer cell 17(3): 225-234

Weckwerth W (2003) Metabolomics in systems biology. Annual review of plant biology 54(1): 669-689

Weckwerth W, Morgenthal K (2005) Metabolomics: from pattern recognition to biological interpretation. Drug discovery today 10(22): 1551-1558

Weljie AM, Newton J, Mercier P, Carlson E, Slupsky CM (2006) Targeted profiling: quantitative analysis of 1H NMR metabolomics data. Analytical chemistry 78(13): 4430-42

Westerhuis J, Hoefsloot H, Smit S, Vis D, Smilde A, Velzen E, Duijnhoven JPM, FD D (2008) Assessment of PLSDA cross validation. Metabolomics 4(1): 81-89

Wishart DS (2015) Is cancer a genetic disease or a metabolic disease? EBioMedicine 2(6): 478-479

Wishart DS, Mandal R, Stanislaus A, Ramirez-Gaona M (2016) Cancer Metabolomics and the Human Metabolome Database. Metabolites 6(1)

Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S, Fung C, Nikolai L, Lewis M, Coutouly MA, Forsythe I, Tang P, Shrivastava S, Jeroncic K, Stothard P, Amegbey G, Block D, Hau DD, Wagner J, Miniaci J, Clements M, Gebremedhin M, Guo N, Zhang Y, Duggan GE, Macinnis GD, Weljie AM, Dowlatabadi R, Bamforth F, Clive D, Greiner R, Li L, Marrie T, Sykes BD, Vogel HJ, Querengesser L (2007) HMDB: the Human Metabolome Database. Nucleic Acids Res 35(Database issue): D521-6

Wishart, D. S., Knox, C., Guo, A. C., Eisner, R., Young, N., Gautam, B., Hau, D. D.,Psychogios, N., Dong, E., Bouatra, S., Mandal, R., Sinelnikov, I., Xia, J., Jia, L.,Cruz, J. a, Lim, E., Sobsey, C. a, Shrivastava, S., Huang, P., Liu, P., Fang, L., Peng, J., Fradette, R., Cheng, D., Tzur, D., Clements, M., Lewis, A., De Souza, A., Zuniga, A., Dawe, M., Xiong, Y., Clive, D., Greiner, R., Nazyrova, A., Shaykhutdinov, R., Li, L., Vogel, H. J., & Forsythe, I. HMDB: a knowledgebase for the human metabolome. Nucleic Acids Research. 2009; 37(Database issue): D603–10.

Xi B, Gu H, Baniasadi H, Raftery D (2014) Statistical analysis and modeling of mass spectrometry-based metabolomics data. Methods Mol Biol 1198: 333-53

Xia J, Bjorndahl TC, Tang P, Wishart DS (2008) MetaboMiner--semi-automated identification of metabolites from 2D NMR spectra of complex biofluids. BMC bioinformatics 9: 507

Yang M, Soga T, Pollard PJ (2013) Oncometabolites: linking altered metabolism with cancer. The Journal of clinical investigation 123(9): 3652-8

Yang W, Wang Y, Zhou Q, Tang H (2008) Analysis of human urine metabolites using SPE and NMR spectroscopy. Science in China Series B: Chemistry 51(3): 218-225

Yin P, Peter A, Franken H, Zhao X, Neukamm SS, Rosenbaum L, Lucio M, Zell A, Haring HU, Xu G, Lehmann R (2013) Preanalytical aspects and sample quality assessment in metabolomics studies of human blood. Clinical chemistry 59(5): 833-45

Zaragoza P, Ruiz-Cerda JL, Quintas G, Gil S, Costero AM, Leon Z, Vivancos JL, Martinez-Manez R (2014) Towards the potential use of (1)H NMR spectroscopy in urine samples for prostate cancer detection. The Analyst 139(16): 3875-8

Zhang J, Bowers J, Liu L, Wei S, Gowda GA, Hammoud Z, Raftery D (2012a) Esophageal cancer metabolite biomarkers detected by LC-MS and NMR methods. PloS one 7(1): e30181

Zhang T, Watson DG, Wang L, Abbas M, Murdoch L, Bashford L, Ahmad I, Lam NY, Ng AC, Leung HY (2013a) Application of Holistic Liquid Chromatography-High Resolution Mass Spectrometry Based Urinary Metabolomics for Prostate Cancer Detection and Biomarker Discovery. PloS one 8(6): e65880

Zhang WC, Shyh-Chang N, Yang H, Rai A, Umashankar S, Ma S, Soh BS, Sun LL, Tai BC, Nga ME, Bhakoo KK, Jayapal SR, Nichane M, Yu Q, Ahmed DA, Tan C, Sing WP, Tam J, Thirugananam A, Noghabi MS, Pang YH, Ang HS, Mitchell W, Robson P, Kaldis P, Soo RA, Swarup S, Lim EH, Lim B (2012b) Glycine decarboxylase activity drives non-small cell lung cancer tumor-initiating cells and tumorigenesis. Cell 148(1-2): 259-272

Zhang X, Xu L, Shen J, Cao B, Cheng T, Zhao T, Liu X, Zhang H (2013b) Metabolic signatures of esophageal cancer: NMR-based metabolomics and UHPLC-based focused metabolomics of blood serum. Biochimica et biophysica acta 1832(8): 1207-16

Zhou W, Capello M, Fredolini C, Racanicchi L, Piemonti L, Liotta LA, Novelli F, Petricoin EF (2012) Proteomic analysis reveals Warburg effect and anomalous metabolism of glutamine in pancreatic cancer cells. Journal of proteome research 11(2): 554-63

Zira AN, Theocharis SE, Mitropoulos D, Migdalis V, Mikros E (2010) (1)H NMR metabonomic analysis in renal cell carcinoma: a possible diagnostic tool. Journal of proteome research 9(8): 4038-44

VII. ANEXOS

Metabolomics (2017) 13:52
DOI 10.1007/s11306-017-1194-y

ORIGINAL ARTICLE

Non-invasive urinary metabolomic profiling discriminates prostate cancer from benign prostatic hyperplasia

Clara Pérez-Rambla[1] · Leonor Puchades-Carrasco[1] · María García-Flores[2] · José Rubio-Briones[3] · José Antonio López-Guerrero[2] · Antonio Pineda-Lucena[1,4]

Received: 26 August 2016 / Accepted: 2 March 2017 / Published online: 9 March 2017

Abstract
Introduction Prostate cancer (PCa) is one of the most common malignancies in men worldwide. Serum prostate specific antigen (PSA) level has been extensively used as a biomarker to detect PCa. However, PSA is not cancer-specific and various non-malignant conditions, including benign prostatic hyperplasia (BPH), can cause a rise in PSA blood levels, thus leading to many false positive results.
Objectives In this study, we evaluated the potential of urinary metabolomic profiling for discriminating PCa from BPH.
Methods Urine samples from 64 PCa patients and 51 individuals diagnosed with BPH were analysed using ^{1}H nuclear magnetic resonance (^{1}H-NMR). Comparative analysis of urinary metabolomic profiles was carried out using multivariate and univariate statistical approaches.
Results The urine metabolomic profile of PCa patients is characterised by increased concentrations of branched-chain amino acids (BCAA), glutamate and pseudouridine, and decreased concentrations of glycine, dimethylglycine, fumarate and 4-imidazole-acetate compared with individuals diagnosed with BPH.
Conclusion PCa patients have a specific urinary metabolomic profile. The results of our study underscore the clinical potential of metabolomic profiling to uncover metabolic changes that could be useful to discriminate PCa from BPH in a clinical context.

Keywords Biomarkers · Metabolomics · Prostate cancer · Benign prostatic hyperplasia · Nuclear magnetic resonance

Abbreviations

PCa	Prostate cancer
PSA	Prostate specific antigen
DRE	Digital rectal examination
^{1}H-NMR	Nuclear magnetic resonance
BPH	Benign prostatic hyperplasia
TSP	Trimethylsilylpropionic acid-d_4 sodium salt
CPMG	Carr-Purcell-Meiboom-Gill
FIDs	Free induction decays
1D	One-dimensional
2D	Two-dimensional
PQN	Probabilistic quotient normalization
PCA	Principal component analysis
OPLS-DA	Orthogonal partial least squares discriminant analysis
VIP	Variable importance in projection
BMI	Body mass index
BCAA	Branched-chain amino acids
LC–MS	Liquid chromatography–mass spectrometry
GC–MS	Gas chromatography–mass spectrometry
GNMT	Glycine-*N*-methyltransferase
SARDH	Sarcosine dehydrogenase

The original version of this article was revised due to a retrospective Open Access order.

Clara Pérez-Rambla and Leonor Puchades-Carrasco are joint first authors with equal contribution.

✉ Antonio Pineda-Lucena
pineda_ant@gva.es

[1] Structural Biochemistry Laboratory, Centro de Investigación Príncipe Felipe, 46012 Valencia, Spain

[2] Laboratory of Molecular Biology, Fundación Instituto Valenciano de Oncología, 46009 Valencia, Spain

[3] Department of Urology, Fundación Instituto Valenciano de Oncología, 46009 Valencia, Spain

[4] Drug Discovery Unit, Instituto de Investigación Sanitaria La Fe, Avda. Fernando Abril Martorell, 106, 46026 Valencia, Spain

DMGDH	Dimethylglycine dehydrogenase
TCA	Tricarboxylic acid
SDH	Succinate dehydrogenase
FH	Fumarate hydratase

1 Introduction

Prostate cancer (PCa) is the most common cancer in men worldwide. The number of PCa cases is increasing, nowadays representing the sixth leading cause of cancer deaths in men (Zhang et al. 2014). Currently, the most frequently used tests for PCa screening include the determination of prostate specific antigen (PSA) serum levels and digital rectal examination (DRE) (Bunting 2002). The introduction of PSA testing revolutionised PCa screening and became widely adopted by the early 1990s. Since then, the European Randomized study of Screening for Prostate Cancer (ERSPC) has reported a small absolute survival benefit with PSA screening (Ilic et al. 2013; Heijnsdijk et al. 2015). However, PCa screening suffers from a number of limitations, due to the poor specificity of PSA test for detecting cancer and for differentiating indolent cancers from high risk ones.

The low specificity of serum PSA has translated into many unnecessary prostate biopsies and overtreatment of tumours with a low malignant potential, or with a low potential for morbidity or death if left untreated (Draisma et al. 2003; Zappa et al. 1998). It has been estimated that the overdiagnosis, and consequently the overtreatment, of PCa ranges between 30 and 84%, depending on the studies (Etzioni et al. 2002; McGregor et al. 1998). Moreover, trans-rectal ultrasound (TRUS)-guided biopsy following histopathology-based Gleason score, the gold standard test providing histological confirmation (Gleason 1977), is also plagued by high false negative rates (Rabbani et al. 1998; Schoenfield et al. 2007). Early-stage PCa is generally not visible on ultrasound, thus meaning that many tumours are missed on initial biopsy and patients are required to undergo repeated prostate biopsies before definitive PCa detection.

Very few biomarkers are currently validated for use in PCa diagnosis. A recent FDA clinical-grade urine-based assay for the non-coding transcript *PCA3* (overexpressed in >95% of PCa) has demonstrated utility when combined with serum PSA for PCa detection (Loeb and Partin 2011). Another potential biomarker is the specific TMPRSS2 and ERG rearrangement at 21q22, which is 100% indicative of PCa (Barbieri et al. 2012). However, it is only present in approximately 50% of PCa cases. Hence, additional clinically robust biomarkers able to differentiate between indolent and aggressive PCa are urgently needed.

In this context, metabolomics could represent an alternative and very powerful approach for the understanding of the biological pathways and molecular mechanisms involved in the onset and progression of PCa. Metabolomics focuses on the characterisation of metabolic signatures in biofluids or tissues and is leading to advanced diagnostic and therapeutic procedures (Nicholson et al. 2005). Recent studies have shown the potential of metabolomic approaches in the PCa field (Kumar et al. 2015; Stabler et al. 2011; Struck-Lewicka et al. 2015; Zhang et al. 2013). However, so far, no comprehensive PCa studies have been performed on urine, the most accessible and least invasive biofluid, using Nuclear Magnetic Resonance (^{1}H-NMR) spectroscopy, a robust and reliable technological platform allowing the simultaneous measurement and quantification of metabolites with minimal sample handling (Duarte and Gil 2012).

To that end, in this study, a thorough analysis of the urinary metabolomic profile of PCa patients was compared with that corresponding to individuals diagnosed with benign prostatic hyperplasia (BPH), a prostatic condition that cannot be easily distinguished from PCa based on the current PSA screening (Roehrborn et al. 1999). Using a metabolomic approach based on ^{1}H-NMR, it was possible to identify a set of specific metabolites that could contribute to a better understanding of the pathophysiological processes involved in the onset and progression of this disease.

2 Materials and methods

2.1 Patient selection

Patient recruitment was carried out through the Department of Urology and the Biobank of the Instituto Valenciano de Oncología (Valencia, Spain), and measurement and analysis of the urinary metabolomic profiles were performed at the Centro de Investigación Príncipe Felipe (Valencia, Spain) and the Instituto de Investigación Sanitaria La Fe (Valencia, Spain). Urine samples were collected from 64 PCa patients and 51 age-matched individuals. Patient recruitment and sampling procedures were performed in accordance with the Declaration of Helsinki and applicable local regulatory requirements and laws and after approval from the Ethics Committee of the Instituto Valenciano de Oncología. Written informed consent was obtained from each participant before being included in this study.

Clinical diagnosis of individuals was performed according to serum PSA, DRE, biopsy results and Gleason score. Biopsy was performed using at least 6 cores and classification of the individuals included in the study was carried out according to the EAU-ESTRO-SIOG Guidelines on Prostate Cancer (Mottet et al. 2016). The control group consisted of men with no proven PCa based on PSA levels, negative findings on DRE and no malignant findings in prostate tissue biopsies. Based on their clinical

Table 1 Characteristics of the individuals included in the study

	BPH group ($n=51$) (median, range)	PCa patients ($n=64$) (median, range)
Age (years)	62.1 (41.4–74.5)	66.2 (50.0–86.3)
BMI (kg m^{-2})	27 (22.8–34)	27.5 (23–33)
Prostate volume (ml)	52 (24–171)	40.5 (2–134)
PSA (ng/mL)	4.86 (1.02–11.29)	5.11 (0.85–71.41)
Number of cores	12 (10–19)	12.5 (6–54)
Positive cores	NA	25% (0.05–100)
Tumor burden	NA	3.71% (0.14–67.74)
Tumor gleason score	NA	6 (5–9)

BPH benign prostatic hyperplasia, *PCa* prostate cancer, *PSA* prostate-specific antigen, *BMI* body mass index, *NA* not applicable

characteristics, all of them were diagnosed with BPH. Clinical and demographics characteristics of the individuals included in the study are shown in Table 1.

2.2 Sample preparation and ^{1}H-NMR acquisition

Urine samples were immediately frozen after collection and stored at −80 °C. At the time of ^{1}H-NMR analysis, urine samples were thawed on ice and centrifuged at 6000 rpm for 5 min at room temperature. 60 µL of 1.5 mol/L potassium phosphate buffer (pH 7.4) containing 0.1% trimethylsilylpropionic acid-d_4 sodium salt (TSP) and 0.05% NaN_3 were added to 540 µL of urine sample supernatant. After this, 500 µL of the mixture were transferred to a 5-mm NMR tube for analysis.

^{1}H-NMR spectra were acquired using a Bruker Avance II 500 MHz spectrometer. ^{1}H-NMR experiments were acquired at 310 K for every sample. Carr-Purcell-Meiboom-Gill (CPMG) spin-echo pulse sequence (Meiboom and Gill 1958), which generates spectra edited by T2 relaxation times with reduced signals from high molecular weight species and giving improved resolution of low molecular weight metabolite resonances, was collected for each sample with a total of 16 accumulations and 72 K data points over a spectral width of 16 ppm. A 4-s relaxation delay was included between free induction decays (FIDs). The total spin–spin relaxation delay was 40 ms. A one-dimensional (1D) NOESY pulse sequence that generates an unedited spectrum with improved solvent peak suppression (Nicholson et al. 1995) was collected using the same parameters as the CPMG experiment, with a 4-s relaxation delay and 10 ms of mixing time. For both experiments, a water presaturation pulse of 25 Hz was applied throughout the relaxation delays to improve solvent suppression. In addition, two-dimensional (2D) J-resolved spectra, homonuclear 2D ^{1}H–^{1}H total correlation spectroscopy and 2D ^{1}H, ^{13}C heteronuclear single quantum correlation were acquired for selected samples to facilitate the identification of biochmemical substances (Beckonert et al. 2007). All spectra were multiplied by a line-broadening factor of 1 Hz and Fourier transformed. Spectra were automatically phased and baseline corrected, and chemical shift internally referenced to the methyl group signal of TSP at 0.00 ppm using TOPSPIN 3.0 (Bruker Biospin).

2.3 Data modelling and statistical analysis

The main steps of the data modelling and statistical analysis procedures followed in this study are shown in Fig. 1. 1D CPMG spectra were binned using Amix 3.9.7 (Bruker Biospin) into 0.001 ppm wide rectangular buckets over the region δ 9.50–0.15 ppm. The water (δ 5.09–4.55 ppm) and urea signal (δ 6.10–5.52 ppm) regions were excluded from the analysis to avoid interferences arising from differences in water suppression and variability from urea signal, respectively. Spectra were aligned using the "Speaq" R package, a hierarchical cluster-based peak alignment algorithm that minimizes chemical shift variations (Vu et al. 2011), and normalization of

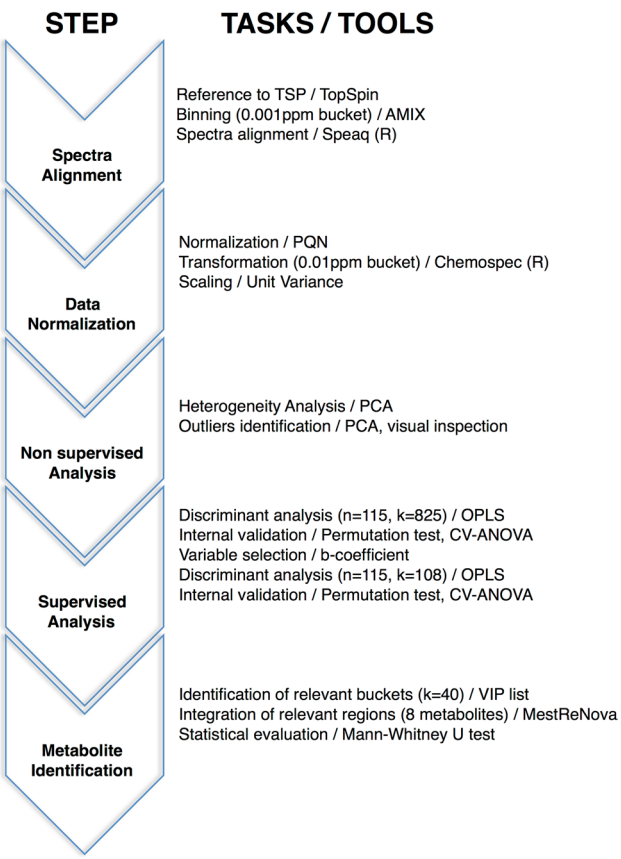


Fig. 1 General scheme of the data modeling and statistical analysis procedures with the main steps highlighted (*n* number of samples, *k* number of variables)

the aligned spectra was performed according to the probabilistic quotient normalization method (PQN) (Dieterle et al. 2006). Finally, the resulting bucket table was transformed into a data matrix containing 0.01 ppm wide rectangular buckets using the "Chemospec" R package (Hanson 2014) to facilitate the statistical analysis.

Multivariate statistical analysis was carried out using SIMCA-P 12.0 (Umetrics AB). Before statistical analysis, data were scaled to unit variance by dividing each variable by 1/SD, where SD represents the standard deviation value of each variable, so that all variables were given equal weight regardless of their absolute value. Principal component analysis (PCA), a nonsupervised statistical approach, was performed on normalized data for finding potential patterns, intrinsic clusters, and outliers. Orthogonal partial least squares discriminant analysis (OPLS-DA) was applied to minimize the possible contribution of inter-group variability and to further improve separation between the groups of samples. The default method of sevenfold internal cross validation was applied, from which Q^2Y (predictive ability parameter, estimated by cross-validation) and R^2Y (goodness of fit parameter) values were extracted. Those parameters, together with the corresponding permutation tests ($n = 100$), were used for the evaluation of the quality of the OPLS-DA models obtained. Variable selection was based on the regression coefficient (b-coefficient) method (Diaz et al. 2013), retaining only those variables with a quotient $|b/b_{cvSE}| > 1.0$, being b_{cvSE} the standard error associated with the b-coefficients.

2.4 Identification and quantification of relevant metabolites

The identification of the variables responsible for the separation between groups of samples in the OPLS-DA models was performed according to the corresponding loading plots and the variable importance in projection (VIP) list of each model. Metabolites of interest were identified using Analysis of MIXtures (AMIX; Bruker) in combination with the Bruker NMR Metabolic Profiling Database BBIOREFCODE 2.0.0 database (Bruker Biospin, Rheinstetten, Germany), as well as other existing public databases and literature reports (Bouatra et al. 2013; Salek et al. 2007). Metabolites contributing to group discrimination in each model were integrated using MestReNova (Cobas and Sardina 2003) to enable comparison between sample groups. Statistical significance of the observed changes was assessed using the Mann–Whitney U test. A p value lower than 0.05 (confidence level 95%) was considered statistically significant.

3 Results

3.1 Urinary metabolomic profile of PCa patients

^{1}H-NMR CPMG spectra were acquired for all urine samples included in the study. Good quality spectra, characterized by the presence of signals with varying degrees of overlapping, were obtained for most of the samples. Figure 2 displays a representative urine ^{1}H-NMR spectrum from a PCa patient and the assignment of the most relevant metabolites identified in these samples. In general, spectra corresponding to this biofluid contain signals from a wide range of low-molecular-weight metabolites of diverse chemical classes (Bouatra et al. 2013), including organic acids, simple sugars and polysaccharides, amino acids, and low-molecular-weight proteins. In particular, urine spectra are dominated by urea, creatinine, trimethylamine-*N*-oxide, dimethylamine, hippuric acid, and citric acid resonances, among others (Fig. 2).

3.2 Non-supervised analysis of the urinary metabolomic profiles

Sample homogeneity within the groups of samples was based on the PCA analysis of the ^{1}H-NMR CPMG urine spectra. Using this approach, it was possible to identify urine samples exhibiting metabolic profiles unusually different to the rest of the samples within their groups. Careful inspection of those samples revealed their spectra contained signals corresponding to several contaminants (e.g., manitol, ethanol, drugs, etc.), or exhibited bad quality due to acquisition problems. These samples were classified as outliers and excluded from the study.

PCA analysis was also used to evaluate the potential influence of different clinical variables on the metabolic profiles obtained for the urine samples of PCa patients and individuals diagnosed with BPH. None of the variables assessed (i.e., age, PSA level, body mass index (BMI), Gleason score) had an impact in the clustering of the samples from both groups. Finally, a non-supervised analysis of the global data did not reveal any significant clustering of the samples based on the urine metabolomic profiles of the two sample groups in this study.

3.3 Supervised analysis of the urinary metabolomic profiles

To better examine potential differences between the groups of samples, an OPLS-DA model aiming to discriminate the urinary profiles from PCa patients and individuals diagnosed with BPH was built. This OPLS-DA model (Fig. 3) showed a reasonable fitting of the data ($R^2 = 0.586$), but it did not exhibit any predictive power ($Q^2 = -0.230$).

