

Development of new bioinformatic tools to improve mass spectrometry-based analysis of the lipidome

PhD Thesis

by

María Isabel Alcoriza Balaguer

May, 2023



VNIVERSITAT
DE VALÈNCIA

Doctorado en Biomedicina y Biotecnología
Departamento de Bioquímica y Biología Molecular
Facultad de Ciencias Biológicas

Supervisor:

Dr. Agustín Lahoz Rodríguez



 Biomarkers and
Precision Medicine Unit

D. Agustín Lahoz Rodríguez, Director de la Unidad de Biomarcadores y Medicina de Precisión del Instituto de Investigación Sanitaria de La Fe,

CERTIFICA:

Que la presente memoria, titulada “Development of new bioinformatic tools to improve mass spectrometry-based analysis of the lipidome.”, corresponde al trabajo realizado bajo su dirección por **Dña. María Isabel Alcoriza Balaguer** para su presentación como Tesis Doctoral en el Programa de Doctorado en Biotecnología y Biomedicina de la Universidad de Valencia.

Y para que conste firma el presente certificado en Valencia, a 10 de mayo de 2023.

Fdo:

Agustín Lahoz Rodríguez

This PhD thesis has been carried out at the Biomarkers and Precision Medicine Unit located at the Health Research Institute - Hospital La Fe in Valencia (Spain). Maria Isabel Alcoriza has been supported by a PFIS contract from the Carlos III Health Institute of the Spanish Ministry of Economy and Competitiveness [FI18/0024].

Agradecimientos

En primer lugar, me gustaría agradecer a todas aquellas personas que han participado de forma directa en la realización de esta tesis.

A Agustín, gracias por confiar en mí, por introducirme en el mundo de la metabolómica y la lipidómica y por darme el espacio y el apoyo para darle forma a esta tesis. Gracias por las infinitas correcciones y por todos los consejos que han hecho posible que llegue hasta aquí.

A mis compañeros de la Unidad de Biomarcadores y Medicina de Precisión. En especial, a Juan Carlos. Gran parte del mérito de esta tesis es tuyo. Gracias por todo lo que me has enseñado, por tus consejos, por darme siempre otro punto de vista y, sobre todo, por enseñarme que, aunque la ciencia no siempre es agradecida, es preciosa cuando las cosas salen bien y todo cobra sentido. Hay que saber cuándo parar, apagar el ordenador, y volver a intentarlo otro día con la mente despejada. Gracias por haberme guiado y por seguir guiándome en este camino.

A las chicas de la Unidad Analítica, gracias por haberme acogido con tanto cariño y por enseñarme todo lo que sé del trabajo en el laboratorio y de espectrometría de masas. Y gracias por rescatarme en innumerables ocasiones para ir a tomar un café. Especialmente a Marta, por escucharme y por hacerme sentir parte del laboratorio. Gracias por toda la ayuda y por tus consejos, me han ayudado mucho más de lo que puedo agradecerme.

Y a Javi y Alfredo del área de informática, por vuestra paciencia y ayuda para poner a punto el servidor y las versiones web de las aplicaciones. No hubiera sido posible sin vosotros.

Por otro lado, también me gustaría agradecer a todas aquellas personas que, sin saberlo, me han ayudado de algún modo a llegar hasta aquí.

A mis amigos, por ser un refugio en el que desconectar de la ciencia. A los de siempre, Juanen, Majo Jorge (y el pequeño Leo), diría que a día de hoy aún no sabéis muy bien qué es lo que hago, pero no habría llegado hasta aquí sin vosotros. Gracias por todas las cervezas, cenas y conversaciones con las que desahogarnos. Siempre estáis ahí. Y Leo, gracias por enseñarme que “ixes estàn ballant i cantant”. A Pablo, por ser mi referente en cómo montártelo bien y por toda la envidia mala que me diste mientras vivíamos juntos, sin duda fue un motor importante para ponerme a tope con la tesis. A *La cúpula*, por hacer de la música un oasis en el que coger fuerzas. Y *Follapo*, aunque con vosotros es más difícil desconectar de la ciencia... Fue una suerte encontraros en la universidad y ver cómo hemos ido creciendo desde entonces.

A mi familia. Gracias por estar siempre ahí y por todo el amor y confianza que me habéis dado. Tengo mucha suerte de teneros. A mi tía Concha, gracias por transmitirme tu tranquilidad y tu paciencia, y por enseñarme que las galletas también se pueden mojar en el té. A José, mi primo favorito, gracias por ser único y por ser un ejemplo de cómo preocuparse por los demás. A mi tío Antonio, gracias por tener siempre un buen vino con el que celebrar cualquier ocasión. Hay que mantener vivas las buenas tradiciones. A mi tía Isa y a José Antonio, por ser un ejemplo del trabajo duro y por estar siempre ahí pese a la distancia.

Gracias por todo el esfuerzo que eso conlleva y por tener siempre las puertas abiertas para recibirnos. A mi tía Amparo, a ti no sé ni por dónde empezar. Gracias por haberme tratado siempre como a una hija, por enseñarme que la impaciencia va en el gen Alcoriza y por tener La Corrala siempre abierta para juntar a la familia. A mi hermano, gracias por estar ahí cuando te he necesitado, incluso cuando no ha sido fácil. Me has enseñado tres grandes lecciones en la vida: cómo hacer agujeros en la pared, cómo no hacer una tortilla de patata y cómo reciclar el papel de regalo. Y especialmente, gracias a mi madre, eres mi referente, la persona más fuerte que conozco. Gracias por todo lo que has hecho por mí, me has convertido en una persona fuerte e independiente, sin miedo a enfrentarme a lo que sea necesario. Nunca podré agradecértelo lo suficiente.

Y, por último, a Amparo y Andrea. Las chicas más importantes de mi vida. Amparo, gracias por haberme hecho crecer como persona y por tu apoyo incondicional. Todo es mejor cuando lo comparto contigo. Y Andrea, llevas poquito en este mundo, pero has sido mi gran motivación para acabar esta tesis, para cerrar este capítulo y seguir avanzando. Os quiero muchísimo.

En definitiva, muchas gracias a todos aquellos que de un modo u otro habéis contribuido a esta tesis y me habéis ayudado en esta etapa.

Table of Contents

| | |
|---|-----------|
| Glossary | 19 |
| Abstract | 23 |
| General Introduction | 27 |
| 1. Metabolomics | 29 |
| 1.1. Analytical techniques used in metabolomics | 30 |
| 1.2. Untargeted LC-MS metabolomics..... | 32 |
| 1.2.1. Experimental design..... | 34 |
| 1.2.2. Sample preparation... .. | 35 |
| 1.2.3. LC-MS analysis..... | 36 |
| 1.2.4. Data processing..... | 38 |
| 1.2.5. Data analysis..... | 43 |
| 2. The human metabolome..... | 45 |
| 3. Lipidomics..... | 49 |
| 3.1. Fatty acyls..... | 50 |
| 3.2. Glycerolipids..... | 52 |
| 3.3. Glycerophospholipids..... | 53 |
| 3.4. Sphingolipids..... | 56 |
| 3.5. Sterol lipids..... | 57 |
| 3.6. Lipids and disease..... | 58 |
| 4. Metabolomics and isotope tracing..... | 61 |
| Objectives | 63 |
| Chapter 1. LipidMS: an R-package and a web-based tool for untargeted LC-MS/MS data processing and lipid annotation | 67 |
| Introduction..... | 69 |
| 1. Lipid identification in LC-MS lipidomics..... | 71 |
| 1.1. Challenges in lipid annotation in LC-MS-based lipidomics..... | 71 |
| 2. Most used bioinformatic tools for LC-MS-based lipidomics..... | 75 |

| | |
|--|-----|
| Methodology..... | 77 |
| 1. Chemicals and reagents..... | 79 |
| 2. Sample preparation..... | 81 |
| 2.1. Preparation of standards..... | 81 |
| 2.2. Lipid extraction from human serum samples..... | 81 |
| 3. LC-MS analysis..... | 82 |
| 3.1. Instrumentation..... | 82 |
| 3.2. Chromatographic separation..... | 82 |
| 3.3. MS detection..... | 83 |
| 4. Data processing and analysis for untargeted LC-MS lipidomic analysis..... | 85 |
| Results and Discussion..... | 89 |
| 1. LipidMS overview..... | 91 |
| 2. Features and implementation..... | 93 |
| 2.1. Data processing..... | 93 |
| 2.1.1. Peak-picking..... | 93 |
| 2.1.2. Peak alignment..... | 95 |
| 2.1.3. Peak grouping..... | 97 |
| 2.1.4. Peak filling..... | 98 |
| 2.2. Lipid annotation..... | 98 |
| 2.2.1. Rationale behind LipidMS annotation..... | 98 |
| 2.2.2. Lipid coverage and building block database customization..... | 105 |
| 2.2.3. LipidMS annotation workflow..... | 106 |
| 2.2.4. Additional functions..... | 112 |
| 2.3. Implementation..... | 114 |
| 2.3.1. R package..... | 115 |
| 2.3.2. Web-based application..... | 116 |
| 3. LipidMS performance evaluation..... | 121 |
| 3.1. Comparison between LipidMS and XCMS data pre-processing.... | 122 |

| | |
|---|------------|
| 3.2. Comparison between LipidMS and MS-DIAL..... | 129 |
| 3.2.1. Data processing and annotation of known lipid standards..... | 129 |
| 3.2.2. Annotation of lipids in human serum pool sample..... | 130 |
| 4. Future improvements of LipidMS..... | 134 |
| Chapter 2. FAMetA: a mass isotopologue-based tool for the comprehensive analysis of fatty acid metabolism..... | 137 |
| Introduction..... | 139 |
| Methodology..... | 145 |
| 1. Chemicals and reagents..... | 147 |
| 2. Cell lines and growth conditions for cell metabolism studies.... | 149 |
| 2.1. Mouse naïve CD8+ T-cells..... | 149 |
| 2.2. A549 cell line..... | 150 |
| 3. Sample preparation..... | 151 |
| 3.1. Preparation of standards..... | 148 |
| 3.2. Saponification and extraction of total FA from cells..... | 151 |
| 4. LC-MS analysis..... | 152 |
| 4.1. Instrumentation..... | 152 |
| 4.2. Chromatographic separation..... | 152 |
| 4.3. MS detection..... | 152 |
| 5. Data processing and analysis for fatty acid analysis..... | 153 |
| Results and Discussion..... | 157 |
| 1. FAMetA overview..... | 159 |
| 2. Features and implementation..... | 164 |
| 2.1. FAMetA workflow..... | 164 |
| 2.2. Implementation of the quasi-multinomial distribution..... | 168 |
| 2.3. Estimation of DNL parameters..... | 170 |
| 2.4. Estimation of elongation parameters..... | 172 |
| 2.5. Estimation of desaturation..... | 174 |
| 2.6. Model assumptions..... | 176 |

| | |
|--|------------|
| 2.7. Data requirements for FA modelling..... | 177 |
| 2.8. Implementation..... | 177 |
| 2.8.1. R package..... | 178 |
| 2.8.2. Web-based tool..... | 178 |
| 3. FAMetA performance evaluation..... | 182 |
| 3.1. In silico validation..... | 182 |
| 3.2. Biological validation..... | 182 |
| 3.2.1. FAMetA enables the analysis of FA metabolism in vitro... | 183 |
| 3.2.2. FAMetA enables the analysis of the FA metabolism in vivo... | 187 |
| 3.3. Comparison between FAMetA and other available tools..... | 189 |
| 4. FAMetA enables the identification of unknown FA in biological samples | 194 |
| 5. Future improvements of FAMetA..... | 200 |
| Conclusions..... | 201 |
| Resumen en castellano..... | 205 |
| References..... | 221 |
| Appendix 1: List of Publications..... | 243 |
| Appendix 2: Additional figures and tables..... | 247 |

List of Figures and Tables

| | |
|--|-----|
| Figure 1. The “ <i>omics</i> cascade” | 29 |
| Figure 2. Systematic literature analysing the use of MS and NMR platforms in metabolomics..... | 31 |
| Figure 3. General workflow in untargeted LC-MS metabolomics | 33 |
| Figure 4. Schematic diagram of LC-MS in untargeted metabolomics... | 36 |
| Figure 5. Scheme of common acquisition modes in untargeted LC-MS metabolomics..... | 38 |
| Figure 6. Main steps in untargeted LC-MS data processing | 39 |
| Figure 7. Analysis of the human metabolome coverage provided by LC-MS | 46 |
| Figure 8. Representation of predicted lipids in main metabolomic databases. | 48 |
| Figure 9. Main lipid classes and key building blocks of the human lipidome. | 49 |
| Figure 10. Main FA biosynthetic reactions..... | 51 |
| Figure 11. Main biosynthetic pathways of TG..... | 53 |
| Figure 12. Main biosynthetic pathways of PL..... | 54 |
| Figure 13. Main biosynthetic pathways of SL..... | 57 |
| Figure 14. Main biosynthetic pathway of CE.. .. | 58 |
| Figure 15. LipidMS v3.0 overview | 92 |
| Figure 16. Clustering algorithm used for peak alignment and grouping in LipidMS | 96 |
| Figure 17. Scheme of sequential partitioning and clustering of peaks executed during alignment and grouping steps | 97 |
| Figure 18. Coelution profile of common fragment 184.074 (phosphocholine) of PC and SM in a LC-MS analysis..... | 100 |
| Figure 19. Example of building block structure of a PC. | 105 |
| Figure 20. Lipid annotation workflow in LipidMS.. .. | 107 |
| Figure 21. Examples of annotation results returned by LipidMS..... | 111 |

| | |
|---|-----|
| Figure 22. Example of graphical output of the <i>plotLipids</i> function for DIA acquired data..... | 112 |
| Figure 23. Example of graphical output of the <i>plotLipids</i> function for DDA acquired data..... | 113 |
| Figure 24. Examples of additional graphical outputs..... | 114 |
| Figure 25. Alternative processing pipelines in LipidMS..... | 116 |
| Figure 26. Data import tab of the LipidMS web tool..... | 118 |
| Figure 27. Peak-picking tab of the LipidMS web tool..... | 118 |
| Figure 28. Batch processing tab of the LipidMS web tool..... | 119 |
| Figure 29. Annotation tab of the LipidMS web tool..... | 119 |
| Figure 30. Run tab of the LipidMS web tool..... | 120 |
| Figure 31. Summary of lipid annotations provided by LipidMS and MS-DIAL for the human serum pool..... | 134 |
| Figure 32. Classical isotope labelling experiment for FA analysis..... | 141 |
| Figure 33. FA metabolism network..... | 160 |
| Figure 34. Example of the FAMetA calculations for FA(16:0) to FA(20:1)n9..... | 161 |
| Figure 35. FAMetA overview..... | 163 |
| Figure 36. Detailed workflow for data pre-processing..... | 165 |
| Figure 37. Detailed FAMetA workflow and output..... | 167 |
| Figure 38. Fitting experimental mass-isotopologue FA data to multinomial and quasi-multinomial distributions..... | 169 |
| Figure 39. Data pre-processing tab of the FAMetA web tool..... | 180 |
| Figure 40. Manual curation tab of the FAMetA web tool..... | 180 |
| Figure 41. Metabolic analysis tab of the FAMetA web tool..... | 181 |
| Figure 42. Biological validation of FAMetA in active mouse CD8 ⁺ T-cells incubated with different U- ¹³ C-tracers..... | 184 |
| Figure 43. Biological validation of FAMetA in active mouse CD8 ⁺ T-cells incubated with U- ¹³ C-glucose and different inhibitors of the FA metabolism..... | 185 |

| | |
|---|-----|
| Figure 43. Biological validation of FAMetA in active mouse CD8 ⁺ T-cells incubated with U- ¹³ C-glucose and different inhibitors of the FA metabolism..... | 185 |
| Figure 44. Biological validation of FAMetA in WT and KHK-C mice after drinking normal or 5% or 10% sucrose water..... | 189 |
| Figure 45. Analysis of the influence of the down-regulation of SCAP on the FA metabolism in the H1299 cells | 191 |
| Figure 46. Analysis of the FA diversity in the human NSCLC cell line A549 incubated for 72 h with U- ¹³ C-glucose induced by the use of different FA metabolism inhibitors (FASNi, SCDi and FADS2i) | 191 |
| Figure 47. Elucidation of the synthesis route of unidentified FA species by combining FAMetA and FA metabolism inhibitors. | 191 |
| Figure 48. The algorithm employed to identify unknown FA by the reconstruction of their biosynthesis route | 191 |
| Figure 49. Confirmation of the identity of 11 unknown FA in the A549 cells with chemical standards | 198 |
| Figure 50. FA biosynthesis routes in the NSCLC cell line A549..... | 199 |

| | |
|---|-----|
| Table 1. Common isobaric and isomeric overlaps in LC-MS lipidomics.. | 73 |
| Table 2. Preferred adducts and fragmentation rules set by defaults to annotate lipids in ESI+..... | 101 |
| Table 3. Preferred adducts and fragmentation rules set by defaults to annotate lipids in ESI-..... | 103 |
| Table 4. Building blocks included by default in the bbDB to build the qDB..... | 106 |
| Table 5. Summary of the expected lipid standard features detected and identified for each software package | 123 |
| Table 6. Lipid standards detected and identified in ESI-..... | 124 |
| Table 7. Lipid standards detected and identified in ESI+. | 127 |
| Table 8. Summary of the lipids identified in ESI-. | 133 |
| Table 9. Summary of lipids identified in ESI+.. | 133 |
| Table 10. Comparison of features implemented within the main available tools for the analysis of FA metabolism. | 190 |

Glossary

| | |
|----------------|----------------------------------|
| ACACA/B | Acetyl-CoA carboxylases A/B |
| ACLY | ATP citrate lyase |
| ACSS1/2 | Acetyl-CoA synthetases 1/2 |
| AGPAT | Acylglycerol acyltransferases |
| BA | Bile acid |
| bbDB | Building block database |
| BSA | Bovine serum albumin |
| Car | Acylcarnitine |
| CCT | CDP-Cho transferase |
| CDP | Cytidyl diphosphate |
| CE | Cholesteryl esters |
| Cer | Ceramide |
| CERK | Ceramide kinase |
| CerP | Ceramides phosphate |
| CERS | Ceramide synthase |
| CES1 | Carboxylesterase 1 |
| Cho | Choline |
| CK | Choline kinase |
| CL | Cardiolipin |
| CLS | Cardiolipin synthase |
| CMP | Cytidyl monophosphate |
| CPT | Choline phosphotransferase |
| CRC | Colorectal cancer |
| CTP | Cytidyl triphosphate |
| DDA | Data dependent acquisition |
| DEGS | Dihydroceramide desaturase |
| DG | Diacylglycerol |
| DGAT | Diacylglycerols acyltransferases |
| DGAT | Diacylglycerol acyltransferases |
| DHAP | Dihydroxyacetone phosphate |
| DIA | Data independent acquisition |
| DNL | <i>De novo</i> lipogenesis |
| ECT | CDP-Ethanolamine transferase |
| EDTA | Ethylenediaminetetraacetic acid |

| | |
|----------------|--|
| EIC | Extracted ion chromatogram |
| EK | Ethanolamine kinase |
| ELOVL | Elongation of very long chain fatty acids protein |
| EMT | Epithelial-to-mesenchymal transition |
| EPT | Ethanolamine phosphotransferase |
| ESI | Electrospray Ionization |
| Et | Ethanolamine |
| FA | Fatty Acids |
| FADS1/2 | FA Desaturases 1/2 |
| FAO | Fatty acid oxidation |
| FASA | Fatty acid source analysis |
| FASN | Fatty acid synthetase |
| FATP | Fatty acid transport proteins |
| FBS | Fetal bovine serum |
| FFA | Free fatty acids |
| G3P | Glycerol-3-phosphate |
| GC | Gas chromatography |
| GL | Glycerolipids |
| GNPS | Global Natural Product Social Molecular Networking |
| GPAT | Glycerol-3-phosphate acyltransferases |
| GPD1 | Glycerol-3-phosphate dehydrogenase 1 |
| HCA | Hierarchical clustering analysis |
| HDL | High density molecules |
| HMDB | Human Metabolome Database |
| HMG-CoA | Hydroxymethylglutaryl-CoA |
| HPLC | High performance liquid chromatography |
| HRMS | High resolution mass spectrometry |
| InChI | International chemical identifier |
| ISA | Isotopomer spectral analysis |
| KEGG | Kyoto Encyclopedia of Genes and Genomes |
| LC | Liquid chromatography |
| LCAT | Lecithin cholesterol acyltransferase |
| LDA | Lipid data analyzer |
| LDL | Low density molecules |
| LPA | Lysophosphatidic acid |

| | |
|-------------------|--|
| LPC | Lysophosphatidylcholines |
| LPIN | Lipins |
| LPL | Lysophospholipids |
| <i>m/z</i> | Mass-to-charge ratio |
| MG | Monoacylglycerol |
| MIDA | Mass isotopomer distribution analysis |
| MoNA | Massbank of north america |
| MS | Mass spectrometry |
| MS/MS | Tandem mass spectrometry |
| MSEA | Metabolite set enrichment analysis |
| MUFA | Monounsaturated fatty acids |
| NA | Not available |
| NMR | Nuclear magnetic resonance |
| NSCLC | Non-small cell lung cancer |
| O-PLS-DA | Orthogonal partial least squares discriminant analysis |
| PA | Phosphatidic acid |
| PBS | Phosphate-buffered Saline |
| PC | Phosphatidylcholine |
| PCA | Principal clustering analysis |
| PCho | Phosphocholine |
| PE | Phosphatidylethanolamine |
| PEMT | Phosphatidylethanolamine N-methyltransferase |
| PEt | Phosphoethanolamine |
| PFCS | Parent-fragment coelution score |
| PG | Phosphatidylglycerol |
| PGP | Phosphatidylglycerol phosphate |
| PGPS | Phosphatidylglycerol phosphate synthase |
| PI | Phosphatidylinositol |
| PIP | Phosphatidylinositol phosphate |
| PIS | Phosphatidylinositol synthase |
| PISD | Phosphatidylserine decarboxylase |
| PL | Phospholipids |
| ppm | Parts per million |
| PS | Phosphatidylserine |
| PSS | Phosphatidylserine synthases |

| | |
|---------------|---|
| PTPMT | Mitochondrial phosphatase |
| PUFA | Polyunsaturated fatty acids |
| QC | Quality control |
| qDB | Query database |
| Q-ToF | Quadrupole time of flight |
| RP | Reversed-phase chromatography |
| RT | Retention time |
| SCD1/5 | Stearoyl-CoA desaturases 1/5 |
| SFA | Saturated fatty acids |
| SL | Sphingolipids |
| SM | Sphingomyelin |
| SMS | Sphingomyelin synthase |
| Sph | Sphingoid base |
| SPHK | Sphingosine kinase |
| SphP | Sphingosines phosphate |
| SPT | Serine palmitoyl transferase |
| TG | Triacylglycerol |
| TIC | Total ion chromatogram |
| UPLC | Ultra performance liquid chromatography |
| VMH | Virtual metabolic human database |
| WOS | Web of science |
| 3KSR | 3-ketosphinganine reductase |

Abstract

The increasing interest in understanding the role of lipids in cell function and disease has promoted great advances in the field of lipidomics during the last decade. Yet lipid identification stands out as the main bottleneck in the lipidomic analysis workflow. Additionally, the biological interpretation of the specific functions of different lipid species remain unknown. The main objective of this thesis was to develop analytical methods and computational tools that facilitate the analysis of the human lipidome and assist the analyst to unravel the complex metabolic network behind fatty acid metabolism. To this end, two different tools have been developed and evaluated in different relevant biological scenarios. LipidMS, a tool for data management and lipid annotation in LC-MS-based lipidomic analysis, and FAMetA a tool aimed to determine the complex fatty acid metabolic network using ^{13}C -isotope tracers and MS-based analysis.

A mi padre,

General Introduction

1. Metabolomics

Bioinformatics and state-of-the-art analytical technologies have promoted the development of the so-called *omics* sciences, which are aimed to study the whole set of biomolecules and biopolymers in a particular biological sample such as genes (genomics) and their epigenetic modifications (epigenomics), mRNA (transcriptomics), proteins (proteomics) and metabolites (metabolomics). Particularly, metabolomics, understood as the unbiased determination of all the small molecules (<1.5kDa) present in a biological system¹, has experienced a continuous growth over the last two decades. Metabolites are the end products of the “*omics* cascade” and their levels constitute a direct reflection not only of the metabolism of the system under study but also of all levels of regulation upstream to the metabolism² (Figure 1).

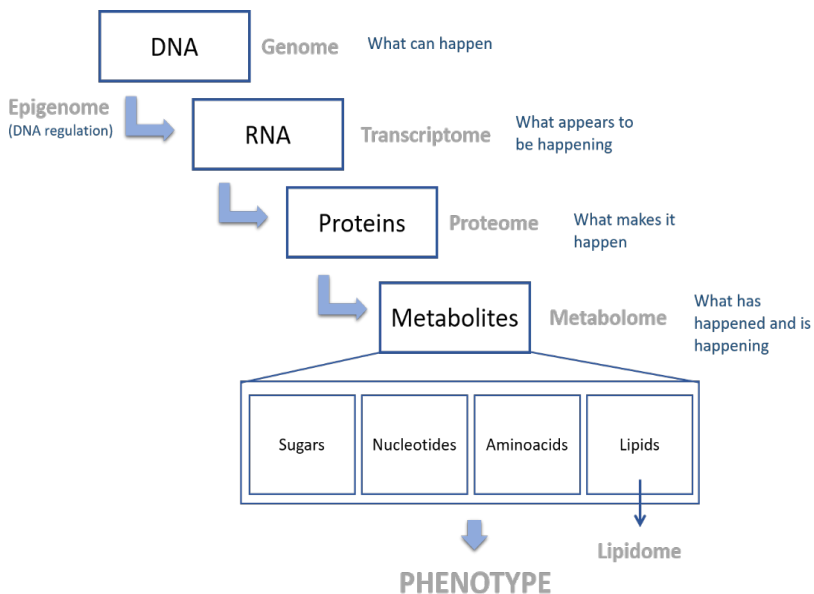


Figure 1. The “*omics* cascade”. Metabolites are the downstream end products of the “*omics* cascade” and, as such, are the closest reflection of the phenotype of the system under study. Adapted from ref.².

Therefore, the metabolome integrates intrinsic (i.e., genome) and extrinsic (i.e., diet, exposure to drugs/xenobiotics...) information at a specific and unique physiological state, providing a closer readout of the phenotype than other *omics*³⁻⁵. Comparison of different metabolomic profiles (e.g., control versus disease, external stimuli, drug intake...) may provide both the discovery of biomarkers and new knowledge of the biochemical mechanisms underlying the pathophysiology of human diseases^{1,6,7}. Improvements in analytical techniques and the development of new bioinformatic tools have allowed the fast evolution and impact of this discipline⁸.

1.1. Analytical techniques used in metabolomics

Currently, nuclear magnetic resonance (NMR) and mass spectrometry (MS) are the most widely used analytical platforms in metabolomics. NMR is a highly reproducible spectroscopic technique, but its metabolome coverage is limited to detect most abundant and highly concentrated metabolites ($\geq 1\mu\text{M}$), depending on the spectral resolution and biospecimen^{9,10}. Conversely, MS has the potential to measure metabolites at very low concentrations (fM to aM) within a wide dynamic range. Additionally, MS can be easily coupled to a variety of separation techniques, which contribute to expand the number of detected metabolites. Furthermore, the combination of spectral resolution, mass accuracy and mass fragmentation have considerably improved MS-based metabolite identification capabilities. These technical breakthroughs have promoted MS as the foremost analytical technique used in metabolomics⁸ (Figure 2).

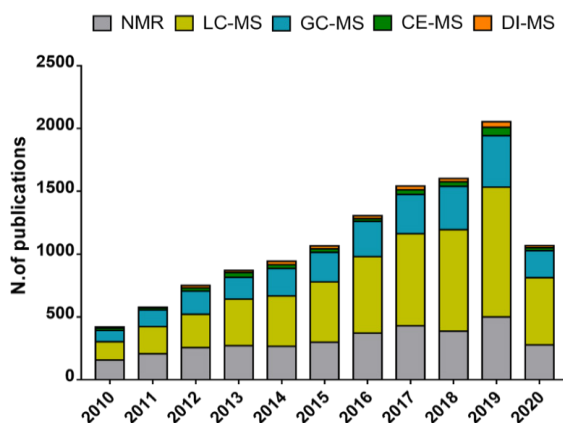


Figure 2. Systematic literature analysing the use of MS and NMR platforms in metabolomics. Bar plot showing the number of published articles per year using NMR (grey) and MS (LC-MS in green, GC-MS in blue, CE-MS in dark green and DI-MS in orange) for metabolomic analysis during the last decade based on the Web of Science. Up to August 2020⁸.

MS-based metabolomics analysis is typically performed by two different but complementary approaches, known as targeted and untargeted metabolomics. The former focuses on the quantification of a pre-determined set of known metabolites, while the latter aims to determine the widest range of metabolites possible. An alternative approach is to combine them into what is called pseudotargeted analyses, in which a pooled sample is first characterized using an untargeted method and then the identified metabolites are quantified in individual samples using a targeted method¹¹. Targeted approaches are usually performed using low-resolution mass spectrometers as triple quadrupole, triple quadrupole ion trap or quadrupole linear ion trap, which offer high sensitivity and usually work in multiple reaction monitoring in which each metabolite of interest is quantified by monitoring one or more characteristic precursor to product transitions¹². Conversely, untargeted approaches requires high-resolution mass spectrometers (HRMS) such as quadrupole time of flight (Q-ToF), (Q)-Orbitrap, or Fourier transform ion cyclotron resonance. In HRMS,

metabolite identification is usually performed by combining accurate mass (<1-10ppm) with mass fragmentation data¹³.

Due to sample complexity, MS is usually hyphenated to different separation techniques such as liquid chromatography (LC), gas chromatography (GC) or capillary electrophoresis. This previous sample separation may reduce matrix interference, where compounds affect each other's ionization efficiency leading mainly to ion suppression effects. In addition, these separation techniques improve the identification of isobaric metabolites (particularly those with very similar or indistinguishable fragmentation patterns) or metabolites with identical fragment ions generated in source (e.g., ATP fragmenting ADP, or glutathione fragmenting glutamate). Between them, LC-MS has enjoyed a growing popularity as the platform for metabolomic studies due to its high throughput, soft ionization (e.g., through electrospray ionization (ESI)), and good coverage of the metabolome (Figure 2).

1.2. Untargeted LC-MS metabolomics

As mentioned above, untargeted metabolomics is focused on global detection and relative quantification of the maximum number of metabolites possible. This approach does not require any previous knowledge and can provide a comprehensive view of the samples under study, what may allow the generation of new hypothesis that will need to be further validated using targeted approaches¹³. Untargeted approaches can be classified into three different categories based on the aim of the study: i) biomarkers discovery, where the objective is to find metabolites for diagnosis, prognosis or treatment response related to a disease progression; ii) pathogenesis studies, which aim to unravel the mechanisms under a disease; and iii) association studies, which are

focused on the search of correlations between the metabolome and physiological or clinical factors¹⁴.