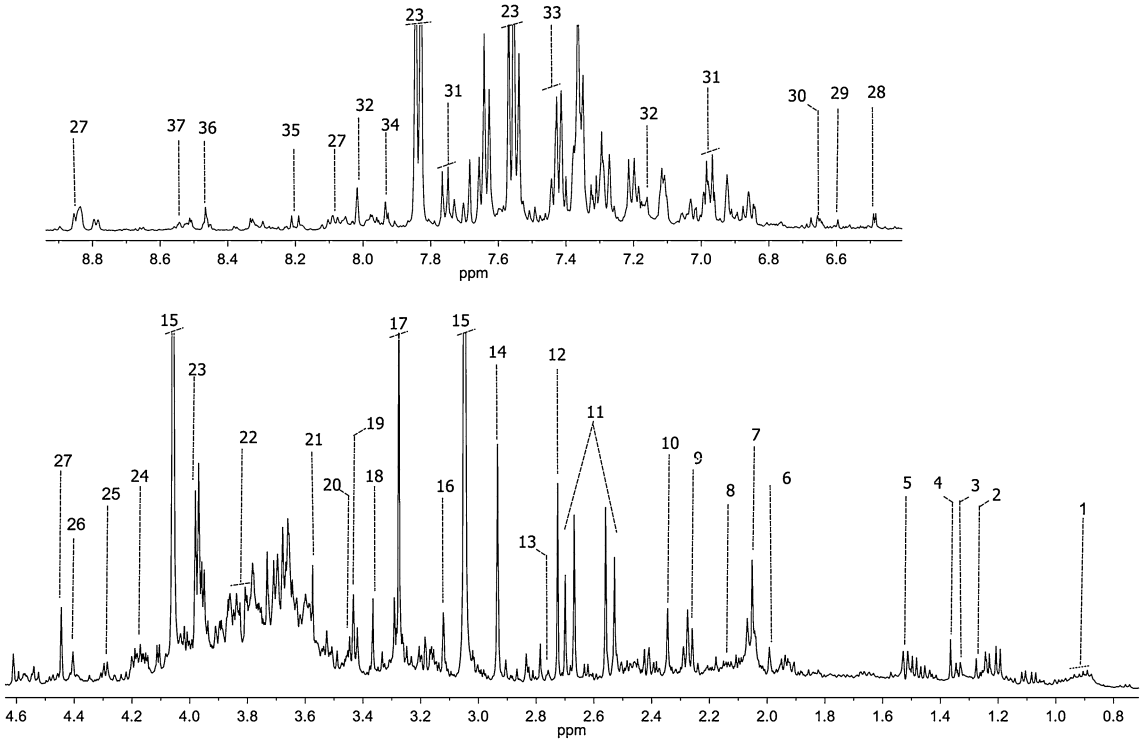


Fig. 2 Representative 500 MHz ^{1}H-NMR spectrum and assignment of a urine sample from a PCa patient. Assigned metabolites: *1* branched-chain amino acids; *2* 3-hydroxyisovalerate; *3* lactate; *4* 2-hydroxyisobutyrate; *5* alanine; *6* acetate; *7* *N*-acetyl groups; *8* glutamate; *9* 2-hydroxy-glutarate; *10* pyruvate; *11* citrate; *12* dimethylamine; *13* sarcosine; *14* dimethylglycine; *15* creatinine; *16* cis-aconitic acid; *17* trimethylamine-*N*-oxide; *18* methanol; *19* trans-aconitic acid; *20* taurine; *21* glycine; *22* serine; *23* hippurate; *24* pseudouridine; *25* threonine; *26* dihydroxyacetone; *27* trigonelline; *28* U1; *29* fumarate; *30* 2-furoylglycine; *31* 4-hydroxybenzoate; *32* 3-methylhistidine; *33* phenylalanine; *34* histidine; *35* hypoxanthine; *36* formate; *37* 4-imidazole-acetate

OPLS-DA model significance was assessed using a cross-validated ANOVA ($p \leq 0.01$ was considered significant) and a permutation test ($n = 100$). The results of this internal validation ($R^2 = 0.600$, $Q^2 = -0.101$; p value >0.01) revealed overfitting of the data, most probably reflecting the elevated number of variables (823) over samples (115) used to build this model (Andersen and Bro 2010).

To overcome this limitation, a variable selection strategy, based on the regression coefficient method (b-coefficient) (Diaz et al. 2013), was followed to remove uninformative variables. The application of this variable selection method reduced the number of variables to 108, and the OPLS-DA then provided a model with significant reduction in sample scores dispersion and improved predictive power ($Q^2 = 0.416$) (Fig. 4). The results of the internal validation of this new OPLS-DA model ($R^2 = 0.358$, $Q^2 = -0.234$; p value <0.01) confirmed its robustness (Szymanska et al. 2012). The value of R^2 of this new model remained unchanged ($R^2 = 0.600$) when compared with the original one, confirming that the discarded variables were not relevant for explaining the differences between the metabolomic profiles of PCa patients and individuals diagnosed with BPH.

3.4 Metabolite identification and quantification

Examination of the corresponding loading plot and VIP list of the new OPLS-DA model facilitated the identification of the most relevant variables that were contributing to the discrimination of the PCa patients and the individuals diagnosed with BPH. Following this strategy, a total of 40 out of the 108 variables were identified as relevant regions in the discrimination, and used to identify the spectral signals corresponding to the altered metabolites in pathological conditions. The metabolites corresponding to those regions were identified through a combination of their ^{1}H chemical shifts in the ^{1}H-NMR CPMG spectra and the spin system patterns obtained from the 2D spectra acquired for representative samples of each group.

Further analysis of the data was carried out with the use of variable-size bucketing to assess if the metabolites associated with the relevant variables were also significant when comparing the two sample groups. This analysis revealed a total of 8 metabolites (Table 2) whose concentrations exhibited statistically significant differences when comparing the urinary metabolomic profiles of PCa patients and individuals diagnosed with BPH. Thus, it was

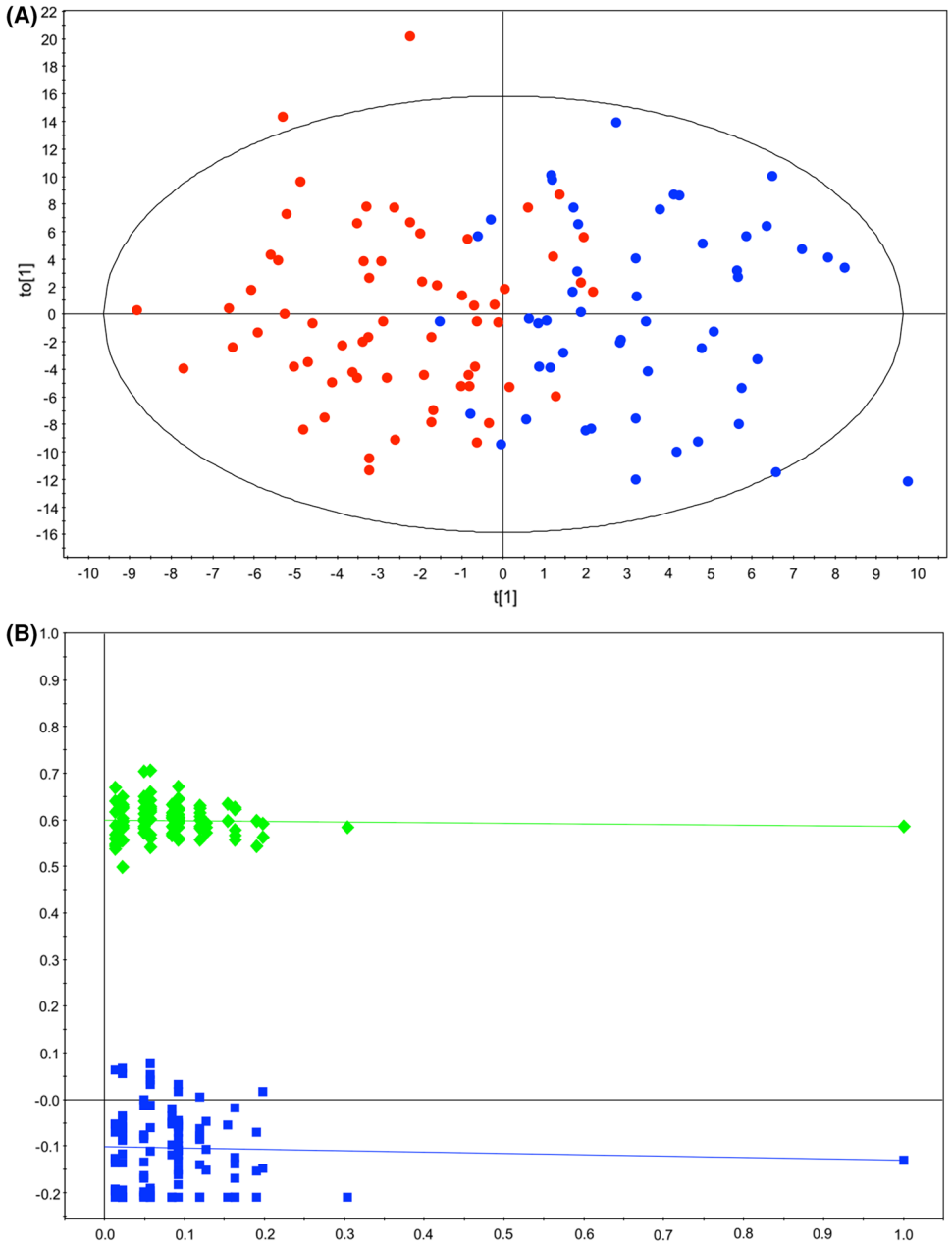


Fig. 3 Multivariate modelling resulting from the analysis of urine ^{1}H-NMR spectra before variable selection (823 variables). **a** OPLS-DA score plot for the comparison between PCa patients (*red circle*) *vs*. individuals diagnosed with BPH (*blue circle*); **b** internal validation of the corresponding OPLS-DA model by permutation analysis (n = 100), R^2 (*green diamond*), Q^2 (*blue square*)

found that urine from PCa patients, compared with individuals diagnosed with BPH, was characterized by increased concentrations of branched-chain amino acids (BCAA), glutamate and pseudouridine, and decreased concentrations of glycine, dimethylglycine, fumarate, 4-imidazole-acetate, and one unknown metabolite (U1).

4 Discussion

Efforts to identify non-invasive PCa biomarkers that can stratify patients with high sensitivity and specificity for screening, diagnosis, prognosis, prediction and monitoring remain a fundamental goal in this area (Thapar and Titus 2014). In this context, our study represents the first comprehensive study focused on the characterisation and comparison of the specific urinary metabolomic profile of PCa patients with that of patients diagnosed with BPH using ^{1}H-NMR. The only other report focused on the analysis of a relatively similar set of urine samples to that included in our study suggested that "fingerprints" (i.e., global profiles) based on the analysis of urinary NMR metabolomic profiles could be a suitable and promising method for PCa detection (Zaragoza et al. 2014). The in-depth analysis carried out in our study, based on non-invasive urinary metabolomic studies, reveals that the discrimination between PCa

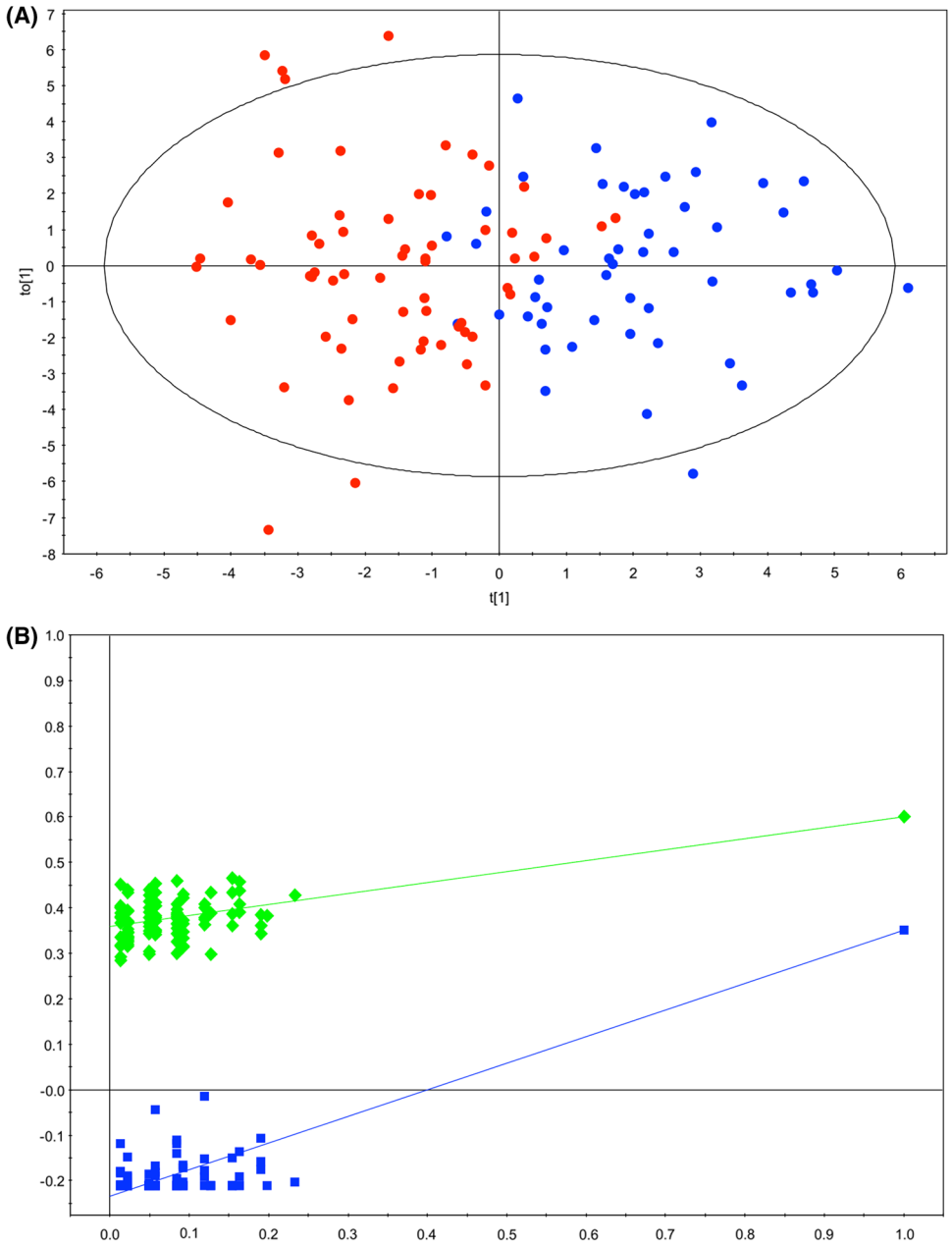


Fig. 4 Multivariate modelling resulting from the analysis of urine ^{1}H-NMR spectra after variable selection (108 variables). **a** OPLS-DA score plot for the comparison between PCa patients (*red circle*) vs. individuals diagnosed with BPH (*blue circle*); **b** internal validation of the corresponding OPLS-DA model by permutation analysis (n = 100), R^2 (*green diamond*), Q^2 (*blue square*)

patients and individuals diagnosed with BPH actually relies on specific urinary metabolites, an information that could eventually contribute to the early diagnosis of PCa. Our results show that the urinary metabolomic profile of PCa patients, compared with individuals diagnosed with BPH, is characterised by statistically significant changes in the concentration of several metabolites. The analysis of those metabolic alterations reveals that PCa is associated with profound changes in energy metabolism.

Thus, in our study, we observed decreased levels of glycine and dimethylglycine when the urinary metabolomic profiles of PCa patients and individuals diagnosed with BPH were compared. This result is in agreement with recent studies performed in serum (Kumar et al. 2015) and urine (Struck-Lewicka et al. 2015) of PCa patients and healthy individuals. Kumar et al. (Kumar et al. 2015) found increased levels of sarcosine and decreased levels of glycine in serum samples of PCa patients compared with healthy individuals. Furthermore, Struck-Lewicka et al. (Struck-Lewicka et al. 2015) have reported decreased levels of glycine, in a study performed by liquid chromatography–mass spectrometry (LC–MS) and gas chromatography–mass spectrometry (GC–MS), when comparing the urinary metabolomic profiles of PCa patients and healthy individuals.

Table 2 Mean intensities and variations for the statistically significant metabolites involved in the discrimination between individuals diagnosed with BPH and PCa patients

Metabolite	δ ^{1}H (ppm)[a]	BPH group (mean ± s.e.m.)[b]	PCA patients (mean ± s.e.m.)[b]	*p* value	% variation
BCAA	0.930–0.842	13.10 ± 0.26	14.33 ± 0.37	0.011*	9.37
Glutamate	2.115–2.081	11.72 ± 0.18	12.42 ± 0.21	0.012*	5.98
Dimethylglycine	2.944–2.922	20.56 ± 0.91	17.17 ± 1.11	0.034*	−16.48
Glycine	3.582–3.567	23.29 ± 1.01	20.49 ± 0.94	0.015*	−12.00
Pseudouridine	4.309–4.277	6.88 ± 0.13	7.68 ± 0.41	0.049*	11.65
U1	6.496–6.478	0.69 ± 0.09	0.45 ± 0.04	0.027*	−34.50
Fumarate	6.551–6.498	0.99 ± 0.05	0.87 ± 0.03	0.021*	−12.54
4-Imidazole-acetate	8.567–8.517	1.77 ± 0.11	1.37 ± 0.62	0.006*	−22.52

BCAA branched-chain amino acids, *ppm* parts per million, *s.e.m.* standard error of mean. *P* values calculated using the Mann–Whitney U test. *$P < 0.05$

[a]Chemical shift range for integration

[b]Spectral intensity in arbitrary units

Glycine is converted to sarcosine, an *N*-methyl derivative of glycine that has been previously linked to PCa (Sreekumar et al. 2009), by the enzyme glycine-*N*-methyltransferase (GNMT). Sarcosine levels are also regulated by sarcosine dehydrogenase (SARDH), the enzyme that converts sarcosine back to glycine, and dimethylglycine dehydrogenase (DMGDH) which generates sarcosine from dimethylglycine (Sreekumar et al. 2009). The involvement of sarcosine in PCa has been the subject of many studies (Khan et al. 2013; Miyake et al. 2012; Issaq 2011; Bianchi et al. 2011; Lucarelli et al. 2012; Sreekumar et al. 2009; Kumar et al. 2015; Jentzmik et al. 2010). However, its role as a potential biomarker of PCa remains controversial and unclear (Ploussard and de la Taille 2010). In our study, we found elevated levels of sarcosine in PCa patients, although this variation was not statistically significant. Taken together, our results would support the idea of an interconversion between glycine/dimethylglycine and sarcosine through the activation of both DMGH and GNMT, and the down-regulation of SARDH.

There are also other mechanisms that could contribute to a reduction in the levels of circulating glycine. Recent work on cancer metabolomics has shown that glycine uptake is associated with cancel cell proliferation through its involvement in one-carbon metabolism (Zhang et al. 2012). This pathway has been traditionally considered a "housekeeping" process, and encompasses a complex metabolic network based on the chemical reactions of folate compounds. Recent findings also suggests that hyperactivation of this pathway could potentially be a driver of oncogenesis and tumor maintenance (Locasale 2013). In this context, glycine metabolism has been reported to be involved in cell transformation and tumorigenesis. This process would be mediated by the activity of glycine dehydrogenase (decarboxylating) (GLDC) that links glycine cleavage with the charging of the folate cycle.

Furthermore, the rapid, dysregulated cell growth found in cancer cells, demands extra sources of energy to sustain proliferation (Zhang et al. 2012). Thus, in addition to pyruvate derived from glycolysis, fatty acids and particularly amino acids can supply substrates to the tricarboxylic acid (TCA) cycle to maintain mitochondrial production in cancer cells (Chen and Russo 2012).

One of the factors contributing to the availability of amino acids is a metabolic syndrome experienced by approximately 60% of PCa patients termed cachexia (Utech et al. 2012). This process involves a net increase in protein catabolism along with activation of proteolysis, and has a tremendous impact in the levels of BCAAs (O'Connell 2013). Under normal conditions, BCAA oxidation in skeletal muscle provides 6–7% of the energy needs, but under highly catabolic circumstances, such as cancer cachexia, the contribution can be as high as 20% (Lam and Poon 2008). In these conditions, it would be expected an increase in circulating BCAAs, thus being in perfect agreement with our observation and other studies carried out in prostate tissue (Giskeødegård et al. 2013; McDunn et al. 2013) and serum samples (Giskeødegård et al. 2015) from PCa patients. It would also explain the results obtained in previous studies showing that the levels of BCAAs are significantly increased in certain neoplastic processes (e.g., gastric and esophageal cancers) (Fan et al. 2012; Zhang et al. 2013). Interestingly, BCAAs can be converted into acetyl-CoA and other organic molecules that enter the TCA cycle. The metabolic flexibility afforded by multiple inputs into the TCA cycle allows cancer cells to adequately respond to the fuels available in the changing microenvironment during the evolution of the tumor (Boroughs and DeBerardinis 2015).

Furthermore, the catabolism of BCAAs also provides an important source for the generation of amino acids, especially glutamine and alanine. Different cancer studies

(Lasagna-Reeves et al. 2010; Gao et al. 2008; Zira et al. 2010) have shown alterations in glutamine levels that are presumably associated with increased metabolic activity derived from the conditions of hypoxia and hypermetabolism observed in the tumor environment (Eigenbrodt et al. 1998). Proliferating cancer cells take up glutamine and convert it to glutamate through a variet of deamidation and transamidation reactions, most notably the mitochondrial amidohydrolase glutaminase (Hensley et al. 2013). It leads to the production of ammonia and glutamate to balance the pH in tumor cells and could explain the increase of glutamate observed in the urine of PCa patients. This result is also in agreement with previous PCa studies performed in serum (Giskeødegård et al. 2015) and prostate tissue (McDunn et al. 2013). Glutamate is subsequently transformed into α-ketoglutarate through a series of biochemical reactions termed glutaminolysis that contribute to replenish depleted intermediates of the TCA cycle (DeBerardinis et al. 2008).

Regarding amino acids metabolism, a significant decrease of 4-imidazole-acetate, a compound linked to histidine metabolism, was also observed in the urine of PCa patients. Interestingly, this metabolite was also identified in a previous study carrried out with serum samples collected up to 20 years prior to PCa diagnosis (Mondul et al. 2015). In this study, it was associated with both the overall risk of PCa (odds ratio 1.33) and aggressive PCa (odds ratio 1.40). Previous studies have also shown that histidine levels are increased in serum (Giskeødegård et al. 2015) and tissue (McDunn et al. 2013) samples from PCa patients, our finding perhaps reflecting a limited ability to process this amino acid by PCa cells. Alterations in histidine metabolism, as well as in BCAA (valine, leucine and isoleucine) metabolism, have also been observed in other cancers (e.g., ovarian cancer, breast cancer) (Ke et al. 2015; Schramm et al. 2010).

An increase in the urinary levels of pseudouridine, an isomer of the nucleoside uridine in which the uracil moiety is attached through a carbon–carbon bond, was found to be elevated in the urine metabolomic profile of PCa patients compared with individuals diagnosed with BPH. Increased levels of uracil, or other uracil-containing metabolites (e.g., 2′-deoxyuridine) (Mondul et al. 2015), have been found in previous PCa studies (Jiang et al. 2010; McDunn et al. 2013; Spur et al. 2013; Sreekumar et al. 2009) suggesting an important role of the metabolism of this compound in this disease. Alterations in the levels of pseudouridine have also been observed in other pathological processes (Rasmuson and Bjork 1995; Vicente-Munoz et al. 2015; Masaki et al. 2006) and have been associated with disease activity, tumor burden, and clinical status (Tamura et al. 1987).

Finally, the analysis of the urinary metabolomic profiles of PCa patients and individuals diagnosed with BPH also revealed significant variations in the levels of fumarate, a key molecule in the TCA cycle. Within this cycle, the succinate dehydrogenase (SDH) complex converts succinate to fumarate, that is further down transformed to malate by the fumarate hydratase (FH). Mutations in these enzymes have been previously linked to renal cell carcinomas, uterine and skin cancer (Tomlinson et al. 2002). Previous studies have also shown decreased levels of other TCA metabolites (isocitrate, aconitate and succinate) in the urine of PCa patients, all the data supporting a disruption in energy metabolism (Struck-Lewicka et al. 2015). Moreover, the decreased levels of fumarate in the urine of PCa patients, compared with individuals diagnosed with BPH, positively correlates with previous studies showing an accumulation of this metabolite in PCa bone metastases (Thapar and Titus 2014) and prostate tissue (McDunn et al. 2013), a process that would lead to a reduction in the levels of circulating fumarate. Interestingly, succinate, another metabolite exhibiting decreased levels in the urine of PCA patients, also tends to accumulate in cancer cells. Both metabolites belong to a family of compounds termed oncometabolites that are known to accumulate in cancer cells and facilitate cancer progression (Yang et al. 2013). In particular, these two oncometabolites have been associated with the aberrant stabilization of HIF-1a (Semenza 2010), a key protein in cancer that is commonly overexpressed in PCa cells (Thomas and Kim 2008).

5 Concluding remarks

In summary, the present study reveals for the first time that the analysis of urinary metabolomic profiles provides a non-invasive tool for characterizing PCa-associated biomarkers and for getting a better understanding of the metabolic alterations underlying this neoplastic process. Although further validation of the results, using an independent set of samples, will be necessary to increase the robustness of this analysis, our data support the idea that multivariate statistical analysis of ^{1}H-NMR urinary metabolomic profiles obtained from PCa patients could be used for objectively discriminating individuals with BPH or PCa.

Acknowledgements The authors thank all the staff members and the Biobank of the Instituto Valenciano de Oncología who contributed to sample collection and handling. Funding for the present study was provided by the Ministerio de Economía y Competitividad (SAF2014-53977-R), the Conselleria de Educación, Investigación, Cultura y Deporte (GVA, PROMETEO/2016/103) and a grant sponsored by Abbott from the Spanish Society of Urology (2012–2013).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the Declaration of Helsinki and applicable local regulatory requirements and laws and after approval from the Ethics Committee of the Instituto Valenciano de Oncología.

Ethical requirements The manuscript is in compliance with ethical requirement of the journal.

Informed consent Written informed consent was obtained from each participant before being included in this study.

References

Andersen, C. M., & Bro, R. (2010). Variable selection in regression—a tutorial. *Journal of Chemometrics, 24*(11-12), 728–737.

Barbieri, C. E., Demichelis, F., & Rubin, M. A. (2012). Molecular genetics of prostate cancer: Emerging appreciation of genetic complexity. *Histopathology, 60*(1), 187–198.

Beckonert, O., Keun, H. C., Ebbels, T. M., Bundy, J., Holmes, E., Lindon, J. C., et al. (2007). Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature Protocols, 2*(11), 2692–2703.

Bianchi, F., Dugheri, S., Musci, M., Bonacchi, A., Salvadori, E., Arcangeli, G., et al. (2011). Fully automated solid-phase microextraction-fast gas chromatography-mass spectrometry method using a new ionic liquid column for high-throughput analysis of sarcosine and *N*-ethylglycine in human urine and urinary sediments. *Analytica Chimica Acta, 707*(1–2), 197–203.

Boroughs, L. K., & DeBerardinis, R. J. (2015). Metabolic pathways promoting cancer cell survival and growth. *Nature Cell Biology, 17*(4), 351–359.

Bouatra, S., Aziat, F., Mandal, R., Guo, A. C., Wilson, M. R., Knox, C., et al. (2013). The human urine metabolome. *PLoS ONE, 8*(9), e73076.

Bunting, P. S. (2002). Screening for prostate cancer with prostate-specific antigen: Beware the biases. *Clinica Chimica Acta; International Journal of Clinical Chemistry, 315*(1–2), 71–97.

Chen, J. Q., & Russo, J. (2012). Dysregulation of glucose transport, glycolysis, TCA cycle and glutaminolysis by oncogenes and tumor suppressors in cancer cells. *Biochimica et Biophysica Acta, 1826*(2), 370–384.

Cobas, J. C., & Sardina, F. J. (2003). Nuclear magnetic resonance data processing. MestRe-C: A software package for desktop computers. *Concepts in Magnetic Resonance Part A, 19A*(2), 80–96.

DeBerardinis, R. J., Lum, J. J., Hatzivassiliou, G., & Thompson, C. B. (2008). The biology of cancer: Metabolic reprogramming fuels cell growth and proliferation. *Cell Metabolism, 7*(1), 11–20.

Diaz, S. O., Barros, A. S., Goodfellow, B. J., Duarte, I. F., Galhano, E., Pita, C., et al. (2013). Second trimester maternal urine for the diagnosis of trisomy 21 and prediction of poor pregnancy outcomes. *Journal of Proteome Research, 12*(6), 2946–2957.

Dieterle, F., Ross, A., Schlotterbeck, G., & Senn, H. (2006). Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in ^{1}H NMR metabonomics. *Analytical Chemistry, 78*(13), 4281–4290.

Draisma, G., Boer, R., Otto, S. J., van der Cruijsen, I. W., Damhuis, R. A., Schroder, F. H., et al. (2003). Lead times and overdetection due to prostate-specific antigen screening: Estimates from the European Randomized Study of Screening for Prostate Cancer. *Journal of the National Cancer Institute, 95*(12), 868–878.

Duarte, I. F., & Gil, A. M. (2012). Metabolic signatures of cancer unveiled by NMR spectroscopy of human biofluids. *Progress in Nuclear Magnetic Resonance Spectroscopy, 62*, 51–74.

Eigenbrodt, E., Kallinowski, F., Ott, M., Mazurek, S., & Vaupel, P. (1998). Pyruvate kinase and the interaction of amino acid and carbohydrate metabolism in solid tumors. Anticancer Res, *18*(5A), 3267–3274.

Etzioni, R., Penson, D. F., Legler, J. M., di Tommaso, D., Boer, R., Gann, P. H., et al. (2002). Overdiagnosis due to prostate-specific antigen screening: Lessons from U.S. prostate cancer incidence trends. *Journal of the National Cancer Institute, 94*(13), 981–990.

Fan, J., Hong, J., Hu, J.-D., & Chen, J.-L. (2012). Ion chromatography based urine amino acid profiling applied for diagnosis of gastric cancer. *Gastroenterology Research and Practice*, 2012, 474907.

Gao, H., Dong, B., Liu, X., Xuan, H., Huang, Y., & Lin, D. (2008). Metabonomic profiling of renal cell carcinoma: High-resolution proton nuclear magnetic resonance spectroscopy of human serum with multivariate data analysis. *Analytica Chimica Acta, 624*(2), 269–277.

Giskeødegård, G. F., Bertilsson, H., Selnæs, K. M., Wright, A. J., Bathen, T. F., Viset, T., et al. (2013). Spermine and citrate as metabolic biomarkers for assessing prostate cancer aggressiveness. *PLoS ONE, 8*(4), e62375.

Giskeødegård, G. F., Hansen, A. F., Bertilsson, H., Gonzalez, S. V., Kristiansen, K. A., Bruheim, P., et al. (2015). Metabolic markers in blood can separate prostate cancer from benign prostatic hyperplasia. *British Journal of Cancer, 113*(12), 1712–1719.

Gleason, D. (1977). Histologic grading and clinical staging of prostatiiccarcinoma. In M. Tannenbaum (Ed.), *Urologic pathology: The prostate* (pp. 171–198). Philadelphia, PA: Lea and Febiger.

Hanson, B. A. (2014). ChemoSpec: An R Package for Chemometric Analysis of Spectroscopic Data. Package Version 2.0–2.

Heijnsdijk, E. A., de Carvalho, T. M., Auvinen, A., Zappa, M., Nelen, V., Kwiatkowski, M., et al. (2015). Cost-effectiveness of prostate cancer screening: A simulation study based on ERSPC data. *Journal of the National Cancer Institute, 107*(1), 366.

Hensley, C. T., Wasti, A. T., & DeBerardinis, R. J. (2013). Glutamine and cancer: Cell biology, physiology, and clinical opportunities. *The Journal of Clinical Investigation, 123*(9), 3678–3684.

Ilic, D., Neuberger, M. M., Djulbegovic, M., & Dahm, P. (2013). Screening for prostate cancer. *Cochrane Database of Systematic Reviews (Online), 1*, CD004720.

Issaq, H. J., & Veenstra, T. D. (2011). Is sarcosine a biomarker for prostate cancer? *Journal of Separation Science, 34*(24), 3619–3621.

Jentzmik, F., Stephan, C., Miller, K., Schrader, M., Erbersdobler, A., Kristiansen, G., et al. (2010). Sarcosine in urine after digital rectal examination fails as a marker in prostate cancer detection and identification of aggressive tumours. *European Urology, 58*(1), 12–18 (**discussion 20–11**).

Jiang, Y., Cheng, X., Wang, C., & Ma, Y. (2010). Quantitative determination of sarcosine and related compounds in urinary samples by liquid chromatography with tandem mass spectrometry. *Analytical Chemistry, 82*(21), 9022–9027.

Ke, C., Hou, Y., Zhang, H., Fan, L., Ge, T., Guo, B., et al. (2015). Large-scale profiling of metabolic dysregulation in ovarian cancer. *International Journal of Cancer, 136*(3), 516–526.

Khan, A. P., Rajendiran, T. M., Bushra, A., Asangani, I. A., Athanikar, J. N., Yocum, A. K., et al. (2013). The role of sarcosine metabolism in prostate cancer progression. *Neoplasia, 15*(5), 491–N413.

Kumar, D., Gupta, A., Mandhani, A., & Sankhwar, S. N. (2015). Metabolomics-derived prostate cancer biomarkers: Fact or fiction? *Journal of Proteome Research, 14*(3), 1455–1464.

Lam, V. W., & Poon, R. T. (2008). Role of branched-chain amino acids in management of cirrhosis and hepatocellular carcinoma. *Hepatology Research, 38*(Suppl 1), 107–115.

Lasagna-Reeves, C., Gonzalez-Romero, D., Barria, M. A., Olmedo, I., Clos, A., Sadagopa Ramanujam, V. M., et al. (2010). Bioaccumulation and toxicity of gold nanoparticles after repeated administration in mice. *Biochemical and Biophysical Research Communications, 393*(4), 649–655.

Locasale, J. W. (2013). Serine, glycine and one-carbon units: cancer metabolism in full circle. *Nature Reviews Cancer, 13*(8), 572–583.

Loeb, S., & Partin, A. W. (2011). Review of the literature: PCA3 for prostate cancer risk assessment and prognostication. *Reviews in Urology, 13*(4), e191–e195.

Lucarelli, G., Fanelli, M., Larocca, A. M., Germinario, C. A., Rutigliano, M., Vavallo, A., et al. (2012). Serum sarcosine increases the accuracy of prostate cancer detection in patients with total serum PSA less than 4.0 ng/ml. *The Prostate, 72*(15), 1611–1621.

Masaki, Y., Itoh, K., Sawaki, T., Karasawa, H., Kawanami, T., Fukushima, T., et al. (2006). Urinary pseudouridine in patients with lymphoma: Comparison with other clinical parameters. *Clinica Chimica Acta, 371*(1–2), 148–151.

McDunn, J. E., Li, Z., Adam, K. P., Neri, B. P., Wolfert, R. L., Milburn, M. V., et al. (2013). Metabolomic signatures of aggressive prostate cancer. *The Prostate, 73*(14), 1547–1560.

McGregor, M., Hanley, J. A., Boivin, J. F., & McLean, R. G. (1998). Screening for prostate cancer: Estimating the magnitude of overdetection. *Canadian Medical Association Journal, 159*(11), 1368–1372.

Meiboom, S., & Gill, D. (1958). Modified spin-echo method for measuring nuclear relaxation times. *The Review of Scientific Instruments, 29*(8), 688–701.

Miyake M, G. G. E., Aguilar Palacios, D., & Rosser, C. J. (2012). Sarcosine, a biomarker for prostate cancer: Ready for prime time? *Biomarkers in Medicine, 6*(4), 513–514.

Mondul, A. M., Moore, S. C., Weinstein, S. J., Karoly, E. D., Sampson, J. N., & Albanes, D. (2015). Metabolomic analysis of prostate cancer risk in a prospective cohort: The alpha-tocolpherol, beta-carotene cancer prevention (ATBC) study. *International Journal of Cancer, 137*(9), 2124–2132.

Mottet, N., Bellmunt, J., Briers, E., Bolla, M., Cornford, P., De Santis, M., et al. (2016). EAU-ESTRO-SIOG guidelines on prostate cancer. *European Association of Urology*. doi:10.1016/j.eururo.2016.08.002.

Nicholson, J. K., Foxall, P. J., Spraul, M., Farrant, R. D., & Lindon, J. C. (1995). 750 MHz ^{1}H and ^{1}H-^{13}C NMR spectroscopy of human blood plasma. *Analytical Chemistry, 67*(5), 793–811.

Nicholson, J. K., Holmes, E., & Wilson, I. D. (2005). Gut microorganisms, mammalian metabolism and personalized health care. *Nature Reviews Microbiology, 3*(5), 431–438.

O'Connell, T. M. (2013). The complex role of branched chain amino acids in diabetes and cancer. *Metabolites, 3*(4), 931–945.

Ploussard, G., & de la Taille, A. (2010). Urine biomarkers in prostate cancer. *Nature Reviews Urology, 7*(2), 101–109.

Rabbani, F., Stroumbakis, N., Kava, B. R., Cookson, M. S., & Fair, W. R. (1998). Incidence and clinical significance of false-negative sextant prostate biopsies. *The Journal of Urology, 159*(4), 1247–1250.

Rasmuson, T., & Bjork, G. R. (1995). Urinary excretion of pseudouridine and prognosis of patients with malignant lymphoma. *Acta Oncologica, 34*(1), 61–67.

Roehrborn, C. G., Boyle, P., Gould, A. L., & Waldstreicher, J. (1999). Serum prostate-specific antigen as a predictor of prostate volume in men with benign prostatic hyperplasia. *Urology, 53*(3), 581–589.

Salek, R. M., Maguire, M. L., Bentley, E., Rubtsov, D. V., Hough, T., Cheeseman, M., et al. (2007). A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat, and human. *Physiological Genomics, 29*(2), 99–108.

Schoenfield, L., Jones, J. S., Zippe, C. D., Reuther, A. M., Klein, E., Zhou, M., et al. (2007). The incidence of high-grade prostatic intraepithelial neoplasia and atypical glands suspicious for carcinoma on first-time saturation needle biopsy, and the subsequent risk of cancer. *BJU International, 99*(4), 770–774.

Schramm, G., Surmann, E. M., Wiesberg, S., Oswald, M., Reinelt, G., Eils, R., et al. (2010). Analyzing the regulation of metabolic pathways in human breast cancer. *BMC Medical Genomics, 3*, 39.

Semenza, G. L. (2010). HIF-1: Upstream and downstream of cancer metabolism. *Current Opinion in Genetics and Development, 20*(1), 51–56.

Spur, E. M., Decelle, E. A., & Cheng, L. L. (2013). Metabolomic imaging of prostate cancer with magnetic resonance spectroscopy and mass spectrometry. *European Journal of Nuclear Medicine and Molecular Imaging, 40*(Suppl 1), 60–71.

Sreekumar, A., Poisson, L. M., Rajendiran, T. M., Khan, A. P., Cao, Q., Yu, J., et al. (2009). Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature, 457*(7231), 910–914.

Stabler, S., Koyama, T., Zhao, Z., Martinez-Ferrer, M., Allen, R. H., Luka, Z., et al. (2011). Serum methionine metabolites are risk factors for metastatic prostate cancer progression. *PLoS ONE, 6*(8), e22486.

Struck-Lewicka, W., Kordalewska, M., Bujak, R., Yumba Mpanga, A., Markuszewski, M., Jacyna, J., et al. (2015). Urine metabolic fingerprinting using LC-MS and GC-MS reveals metabolite changes in prostate cancer: A pilot study. *Journal of Pharmaceutical and Biomedical Analysis, 111*, 351–361.

Szymanska, E., Saccenti, E., Smilde, A. K., & Westerhuis, J. A. (2012). Double-check: Validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics, 8*(Suppl 1), 3–16.

Tamura, S., Fujioka, H., Nakano, T., Hada, T., & Higashino, K. (1987). Serum pseudouridine as a biochemical marker in small cell lung cancer. *Cancer Research, 47*(22), 6138–6141.

Thapar, R., & Titus, M. A. (2014). Recent advances in metabolic profiling and imaging of prostate cancer. *Current Metabolomics, 2*(1), 53–69.

Thomas, R., & Kim, M. H. (2008). HIF-1 alpha: A key survival factor for serum-deprived prostate cancer cells. *The Prostate, 68*(13), 1405–1415.

Tomlinson, I. P., Alam, N. A., Rowan, A. J., Barclay, E., Jaeger, E. E., Kelsell, D., et al. (2002). Germline mutations in FH predispose to dominantly inherited uterine fibroids, skin leiomyomata and papillary renal cell cancer. *Nature Genetics, 30*(4), 406–410.

Utech, A. E., Tadros, E. M., Hayes, T. G., & Garcia, J. M. (2012). Predicting survival in cancer patients: The role of cachexia and

hormonal, nutritional and inflammatory markers. *Journal of Cachexia, Sarcopenia and Muscle, 3*(4), 245–251.

Vicente-Munoz, S., Morcillo, I., Puchades-Carrasco, L., Paya, V., Pellicer, A., & Pineda-Lucena, A. (2015). Nuclear magnetic resonance metabolomic profiling of urine provides a noninvasive alternative to the identification of biomarkers associated with endometriosis. *Fertility and Sterility, 104*(5), 1202–1209.

Vu, T. N., Valkenborg, D., Smets, K., Verwaest, K. A., Dommisse, R., Lemiere, F., et al. (2011). An integrated workflow for robust alignment and simplified quantitative analysis of NMR spectrometry data. *BMC Bioinformatics, 12*, 405.

Yang, M., Soga, T., & Pollard, P. J. (2013). Oncometabolites: Linking altered metabolism with cancer. *The Journal of Clinical Investigation, 123*(9), 3652–3658.

Zappa, M., Ciatto, S., Bonardi, R., & Mazzotta, A. (1998). Overdiagnosis of prostate carcinoma by screening: An estimate based on the results of the Florence Screening Pilot Study. *Annals of Oncology: Official Journal of the European Society for Medical Oncology/ESMO, 9*(12), 1297–1300.

Zaragoza, P., Ruiz-Cerda, J. L., Quintas, G., Gil, S., Costero, A. M., Leon, Z., et al. (2014). Towards the potential use of (1)H NMR spectroscopy in urine samples for prostate cancer detection. *The Analyst, 139*(16), 3875–3878.

Zhang, A., Yan, G., Han, Y., & Wang, X. (2014). Metabolomics approaches and applications in prostate cancer research. *Applied Biochemistry and Biotechnology, 174*(1), 6–12.

Zhang, J., Bowers, J., Liu, L., Wei, S., Gowda, G. A., Hammoud, Z., et al. (2012). Esophageal cancer metabolite biomarkers detected by LC-MS and NMR methods. *PLoS ONE, 7*(1), e30181.

Zhang, T., Watson, D. G., Wang, L., Abbas, M., Murdoch, L., Bashford, L., et al. (2013). Application of holistic liquid chromatography-high resolution mass spectrometry based urinary metabolomics for prostate cancer detection and biomarker discovery. *PLoS ONE, 8*(6), e65880.

Zhang, W. C., Shyh-Chang, N., Yang, H., Rai, A., Umashankar, S., Ma, S., et al. (2012). Glycine decarboxylase activity drives non-small cell lung cancer tumor-initiating cells and tumorigenesis. *Cell, 148*(1–2), 259–272.

Zhang, X., Xu, L., Shen, J., Cao, B., Cheng, T., Zhao, T., et al. (2013). Metabolic signatures of esophageal cancer: NMR-based metabolomics and UHPLC-based focused metabolomics of blood serum. *Biochimica et Biophysica Acta, 1832*(8), 1207–1216.

Zira, A. N., Theocharis, S. E., Mitropoulos, D., Migdalis, V., & Mikros, E. (2010). (1)H NMR metabonomic analysis in renal cell carcinoma: A possible diagnostic tool. *Journal of Proteome Research, 9*(8), 4038–4044.

Serum metabolomic profiling facilitates the non-invasive identification of metabolic biomarkers associated with the onset and progression of non-small cell lung cancer

Leonor Puchades-Carrasco[1], Eloisa Jantus-Lewintre[2], Clara Pérez-Rambla[1,2,3], Francisco García-García[4], Rut Lucas[2], Silvia Calabuig[2], Ana Blasco[5], Joaquín Dopazo[4,6,7], Carlos Camps[2,5,8] and Antonio Pineda-Lucena[1,3]

[1] Structural Biochemistry Laboratory, Centro de Investigación Príncipe Felipe, Valencia, Spain

[2] Molecular Oncology Laboratory, Fundación para la Investigación del Hospital General Universitario, Valencia, Spain

[3] Instituto de Investigación Sanitaria La Fe, Hospital Universitario i Politécnico La Fe, Valencia, Spain

[4] Computational Genomics Department, Centro de Investigación Príncipe Felipe, Valencia, Spain

[5] Department of Medical Oncology, Consorcio Hospital General Universitario, Valencia, Spain

[6] Bioinformatics of Rare Diseases (BIER), CIBER de Enfermedades Raras (CIBERER), Valencia, Spain

[7] Functional Genomics Node, Instituto Nacional de Bioinformática / Centro de Investigación Príncipe Felipe, Valencia, Spain

[8] Department of Medicine, Universitat de València, Valencia, Spain

***Correspondence to**: Antonio Pineda-Lucena, **email**: pineda_ant@gva.es*

Keywords: NSCLC, metabolomics, biomarkers, early diagnosis, prognosis

***Received**: January 07, 2015* ***Accepted**: January 27, 2016* ***Published**: February 12, 2016*

ABSTRACT

Lung cancer (LC) is responsible for most cancer deaths. One of the main factors contributing to the lethality of this disease is the fact that a large proportion of patients are diagnosed at advanced stages when a clinical intervention is unlikely to succeed. In this study, we evaluated the potential of metabolomics by ^{1}H-NMR to facilitate the identification of accurate and reliable biomarkers to support the early diagnosis and prognosis of non-small cell lung cancer (NSCLC).

We found that the metabolic profile of NSCLC patients, compared with healthy individuals, is characterized by statistically significant changes in the concentration of 18 metabolites representing different amino acids, organic acids and alcohols, as well as different lipids and molecules involved in lipid metabolism. Furthermore, the analysis of the differences between the metabolic profiles of NSCLC patients at different stages of the disease revealed the existence of 17 metabolites involved in metabolic changes associated with disease progression.

Our results underscore the potential of metabolomics profiling to uncover pathophysiological mechanisms that could be useful to objectively discriminate NSCLC patients from healthy individuals, as well as between different stages of the disease.

INTRODUCTION

Lung cancer (LC) is the most common cause of cancer death worldwide, accounting for approximately 12% of all cases of cancer, with an incidence of almost two million new cases annually worldwide [1]. The average five-year LC survival rate in early-stage, operable, non-small cell lung cancer (NSCLC) is approximately 50-70%. However, the five-year survival rate drops to 2-5% for patients diagnosed after their tumors have spread distantly [2]. At present, the diagnosis is primarily based on symptoms and detection often occurs at late stages, thus resulting in a very poor prognosis. If the diagnosis could be shifted to early stages, then the overall morbidity for this disease could be dramatically altered.