A general workflow for untargeted LC-MS metabolomics analysis is shown in Figure 3, which includes experimental design, sample

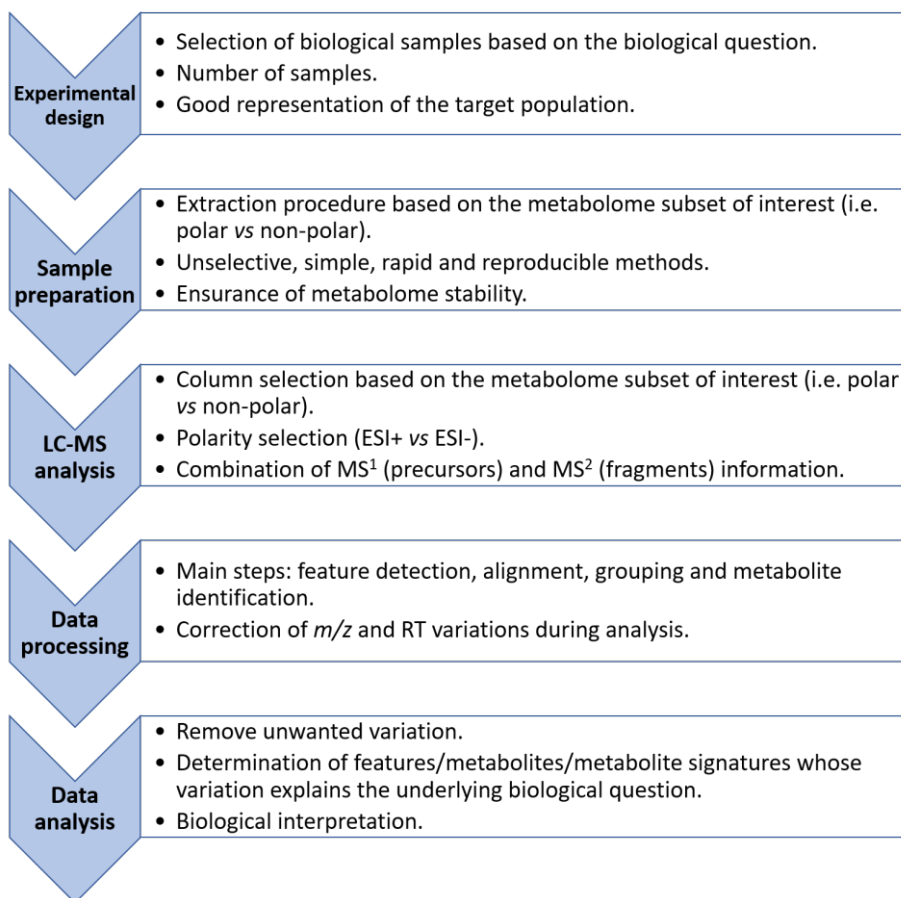


Figure 3. General workflow in untargeted LC-MS metabolomics. It comprises 5 main steps: i) experimental design, which consists of selecting the required number and type of samples to answer the biological question; ii) sample preparation, which implies extracting the metabolites of interest from the samples; iii) LC-MS analysis, which consists of separating and detecting those metabolites; iv) data processing, which is transforming the three dimensional raw data into features that can be compared between samples; and v) data analysis, which consists of extracting the biological variation, identifying the metabolites of interest and interpreting the results.

preparation, LC-MS analysis, data processing and data analysis. Since there is no predefined set of metabolites of interest, each step of this workflow should aim to detect as many compounds as possible without introducing any kind of bias. The following sections will briefly present the main aspects to be considered in each step of this workflow, paying special attention into data processing, since the main objective of this thesis is related to this step.

1.2.1. Experimental design

The very first step in any biological experiment is the study design, which must be focused in the biological question to be answered. In this regard, the selected samples (e.g., *in vivo* vs *in vitro*, biological matrix, cell cultures...) need to be representative of the population under study. For example, although non- or minimally-invasive samples as urine and serum are easily obtained and are highly recommended, the information provided by them would not be necessarily informative of the specific condition under study (e.g., metabolic reprogramming in tumors), and in this situation maybe a biopsy of the tissue could be more suitable.

On the other hand, the number of samples must ensure a good representation of the target population, sufficient homogeneity (e.g., physiological and/or demographic factors) and provide enough statistical power based on the aim of the study¹⁴. For example, when further cross-validation procedures are going to be applied in order to assess the predictive power of a biomarker, a training and validation datasets will be needed and they must be representative of the population and large enough to construct a robust model.

Additionally, the use of quality control (QC) pooled samples also need to be considered in order to ensure good quality and reproducibility of the results. QC samples, prepared from a small aliquot of all the samples included in a study, allow the evaluation and correction of the technical variation occurred along all the steps of the metabolomic analysis. In addition, QC samples can also be used for signal correction and data normalization during the data processing step¹⁴⁻¹⁶.

1.2.2. Sample preparation

Sample preparation is a critical step within the metabolomics workflow, which affects both the eventual metabolome coverage and the quality of the obtained data¹⁷. It must guarantee the stability of the metabolome, avoiding metabolite losses and the occurrence of artefacts. In addition, to obtain the closest snapshot of the whole metabolome underlying a specific biological condition, the sample preparation approach should enable the extraction of the largest number of metabolites possible, without introducing any kind of bias toward certain chemical families or physical localizations. Briefly, an ideal sample preparation method for untargeted metabolomics should be: i) unselective; ii) simple and rapid with a minimum number of steps; and iii) reproducible¹⁸. Unfortunately, there is not a common method for all the biospecimens, thus, sample preparation should be carefully adapted to the nature (e.g., liquid, solid, soft, hard...) and the chemical properties (e.g., protein content, salts, expected metabolome composition...) of the available biological sample¹⁹.

1.2.3. LC-MS analysis

In LC-MS, metabolites are first separated by liquid chromatography based on their affinity to the chromatographic column and the mobile phases employed. Then, metabolites are ionized and detected by the mass spectrometer, what results in three-dimensional raw data characterized by: retention time (RT), mass-to-charge ratio (m/z) and intensity (Figure 4). The intensity values represent the counts in a short time frame (i.e., a scan) of each ionized molecule, which are characterized by a unique m/z value.

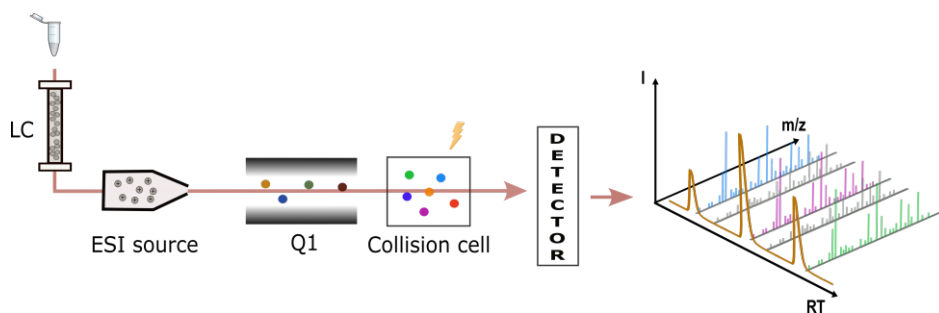


Figure 4. Schematic diagram of LC-MS in untargeted metabolomics. Metabolites are separated based on their affinity to the column, ionized and detected, which results in three dimensional raw data characterized by m/z , RT and intensity. Optionally, the ionized molecules can be filtered at the quadrupole (Q1) and fragmented at the collision cell.

The most commonly used LC approaches are: i) reversed-phase (RP), which depending on the stationary and mobile phases is able to separate a wide range of metabolites in the medium to non-polar range; ii) hydrophilic interaction liquid chromatography, which properly separates polar compounds; and iii) ion-pairing, which is most suited to separate ionic and ionizable metabolites using typical RP chromatography configuration thanks to the use of an ion pairing agent.

After LC separation, molecules in the liquid phase have to be ionized into gas-phase ions in order to be MS-detected. Currently, ESI is the most used atmospheric ionization technique, as it provides a soft ionization of small semi-polar and polar molecules, what allows the detection of intact molecules, in both positive (mainly as $[M+H]^+$) and negative (mainly as $[M-H]^-$) ionization modes^{20,21}. The analysis of LC-MS spectral databases shows that most metabolites can be detected either in positive or in both positive (ESI+) and negative (ESI-) modes²²⁻²⁷. Furthermore, ESI- provides additional information in the case of organic acids, lipids and lipid-like molecules and carbohydrates and conjugates.

Finally, MS analysis is usually performed using high-resolution mass spectrometers such as Q-ToF or Q-Orbitrap which reach a mass accuracy below 1-10ppm. Usually, untargeted LC-MS metabolomics require the combination of full scan acquisitions (MS^1 level), which provide information about the nominal mass and formula of the metabolites, and MS/MS acquisitions (MS^2 level), where parent ions are fragmented by a collision energy providing information about the structure of the metabolites. Both levels of information will be required for subsequent metabolite annotation. Currently, there are two main approaches to generate LC-MS/MS data (Figure 5): data-dependent acquisition (DDA), in which some precursors (i.e., top N most intense precursors of each MS^1 scan or precursors present in a inclusion list) from MS^1 are selected and immediately fragmented to provide a clean spectrum that can be directly queried against a spectra database²⁸; and data-independent acquisition (DIA), where no precursors from MS^1 are isolated and all the ions are subsequently fragmented²⁹. This avoids missing information about less abundant precursors or closely coeluting isomers that would not be selected for fragmentation in DDA. While DDA approaches offer easier-to-interpret but more limited data, DIA approaches return more complex data sets containing the MS^2

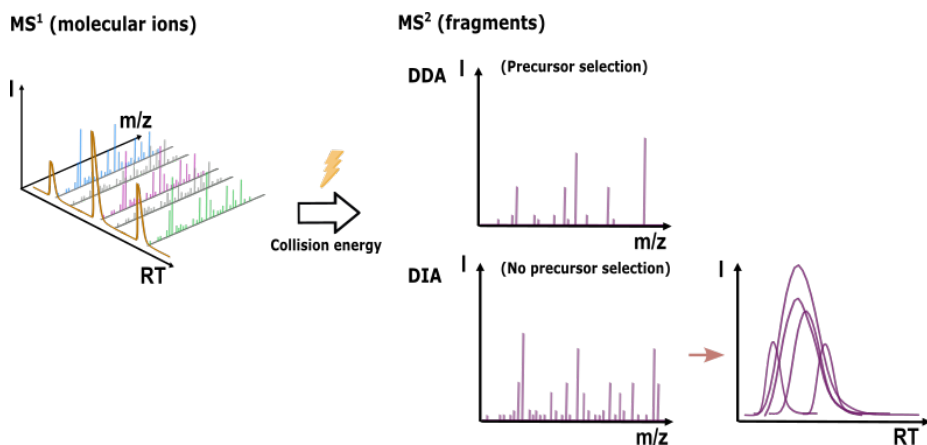


Figure 5. Scheme of common acquisition modes in untargeted LC-MS metabolomics. In DDA, specific precursors (MS¹) are selected to be fragmented (MS²), while in DIA, all precursors are fragmented at the same time. While in DDA, the link between MS¹ precursors and MS² fragments is straightforward, in DIA, a deconvolution process is required.

information for all coeluting precursors present in the sample. Unlike for DDA acquisitions, in DIA approaches the precursor-fragments links are lost, so it is necessary to reconstruct these relationships by a deconvolution process based on the peak shape of precursors and fragments to be able of combining MS¹ and MS² information for metabolite identification²⁹.

1.2.4. Data processing

Once the LC-MS analysis has been performed and in order to obtain meaningful results, raw LC-MS data need to be processed to extract the levels of the metabolites that are present in the samples of interest. The objective of this data processing step is to build a numerical matrix that contains the intensity of each detected peak for all the samples analyzed. This matrix will be used subsequently to perform the required downstream analysis (e.g., normalization and statistical analysis). This specific data processing workflow comprises

several steps: i) extraction and quantification of all features or peaks present in each sample (peak-picking); ii) time-drift correction occurred between samples along the analytical sequence (alignment); iii) grouping of those signals from different samples that belong to the same feature (peak grouping); iv) peak filling for retrieving missing peaks; and v) identification of the detected metabolites (Figure 6).

Main steps in LC-MS data processing

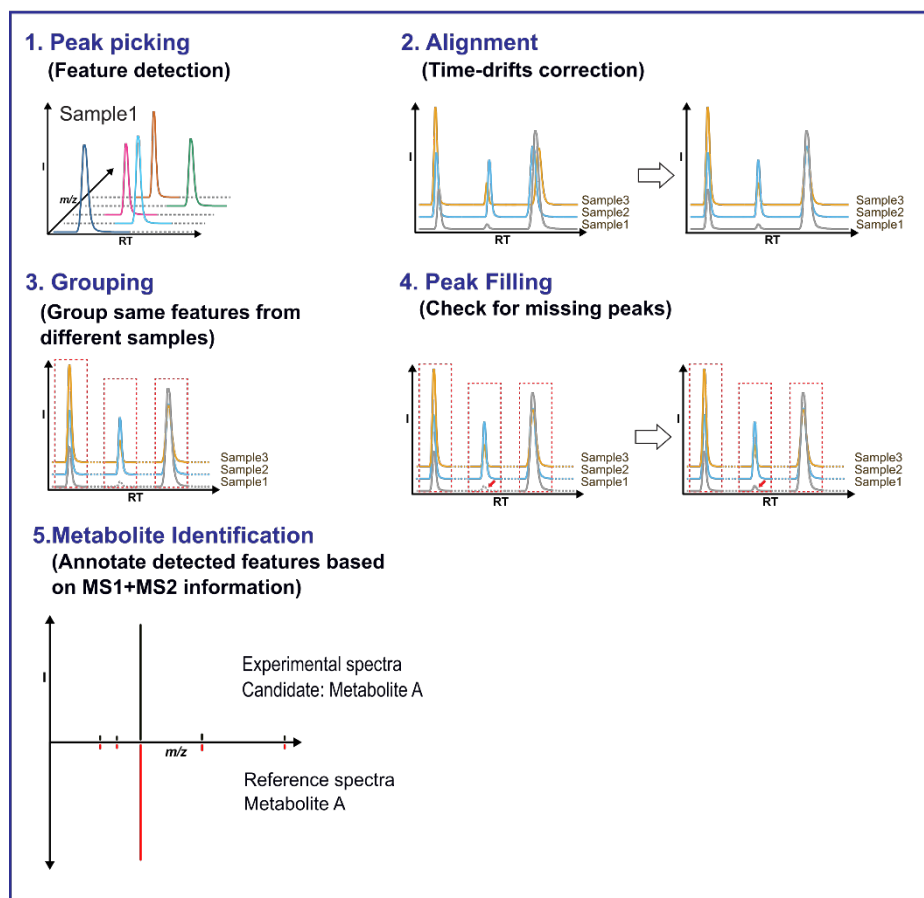


Figure 6. Main steps in untargeted LC-MS data processing. i) Peak-picking, which extracts features from the raw data for each sample; ii) alignment, which corrects the time drifts occurred along the analytical sequence; iii) grouping, which matches peaks from different samples that correspond to the same feature; iv) peak filling, which retrieves missing peaks; and v) metabolite identification, which annotates the detected features based on the MS¹ + MS² information.

1.2.4.1. Peak-picking

The first step in LC-MS data processing is peak-picking, which consists of extracting and quantifying the features present in each sample. Each feature refers to a peak defined by a unique combination of m/z and RT, and it can be a molecular ion, adduct, isotope or fragment of a metabolite, so several features could come from a unique metabolite.

Most common peak-picking algorithms, comprise two steps: i) construction of extracted ion chromatograms (EIC) and ii) detection of peaks from those EIC. The EIC construction consists of bucketing all m/z values found along the time axis within a certain m/z range as a unique m/z signal³⁰, in order to reduce data dimensions from 3D to 2D (RT and intensity). Several methods based on binning³¹ and clustering have been developed to this end^{30,32-34}. After EIC construction, peak detection algorithms try to find the time bounds that define each peak. For this purpose, different algorithms based on peak shape, peak width, signal-to-noise ratio and/or signal-to-baseline ratio have been developed to extract unique peaks from the previously defined EIC³⁴⁻³⁶. At the end of this step, each peak will represent a feature.

1.2.4.2. Alignment

When multiple samples are analyzed within a study, it is common to observe m/z and RT drifts throughout the sample batch, which are attributed to different factors that affect the LC-MS instrument (e.g., temperature, changes in the mobile phases or ion suppression, among others). Unfortunately, these drifts in m/z and RT are translated into changes in the extracted features for each sample, what may difficult the subsequent data analysis. Alignment consists of

correcting those RT drifts between samples so that peaks that represent the same feature have a similar RT.

Most common alignment algorithms are based on warping, which consists on shifting, stretching or squeezing different parts of the chromatogram to minimize differences between runs³⁷. Correlation Optimized Warping^{38,39} and Dynamic Time Warping⁴⁰ algorithms try to minimize differences in the total ion chromatogram (TIC) between a reference sample and the others, while OBI-warp algorithm⁴¹ uses EICs. On the other hand, mzMine^{42,43} and XCMS³¹ alignment algorithms, among others, minimize RT deviations between matched features from different samples.

1.2.4.3. Grouping

As well as RT drifts, m/z variations along the sample batch need to be considered to compare the features obtained for different samples. Grouping or matching consists of linking peaks from different samples that correspond to a unique feature considering the small variations in m/z and RT between samples³⁷. Most grouping methods are based on clustering algorithms that use predefined m/z and RT tolerances^{31,42}. Once all peaks have been grouped, only those features that are present in a sufficient number of samples, which is usually customizable, will be kept.

1.2.4.4. Peak filling

Due to the analytical variations or the parameter setting used for data processing, some peaks may not have been extracted correctly in the previous steps. Once alignment and grouping have been performed, all the features present in the sample batch have been defined (m/z and

RT), and this information can be used, optionally, to search for those missing peaks in a targeted manner or to improve the area integration of all peaks.

1.2.4.5. Metabolite identification

Metabolite identification, which is the major bottleneck of untargeted metabolomics, is required to give biological meaning to the results of any untargeted metabolomic analysis. Most common approaches for metabolite identification rely on querying metabolomic databases with a given tolerance to find candidates for the features extracted. According to the Metabolomics Standards Initiative there are four levels of metabolite identification⁴⁴: level 1, the highest, where the metabolite identification is confirmed by using a chemical standard that is detected using the same analytical conditions and comparing its m/z , isotope pattern, RT and MS/MS spectrum with the feature; levels 2 and 3, when no standard is available or employed but MS/MS spectra is matched by similarity against a spectral library (in case of level 3, only the chemical class is confirmed); and level 4, which comprises unknown compounds.

Generally, MS¹ and MS² levels of information are required to annotate unknown metabolites. In MS¹, molecular ions m/z depend on several factors such as ionization mode (i.e., ESI+ or ESI-), adducts formation (i.e., Na⁺, NH₄⁺) and neutral losses (i.e., H₂O, HCOOH), so that multiple features may represent a unique metabolite. Thus, the use of a unique m/z value may lead to false positive annotations. In addition, the existence of isobaric and isomeric compounds, makes necessary the use of the MS² information to further elucidate the metabolite structure. As mentioned above, querying MS/MS spectra against metabolomics databases has become the most common method to annotate unknown metabolites and comprises two steps: i) precursor ion m/z is used to filter

candidates from the whole database, and ii) MS/MS similarity between the unknown feature and the selected candidates are scored and ranked based on m/z values and product ion intensities⁴⁵. Several databases such as Human Metabolome Database (HMDB)²², METLIN²⁵, MassBank²⁴ or The Massbank of North America (MoNA)⁴⁶, among others, contain thousands of experimental and *in silico* generated MS/MS spectra, yet the known metabolome is far of being complete.

1.2.5. Data analysis

The final step in untargeted LC-MS metabolomics consists of finding those metabolites or features whose variation may explain the differences between the sample groups and interpreting their biological meaning. To this end, experimental variation introduced at different levels of the metabolomic workflow (e.g., sample collection and preparation, metabolite extraction, analytical platforms) need to be firstly removed⁴⁷. In addition, data analysis has to deal with some challenges such as high noise levels, highly correlated features due to the presence of isotopologues, adducts and in-source fragments and variability in signal sensitivity (i.e., intensity), mass accuracy (i.e., m/z) and RT due to long periods of analysis⁴⁸. First, processed peak tables are corrected, normalized and filtered to remove unwanted variation^{47,49-51}. Then, proper statistical analysis can be performed. Two different strategies can be differentiated in untargeted metabolomics data analysis: “bottom-up” and “top-down”⁵². In “bottom-up” approaches, also known as “metabolite profiling”, predefined sets of metabolites, which usually have been previously identified, are analyzed in order to extract the most significant variables that explain the underlying research questions. In this strategy, features are treated individually. Univariate statistical methods, Hierarchical Clustering Analysis (HCA) or unsupervised Principal Component Analysis (PCA) are commonly used methods for this kind of analysis⁵³. On the other hand, “top-down”

approaches, also known as “metabolite fingerprint”, aim to extract metabolite or feature signatures underlying the differences between the groups under study. In this case, all features are usually analyzed before the identification step. Supervised Multivariate methods such as (Orthogonal) Partial Least Squares Discriminant Analysis or machine learning methods such as Random Forest, Support Vector Machines or Neural Networks are applied in this case to reduce high-dimensional data^{54,55}. In addition, feature selection methods such as Recursive Feature Elimination, Lasso, Elastic Net, Ridge Regression or Sparse N-way Partial Least Squares can also be employed^{56,57}. Finally, once data analysis has been performed, biological interpretation is required to understand the results. Recently, a great number of algorithms have been developed to facilitate this functional analysis using predefined sets of related metabolites based on prior knowledge of metabolic pathways or biological functions instead of treating them as single units. Most common algorithms are Metabolite Set Enrichment Analysis⁵⁸, Metabolic Pathway Analysis⁵⁹ or Metabolic Network Analysis⁶⁰. In this sense, MetaboAnalyst⁶¹, which is one of the most commonly used tools for data analysis, includes a wide variety of these methods for data processing, data analysis, functional analysis and data integration from multiple *omics*.

2. The human metabolome

The human metabolome can be defined as the collection of all metabolites present in the human body under a particular condition and it includes a heterogeneous group of intermediates and end products of the metabolism such as lipids, amino acids, peptides, carbohydrates and their conjugates, nucleic acids and amines among others. The human metabolome has been estimated to comprise from several thousands to a few millions of metabolites, but only a few thousands of them have actually been detected and quantified^{8,22,62}. Although improvements in analytical techniques and the development of theoretical metabolites databases and bioinformatics tools have allowed to notably extend the metabolome coverage, a large part of it still remains unknown or undetected⁸. This thesis is mainly focused on the use of LC-MS, therefore, we first attempted to ascertain which is the current and actual human metabolome coverage that can be obtained when LC-MS is used and which part of it correspond to lipids. To this end, we first analyzed metabolite data from four of the main publicly and downloadable databases comprising; i) the human metabolome database (HMDB)²², ii) the Virtual Metabolic Human database (VMH)⁶³, iii) HumanCyc⁶⁴ and iv) KEGG database⁶⁵. InChI codes⁶⁶ were used to identify unique compounds across different databases⁶⁶. Our analysis estimated that the current known human metabolome is around 118,000 compounds⁸ (Figure 7). However, it should be noted that such figure and the number of metabolites represented are biased due to the own nature of our current knowledge and the methodology used to build the databases. On the one hand, many lipid classes have a backbone structure that defines the class to which various fatty acyl moieties are attached. This modular nature allows the prediction of thousands of lipids, which in turn causes their artificial overrepresentation in the expected human metabolome.

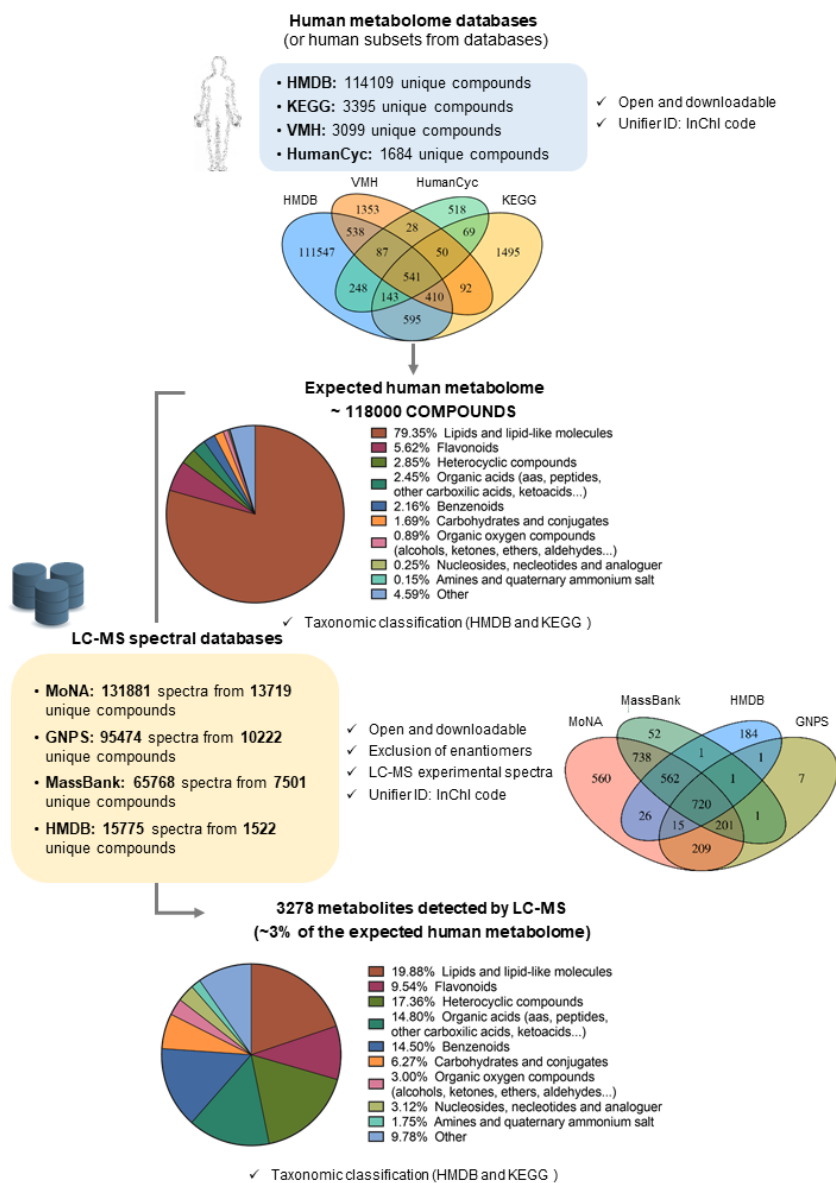


Figure 7. Analysis of the human metabolome coverage provided by LC-MS. The workflow followed to define the expected human metabolome is based on freely available metabolome databases, and the fraction of this metabolome that can be analyzed by LC-MS is based on open spectral databases. The taxonomic classification of metabolites was performed with a combination of chemical structures and biological functions based on HMDB and KEGG classification, as described in the Supplementary Information.

On the other hand, most of the metabolites comprised in those 118,000 are endogenous metabolites. However, within humans a huge variety of exogenous compounds (e.g., diet, chemicals, microbiota, drugs...) are also present and have important biological roles. In addition, such exogenous compounds can be transformed through a variety of enzymatic reactions (e.g., oxidation, reduction, hydrolysis, conjugation...), which importantly increases their number and diversity. In this regard, Wishart and colleagues have recently introduced BioTransformer, a computational tool that predicts metabolite biotransformation⁶². Based on the current figures of metabolites across several databases (including endogenous and exogenous metabolites), the authors estimated that the metabolome size could reach around 5,000,000 compounds. However, it remains difficult to ascertain how many of them would be actually found in human samples and most likely only a small fraction of them would be present at enough concentration to be detected with currently available technology. Then, to estimate which part of this human metabolome has actually been detected by LC-MS, we queried the estimated human metabolome against a variety of open downloadable spectral databases, including MoNA⁴⁶, GNPS²³, Massbank²⁴ and HMDB²². Common databases such as METLIN²⁵, mzCloud²⁷ or LIPID MAPS²⁶ were not considered for the survey because they did not meet the criteria of being free, downloadable and experimental. Again, InChI codes were used as unique compound identifiers⁶⁶. Standard untargeted LC-MS techniques are not suitable for enantiomers separation and in many cases only a particular enantiomer is expected, active or abundant enough, thus it was decided to consider only one enantiomer of each pair. Such criterion reduced the figure into around 14,000 compounds. Intriguingly, for the remaining 104,000 metabolites, only around 3,200 have an experimental LC-MS spectra in these databases, which only

accounts for a 3% of the estimated human metabolome (Figure 7). As pointed above, such low percentage can be attributed to the overrepresentation of predicted lipids in the databases. For example, in the HMDB, 80% of the listed metabolites are lipids and 92% of them are predicted (Figure 8), that is, without experimental spectra and in most cases without known biological role. In addition, the lack of lipid standards hinders the availability of experimental spectra in the databases. In this respect, the use of bioinformatics tools to analyze LC-MS data has considerably extended the lipidome coverage in a variety of biological samples^{62,67-69}, but there is still an important disagreement between the expected number of lipids and those actually detected in untargeted LC-MS studies. Apart from the issues related to the nature of lipids, we have to consider that databases are not always updated at the same pace that new LC-MS methods for extending human metabolome appear or new metabolites are proposed. Incorporation of experimental LC-MS data into databases requires huge efforts, thus, a gap between new metabolomics achievements and incorporation of such information into databases will always be present.

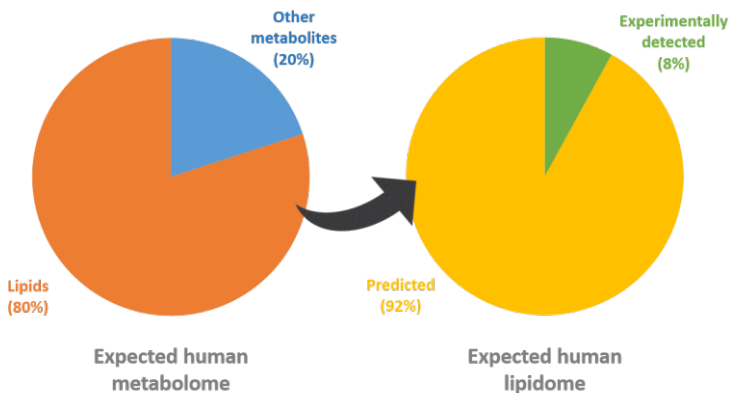


Figure 8. Representation of predicted lipids in main metabolomic databases.

3. Lipidomics

As mentioned in the previous section, lipids comprise a large and heterogeneous class of metabolites composed by the combination of different building blocks (Figure 9). Lipids are involved in many biological functions as intermediates or products in signaling pathways, structural components of cell membranes and energy store sources,

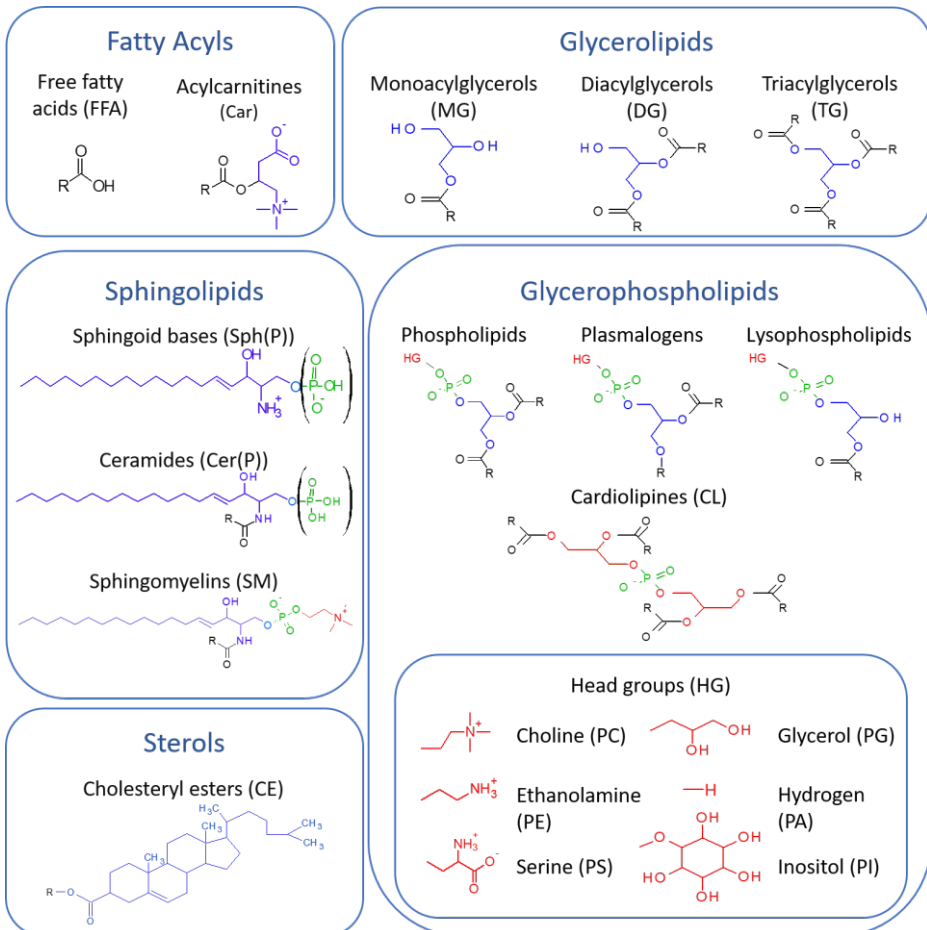


Figure 9. Main lipid classes and key building blocks of the human lipidome. Blue represents the main building block for each group (e.g., carnitine, glycerol, sphingoid base, cholesterol...), green shows the phosphate groups and red shows the different head groups in the glycerophospholipids.

among others⁷⁰. For all these reasons, lipidomics has emerged as a subdiscipline of metabolomics by its self, which refers to the systemic-scale analysis of all lipids present in a biological sample⁷¹. The LIPID MAPS Consortium, classifies lipids into eight classes: fatty acyls, glycerolipids, glycerophospholipids, sphingolipids, sterol lipids, prenol lipids, saccharolipids and poliketides, with a heterogeneous distribution within species and tissues^{72,73}, being the first five categories the most common and abundant lipid classes in human samples^{22,74}.

3.1. Fatty acyls

Fatty acids (FA) are the key building blocks of main lipid classes and are composed by a linear alkyl chain of variable length with a terminal carboxyl group. They usually have an even number of carbons (commonly from 14C to 24C) and a variable number of desaturations giving rise to a wide variety of saturated (with no double bounds within the alkyl chain), monounsaturated (with one double bound) and polyunsaturated (with two or more double bounds) FA, also referred as SFA, MUFA and PUFA, respectively. While free fatty acids (FFA) play a central role in cellular biology, they are mainly found in their esterified form as part of complex lipids⁷⁵. Oleic acid (FA(18:1)n9), where 18 informs about the number of carbons of the alkyl chain, 1 represents the number of desaturations and n9, also referred as omega 9 (ω 9), indicates the position of the last double bound, followed by palmitic (FA(16:0)) and stearic acid (FA(18:0)) are the most abundant FA in human samples, representing about an 80% of all FFA of human plasma⁷⁴.

FA can be either synthesized *de novo* inside cells or imported from external sources (Figure 10). The main product of *de novo*

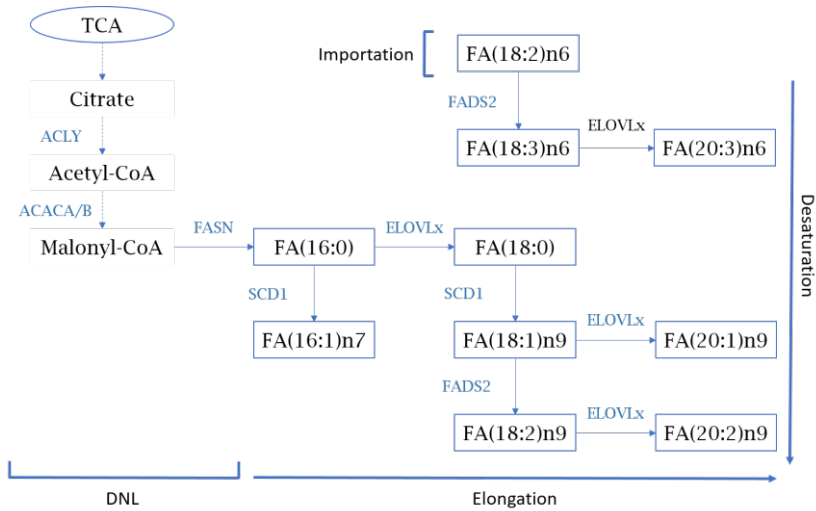


Figure 10. Main FA biosynthetic reactions. End product of the de novo lipogenesis (DNL) is the palmitic acid (FA(16:0)) and essential FA (n3 and n6 series) are imported from external sources. From these FA, elongation and desaturation reactions give rise to the whole FA variability. ELOVLx represent different elongases.

lipogenesis (DNL) is FA(16:0), which results from the condensation of acetyl-CoA molecules through the enzymatic action of acetyl-CoA carboxylase (ACACA/B) and FA synthase (FASN). The acetyl-CoA pool is generated via ATP citrate lyase (ACLY) from citrate which can, in turn, be produced from several carbon sources (i.e., glucose, glutamine, amino acids, FA), or from acetate via acetyl-CoA synthetases (ACSS1/2)⁷⁶. Linoleic (FA(18:2n6)) and γ -linolenic acid (FA(18:3n3)) are essential FA that must be exogenously acquired. Free FA import occurs by either passive diffusion or the action of translocases like CD36 and FA transport proteins (FATP). FA can be elongated via the elongation of very long-chain FA proteins (ELOVL1-7). They can also be unsaturated via the action of stearoyl-CoA desaturases 1/5 (SCD1/5) and FA desaturases 1/2 (FADS1/2) enzymes^{77,78}. All these transformations produce the wide variety of FA required for the cellular functioning.

3.2. Glycerolipids

Glycerolipids (GL) are simple lipids composed by a variable number of FA molecules (from 1 to 3) esterified to a glycerol backbone resulting in monoacylglycerols (MG), esterified by a unique FA, diacylglycerols (DG), esterified by 2 FA molecules, and triacylglycerols (TG), containing 3 FA molecules (Figure 9). TG are the most abundant subclass of glycerolipid and serve as an energy storage for the organism and as precursors for membrane lipid synthesis (FA and DG). In addition, cellular lipid droplets can play an important role in lipid mobilization and membrane trafficking. Conversely, while MG and DG are considered partial glycerides or intermediates of TG synthesis and degradation, they also have important biological functions as cellular messengers, surfactants and key intermediates for the synthesis of glycerophospholipids, among others⁷⁹.

TG are mainly synthesized at the liver or the adipose tissue from several FA and glycerol-3-phosphate (G3P), which may come from glycerol, via glycerol kinase, or from dihydroxyacetone phosphate (DHAP), via glycerol-3-phosphate dehydrogenase 1 (GPD1) (Figure 11). FA need to be activated as fatty acyl-CoA in order to be attached to the glycerol backbone. This activation is performed by different acyl-CoA synthases at the endoplasmic reticulum. The first acylation is performed by glycerol-3-phosphate acyltransferases (GPAT), which add a fatty acyl-CoA molecule to the sn1 position of the G3P, resulting in lysophosphatidic acid (LPA). Then, the acylglycerol acyltransferases (AGPAT) add a second fatty acyl-CoA unit at the sn2 position, resulting in phosphatidic acid (PA). At this point, the phosphate group is removed by phosphatidic acid phosphatases, also known as lipins (LPIN), to generate DG. Finally, a third fatty acyl-CoA unit is added by diacylglycerols acyltransferases (DGAT)⁸⁰.

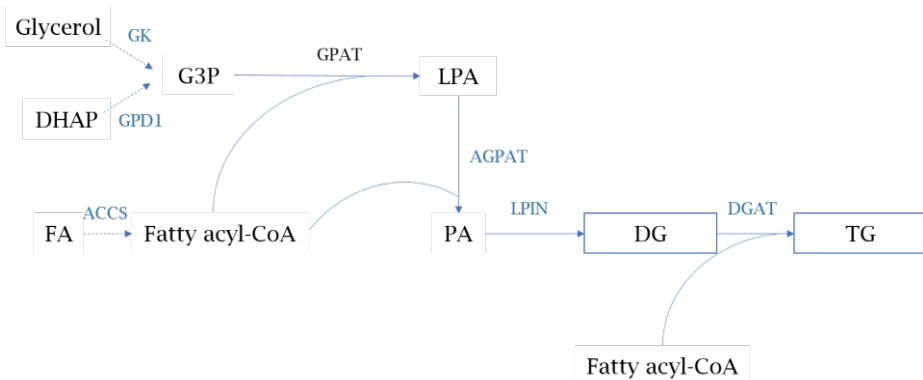


Figure 11. Main biosynthetic pathways of TG. Enzymes are shown in blue.

3.3. Glycerophospholipids

Glycerophospholipids, also known as phospholipids (PL), are the key constituents of cell membranes and are synthesized from phosphatidic acids (PA) and DG, intermediates of the TG biosynthetic pathways. They are composed by a glycerol backbone joined to different polar head groups (phosphocholine, phosphoethanolamine, etc.) and esterified by two FA molecules, resulting in a variety of phospholipid subclasses with different locations and functions within the cell membranes (Figure 9). In addition to being linked by two ester bounds, which represent the vast majority of phospholipid bounds in human and animal cells⁸¹, phospholipid FA chains can also be linked to the glycerol molecule by ether or vinyl ether bounds, usually at the sn1 position. These phospholipids are known as plasmalogen and plasmalogen PL, respectively⁷⁵. Lysophospholipids (LPL) are intermediates of the PL biosynthesis which contain a unique FA chain and have important roles as surfactants and as signaling molecules.

Phosphatidylcholines (PC) and phosphatidylethanolamines (PE), which contain a phosphocholine (PCho) or phosphoethanolamine (PEt)

group linked to the glycerol backbone as head group, are the two most abundant structural components of animal cell membranes. While PC are usually found in the external face of the bilayer, PE are mostly found in the inner leaflet⁸². PC can be *de novo* synthesized through the Kennedy pathway, which involves phosphorylation of choline (Cho) to PCho by a choline kinase (CK), which is then activated by condensation with cytidyl triphosphate (CTP) to generate cytidyl diphosphate (CDP)-Cho by a choline kinase (CK), which is then activated by condensation with cytidyl triphosphate (CTP) to generate cytidyl diphosphate (CDP)-Cho by a CDP-Cho transferase (CCT). This CDP-Cho is finally transferred to a DG molecule releasing cytidyl monophosphate (CMP) to give rise to PC by a choline phosphotransferase (CPT) (Figure 12). Alternatively, PC can also be synthesized through methylation of PE by a phosphatidylethanolamine N-methyltransferase (PEMT)^{83,84} (Figure 12). Otherwise, PE can also be *de novo* synthesized through de Kennedy pathway by phosphorylation of ethanolamine (Et) to PEt by a ethanolamine kinase (EK), which is then activated as a CDP-Et by a CDP-Et transferase (ECT). Finally, this CDP-Et is condensed by a ethanolamine phosphotransferase (EPT) with a DG to form PE (Figure 12). Alternatively, PE can also be synthesized from phosphatidylserines (PS) through decarboxylation by a phosphatidylserine decarboxylase proenzyme (PISD)⁸³⁻⁸⁵ (Figure 12).

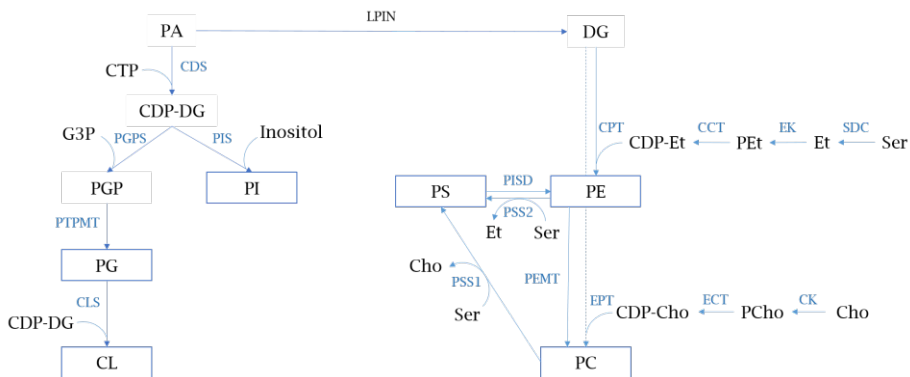


Figure 12. Main biosynthetic pathways of PL. Enzymes are shown in blue.

Phosphatidylinositols (PI) are acidic PL with inositol at the polar head and are intermediates of other lipids such as phosphatidylinositol phosphates (PIP) and DG, which have important signaling functions related to cell growth and survival among others^{79,86}. In addition, PI can serve as an anchor between the external membrane and different proteins⁸⁷. Similarly, phosphatidylserines (PS), which are also acidic PL and are usually found in the inner leaflet of the membrane, present important signaling functions related to apoptosis or as a cofactor for the activation of the protein kinase C⁸⁸, and phosphatidic acids (PA) are important precursors for other PL and GL and they are related to signaling functions⁸⁹. Finally, phosphatidylglycerols (PG), which are minor components of animal cell membranes, are key intermediates of cardiolipins (CL), PL with four FA chains that are mainly found at the mitochondrial membrane and are essential for its function (i.e. mitochondrial protein transport, morphology, signaling and oxidative phosphorylation)⁹⁰. PI and PG are synthesized through the activation of DG with CTP to generate CDP-DG followed by the displacement of CMP by specific groups depending on the PL subclass (Figure 12). In the case of PI, inositol is added by the phosphatidylinositol synthase (PIS)⁹¹, while for the synthesis of PG, G3P is first condensed with CDP-DG by a phosphatidylglycerol phosphate synthase (PGPS) to form phosphatidylglycerol phosphate (PGP), which is then dephosphorylated by a mitochondrial phosphatase (PTPMT) to generate PG⁹². An additional step is required for the synthesis of CL, where a CDP-DG molecule is added to a PG by a cardiolipin synthase (CLS)⁹³ (Figure 12). Finally, PS is synthesized from PC or PE by different phosphatidylserine synthases (PSS)⁹⁴ (Figure 12).

3.4. Sphingolipids

Sphingolipids (SL) are composed by a sphingoid base (Sph) linked to a FA chain by an amide bound (Figure 9) and, usually, to different phosphoryl or carbohydrate moieties. Free Sph can be found at trace levels in animal tissues and they present important signaling functions such as stimulation of cell proliferation⁹⁵. Ceramides (Cer), which present only a FA chain esterified to the Sph, have important functions in cellular signaling related to cell differentiation and proliferation and they are precursors of more complex sphingolipids, such as sphingomyelins (SM)⁹⁶. SM contain a glycerophosphocholine group linked to the sphingoid base and are structural components of the cell membrane mainly located at the lipid rafts⁹⁷. They are the most abundant sphingolipids in animal tissues.

SL synthesis (Figure 13) begin with the conversion of serine, or other aminoacids in a lesser extent, and a fatty acyl-CoA to 3-ketosphinganine by a serine palmitoyl transferase (SPT), which is then reduced to dihydrosphingosine by a 3-ketosphinganine reductase (3KSR) in a NADPH-dependent manner to give rise to different Sph⁹⁶. Then, different fatty acyl-CoA can be N-acylated to those Sph by a Cer synthase (CERS) to form a dihydroceramide, which is then desaturated by a dihydroceramide desaturase (DEGS) on the 4,5-bound of the sphingoid part to generate a Cer^{95,96}. From Cer, different SL can be obtained such as SM, CerP or glycosphingolipids. SM are synthesized by the action of a SM synthase (SMS) which transfers the PCho group of a PC to a ceramide to form SM and DG⁹⁸. Alternatively, Cer can be phosphorylated by a ceramide kinase (CERK) to form a ceramide phosphate (CerP)⁹⁶. Sph can also be phosphorylated by a sphingosine kinase (SPHK) to form sphingosines phosphate (SphP)^{95,96}. Finally, Cer can be glycosylated to give rise to a great variety of glycosphingolipids⁹⁶.

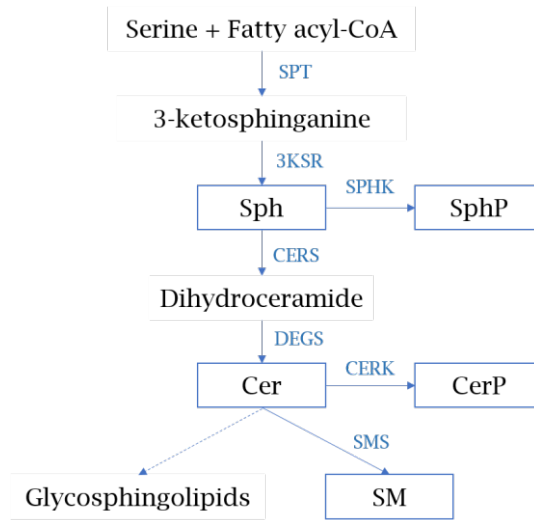


Figure 13. Main biosynthetic pathways of SL. Enzymes are shown in blue.

3.5. Sterol lipids

Sterols are polycyclic lipids derived from sterane, which play an important role modulating the membranes fluidity and stability by interacting to other structural components such as phospholipids, sphingomyelins or lipoproteins⁹⁹. Cholesterol is the most common sterol in animal tissues and it can be used to synthesize hormone and bile acids, followed by 7-dehydrocholesterol, which is used to synthesize vitamin D. Cholesterol is responsible for the order of the fatty acyl chains of phospholipids in cellular membranes, and it is a key component of lipid rafts⁹⁹. In blood, cholesterol can be found free or esterified to FA⁷⁵. Cholesterol biosynthesis starts with the condensation of three acetyl-CoA molecules to form hydroxymethylglutaryl-CoA (HMG-CoA), in a NADPH-dependent manner, which is then reduced by a HMG-CoA reductase to generate mevalonate. Then, through a series of reactions, mevalonate is converted into cholesterol. Cholesterol is then

transported to high and low density molecules (HDL and LDL) where it can be esterified to a FA chain (Figure 9) from a PC through the action of a lecithin cholesterol acyltransferase (LCAT) giving rise to cholesteryl esters (CE) and lysophosphocholines (LPC)¹⁰⁰ (Figure 14).

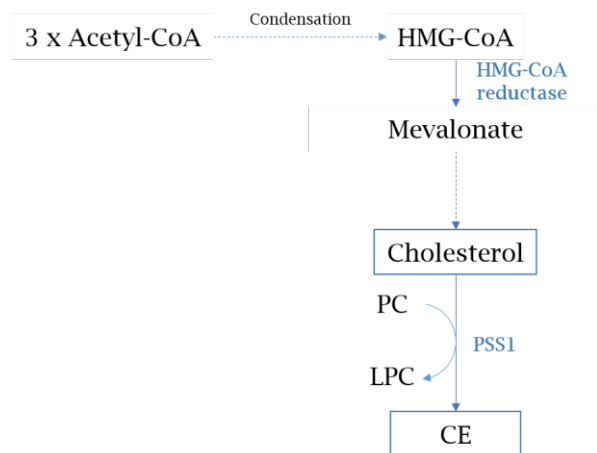


Figure 14. Main biosynthetic pathway of CE. Enzymes are shown in blue.

3.6. Lipids and disease

Lipid metabolism plays a central role in biological systems and its study may contribute to the understanding of mechanisms underlying different pathological conditions. In recent years, alterations in general lipid profiles and in particular lipid species have been identified in highly prevalent diseases as cancer^{101,102}, non-alcoholic fatty liver disease^{103,104}, diabetes¹⁰⁵, heart disease¹⁰⁶, and neurological diseases¹⁰⁷. Currently, great efforts are being directed to ascertain not only lipid-related mechanism underlying diseases but also to find new biomarkers that allow for prediction in diagnosis, prognosis or treatment response¹⁰⁸.

In the context of cancer, changes in lipid metabolism allow the adaptation of cancer cells to the harsh changing tumor microenvironment by supporting tumor cell requirements such as energy production, cell proliferation and growth, resistance to oxidative stress, intercellular communication and evasion of the immune system^{109,110}. For example, several studies have associated phospholipid composition of cell membranes and fatty acid metabolism with better survival and prognosis in breast cancer patients^{101,111}. Changes in the uptake and use of FA have also been observed in different types of cancer¹¹² such as lung¹¹³, ovarian¹¹⁴, colorectal¹¹⁵ and breast cancer^{116,117}, and FA translocase CD36 has been related to metastasis^{114,118}. In colorectal carcinoma cells, an increase in fatty acid uptake and oxidation (FAO) has been found to promote epithelial-to-mesenchymal transition, angiogenesis, tissue invasion and therapy resistance, and this dependency on FAO could be exploited by the inhibition of different enzymes involved in this pathway¹¹⁵. In this case, the authors propose to target CES1, which promotes TG breakdown to fuel FAO and oxidative phosphorylation to prevent toxic lipid accumulation. FA desaturation has also been proposed as a potential target in different tumors due to its relevance in supporting cell proliferation, but only some cancer cells are sensitive to this approach due to the plasticity of the fatty acid metabolism in tumors. In this sense, the biosynthesis of sapienic acid (FA(16:1)n10) from FA(16:0) by FADS2 allows some cancer cells to bypass the inhibition of SCD¹¹⁹.

Despite promising results are being published in clinical lipidomics research, most of the proposed lipid biomarkers are not validated or are not useful as clinical biomarkers due to the lack of specificity or sensitivity of these molecules. In addition, biological interpretation of lipid metabolism alterations is limited because specific functions of most of lipid species are still unknown. In most cases, only

global levels of lipid classes and total FFA are used for interpretation overlooking FA composition of complex lipids. Therefore, advances in analytical methods and bioinformatic tools that improve the analysis of the lipidome are still required to fully understand lipid metabolism and its implications in human disease¹¹².

4. Metabolomics and isotope tracing

The main objective of metabolomics is the characterization and measurement of the metabolites present in a biological sample. However, metabolite abundances are not enough to understand pathway activities as concentrations depend both on production and consumption rates¹²⁰. In this sense, the use of stable-isotope tracers has enabled the study of the metabolism dynamics and the assessment of the contribution of different reactions to the production or consumption of specific metabolites¹²¹. Stable isotopes are non-radioactive different forms of the same element that differ in their mass due to a different number of neutrons in their nucleus. Besides this difference in their mass, which can be distinguished by MS, these elements are chemically identical and have the same functionalities¹²². These properties have allowed the use of stable isotope tracers, molecules in which one or several atoms have been replaced by their heavier less abundant isotopic equivalent (¹²C-¹³C, ¹⁴N-¹⁵N, ¹⁶O-¹⁸O, ¹H-²H or D)¹²², to map metabolic routes by following the incorporation of these isotopic labels into downstream products¹²¹. Usually, these tracers are organic compounds such as FA, amino acids or sugars that are supplied to a biological system and are used to “trace” the metabolic fate of these compounds within that biological system¹²².

Regarding lipid metabolism, many aspects such as building block sources, function of related enzymes and transporters or lipid-lipid interactions, remain poorly understood. In order to clarify some of these aspects, stable isotope experiments can be performed. Some efforts have been made in studying the metabolism of complex lipids (e.g., phospholipids, glycosphingolipids)¹²³⁻¹²⁵, although most studies have been mainly focused on the FA metabolism^{119,126-130}.

Objectives

The general objective of this thesis was to develop new methods and freely available computational tools that facilitate the characterization of the lipidome and the study of lipid metabolism, with particular focus on FA. To this end, two main objectives were proposed:

- 1) Development of a new bioinformatic tool that improves lipid annotation in untargeted LC-MS lipidomics. This tool should cover the whole workflow required for data processing and implement rule-based identification for both DIA and DDA acquisition modes.
- 2) Development of a method that allows the study of the whole set of reactions involved in fatty acid biosynthesis based on the combined use of LC-MS and ^{13}C -tracers.

Chapter 1

LipidMS: an R-package and a web-based tool for untargeted LC-MS/MS data processing and lipid annotation

Introduction

1. Lipid identification in LC-MS lipidomics

As mentioned above, lipids can be described as a combination of different building blocks, usually a core structure that defines their class (e.g., glycerol, sphingoid bases or cholesterol) and subclass (e.g. polar head groups of phospholipids as phosphocholine and phosphoethanolamine) and a variable number of FA chains attached to the core structure¹³¹ (Figure 9). As a result of the different structural arrangements of FA and core structures, a great number of isobaric, isomeric and adduct overlaps can be found¹³² (e.g., PC(18:1/18:1) vs PC(18:0/18:2), PC(16:0/20:4) vs PC(20:4/16:0), PC(34:1) as $[M+Na]^+$ vs PC(36:4) as $[M+H]^+$ or PC(32:0) as $[M+H]^+$ vs PS(32:1) as $[M+H]^+$), which evidences the complexity of analysing the lipidome. Therefore, lipid annotation in untargeted LC-MS based lipidomics requires the accurate determination of the detected adducts and the particular building blocks that compose a given lipid and the way in which those blocks are arranged¹³¹, which in turn requires fragmentation of the precursor ions.

1.1. Challenges in lipid annotation in LC-MS-based lipidomics

The precise identification of any metabolite in LC-MS, level 1 of the Metabolomics Standard Initiative classification, requires the match of its RT, m/z and MS/MS spectra between the candidate feature and a commercially available standard⁴⁴. In case of lipids, due to the huge variety of lipid species and the reduced number of available standards, this strategy cannot be fully implemented. In this regard, the definition of fragmentation patterns for different lipid classes has allowed the construction of *in silico* MS/MS spectra libraries^{22,25,67}, which are used for lipid annotation based on spectral matching algorithms^{133,134}. Yet this strategy still has some limitations⁴⁵. First, a unique

m/z value from a precursor is not enough to identify the molecular ion (i.e., its chemical formula) due to the great amount of overlaps between isomeric and isobaric species. Table 1 shows the most common overlaps¹³². For this reason, working with high-resolution mass spectrometers and a correct isotope and adduct annotation is of utmost importance in untargeted lipidomics. In addition, although MS² information might help to distinguish some of these overlaps, it is not enough in many cases where common fragments are obtained. Moreover, if the MS/MS spectra contains a low number of fragments with high intensities, similarity scores can be skewed, and equal results can be obtained for those isobaric and isomeric species. This is very frequent in lipids, where class specific fragments that only inform about the subclass of a lipid (e.g., head group fragments) or fatty acyl chain fragments that only inform about the fatty acyl composition but not about the class or subclass of the lipid specie of interest, are common to a great number of species. Otherwise, when isobaric or isomeric compounds coelute during the chromatographic separation, which is also common due to the building block nature of lipids, complex MS/MS spectra are obtained for both DDA and DIA data, which hinders lipid annotations.

As an alternative, lipid identification based on fragmentation rules and the presence or absence of the expected fragments for each lipid class have been implemented in a few number of bioinformatic tools^{68,135}. These rules comprise class specific fragments that will only allow the annotation of a lipid class and its sum composition of carbons and double bound (e.g., PC(34:1)), chain specific fragments that will inform about the composition of its FA chains (e.g., PC(16:0_18:1)), and relative ratios of those chain fragments that will inform about the specific position of each FA (e.g., PC(18:1/16:0)). Although important efforts have been made regarding these fragmentation rules, most tools were developed only for DDA data or did not cover the whole data processing workflow required for untargeted LC-MS analysis (i.e., from peak-picking to lipid annotation).

Table 1. Common isobaric and isomeric overlaps in LC-MS lipidomics. O and P refer to plasmanyl and plasmeyl phospholipids; DB means double bound; and d and t in sphingolipids refer to dihydroxy and trihydroxy sphingoid bases.

| Polarity | Lipid classes involved | Overlap | m/z difference | Common fragments | Example |
|----------|--------------------------|--|----------------|--|--|
| Any | PC O, LPC | PC O = LPC | - | Class and chain specific fragments | PC(O-24:1) = LPC(24:1) |
| | PC P, PC O | PC P = PC O + DB | - | Class and chain specific fragments | PC(P-36:1) = PC(O-36:2) |
| | PE O, LPE | PE O = LPE | - | Class and chain specific fragments | PE(O-24:1) = LPE(24:1) |
| | PE P, PE O | PE P = PE O + DB | - | Class and chain specific fragments | PE(P-36:1) = PE(O-36:2) |
| | PC, PE | PC = PE + 3CH ₂ | - | FA chains | PC(34:1) = PE(37:1) |
| | PL, PL O | PL ≈ PL O + CH ₂ | 0.036 | Class and chain specific fragments | PC(33:1) ≈ PC(O-34:1) |
| | SL | SL ≈ SL + CH ₂ + DB - OH | 0.036 | Class specific fragments and FA chains | SM(t42:2) ≈ SM(d43:1) |
| | PI, PS | PI ≈ PS + 6CH ₂ + 5DB + ¹³ C | 0.002 | FA chains | P(34:1) ≈ PS(40:6) [M+1] |
| | PC, SM | PC [M+1] ≈ SM - H ₂ O + 4CH ₂ - DB | 0.065 | PCho related fragments and FA chains | PC(38:3) [M+1] ≈ SM(d42:2) |
| | Any specie containing DB | M ≈ M + DB + 2 ¹³ C | 0.009 | Class and chain specific fragments | PC(34:0) ≈ PC(34:1) [M+2] |
| ESI+ | PC, PA | PC [M+H] ⁺ = PA [M+NH ₄] ⁺ + 5CH ₂ + DB | - | - | PC(34:0) [M+H] ⁺ = PA(37:1) [M+NH ₄] ⁺ |
| | PE, PA | PE [M+H] ⁺ = PA [M+NH ₄] ⁺ + 2CH ₂ + DB | - | - | PE(34:0) [M+H] ⁺ = PA(36:1) [M+NH ₄] ⁺ |
| | PS, PG | PS [M+H] ⁺ = PG [M+NH ₄] ⁺ + 2DB | - | - | PS(36:0) [M+H] ⁺ = PG(36:2) [M+NH ₄] ⁺ |
| | PC, PS | PC [M+H] ⁺ ≈ PS [M+H] ⁺ + DB | 0.073 | - | PC(34:0) [M+H] ⁺ ≈ PS(34:1) [M+H] ⁺ |
| | DG, CE | CE(X:Y) ≈ DG (X+20:Y) | 0.015 | - | CE(18:1) ≈ DG(38:1) |
| | Any specie containing DB | M [M+Na] ⁺ ≈ M [M+H] ⁺ + 2CH ₂ + 3DB | 0.002 | Class and chain specific fragments | PC(34:1) [M+Na] ⁺ ≈ PC(36:4) [M+H] ⁺ |

| Polarity | Lipid classes involved | Overlap | m/z difference | Common fragments | Example |
|----------|------------------------|--|----------------|------------------|---|
| ESI- | PC, PS | PC [M+HCOO] ⁺ = PS [M-H] ⁻ + 3CH ₂ -DB | - | FA chains | PC(33:1) [M+HCOO] ⁺ = PS(36:0) [M-H] ⁻ |
| | PC, PS | PC [M+CH ₂ COO] ⁺ = PS [M-H] ⁻ + 4CH ₂ -DB | - | FA chains | PC(32:1) [M+CH ₂ COO] ⁺ = PS(36:0) [M-H] ⁻ |
| | PA, CL | PA(X:Y) [M-H] ⁻ ≈ CL(2X:2Y) [M-2H] ²⁻ | 0.018 | - | PA(34:2) [M-H] ⁻ ≈ CL(64:4) [M-2H] ²⁻ |

2. Most used bioinformatic tools for LC-MS-based lipidomics

Bioinformatic tools used in untargeted LC-MS lipidomics can be divided into three categories based on their functionality: generalist tools, data processing tools and specific lipid annotation tools⁷⁰. Generalist tools cover all the steps for the data processing workflow, from peak-picking to lipid identification, such as MS-DIAL^{133,136}, Lipid Data Analyzer (LDA)⁶⁸, Liquid¹³⁷, LipidHunter¹³⁸, LPPTiger¹³⁹, LipidSearch (Thermo Scientific), SimLipid (Premier Biosoft) or Lipostar¹⁴⁰. From those, MS-DIAL, LDA, Liquid, LipidHunter and LPPTiger are platform-independent and freely available tools aimed to process LC-MS DDA data and DIA only in the case of MS-DIAL. In addition, LipidHunter was initially designed only for phospholipids analysis and LPPTiger for oxidized lipids, although LipidHunter has recently included DG and TG identification. Otherwise, MS-DIAL and Liquid were designed to annotate lipids based on spectral similarity using *in silico*-generated libraries such as LipidBlast⁶⁷, while LDA, LipidHunter and LPPTiger used rule-based identification. Recently, MS-DIAL has also included lipid annotation based on fragmentation rules¹³⁶. The second group encompasses data processing tools, which cover most of the steps of the required workflow in lipidomics and generates feature/peak tables, but do not include lipid identification based on MS² information. Most common tools are XCMS^{31,134} and mzMine⁴³. Finally, the third group is comprised by specific tools for lipid annotation such as LipidMatch¹³⁵, LipiDex¹⁴¹ or LipidFinder^{142,143}. Usually, these tools use the outputs of the two previous groups to perform the eventual lipid identification step. From those, LipidFinder only annotates lipids putatively (i.e., based on MS¹ information) while LipidMatch and LipiDex annotate lipids based on MS² data. In case of LipiDex, it uses *in silico*-

generated libraries to identify lipids by spectra similarity only for DDA data, while LipidMatch, performs lipid annotation based on fragmentation rules for DIA and DDA acquisition modes, although initially DIA data was only supported for Thermo (.raw) data. Attending to the number of cites, from all these tools, MS-DIAL and XCMS combined with specific tools for lipid annotation are the most common freely available tools used in untargeted LC-MS lipidomics.