Recent studies have shown that LC screening using Low Dose Computed Tomography (LDCT) is effective in reducing mortality [3]. However, the large proportion of individuals with indeterminate nodules, the high costs involved and the limited resources available, demand the

identification of more accurate risk profiles, ideally based on non-invasive or minimally invasive techniques (i.e., blood, sputum, exhaled air, etc.) [4], in combination with other clinical, epidemiological, imaging, and life-style information. This strategy could be particularly relevant for individuals at-risk for LC as they may have subclinical disease for years before presentation.

Metabolomics, an analytical tool used in combination with pattern recognition approaches, is a very promising approach in systems biology; its objective being the comprehensive analysis of low-molecular weight metabolites in biological samples [5]. It represents a very powerful approach to the understanding of the biological pathways involved in the onset and progression of diseases, providing valuable insights into the molecular mechanisms of pathological processes [6]. The most commonly employed analytical techniques used for metabolic profiling are Nuclear Magnetic Resonance (^{1}H-NMR) spectroscopy and Mass Spectrometry (MS). High-resolution ^{1}H-NMR spectroscopy provides quantitative analysis of metabolite concentrations and reproducible information with minimal sample handling.

Monitoring specific metabolite levels in serum/plasma, the most commonly used biofluids in clinical metabolomics, has become an important tool for detecting early stages of some oncological diseases [7]. Thus, metabolomics by ^{1}H-NMR spectroscopy has been applied for the identification of different biomarkers in renal [8, 9], colorectal [10], pancreatic [11], ovarian [12, 13], and oral cancers [14], as well as in some hematological diseases [15, 16], among others.

In recent years, a number of studies have reported promising results in the characterization of the metabolic profile of LC patients [17-22]. However, a comprehensive approach to the early detection of this disease requires the extensive analysis of a more representative set of samples that could lead to the identification of specific and reliable clinical biomarkers. To that end, in this study, a thorough analysis of the serum metabolic profile of NSCLC patients at different stages of the disease was compared with that corresponding to healthy individuals and patients diagnosed with other benign pulmonary diseases (BPDs). Using a metabolomics approach based on ^{1}H-NMR, it was possible to identify and independently validate a set of selective and specific metabolites that could be useful for the early detection of LC in the clinical context. Taken together, the results provide an opportunity for improving current risk stratification models.

RESULTS

Non-supervised analysis of the serum samples from the training set reveals that disease status contributes to the metabolic discrimination of healthy individuals and NSCLC patients

Non-supervised analysis of the ^{1}H-NMR spectra (Supplementary Figure 1) was carried out to evaluate the potential influence of different clinical variables on the metabolic profiles obtained for the serum samples from the training set. Among all the variables assessed, only classification of the samples according to disease status had an impact in the clustering of the samples (Supplementary Figure 2).

The unsupervised analysis also highlighted the existence of differences between the two independent, training and validation, sets of samples included in the study (data not shown). An analysis of these differences revealed that they were attributed to technical variability, most likely reflecting the existence of differences in the suppression of the residual water signal of the spectra at the time of measurement.

Supervised analysis of the data reveals the existence of statistically significant differences between the metabolic profiles of NSCLC patients and healthy individuals, as well as between different disease stages

Discriminant statistical models (OPLS-DA) were built based on the comparisons between the different groups of samples included in the training set (Figure 1). This analysis revealed that serum samples from NSCLC patients, compared with healthy individuals, exhibit a specific serum metabolic profile ($R^2 = 0.931$; $Q^2 = 0.873$) characterized by statistically significant differences in the concentrations of a number of metabolites (Figure 1A). A similar analysis performed to compare the serum metabolomics profile of NSCLC patients at early and advanced stages of the disease ($R^2 = 0.779$; $Q^2 = 0.592$) showed that disease progression has also a reflection in the metabolic profile of patients (Figure 1D).

An analysis of the metabolic differences based on the results of the corresponding shared and unique structures plots (SUS-plots) (Figure 1E,1F) revealed that the most significant variations between the serum metabolic profile of NSCLC patients and the healthy individuals were shared regardless of the stage of the disease (Figure 1A,1B,1C), and that they were different from those found between early and advanced stages of NSCLC (Figure 1D).

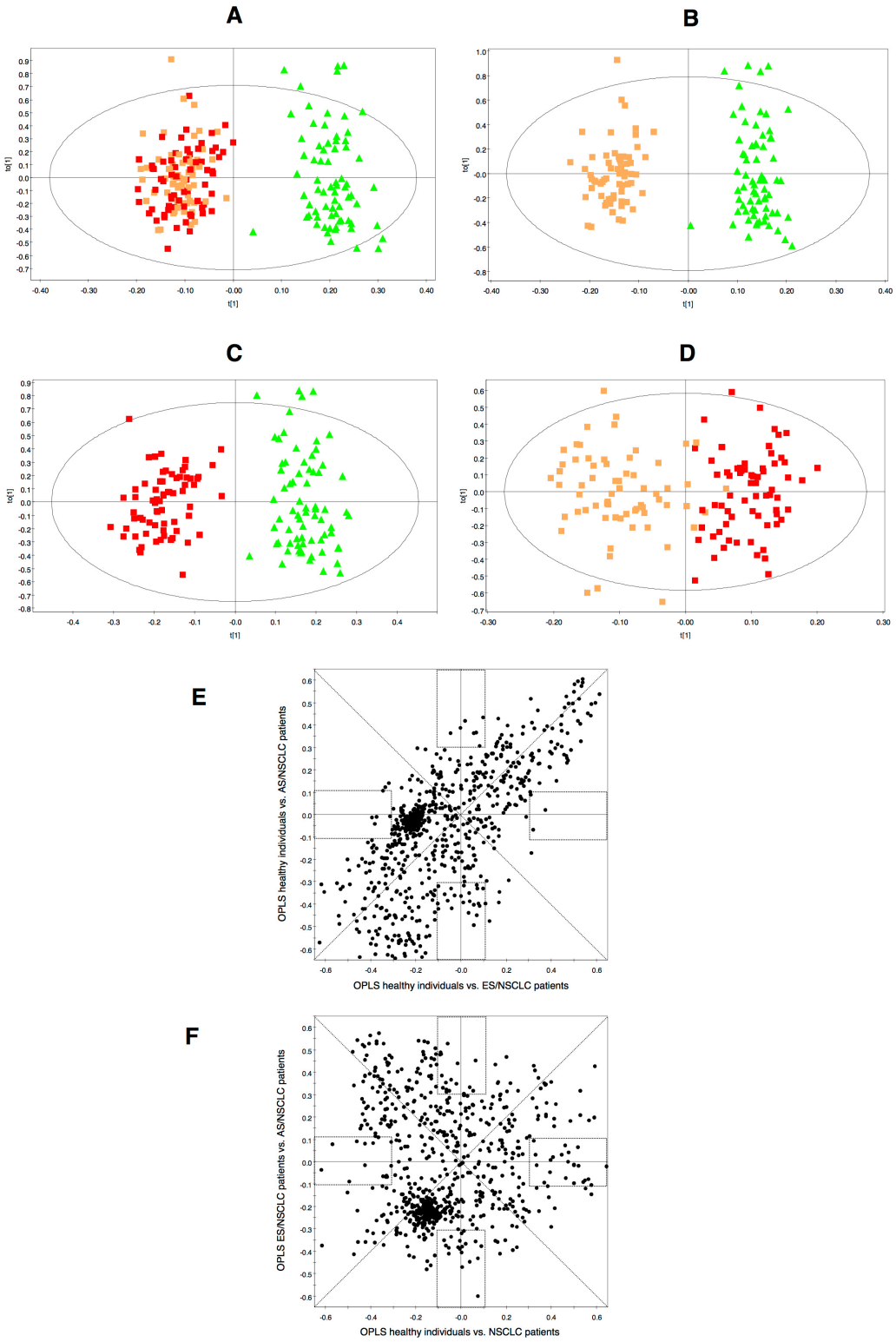


Figure 1: Multivariate modelling resulting from the analysis of serum ^{1}H-NMR spectra. OPLS-DA score plots for the comparisons between: **A.** healthy individuals (▲) *vs.* NSCLC patients (early-stage and advanced-stage, ■ and ■, respectively); **B.** healthy individuals (▲) *vs.* early-stage NSCLC patients (■); **C.** healthy individuals (▲) *vs.* advanced-stage NSCLC patients (■) and **D.** early-stage NSCLC patients (■) *vs.* advanced-stage NSCLC patients (■). SUS-plots derived from the OPLS-DA models between: **E.** healthy individuals *vs*. early-stage NSCLC patients *(model B, horizontal axis)* and healthy individuals *vs*. advanced-stage NSCLC patients *(model C, vertical axis)*; **F.** healthy individuals *vs*. NSCLC patients *(model A, horizontal axis)* and early-stage NSCLC *vs*. advanced-stage NSCLC patients *(model D, vertical axis)*. Rectangles indicate unique biomarkers for each model.

Table 1A: Mean intensities and variations of the most significant metabolites found in the comparison between healthy individuals and NSCLC patients

Metabolite	δ ^{1}H (ppm)[a]	Mean spectral intensity ± s.e.m. (arbitrary units)		% variation	*P*-value[b]
		Healthy Group	NSCLC patients		
HDL (CH_3)	0.85-0.79	55.14 ± 1.50	45.55 ± 0.73	-17.39	0.0000*
VLDL (CH_3)	0.91-0.85	66.10 ± 1.49	65.17 ± 0.78	-1.40	0.9070
Leucine/Isoleucine	0.97-0.91	14.40 ± 0.25	16.67 ± 0.26	15.75	0.0000*
3-hydroxybutyrate	1.20-1.18	8.22 ± 0.15	8.60 ± 0.20	4.59	0.8585
LDL/VLDL (CH_2)$_n$	1.31-1.20	229.06 ± 6.10	197.77 ± 3.26	-13.66	0.0001*
Unknown 1	1.40-1.36	4.01 ± 0.11	5.04 ± 0.14	25.65	0.0000*
Adipic acid	1.60-1.52	15.97 ± 0.69	13.79 ± 0.34	-13.64	0.0406*
Acetate	1.91-1.90	1.29 ± 0.04	1.56 ± 0.04	21.22	0.0001*
Lipids (CH_2-C=C)	2.02-1.93	40.15 ± 0.46	35.56 ± 0.31	-11.44	0.0000*
N-Acetyl-cysteine	2.05-2.02	23.50 ± 0.32	27.01 ± 0.44	14.94	0.0000*
Lipids (CH_2-CO)	2.26-2.19	14.19 ± 0.65	13.17 ± 0.30	-7.15	0.5775
Glutamate	2.36-2.33	2.14 ± 0.07	2.95 ± 0.07	37.65	0.0000*
Glutamine	2.47-2.41	5.82 ± 0.19	4.98 ± 0.11	-14.37	0.0002*
Choline–N(CH_3)$_3^+$	3.21-3.18	23.56 ± 0.67	17.60 ± 0.34	-25.30	0.0000*
Methanol	3.36-3.35	1.01 ± 0.05	1.81 ± 0.05	78.81	0.0000*
Glycerol	3.80-3.78	2.90 ± 0.09	3.63 ± 0.08	25.16	0.0000*
Creatine	3.92-3.91	0.90 ± 0.04	1.20 ± 0.03	33.82	0.0000*
Lactate	4.13-4.08	10.45 ± 0.38	13.74 ± 0.42	31.53	0.0000*
Threonine	4.30-4.21	6.92 ± 0.12	5.86 ± 0.09	-15.34	0.0000*
Glucose	5.24-5.21	11.92 ± 0.35	12.63 ± 0.26	5.92	0.2202
Lipids (CH=CH)	5.37-5.24	32.92 ± 0.75	28.08 ± 0.45	-14.70	0.0000*
Histidine	7.78-7.74	0.37 ± 0.01	0.29 ± 0.01	-20.76	0.0000*

[a]Chemical shift region used for quantification
[b]*P*-value calculated using the Mann–Whitney-Wilcoxon test
*Statistically significant ($P < 0.05$)

Specific combinations of metabolites are involved in the discrimination between healthy individuals and NSCLC patients, and between different stages of NSCLC

An inspection of the contribution to the separation between groups resulted in the identification of the spectral signals, and eventually the metabolites, that contributed more to the discrimination between the groups of samples being compared. Using this approach, a total of 18 metabolites showed statistically significant differences when comparing the serum metabolite levels of NSCLC patients and healthy individuals (Table 1A, Supplementary Figure 3A), and 17 when comparing NSCLC patients at early and advanced stages of the disease (Table 1B, Supplementary Figure 3B).

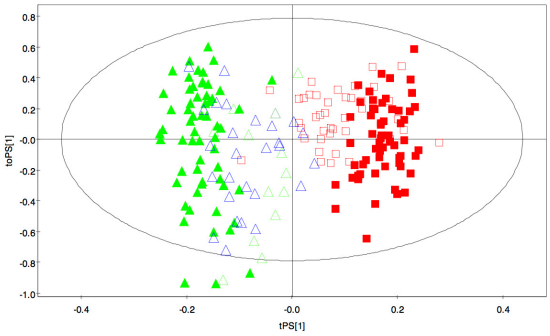
A logistic regression analysis of the data identifies a minimal set of metabolites involved in the discrimination of NSCLC patients and healthy individuals

A logistic regression equation was obtained based on the analysis of the metabolites exhibiting significant statistical differences between both groups (Table 1A). Using this approach, characteristic higher levels of lactate and methanol and lower levels of glutamine, choline and threonine were found in serum samples from NSCLC patients compared with healthy individuals (Table 2). Internal validation of the logistic regression equation was performed by evaluating the AUC values of the ROC curves for each individual metabolite (Supplementary Figure 4).

A

Validation set	Predicted as NSCLC	Predicted as healthy	Correctly classified
NSCLC patients (40)	38	2	95%
Healthy group (13)	1	12	92.31%
BPD group (27)	4	23	

B



C

Validation set	Predicted as NSCLC	Predicted as healthy	Correctly classified
NSCLC patients (40)	**31**	9	77.50%
Healthy group (13)	3	**10**	76.90%
BPD group (27)	9	18	

D

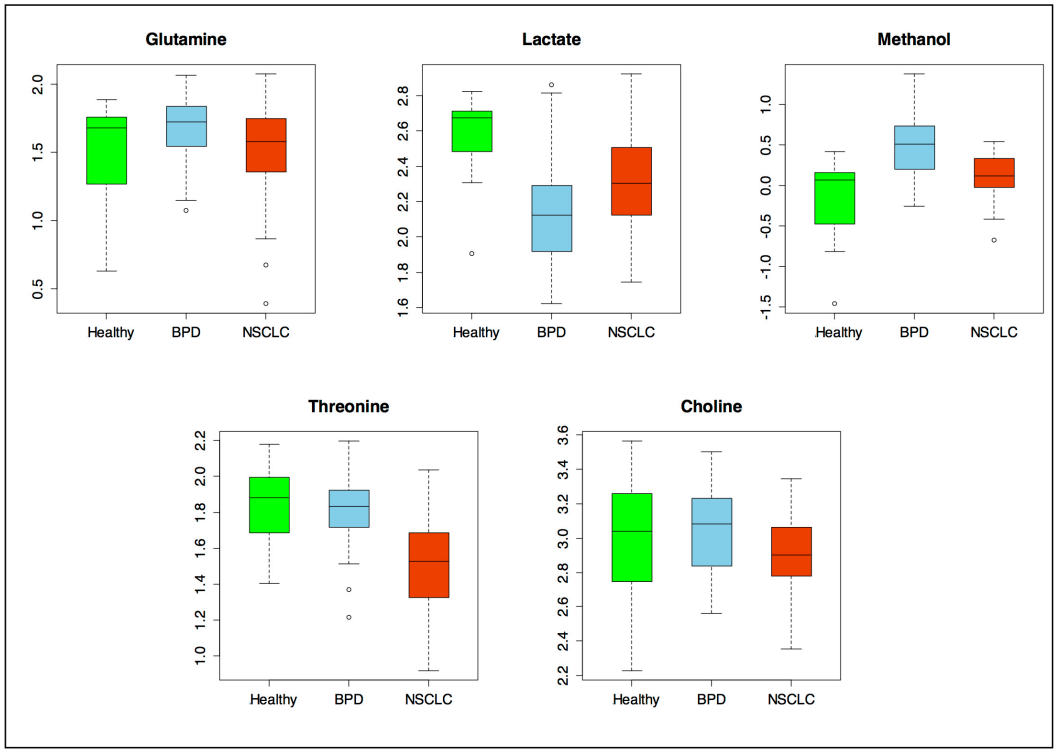


Figure 2 : A. Prediction results derived from the OPLS-DA model corresponding to the comparison between healthy individuals and NSCLC patients (training set). **B.** OPLS-DA score plot displaying the prediction of the samples included in the validation set based on the model corresponding to the training set (△: healthy individuals -validation set-; □: NSCLC patients -validation set-; △: BPD patients -validation set-; ▲: healthy individuals -training set-; ■: NSCLC patients -training set-). **C.** Misclassification table based on the logistic regression equation. **D.** Boxplot (log scale) representing the intensities of the metabolites included in the logistic regression equation for the different groups. For each box, the central line is the median, the edges of the box are the upper and lower quartiles, the whiskers extend the box by a further ±1.5 interquartile range (IQR), and outliers are plotted as individual points.

Table 1B: Mean intensities and variations of the most significant metabolites found in the comparison between early-stage and advanced-stage NSCLC patients

Metabolites	δ ^{1}H (ppm)[a]	Mean spectral intensity ± s.e.m. (arbitrary units)		% variation	*P*-value[b]
		ES/NSCLC	AS/NSCLC		
HDL (CH_3)	0.85-0.79	45.72 ± 1.08	45.70±1.01	-0.06	0.9054
VLDL (CH_3)	0.91-0.85	65.66 ± 0.94	64.71 ± 1.23	-1.43	0.8602
Leucine/Isoleucine	0.97-0.91	15.89 ± 0.34	17.41 ±0.37	9.59	0.0039*
3-hydroxybutyrate	1.20-1.18	9.38 ± 0.31	7.85 ± 0.21	-16.26	0.0000*
LDL/VLDL $(CH_2)_n$	1.31-1.20	203.92 ± 4.48	191.89 ± 4.65	-5.90	0.1151
Adipic acid	1.60-1.52	14.61 ± 0.50	13.00 ± 0.44	-11.01	0.0248*
Lysine	1.75-1.65	9.15 ± 0.26	10.36 ± 0.26	13.16	0.0025*
Lipids (CH_2-C=C)	2.02-1.93	35.27 ± 0.40	35.84 ±0.48	1.60	0.1403
N-Acetyl-cysteine	2.05-2.02	24.51 ± 0.50	29.39 ± 0.59	19.90	0.0000*
Glutamate	2.09-2.05	10.09 ± 0.24	12.30 ± 0.27	21.86	0.0000*
Lipids (CH_2-CO)	2.26-2.19	14.55 ± 0.42	11.85 ± 0.37	-18.54	0.0000*
Glutamine	2.47-2.41	5.91 ± 0.17	5.03 ± 0.15	-14.84	0.0005*
Citrate	2.96-2.64	4.66 ± 0.11	5.41 ± 0.11	16.26	0.0000*
Choline–$N(CH_3)_3^+$	3.21-3.18	18.05 ± 0.50	17.17 ± 0.46	-4.86	0.2081
Unknown 2	3.55-3.54	3.22 ± 0.09	3.40 ± 0.09	5.76	0.1852
Unknown 3	3.58-3.56	4.98 ± 0.17	5.97 ± 0.16	19.71	0.0000*
Valine	3.61-3.59	3.55 ± 0.10	4.30 ± 0.10	21.05	0.0000*
Glycerol	3.67-3.63	13.49 ± 0.38	15.76 ± 0.38	16.82	0.0001*
Creatine	3.92-3.91	1.25 ± 0.04	1.43 ± 0.04	14.43	0.0031*
H α/β amino acids	4.02-3.92	8.86 ± 0.26	10.54 ± 0.27	18.92	0.0000*
Lactate	4.13-4.08	13.01 ± 0.52	14.59 ± 0.64	12.15	0.0274*
Glucose	5.24-5.21	13.06 ± 0.37	11.86 ± 0.37	-9.19	0.0058*
Lipids (CH=CH)	5.37-5.24	28.89 ± 0.58	27.51 ± 0.70	-4.79	0.2948
Phenylalanine	7.43-7.40	0.31 ± 0.02	0.47 ± 0.03	52.10	0.0000*

[a]Chemical shift region used for quantification
[b]*P*-value calculated using the Mann–Whitney-Wilcoxon test
*Statistically significant ($P < 0.05$)

Table 2: Characteristics of the logistic regression equation obtained for the discrimination between healthy individuals and NSCLC patients

Metabolite	β[a]	OR[b]	1/OR	*P*-value
Glutamine	-1.70	0.18	5.47	0.0032*
Choline	-0.41	0.66	1.51	0.0011*
Methanol	4.60	99.63	0.01	0.0001*
Lactate	0.34	1.41	0.71	0.0347*
Threonine	-1.82	0.16	6.19	0.0036*
Constant	17.70	4.89E+07	2.05E-08	0.0032*

[a]β: Coefficient of logistic regression
[b]OR: odds ratio
*Statistically significant ($P < 0.05$)

Table 3: Variations of the mean intensities of the most relevant metabolites involved in the discrimination between BPD patients and healthy individuals or NSCLC patients

Metabolites	δ ^{1}H (ppm)[a]	BPD *vs.* Healthy		BPD *vs.* NSCLC	
		% variation	*P*-value[b]	% variation	*P*-value[b]
HDL (CH_3)	0.85-0.79	-6.55	0.3602	-9.37	0.1073
VLDL (CH_3)	0.91-0.85	-0.19	0.7539	-16.47	0.0002*
Leucine/Isoleucine	0.97-0.91	2.52	0.9319	8.54	0.0360*
3-hydroxybutyrate	1.20-1.18	-2.98	0.5490	20.64	0.0735
LDL/VLDL $(CH_2)_n$	1.31-1.20	1.61	0.9319	-23.28	0.0000*
Unknown 1	1.40-1.36	34.21	0.1424	-7.80	0.0676
Adipic acid	1.60-1.52	4.30	0.8867	-28.57	0.0001*
Acetate	1.91-1.90	-15.53	0.7979	6.96	0.0078*
Lipids (CH_2-C=C)	2.02-1.93	-6.28	0.1424	-15.00	0.0000*
N-Acetyl-cysteine	2.05-2.02	-8.02	0.0392*	-6.80	0.0152*
Lipids (CH_2-CO)	2.26-2.19	4.63	0.8867	-12.33	0.1674
Glutamate	2.36-2.33	-7.84	0.2175	43.35	0.0000*
Glutamine	2.47-2.41	-15.34	0.1065	-14.08	0.0136*
Lipids (CH=CH-CH2-CH=CH)	2.79-2.68	-0.06	0.6077	-15.40	0.0000*
Choline –$N(CH_3)_3^+$	3.21-3.18	-3.46	0.8197	-14.45	0.0461*
Proline	3.30-3.34	-27.10	0.0130*	-15.46	0.0434*
Methanol	3.36-3.35	-45.45	0.0010*	-34.12	0.0002*
Unknown 2	3.55-3.54	-10.03	0.3166	17.93	0.0091*
Unknown 3	3.58-3.56	-13.86	0.2765	9.10	0.1714
Valine	3.61-3.59	-11.64	0.2068	9.14	0.2150
Glycerol	3.80-3.78	-7.20	0.4406	17.94	0.0027*
Creatine	3.92-3.91	-12.39	0.2175	22.52	0.0005*
Creatinine	4.04-4.03	-0.24	0.6898	-24.18	0.0000*
Myo-inositol	4.07-4.05	10.28	0.4578	-33.58	0.0000*
Lactate	4.13-4.08	49.44	0.0002*	16.92	0.0136*
Threonine	4.30-4.21	4.85	0.5878	-26.09	0.0000*
Glucose	5.24-5.21	-5.59	0.2639	42.78	0.0000*
Lipids (CH=CH)	5.37-5.24	2.23	0.3602	-21.01	0.0000*

[a]Chemical shift region used for quantification
[b]*P*-value calculated using the Mann–Whitney U test
*Statistically significant ($P < 0.05$)

External validation of the predictive ability of the OPLS-DA and the logistic regression models

Samples included in the validation set were used to assess the predictive ability of the orthogonal projection to latent structures discriminant analysis (OPLS-DA) and the logistic regression models based on the training set. To remove the variation between the NMR data obtained for the two sets of samples, standardization of NMR signal intensities for the training and the validation sets of samples was achieved using the *ComBat* method [23].