Despite the wide variety of free available bioinformatic tools have been developed for lipid annotation, lipid identification still remains as the most challenging step in untargeted LC-MS-based lipidomics workflow. Up the moment this thesis started, only MS-DIAL covered the whole lipidomics workflow and did not include rule-based annotation, whereas specific tools for lipid annotation based on fragmentation rules such as LDA and LipidMatch, were mainly developed for DDA data.

Methodology

1. Chemicals and reagents

Solvents for sample processing and LC-MS analysis were isopropanol, ammonium formate and ammonium acetate, all obtained from Sigma-Aldrich, and acetonitrile from Fisher Scientific. Commercial human serum was also obtained from Sigma-Aldrich (reference P2918).

Lipid standards, obtained from Avanti Polar Lipids, Sigma-Aldrich/Fluka, Larodan and Caiman Chemicals, were 1-O-oleoyl-N-heptadecanoyl-D-sphingosine (AcylCer(18:1;18:1/17:0)), cholest-5-en-3 β -yl heptadecanoate (CE(17:0)), N-heptadecanoyl-D-sphingosine (Cer(d18:1/17:0)), N-palmitoyl-D-sphingosine-1-phosphate (CerP(d18:1/16:0)), 1,3-bis-(1,2-di-octadecenoyl-sn-glycero-3-phospho)-sn-glycerol (CL(18:1/18:1/18:1/18:1)), diheptadecanoylglycerol (DG(17:0/17:0)), capric acid (FA(10:0)), lauric acid (FA(12:0)), myristic acid (FA(14:0)), myristoleic acid (FA(14:1)n5), pentadecanoic acid (FA(15:0)), palmitic acid (FA(16:0)), palmitoleic acid (FA(16:1)n7), margaric acid (FA(17:0)), stearic acid (FA(18:0)), trans-vaccenic acid (FA(18:1)n7t), oleic acid (FA(18:1)n9), linoleic acid (FA(18:2)n6), alpha-linolenic acid (FA(18:3)n3), gamma-linolenic acid (FA(18:3)n6), nonadecanoic acid (FA(19:0)), arachidic acid (FA(20:0)), gondoic acid (FA(20:1)n9), 11,14-eicosadienoic acid (FA(20:2)n6), dihom-alpha-linolenic acid (FA(20:3)n3), arachidonic acid (FA(20:4)n6), eicosapentaenoic acid (FA(20:5)n3), behenic acid (FA(22:0)), erucic acid (FA(22:1)n9), docosadienoic acid (FA(22:2)n6), 10,13,16-docosatrienoic acid (FA(22:3)n6), adrenic acid (FA(22:4)n6), clupanodonic acid (FA(22:5)n3), cervonic acid (FA(22:6)n3), lignoceric acid (FA(24:0)), nervonic acid (FA(24:1)n9), cerotic acid (FA(26:0)), heptadecanoyl-sn-glycero-3-phosphocholine (LPC(17:0)), monoheptadecanoylglycerol (MG(17:0)), 1-hexadecanoyl-2-octadecenoyl-sn-glycero-3-phosphocholine (PC(16:0/18:1)), 1,2-diheptadecanoyl-sn-glycero-3-phosphatidylcholine (PC(17:0/17:0)), 1-octadecanoyl-2-octadecadienoyl-sn-glycero-3-phosphocholine, (PC(18:0/18:2)), 1-hexadecyl-2-(5Z,8Z,11Z,14Z,17Z-

eicosapentaenoyl)-sn-glycero-3-phosphocholine, (PC(O-16:0/20:5)), 1-(1Z-octadecenyl)-2-(5Z,8Z,11Z,14Z-eicosatetraenoyl)-sn-glycero-3-phosphocholine (PC(P-18:0/20:4)), 1,2-diheptadecanoyl-sn-glycero-3-phosphoethanolamine (PE(17:0/17:0)), 1-hexadecanoyl-2-octadecenoyl-sn-glycero-3-phosphoethanolamine (PE(16:0/18:1)), 1-hexadecyl-2-(9Z-octadecenoyl)-sn-glycero-3-phosphoethanolamine (PE(O-16:0/18:1)), 1-(1Z-octadecenyl)-2-(4Z,7Z,10Z,13Z,16Z,19Z-docosahexaenoyl)-sn-glycero-3-phosphoethanolamine (PE(P-18:0/22:6)), 1-hexadecanoyl-2-octadecenoyl-sn-glycero-3-phosphoglycerol (PG(16:0/18:1)), 1,2-diheptadecanoyl-sn-glycero-3-phosphoglycerol (PG(17:0/17:0)), 1-heptadecanoyl-2-(9Z-tetradecenoyl)-sn-glycero-3-phospho-(1'-myo-inositol) (PI(17:0/14:1)), 1-hexadecanoyl-2-octadecenoyl-sn-glycero-3-phosphoserine (PS(16:0/18:1)), 1,2-diheptadecanoyl-sn-glycero-3-phosphoserine (PS(17:0/17:0)), N-palmitoyl-D-sphingomyelin (SM(18:1/16:0)), N-heptadecanoyl-D-sphingomyelin (SM(18:1/17:0)), 1,2,3-octanoylglycerol (TG(8:0/8:0/8:0)), 1,2,3-tridecanoylglycerol (TG(10:0/10:0/10:0)), 1,2,3-tridodecanoylglycerol (TG(12:0/12:0/12:0)), 1,2,3-tritetradecanoylglycerol (TG(14:0/14:0/14:0)), 1,2,3-trihexadecanoylglycerol (TG(16:0/16:0/16:0)), 1,2-dipalmitoyl-3-oleoylglycerol (TG(16:0/16:0/18:1)), ,2,3-triheptadecanoylglycerol (TG(17:0/17:0/17:0)), 1,3-dioleoyl-2-palmitoylglycerol (TG(18:1/16:0/18:1)).

2. Sample preparation

2.1. Preparation of standards

Individual stocks for each compound were prepared at 2mg/mL following the recommendations of the suppliers. Working solutions for the elucidation of the fragmentation patterns of lipid standards were prepared at 5µg/mL in isopropanol/water (80:20). A mixed solution containing all the lipid standards was prepared in isopropanol at 30µg/mL each and subsequently diluted at the suitable final concentrations.

2.2. Lipid extraction from human serum samples

For lipid extraction, 50µL of human serum were mixed with 10µL of solvent or a mixture of lipid standards at 20µg/mL and 150µL of isopropanol. After vortexing, samples were left for 20min at -20°C and then centrifuged for 15min at 15000g and 4°C. Finally, 100µL of the supernatants were transferred to an HPLC vial for their LC-MS-based analysis.

3. LC-MS analysis

3.1. Instrumentation

LC-MS instruments used in this chapter were an Agilent 1290 Infinity LC system coupled to an Agilent 6550 Q-ToF mass spectrometer equipped with an ESI source (Agilent Technologies, Santa Clara, CA, USA), an Acquity Ultra Performance LC (UPLC) system coupled to a Synapt G2-Si Q-ToF mass spectrometer equipped with an ESI source (Waters, Milford, MA, USA), and a Q-orbitrap mass spectrometer (Q-Exactive, Thermo-Fisher Scientific) coupled to RP chromatography through an ESI source. Lipid fragmentation patterns were obtained using all three instruments, while comparison for untargeted LC-MS lipidomic analysis between MS-DIAL and LipidMS was performed using the Q-Exactive instrument.

3.2. Chromatographic separation

Lipids were separated on an Acquity UPLC CSH C18 column (100 x 2.1mm; 1.7 μ m) (Waters, Milford, MA, USA). The mobile phases consisted of (A) 10mM ammonium formate for ESI+ or ammonium acetate for ESI- in 60:40 (v/v) acetonitrile:water and (B) 10mM ammonium formate for ESI+ or ammonium acetate for ESI- in 90:10 (v/v) isopropanol:acetonitrile. The separation was conducted under the following gradient at a flow of 0.4mL/min (adapted from reference¹⁴⁴): 0min 20% (B); 0-2min 40% (B); 2-4min 43% (B); 4-4.1min 50% (B); 4.1-14min 54% (B); 14-14.1min 70% (B); 14.1-20min 99% (B); 20-24min 99% (B); 24-24.5min 20% (B); 24.5-27.5min 20% (B). Sample and column temperatures were maintained at 4°C and 65°C, respectively. The injection volume was 5 μ L.

3.3. MS detection

For the Agilent Q-ToF 6550, the following conditions were employed for ESI+ and ESI- ionization modes, respectively: the capillary voltage was 3.5kV for ESI+ and 4.0kV for ESI-; the nozzle voltage was 0.5kV for ESI+ and 1.5kV for ESI-; the gas temperature was 150°C for ESI+ and 275°C for ESI-; the drying gas (nitrogen) was 14L/min for ESI+ and 12L/min for ESI-; the nebulizer gas (nitrogen) was 35psi; the sheath gas temperature was 250°C for ESI+ and 350°C for ESI-; the sheath gas flow (nitrogen) was 11L/min for ESI+ and 12L/min for ESI-; and the fragmentor voltage was 200V for ESI+ and 150V for ESI-. Data was acquired in centroid mode using either full scan, DDA (using the Auto MS/MS mode), and DIA (using the all ions mode), and in both cases, using 0eV (full scan), 20eV and 40eV. For DDA mode, acquisition rate was set at 6 spectra/s in all cases.

Otherwise, for the Synapt G2-Si Q-ToF, the following conditions were employed for ESI+ and ESI- ionization modes, respectively: the capillary voltage was 3.0kV for ESI+ and 2.5kV for ESI-; the sampling cone was 40V; the source offset was 80V; the source temperature was 100°C; the desolvation temperature was 250°C; the cone gas flow was 50L/h; the desolvation gas flow was 600L/h; and the nebulizer gas was 6.5bar. Data was acquired in centroid mode using full scan, DDA and DIA (using MSe mode) with extended dynamic range, and using 0 and 40V as collision energies. Scan time was set at 0.3 seconds.

Finally, for the Q-Exactive instrument, the following conditions were employed for the ESI+ and ESI- ionization modes, respectively: the sheath gas flow rate was 25 (0-14 min) and 80 (14-27min) for ESI+ and 25 for ESI-; the auxiliary gas flow rate was 10 (0-14min) and 25 (14-27min) for ESI+ and 25 for ESI-; the spray voltage was 3kV for ESI+ and

2.5kV for ESI-; the capillary temperature was 215°C for ESI+ and 400°C for ESI-; the S-lens RF-level was 95 for ESI+ and 65 for ESI-; and the auxiliary gas heater temperature was 215°C for ESI+ and 350°C for ESI-. The pooled samples were acquired in full scan, DDA and DIA modes, while individual samples were acquired only in the MS scan mode. For MS scan acquisition purposes, resolution was set at 70000, the AGC target at 1000000, the maximum IT at 100ms, the scan range was 113-1700 and data type was centroid. For DDA acquisition purposes, the full scan parameters were as in the MS acquisition, while the MS² parameters were: resolution 70000, AGC target 1000000, maximum IT 200ms, loop count 5, MSX count 1, isolation window 0.4 *m/z*, isolation offset 0.4 *m/z*, collision energies 30V and 40V, data type centroid, minimum AGC target 1000 and dynamic exclusion 5sec. Finally, for DIA acquisition purposes, the full scan parameters were as in the MS acquisition, while the MS² parameters were: resolution 70000, AGC target 1000000, maximum IT 200ms, collision energies 30 and 40V, scan range from *m/z* 80 to *m/z* 1200 and data type centroid.

4. Data processing and analysis for untargeted LC-MS lipidomic analysis

For the performance evaluation of LipidMS, the lipidomic profile of additivated and non-additivated commercial serum (Sigma-Aldrich reference P2918) were analyzed and processed using three different workflows: i) LipidMS workflow; ii) MS-DIAL workflow¹³⁶; iii) data pre-processing using XCMS¹³⁴, isotope annotation using CAMERA¹⁴⁵. The parameters employed for these software were:

- LipidMS 3.0: all the samples were processed together (full scan, DDA and DIA) using the LipidMS R package. Files were previously converted into the mzXML format using the msConvert software (ProteoWizard 3.0.10800).
 - o Peak-picking parameters:
 - dmzagglom: 15
 - drtagglom: 200
 - drtclust: 25
 - minpeak: 5
 - drtgap: 5
 - drtminpeak: 10
 - drtmaxpeak: 200
 - recurs: 5 for MS¹ and 10 for MS².
 - sb: 3 for MS¹ and 2 for MS².
 - sn: 3 for MS¹ and 2 for MS².
 - minint: 1000 for MS¹ and 100 for MS².
 - weight: 2 for MS¹ and 3 for MS².
 - dmzIso: 5
 - drtIso: 5

- Batch processing parameters (alignment and grouping):
 - dmzalign: 10
 - drtalign: 100
 - span: 0.4
 - minsamplesfracalign: 0.75
 - dmzgroup: 10
 - drtagglomgroup: 50
 - drtgroup: 15
 - minsamplesfracgroup: 0.30
 - Lipid annotation parameters:
 - dmz for precursors: 5,
 - dmz for products: 10
 - rttol: 6,
 - coelCutoff: 0.6
- XCMS 3.16: all the samples were pre-processed together for MS level 1 (full scan, DDA and DIA) by the XCMS R package. Files were previously converted into the mzXML format using the msConvert software (ProteoWizard 3.0.10800). Several values for the bandwidth and binSize parameters were tested to optimize the extraction of isomeric peaks.
- Peak-picking:
 - peakwidth: between 5 and 30 seconds
 - noise: 1000
 - ppm: 15
 - snthres: 3
 - prefilter: 5 scans with a minimum intensity of 1000

- Alignment and grouping:
 - Alignment method: based on the peak groups.
 - Grouping:
 - minFraction: 0.3
 - bw = 2
 - binSize = 0.005
- Isotope annotation using CAMERA 3.15:
 - perfwHM = 0.6
 - cor_eic_th = 0.75
 - maxcharge = 3
 - ppm = 5
 - mzabs = 0.01
 - filter (C12/C13) = TRUE
- MS-DIAL 4.80: the full scan and DDA acquired samples were processed together, while the DIA files were analyzed in a different batch. Files were previously converted into the abf format using Reifycs Abf (Analysis Base File) Converter 4.0. Then the DIA identifications were added to the feature matrix obtained for the full scan and DDA files using an m/z tolerance of 0.005 and an RT tolerance of 10 seconds.
 - Data collection:
 - MS¹ tolerance: 0.005
 - MS² tolerance: 0.01
 - Peak detection:
 - Minimum peak height: 1000
 - Mass slice width: 0.1 Da
 - Smoothing method: Linear-weighted moving average

- Smoothing level 5 scan
- Minimum peak width: 5 scan
- Adducts:
 - ESI-: M-H, M-H-H₂O, M+Na-2H, M+Hac-H, M+FA-H, 2M-H and M-2H.
 - ESI+: M+H, M+NH₄, M+Na, M+H-H₂O, M+H-2H₂O, 2M+NH₄ and 2M+Na.

For all the three workflows, features were filtered and normalized based on QC samples. Only the features present in at least 70% of the QC samples were kept. Then data were normalized using a LOESS function, which was fitted to the QC samples based on the injection order. The resulting interpolated curve for each feature was used to normalize its response¹⁶. Finally, a differential analysis between the additivated and non-additivated serum samples was performed using a Student's t-test, corrected for multiple testing and magnitude of change. The features with an adjusted *p-value* < 0.05 and a fold change > 1.5 were considered to be differential variables between both groups.

Results and Discussion

1. LipidMS overview

As mentioned above, the size, complexity and heterogeneity of the lipidome and the lack of available lipid standards makes lipid annotation hard and time-consuming. Additionally, most of the current MS/MS lipid annotation tools are restricted to DDA. To complement the information provided by DDA, LipidMS was initially conceived to annotate lipids in single samples using DIA and rule-based annotation, but it required the use of external data processing tools for conducting batch data processing¹⁴⁶. To overcome this limitation, new releases of LipidMS package have incorporated the needed functionalities to cover the whole data processing workflow¹⁴⁷. Furthermore, and since most of the LC-MS-based lipidomic studies are still conducted using DDA and that both approaches provide complementary information, we decided to implement the analysis of DDA within LipidMS. The last version of LipidMS's workflow is depicted in Figure 15. Briefly, raw data files in mzXML format and a csv metadata file (sample, acquisition mode, which can be full scan, DDA or DIA, and sample type) are used as input. Data processing, including peak-peaking, alignment, grouping and peak filling, is executed based on the MS¹ level information from all the samples to obtain a feature matrix that contains the peak intensities. Then, lipids are annotated based on the established fragmentation rules for those samples acquired in DIA or DDA using both MS¹ and MS² levels of information, and the identifications are incorporated into the feature matrix generated in step 2. Finally, two main outputs can be obtained: a data matrix containing the peak areas and lipid annotations, if obtained, for all the features and samples found in the dataset, and plots showing the fragments that support the proposed lipid identifications. The details of all these steps are described in the following sections. Furthermore, LipidMS also supports the simultaneous processing of all the following combinations of MS acquisitions modes: all the samples

in DIA; all the samples in DDA; combination of DIA and DDA samples; combination of full scan and DIA; combination of full scan and DDA; and combination of full scan, DIA and DDA.

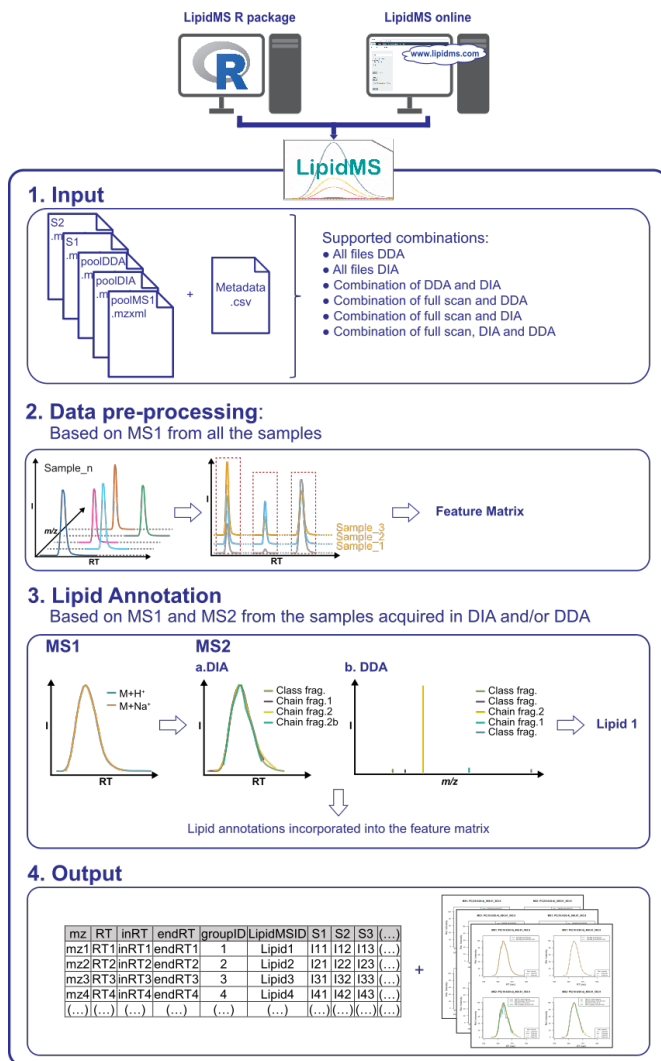


Figure 15. LipidMS v3.0 overview. Briefly, LipidMS uses raw data files in mzXML format and a metadata csv file as input. Then, raw data are pre-processed to extract all features for each sample and build a feature matrix that contains the peak area for all features and samples. Finally, lipids are annotated based on the MS¹ and MS² information of DIA and DDA acquired samples. Several tables and graphical outputs can be obtained.

2. Features and implementation

2.1. Data processing

LipidMS v3.0 covers the whole workflow required to process untargeted LC-MS lipidomics datasets, including peak-picking, alignment, grouping and filling missing peaks, before performing the actual lipid annotation.

2.1.1. Peak-picking

The first step of the LipidMS workflow, which is executed using the *dataProcessing* function, extracts all the peaks from each sample in the dataset. Peak-picking is performed for MS¹ in all the samples, and MS² in those files acquired in the DIA mode. This function is based on the *enviPick* algorithm³⁴, which has been implemented into LipidMS. Briefly, *enviPick* uses a clustering-based algorithm, which extracts peaks in three steps:

- 1) **Partitioning.** First, in order to accelerate the following steps, data is divided into multiple partitions or bins based on large user-defined tolerances for m/z and RT so that data points from different partitions do not overlap. To this end, data is ordered by increasing m/z and RT and each point is initialized as a partition. Then, each partition is evaluated to decide whether it can be joined to the previous partition or not. If the m/z and RT of a partition match the tolerances of any point in the previous partition, it is reassigned.
- 2) **Clustering.** For each partition, EIC are extracted based on smaller user-defined tolerances for m/z and RT. Data is ordered by intensity and first point is initialized as a cluster. Then, for each point, if it is assignable to any cluster based on the predefined

tolerances and no points with identical RT are already in the cluster, it is assigned to the cluster with the closest m/z mean. Otherwise, a new cluster is initialized.

- 3) **Peak-picking.** Finally, peaks are searched within each EIC. To this end, data is ordered by RT and the most intense point is taken as the first peak apex candidate. From this point, lower and upper RT bounds are searched. Difference between the cumulative sum of intensity increases and decreases between the peak apex and its neighbour points towards lower and higher RT are calculated. Then, bounds are set where maximum differences between increases and decreases are reached. This process is repeated n times based on user-defined parameters. Finally, peaks are filtered based on signal-to-noise and signal-to-baseline ratios and intensity threshold. Peak areas are estimated by the sum of peak intensities within the RT bounds and baseline correction is performed.

At this point, ^{13}C isotopologues are also annotated based on CAMERA algorithm¹⁴⁵. The following criteria must be met by a peak to be considered a ^{13}C isotopologue: i) mass difference of 1.0033 between the ^{12}C and the ^{13}C isotopologues; ii) relative intensity between isotopologues, consistently with the known natural abundances of ^{12}C and ^{13}C isotopes; iii) co-elution, calculated using Pearson correlation based on peak shape¹⁴⁸:

$$P_{P1,P2} = \frac{\sum_{i=1}^n (I_{P1i} - \widetilde{I}_{P1})(I_{P2i} - \widetilde{I}_{P2})}{\sqrt{\sum_{i=1}^n (I_{P1i} - \widetilde{I}_{P1})^2} \sqrt{\sum_{i=1}^n (I_{P2i} - \widetilde{I}_{P2})^2}} \quad (\text{Equation 1})$$

, where P1 refers to the peak of the M+0 isotopologue and P2 refers to a heavier isotopologue, I_{P1i} or I_{P2i} refer to the intensity of each scan of the aligned and smoothed peak, and \widetilde{I}_{P1} and \widetilde{I}_{P2} refer to the sum of the

intensity of all scans of each peak. Before calculating the peak coelution score, peaks are smoothed using the *smooth.spline* function from the R core package.

The output of the *dataProcessing* function is an *mobject* containing a peaklist for MS¹ and another for MS² in case of DIA acquisition, and raw data for MS¹ and MS² when available. In addition, it contains all metadata required to perform the subsequent steps such as polarity, MS level for each scan or precursors for MS² scans in DDA, among others. From LipidMS v3.0, the *batchdataProcessing* function can also be used, which returns an *mobject* for each sample and wraps all of them into an *msbatch*, which will be subsequently used for the following data processing steps.

2.1.2. Peak alignment

Once all the peaks for each sample have been extracted, time drifts during the acquisition queue need to be corrected. The *alignmsbatch* function performs peak alignment based on the MS¹ information obtained for all the samples. First, peak partitions are created based on the *enviPick* algorithm³⁴ described above to speed up the following clustering algorithm. Then, the clustering algorithm (Figure 16) is executed to group peaks based on their RT for each partition as follows:

- 1) Each peak in the partition is initialized as a new cluster. For each cluster, the minimum, maximum and mean values of the RT, which, at this point have the same values, are kept.
- 2) Calculate a distance matrix between all the clusters. This distance will be the greatest difference between the minimum and maximum values of each cluster. Distances between the clusters containing peaks from the same samples will be set at not

available (NA) (i.e., if a sample has two peaks in two different clusters, these two clusters cannot be merged).

- 3) If any distance differs from NA, search the minimum distance between two clusters.
- 4) If distance is below the maximum distance allowed, join clusters and update the minimum, maximum and mean values. Otherwise, set the distance at NA and go back to point 3.

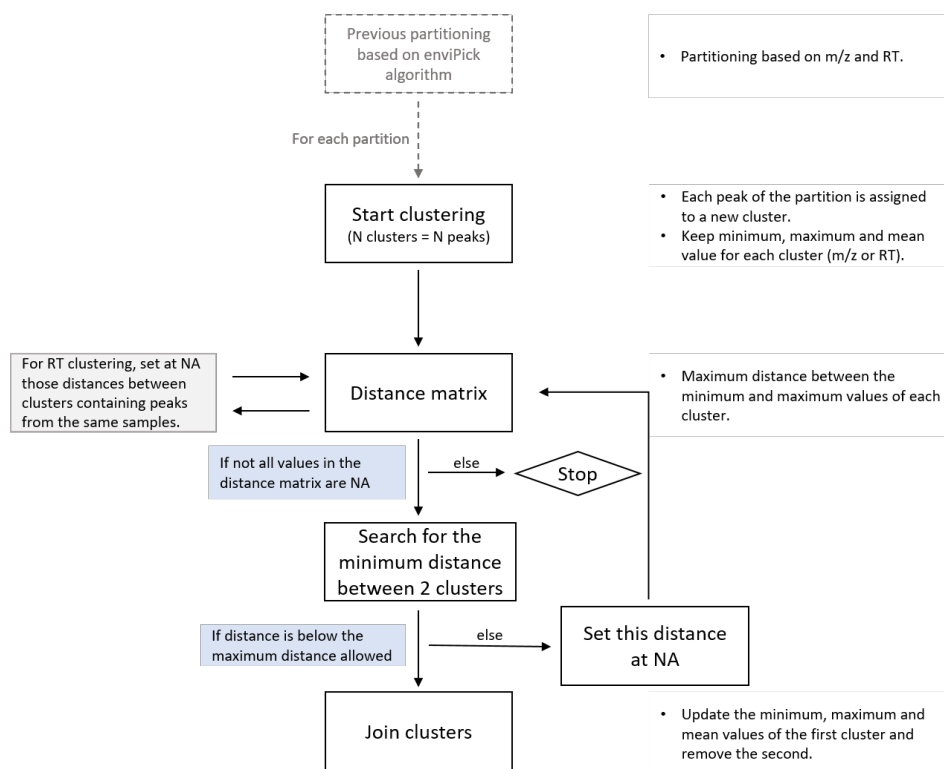


Figure 16. Clustering algorithm used for peak alignment and grouping in LipidMS.

Then, clusters with a sample representation over a defined minimum will be used for alignment. To this end, a matrix that contains the RT of the peaks for each sample from the selected clusters is built. The median RT is calculated for each cluster and an RT deviation matrix is

obtained. Finally, time drifts for each sample are corrected using LOESS regression by constructing a function based on RT deviation and the median. This same function is used to correct the time drifts of the MS² level for those samples acquired in DIA or DDA mode.

2.1.3. Peak grouping

Once alignment has been performed, peaks from the different samples that belong to the same feature are grouped using the *groupmsbatch* function (Figure 17). To this end, the same algorithms as those employed for alignment are applied in the following order: peak partitions are created based on the *m/z* and RT values using the *enviPick* algorithm³⁴; *m/z* clustering is applied to each partition as described previously for RT; then, peaks are grouped by RT using the same clustering algorithm; and finally, clusters with a sample representation over the defined minimum are selected to build the feature table. An example of sequential partitioning and clustering executed for the alignment and grouping steps is summarized in Figure 17.

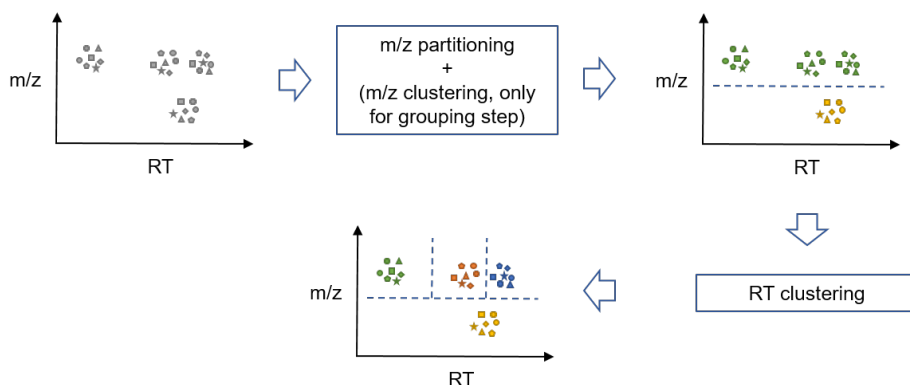


Figure 17. Scheme of sequential partitioning and clustering of peaks executed during alignment and grouping steps. Each point represents a peak, each shape denotes a sample and each colour depicts a cluster.

2.1.4. Peak filling

Once all feature peaks have been defined, areas are extracted again for each peak and sample based on the peak parameters defined for each feature (m/z tolerance, initial and final RT) using the *fillpeaksmsbatch* function. This step avoids missing peaks that were not found in the peak-picking step and improves the area estimation by homogenizing peak bounds for all samples.

2.2. Lipid annotation

Lipid annotation based on the DIA or DDA acquired samples is performed with the *annotatemsbatch* function to search for lipids in the *msbatch* based on a set of predefined fragmentation rules, which are detailed below.

2.2.1. Rationale behind LipidMS annotation

The building block nature of most lipid species enables the establishment of generic structure-derived fragmentation rules that can be used for MS-based identification and structure elucidation. This strategy has been satisfactorily implemented for lipid identification in both DDA and DIA approaches⁶⁷⁻⁶⁹. However, to accomplish lipid identification these methods commonly relied on the use of most intense fragments, which can generate false positives due to the poor selectivity of these ions when coelution is present. In RP, among other factors, lipids elution depends both on the lipid class and their FA composition, thus each lipid class usually elutes within a narrow RT window, eluting first those with shorter FA chains and more unsaturations. As a result, many common fragments, as those corresponding to head groups, are poorly chromatographically

resolved as represented in Figure 18, which can strongly affect their selectivity for lipid annotation. This issue becomes particularly relevant when complex biological samples are analyzed. To overcome these drawbacks, lipid annotation in LipidMS is based on combining two complementary approaches. First, for DIA data, a parent-fragment coelution score (PFCS) is calculated in a predefined RT window around the parent ion RT to modulate the stringency in the association coeluting ions (adducts in MS¹ or fragments in MS²). The PFCS score is formally defined as a Pearson correlation coefficient calculated based on the peak shape (distribution of intensities over elution time) of two peaks and it tests the similarity of the ion chromatograms between them. For each fragment ion or adduct coeluting with the parent ion in the predefined RT window, a PFCS is calculated based on Equation 1, where P1 refers to the parent/precursor peak (MS¹) and P2 refers to an adduct (MS¹) or fragment (MS²) peak. This procedure has been successfully applied in metabolomics¹⁴⁸. In case of DDA data, only the closest MS² scan to the RT of the MS¹ peak is selected for annotation, which improves differentiation between coeluting isomeric lipid species. Second, and most importantly, LipidMS takes advantage of the use of fragmentation and intensity rules. The last are defined based on the relative intensity of different fragment ions and are used to elucidate the position of the different FA into the lipid backbone structure. Both fragmentation and intensities rules have been manually curated by using publicly available spectral information (i.e., LipidMaps²⁶, METLIN²⁵, LipidBlast⁶⁷, HMDB²²) and in-house generated MS/MS spectra for DDA and DIA in three different MS/MS platforms (Thermo Q-Exactive, Waters Synapt G2-Si Q-ToF and Agilent Q-ToF 6550). In the fragmentation rules curation procedure, the use of highly intense fragments common to several lipid classes has been avoided when possible and specific well-characterized fragments and adducts have been selected instead. Specific adducts selected fragments as well as the preferred

acquisition mode (i.e., ESI+ and ESI-) for each lipid class are summarized in Tables 2-3. Additionally, the experimental data supporting the selection of the fragmentation rules used by LipidMS are represented in Additional Figures S1-S28 (Appendix 2).

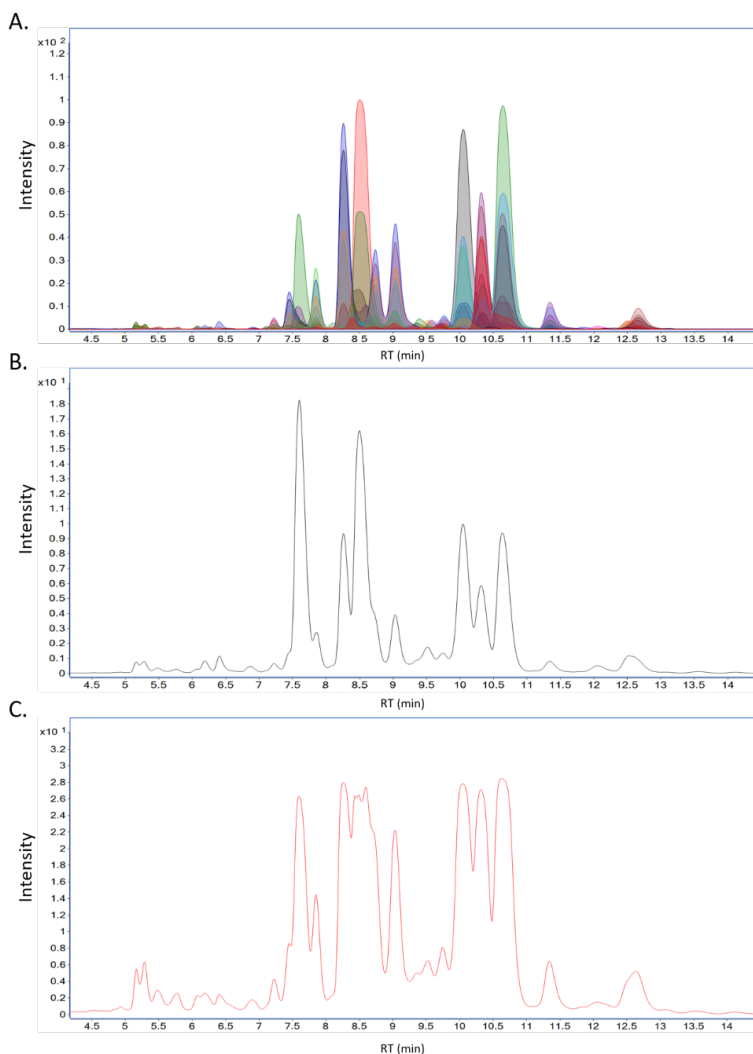


Figure 18. Coelution profile of common fragment 184.074 (phosphocholine) of PC and SM in a LC-MS analysis. A) Chromatographic peaks of all PC and SM detected in a sample using ESI+. B) Chromatographic profile of fragment 184.074 at collision energy 20eV. C) Chromatographic profile of fragment 184.074 at collision energy 40eV.

Table 2. Preferred adducts and fragmentation rules set by default to annotate lipids in ESI+. (*) In case of TG, DG, snX refers to the fragment ion resulting from the loss of the FA chain of the snX position, which is a DG.

| Class | Adducts | Class fragments | Chain fragments | Intensity rules |
|-------------------|---|--|---|--|
| Carnitines | M+H, M+Na | 60.0807, 85.0295 FA as M+H-H ₂ O | FA as M+H-H ₂ O | - |
| CE | 2M+NH ₄ , M+NH ₄ , 2M+Na, M+Na | 369.3516 FA as M+H-H ₂ O | FA as M+H-H ₂ O | - |
| LPC | M+H, M+Na | 104.1075, 184.0739 | MG as M+H-H ₂ O | - |
| LPE | M+H, M+Na | NL of 141.01909 | MG as M+H-H ₂ O | - |
| PC | M+H, M+Na | 104.1075, 184.0739 NL of 183.06604 | Sn1: LPC as M+H or M+H-H ₂ O Sn2: LPC as M+H or M+H-H ₂ O or Precursor - sn1 | LPC sn1 > 2 * LPC sn2 |
| PE | M+H, M+Na | DG as M+H-H ₂ O | Sn1: LPE or MG as M+H-H ₂ O Sn2: FA or MG as M+H-H ₂ O | LPE sn1 > 3 * LPE sn2 MG sn2 > 2 * MG sn1 |
| oPC | M+H, M+Na | 104.1075, 184.0739 NL of 183.06604 | Sn1: oLPC as M+H or M+H-H ₂ O Sn2: LPC as M+H or M+H-H ₂ O or Precursor - sn1 | oLPC sn1 > 2 * LPC sn2 |
| pPC | M+H, M+Na | 104.1075, 184.0739 NL of 183.06604 | Sn1: pLPC as M+H or M+H-H ₂ O Sn2: LPC as M+H or M+H-H ₂ O or Precursor - sn1 | pLPC sn1 > 2 * LPC sn2 |
| oPE | M+H, M+Na | NL of 140.012 | Sn1: oLPE as M+H or M+H-H ₂ O Sn2: MG as M+H-H ₂ O | oLPE sn1 > 2 * MG sn2 |
| pPE | M+H, M+Na | NL of 140.012 | Sn1: pLPE as M+H or M+H-H ₂ O Sn2: MG as M+H-H ₂ O | pLPE sn1 > 2 * MG sn2 |

| Class | Adducts | Class fragments | Chain fragments | Intensity rules |
|---------|---|--|---|-----------------------------|
| DG | M+H-H ₂ O, M+NH ₄ , M+Na | - | Sn1: MG as M+H-H ₂ O Sn2: MG as M+H-H ₂ O | MG sn1 > MG sn2 |
| TG | M+NH ₄ , M+Na | - | Sn1: Precursor - DG as M+H-H ₂ O Sn2: Precursor - DG as M+H-H ₂ O Sn3: Precursor - DG as M+H-H ₂ O | (*)DG sn2 > DG sn1 > DG sn3 |
| Sph | M+H | - | Sph as M+H-H ₂ O or M+H-2H ₂ O | - |
| SphP | M+H | - | Sph as M+H-H ₂ O, M+H-2H ₂ O or M+H-H ₂ O-NH ₄ | - |
| AcylCer | M+H, M+H-H ₂ O, M+Na | - | NL of acyl chain (Cer as M+H, M+H-H ₂ O or M+H-2H ₂ O) Sph as M+H-H ₂ O or M+H-2H ₂ O FA as M+H | - |
| Cer | M+H-H ₂ O, M+H, M+Na | - | Sph as M+H-2H ₂ O Precursor - Sph | - |
| CerP | M+H | NL of phosphate group (Cer as M+H-H ₂ O or M+H-2H ₂ O) | Sph as M+H-2H ₂ O Precursor - Sph | - |
| SM | M+H, M+Na | 104.1075, 184.0739 NL of 183.06604 | Sph as M+H-2H ₂ O Precursor - Sph | - |

Table 3. Preferred adducts and fragmentation rules set by default to annotate lipids in ESI⁻.

| Class | Adducts | Class fragments | Chain fragments | Intensity rules |
|-------|---|---|--|--|
| FA | M-H, 2M-H | FA as M-H or M-H-H ₂ O | - | - |
| FAHFA | M-H | - | HFA as M-H FA as M-H | - |
| LPC | M+CH ₃ COO, M-CH ₃ , M+CH ₃ COO-CH ₃ | 168.0426, 224.0688, LPA as M-H or LPC as M-CH ₃ | FA as M-H | - |
| LPE | M-H | 140.0115, 196.038, 214.048, No presence of NL of CH ₃ | FA as M-H | - |
| LPG | M-H | 152.9958, 209.022, 227.0326, NL of 74.0359 | FA as M-H | - |
| LPI | M-H | 223.0008, 241.0115, 259.0219, 297.0375 | FA as M-H | - |
| LPS | M-H, M+Na-2H | NL of 87.032 | FA as M-H | - |
| PC | M+CH ₃ COO, M-CH ₃ , M+CH ₃ COO-CH ₃ | NL of CH ₃ , 168.0426, 224.0688 | Sn1: LPC as M-CH ₃ Sn2: LPC as M-CH ₃ or FA as M-H | LPC sn1 > 3 * LPC sn2 |
| PE | M-H | 140.0115, 196.038, 214.048, No presence of NL of CH ₃ | Sn1: LPE as M-H Sn2: LPE as M-H or FA as M-H | LPE sn1 > 3 * LPE sn2 |
| PG | M-H | 152.9958, 209.022, 227.0326, NL of 74.0359 | Sn1: LPG as M-H Sn2: LPG as M-H or FA as M-H | LPG sn1 > 3 * LPG sn2 |
| PI | M-H | 223.0008, 241.0115, 259.0219, 297.0375 | Sn1: LPI or LPA as M-H Sn2: LPI as M-H or FA as M-H | LPI sn1 > 3 * LPI sn2 LPA sn1 > 3 * LPA sn2 |
| PS | M-H, M+Na-2H | NL of 87.032 | Sn1: LPA as M-H or M-H-H ₂ O Sn2: LPA as M-H or M-H-H ₂ O or FA as M-H | LPA sn1 > 3 * LPA sn2 |

| Class | Adducts | Class fragments | Chain fragments | Intensity rules |
|----------------|---|--|--|-------------------------------------|
| oPC | M+CH ₃ COO, M-CH ₃ , M+CH ₃ COO-CH ₃ | NL of CH ₃ , 168.0426, 224.0688 | Sn1: oLPC as M-CH ₃ or M-CH ₃ -H ₂ O Sn2: FA as M-H or M-CO ₂ -H | FA sn2 > 3 * oLPC sn1 |
| pPC | M+CH ₃ COO, M-CH ₃ , M+CH ₃ COO-CH ₃ | NL of CH ₃ , 168.0426, 224.0688 | Sn1: pLPC as M-CH ₃ or M-CH ₃ -H ₂ O Sn2: FA as M-H or M-CO ₂ -H | FA sn2 > 3 * pLPC sn1 |
| oPE | M-H, M+NaCH ₃ COO | 140.0115, 196.038, 214.048 | Sn1: oLPE as M-H or M-H-H ₂ O Sn2: FA as M-H | FA sn2 > 3 * oLPE sn1 |
| pPE | M-H, M+NaCH ₃ COO | 140.0115, 196.038, 214.048 | Sn1: pLPE as M-H or M-H-H ₂ O Sn2: FA as M-H | FA sn2 > 3 * pLPE sn1 |
| CL | M-H, M-2H | 78.9585, 152.9958 | Sn1: LPA as M-H-H ₂ O Sn2: LPA as M-H-H ₂ O Sn3: LPA as M-H-H ₂ O Sn4: LPA as M-H-H ₂ O | - |
| Sph | M-H | - | Sph as M-H-H ₂ O or M-H-2H ₂ O | - |
| SphP | M-H | 78.9585, 96.9691 | Sph as M-H-H ₂ O | - |
| AcylCer | M+CH ₃ COO, M-H | - | | Acyl chain > 5 * NL of Sph > 2 * FA |
| Cer | M+CH ₃ COO, M-H | - | Sph as M-H-H ₂ O or NL of Sph (partial) FA as M-H | - |
| CerP | M-H | | | - |
| BA | M-H | BA as M-H-H ₂ O Conjugate fragment | - | - |

2.2.1. Lipid coverage and building block database customization

As previously mentioned, most of the lipids can be defined by a backbone structure, which defines the lipid class and subclass, and a number of acyl residues attached to that core structure (Figure 19). Thanks to this feature, a lipid database can be built by combining both the lipid core and the set of acyl chains to be incorporated¹³¹. By default, the building block database (bbDB) of LipidMS comprises a set of 30 fatty acyl chains and 4 sphingoid bases (Table 4), which were selected based on their biological relevance⁷³.

These chains are combined with the core structures of 29 lipid classes (Tables 2-3) to build a query database (qDB) that will be eventually used to interrogate the MS¹ data. This qDB contains, by default, 3726 unique putative lipids regardless of the composition of each fatty acyl chain, even though the bbDB can be customized to add, for example, additional odd fatty acyl chains as FA(19:0).

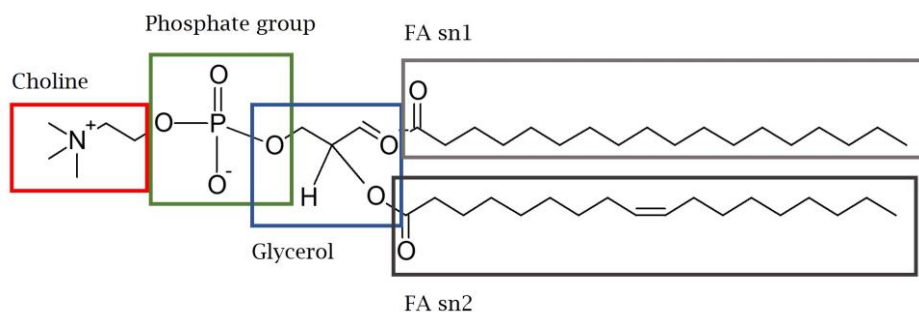


Figure 19. Example of building block structure of a PC. Glycerophosphocholine is the core structure to which two FA chains are attached, FA(16:0) at the sn1 position and FA(18:1) at the sn2 position.

Table 4. Building blocks included by default in the bbDB to build the qDB.

| FA chains | | | Sphingoid bases |
|-----------|----------|----------|-----------------|
| FA(8:0) | FA(18:1) | FA(22:0) | Sph(16:0) |
| FA(10:0) | FA(18:2) | FA(22:1) | Sph(16:1) |
| FA(12:0) | FA(18:3) | FA(22:2) | Sph(18:0) |
| FA(14:0) | FA(18:4) | FA(22:3) | Sph(18:1) |
| FA(14:1) | FA(20:0) | FA(22:4) | |
| FA(15:0) | FA(20:1) | FA(22:5) | |
| FA(16:0) | FA(20:2) | FA(22:6) | |
| FA(16:1) | FA(20:3) | FA(24:0) | |
| FA(17:0) | FA(20:4) | FA(24:1) | |
| FA(18:0) | FA(20:5) | FA(26:0) | |

2.2.2. LipidMS annotation workflow

LipidMS v3.0 contains 45 individual annotation functions (e.g., *idPCpos*, *idPCneg*) wrapped into two general functions aimed to annotate *msobjects* for ESI+ or ESI- (*idPOS* and *idNEG*). In addition, *annotatemsbatch* function can be used to automatically annotate all the DIA and DDA *msobjects* contained in the *msbatch* and to dump the results into the dataset feature table. To exemplify the LipidMS annotation workflow, the identification procedure for a PG(16:0/18:1) is described in Figure 20.

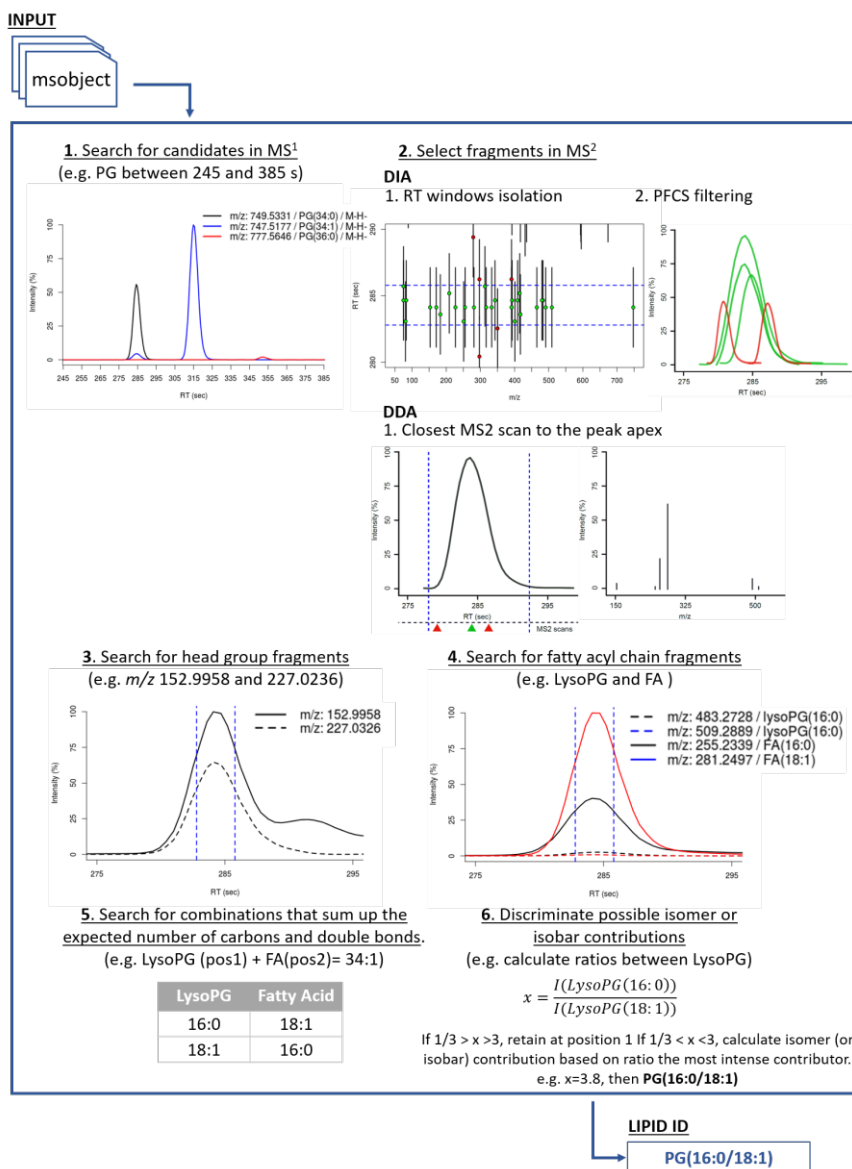


Figure 20. Lipid annotation workflow in LipidMS. The steps for the identification of m/z . 747.5177 in a RT of 285 seconds are shown as an example: 1) Search for PG candidates as [M-H] based on MS¹ information; 2) isolation of MS² fragments based on RT windows and PFCS score for DIA data or from the corresponding MS² scan in case of DDA data; 3) identification of class specific fragments; 4) identification of chain specific fragments; 5) selection of combinations of FA chains that sum up the total carbon and double bond composition; and 6) confirmation of FA chains position based on relative intensities between the chain fragments.

Overall, the following steps are executed internally within each identification function survey for lipid annotation (i.e., *idPGneg*):

- 1) On the basis of the set of chemical entities included in the bbDB (Table 4) and on the ionization properties selected for each lipid class (Tables 2-3), a target ion list is used to interrogate the MS¹ data within defined *m/z* and RT tolerances (*findCandidates*). These parameters can be easily set up by the user. At this step, putatively annotated lipids are identified based on the lipid class, and the number of carbons and double bonds is determined. This level of survey is not reported by LipidMS by default, as we considered it as non-informative. However, this information can be easily recovered by the *findCandidates* function, or the class identification functions (e.g., *idPGneg*). In this step, only those features confirmed as M+0 in the MS¹ level (those for which at least an M+1 isotopologue has been detected) are used for annotation to reduce false-positive annotations.
- 2) Then, fragment ions (MS²) related to the selected parent ions (MS¹) need to be isolated. In case of DIA data, coeluting fragment ions for each putatively annotated lipid are selected based on the defined RT window. Optionally, a PFCS is then calculated for each of the pair precursor-fragment ions and only those fragments above a previously defined threshold are retained. To minimize false positives, a value of 5 seconds for the RT window and a PFCS value of 0.8 are set by default (*coelutingFrag*s). However, these values can be easily changed by the user. In the case of DDA data, where a direct link between parent and fragment ions is available, the algorithm searches for the MS² scans that fall within the limits of the MS¹ peak and for which the precursor of interest has been selected. If multiple MS² scans meet

the requirements, only the closest to the RT of the MS¹ peak is selected for annotation, which improves differentiation between isomeric lipid species (*ddaFragments*).

- 3) On the basis of the established fragmentation rules (Tables 2-3) and the defined *m/z* tolerance (10 ppm by default for product ions), a survey for informative fragment ions of the lipid class (e.g., head groups) is performed among those ions extracted in step 2 (*checkClass*).
- 4) Then, the same procedure is applied for searching informative fragments of the fatty acyl components (*chainFragments*).
- 5) Based on the proposed fatty acyl components, combinations that sum up the expected total number of carbons and double bonds determined in step 1 are searched (*combineChains*).
- 6) Once the fatty acyl components have been determined, intensity rules, which are based on the relative intensities ratio between the fragments, are applied to elucidate the position of those chains (*checkIntensityRules*). For further details regarding intensity rules, see Tables 2-3.

Depending on the structural evidence reached for annotation, lipids are identified with four different confidence levels: i) “MS-only”, when no clear fragmentation pattern is known (this level is available only for MG and FA as they do not have defined fragmentation patterns); ii) “subclass level”, when specific subclass fragments are found, but only the total number of carbons and double bonds of the chains can be proposed based on the precursor ion. At this level, LipidMS cannot differentiate which fatty acids are linked to the backbone and a sum of several isobaric/isomeric compounds is proposed (e.g. PG(34:1)); iii) “fatty acyl level”, when specific chain fragments inform about the composition of fatty acyl chains, but no positional information can be provided (e.g., PG(16:0_18:1)); iv) “fatty acyl position level”, when the specific chain

position can be elucidated based on the chain fragments' intensity ratios (e.g., PG(16:0/18:1)).

As a result of the execution of lipid identification functions (*idPOS* or *idNEG*), two separate items, which can be easily saved as tables, are generated (i.e., 'results peak table' and 'annotated peak table'). On the one hand, the 'results peak table' (Figure 21A) contains the following information for each annotated lipid: i) feature identity, annotated as lipid class, total number of carbons, double bonds and fatty acid composition, ii) peak properties, including *m/z*, RT, peak intensity and peakID information and iii) identification criteria used, reporting information about the detected adduct/s, *m/z* error, structural annotation level, and score (mean value of PFCS of all fragments used for annotation in DIA or sum of the relative intensity of the fragments in DDA). On the other hand, the 'annotated peak table' (Figure 21B) links the original MS¹ peak table with the 'results peak table', providing the following information for each feature: *m/z*, RT, peak intensity, peakID, all the possible identities ranked by the annotation level, ion adducts and the mean value of the PFCS used in each lipid identification. Further information about the fragments that support each identification can be explored using class-specific identification functions (i.e., *idPGneg*).

In case of batch processing, after all DIA/DDA *msubjects* have been annotated individually, the whole set of potential identities are automatically dumped into the dataset feature matrix (*joinAnnotationResults* function) (Figure 21C). If different annotation levels are obtained for a given lipid, e.g. feature X is identified as PC(34:1) based on DIA and PC(16:1/18:0) based on DDA, the identification with the highest degree of structural information (i.e., PC(16:1/18:0)) is maintained. If several identifications with the same annotation level are obtained (e.g., PC(18:1/16:0), PC(16:1/18:0)), they are all maintained.

A. Results table of each *msobject*.

| ID | Class | CDB | FComp | mz | RT | int | Adducts | ppm | confidenceLevel | peakID | Score |
|--------------------|-------|------|----------------|----------|---------|-------------|---|-------|-----------------|------------|-------|
| MG(16:0) | MG | 16.0 | 16.0 | 313.2735 | 113.01 | 5515012 | M+H ₂ O | 2.093 | MS-only | MS1_0_627 | 0 |
| MG(18:1) | MG | 18.1 | 18.1 | 339.2892 | 914.88 | 7950690 | M+H ₂ O | 1.754 | MS-only | MS1_0_732 | 0 |
| MG(20:4) | MG | 20.4 | 20.4 | 396.3111 | 67.53 | 795318 | M+NH ₄ | 0.698 | MS-only | MS1_0_996 | 0 |
| LPC(16:0) | LPC | 16.0 | 16.0 | 496.3398 | 111.20 | 20668285805 | M+H ₂ M+Na | 1.025 | FA | MS1_0_1507 | 0.988 |
| LPC(18:0) | LPC | 18.0 | 18.0 | 524.3712 | 184.67 | 10888241030 | M+H ₂ M+Na | 0.774 | FA | MS1_0_1742 | 0.99 |
| LPC(18:1) | LPC | 18.1 | 18.1 | 522.3555 | 124.53 | 4171665348 | M+H ₂ M+Na | 0.948 | FA | MS1_0_1720 | 0.986 |
| LPC(20:4) | LPC | 20.4 | 20.4 | 544.3397 | 82.71 | 81919226 | M+H ₂ M+Na | 1.222 | Subclass | MS1_0_1929 | 0.907 |
| PC(16:0/14:0) | PC | 30.0 | 16.0/14.0 | 706.5383 | 492.62 | 475061084 | M+H | 0.558 | FA position | MS1_0_3967 | 0.991 |
| PC(16:0/18:1) | PC | 34.1 | 16.0/18.1 | 760.5846 | 607.59 | 39091671838 | M+H | 1.294 | FA position | MS1_0_5123 | 0.992 |
| PC(16:0/18:2) | PC | 34.2 | 16.0/18.2 | 758.5693 | 527.50 | 57925316042 | M+H ₂ M+Na | 0.842 | FA position | MS1_0_5077 | 0.982 |
| PC(16:0/18:3) | PC | 34.3 | 16.0/18.3 | 756.5534 | 353.25 | 319091491 | M+H | 1.262 | FA position | MS1_0_5030 | 0.912 |
| PC(18:0/18:1) | PC | 36.1 | 18.0/18.1 | 788.6157 | 732.50 | 6729390612 | M+H | 1.543 | FA position | MS1_0_5808 | 0.973 |
| PC(18:0/18:2) | PC | 36.2 | 18.0/18.2 | 786.6007 | 639.58 | 46500125550 | M+H | 0.766 | FA position | MS1_0_5807 | 0.979 |
| PCp(16:0/20:4) | PCp | 36.4 | 16.0/20.4 | 788.5565 | 552.16 | 50712514 | M+Na+M+H | 0.014 | FA position | MS1_0_5862 | 0.902 |
| PE(36:1) | PE | 36.1 | 36.1 | 746.5669 | 787.84 | 46883859 | M+H | 4.186 | Subclass | MS1_0_4796 | 0.982 |
| PE(38:3) | PE | 38.3 | 38.3 | 770.5692 | 709.32 | 17650256 | M+H | 1.046 | Subclass | MS1_0_5409 | 0.987 |
| PI(36:1) | PI | 36.1 | 36.1 | 882.6056 | 561.78 | 77354410 | M+NH ₄ | 1.745 | Subclass | MS1_0_8536 | 0.989 |
| PI(6:2) | PI | 36.2 | 36.2 | 880.5904 | 492.02 | 291587237 | M+NH ₄ M+H ₂ M+Na | 1.295 | Subclass | MS1_0_8480 | 0.981 |
| Cer(d18:0/22:0) | Cer | 40.0 | 18.0/22.0 | 606.6181 | 964.39 | 13956746 | M+H ₂ O | 1.144 | FA position | MS1_0_2533 | 0.871 |
| Cer(d18:0/24:0) | Cer | 42.0 | 18.0/24.0 | 634.6488 | 1000.21 | 17297148 | M+H ₂ O | 1.986 | FA position | MS1_0_2870 | 0.967 |
| SM(d18:1/16:0) | SM | 34.1 | 18.1/16.0 | 725.5560 | 476.95 | 388794322 | M+Na+M+H | 1.045 | FA position | MS1_0_4372 | 0.934 |
| TG(16:0/16:0/18:1) | TG | 50.1 | 16.0/16.0/18.1 | 850.7853 | 1140.92 | 6097968495 | M+NH ₄ M+Na | 1.133 | FA position | MS1_0_7678 | 0.982 |