Thus, it was found that, based on the OPLS-DA model obtained for the training set, 95% of patients diagnosed with NSCLC, as well as all but one of the healthy individuals included in the validation set, were correctly classified. An evaluation of the behavior of the serum samples obtained from patients diagnosed with BPDs was also carried out. In this case, it was found that 23 out of the 27 BPD samples (85.2%) were classified as healthy individuals. Overall, the multivariate statistical model obtained for the discrimination between NSCLC patients and healthy individuals was 95% specific and 92.31% sensitive (87.50% for all non-cancer samples) (Figure 2A, 2B).

The probability of belonging to the group of NSCLC patients for the samples included in the validation set was also evaluated using the logistic regression equation.

Unsurprisingly, the ability of the prediction based on the logistic regression model was lower (Figure 2C) than that based on the OPLS-DA multivariate statistical model, since the latter includes information about all the regions of the spectra. Overall, 77.3% of samples in the validation set were correctly classified; 77.5% of the NSCLC patients and 76.9% of samples of the healthy individuals included in the validation set were correctly classified (70% for all non-cancer samples).

The decrease in the percentage of BPD samples classified as healthy individuals (85.2% (OPLS-DA model) *versus* 66.6% (logistic regression model)) was further investigated by evaluating the levels of the five metabolites from the logistic regression equation in the samples from the validation set (Figure 2D). This analysis revealed that patients diagnosed with BPDs, compared with healthy individuals, exhibit statistically significant higher levels of methanol. These results prompted the analysis of the potential differences existing between patients diagnosed with BPDs and healthy volunteers or NSCLC patients to get a better understanding of the metabolic changes that were specific of NSCLC patients and those ones shared with BPDs.

Patients diagnosed with BPDs have a metabolic profile different from both healthy individuals and NSCLC patients

OLPS-DA statistical models were thus generated to compare the metabolic profiles of patients diagnosed with BPDs and NSCLC patients (R^2 = 0.972; Q^2 = 0.856) or healthy individuals (R^2 = 0.963; Q^2 = 0.782) (Supplementary Figure 5A,5B). The analysis of the contribution coefficients of each spectral region in these two statistical models, together with the SUS-plots obtained for the comparison of the three OPLS-DA models (Supplementary Figure 5A,5B,5C) allowed the identification of the shared and unique metabolic differences relevant in each statistical model (Supplementary Figure 5D, 5E, 5F). This analysis revealed that the serum metabolic profile of BPD patients is characterized by statistically significant higher levels of methanol and lower levels of lactate compared with healthy individuals, and statistically significant higher levels of methanol, choline and LDL/VLDL and lower levels of lactate and glucose compared with NSCLC patients (Table 3).

DISCUSSION

Efforts to identify NSCLC biomarkers that could help to better understand disease pathogenesis and to effectively identify patients at early stages of the disease remain a fundamental goal in this area [24]. In this context, our study represents the first attempt, based on the analysis of a significant number of samples, to characterize and compare the specific serum metabolic profile of NSCLC patients at different stages of the disease with those of healthy individuals and patients diagnosed with different BPDs.

The presence of lower levels of high-density lipoprotein / low-density lipoprotein / very-low-density lipoprotein (HDL/LDL/VLDL) in NSCLC patients correlates well with the relationship between decreased serum lipid levels and the development of some oncological processes [25]. Variations in the lipid levels in oncological patients have been previously associated with an increased uptake of cholesterol, an essential component of cell membranes, by tumor cells [16]. Changes in lipid metabolism could also explain the variations in the concentration of adipic acid, a metabolite that is associated with abnormalities in the metabolism of fatty acids, in patients at different stages of the disease [26]. The lower level of serum choline, a precursor of membrane phospholipids, observed in the group of patients with NSCLC could also be associated with the increased demand of this metabolite by tumor cells due to the their high proliferation rate [27, 28]. Serum metabolic profile of NSCLC patients is also characterized by significantly higher levels of lactate and lower levels of glucose. These results are consistent with the increased uptake of glucose and its conversion to lactate described in various tumor tissues [29, 30], a phenomenon associated with the well-known Warburg effect [31].

Previous studies have reported significant alterations in the serum amino acid profile of cancer patients, most probably reflecting the hypermetabolic state and increased demand of amino acids during tumor development [32-34]. Our data show that NSCLC patients, compared with healthy individuals, exhibit higher serum levels of leucine/isoleucine (15.75%), N-acetyl-cysteine (14.94%) and glutamate (37.65%), and lower levels of glutamine (-14.37%), threonine (-15.34%) and histidine (-20.76%). Increased concentrations of leucine/isoleucine, N-acetyl-cysteine and glutamate and decreased concentrations of glutamine are also observed when the serum metabolic profiles of NSCLC patients at early and advanced stages of the disease are compared. The specific decrease of serum threonine and histidine levels observed in NSCLC patients, compared with healthy individuals, most probably reflects the up-regulation of the glycine/serine/threonine and pyrimidine metabolic pathways, respectively, that have been described as metabolic hallmarks of NSCLC tumor-initiating cells [35]. Furthermore, disease progression is characterized by a specific increase in the serum concentrations of lysine (13.16%), valine (21.05%) and phenylalanine (52.10%). The significant increase in the serum concentration of phenylalanine is in agreement with the down-regulation of gene modules involved in phenylalanine metabolism observed in tissue samples from NSCLC patients [36], and could reflect a limited ability of

Table 4: Clinical and demographic characteristics of the individuals included in the study

	Training data set			Validation set			
	Healthy	**ES/NSCLC**	**AS/NSCLC**	**Healthy**	**BPD**	**ES/ NSCLC**	**AS/NSCLC**
Total number	74	72	70	13	27	20	20
Gender							
Female	13	8	12	2	13	5	7
Male	61	64	58	11	14	15	13
Age ± s.e.m[a]	56 ± 1.55	63 ± 1.17	63 ± 1.29	47 ± 1.78	52 ± 2.88	68 ± 1.48	61 ± 2.28
Smoking habits							
Ex-smoker	22	25	21	1	8	9	10
Smoker	31	42	20	8	7	8	4
Non-smoker	21	5	5	3	11	3	6
Unknown			24	1	1		
Histology							
Adenocarcinoma		24	32			9	16
Large-cell carcinoma		2	4			1	1
Squamous-cell carcinoma		38	27			9	
Other or unspecified		8	7			1	3
Stage							
IA		10				5	
IB		27				4	
IIA		1				5	
IIB		17				4	
IIIA		17				2	
IIIB			18				
IV			52				20
Other pathology							
COPD					9		
TBC					3		
Pneumonia					4		
CB					2		
Other					9		

Abbreviations: ES/NSCLC, early-stage non-small cell lung cancer; AS/NSCLC, advanced-stage non-small cell lung cancer; BPD, benign pulmonary diseases; COPD, chronic obstructive pulmonary disease; TBC, tuberculosis; CB, chronic bronchitis.

[a]Age=mean years at time of sample collection ± s.e.m (standard error of mean).

lung cancer cells to process this amino acid at advanced stages of the disease. The decrease in serum glutamine levels in NSCLC patients is consistent with other cancer studies [11, 37, 38] where it has been associated with increased metabolic activity derived from the conditions of hypoxia and hypermetabolism observed in the tumor environment [39]. A recent study has also revealed the important role that glutamine, as a nitrogen source for the synthesis of nucleotides and amino acids, plays in these conditions [40]. In this context, the hydrolysis of glutamine for the production of ammonia and glutamate to balance the pH in tumor cells could explain the high serum levels of glutamate observed in the group of NSCLC patients. Interestingly, the sustained increase in N-acetyl-cysteine levels suggest that metabolic pathways leading to the production of antioxidant species are up-regulated in NSCLC patients. This finding provides further support to a recent study conducted in a genetically engineered mouse model that mimics early human NSCLC [41]. In this study, authors concluded that antioxidants play an important role in LC progression by reducing the expression of p53, a key tumor suppressor protein.

The increase in creatine levels deserves special attention as a chemically related metabolite, creatine riboside, was recently associated [42] with NSCLC in a urinary metabolomics study. This metabolite was found to be elevated in the urine of NSCLC patients and associated with poor prognosis. Creatine is transformed into phosphocreatine, an energy reservoir, by creatine kinase isoenzyme BB, an enzyme that has been shown to exhibit high serum levels in NSCLC patients [43, 44]. Therefore, the observation of elevated levels of creatine in NSCLC

patients compared with healthy individuals, as well as between different disease stages, could be associated with the high metabolic activity of this neoplastic process.

Finally, our metabolomics study reveals that there are significant statistical differences between the serum metabolic profile of patients diagnosed with BPDs and healthy individuals ($R^2 = 0.963$; $Q^2 = 0.782$) or NSCLC patients ($R^2 = 0.972$; $Q^2 = 0.856$). Our results are partially in agreement with a recent NMR metabolomics study carried out using serum samples from NSCLC and chronic obstructive pulmonary disease (COPD) patients [20]. Thus, Deja *et al*. reported that the serum metabolic profile of NSCLC patients, compared with patients diagnosed with COPD, was characterized by higher levels of lactate and lower levels of methanol [42]. Our results also show that the serum metabolic profile of BPD patients, compared with NSCLC patients, is effectively characterized by higher levels of methanol and lower levels of lactate. In contrast, they observed higher levels of choline in serum samples from NSCLC patients, and we report lower levels of choline in serum of NSCLC patients, our results being in agreement with previous results obtained from the analysis of tissue samples from LC tumors [45].

Overall, our results show that NSCLC patients, compared with healthy individuals and patients diagnosed with BPDs, exhibit characteristic serum metabolic profiles, and that disease stage has also a significant impact in the serum metabolic profile of patients. A specific combination of five metabolites, based on a logistic regression analysis, is presented, enabling the discrimination between healthy individuals, BPD patients and NSCLC patients with a 77.5% specificity and a 76.9% sensitivity (70% for all non-cancer samples). The combination of all the metabolites involved in the discrimination between healthy individuals and NSCLC patients should also be explored as they provide a specific signature, both in terms of magnitude and change direction, of the metabolic alterations responsible for the onset/progression of NSCLC with a 95% specificity and 92.31% sensitivity (87.50% for all non-cancer samples). The strategy described in this work provides a sensitive, specific, and minimally invasive method that may aid in the early diagnosis and staging of NSCLC and the optimization of current risk stratification models.

MATERIALS AND METHODS

Patient cohorts

A total of 296 serum samples were analyzed by ^{1}H-NMR (Table 4). Samples from NSCLC patients were classified into two groups [46]:

- Advanced stage NSCLC: Patients diagnosed with advanced NSCLC (stage IIIB with pleural effusion or stage IV, non-squamous histologies) with no other concomitant malignancies [47, 48]. Samples were obtained prior to chemotherapy.

- Early stage NSCLC: Newly diagnosed patients with resectable NSCLC (stage IA-IIIA) without prior chemotherapy. A pre-surgery serum sample was collected from each patient.

Furthermore, the study included two control groups: 87 serum samples from healthy individuals without any acute or chronic inflammatory conditions, and a group 27 serum samples from patients diagnosed with BPDs in the validation set.

Patient recruitment and sampling procedures were performed in accordance with the Declaration of Helsinki and applicable local regulatory requirements and laws and after approval from the Ethics Committees of all participating institutions.

Sample preparation and ^{1}H-NMR acquisition

Serum samples were immediately stored at -80ºC after collection. At the time of NMR analysis, samples were thawed on ice. 300 µL of 10% D_2O buffer (5 mM TSP, 140 mM Na_2HPO_4, 0.04% NaN_3, pH 7.4) were added to 300 µL of serum. ^{1}H-NMR spectra were acquired using a Bruker Avance II 600 MHz spectrometer equipped with triple resonance cryo-probe with a cooled ^{13}C preamplifier (TCI) at 310 K (37ºC) [49, 50]. Metabolites of interest were identified using Amix v 3.9.7 in combination with the Bruker NMR Metabolic Profiling Database BBIOREFCODE 2.0.0 database (Bruker Biospin, Rheinstetten, Germany), as well as other existing public databases and literature reports [12, 22, 51]. NMR experiments for each set were independently acquired at two different times.

Multivariate statistical analysis

^{1}H-NMR spectra were binned using Amix 3.9.7 (Bruker Biospin, Rheinstetten, Germany) over the region δ 9.02-0.14 ppm. The water (δ 5.06-4.30 ppm) and urea signal (δ 5.85-5.60 ppm) regions were excluded from the analysis to avoid interference arising from differences in water suppression and variability from urea signal, respectively. All bucket intensities were normalized to the total area of the corresponding spectra. Bucket tables generated were imported into SIMCA-P 12.0 software (Umetrics AB, Sweden). Prior to statistical analysis, data were Pareto scaled. The *ComBat* method, included in the "sva" R package [52], was applied to compensate differences due to batch effects.

PCA was used to examine the intrinsic variability within the data set, to observe clustering or separation trends and for the identification of outliers. OPLS-DA

was applied to minimize the possible contribution of inter-group variability and to further improve separation between the groups of samples. The default method of 7-fold internal cross validation was applied, from which Q^2Y (predictive ability parameter, estimated by cross-validation) and R^2Y (goodness of fit parameter) values were extracted. Those parameters, together with the corresponding permutation tests (n = 100), were used for the evaluation of the quality of the OPLS-DA models obtained. SUS-plots were also obtained to evaluate the shared (metabolites aligned with the diagonals) and unique differences (metabolites aligned with the axes) found when comparing two OPLS-DA statistical models.

Quantitative analysis of selected metabolites

The main metabolites contributing to group discrimination in each model were integrated using Amix 3.9.7. Normality in variable distribution was assessed using the Kolmogorov-Smirnov test. Statistical significance was assessed using the Mann-Whitney U test. A P-value < 0.05 (confidence level 95%) was considered statistically significant.

Logistic regression

Logistic regression analysis was performed using the "stats" R package [53]. Univariate logistic regression was carried out with the *Introduction* method, and the *Forward stepwise regression* method was used for the multivariate logistic regression. Odds ratio (OR) values were calculated for all the variables included in the equation. A P-value < 0.05 (confidence level 95%) was considered statistically significant.

ACKNOWLEDGMENTS

The authors thank Mr Jacobo Martínez (Red de Biobancos de Valencia) for his involvement in the selection of healthy volunteers and sample collection, and the Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERES) for providing serum samples from patients diagnosed with benign pulmonary diseases. We also thank Umetrics support for technical advice.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

GRANT SUPPORT

This study was supported by grants from the Spanish Ministerio de Economía y Competitividad (SAF2014-53977-R, RD12/0036/0025), the Red Temática de Investigación Cooperativa en Cáncer (RTICC), the Instituto de Salud Carlos III (ISCIII), and Fundación Mutua Madrileña.

REFERENCES

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D and Bray F. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. International journal of cancer. 2015; 136:E359-386.
2. Goldstraw P, Crowley J, Chansky K, Giroux DJ, Groome PA, Rami-Porta R, Postmus PE, Rusch V, Sobin L, International Association for the Study of Lung Cancer International Staging C and Participating I. The IASLC Lung Cancer Staging Project: proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edition of the TNM Classification of malignant tumours. Journal of thoracic oncology. 2007; 2:706-714.
3. National Lung Screening Trial Research T, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM and Sicks JD. Reduced lung-cancer mortality with low-dose computed tomographic screening. The New England journal of medicine. 2011; 365:395-409.
4. Jantus-lewintre E, Usó M, Sanmartín, E, Camps C. Update on biomarkers for the detection of lung cancer. Lung Cancer: Targets and Therapy. 2012; :1-9.
5. Holmes E, Wilson ID and Nicholson JK. Metabolic phenotyping in health and disease. Cell. 2008; 134:714-717.
6. Puchades-Carrasco L and Pineda-Lucena A. Metabolomics in pharmaceutical research and development. Curr Opin Biotechnol. 2015; 35:73-77.
7. Kobayashi T, Nishiumi S, Ikeda A, Yoshie T, Sakai A, Matsubara A, Izumi Y, Tsumura H, Tsuda M, Nishisaki H, Hayashi N, Kawano S, Fujiwara Y, Minami H, Takenawa T, Azuma T, et al. A novel serum metabolomics-based diagnostic approach to pancreatic cancer. Cancer epidemiology, biomarkers & prevention. 2013; 22:571-579.
8. Kind T, Tolstikov V, Fiehn O and Weiss RH. A comprehensive urinary metabolomic approach for identifying kidney cancer. Analytical biochemistry. 2007; 363:185-195.
9. Kim K, Aronov P, Zakharkin SO, Anderson D, Perroud B, Thompson IM and Weiss RH. Urine metabolomics analysis for kidney cancer detection and biomarker discovery. Molecular & cellular proteomics. 2009; 8:558-570.
10. Zhu J, Djukovic D, Deng L, Gu H, Himmati F, Chiorean EG and Raftery D. Colorectal cancer detection using targeted serum metabolic profiling. Journal of proteome research. 2014; 13:4120-4130.
11. Urayama S, Zou W, Brooks K and Tolstikov V. Comprehensive mass spectrometry based metabolic

profiling of blood plasma reveals potent discriminatory classifiers of pancreatic cancer. Rapid communications in mass spectrometry. 2010; 24:613-620.

12. Guan W, Zhou M, Hampton CY, Benigno BB, Walker LD, Gray A, McDonald JF and Fernandez FM. Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines. BMC bioinformatics. 2009; 10:259.
13. Odunsi K, Wollman RM, Ambrosone CB, Hutson A, McCann SE, Tammela J, Geisler JP, Miller G, Sellers T, Cliby W, Qian F, Keitz B, Intengan M, Lele S and Alderfer JL. Detection of epithelial ovarian cancer using ^{1}H-NMR-based metabonomics. International journal of cancer. 2005; 113:782-788.
14. Tiziani S, Lopes V and Gunther UL. Early stage diagnosis of oral cancer using ^{1}H NMR-based metabolomics. Neoplasia. 2009; 11:269-276, 264p following 269.
15. MacIntyre DA, Jimenez B, Lewintre EJ, Martin CR, Schafer H, Ballesteros CG, Mayans JR, Spraul M, Garcia-Conde J and Pineda-Lucena A. Serum metabolome analysis by ^{1}H-NMR reveals differences between chronic lymphocytic leukaemia molecular subgroups. Leukemia. 2010; 24:788-797.
16. Puchades-Carrasco L, Lecumberri R, Martinez-Lopez J, Lahuerta JJ, Mateos MV, Prosper F, San-Miguel JF and Pineda-Lucena A. Multiple myeloma patients have a specific serum metabolomic profile that changes after achieving complete remission. Clinical cancer research. 2013; 19:4770-4779.
17. Nobakht MGBF, Aliannejad R, Rezaei-Tavirani M, Taheri S and Oskouie AA. The metabolomics of airway diseases, including COPD, asthma and cystic fibrosis. Biomarkers. 2015; 20:5-16.
18. Rocha CM, Carrola J, Barros AS, Gil AM, Goodfellow BJ, Carreira IM, Bernardo J, Gomes A, Sousa V, Carvalho L and Duarte IF. Metabolic signatures of lung cancer in biofluids: NMR-based metabonomics of blood plasma. Journal of proteome research. 2011; 10:4314-4324.
19. Carrola J, Rocha CM, Barros AS, Gil AM, Goodfellow BJ, Carreira IM, Bernardo J, Gomes A, Sousa V, Carvalho L and Duarte IF. Metabolic signatures of lung cancer in biofluids: NMR-based metabonomics of urine. Journal of proteome research. 2011; 10:221-230.
20. Deja S, Porebska I, Kowal A, Zabek A, Barg W, Pawelczyk K, Stanimirova I, Daszykowski M, Korzeniewska A, Jankowska R and Mlynarz P. Metabolomics provide new insights on lung cancer staging and discrimination from chronic obstructive pulmonary disease. Journal of pharmaceutical and biomedical analysis. 2014; 100:369-380.
21. Jordan KW, Adkins CB, Su L, Halpern EF, Mark EJ, Christiani DC and Cheng LL. Comparison of squamous cell carcinoma and adenocarcinoma of the lung by metabolomic analysis of tissue-serum pairs. Lung cancer. 2010; 68:44-50.
22. Wang L, Tang Y, Liu S, Mao S, Ling Y, Liu D, He X and Wang X. Metabonomic profiling of serum and urine by (1)H NMR-based spectroscopy discriminates patients with chronic obstructive pulmonary disease and healthy individuals. PloS one. 2013; 8:e65675.
23. Johnson WE, Li C and Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007; 8:118-127.
24. Hassanein M, Callison JC, Callaway-Lane C, Aldrich MC, Grogan EL and Massion PP. The state of molecular biomarkers for the early detection of lung cancer. Cancer prevention research. 2012; 5:992-1006.
25. Body JJ. Metabolic sequelae of cancers (excluding bone marrow transplantation). Curr Opin Clin Nutr Metab Care. 1999; 2:339-344.
26. Muntoni S, Atzori L, Mereu R, Satta G, Macis MD, Congia M, Tedde A, Desogus A and Muntoni S. Serum lipoproteins and cancer. Nutrition, metabolism, and cardiovascular diseases. 2009; 19:218-225.
27. Banez-Coronel M, Ramirez de Molina A, Rodriguez-Gonzalez A, Sarmentero J, Ramos MA, Garcia-Cabezas MA, Garcia-Oroz L and Lacal JC. Choline kinase alpha depletion selectively kills tumoral cells. Current cancer drug targets. 2008; 8:709-719.
28. Gallego-Ortega D, Ramirez de Molina A, Ramos MA, Valdes-Mora F, Barderas MG, Sarmentero-Estrada J and Lacal JC. Differential role of human choline kinase alpha and beta enzymes in lipid metabolism: implications in cancer onset and treatment. PloS one. 2009; 4:e7819.
29. Chan EC, Koh PK, Mal M, Cheah PY, Eu KW, Backshall A, Cavill R, Nicholson JK and Keun HC. Metabolic profiling of human colorectal cancer using high-resolution magic angle spinning nuclear magnetic resonance (HR-MAS NMR) spectroscopy and gas chromatography mass spectrometry (GC/MS). Journal of proteome research. 2009; 8:352-361.
30. Yang Y, Li C, Nie X, Feng X, Chen W, Yue Y, Tang H and Deng F. Metabonomic studies of human hepatocellular carcinoma using high-resolution magic-angle spinning ^{1}H NMR spectroscopy in conjunction with multivariate data analysis. Journal of proteome research. 2007; 6:2605-2614.
31. Warburg O. On the origin of cancer cells. Science. 1956; 123:309-314.
32. Lai HS, Lee JC, Lee PH, Wang ST and Chen WJ. Plasma free amino acid profile in cancer patients. Seminars in cancer biology. 2005; 15:267-276.
33. Maeda J, Higashiyama M, Imaizumi A, Nakayama T, Yamamoto H, Daimon T, Yamakado M, Imamura F and Kodama K. Possibility of multivariate function composed of plasma amino acid profiles as a novel screening index for non-small cell lung cancer: a case control study. BMC cancer. 2010; 10:690.
34. Pisters PW and Pearlstone DB. Protein and amino acid metabolism in cancer cachexia: investigative techniques

and therapeutic interventions. Critical reviews in clinical laboratory sciences. 1993; 30:223-272.

35. Zhang WC, Shyh-Chang N, Yang H, Rai A, Umashankar S, Ma S, Soh BS, Sun LL, Tai BC, Nga ME, Bhakoo KK, Jayapal SR, Nichane M, Yu Q, Ahmed DA, Tan C, et al. Glycine decarboxylase activity drives non-small cell lung cancer tumor-initiating cells and tumorigenesis. Cell. 2012; 148:259-272.
36. Long F, Su JH, Liang B, Su LL and Jiang SJ. Identification of Gene Biomarkers for Distinguishing Small-Cell Lung Cancer from Non-Small-Cell Lung Cancer Using a Network-Based Approach. BioMed research international. 2015; 2015:685303.
37. Gao H, Dong B, Liu X, Xuan H, Huang Y and Lin D. Metabonomic profiling of renal cell carcinoma: high-resolution proton nuclear magnetic resonance spectroscopy of human serum with multivariate data analysis. Analytica chimica acta. 2008; 624:269-277.
38. Zira AN, Theocharis SE, Mitropoulos D, Migdalis V and Mikros E. (1)H NMR metabonomic analysis in renal cell carcinoma: a possible diagnostic tool. Journal of proteome research. 2010; 9:4038-4044.
39. Eigenbrodt E, Kallinowski F, Ott M, Mazurek S and Vaupel P. Pyruvate kinase and the interaction of amino acid and carbohydrate metabolism in solid tumors. Anticancer research. 1998; 18:3267-3274.
40. Zhou W, Capello M, Fredolini C, Racanicchi L, Piemonti L, Liotta LA, Novelli F and Petricoin EF. Proteomic analysis reveals Warburg effect and anomalous metabolism of glutamine in pancreatic cancer cells. Journal of proteome research. 2012; 11:554-563.
41. Sayin VI, Ibrahim MX, Larsson E, Nilsson JA, Lindahl P and Bergo MO. Antioxidants accelerate lung cancer progression in mice. Sci Transl Med. 2014; 6:221ra215.
42. Mathe EA, Patterson AD, Haznadar M, Manna SK, Krausz KW, Bowman ED, Shields PG, Idle JR, Smith PB, Anami K, Kazandjian DG, Hatzakis E, Gonzalez FJ and Harris CC. Noninvasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer. Cancer research. 2014; 74:3259-3270.
43. Neri B, Bartalucci S, Gemelli MT, Tommasi M and Bacalli S. Creatine kinase isoenzyme BB: a lung cancer associated marker. Int J Biol Markers. 1988; 3:19-22.
44. Gazdar AF, Zweig MH, Carney DN, Van Steirteghen AC, Baylin SB and Minna JD. Levels of creatine kinase and its BB isoenzyme in lung cancer specimens and cultures. Cancer research. 1981; 41:2773-2777.
45. Rocha CM, Barros AS, Gil AM, Goodfellow BJ, Humpfer E, Spraul M, Carreira IM, Melo JB, Bernardo J, Gomes A, Sousa V, Carvalho L and Duarte IF. Metabolic profiling of human lung cancer tissue by ^{1}H high resolution magic angle spinning (HRMAS) NMR spectroscopy. Journal of proteome research. 2010; 9:319-332.
46. Greene FL, Page D.L, Fleming I.D, Fritz A.G, Balch C.M, Haller D.G, Morrow, M. (2003). AJCC Cancer Staging Manual, 6th edition. (Philadelphia: Lippincott Raven Publishers).
47. Iranzo V, Sirera R, Bremnes RM, Blasco A, Jantus-Lewintre E, Taron M, Berrocal A, Blasco S, Caballero C, Del Pozo N, Rosell R and Camps C. Chemotherapy-induced neutropenia does not correlate with DNA repair gene polymorphisms and treatment efficacy in advanced non-small-cell lung cancer patients. Clinical lung cancer. 2011; 12:224-230.
48. Jantus-Lewintre E, Sirera R, Cabrera A, Blasco A, Caballero C, Iranzo V, Rosell R and Camps C. Analysis of the prognostic value of soluble epidermal growth factor receptor plasma concentration in advanced non-small-cell lung cancer patients. Clinical lung cancer. 2011; 12:320-327.
49. Meiboom S GD. Modified spin-echo method for measuring nuclear relaxation times. The Review of scientific instruments. 1958; 29:688-701.
50. Nicholson JK, Foxall PJ, Spraul M, Farrant RD and Lindon JC. 750 MHz ^{1}H and ^{1}H-^{13}C NMR spectroscopy of human blood plasma. Analytical chemistry. 1995; 67:793-811.
51. Lindon JC TG, Koppenaal D. (2010). Encyclopedia of spectroscopy and spectrometry. (London: Academic Press).
52. Leek JT, Johnson WE, Parker HS, Jaffe AE and Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012; 28:882-883.
53. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013; http://www.R-project.org/

OXFORD

Briefings in Bioinformatics, 2015, 1–12

doi: 10.1093/bib/bbv077
Paper

Bioinformatics tools for the analysis of NMR metabolomics studies focused on the identification of clinically relevant biomarkers

Leonor Puchades-Carrasco, Martina Palomino-Schätzlein, Clara Pérez-Rambla and Antonio Pineda-Lucena

Corresponding author. Antonio Pineda-Lucena, Instituto de Investigación Sanitaria La Fe, Avenida Fernando Abril Martorell 106, 46026 Valencia, Spain. Tel.: +34 961246600. E-mail: pineda_ant@gva.es

Abstract

Metabolomics, a systems biology approach focused on the global study of the metabolome, offers a tremendous potential in the analysis of clinical samples. Among other applications, metabolomics enables mapping of biochemical alterations involved in the pathogenesis of diseases, and offers the opportunity to noninvasively identify diagnostic, prognostic and predictive biomarkers that could translate into early therapeutic interventions. Particularly, metabolomics by Nuclear Magnetic Resonance (NMR) has the ability to simultaneously detect and structurally characterize an abundance of metabolic components, even when their identities are unknown. Analysis of the data generated using this experimental approach requires the application of statistical and bioinformatics tools for the correct interpretation of the results. This review focuses on the different steps involved in the metabolomics characterization of biofluids for clinical applications, ranging from the design of the study to the biological interpretation of the results. Particular emphasis is devoted to the specific procedures required for the processing and interpretation of NMR data with a focus on the identification of clinically relevant biomarkers.

Key words: Nuclear Magnetic Resonance; metabolomics; biomarker

Introduction

Metabolomics, a systems biology approach focused on the analysis of metabolites, is a promising tool for characterizing clinically relevant biomarkers in different clinical settings [1, 2]. Metabolomics provides a close view of an individual's phenotype, thus explaining the potential of this technology for identifying biomarkers that could be used for the early detection and diagnosis of different pathologies, monitoring disease progression and predicting therapeutic outcomes [3]. In particular, oncological processes, which are characterized by the dysregulation of multiple biochemical pathways, are extremely amenable to metabolomic studies. In this context, metabolomics has been shown to be useful in the identification of biomarkers associated to the diagnosis/prognosis of different oncological processes [4–6] and the response to treatment [7, 8].

Compared with other omics approaches, metabolomics exhibits a number of characteristics that make necessary the application of distinctive bioinformatics tools [9]. First of all, in a metabolomics study, the potential metabolic targets and associated effect sizes are initially unknown. This is in contrast with

Leonor Puchades Carrasco is a postdoctoral researcher at the Structural Biochemistry Laboratory, Centro de Investigación Príncipe Felipe. Her research projects include the evaluation of clinical applications of metabolomics in the oncology area.
Martina Palomino-Schätzlein is a postdoctoral researcher at the Structural Biochemistry Laboratory, Centro de Investigación Príncipe Felipe. Her research projects include the metabolomic analysis of different biological matrices.
Clara Pérez-Rambla is a predoctoral student at the Structural Biochemistry Laboratory, Centro de Investigación Príncipe Felipe. Her research projects include the characterization of pre-analytical variations in metabolomics studies.
Antonio Pineda-Lucena is the head of the Structural Biochemistry Laboratory, Centro de Investigación Príncipe Felipe. His research interests include structure-based drug design, computational chemistry and metabolomics.
Submitted: 7 April 2015; **Received (in revised form):** 29 July 2015

other omics approaches, such as genomics, transcriptomics or proteomics, where variables are, at least partially, known before the omics analysis is performed. These uncertainties have important implications in the estimation of an appropriate sample size to ensure reliable metabolomics results [10]. Second, the metabolome is considered to be more diverse and complex than the expected repertoire of molecules studied by other omics approaches [11, 12]. Metabolomics focuses on the characterization of small molecules from different chemical classes (e.g. amino acids, organic acids and lipids), each of them with different biophysical and biochemical properties. Small modifications in the chemical structure can dramatically change the function of a metabolite [9]. Therefore, the identification and biological evaluation of these compounds in a metabolomics study should be guided by robust and specific tools [13]. Third, metabolomics data are characterized by strong correlations between the variables. It is not uncommon that a single metabolite can display several signals in metabolomics phenotyping studies, a situation that is further complicated for the existence of connections between different biochemical pathways. Metabolic variables also exhibit different statistical variances, thus invalidating the application of many statistical methods that are currently used for the analysis of other omics approaches [10].

Mass Spectrometry and Nuclear Magnetic Resonance (NMR) spectroscopy are the two main analytical platforms used in metabolomics studies, and each of them has their own advantages and applications [9, 14]. In particular, NMR has the ability to provide a high degree of structural information in a short time. NMR is also a nondestructive and highly reproducible technique, and does not require an extensive sample preparation. However, NMR spectra of biofluids contain many signals, which can be overlapped and noisy, owing to the presence of hundreds of metabolites. In these conditions, the identification and quantification of metabolites is a complicated task that can not be easily automated. Therefore, careful data processing and statistical analysis are required to derive useful and reliable information from these profiles [15].

The specific nature of the metabolomics data and the NMR methods used for the analysis of biological samples require the application of different bioinformatics tools following a specific workflow (Figure 1). The aim of this review is to provide a summary of current available software and databases used in the analysis of NMR-based metabolomics data with a focus on the identification of clinically relevant biomarkers.

Design of a metabolomics study

An appropriate experimental design, which takes into account the specific characteristics of the disease and the objectives to be achieved, is essential in any clinical metabolomics study. Misleading statistical outcomes and subsequent erroneous biological and biochemical interpretations typically result from poor study design. Human biomarkers discovery studies are performed using a variety of experimental designs [16]. Retrospective case-control studies is the type of epimediological study most frequently used to identify biomarkers, by comparing patients who have a specific medical condition (cases) with individuals who do not have this condition but have other similar phenotypic and patient-specific characteristics (controls). On the other hand, longitudinal cohort studies allow patients to serve as their own biological control, which reduces the interindividual variability observed in multiple cohort studies [17].

The experimental design of a metabolomics study should also include the selection of a technological platform for performing the measurements and the consideration of different analytical aspects, such as instrument settings and calibration procedures, that can be the source of unwanted variations [9]. Finally, special attention should be paid to inconsistencies in sample handling and storage, and technical control of preanalytical sample variability that can lead to unexpected results [18].

The most critical factor, however, in any metabolomics study is the selection of the sample size, as it will have a tremendous impact on the statistical power of the results that can be derived from the analysis of the data. Metabolomics often relies on the identification of molecular signatures derived from a small set of samples relative to the number of molecular measurements [19]. This limitation makes metabolomics studies more sensitive to technical and biological sources of noise and variation, and less likely to capture information associated with the phenotype of interest [20]. The situation is further complicated by the intrinsic variability of metabolite levels over time for a given individual, a situation that can have tremendous impact on the interpretation of the results of a metabolomics study, and requires larger sample sizes to compensate [21]. Metabolomics is also characterized by weak effect sizes (e.g. odds ratios) as biomarkers are likely to occur in low relative abundance and be massively diluted in circulation [22]. Therefore, special attention should be paid during the experimental design to the detection of small effect sizes, and the potential integration of the metabolomics results with other omics technologies to produce robust models with larger effect sizes.

There are, however, inherent limitations to the selection of human samples for metabolomics studies, the most important being the prevalence of the disease. The heterogeneity of the different sample groups included in the studies also adds another layer of complexity to the experimental design of a metabolomics study. Thus, factors such as lifestyle, diet, smoking habit, drug treatments, exercise can lead to important biochemical differences, not associated to particular pathologies, between the patient groups [3]. Therefore, the design of a metabolomics study requires stringent definitions of inclusion/exclusion criteria for the individuals participating in the study, as well as a detailed description of their clinical data.

The final decision on the appropriate sample size will mainly depend on the expected statistical performance of the biomarkers. There exist several methods for sample size selection in high-throughput experiments. However, most of these methods are not suitable for metabolomics studies because they assume variables to have equal variance or to be independent. Multivariate analysis of variance, data simulation, as well as other approaches based on the calculation of receiver operating characteristic (ROC) curves [23], are often used to estimate sample sizes [24] in omics studies. Considering that many times the expectation is to achieve a fixed specificity with a minimum sensitivity, minimum sample sizes can also be estimated using simple inferential approaches [25]. The current version of MetaboAnalyst [26], a web-based server for the comprehensive analysis of metabolomics data, includes a new module, based on the Bioconductor package SSPA [27, 28], to estimate the effect size distribution, the statistical power and the sample size. Based on the results obtained from a pilot data, it is possible to evaluate the predicted statistical powers for sample sizes ranging from 3 to 1000 samples. Perhaps more interesting are two other methods, the so-called data-driven sample size determination (DSD) [29] and MetSizeR [30], that have been specifically

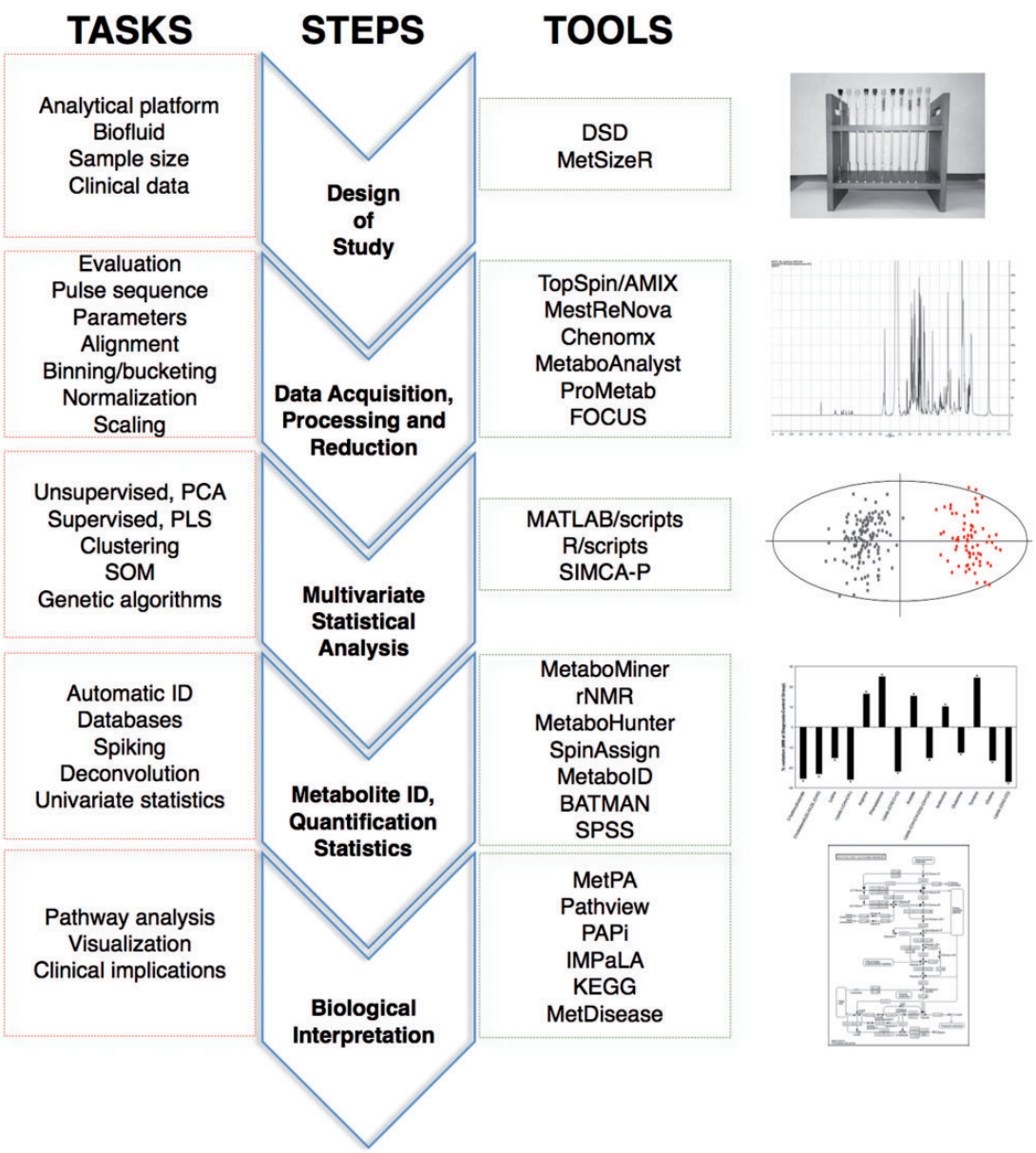


Figure 1. Typical workflow of a clinical metabolomics study carried out by NMR including the necessary steps, tasks and available bioinformatics tools.

designed for estimating sample sizes in metabolic phenotyping studies. Both methods focus on controlling the false discovery rate and rely on slightly different statistical approaches that can be easily modified to address specific analyses. MetSizeR has the ability to determine sample size without requiring previous pilot data and without assuming variable independence. However, it requires a preprocessing step of the raw data, as it can only deal with a limited number of variables. In contrast, DSD can deal with complete raw data but requires the use of pilot data for estimating the sample size.

Data acquisition, processing and reduction

From an experimental point of view, it is important to select appropriate pulse sequences in NMR metabolomics studies for each kind of sample according to its specific nature. Most biological samples are aqueous solutions and therefore require a water suppression element in the pulse sequence, presaturation [31] being the optimal choice in the case of quantitative analysis. Furthermore, an additional NOESY mixing block can also be included in the NMR experiments to improve baseline and water suppression [32]. Samples containing high-molecular weight species are better measured using NMR experiments including relaxation filter to better visualize the signals of metabolites [33]. NMR acquisition requires the adjustment of different parameters (e.g. number of scans, relaxation delay and spectral width) that have a strong impact on the signal to noise (S/N) ratio and the quality of the final NMR spectra. At this stage, software packages are limited to those supplied by the manufactures of NMR instruments [TopSpin (Bruker, Billerica, MA, USA), VnmrJ (Agilent Technologies, Santa Clara, CA, USA) and Delta (Jeol, Peabody, MA, USA)]. This is in contrast, as it will be described herein, with the different software options available for other steps associated with the handling of NMR data. Following the acquisition of the raw data, a

typical NMR metabolomics study encompasses the following steps:

Step 1: Processing

NMR spectra are usually subjected to several transformations before the statistical analysis can be performed. Thus, raw NMR data are Fourier transformed using adequate apodization functions, phased and baseline corrected [19]. Many different NMR software packages can perform these operations (e.g. NMRPipe [34], Sparky [35], MestReNova [36], MatNMR [37], NUTS [38], GIfa [39] and ACD Labs [40]). MestReNova and NMRPipe, in particular, are convenient tools for the analysis of NMR metabolomics data, as they can be easily adapted to the batch processing of large data sets of spectra. The scripting capabilities of these two software packages facilitate the automatic full processing (phase and baseline corrections, solvent suppression, etc.) of 1D and 2D NMR metabolomics spectra. Although both tools provide comparable possibilities for the initial transformations of the spectra, MestReNova has the advantage of providing a number of specific features for the downstream analysis of metabolomics data (e.g. binning/bucketing, peak deconvolution and signal quantification).

In this context, it is worth considering, when selecting a software package for NMR metabolomics studies, the possibility of reducing the number of tools involved in the analysis of the NMR metabolomics data. Thus, HiRes [41], Metabonomic [42], Automics [43], MVPACK [44] or Spectrum Miner [45] represent valuable alternatives for performing NMR data processing and exploratory statistical analyses. Among them, MVAPACK provides a free and robust alternative for both handling of NMR metabolomics data and chemometrics studies. The modular, open-source design of MVAPACK accepts new functionality, thus providing the possibility to integrate new chemometrics techniques.

Another alternative includes the selection of software packages integrating tools for spectra processing and compound identification and quantification, such as Chenomx [46], FOCUS [47], KnowItAll (BioRad, Philadelphia, PA, USA) or the freely accessible web-based option Bayesil [48]. In particular, Chenomx, a commercially available software package, provides, in addition to the capabilities for processing NMR spectra, a semi-automated tool for spectral deconvolution, which allows interactive fitting of metabolite peaks to reference metabolite spectra, and for quantifying their concentrations, significantly facilitating this time-consuming task. One of the major advantages of this NMR Suite is that, in addition to the automatic fitting, it also allows interactive adjustment of metabolite concentrations and peak frequencies, making it more flexible for experienced users.

Step 2: Data alignment and reduction

Depending on the physicochemical properties of the sample (i.e. pH, protein content, ionic strength, etc.), it is sometimes required to perform a spectral alignment to compensate for small variations in the peak positions between the different spectra. This process is usually applied over the whole spectrum and performed by referencing all spectra to an external compound or an internal metabolite. However, a global alignment of the spectra is not always possible for all the samples. This is particularly relevant when working with urine samples that are characterized by large pH differences, and requires the application of specific algorithms that facilitate the local alignment of specific spectral regions, a procedure that is more sensitive to the nonlinear nature of complex misalignments. Among the different software options, MetaboLab [49], Automics [43], Icoshift [50] and speaq [51] represent good examples of software tools for the alignment of peaks using a reference spectrum (Figure 2). An alternative alignment method, implemented in FOCUS [47], is RUNAS (Recursive Unreferenced Alignment of Spectra), that relies on the calculation of a cross-correlation function between spectra for optimizing the alignment. This approach avoids potential analytical biases derived from the use of a reference spectrum that may not be representative of the spectral diversity present in the samples. Furthermore, the alignment procedure in RUNAS is based on a spectral transformation, the so-called intensity weight slope transform, that enhances peak shapes and reduces the alignment bias owing to the presence of multiple peaks in the same alignment window.

Once the NMR spectra have been transformed and aligned, data are usually arranged into a matrix in which each row corresponds to a particular sample and each column to the intensity of the signal at a particular chemical shift. Statistical analysis of the data requires the fragmentation of the full spectra into smaller segments, a process called binning or bucketing, that can be performed using different approaches. This process leads to a significant data reduction, which simplifies subsequent data analysis [52]. This data table can be easily generated by many of the previously described NMR software packages, such as MestReNova [36] or NUTS [38]. Bucket width is usually fixed to 0.04 ppm, resulting in the reduction of a typical NMR spectrum to an average of 250 buckets. The main drawback of the selection of a fixed size for the binning/bucketing is that often the fragmentation of the spectra leads to the separation of a specific peak into several buckets. On the contrary, variable bucketing, which usually requires the manual intervention of the user, is specifically designed to take into account the precise spectral region covered by individual peaks. Available software packages offer different solutions for this procedure. Thus, in AMIX (Bruker, Billerica, MA, USA), variable size bucketing is based on the generation of graphical patterns, whereas Automics [43] relies on a feature called intelligent adaptive binning, and MVAPACK [44] on an optimized binning algorithm. Binning/bucketing tends to exclude uninformative regions, such as the residual water signal, urea, solvent, contaminants and broad noise regions. Some software packages (e.g. MetaboLab [49] and MetaboAnalyst [26]) include specific filters that offer the possibility to automatically select and eliminate noisy regions.