B. Annotated peak table of each *msobject*.

| mz | RT | int | minRT | maxRT | peakID | isotope | isoGroup | LipidMSid | Adduct | confidenceLevel | Score |
|----------|--------|-------------|--------|--------|------------|---------|----------|---------------------|--------------------|-----------------|-------------|
| 116.0368 | 13.0 | 13371103 | 12.2 | 18.3 | MS1_0_8 | | 0 | | | | |
| 116.0705 | 36.0 | 7004341 | 34.1 | 42.0 | MS1_0_9 | [M+0] | 1 | | | | |
| 175.1190 | 37.8 | 29380745 | 34.1 | 46.9 | MS1_0_207 | [M+0] | 6 | | | | |
| 176.1223 | 36.6 | 1866761 | 32.3 | 45.0 | MS1_0_210 | [M+1] | 6 | | | | |
| 313.2735 | 113.0 | 5515012 | 113.1 | 121.0 | MS1_0_627 | [M+0] | 27 | MG(16:0) | M+H ₂ O | MS-only | 0 |
| 339.2892 | 914.9 | 7950690 | 906.8 | 917.7 | MS1_0_732 | [M+0] | 40 | MG(18:1) | M+H ₂ O | MS-only | 0 |
| 341.2085 | 74.2 | 7219435 | 71.8 | 80.9 | MS1_0_740 | [M+0] | 41 | | | | |
| 396.3111 | 67.5 | 795318 | 63.9 | 68.2 | MS1_0_996 | [M+0] | 75 | MG(20:4) | M+NH ₄ | MS-only | 0 |
| 396.3320 | 95.4 | 118616577 | 87.6 | 104.6 | MS1_0_997 | [M+0] | 76 | | | | |
| 396.3471 | 329.6 | 9595242 | 328.5 | 332.1 | MS1_0_998 | | 0 | | | | |
| 397.3140 | 66.9 | 93087 | 65.7 | 69.4 | MS1_0_999 | [M+1] | 75 | | | | |
| 397.3352 | 95.4 | 25339150 | 93.1 | 101.0 | MS1_0_1000 | [M+1] | 76 | | | | |
| 496.3398 | 111.2 | 20668285805 | 106.5 | 146.6 | MS1_0_1507 | [M+0] | 164 | LPC(16:0) | M+H | FA | 0.988 |
| 518.3218 | 113.0 | 268482290 | 112.5 | 130.8 | MS1_0_1688 | [M+0] | 194 | LPC(16:0) | M+Na | FA | 0.988 |
| 524.3712 | 184.7 | 10888241030 | 178.3 | 230.6 | MS1_0_1742 | [M+0] | 205 | LPC(18:0) | M+H | FA | 0.99 |
| 525.2881 | 33.6 | 113170499 | 32.3 | 41.4 | MS1_0_1755 | [M+0] | 207 | | | | |
| 525.2890 | 448.6 | 426324 | 444.3 | 455.8 | MS1_0_1756 | | 0 | | | | |
| 544.3372 | 124.5 | 183695112 | 118.6 | 130.8 | MS1_0_1928 | [M+0] | 242 | LPC(18:1)/PCp(18:0) | M+Na/M+Na | FA/Subclass | 0.993 0.993 |
| 544.3397 | 82.7 | 81919226 | 75.5 | 84.0 | MS1_0_1929 | [M+0] | 244 | LPC(20:4)/PCp(18:0) | M+H/M+Na | FA/Subclass | 0.984 0.907 |
| 546.3525 | 187.7 | 613201759 | 177.7 | 200.8 | MS1_0_1937 | [M+0] | 247 | LPC(18:0) | M+Na | FA | 0.99 |
| 548.4343 | 449.8 | 195193209 | 446.1 | 458.2 | MS1_0_1962 | [M+0] | 253 | | | | |
| 548.4758 | 1096.7 | 18031029 | 1092.2 | 1099.4 | MS1_0_1963 | [M+1] | 251 | | | | |
| 606.6181 | 964.4 | 13936746 | 962.0 | 968.0 | MS1_0_2533 | [M+0] | 390 | | | | |
| 743.5586 | 917.9 | 37631513 | 915.9 | 925.0 | MS1_0_4739 | | 0 | Cer(d18:0/22:0) | M+H ₂ O | FA position | 0.871 |

C. Feature table of the *msbatch*.

| mz | minmz | maxmz | RT | minRT | maxRT | intRT | endRT | npeaks | group | isotope | LipidMSid | Adduct | confidenceLevel | Score | Samples... |
|----------|----------|----------|-------|-------|-------|-------|-------|--------|-------|---------|-------------------------------------|-----------------------|-----------------------------|-------------|-------------|
| 313.2733 | 313.2731 | 313.2735 | 106.3 | 102.1 | 109.6 | 106.8 | 109.0 | 11 | 825 | [M+0] | MG(16:0) | M+H ₂ O | MS-only | 0 | 891.75 |
| 316.2480 | 316.2479 | 316.2481 | 42.3 | 41.7 | 42.9 | 38.9 | 46.2 | 26 | 850 | [M+0] | Carnitine(10:0) | M+H | FA | 0.061 | 28084702 |
| 319.2385 | 319.2384 | 319.2387 | 55.2 | 54.6 | 55.9 | 53.7 | 61.9 | 17 | 878 | [M+0] | | | | | 368454 |
| 328.2479 | 328.2477 | 328.2481 | 42.0 | 41.2 | 42.8 | 39.1 | 47.0 | 26 | 917 | [M+0] | | | | | 5314619 |
| 339.2890 | 339.2889 | 339.2892 | 915.0 | 914.1 | 915.6 | 911.4 | 916.9 | 24 | 978 | [M+0] | MG(18:1) | M+H ₂ O | MS-only | 0 | 6226475 |
| 454.2928 | 454.2925 | 454.2931 | 113.9 | 113.3 | 114.7 | 108.3 | 122.4 | 26 | 1757 | [M+0] | LPE(16:0) | M+H | FA | 0.362 | 92726411 |
| 480.3087 | 480.3084 | 480.3088 | 326.7 | 325.2 | 327.4 | 323.1 | 341.8 | 26 | 1941 | [M+0] | LPC(18:1) | M+H | FA | 0.327 | 8885520 |
| 496.3397 | 496.3395 | 496.3399 | 110.7 | 109.6 | 111.8 | 106.2 | 139.2 | 26 | 2150 | [M+0] | Carnitine(22:5) | M+H ₂ M+Na | FA/Subclass | 0.988 0.02 | 20607628835 |
| 497.2357 | 497.2355 | 497.2358 | 33.5 | 33.3 | 34.2 | 32.5 | 41.1 | 24 | 2189 | [M+0] | | | | | 5344877 |
| 498.8010 | 498.8007 | 498.8011 | 34.8 | 35.5 | 33.0 | 38.4 | 24 | 2182 | [M+0] | | | | | | 14748318 |
| 500.2770 | 500.2766 | 500.2774 | 71.2 | 70.3 | 71.9 | 68.0 | 76.7 | 25 | 2196 | [M+0] | | | | | 3669785 |
| 518.3214 | 518.3209 | 518.3218 | 112.7 | 110.8 | 113.4 | 111.6 | 116.3 | 8 | 2431 | [M+0] | LPC(16:0) | M+Na | FA | 0.988 | 42274910 |
| 520.3399 | 520.3389 | 520.3402 | 90.5 | 89.8 | 90.9 | 86.9 | 108.8 | 26 | 2460 | [M+0] | LPC(18:2) | M+H | FA | 0.917 | 575356222 |
| 521.4203 | 521.4201 | 521.4207 | 296.8 | 296.2 | 297.4 | 293.7 | 299.7 | 17 | 2476 | [M+0] | | | | | 2367044 |
| 522.3551 | 522.3548 | 522.3554 | 112.7 | 111.5 | 114.0 | 106.6 | 117.1 | 26 | 2487 | [M+0] | LPC(18:1)/PCp(18:0) | M+H ₂ M+H | FA/Subclass | 0.835 0.839 | 243509768 |
| 524.3713 | 524.3711 | 524.3715 | 184.1 | 183.1 | 185.9 | 177.3 | 212.0 | 26 | 2520 | [M+0] | LPC(18:0) | M+H | FA | 0.993 | 10810843369 |
| 525.0432 | 525.0429 | 525.0436 | 287.0 | 286.5 | 287.5 | 284.9 | 289.6 | 18 | 2532 | [M+0] | | | | | 1461794 |
| 578.5865 | 578.5852 | 578.5874 | 923.0 | 922.5 | 923.7 | 918.8 | 928.8 | 25 | 3346 | [M+0] | Cer(d18:0/20:0) | M+H ₂ O | FA position | 0.853 | 3891790 |
| 581.4285 | 581.4283 | 581.4287 | 113.6 | 113.1 | 114.4 | 112.7 | 118.0 | 8 | 3407 | [M+0] | | | | | 3848702 |
| 594.5821 | 594.5818 | 594.5824 | 918.3 | 917.7 | 918.9 | 915.1 | 921.9 | 24 | 3588 | [M+0] | Cer(d18:1/20:0); Cer(d16:1/22:0) | M+H ₂ M+H | FA position; FA position | 0.959 0.896 | 8880180 |
| 630.6183 | 630.6175 | 630.6186 | 959.3 | 958.6 | 959.8 | 955.6 | 963.2 | 25 | 4220 | [M+0] | Cer(d18:1/24:1) | M+H ₂ O | FA position | 0.998 | 20980884 |
| 634.4807 | 634.4800 | 634.4811 | 395.3 | 394.8 | 395.8 | 393.4 | 397.2 | 14 | 4269 | [M+0] | | | | | 552941 |

Figure 21. Examples of annotation results returned by LipidMS.

2.2.3. Additional functions

Besides providing the feature table that summarizes the intensity and identity of all the detected lipids across samples, LipidMS allows to obtain different graphical outputs that improve the interpretation of the results. On the one hand, the *plotLipids* function allows depicting information that supports the proposed lipid identities, as well as the achieved level of confidence for each identification at the *msubject* level so that the user could have a file with all lipid identifications for each sample acquired in DIA (Figure 22) or DDA (Figure 23).

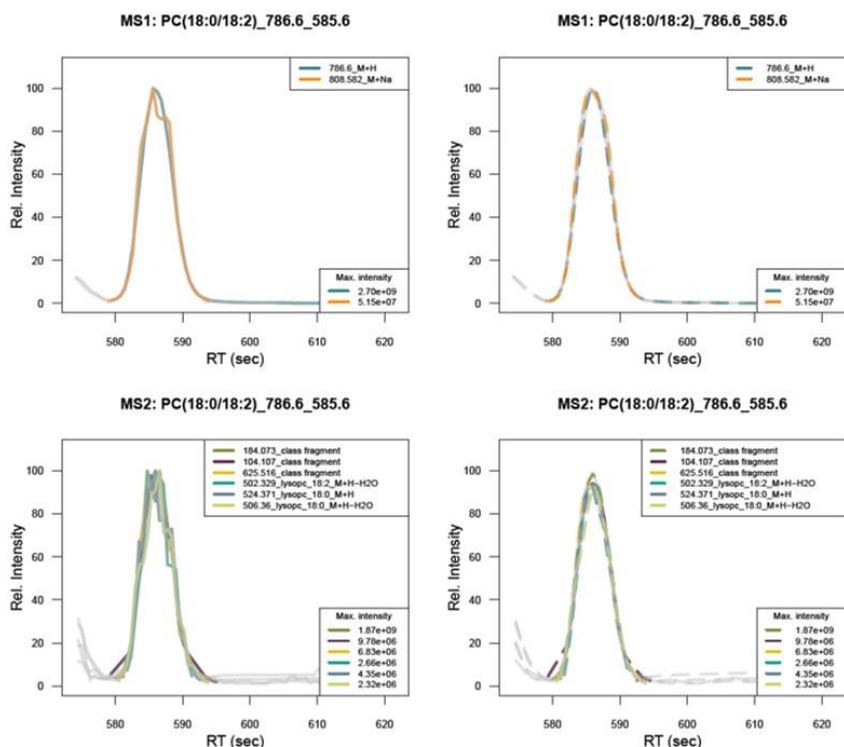


Figure 22. Example of graphical output of the *plotLipids* function for DIA acquired data. MS¹ plots show the adducts found for the annotated lipid specie and MS² plots show the fragments that support the identification. Plots displayed on the left side show raw data, while smoothed peaks are displayed on the right side.

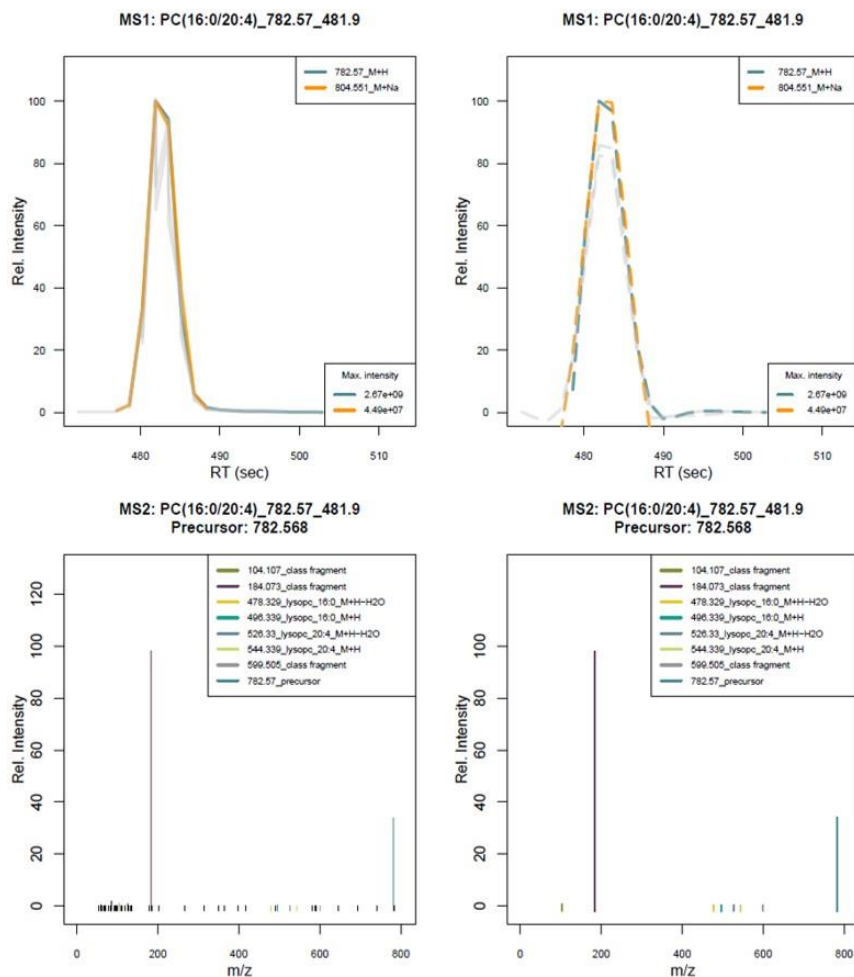
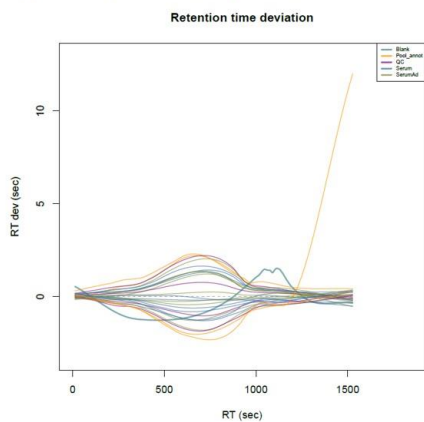


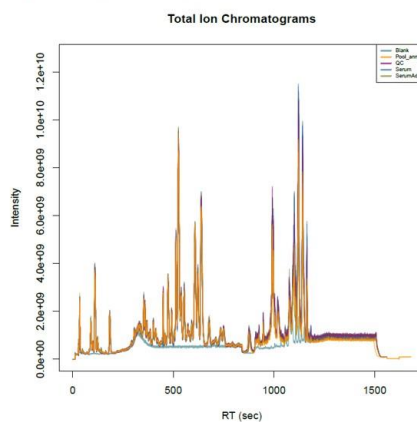
Figure 23. Example of graphical output of the *plotLipids* function for DDA acquired data. MS¹ plots show the adducts found for the annotated lipid specie and MS² plots show the fragments that support the identification. Plots displayed on the left side show raw data, while plots on the right side show the smoothed peak from MS¹ (up) and the clean MS/MS spectra (down).

On the other hand, functions such as *rtdevplot*, *plotticmsbatch* and *ploteicmsbatch*, allow users to check that peak-picking and alignment steps have worked properly by visualizing the whole dataset (Figure 24).

A. Graphical output of the *rtdevplot* function.



B. Graphical output of the *plotticmsbatch* function.



C. Graphical output of the *ploteicmsbatch* function.

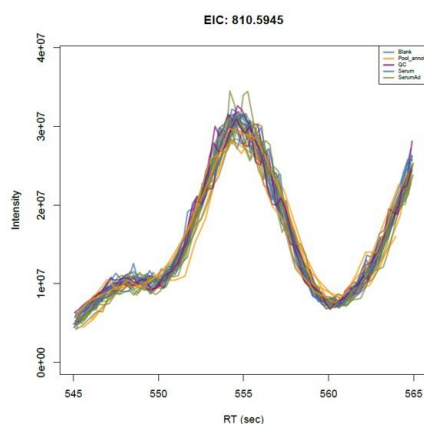


Figure 24. Examples of additional graphical outputs. A) *rtdevplot* function, B) *plotticmsbatch* function and C) *ploteicmsbatch* function.

2.3. Implementation

LipidMS has been developed in an R programming environment and is available via CRAN (<https://CRAN.R-project.org/package=LipidMS>). The source code and development

version are also available at <https://github.com/maialba3/LipidMS>. In addition, a web-based implementation of LipidMS has been built using the Shiny R-package¹⁴⁹, which is accessible at www.lipidms.es. Example data files, scripts, tutorials for the R package and web application and links to the development version can be found at <http://www.lipidms.es> via the “Resources” tab.

2.3.1. R package

LipidMS package works with two main types of objects: *msubject* and *msbatch*. On the one hand, the *msubject* is a list that contains all the raw and processed data for a single sample (raw scans, scans metadata, processing parameters, peaks properties and annotation results), while the *msbatch* consists of a list of all the *msubjects* that belong to a dataset and the information regarding to the alignment and grouping steps (i.e., processing parameters, clustering results, peak groups and feature matrix). With these two types of objects, LipidMS may be used to process and annotate single sample files acquired in DIA or DDA (*msubjects*) or to work with larger datasets that combine full scan, DIA and DDA acquisition (*msbatch*) (Figure25). In this case, several pre-processed *msubjects* are wrapped into a single *msbatch* to be aligned and grouped and then, DIA and DDA *msubjects* are annotated individually. Results are finally dumped into the general feature matrix. These two optional workflows are outlined in Figure25.

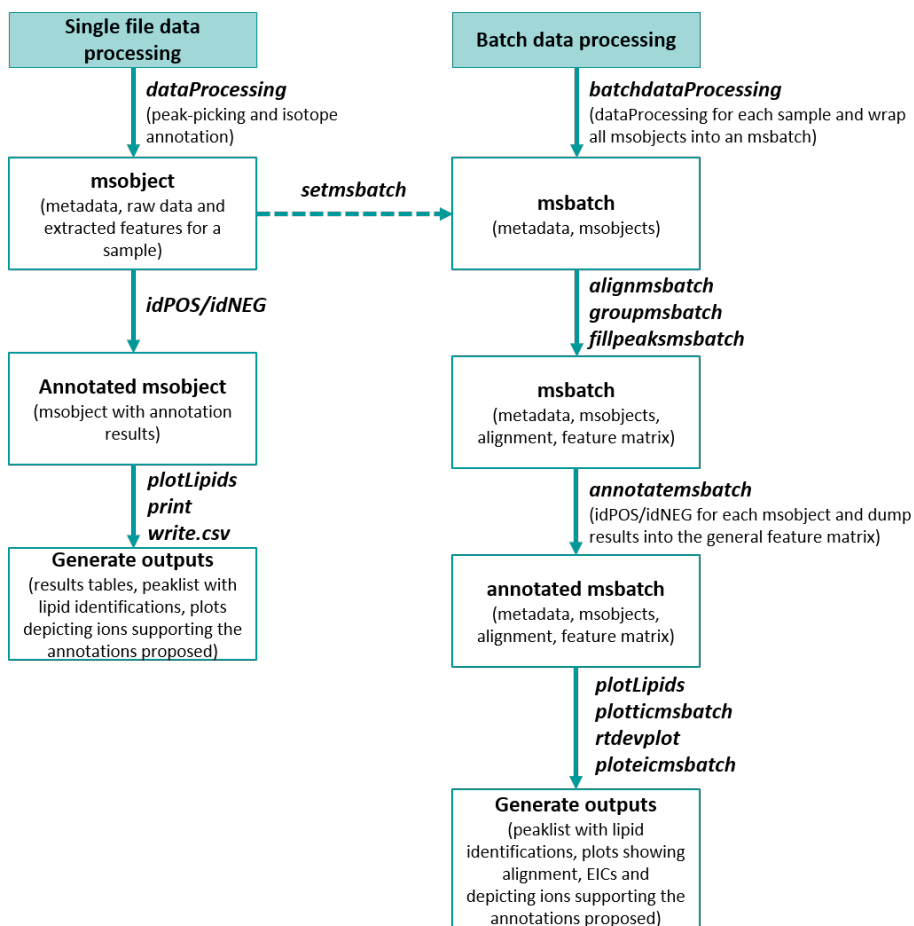


Figure 25. Alternative processing pipelines in LipidMS. A) Single files workflow, and B) batch data processing workflow.

2.3.2. Web-based application

In order to provide a user-friendly GUI interface, LipidMS has also been implemented as a web-based tool using Shiny¹⁴⁹, which is accessed at www.lipidms.es. After accessing the tool, the following tabs will take users through the LipidMS workflow:

- **Data import.** On the first tab (Figure 26), users must choose polarity and upload all the mzXML files and a metadata file in the csv format with three columns: sample (mzXML file names); acquisition mode (MS for full scan, DIA or DDA); sample type (e.g., QC, group1, group2, etc.).
- **Peak-picking.** Then all the parameters required for peak-picking can be tuned. On this tab (Figure 27), the MS¹ and MS² values correspond to those parameters used to process the MS¹ level in all cases and the MS² level for the DIA data files, respectively.
- **Batch processing.** The third tab (Figure 28) contains the parameters required for alignment, grouping and filling peak steps.
- **Annotation.** On the annotation tab (Figure 29), the lipid classes to be searched, and the *m/z* and RT tolerances, can be defined.
- **Run.** Finally, users can run their job (Figure 30). The results will be sent to the email provided by the user and will contain two or three csv files with the results tables (feature matrix if batch processing is performed, summary tables and the whole peak tables with annotations) and the pdf files with plots of the peaks supporting the lipid identifications for all the files.

Extra documentation, examples and links to the source code in github or CRAN can be found by clicking on the “Resources” tab.

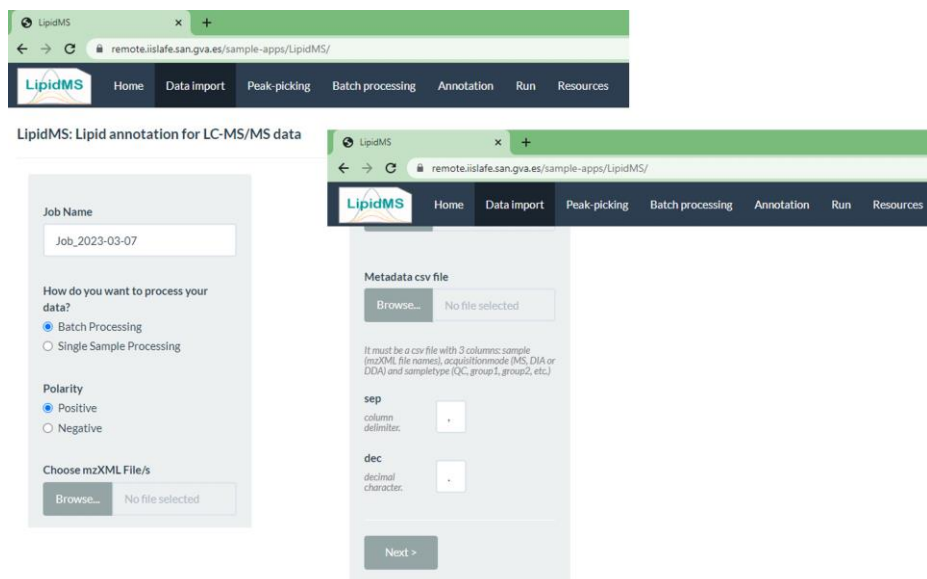


Figure 26. Data import tab of the LipidMS web tool. On this tab, users can upload mzXML data files and the csv metadata file.

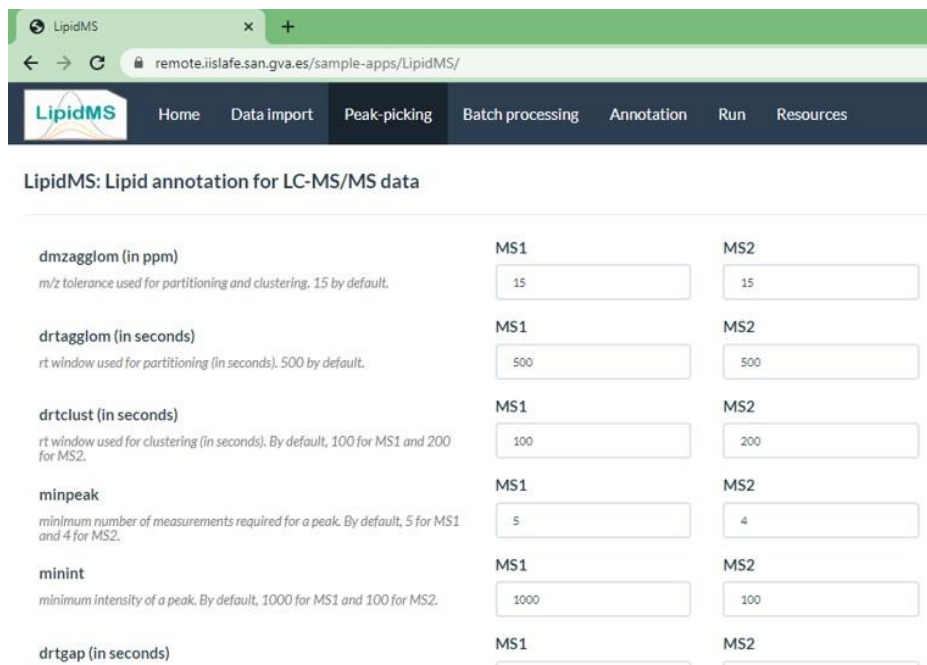


Figure 27. Peak-picking tab of the LipidMS web tool. On this tab, users can tune the peak-picking processing parameters for peak extraction in MS¹ and MS² levels separately.

LipidMS

Home Data import Peak-picking **Batch processing** Annotation Run Resources

LipidMS: Lipid annotation for LC-MS/MS data

dmzalign
mass tolerance between peak groups for alignment (in ppm). 5 by default.

drtalign
maximum rt distance between peaks for alignment (in seconds). 30 by default.

span
span parameter for loess rt smoothing. 0.4 by default.

minsamplesfracalign
minimum samples fraction represented in each cluster used for alignment. 0.75 by default.

Figure 28. Batch processing tab of the LipidMS web tool. On this tab, users can tune the parameters used for alignment and grouping.

LipidMS

Home Data import Peak-picking Batch processing **Annotation** Run Resources

LipidMS: Lipid annotation for LC-MS/MS data

dmzprecursor
mass tolerance for precursor ions. 5 by default.

dmzproducts
mass tolerance for product ions. 10 by default.

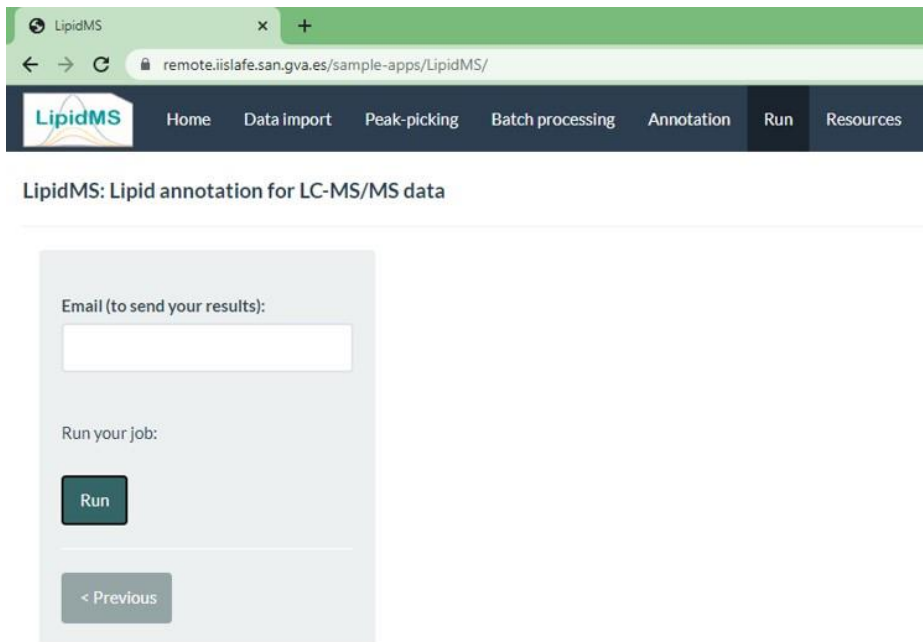
rttol
total rt window for coelution between precursor and product ions. 5 by default.

coelcutoff
coelution score threshold between parent and fragment ions. Only applied if rawData info is supplied. 0.7 by default.

Lipid classes to annotate for ESI+ :
 MG

Lipid classes to annotate for ESI- :
 FA

Figure 29. Annotation tab of the LipidMS web tool. On this tab, users can tune the parameters used for lipid annotation.



LipidMS is intended to be used for research purposes only, without any medical objective.

[Return to www.iislafes.es](http://www.iislafes.es)

Figure 30. Run tab of the LipidMS web tool. On this tab, users can tune the parameters used for alignment and grouping.

3. LipidMS performance evaluation

To evaluate LipidMS performance, a commercial human serum pool (Sigma-Aldrich reference P2918) was analyzed by LC-MS using both the ESI+ and ESI- ionization modes and in full scan, DIA and DDA acquisition modes. To provide an objective qualifier for the comparison, serum was extracted with or without the addition of 68 lipid standards. Raw data files and results are available at Zenodo (<https://doi.org/10.5281/zenodo.6645498>). Three different workflows were compared: i) LipidMS workflow; ii) MS-DIAL workflow¹³⁶; iii) data pre-processing using XCMS³¹ followed by isotope annotation using CAMERA¹⁴⁵¹⁴⁵.

When employing the LipidMS workflow (i.e., data pre-processing and lipid annotation), the samples acquired in full scan, DIA and DDA were simultaneously processed, which was not possible using MS-DIAL. For MS-DIAL, the MS and DDA files were processed together and the DIA files separately. Then DIA identifications were added to the feature matrix obtained for the full scan and DDA files using a m/z tolerance of 0.005 and an RT tolerance of 10 seconds. XCMS was exclusively applied to pre-process the MS¹ level of all the samples and CAMERA for the annotation of the isotopes in the generated feature matrix. In all cases, the reported features referred to the combination of the positive and negative ionization modes for the MS¹ level, and lipid identities are based on the information obtained for MS² from the DDA and DIA acquired samples. In all three cases, features were filtered and normalized based on quality control (QC) samples. Only the features that appeared in at least 70% of the QC samples were kept. Then, data was normalized using a LOESS function, which was fitted to the QC samples based on the injection order. Finally, a differential analysis between the spiked and non-spiked serum samples was performed using a Student's t-test.

The performance of each pre-processing software was evaluated by comparing the number of detected and identified lipid standard features (a single lipid can be annotated using different features that could result from the ionization of many adducts, e.g. $[M+H]^+$ and $[M+Na]^+$, thus, 129 features were expected from the 68 lipids) and the differences in these lipid standard features between spiked and non-spiked samples. Additionally, for the comparison between MS-DIAL and LipidMS workflows, the number of correctly and wrongly annotated lipids was evaluated.

3.1. Comparison between LipidMS and XCMS data pre-processing

First LipidMS version was designed for lipid annotation of single samples, which required the use of external software as XCMS to perform data pre-processing (i.e., from peak-picking to fill missing peaks). From LipidMS v3.0, it includes the whole workflow for batch processing. To evaluate the performance of these pre-processing steps incorporated into LipidMS v3.0, we compared the results obtained by the LipidMS workflow to those obtained by employing one of the most widely used platforms in MS data processing: XCMS³¹.

Despite the fact that XCMS found a larger total number of features than LipidMS (33352 in XCMS vs. 19382 in LipidMS) (Table 5), both software provide similar numbers in terms of the expected lipid standard features detected and significant changes in them (Table 5-7). These results validate the LipidMS pre-processing workflow, which for lipidomic studies provided results that were comparable to those obtained by XCMS.