Step 3: Data normalization

Normalization of the data matrix obtained after binning/bucketing of the NMR spectra is an important step to improve the performance of the subsequent statistical analysis. Two different normalization processes are usually performed, a row-wise and a column-wise normalization. Row-wise is required to remove undesired variations associated with the concentration of the samples [53]. The most common procedure is normalization to total area in which the intensity of each bucket is divided by the sum of all the values in the same row. However, depending on the characteristics of the biological matrix, other procedures can be more suitable. Thus, the normalization of NMR spectra obtained from biofluids containing metabolites that can vary by several orders of magnitude (e.g. urine) requires the application of specific normalization methods, such as those based on the

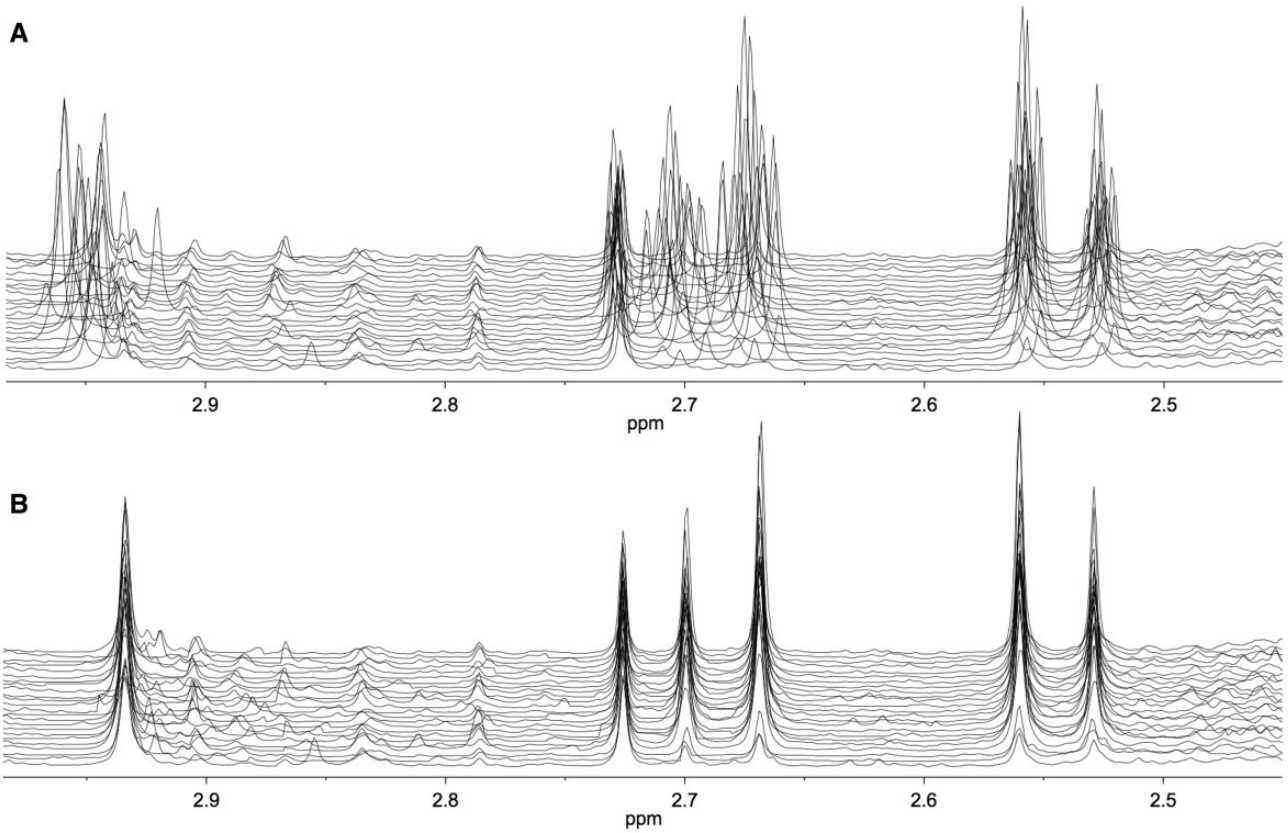


Figure 2. Alignment of NMR spectra is critical for the comparison of the data. (A) Collection of 1D NMR spectra corresponding to a set of urine samples; (B) same set of NMR spectra after the application of speaq [20]. The application of this bioinformatics tool translates into a better alignment of the spectra, thus overcoming the impact of chemical and physical variations on the chemical shifts of the metabolites present in those samples.

intensity of one or more peaks of known concentration. The so-called probabilistic quotient normalization method [54], where each bucket is divided by the most common scale factor between spectra, is also a popular normalization procedure for the analysis of biofluids containing highly abundant metabolites. Two other methods, quantile [55] and cubic-spline normalization [56], originally developed for DNA microarray analysis, have recently been shown [57] to outperform other normalization procedures in reducing unwanted biases and experimental variance. Both methods aim to obtain a similar distribution of feature intensities across spectra and provide reliable results reproducing fold changes. Cubic spline normalization, in particular, is able to correctly classify samples irrespectively of the data set size. Most frequent normalization procedures are available through different software packages and tools (AMIX, MetaboAnalyst [26], mQTL.NMR [58], Metabnorm [59], etc.).

A second normalization step includes a column-wise mean centering [60] and scaling [61] of the data matrix elements. There are different approaches to perform this task, although most metabolomics studies refer to unit variance scaling or pareto scaling, that can be performed with several NMR processing tools, including MetaboAnalyst [26], MetaboLab [49] or MVAPACK [44]. In unit variance scaling, each variable (i.e. column in the data matrix) is divided by its standard deviation, resulting in a variance of unity for the scaled variable. A known limitation of this approach is that it can amplify the effect of noisy variables. On the contrary, pareto scaling relies on the division of each variable by the square root of the standard deviation. In this case, the influence of small peaks is increased without amplifying uninformative variables [62]. A more sophisticated scaling method is the glog transformation implemented in ProMetab [63] that has been shown to estabilize the technical variance in NMR metabolomics data, but has the drawback of requiring the calculation of a specific parameter for each type of biological sample and set of NMR conditions, making it more time-consuming than other methods.

Finally, variability between NMR data can also be caused by technical reasons, one of the most important being batch-effects due to measurements performed at different times or by different people. To that end, the ComBat function, within the sva package in R [64], has been shown to be useful for removing such effects (Figure 3).

Multivariate statistical analysis

The overall outcome of the previous procedures is a data matrix consisting of rows (samples) and columns (bins/buckets) that is usually subjected to two different multivariate statistical analyses [65]. The first one, based on the application of unsupervised methods that do not take into account any information about the structure of the data, is used for dimension reduction and visualization of the data. This analysis also facilitates the assessment of the homogeneity of the data, with a focus on the evaluation of data quality, the identification of hidden biases in the study and the characterization of sample outliers (Figure 4A). In this context, an outlier is a sample that does not behave as the rest of samples within its group. Outliers are usually excluded from the study although there has to be a valid

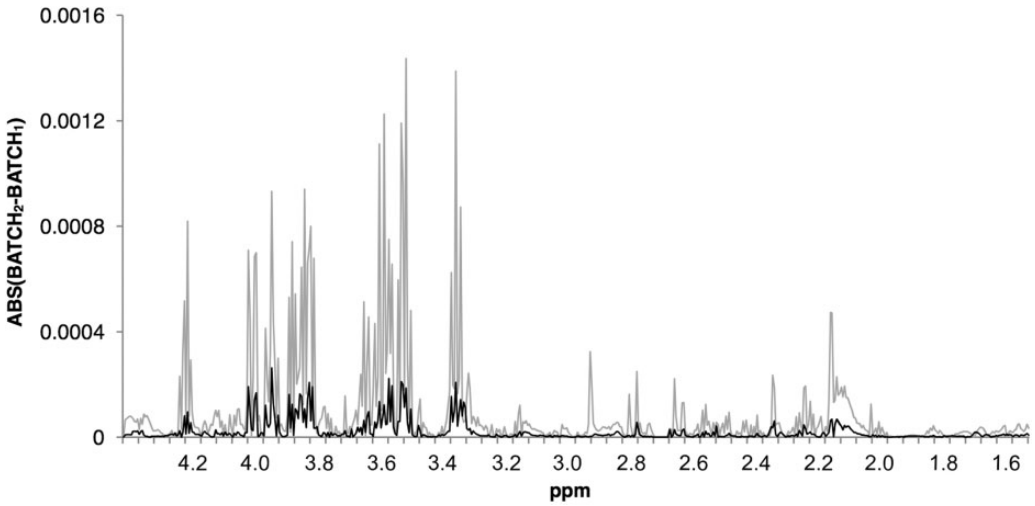


Figure 3. Variations between NMR data owing to batch effects can be minimized using appropriate software tools. Gray, absolute values of the mean differences at different chemical shifts between two batches of samples acquired at different times; black, same values after the application of the ComBat function [25]. It can be observed that the differences between batches are drastically reduced, thus facilitating the combination of data sets in clinical metabolomics studies.

analytical or biological reason to remove them. The second analysis, based on the so-called supervised methods, pursues the building of mathematical rules that use the metabolic profile for predicting a response variable, such as the presence of a disease or the response to a particular treatment [14] (Figure 4B).

Principal Component Analysis (PCA) and Partial Least Squares (PLS) regression are the most popular unsupervised and supervised methods, respectively, for the multivariate statistical analysis of metabolomics data [15]. These methods allow the user to project the NMR data (i.e. data matrix) into a reduced dimensional space for easier interpretation and visualization [66]. In PCA, the model describes the space corresponding to the highest variance of the data, while in PLS the space corresponds to that with the highest covariance between the NMR data and the response variable [67].

PLS models are seriously influenced by systematic variation in the data matrix that is not related to the response variable. It leads to pitfalls regarding the interpretation and selection of metabolite biomarkers [68]. To deal with this problem, orthogonal PLS (O-PLS) [69] were developed. O-PLS models are characterized by a first component correlated with the variable of interest and a second uncorrelated component, orthogonal to the first one. Furthermore, a limitation of PLS and other multivariate statistical approaches, applied in metabolomics studies relying on small sample sizes, is the possibility of overfitting, thus leading to apparent class separation when no real class differences exist [70]. An appropriate selection of variables can offer a solution to this problem, as it provides significant improvements in both model performance and in the interpretability of models by reducing complexity and the likelihood of model overfitting [71].

Other unsupervised statistical methods for metabolomics studies include hierarchical clustering algorithms (HCA) and self-organizing maps (SOM). These methods are particularly suitable for detecting nonlinear trends in the data that are not conveniently covered by PCA [72]. In particular, HCA has the ability to group profiles according to their similarity without any prior knowledge. An advantage of this method over PCA is that similarities in the multidimensional space can be visualized in a single plot. However, a drawback of this method is that relationships between elements are 1D, as similarity is a univariate parameter, and trees are hard to follow when they are crowded [15]. SOM is another unsupervised clustering and visualization method that is based on mapping vectors in a multidimensional space to a regular array of nodes in a low dimensionality space, a 2D map. An important advantage of this method is that it provides a good visualization tool and is less susceptible to select variables based on high variance, as it often happens in PCA and PLS, although it is not always easy to interpret the individual variables responsible for the classification [52]. Among the supervised methods, genetic algorithms (GA) are often used in metabolomics studies to refine the large number of metabolic variables down to a small subset of metabolites that are highly predictive of the clinical condition under study [73]. GA algorithms also tend to avoid the selection of variables with high variance and are particularly amenable for coupling with other supervised data analysis methods. However, a limitation of these models is that often there are many subsets of metabolites well suited for the classification and it is not always possible to select one with biological meaning [52].

The complexity of the NMR data and the potential biases associated with the multivariate statistical analysis demand the application of robust and versatile bioinformatics tools that can facilitate a rigorous analysis of the metabolomics studies. The most popular platforms for performing multivariate statistical analyses are R [74], MATLAB (The MathWorks, Natick, MA, USA) and SIMCA-P (Umetrics, Umea/Malmo, Sweden). Several software packages are also well suited for the multivariate statistical analysis of NMR metabolomics data. In particular, MetaboAnalyst [26] provides an extremely convenient, and freely accessible, web-based server for performing multivariate statistical analysis of NMR metabolomics data. The majority of the backend calculations in MetaboAnalyst are carried out by R functions and the data analysis functionality within the package includes an increasing number of statistical tests. Some other free bioinformatics tools implementing analysis techniques that are commonly used for the analysis of NMR metabolomics data are:

- **ChemoSpec** (R): PCA, HCA and Statistical Total Correlation Spectroscopy (STOCSY) [75] analysis, among others. STOCSY takes advantage of the multicollinearity of the intensity

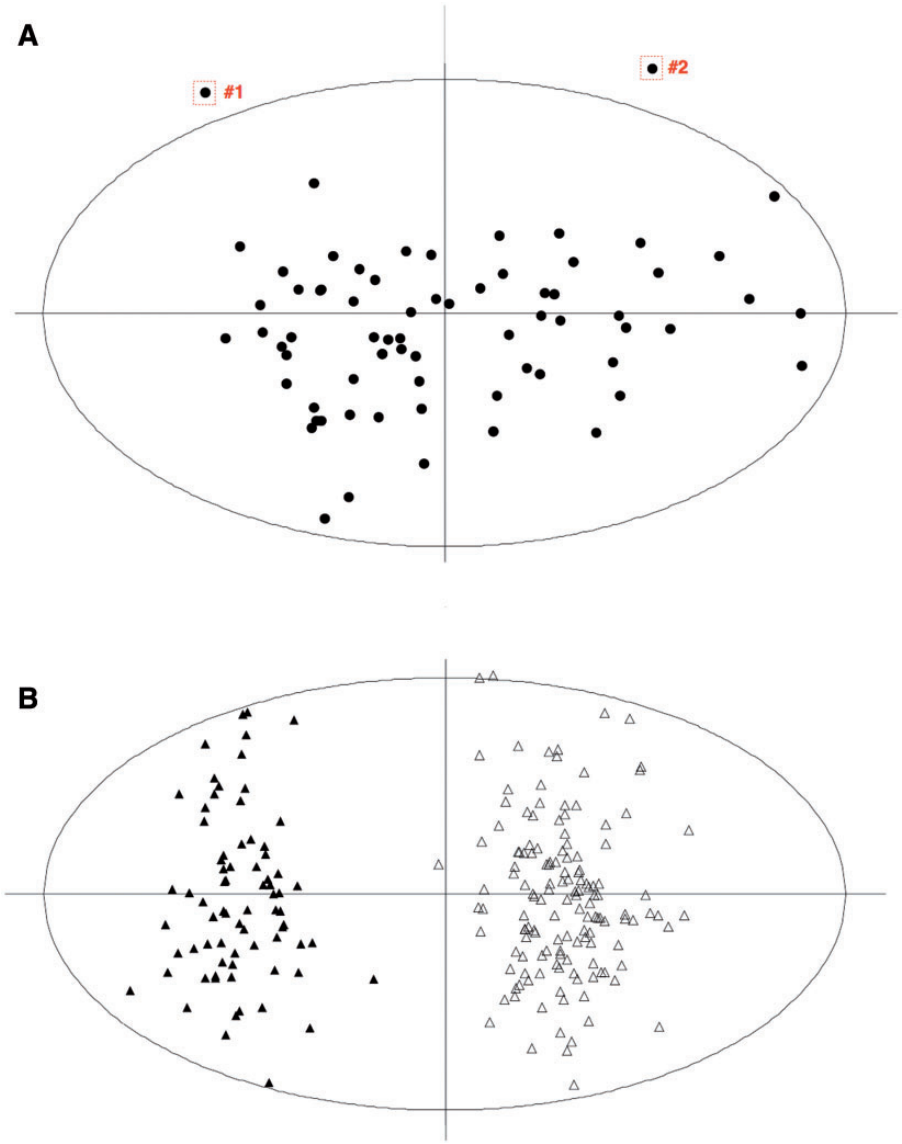


Figure 4. Multivariate statistical analysis of NMR metabolomics data relies on the application of unsupervised and supervised methods. (**A**) PCA analysis of the data matrix corresponding to a particular group of samples in a metabolomics study reveals, for example, the existence of outliers (#1, #2) that do not statistically behave as the rest of the group; (**B**) PLS analysis (e.g. healthy individuals versus diseased patients) can be used to detect the existence of underlying metabolic differences between groups of samples.

variables in a set of NMR spectra to generate a pseudo-2D NMR spectrum that facilitates the study of correlations between signals corresponding to different metabolites [76].

- **muma** (R): PCA, discriminant analysis versions of PLS (PLS-DA [77]) and O-PLS (OPLS-DA [78]) and the possibility of performing the selection of the best-separating principal components, the automatic testing of outliers and the automatic univariate analysis of data [79]. Furthermore, methods for preprocessing and analyzing NMR data, such as STOCSY and Ratio Analysis NMR Spectroscopy [80], a statistical approach for identifying resonances of the same molecule from a series of NMR spectra, can also be carried out using this tool.
- **MVAPACK** (Octave): In addition to the capabilities for NMR data loading, preprocessing and pretreatment, this tool offers the possibility to perform statistical analysis (PCA, PLS, OPLS-DA) of the data [44]. Further validation of the multivariate statistical models, a necessary step to estimate how well the classification models will perform when applied to new samples, is based on Monte Carlo *n*-fold internal cross-validation [81]. Overall, MVAPACK provides a powerful tool for the handling of NMR metabolomics data.
- **speaq** (R): Provides an integrated workflow for robust alignment and quantitative analysis of NMR metabolomics data. For each aligned NMR data point, the ratio of the between-group and within-group sum of squares (between-within (BW) ratio, Figure 5) is calculated to quantify the difference in variability between and within predefined groups of NMR spectra [51].

Although these tools contain different features for the analysis of metabolomics data, a combination of them is often used for the complete analysis of the clinical metabolomics data. In particular, AMIX, muma or MVAPACK, which contain features

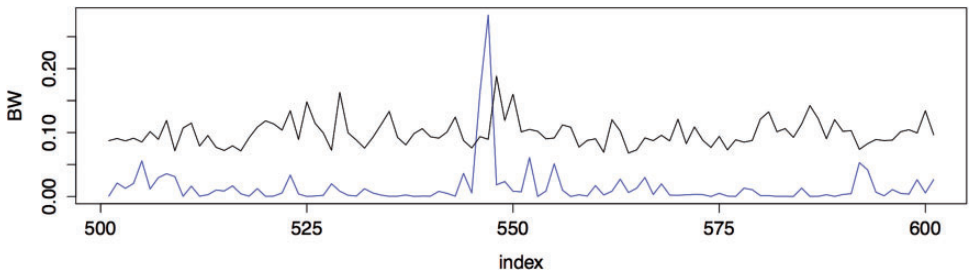


Figure 5. Variability between and within predefined groups of NMR spectra can be examined using specific bioinformatics tools, such as the BW ratio in speaq [20]. The blue line represents the BW statistic for two groups of samples at different indexes (i.e. bins/buckets). The upper line corresponds to the critical BW-value for rejecting the null hypothesis. Therefore, this tool can be used to identify specific regions (bottom line surpassing the upper line) exhibiting statistically significant differences between predefined groups of samples.

Table 1. Summary of NMR metabolomics databases

Database name	Availability	Comments	Access
Human Metabolome Database	Free	3000 metabolites, biological information for 40 000 metabolites	http://www.hmdb.ca/
Biological Magnetic Resonance Data Bank	Free	1000 compounds, no biological information	http://www.bmrb.wisc.edu/metabolomics/
Spectral Database for Organic Compounds	Free	15 000 compounds, not limited to metabolites	http://sdbs.db.aist.go.jp/
Madison-Qingdao Metabolomics Consortium Database	Free	794 compounds, literature information from 1156 compounds	http://www.mmcd.nmrfam.wisc.edu/
Platform for RIKEN Metabolomics	Free	80 compounds	http://prime.psc.riken.jp/
Birminghan Metabolite Library	Free	208 compounds at different pHs	http://www.bml-nmr.org/
BBIOREFCODE	AMIX software/license required	600 metabolites	https://www.bruker.com/
Chenomx	Chenomx software required	500 metabolites	http://www.chenomx.com/

for preprocessing NMR metabolomics data but limited options for statistical analysis, are usually combined with more sophisticated statistical packages (e.g. MetaboAnalyst and SIMCA-P). In this context, it would be ideal to develop NMR software packages that provide a full coverage, in an exhaustive fashion, of the different steps required for the analysis of NMR metabolomics data.

Metabolite identification, quantification and statistical evaluation

One of the most challenging steps of any NMR-based metabolomics study is the univocal assignment of the relevant regions identified in the statistical analysis to their corresponding metabolites. To that end, a common procedure involves the comparison of the NMR spectra acquired during the study with those of reference compounds available from literature or databases (Table 1). This assignment step can be accelerated with the use of specific programs that facilitate the automatic identification of metabolites, such as Chenomx [46], FOCUS [47], MetaboMiner [82], rNMR [83], MetaboHunter [84], SpinAssign [85], MetaboID [86]. Among them, Chenomx represents a powerful and user-friendly platform for the identification of a wide range of biologically and clinically relevant compounds. However, and despite the introduction of new and more powerful software packages, this process is still far from being completely automatic owing to the high degree of overlapping between signals and the contribution of different experimental factors (i.e. pH, salt, solvent effects). For this reason, it is sometimes advisable to confirm the postulated assignments using spiking experiments. They involve the addition of a specific metabolite to a previously measured biofluid sample and in the subsequent evaluation of the consistency between the newly appearing signals with the proposed assignment (Figure 6).

The overlapping problem and the typically low S/N ratio of the signals present in the NMR spectra make difficult the quantification of intensities and concentrations directly from the data matrices obtained from raw data. This issue can be alleviated using specific software tools (BATMAN [87], Newton [88], MetaboQuant [89], Chenomx, FOCUS, etc.) that apply different algorithms to facilitate an accurate integration of the signals. In particular, BATMAN and Newton are based on the application of different deconvolution approaches for the quantification of NMR spectra of complex mixtures. BATMAN, an R package, uses a Bayesian model of 1D NMR spectra to automatically deconvolve and quantify metabolites. The Bayesian model makes extensive use of prior information on the characteristic pattern of each metabolite, allowing it to recognize overlapped signals while taking account of shifting peaks. This is in contrast to other publically available tools, such as BQuant [90], that also models 1D NMR spectra by using a Bayesian approach. In this case, the program requires alignment and peak picking of the spectra before the application of the model, and the Bayesian

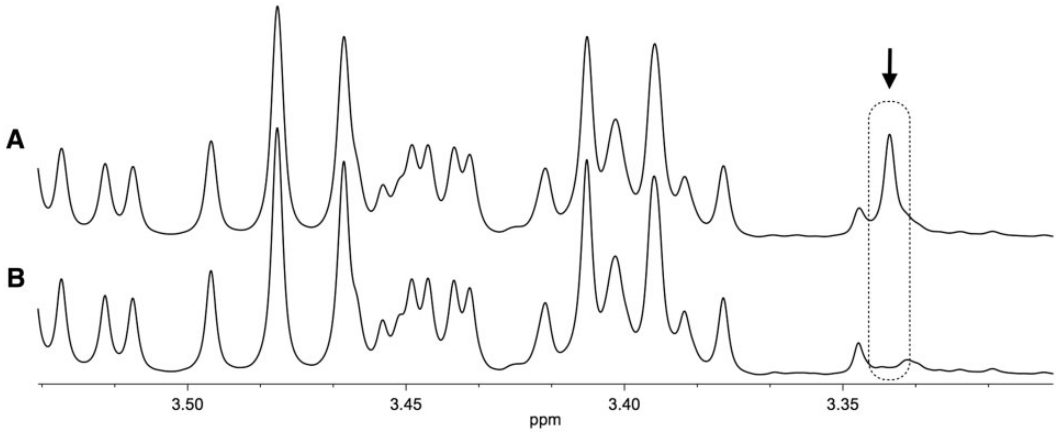


Figure 6. Spiking experiments are useful for validating metabolite assignments. The comparison of the NMR spectra obtained for a biological sample spiked-in with a known metabolite (**A**) with that obtained in the absence of the metabolite (**B**) can be used to confirm/discard specific metabolite assignments.

model itself does not account for peak shifts. On the other hand, Newton, a Java-based framework, is used for spectral analysis of multidimensional NMR data. The program implements the fast maximum likelihood reconstruction algorithm, a form of spectral deconvolution that constructs the simplest time-domain model whose Fourier processed spectrum most closely matches the spectrum of the identically processed free induction decay. Both methods have been shown to provide reliable results in the analysis of complex mixtures [87, 88]. MetaboQuant also calculates accurate compound concentration values from 1D and 2D NMR peak intensities using individual calibration factors and different outlier detection algorithms. However, it requires the application of other software packages for performing peak picking, fitting and integration on NMR spectra [89]. These procedures are generally time-consuming and are therefore often performed on those signals associated to statistically relevant regions. The proper evaluation of metabolite concentrations also requires the comparison of the intensity levels with those of a reference compound present in the mixture at a known concentration (e.g. Trymethylsilyl propanoic acid (TSP) and Tetramethylsilane (TMS)). A potential drawback of this approach is the possibility of nonspecific interactions between the reference compounds and components of the biological matrix, thus leading to inaccurate concentration values. Other options include the possibility of adding an artificial signal that has been previously calibrated with a reference sample (Eretic [91]).

The final step of this process involves the univariate statistical analysis of the potential differences found in the metabolite levels between the different groups of samples included in the study. This analysis can be easily performed using standard statistical software packages [SPSS (IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.), R [74], MATLAB, etc.]. Furthermore, results are usually validated using different mathematical approaches. A commonly used approach to achieve this task in clinical metabolomics studies is an analysis based on the ROC. ROC curve estimation is a nonparametric procedure consisting of the comparison of specificity (true-negative rate) against sensitivity (true-positive rate) according to specific decision boundaries. This evaluation can be performed using different tools, including the R packages ROCR [92] and pROC [93] or ROCCET [94], a web-based tool that offers the possibility of performing univariate and multivariate ROC analysis for all the relevant metabolites identified in the study. This option is particularly relevant in metabolomics studies focused on the identification of clinical biomarkers, as it allows the evaluation of the predictive ability of different combinations of metabolites.

Biological interpretation

A major outcome of any metabolomics study is a list of metabolites that have changed significantly under a specific condition (i.e. healthy individuals versus diseased patients). This information can be used to identify and interrogate the underlying biochemical pathways through different analyses [95]. However, the nature of cell metabolism, where the same metabolite can be involved in many different pathways, makes the pathway activity analysis a difficult task [96]. Thus, although the information obtained from metabolomics studies can potentially lead to important insights into the pathophysiological mechanisms of many diseases, the biological interpretation of metabolomics data remains a major bottleneck in these studies.

Over the past few years, different pathway mapping and visualization tools have been proposed for the analysis of metabolomics data, such as the freely available Pathview [97], MetScape [98] or MetaboAnalyst [26], and the commercial Ingenuity Pathway Analysis (IPA [99]) or MetaCore [100]. Most of these bioinformatics tools extract metabolic information from databases like KEGG [101] or ConsensusPathDB [102] and provide capabilities to facilitate the identification of specific metabolic pathways altered in the clinical situation under study.

Among the different public tools, MetScape offers some unique features, including the ability to generate network graphs using both user-defined sets of input nodes and a set of canonical metabolic pathways, as well as the possibility of linking isolated network graphs [98]. MetaboAnalyst, a software package that has been extensively discussed in this review, can also perform functional pathway analysis following two different approaches, the so-called Metabolite Set Enrichment Analysis [103] and the metabolic pathway analysis (MetPA [104]). Required inputs for these analyses are metabolite concentration data or a list of metabolite names that are mapped to specific IDs using the Human Metabolome Database (HMDB [105]). Interestingly, the current version of MetaboAnalyst takes advantage of a recently updated version of HMDB, thus considerably extending the possibilities of these analyses. On the other hand, there are a number of commercial bioinformatics

solutions (e.g. IPA [99] and MetaCore [100]) that provide useful features for performing molecular pathway and network analyses. Most of them integrate information from different omics approaches (genomics, transcriptomics, proteomics) and focus on providing solutions that facilitate a deeper insight into the role of metabolites in the molecular mechanisms of pathophysiological processes.

Although significant advances in the functional analysis of metabolomics data have been achieved in recent years, biological interpretation of metabolomics experiments is still hindered by relatively low coverage of experimentally identified metabolites in pathway databases [106]. Most metabolite reconstructions cover primary metabolites, but not much information is available about metabolites from different organisms (e.g. microbiome) or drug metabolites. In this context, an interesting tool termed MetDisease was recently proposed [107] to explore the possibility of expanding metabolite annotation through biomedical literature. It links PubChem compounds to MeSH terms via substances that are annotated to PubMed articles [108]. MetDisease does not predict novel associations but can be useful for identifying disease associations that would otherwise be difficult to find. Furthermore, this tool can be used to annotate a broader range of compounds, including drugs, nutritional compounds and environmental toxins.

Key Points

- This review summarizes the different steps involved in the metabolomic characterization of clinical samples by NMR.
- A number of useful considerations, bioinformatics tools and databases are proposed for the design of metabolomics studies and the analysis of the generated data.
- Clinical applications of metabolomics include the diagnosis/prognosis of diseases, the evaluation of treatment response and the monitoring of patients.

Acknowledgments

We thank the Spanish Ministry of Economy and Competitiveness (SAF2014-53977-R) and the Centro de Investigación Príncipe Felipe for financial support.

Funding

Spanish Ministry of Economy and Competitiveness (SAF2014-53977-R).

References

1. Zhang A, Sun H, Wang X. Serum metabolomics as a novel diagnostic approach for disease: a systematic review. *Anal Bioanal Chem* 2012;**404**:1239–45.
2. Wang X, Zhang A, Han Y, *et al*. Urine metabolomics analysis for biomarker discovery and detection of jaundice syndrome in patients with liver disease. *Mol Cell Proteomics* 2012;**11**:370–80.
3. Gowda GA, Zhang S, Gu H, *et al*. Metabolomics-based methods for early disease diagnostics. *Expert Rev Mol Diagn* 2008;**8**:617–33.
4. Mamas M, Dunn WB, Neyses L, *et al*. The role of metabolites and metabolomics in clinically applicable biomarkers of disease. *Arch Toxicol* 2011;**85**:5–17.
5. MacIntyre DA, Jimenez B, Lewintre EJ, *et al*. Serum metabolome analysis by ^{1}H-NMR reveals differences between chronic lymphocytic leukaemia molecular subgroups. *Leukemia* 2010;**24**:788–97.
6. Puchades-Carrasco L, Lecumberri R, Martinez-Lopez J, *et al*. Multiple myeloma patients have a specific serum metabolomic profile that changes after achieving complete remission. *Clin Cancer Res* 2013;**19**:4770–9.
7. Clayton TA, Baker D, Lindon JC, *et al*. Pharmacometabonomic identification of a significant host-microbiome metabolic interaction affecting human drug metabolism. *Proc Natl Acad Sci USA* 2009;**106**:14728–33.
8. Nicholson JK, Wilson ID, Lindon JC. Pharmacometabonomics as an effector for personalized medicine. *Pharmacogenomics* 2011;**12**:103–11.
9. Beisken S, Eiden M, Salek RM. Getting the right answers: understanding metabolomics challenges. *Expert Rev Mol Diagn* 2015;**15**:97–109.
10. Billoir E, Navratil V, Blaise JB. Sample size calculation in metabolic phenotyping studies. *Brief Bioinform* 2015.
11. Manach C, Hubert J, Llorach R, *et al*. The complex links between dietary phytochemicals and human health deciphered by metabolomics. *Mol Nutr Food Res* 2009;**53**(10):1303–15.
12. Davis WV, Bathe OF, Schiller DE, *et al*. Metabolomics and surgical oncology: potential role for small molecule biomarkers. *J Surg Oncol* 2011;**103**(5):451–9.
13. Jeong K, Kim S, Bandeira N. False discovery rates in spectral identification. *BMC Bioinformatics* 2012;**13**(Suppl 16):S2.
14. Ebbels TM, Cavill R. Bioinformatics methods in NMR-based metabolic profiling. *Prog Nucl Magn Reson Spectrosc* 2009;**55**:361–74.
15. Ebbels TM, Lindon JC, Coen M. Processing and modeling of nuclear magnetic resonance (NMR) metabolic profiles. *Methods Mol Biol* 2011;**708**:365–88.
16. Baumgartner C, Osl M, Netzer M, *et al*. Bioinformatic-driven search for metabolic biomarkers in disease. *J Clin Bioinforma* 2011;**1**:2.
17. Baumgartner C, Lewis GD, Netzer M, *et al*. A new datamining approach for profiling and categorizing kinetic patterns of metabolic biomarkers after myocardial injury. *Bioinformatics* 2010;**26**:1745–51.
18. Fassbender A, Vodolazkaia A, Saunders P, *et al*. Biomarkers of endometriosis. *Fertil Steril* 2013;**99**(4):1135–45.
19. Dougherty ER. Small sample issues for microarray-based classification. *Comp Funct Genomics* 2001;**2**:28–34.
20. Sung J, Wang Y, Chandrasekaran S, *et al*. Molecular signatures from omics data: from chaos to consensus. *Biotechnol J* 2012;**7**(8):946–57.
21. Sampson JN, Boca SM, Shu XO, *et al*. Metabolomics in epidemiology: sources of variability in metabolite measurements and implications. *Cancer Epidemiol Biomarkers Prev* 2013;**22**:631–40.
22. Rifai N, Gillette MA, Carr SA. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat Biotechnol* 2006;**24**:971–83.
23. Eng J. Sample size estimation: a glimpse beyond simple formulas. *Radiology* 2004;**230**(3):606–12.
24. Imaizumi A, Nishikata N, Yoshida H, *et al*. Clinical implementation of metabolomics, Chapter 12. In: U Roessner (ed), *Metabolomics*. INTECH Open Access Publisher, 2012.
25. Arkin CF, Wachtel MS. How many patients are necessary to assess test performance? *JAMA* 1990;**263**(2):275–8.

26. Xia J, Sinelnikov IV, Han B, *et al*. MetaboAnalyst 3.0-making metabolomics more Meaningful. *Nucleic Acids Res* 2015;**43**(W1):W251–7.
27. van Iterson M, 't Hoen PA, Pedotti P, *et al*. Relative power and sample size analysis on gene expression profiling data. *BMC Genomics* 2009;**10**:439.
28. van Iterson M, van de Wiel MA, Boer JM, *et al*. General power and sample size calculations for high-dimensional genomic data. *Stat Appl Genet Mol Biol* 2013;**12**:449–67.
29. Blaise BJ. Data-driven sample size determination for metabolic phenotyping studies. *Anal Chem* 2013;**85**:8943–50.
30. Nyamundanda G, Gormley IC, Fan Y, *et al*. MetSizeR: selecting the optimal sample size for metabolomic studies using an analysis based approach. *BMC Bioinformatics* 2013;**14**:338.
31. Hoult D. Solvent peak saturation with single phase and quadrature Fourier transformation. *J Magn Reson* 1976;**21**:337.
32. McKay RT. How the 1D-NOESY suppresses solvent signal in metabolomics NMR spectroscopy: an examination on the pulse sequence components and evolution. *Concepts Magn Reson* 2011;**38A**:197–220.
33. Beckonert O, Keun HC, Ebbels TM, *et al*. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat Protoc* 2007;**2**:2692–703.
34. Delaglio F, Grzesiek S, Vuiter GW, *et al*. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* 1995;**6**(3):277–93.
35. Lee W, Tonelli M, Markley JL. NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics* 2015;**31**(8):1325–7.
36. Cobas JC, Sardina FJ. Nuclear magnetic resonance data processing. MestRe-C: a software package for desktop computers. *Concept Magn Reson* 2003;**19A**:80–96.
37. van Beek JD. matNMR: a flexible toolbox for processing, analyzing and visualizing magnetic resonance data in Matlab. *J Magn Reson* 2007;**187**(1):19–26.
38. NUTS. NMR Data Processing Software. Acorn NMR Inc, 2009. http://www.AcornNMR.com
39. Malliavin TE, Pons JL, Delsuc MA. An NMR assignment module implemented in the Gifa NMR processing program. *Bioinformatics* 1998;**14**(7):624–31.
40. ACD/Labs. http://www.acdlabs.com (1 April 2015, date last accessed).
41. Zhao Q, Stoyanova R, Du S, *et al*. HiRes-a tool for comprehensive assessment and interpretation of metabolomic data. *Bioinformatics* 2006;**22**(20):2562–4.
42. Izquierdo-García JL, Rodríguez I, Kyriazis A, *et al*. A novel R-package graphic user interface for the analysis of metabonomic profiles. *BMC Bioinform* 2009;**10**:363.
43. Wang T, Shao K, Chu Q, *et al*. Automics: an integrated platform for NMR-based metabonomics spectral processing and data analysis. *BMC Bioinformatics* 2009;**10**:83.
44. Worley B, Powers R. MVAPACK: a complete data handling package for NMR metabolomics. *ACS Chem Biol* 2014;**9**(5):1138–44.
45. Spectrum Miner, One Moon Scientific, Inc. 2013. http://www.onemoonscientific.com/datachord-spectrum-miner.
46. Weljie AM, Newton J, Mercier P, *et al*. Targeted profiling: quantitative analysis of ^{1}H NMR metabolomics data. *Anal Chem* 2006;**78**:4430–42.
47. Alonso A, Rodriguez MA, Vinaixa M, *et al*. Focus: a robust workflow for one-dimensional NMR spectral analysis. *Anal Chem* 2014;**86**:1160–9.
48. Ravanbakhsh S, Liu P, Mandal R, *et al*. Accurate, fully-automated NMR spectral profiling for metabolomics. *arXiv* 2014;1409–56.
49. Ludwig C, Günther UL. MetaboLab-advanced NMR data processing and analysis for metabolomics. *BMC Bioinformatics* 2011;**12**:366.
50. Savorani F, Tomasi G, Engelsen SB. Icoshift: a versatile tool for the rapid alignment of 1D NMR spectra. *J Magn Reson* 2010;**202**:190–202.
51. Vu TN, Valkenborg D, Smets K, *et al*. An integrated workflow for robust alignment and simplified quantitative analysis of NMR spectrometry data. *BMC Bioinformatics* 2011;**12**:405.
52. Blekherman G, Laubenbacher R, Cortes DF, *et al*. Bioinformatics tools for cancer metabolomics. *Metabolomics* 2011;**7**:329–43.
53. Torgrip RJO, Aberg KM, Alm E, *et al*. A note on normalization of biofluid 1D ^{1}H NMR data. *Metabolomics* 2008;**4**:114–12.
54. Dieterle F, Ross A, Schlotterbeck G, *et al*. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in ^{1}H NMR metabonomics. *Anal Chem* 2006;**78**:4281–90.
55. Bolstad BM, Irizarry R A, Astrand M, *et al*. A comparison of normalization methods for high densityoligonucleotide array data based on variance and bias. *Bioinformatics* 2003;**19**:185–93.
56. Workman C, Jensen LJ, Jarmer H, *et al*. A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol* 2002;**3**(9):research0048.
57. Kohl SM, Klein MS, Hochrein J, *et al*. State-of-the art data normalization methods improve NMR-based metabolomic analysis. *Metabolomics* 2012;**8**(Suppl 1):146–60.
58. Hedjazi L, Gauguier D, Zalloua P, *et al*. mQTL.NMR: an integrated suite for genetic mapping of quantitative variations of ^{1}H NMR-based metabolic profiles. *Anal Chem* 2015;**87**(8):4377–84.
59. Jauhiainen A, Madhu B, Narita M, *et al*. Normalization of metabolomics data with applications to correlation maps. *Bioinformatics* 2014;**30**(15):2155–61.
60. Smolinska A, Blanchet L, Buydens LM, *et al*. NMR and pattern recognition methods in metabolomics: from data acquisition to biomarker discovery: a review. *Anal Chim Acta* 2012;**750**:82–97.
61. Van den Berg RA, Hoefsloot HC, Westerhuis JA, *et al*. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* 2006;**7**:142.
62. Craig A, Cloarec O, Holmes E, *et al*. Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Anal Chem* 2006;**78**(7):2262–7.
63. Parsons HM, Ludwig C, Günter UL, *et al*. Improved classification accuracy in 1- and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation. *BMC Bioinformatics* 2007;**8**:234.
64. Leek JT, Johnson WE, Parker HS, *et al*. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 2012;**28**:882–3.
65. Kell DB, Oliver SG. Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays* 2004;**26**:99–105.
66. Brereton RG. *Chemometrics for Pattern Recognition*. Chichester: John Wiley & Sons, 2009.

67. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemometr Intell Lab Syst* 2001;**58**:109–30.
68. Lindon JC, Nicholson JK, Holmes E. *The Handbook of Metabonomics and Metabolomics*. Elsevier/Oxford, 2011.
69. Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). *J Chemom* 2002;**16**;119–28. doi:10.1002/cem.695
70. Westerhuis JA, Hoefsloot HCJ, Smit S, *et al*. Assessment of PLSDA cross validation. *Metabolomics* 2008a;**4**:81–9.
71. Quintas G, Portillo N, Garcia-Cañaveras JC, *et al*. Chemometric approaches to improve PLSDA model outcomefor predicting human non-alcoholic fatty liver disease using .UPLC-MS as a metabolic profiling tool. *Metabolomics* 2012;**8**:86–98.
72. Arnald A, Marsal S, Julià A. Analytical methods in untargeted metabolomics: state of the art in 2015. *Bioeng Biotechnol* 2015;**3**:23.
73. Mitchell M. *An Introduction to Genetic Algorithms*. Cambridge, MA: MIT Press, 1996.
74. The Comprehensive R Archive Network. http://cran.r-project.org/web/packages (1 April 2015, date last accessed).
75. Cloarec O, Dumas ME, Craig A, *et al*. Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic ^{1}H NMR data sets. *Anal Chem* 2005;**77**(5):1289–9.
76. Blaise BJ, Navratil V, Emsley L, *et al*. Orthogonal filtered recoupled-STOCSY to extract metabolic networks associated with minor perturbations from NMR spectroscopy. *J Proteome Res* 2011;**10**:4342–8.
77. Kemsley EK. Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods. *Chemometr Intell Lab Syst* 1996;**33**, 47–61.
78. Bylesjö M, Rantalainen M, Cloarec O, *et al*. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *J Chemom* 2006;**20**, 341–51.
79. Gaude E, Chignola F, Spiliotopoulos D, *et al*. muma, An R package for metabolomics univariate and multivariate statistical analysis. *Curr Metabolomics* 2013;**1**:180–9.
80. Wei S, Zhang J, Liu L, *et al*. Ratio analysis nuclear magnetic resonance spectroscopy for selective metabolite identification in complex samples. *Anal Chem* 2011;**83**(20):7616–23.
81. Xu QS, Liang YZ, Du YP. Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. *J Chemom* 2004;**18**:112–20.
82. Xia J, Bjorndahl TC, Tang P, *et al*. MetaboMiner-semi-automated identification of metabolites from 2D NMR spectra of complex biofluids. *BMC Bioinformatics* 2008;**9**:507.
83. Lewis IA, Schommer SC, Markley JL. rNMR: open source software for identifying and quantifying metabolites in NMR spectra. *Magn Reson Chem* 2009;**47**(Suppl 1):S123–6.
84. Tulpan D, Leger S, Belliveau L, *et al*. MetaboHunter: an automatic approach for identification of metabolites from ^{1}H-NMR spectra of complex mixtures. *BMC Bioinformatics* 2011;**12**:400.
85. Chikayama E, Sekiyama Y, Okamoto M, *et al*. Statistical indices for simultaneous large-scale metabolite detections for a single NMR spectrum. *Anal Chem* 2010;**82**:1653–8.
86. MacKinnon N, Somashekar BS, Tripathi P, *et al*. MetaboID: a graphical user interface package for assignment of ^{1}H NMR spectra of bodyfluids and tissues. *J Magn Reson* 2013;**226**:93–9.
87. Hao J, Liebeke M, Astle W, *et al*. Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN. *Nat Protoc* 2014;**9**:1416–27.
88. Chylla RA, Hu K, Ellinger JJ, *et al*. Deconvolution of two-dimensional NMR spectra by fast maximum likelihood reconstruction: application to quantitative metabolomics. *Anal Chem* 2011;**83**(1):4871–80.
89. Klein MS, Oefner PJ, Gronwald W. MetaboQuant: a tool combining individual peak calibration and outlier detection for accurate metabolite quantification in 1D ^{1}H and ^{1}H-^{13}C HSQC NMR spectra. *Biotechniques* 2013;**54**:251–6.
90. Zheng C, Zhang S, Ragg S, *et al*. Identification and quantification of metabolites in (1)H NMR spectra by Bayesian model selection. *Bioinformatics* 2011;**27**:1637–44.
91. Nuzzo G, Gallo C, d'Ippolito G, *et al*. Composition and quantitation of microalgal lipids by ERETIC ^{1}H NMR method. *Mar Drugs* 2013;**11**:3742–53.
92. Sing T, Sander O, Beerenwinkel N, *et al*. ROCR: visualizing classifier performance in R. *Bioinformatics* 2005;**21**, 3940–3941.
93. Robin X, Turck N, Hainard A, *et al*. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011;**12**:77.
94. Xia J, Broadhurst DI, Wilson M, *et al*. Translational biomarker discovery in clinical metabolomics: an introductory tutorial. *Metabolomics* 2013;**9**:280–99.
95. Tzoulaki I, Ebbels TM, Valdes A, *et al*. Design and analysis of metabolomics studies in epidemiologic research: a primer on -omic technologies. *Am J Empidemiol* 2014;**180**(2):129–39.
96. Aggio RB, Ruggiero K, Villas-Boas SG. Pathway activity profiling (PAPi): from the metabolite profile to the metabolic pathway activity. *Bioinformatics* 2010;**26**:2969–76.
97. Luo W, Brouwer C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* 2013;**29**:1830–1.
98. Karnovsky A, Weymouth T, Hull T, *et al*. Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. *Bioinformatics* 2012;**28**(3):373–80.
99. Ingenuity Pathway Analysis (IPA). QIAGEN, Silicon Valley. http://www.ingenuity.com/products/ipa.
100. MetaCore. GeneGo, Inc. https://portal.genego.com.
101. Kanehisa M. The KEGG database. *Novartis Found Symp* 2002;**247**:91–101; discussion 101–3, 119–28, 244–52.
102. Kamburov A, Stelzl U, Lehrach H, *et al*. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res* 2013;**41**:D793–800.
103. Xia J, Wishart DS. MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res* 2010;**38**:W71–7.
104. Xia J, Wishart DS. MetPA: a web-based metabolomics tool for pathway analysis and visualization. *Bioinformatics* 2010;**26**:2342–4.
105. Wishart DS, Jewison T, Guo AC, *et al*. HMDB 3.0-The Human Metabolome Database in 2013. *Nucleic Acids Res* 2013;**41**:D801–7.
106. Barupal DK, Haldiya PK, Wohlgemuth G, *et al*. MetaMapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity. *BMC Bioinformatics* 2012;**13**:99.
107. Duren W, Weymouth T, Hull T, *et al*. MetDisease–connecting metabolites to diseases via literature. *Bioinformatics* 2014;**30**:2239–41.
108. Sartor MA, Ade A, Wright Z, *et al*. Metab2MeSH: annotating compounds with medical subject headings. *Bioinformatics* 2012;**28**:1408–10.