Table 5. Summary of the expected lipid standard features detected and identified for each software package. Significant features were defined by an adjusted p -value < 0.05 and fold change < 1.5 .

| | Expected lipid standards | Expected lipid standard features | Total number of features | | | Expected features found | | | Significant features detected | | | Identified features | | Identified lipid standards | |
|--------------|--------------------------|----------------------------------|--------------------------|--------------|--------------|-------------------------|----------------|----------------|-------------------------------|----------------|---------------|---------------------|---------------|----------------------------|--------------|
| | | | LipidMS | XCMS | MS-DIAL | LipidMS | XCMS | MS-DIAL | LipidMS | XCMS | MS-DIAL | LipidMS | MS-DIAL | LipidMS | MS-DIAL |
| ESI - | 56 | 69 | 10220 | 19776 | 39366 | 68/69 | 56/69 | 56/69 | 59/69 | 56/69 | 54/69 | 55/69 | 43/69 | 48/56 | 42/56 |
| ESI + | 29 | 60 | 15354 | 24486 | 35897 | 54/60 | 58/60 | 59/60 | 41/60 | 46/60 | 42/60 | 43/60 | 33/60 | 22/29 | 21/29 |
| Total | 68 | 129 | 25574 | 44262 | 75263 | 122/129 | 114/129 | 115/129 | 100/129 | 102/129 | 96/129 | 98/129 | 76/129 | 60/68 | 56/68 |

Table 6. Lipid standards detected and identified in ESI-. Significant features were defined by an adjusted *p*-value < 0.05 and fold change < 1.5. (*) Denotes incorrect annotations.

| Compound | Adduct | m/z | RT (s) | Expected features found | | | Significant features detected | | | Lipid standards identified | |
|------------|--------|----------|--------|-------------------------|------|---------|-------------------------------|------|---------|----------------------------|----------|
| | | | | Lipids | XCMS | MS-DIAL | Lipids | XCMS | MS-DIAL | Lipids | MS-DIAL |
| FA(10:0) | M+H | 171.1385 | 54 | y | y | y | y | y | y | - | - |
| FA(12:0) | M+H | 199.1698 | 74 | y | y | y | y | y | y | FA(12:0) | FA(12:0) |
| FA(14:0) | M+H | 227.2011 | 116 | y | y | y | y | y | y | FA(14:0) | FA(14:0) |
| FA(14:1) | M+H | 225.1855 | 83 | y | y | y | y | y | y | FA(14:1) | - |
| FA(15:0) | M+H | 241.2167 | 149 | y | y | y | y | y | y | FA(15:0) | - |
| FA(16:0) | M+H | 255.2324 | 195 | y | y | y | n | n | n | - | FA(16:0) |
| FA(16:1) | M+H | 253.2167 | 133 | y | y | y | y | y | y | FA(16:1) | FA(16:1) |
| FA(17:0) | M+H | 269.2481 | 245 | y | y | y | y | y | y | FA(17:0) | FA(17:0) |
| FA(18:0) | M+H | 283.2637 | 291 | y | y | y | n | n | n | FA(18:0) | FA(18:0) |
| FA(18:1)n9 | M+H | 281.2481 | 218 | y | y | y | n | n | n | FA(18:1) | FA(18:1) |
| FA(18:1)n7 | M+H | 281.2481 | 229 | y | y | y | y | y | y | FA(18:1) | FA(18:1) |
| FA(18:2) | M+H | 279.2324 | 170 | y | y | y | y | y | y | FA(18:2) | FA(18:2) |
| FA(18:3)n3 | M+H | 277.2168 | 111 | y | n | y | y | - | n | - | - |
| FA(18:3)n6 | M+H | 277.2168 | 115 | y | y | y | y | y | y | FA(18:3) | FA(18:3) |
| FA(19:0) | M+H | 297.2799 | 322 | y | y | y | y | y | y | - | FA(19:0) |
| FA(20:0) | M+H | 311.295 | 347 | y | y | y | y | y | y | FA(20:0) | FA(20:0) |
| FA(20:1) | M+H | 309.2794 | 302 | y | y | y | y | y | y | FA(20:1) | - |
| FA(20:2) | M+H | 307.2637 | 241 | y | y | y | y | y | y | FA(20:2) | FA(20:2) |
| FA(20:3) | M+H | 305.2481 | 182 | y | y | y | y | y | y | FA(20:3) | FA(20:3) |
| FA(20:3) | M+H | 305.2481 | 198 | y | y | y | y | y | y | FA(20:3) | FA(20:3) |
| FA(20:4)n3 | M+H | 303.2324 | 139 | y | n | y | y | - | y | - | - |
| FA(20:4)n6 | M+H | 303.2324 | 139 | y | y | y | y | y | y | FA(20:4) | FA(20:4) |
| FA(20:5) | M+H | 301.2168 | 103 | y | y | y | y | y | y | FA(20:5) | FA(20:5) |
| FA(22:0) | M+H | 339.3263 | 406 | y | y | y | y | y | y | FA(22:0) | FA(22:0) |

| Compound | Adduct | m/z | RT (s) | Expected features found | | | Significant features detected | | | Lipid standards identified | |
|-----------------|-----------------------|----------|--------|-------------------------|------|---------|-------------------------------|------|---------|----------------------------|--------------------------|
| | | | | LipidMS | XCMS | MS-DIAL | LipidMS | XCMS | MS-DIAL | LipidMS | MS-DIAL |
| FA(22:1) | M+H | 337.3107 | 352 | y | y | y | y | y | y | FA(22:1) | FA(22:1) |
| FA(22:2) | M+H | 335.295 | 315 | y | y | y | y | y | y | FA(22:2) | FA(22:2) |
| FA(22:3) | M+H | 333.2794 | 270 | y | y | y | y | y | y | FA(22:3) | FA(22:3) |
| FA(22:4)n3 | M+H | 331.2637 | 212 | y | y | y | y | n | y | - | - |
| FA(22:4)n6 | M+H | 331.2637 | 217 | y | n | y | y | y | - | FA(22:4) | FA(22:4) |
| FA(22:5)n3 | M+H | 329.2481 | 158 | y | y | y | y | y | y | FA(22:5) | FA(22:5) |
| FA(22:5)n6 | M+H | 329.2481 | 165 | y | y | y | y | y | y | FA(22:5) | FA(22:5) |
| FA(22:6) | M+H | 327.2324 | 121 | y | y | y | y | y | y | FA(22:6) | FA(22:6) |
| FA(24:0) | M+H | 367.3576 | 483 | y | y | y | y | y | y | FA(24:0) | FA(24:0) |
| FA(24:1) | M+H | 365.3419 | 408 | y | y | y | y | y | y | FA(24:1) | FA(24:1) |
| FA(26:0) | M+H | 395.3889 | 585 | y | y | y | y | y | y | FA(26:0) | FA(26:0) |
| LPC(17:0) | M+CH ₃ COO | 568.3621 | 135 | y | y | y | y | y | y | LPC(17:0) | LPC(17:0) |
| LPC(17:0) | M-CH ₃ | 494.3247 | 135 | y | y | n | y | y | - | LPC(17:0) | - |
| PC(17:0/17:0) | M+CH ₃ COO | 820.6074 | 637 | y | y | n | y | y | - | PC(17:0/17:0) | - |
| PC(17:0/17:0) | M-CH ₃ | 746.57 | 637 | y | y | y | y | y | y | PC(17:0/17:0) | *PE(18:0_18:0) |
| PC(16:0/18:1) | M+CH ₃ COO | 818.5917 | 544 | y | y | y | n | n | n | PC(16:0/18:1) | *PC(O-17:0_17:2) |
| PC(16:0/18:1) | M-CH ₃ | 744.5544 | 544 | y | y | y | n | n | n | PC(16:0/18:1) | - |
| PC(18:0/18:2) | M+CH ₃ COO | 844.6074 | 567 | y | y | y | n | n | n | PC(18:0/18:2) | - |
| PC(18:0/18:2) | M-CH ₃ | 770.57 | 567 | y | y | y | n | n | n | PC(18:0/18:2) | *CL(18:0_18:0_22:0_20:3) |
| PC(O-16:0/20:5) | M+CH ₃ COO | 824.5812 | 462 | y | y | y | y | y | y | PC(O-16:0/20:5) | PC(O-16:0_20:5) |
| PC(O-16:0/20:5) | M-CH ₃ | 750.5438 | 462 | y | y | y | y | y | y | - | *PE(O-18:0_20:5) |
| PC(P-18:0/20:4) | M+CH ₃ COO | 852.6125 | 595 | y | y | y | y | y | y | PC(P-18:0/20:4) | *PC(O-18:1_20:4) |
| PC(P-18:0/20:4) | M-CH ₃ | 778.3751 | 595 | n | n | n | - | - | - | - | - |
| PE(17:0/17:0) | M+H | 718.5387 | 659 | y | y | y | y | y | y | - | PE(17:0_17:0) |
| PE(16:0/18:1) | M+H | 716.5231 | 561 | y | y | y | y | y | y | PE(16:0/18:1) | PE(16:0_18:1) |
| PE(O-16:0/18:1) | M+H | 702.5438 | 630 | y | y | y | y | y | y | - | PE(O-16:0_18:1) |
| PE(P-18:0/22:6) | M+H | 774.5438 | 588 | y | y | y | y | y | y | PE(P-18:0/22:6) | *PE(O-18:1_22:6) |

| Compound | Adduct | m/z | RT (s) | Expected features found | | | Significant features detected | | | Expected features found | |
|--------------------------|-----------------------|----------|--------|-------------------------|------|---------|-------------------------------|------|---------|-------------------------|---------|
| | | | | Lipids | XCMS | MS-DIAL | Lipids | XCMS | MS-DIAL | Lipids | MS-DIAL |
| PG(17:0/17:0) | M+H | 749.5333 | 480 | y | y | y | y | y | y | PG(17:0_17:0) | MS-DIAL |
| PG(16:0/18:1) | M+H | 747.5177 | 426 | y | y | y | y | y | y | PG(16:0_18:1) | MS-DIAL |
| PI(17:0/14:1) | M+H | 793.4867 | 350 | y | y | y | y | y | y | PI(17:0_14:1) | MS-DIAL |
| PS(17:0/17:0) | M+H | 762.5285 | 473 | y | y | y | y | y | y | PS(17:0_17:0) | MS-DIAL |
| PS(17:0/17:0) | M+Na-2H | 784.5099 | 473 | y | y | y | y | y | y | - | - |
| PS(16:0/18:1) | M+H | 760.5129 | 420 | y | y | y | y | y | y | PS(16:0_18:1) | MS-DIAL |
| PS(16:0/18:1) | M+Na-2H | 782.4943 | 420 | y | y | n | y | y | y | - | - |
| CL(18:1/18:1/18:1/18:1) | M+H | 1456.027 | 1014 | y | y | y | y | y | y | CL(18:1/18:1/18:1/18:1) | MS-DIAL |
| CL(18:1/18:1/18:1/18:1) | M+Na-2H | 1478.009 | 1014 | y | y | y | y | y | y | CL(18:1/18:1/18:1/18:1) | MS-DIAL |
| SM(d18:1/17:0) | M+CH ₃ COO | 775.5972 | 476 | y | y | y | y | y | y | SM(18:1/17:0) | MS-DIAL |
| SM(d18:1/17:0) | M-CH ₃ | 701.5598 | 476 | y | y | y | y | y | y | SM(35:1) | MS-DIAL |
| SM(d18:1/16:0) | M+CH ₃ COO | 761.5815 | 440 | y | y | y | n | n | n | SM(18:1/16:0) | MS-DIAL |
| SM(d18:1/16:0) | M-CH ₃ | 687.5441 | 440 | y | y | y | n | n | n | SM(18:1/16:0) | MS-DIAL |
| Cer(d18:1/17:0) | M+CH ₃ COO | 550.5199 | 603 | y | y | y | y | y | y | Cer(d18:1/17:0) | MS-DIAL |
| Cer(d18:1/17:0) | M-CH ₃ | 610.5416 | 603 | y | y | y | y | y | y | Cer(d18:1/17:0) | MS-DIAL |
| CerP(d18:1/16:0) | M+H | 616.4707 | 384 | y | y | y | y | y | y | CerP(d34:1) | MS-DIAL |
| AcylCer(18:1-d18:1/17:0) | M+H | 814.7652 | 1030 | y | y | y | y | y | y | - | - |
| AcylCer(18:1-d18:1/17:0) | M+CH ₃ COO | 874.7869 | 1030 | y | y | y | y | y | y | - | - |

Table 7. Lipid standards detected and identified in ESI+. Significant features were defined by an adjusted *p*-value < 0.05 and fold change < 1.5. (*) Denotes incorrect annotations.

| Compound | Adduct | m/z | RT (s) | Expected features found | | | Significant features detected | | | Lipid standards identified | |
|-----------------|--------|----------|--------|-------------------------|------|---------|-------------------------------|------|---------|----------------------------|-----------------|
| | | | | LipidMS | XCMS | MS-DIAL | LipidMS | XCMS | MS-DIAL | LipidMS | MS-DIAL |
| LPC(17:0) | M+H | 510.356 | 142 | y | y | y | y | y | y | LPC(17:0) | LPC(17:0) |
| LPC(17:0) | M+Na | 532.3372 | 142 | y | y | y | y | y | y | LPC(17:0) | *LPC(19:3) |
| PC(17:0/17:0) | M+H | 762.6013 | 730 | y | y | y | y | y | n | PC(17:0/17:0) | PC(17:0_17:0) |
| PC(17:0/17:0) | M+Na | 784.5827 | 730 | y | y | y | y | y | y | - | - |
| PC(16:0/18:1) | M+H | 760.5856 | 609 | y | y | y | y | n | n | PC(16:0/18:1) | PC(16:0_18:1) |
| PC(16:0/18:1) | M+Na | 782.567 | 609 | y | y | y | y | n | n | PC(16:0/18:1) | PC(34:1) |
| PC(18:0/18:2) | M+H | 786.6013 | 640 | y | y | y | y | n | n | PC(18:0/18:2) | PC(18:0_18:2) |
| PC(18:0/18:2) | M+Na | 808.5827 | 640 | y | y | y | y | n | n | *PC(18:1/18:1) | PC(36:2) |
| PC(O-16:0/20:5) | M+H | 766.5751 | 505 | y | y | y | y | y | y | PC(O-16:0/20:5) | PC(O-36:5) |
| PC(O-16:0/20:5) | M+Na | 788.5565 | 505 | y | y | y | y | y | y | PC(O-16:0/20:5) | - |
| PC(P-18:0/20:4) | M+H | 794.6064 | 675 | y | y | y | y | y | y | PC(P-18:0/20:4) | *PC(O-38:5) |
| PC(P-18:0/20:4) | M+Na | 816.5878 | 675 | y | y | y | y | y | n | PC(P-18:0/20:4) | *PC(O-40:8) |
| PE(17:0/17:0) | M+H | 720.5543 | 762 | y | y | y | y | y | y | PE(17:0/17:0) | PE(17:0/17:0) |
| PE(17:0/17:0) | M+Na | 742.5357 | 762 | n | n | n | - | - | - | - | - |
| PE(16:0/18:1) | M+H | 718.5387 | 635 | y | y | y | y | y | y | - | - |
| PE(16:0/18:1) | M+Na | 740.5201 | 635 | n | n | n | - | - | - | - | - |
| PE(O-16:0/18:1) | M+H | 704.5594 | 725 | y | y | y | y | y | y | PE(O-16:0/18:1) | PE(O-34:1) |
| PE(O-16:0/18:1) | M+Na | 726.5408 | 725 | n | n | n | - | - | - | - | - |
| PE(P-18:0/22:6) | M+H | 776.5594 | 675 | y | y | y | y | y | y | *PE(O-18:1_22:6) | PE(P-18:0_22:6) |
| PE(P-18:0/22:6) | M+Na | 798.5408 | 675 | n | n | n | - | - | n | - | - |
| PG(17:0/17:0) | M+H | 751.5489 | 587 | y | y | y | y | y | y | - | - |
| PG(17:0/17:0) | M+Na | 773.5303 | 587 | n | n | n | - | - | - | - | - |
| PG(16:0/18:1) | M+H | 749.5333 | 503 | y | y | y | y | y | y | - | - |

| Compound | Adduct | m/z | RT (s) | Expected features found | | | Significant features detected | | | Lipid standards identified | |
|-------------------------|--------------------|----------|--------|-------------------------|------|---------|-------------------------------|------|---------|----------------------------|--------------------|
| | | | | Lipids | KCMS | MS-DIAL | Lipids | KCMS | MS-DIAL | Lipids | MS-DIAL |
| PG(16:0/18:1) | M+Na | 771.5147 | 503 | Y | Y | Y | Y | Y | Y | - | - |
| PI(17:0/14:1) | M+H | 795.5023 | 384 | Y | Y | Y | Y | Y | Y | - | - |
| PI(17:0/14:1) | M+NH ₄ | 812.5289 | 384 | Y | Y | Y | Y | Y | Y | PI(31:1) | PI(31:1) |
| PI(17:0/14:1) | M+Na | 817.4837 | 384 | Y | Y | Y | Y | Y | Y | PI(31:1) | - |
| SM(dl18:1/17:0) | M+H | 717.5911 | 520 | Y | Y | Y | Y | Y | Y | SM(dl18:1/17:0) | *SM(dl17:1/18:0) |
| SM(dl18:1/17:0) | M+Na | 739.5725 | 520 | Y | Y | Y | Y | Y | Y | SM(d35:1) | SM(d35:1) |
| SM(dl18:1/16:0) | M+H | 703.5754 | 476 | Y | Y | Y | Y | n | n | SM(d18:1/16:0) | SM(d18:1/16:0) |
| SM(dl18:1/16:0) | M+Na | 725.5568 | 476 | Y | Y | Y | Y | n | n | SM(d18:1/16:0) | SM(d34:1) |
| Cer(d18:1/17:0) | M+H | 552.5355 | 682 | Y | Y | Y | Y | Y | Y | Cer(d18:1/17:0) | Cer(d18:1/17:0) |
| Cer(d18:1/17:0) | M+H ₂ O | 534.5249 | 682 | Y | Y | Y | Y | Y | Y | Cer(d18:1/17:0) | Cer(d18:1/17:0) |
| Cer(d18:1/17:0) | M+Na | 574.5169 | 682 | Y | Y | Y | Y | Y | n | Cer(d18:1/17:0) | *Cer(d18:1/19:2) |
| CerP(d18:1/16:0) | M+H | 618.4863 | 377 | Y | Y | Y | Y | n | n | - | - |
| AcylCer(18:1:18:1/17:0) | M+H | 816.7808 | 1077 | Y | Y | Y | Y | Y | Y | AcylCer(18:1:dl18:1/17:0) | *Cer(53:3:30) |
| AcylCer(18:1:18:1/17:0) | M+H ₂ O | 798.7702 | 1077 | Y | Y | Y | Y | Y | Y | AcylCer(18:1:dl18:1/17:0) | *Cer(53:3:30) |
| AcylCer(18:1:18:1/17:0) | M+Na | 838.7622 | 1077 | Y | Y | Y | Y | Y | Y | AcylCer(18:1:dl18:1/17:0) | *Cer(55:6:30) |
| MG(17:0) | M+Na | 367.2818 | 278 | Y | Y | Y | Y | Y | Y | - | - |
| DG(17:0/17:0) | M+H ₂ O | 579.5351 | 953 | Y | Y | Y | Y | Y | Y | DG(17:0/17:0) | - |
| DG(17:0/17:0) | M+NH ₄ | 614.5723 | 953 | Y | Y | Y | Y | Y | Y | DG(17:0/17:0) | DG(17:0_17:0) |
| DG(17:0/17:0) | M+Na | 619.5271 | 953 | Y | Y | Y | Y | Y | Y | DG(17:0/17:0) | DG(34:0) |
| TG(8:0/8:0/8:0) | M+NH ₄ | 488.3951 | 391 | Y | Y | Y | Y | Y | Y | TG(8:0/8:0/8:0) | TG(8:0_8:0_8:0) |
| TG(8:0/8:0/8:0) | M+Na | 493.3499 | 391 | Y | Y | Y | Y | Y | Y | TG(8:0/8:0/8:0) | TG(8:0_8:0_8:0) |
| TG(10:0/10:0/10:0) | M+NH ₄ | 572.489 | 648 | Y | Y | Y | Y | Y | Y | TG(10:0/10:0/10:0) | TG(10:0_10:0_10:0) |
| TG(10:0/10:0/10:0) | M+Na | 577.4438 | 648 | Y | Y | Y | Y | Y | Y | TG(10:0/10:0/10:0) | TG(10:0_10:0_10:0) |
| TG(12:0/12:0/12:0) | M+NH ₄ | 656.5829 | 966 | Y | Y | Y | Y | Y | Y | TG(12:0/12:0/12:0) | TG(12:0_12:0_12:0) |
| TG(12:0/12:0/12:0) | M+Na | 661.5376 | 966 | Y | Y | Y | Y | Y | Y | TG(12:0/12:0/12:0) | TG(12:0_12:0_12:0) |
| TG(14:0/14:0/14:0) | M+NH ₄ | 740.6767 | 1063 | Y | Y | Y | Y | Y | Y | TG(14:0/14:0/14:0) | TG(14:0_14:0_14:0) |
| TG(14:0/14:0/14:0) | M+Na | 745.6315 | 1063 | Y | Y | Y | Y | Y | Y | TG(14:0/14:0/14:0) | TG(14:0_14:0_14:0) |

| Compound | Adduct | m/z | RT (s) | Expected features found | | | Significant features detected | | | Lipid standards identified | |
|--------------------|--------------------|----------|--------|-------------------------|------|---------|-------------------------------|------|---------|----------------------------|--------------------|
| | | | | LipidMS | KCMS | MS-DIAL | LipidMS | KCMS | MS-DIAL | LipidMS | MS-DIAL |
| TG(16:0/16:0/16:0) | M+NH ₄ | 824.7707 | 1140 | y | y | y | n | n | n | TG(16:0_16:0_16:0) | MS-DIAL |
| TG(16:0/16:0/16:0) | M+Na | 829.7255 | 1140 | y | y | y | y | y | y | TG(16:0/16:0/16:0) | - |
| TG(17:0/17:0/17:0) | M+NH ₄ | 866.8176 | 1173 | y | y | y | y | y | y | TG(17:0/17:0/17:0) | TG(17:0_17:0_17:0) |
| TG(17:0/17:0/17:0) | M+Na | 871.7724 | 1173 | y | y | y | y | y | y | TG(17:0/17:0/17:0) | TG(17:0_17:0_17:0) |
| TG(18:1/16:0/18:1) | M+NH ₄ | 876.802 | 1142 | y | y | y | n | n | n | TG(18:1/16:0/18:1) | TG(16:0_18:1_18:1) |
| TG(18:1/16:0/18:1) | M+Na | 881.7568 | 1142 | y | y | y | n | n | n | TG(18:1/16:0/18:1) | TG(16:0_18:1_18:1) |
| TG(16:0/16:0/18:1) | M+NH ₄ | 850.7863 | 1140 | y | y | y | n | n | n | TG(16:0/16:0/18:1) | TG(16:0_16:0_18:1) |
| TG(16:0/16:0/18:1) | M+Na | 855.7411 | 1140 | y | y | y | n | n | n | TG(16:0/16:0/18:1) | TG(16:0_16:0_18:1) |
| CE(17:0) | 2M+NH ₄ | 1295.235 | 1157 | y | y | y | y | y | y | - | - |
| CE(17:0) | 2M+Na | 1300.189 | 1157 | n | y | y | - | y | y | - | - |

3.1. Comparison between LipidMS and MS-DIAL

3.1.1. Data processing and annotation of known lipid standards

The whole LipidMS workflow from data processing to lipid annotation was compared with one of the most employed tools in MS-based lipidomics, which is MS-DIAL¹³⁶. MS-DIAL found a larger number of features than LipidMS (75263 vs. 25574) (Table 5), but both provided similar numbers in terms of the expected lipid standard features that were detected and significant changes between additivated and non-additivated serum samples (Table 5-7). However, regarding lipid identification, LipidMS provided a larger number of both identified features (98/129 vs. 76/129) and identified lipid species (60/68 vs. 56/68) (Table 5-7). Most of the differences in the proposed identities are attributed to MS-DIAL incorrect assignation of some adducts, where ions $[M+H]^+$ and $[M-H]^-$ were correctly annotated, but adducts like $[M+Na]^+$, $[M+CH_3COO]^-$, $[M-CH_3]^-$ or $[M+Na-2H]^+$ were not annotated or incorrectly identified (Table 6-7). Thus, by means of this strategy, which focused on a subset of lipid classes covered by MS-DIAL and LipidMS, both software packages provided comparable results or LipidMS slightly outperformed MS-DIAL in some aspects. The improved LipidMS annotation of adducts compared to MS-DIAL was due to its underlying lipid annotation strategy in which features are first assigned as being related (e.g., putative $[M+H]^+$ and $[M+Na]^+$ ions of a given lipid), and then their lipid identity is proposed. This approach reduces the possibility of proposing different lipid identities for different adducts from a single lipid.

3.1.2. Annotation of lipids in human serum pool sample

Finally, LipidMS performance was compared to MS-DIAL in relation to the total number of lipid annotations provided for the aforementioned commercial human serum pool (Sigma-Aldrich reference P2918) analyzed by LC-MS in both the ESI+ and ESI- ionization modes and by full scan, DIA and DDA acquisition modes. All the annotations provided by MS-DIAL and LipidMS were manually curated and their results compared. The raw data files and an Excel file containing all the curated lipid identities, the annotations proposed by LipidMS and the annotations proposed by MS-DIAL can be accessed at Zenodo (<https://doi.org/10.5281/zenodo.6645498>).

For both polarities, MS-DIAL provided a larger number of correct lipid annotations (580 vs. 387 in ESI- and 588 vs. 445 in ESI+) (Tables 8-9). The main reasons for this increased coverage can be attributed to the following reasons: i) MS-DIAL covers more lipid classes than LipidMS; ii) MS-DIAL databases have a higher diversity of fatty acyl chains in terms of chain length and double bonds, including oxidized and hydroxylated fatty acyl chains; and iii) MS-DIAL presents a higher diversity of ceramides and sphingomyelins species than LipidMS. These results evidence that new releases of LipidMS should incorporate new lipid classes, sphingoid bases and fatty acyl moieties to fill this gap. Despite this lower number of identification, LipidMS provided higher structural information compared to MS-DIAL (i.e., a bigger proportion of lipids where the structural information level achieves an FA position). This improvement is because LipidMS uses the ratio between the intensity of fragments to elucidate the position of fatty acyl chains for most lipid classes, whereas MS-DIAL only discloses the fatty acyl position for ceramides and sphingomyelins. Additionally, MS-DIAL also provided more incorrect annotations in both polarities (669 vs. 79 in

ESI- and 897 vs. 50 in ESI+). For LipidMS, incorrect annotations represented less than the 20% of the proposed annotations, but added up to 60% of the proposed identities for MS-DIAL (Tables 8-9 and Figure 31). Most of the incorrect annotations in MS-DIAL came from the DIA data, where many annotations were based on noisy spectra and the majority of the reference ions were not present in the samples. As previously mentioned, many incorrect annotations were due to the incorrect assignment of adducts. This was particularly relevant for cardiolipins because almost all of them were annotated incorrectly, and some phosphatidylcholines and phosphatidylethanolamines were erroneously assigned to a particular class due to the miss-annotation of $[M+Na]^+$ as $[M+H]^+$ or $[M+CH_3COO]^-$ as $[M-H]^-$.

In short, when the performance of LipidMS and MS-DIAL to annotate lipids in a complex biological sample was compared, MS-DIAL annotated more lipids but also provided more incorrect annotations, (i.e., peaks that did not correspond to known lipids and were annotated or lipids with a miss-annotation), whereas LipidMS provided fewer annotated lipids, as well as lower false-positives and higher level of structural information (Tables 8-9 and Figure 31). In addition, while MS-DIAL separately processes DDA and DIA files, what requires subsequent merging of the results and is time consuming, any combination of the different modes of MS acquisition can be simultaneously processed using LipidMS.

Table 8. Summary of the lipids identified in ESI-. Total: total number of annotated lipids. Correct: lipids whose proposed annotation is correct based on the observed spectra. Class: specific subclass fragments are found, but only the total number of carbons and double bonds of the chains can be proposed based on the precursor ion. FA: the specific chain fragments that inform about the composition of fatty acyl chains are found. FA position: when the specific fatty acyl chain position can be elucidated based on chain fragments intensity ratios. Incorrect: lipids for which an incorrect annotation is provided, or non-lipidic features that are annotated as lipids. Unique: lipids that are annotated exclusively by one of the software packages. Missing: the lipids with confirmed lipid identity but are not annotated by one of the software packages.

| | Total | Correct | Class | FA | FA position | Incorrect | Unique | Missing |
|----------------|-------|-----------|-------|-----|-------------|-----------|--------|---------|
| LipidMS | 466 | 387 (83%) | 183 | 2 | 202 | 79 (17%) | 152 | 342 |
| MS-DIAL | 1249 | 580 (46%) | 223 | 217 | 140 | 669 (54%) | 345 | 92 |

Table 9. Summary of lipids identified in ESI+. Total: total number of annotated lipids. Correct: lipids whose proposed annotation is correct based on the observed spectra. Class: specific subclass fragments are found, but only the total number of carbons and double bonds of the chains can be proposed based on the precursor ion. FA: the specific chain fragments that inform about the composition of fatty acyl chains are found. FA position: when the specific fatty acyl chain position can be elucidated based on chain fragments intensity ratios. Incorrect: lipids for which an incorrect annotation is provided, or non-lipidic features that are annotated as lipids. Unique: lipids that are annotated exclusively by one of the software packages. Missing: the lipids with confirmed lipid identity but are not annotated by one of the software packages.

| | Total | Correct | Class | FA | FA position | Incorrect | Unique | Missing |
|----------------|-------|-----------|-------|-----|-------------|-----------|--------|---------|
| LipidMS | 495 | 445 (90%) | 95 | 22 | 328 | 50 (10%) | 140 | 297 |
| MS-DIAL | 1485 | 588 (40%) | 270 | 249 | 69 | 897 (60%) | 283 | 124 |

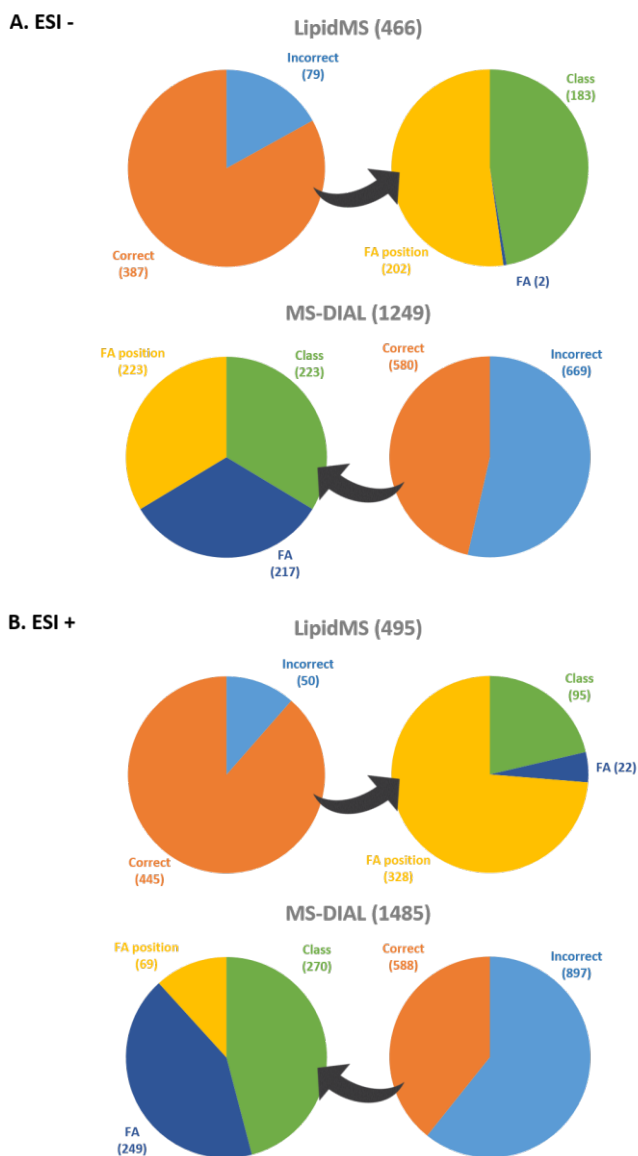


Figure 31. Summary of lipid annotations provided by LipidMS and MS-DIAL for the human serum pool. Levels of structural information provided: Class, the detected fragments allow to provide information only about the total number of carbons and double bonds and the lipid class; FA, the identity of the fatty acyl moieties can be identified; FA position, the actual position of the fatty acyl moieties within the lipid structure can be deduced.

4. Future improvements of LipidMS

In order to expand the lipidome coverage of LipidMS, future releases should incorporate new lipid classes such as oxidized and glycosylated lipids and a wider variety of fatty acyl chains and sphingoid bases (e.g., odd-chain fatty acids). In addition, it might be interesting to standardize LipidMS formats and workflows to make it compatible with other R packages such as those from the R for Mass Spectrometry Initiative (<https://www.rformassspectrometry.org/>). Additionally, most of the lipidomic studies only provide a static snapshot of the lipid profiles, lacking dynamic information about lipid metabolism such as building block sources, function of related enzymes and transporters or lipid-lipid interactions. For this reason, our next challenge is to adapt LipidMS to make it capable of analyzing DDA and DIA data when stable isotope tracers (e.g., ^{13}C) are used. The modelling of isotopic patterns of complex lipids such as phospholipids, glycerolipids or sphingolipids and their building blocks (e.g. polar head groups, fatty acyl chains, sphingoid bases) may allow the estimation of the turnover of the different blocks used to synthesize an individual lipid and to better understand the complex metabolic reactions in which lipids are involved in.

Chapter 2

FAMetA: a mass isotopologues-based tool for the comprehensive analysis of fatty acid metabolism

Introduction

Stable-isotope tracing combined with MS-based has been extensively used for interrogating FA metabolism. An example of a common experimental design for the study of FA biosynthesis using ^{13}C -tracers is shown in Figure 32. The total FA synthesis rate can be estimated by using D_2O , which labels FA through direct solvent incorporation and NADPH-mediated hydrogen transfer^{150,151}, while employing ^{13}C -labelled tracer nutrients (e.g., $\text{U-}^{13}\text{C}$ -glucose, $\text{U-}^{13}\text{C}$ -glutamine, $\text{U-}^{13}\text{C}$ -acetate, etc.) allows the total FA synthesis rate and the relative contribution of a given nutrient to be estimated¹⁵². The framework for FA synthesis data analysis using ^{13}C -labelled tracers and MS was initially set up by Isotopomer Spectral Analysis (ISA)¹⁵³ and Mass Isotopomer Distribution Analysis (MIDA)¹⁵⁴, which model FA synthesis following the incorporation of n 2-carbon units. In the ISA model, each isotopologue is modeled by an equation composed of the linear sum of two multinomial distributions representing the preexisting or imported fraction of the FA and the newly synthesized fraction. For each step in the synthesis process, a molecule of acetate (2 carbon) is added to the FA chain containing 0, 1 or 2 ^{13}C atoms with a

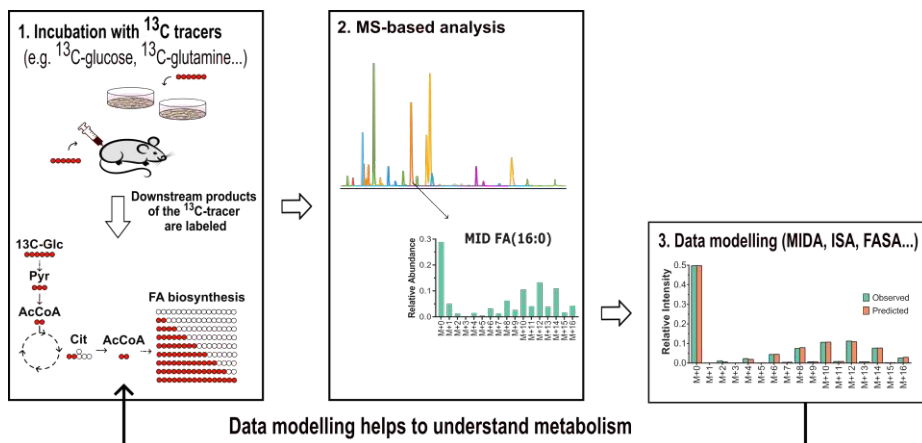


Figure 32. Classical isotope labelling experiment for FA analysis. After incubation with ^{13}C -tracers such as ^{13}C -glucose or ^{13}C -glutamine, isotopic label is incorporated to FA via acetyl-CoA. FA isotopologues distributions are obtained by LC-MS analysis and modelled using different algorithms based on multinomial distributions.

probability D_0 , D_1 or D_2 , respectively, depending on the contribution of the ^{13}C -tracer to the pool of acetate (D). When data is previously corrected for the natural abundance of ^{13}C , for FA up to 16 carbons, the model can be simplified to:

$$P(M + 0) = S * D_0^{x_0} + I \quad (\text{Equation 2})$$

$$P(M + X)_{1 \leq X \leq M} = S * \binom{M}{X} * \sum D_0^{x_0} * D_1^{x_1} * D_2^{x_2} \quad (\text{Equation 3})$$

, given that:

$$S + I = 1 \quad (\text{Equation 4})$$

$$D_0 + D_1 + D_2 = 1 \quad (\text{Equation 5})$$

$$x_0 + x_1 + x_2 = M/2 \quad (\text{Equation 6})$$

$$x_1 + 2 * x_2 = X \quad (\text{Equation 7})$$

M is the total number of carbons in the FA chain, which may contain from 0 to M ^{13}C atoms (X); S is the newly synthesized FA fraction and I is the preexisting or imported fraction, which only contributes to the $M+0$ abundance; and x_0 , x_1 and x_2 refer to the number of acetate molecules containing 0, 1 or 2 ^{13}C atoms with a probability D_0 , D_1 and D_2 , respectively. Unfortunately, these mass isotopologue modelling methods only provide information about the DNL of FA for which the contribution of elongation is minimal (i.e., FA of 14 or 16 carbons). ConvISA incorporated one elongation step to, thus, extending the analysis to 18-carbon FA¹⁵⁵, and

recently, Fatty Acid Source Analysis (FASA) included many elongation steps, which extend the FA species that can be properly modelled to 26C¹²⁹. In the FASA model, each isotopologue is modeled with a sum of several multinomial distributions which represent the preexisting or imported fraction of the FA, the fraction that is newly synthesized and those fractions that are imported and elongated n times (IE_n). For example, for FA of 20 carbons, four different fractions (which are modeled by a different multinomial distribution) are distinguished:

$$S + IE_1 + IE_2 + I = 1$$

(Equation 8)

, where S represents the fraction that comes from newly synthesized FA(16:0) and elongated twice; IE_i represents the fraction of imported FA(18:0) and elongated once, and IE_2 , the fraction of imported FA(16:0) and elongated twice. In addition, for FA from $n3$ and $n6$ series, which come from imported linoleic (FA(18:2) $n6$) and α -linolenic (FA(18:3) $n3$) acids, S is fixed to 0 and elongation is assumed from those 18-carbon species. However, FASA has some limitations as it assumes *de novo* synthesis up to 26-carbon FA (S) and it calculates multiple import-elongation terms, which does not accurately represent the actual biological process. Finally, a simple strategy for estimating the desaturation of FA(18:0) to FA(18:1) $n9$ has also been described by Kamphorst and colleagues¹⁵⁶. Yet this approach is based on the total labelling of precursor and product FA, and its application to the complete array of desaturations has not yet been explored. Despite these valuable advances, reliable FA elongation calculations are still to be fully addressed, whereas systematic desaturation estimations remain unresolved. Additionally, the above-mentioned algorithms have been developed for platforms that require computational skills and commercial software, thus, they are not readily accessible to the broad metabolism community.

Methodology

1. Chemicals and reagents

Solvents for sample processing and LC-MS analysis were isopropanol, isooctane, potassium hydroxide, formic acid and ammonium acetate obtained from Sigma-Aldrich/Fluka, and acetonitrile, methanol and heptane, from Fisher Scientific.

FA standards and internal standards, obtained from Avanti Polar Lipids, Sigma-Aldrich/Fluka, Larodan and Caiman Chemicals, were capric acid (FA(10:0)), lauric acid (FA(12:0)), myristic acid (FA(14:0)), deuterated myristic acid (FA(14:0)D27), myristoleic acid (FA(14:1)n5), pentadecanoic acid (FA(15:0)), palmitic acid (FA(16:0)), 11-hexadecanoic acid (FA(16:1)n5), palmitoleic acid (FA(16:1)n7), 7-hexadecenoic acid (FA(16:1)n9), sapienic acid (FA(16:1)n10), margaric acid (FA(17:0)), stearic acid (FA(18:0)), vaccenic acid (FA(18:1)n7), oleic acid (FA(18:1)n9), 6-octadecenoic acid (FA(18:1)n12), 5-octadecenoic acid (FA(18:1)n13), linoleic acid (FA(18:2)n6), 6,9-octadecadienoic acid (FA(18:2)n9), 5,8-octadecadienoic acid (FA(18:2)n10), alpha-linolenic acid (FA(18:3)n3), gamma-linolenic acid (FA(18:3)n6), nonadecanoic acid (FA(19:0)), arachidic acid (FA(20:0)), 13-eicosenoic acid (FA(20:1)n7), gondoic acid (FA(20:1)n9), 8-eicosenoic acid (FA(20:1)n12), 11,14-eicosadienoic acid (FA(20:2)n6), dihomo-alpha-linolenic acid (FA(20:3)n3), dihomo-gamma-linolenic acid (FA(20:3)n6), 5,8,11-eicosatrienoic acid (FA(20:2)n9), 8,11,14,17-eicosatetraenoic acid (FA(20:4)n3), arachidonic acid (FA(20:4)n6), eicosapentaenoic acid (FA(20:5)n3), behenic acid (FA(22:0)), erucic acid (FA(22:1)n9), docosadienoic acid (FA(22:2)n6), 10,13,16-docosatrienoic acid (FA(22:3)n6), 10,13,16,19-docosatetraenoic acid (FA(22:4)n3), adrenic acid (FA(22:4)n6), clupanodonic acid (FA(22:5)n3), 4,7,10,13,16-docosapentaenoic acid (FA(22:5)n6), cervonic acid (FA(22:6)n3), lignoceric acid (FA(24:0)), nervonic acid (FA(24:1)n9), 9,12,15,18-tetracosatetraenoic acid (FA(24:4)n6), 9,12,15,18,21-tetracosapentaenoic acid (FA(24:5)n3), 6,9,12,15,18-

tetracosapentaenoic acid (FA(24:5)n6), 6,9,12,15,18,21-tetracosahexaenoic acid (FA(24:6)n3), cerotic acid (FA(26:0)) and 1,2-dipalmitoyl-d62-sn-glycero-3-phosphocholine (PC(16:0/16:0)D62).

Commercial human serum was obtained from Sigma-Aldrich (reference P2918). RPMI 1640 with stable glutamine (ref L0498) and 100x streptomycin/penicillin solution (ref L0010) were obtained from Biowest. Fetal bovine serum (FBS, ref A3160502) and dialyzed FBS (ref 26400044) were obtained from Gibco. RPMI 1640 media without glucose, glutamine, and amino acids (ref R9010-01) were supplied by USBiological. U-¹³C-glucose, U-¹³C-glutamine and U-¹³C-glutamine were obtained from Cambridge Isotope Laboratories. FASN inhibitors (GSK2194069)¹⁵⁷ and FADS2 inhibitor (SC26196)¹⁵⁸ were purchased from Sigma-Aldrich; SCD inhibitor A93572^{159,160} was obtained from MedChemExpress. Antibodies anti-CD3 (ref BE0001-1) and anti-CD28 (ref BE0015-1) were provided by BioXCell. Recombinant IL-2 (ref 212-12) was obtained from Peprotech.

2. Cell lines and growth conditions for cell metabolism studies

2.1. Mouse naïve CD8⁺ T-cells

For all the experiments with CD8⁺ T-cells, 8-10-week-old female mice were used. 6-week-old wild-type C57BL/6 were purchased from Charles River Laboratories. Mice were left in a normal light cycle (08:00–20:00h) and had free access to water and a standard chow diet. Animals were housed in the Health Research Institute–Hospital La Fe Valencia facilities. Mouse studies followed the protocols approved by the Health Research Institute–Hospital La Fe Valencia Ethics and Animal Care and Use Committee (Protocol number 2020/VSC/PEA/0048). To isolate naïve CD8⁺ T-cells, spleens were harvested. Single-cell suspensions were prepared by manual disruption and passage through a 70µm cell strainer in PBS supplemented with 0.5% BSA and 2mM EDTA. After RBC lysis, naïve CD8⁺ T-cells were purified by magnetic bead separation using commercially available kits following manufacturers' instructions (naïve CD8a⁺ T-Cell Isolation Kit, mouse, Miltenyi Biotec Inc.)¹⁶¹.

Cells were cultured in complete RPMI media (RPMI 1640 supplemented with 10% FBS, 100U/mL penicillin, 100µg/mL streptomycin, 55µM 2-mercaptoethanol). Naïve T-cells were stimulated for 48h with plate-bound anti-CD3 (10µg/mL) and anti-CD28 (5µg/mL) in complete RPMI media supplemented with recombinant IL-2 (100U/mL). All the experiments on 'active' T-cells were performed on day 4–5 postactivation¹⁶¹.

For FA metabolism studies, isotopically-labelled media were prepared from glucose, glutamine and amino acids-free RPMI media, and were supplemented with 10% dialyzed FBS, 100U/ml penicillin, 100µg/ml streptomycin, recombinant IL-2 (100U/mL) and 55µM 2-

mercaptoethanol. U-¹³C-glucose and U-¹³C-glutamine were added at the normal concentration found in RPMI 1640 media. U-¹³C-acetate was added at 100µM. U-¹³C-lactate was added at 11mM¹⁶¹⁻¹⁶⁵. The CD8⁺ T-cells were seeded at 0.8 x 10⁶cells/mL and incubated for 72h with labelled media (and inhibitors). At 24h and 48h, cells were counted using the Countess II automated cell counter (Thermo Fischer Scientific) and density was adjusted to 0.8x10⁶cells/mL with complete fresh labelled media (and inhibitors). At 72h, the final cell density was determined. Then, cells were transferred to 1.5mL Eppendorf tubes and pelleted (500g, 3min). Media were removed. Cells were washed once with cold PBS 1x, resuspended in 500µL of cold PBS 1x and stored at -80°C^{161,163}.

2.2. A549 cell line

The KRAS-mutant non-small cell lung cancer (NSCLC) cell line A549 was originally obtained from ATCC. The A549 cells were maintained in RPMI-1640 media supplemented with 10% FBS, 100U/mL penicillin and 100µg/mL streptomycin, and were routinely screened for mycoplasma contamination. Identity was confirmed by STR sequencing. For FA metabolism studies, isotopically-labelled media were prepared from glucose, glutamine and amino acids-free RPMI media, and were supplemented with 10% dialyzed FBS, 100U/ml penicillin and 100µg/ml streptomycin. The NSCLC cell line A549, cells were seeded at 7x10⁴ cells/well in 6-well plates. After 24h, media were replaced with labelled media (and inhibitors). Cells were incubated for 48-72h until 80-90% confluence, the media was replaced with fresh media (and inhibitors) every 24h. At the end of the incubation, media were removed, cells were washed once with cold PBS 1x, scraped with 500 µL of cold PBS 1x, transferred to 1.5mL Eppendorf tubes and stored at -80°C¹⁶³.

3. Sample preparation

3.1. Preparation of standards

Individual stocks for each compound were prepared at 2mg/mL following the recommendations of the suppliers. Working solutions for FA standards were prepared at 1µg/mL in methanol/water/acetonitrile (25:25:50). A mixed solution containing all the fatty acid standards was prepared in methanol/water/acetonitrile (25:25:50) at 30µg/mL each and subsequently diluted at the suitable final concentrations.

3.2. Saponification and extraction of total FA from cells

To analyze the total FA, 450µL of cell suspension were transferred to a glass vial, and 1000µL of a 9:1 methanol:hydroxide potassium (3M in H₂O) solution containing PC(16:0/16:0)D62 at 3ppm were added. Saponification was performed for 1h at 80°C in a water bath. After saponification, samples were cooled on ice and acidified by adding 100µL of formic acid. FA were extracted with 2mL of heptane:isooctane (1:1) (2x), dried in a nitrogen flow, resuspended in 200µL of mobile phase A containing FA(14:0)D27 at 1ppm and transferred to a glass HPLC vial¹²⁷.

4. LC-MS analysis

4.1. Instrumentation

All the experiments conducted for fatty acid analysis were performed using a Q-orbitrap mass spectrometer (Q-Exactive, Thermo-Fisher Scientific) coupled to RP chromatography through an ESI source.

4.2. Chromatographic separation

Liquid chromatography separation was performed in a Cortecs C18 column (2.1mm × 150mm, 1.6µm particle size; Waters). Solvent (A) was 2.5mM ammonium acetate in 60:40 water:methanol. Solvent (B) was 2.5mM ammonium acetate in 95:5 acetonitrile:isopropanol. The flow rate was 0.3mL/min, the column temperature was 45°C, the autosampler temperature was 5°C and the injection volume was 5µL. The liquid chromatography gradient was: 0 min, 45% B; 0.5 min, 45% B; 19min, 55% B; 23min, 99% B; 34min, 99% B. Between injections, the column was washed for 2min with 50:50 acetonitrile:isopropanol before being equilibrated to the initial conditions.

4.3. MS detection

The Q-Exactive instrument operated in the ESI- with the following conditions: the sheath gas flow rate was 60; the auxiliary gas flow rate was 20; the spray voltage was 1.50kV; the capillary temperature was 300°C; the S-lens RF-level was 75; and the auxiliary gas heater temperature was 300°C. Data was acquired in centroid mode using the full scan method with the following parameters: resolution 140000, AGC target 1000000, maximum IT 100ms, scan range from m/z 100 to m/z 450 and data type centroid.

5. Data processing and analysis for fatty acid analysis

For the performance evaluation of FAMetA, all data was processed using FAMetA (using LipidMS for data pre-processing) except for the comparison between FAMetA and FASA, for which both software were employed. Scripts, parameters and files used for all experiments are available at Zenodo (accession number 6511248), but general parameters employed for FAMetA processing are described below:

- FAMetA: data pre-processing was performed with the above LipidMS parameters, and then, FAMetA was employed for the FA metabolic analysis:

- Peak-picking parameters:
 - dmzagglom: 15
 - drtagglom: 200
 - drtclust: 100
 - minpeak: 8
 - drtgap: 5
 - drtminpeak: 8
 - drtmaxpeak: 30
 - recurs: 10
 - sb: 5
 - sn 5
 - minint: 100000
 - weight: 2
 - dmzIso: 5
 - drtIso: 5

- Batch processing parameters (alignment and grouping):
 - dmzalign: 10
 - drtalign: 60
 - span: 0.2
 - minsamplesfracalign: 0.50
 - dmzgroup: 10
 - drtagglomgroup: 50
 - drtgroup: 10
 - minsamplesfracgroup: 0.20
- FA annotation:
 - dmz: 5
 - adduct: M-H
- Isotope annotation:
 - dmzIso: 10
 - coelCutoffIso: 0.2
- Data correction:
 - correct13C: TRUE
 - resolution: 140000
 - purity13C: 0.99
 - externalnormalization: (keep empty)
- Synthesis analysis: in case of experiments using inhibitors, S may decrease below the confidence interval and D_2 parameter can be misestimated. To avoid this problem D_2 values were fixed using the control group (misestimated values were replaced with the mean value of the control group for palmitic acid).
 - R2Thr: 0.95
 - maxiter: 1000
 - maxconvergence: 100
 - startpoints: 5

- propagated: TRUE
- Elongation and desaturation analysis:
 - R2Thr: 0.95
 - maxiter: 10000
 - maxconvergence: 100
 - startpoints: 5
 - D2Thr: 0.1
 - SEThr: 0.05

Results and Discussion

1. FAMetA overview

The use of ^{13}C -tracers and MS is the gold standard method for the analysis of the FA metabolism. This method relies on the incorporation of ^{13}C atoms, through acetyl-CoA, to FA during synthesis and elongation reactions and the subsequent analysis of their mass isotopologue distributions (MID). Despite several algorithms and tools have been developed in order to extract information about FA metabolism by modelling these MID, they still provide a limited and difficult-to-interpret snapshot of FA metabolism. Most of these methods only provide information about *de novo* lipogenesis (DNL) for FA up to 16 or 18C^{155} or do not reflect the actual biological steps of the elongation processes¹²⁹. In addition, desaturation is not considered for the complete FA network¹⁵⁶. In order to overcome these limitations and motivated by an increasing body of evidence that suggest the key role of FA in cancer we decided to develop a tool that use ^{13}C mass isotopologue profiles to estimate most of the biosynthetic reactions involved in FA metabolism: DNL, elongation, desaturation and FA import.

FAMetA is an R package (<https://CRAN.R-project.org/package=FAMetA>) and a web-based platform (<https://www.fameta.es>) that relies on the MID obtained from ^{13}C -labelled FA analyzed by LC-MS or GC-MS to estimate import (I), synthesis of FA up to 16C (S), fractional contribution of the ^{13}C -tracer (D_0 , D_1 , D_2 , which represent the acetyl-CoA fraction with 0, 1 or 2 atoms of ^{13}C , respectively), elongation (E) and desaturation (Δ) parameters for the expected biosynthetic network of FA up to 26C^{155} (Figure 33). FAMetA has been designed to model the actual reactions of the biosynthetic network so that each step of the reactions is represented as a unique parameter (Figure 34).

saturated fatty acids

| | | | | | | | | | | | | |
|---|------|---|----------|--|------------|--|------------|--|------------|--|------------|--|
| 14:0 S D \emptyset I _{14:0} | FASN | 16:0 S D \emptyset I _{16:0} | ELOVL3,6 | 18:0 E ₁ I _{18:0} | ELOVL1,3,7 | 20:0 E ₂ I _{20:0} | ELOVL1,3,7 | 22:0 E ₃ I _{22:0} | ELOVL1,3,7 | 24:0 E ₄ I _{24:0} | ELOVL1,3,7 | 26:0 E ₅ I _{24:0} |
|---|------|---|----------|--|------------|--|------------|--|------------|--|------------|--|

n5 series

| | |
|--|------------|
| 14:0 S D \emptyset I _{14:0} | |
| SCD1 | |
| 14:1n5 Δ I _{14:1n5} | ELOVL1,3,7 |
| 16:1n5 E ₀ I _{16:1n5} | |

n9 series

| | | |
|---|-------------------|--|
| 16:0 S D \emptyset I _{16:0} | ELOVL3,6 | 18:0 E ₁ I _{18:0} |
| | SCD1,5 | |
| 16:1n9 S I _{16:1n9} | β oxidation | 18:1n9 Δ I _{18:1n9} |
| | ELOVL1,3,7 | 20:1n9 E ₂ I _{20:1n9} |
| | FADS2 | 22:1n9 E ₃ I _{22:1n9} |
| | FADS2* | 24:1n9 E ₄ I _{24:1n9} |
| 18:2n9 Δ I _{18:2n9} | ELOVL5 | 20:2n9 Δ I _{20:2n9} |
| | FADS1 | 22:2n9 E ₃ I _{22:2n9} |
| | FADS1 | 24:2n9 E ₄ I _{24:2n9} |
| | ELOVL5 | 20:3n9 Δ I _{20:3n9} |
| | ELOVL2,5 | 22:3n9 E ₃ I _{22:3n9} |
| | ELOVL2,5 | 24:3n9 E ₄ I _{24:3n9} |

n7 series

| | |
|--|------------|
| 16:0 S D \emptyset I _{16:0} | |
| SCD1 | |
| 16:1n7 Δ I _{16:1n7} | ELOVL1,3,7 |
| 18:1n7 E ₁ I _{18:1n7} | ELOVL1,3,7 |
| 20:1n7 E ₂ I _{20:1n7} | |

n10 series

| | |
|--|------------|
| 16:0 S D \emptyset I _{16:0} | |
| FADS2 | |
| 16:1n10 Δ I _{16:1n10} | ELOVL1,3,7 |
| 18:1n10 E ₁ I _{18:1n10} | |

n6 series

| | | | | | | | | |
|--|--------|--|----------|--|----------|--|----------|--|
| 16:2n6 | | 18:2n6 I _{18:2n6} | ELOVL5 | 20:2n6 E ₂ I _{20:2n6} | ELOVL2,5 | 22:2n6 E ₃ I _{22:2n6} | ELOVL2,4 | 24:2n6 E ₄ I _{24:2n6} |
| | FADS2* | | FADS2 | | | | | |
| 18:3n6 Δ I _{18:3n6} | | 20:3n6 E ₂ I _{20:3n6} | ELOVL5 | 22:3n6 E ₃ I _{22:3n6} | ELOVL2,5 | 24:3n6 E ₄ I _{24:3n6} | ELOVL2,4 | |
| | FADS1 | | FADS1 | | | | | |
| | | 20:4n6 Δ I _{20:4n6} | ELOVL2,5 | 22:4n6 E ₃ I _{22:4n6} | ELOVL2,5 | 24:4n6 E ₄ I _{24:4n6} | ELOVL2,4 | |
| | | | | | | | | FADS2 |
| | | 22:5n6 E ₃ I _{22:5n6} | | 24:5n6 Δ I _{24:5n6} | | | | |
| | | | | | | | | β oxidation |

n3 series

| | | | | | | | | |
|---------------|--------|--|----------|--|----------|--|----------|--|
| 16:3n3 | | 18:3n3 I _{18:3n3} | ELOVL5 | 20:3n3 E ₂ I _{20:3n3} | ELOVL2,5 | 22:3n3 E ₃ I _{22:3n3} | ELOVL2,4 | 24:3n3 E ₄ I _{24:3n3} |
| | FADS2* | | FADS2 | | | | | |
| | | 18:4n3 Δ I _{18:4n3} | ELOVL5 | 20:4n3 E ₂ I _{20:4n3} | ELOVL2,5 | 22:4n3 E ₃ I _{22:4n3} | ELOVL2,4 | 24:4n3 E ₄ I _{24:4n3} |
| | | | FADS1 | | | | | |
| | | 20:5n3 Δ I _{20:5n3} | ELOVL2,5 | 22:5n3 E ₃ I _{22:5n3} | ELOVL2,5 | 24:5n3 E ₄ I _{24:5n3} | ELOVL2,4 | |
| | | | | | | | | FADS2 |
| | | 22:6n3 E ₃ I _{22:6n3} | | 24:6n3 Δ I _{24:6n3} | | | | |
| | | | | | | | | β oxidation |

Figure 33. FA metabolism network. Summary of the FA interconversions covered by FAMetA and the parameters that can be estimated for each one. In red, the FA for which no parameter can be estimated because they are either solely imported or result from desaturation being performed on them. Horizontal transitions denote elongations and vertical transitions depict desaturations. The responsible enzymes are indicated in both cases. We assume DNL up to FA(16:0), although the calculation of the DNL parameters can be estimated for both FA(14:0) and FA(16:0). For the transformations of FA(18:2)n6 into FA(20:3)n6 and FA(18:3)n3 to FA(20:4)n3, the preferred route is desaturation, followed by elongation. The asterisk denotes a secondary route.

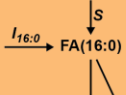
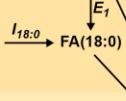
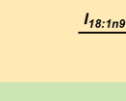
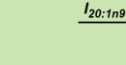
| Synthesis Route | Calculation of FA sources | Reported Endogenous Synthesis | Calculated Parameters |
|--|---|-------------------------------|--|
|  <p>$I_{16:0} \rightarrow$ FA(16:0) \xrightarrow{S} FA(18:0)</p> | $I_{16:0} + S_{16:0}$ | S | $S D_0 D_1 D_2 \Phi I_{16:0}$ |
|  <p>$I_{18:0} \rightarrow$ FA(18:0) $\xrightarrow{E_1}$ FA(18:1n9)</p> | $I_{18:0} + E_1 =$ $= I_{18:0} + E_1' (I_{16:0} + S_{16:0})$ | E_1 | $E_1 S I_{18:0}$ |
|  <p>$I_{18:1n9} \rightarrow$ FA(18:1n9) $\xrightarrow{\Delta}$ FA(20:1n9)</p> | $I_{18:1n9} + \Delta =$ $= I_{18:1n9} + \Delta' (I_{18:0} + E_1' (I_{16:0} + S_{16:0})) =$ $= I_{18:1n9} + E_1' (I_{16:0} + S_{16:0})$ $E_1' = \Delta' E_1$ | Δ | $E_1' S I_{18:1n9}$ <i>indirectly: $\Delta = E_1'/E_1$</i> |
|  <p>$I_{20:1n9} \rightarrow$ FA(20:1n9)</p> | $I_{20:1n9} + E_2 =$ $= I_{20:1n9} + E_2' (I_{18:1n9} + \Delta' (I_{18:0} + E_1' (I_{16:0} + S_{16:0}))) =$ $= I_{20:1n9} + E_2' (I_{18:1n9} + E_1' (I_{16:0} + S_{16:0}))$ $E_1' = \Delta' E_1$ | E_2 | $E_2 E_1' S I_{20:1n9}$ |

Figure 34. Example of the FAMetA calculations for FA(16:0) to FA(20:1)n9. A detailed description of the calculation of FA sources, reported endogenous synthesis and the parameters calculated for the FA FA(16:0), FA(18:0), FA(18:1)n9 and FA(20:1)n9.

For FA up to 16C, DNL is modelled by using quasi-multinomial distributions, which allow the estimation of the following parameters: I , S and D_0 , D_1 , D_2 ; apart from accounting for data overdispersion (Φ). The equations employed to fit the experimental isotopologue distribution are equivalent to those employed by the ISA algorithm^{153,154} when both the Φ parameter is set to 0 and data is corrected by the natural abundance of ^{13}C . For FA of 18 to 26C, apart from the parameters S and I , up to five elongation terms (E_n , $n=1$ for 18C to $n=5$ for 26C FA) are estimated. Each elongation term represents the direct estimation of the fraction that comes from the elongation of the total pool of the precursor FA (Figure 34). Compared to previous tools (i.e., FASA, where the synthesis of a FA longer than 16C is described as DNL up to the total length and multiple import-elongation terms¹²⁹), the way in which elongations are calculated by FAMetA better reflects how FA are elongated within the cells, which permits the straightforward

biological interpretation of the reported elongation parameters. For FA that result from the direct desaturation of a precursor FA, Δ is indirectly estimated based on the calculated synthesis parameters of the precursor (S or E) and the FA of interest (S' or E') (i.e. $\Delta = S'/S$ or $\Delta = E'/E$) (Figure 34). The strategy proposed here is based on the simple approach previously described by Kamphorst *et al.*^{126,156}, where desaturation of FA(18:1)n9 is calculated based on the total labelling found in FA(18:0) and FA(18:1)n9. In FAMetA, we extend this strategy to the complete set of desaturations within the FA metabolic network and refine the calculation by using an approach that uses the estimated synthesis parameter of interest instead of total labelling.

As in previous tools (i.e., ISA, ConvISA and FASA^{129,153-155,166,167}) the *de novo* synthesis parameters (S , E , Δ) are time-dependent. Therefore, at any given time, such parameters correspond to the fraction of a particular FA that has been *de novo* synthesized up-to-the moment of the sampling, and it corresponds to the actual portion of FA that comes from *de novo* synthesis only if the steady state has been reached. Accordingly, the import term ($I=1-S$ or $I=1-E_n$) accounts for both import and pre-existing FA at any given time and to the actual fraction that is acquired from the exogenous pool when the steady state has been reached. The conditions of metabolic and isotopic steady states are only achieved, or can be closely approximated, if the cells are cultured during a long-enough time to ensure that the pre-existing FA pools can be diluted out while ensuring a nutrient supply that maintains relatively stable concentrations^{129,168}. Finally, the FAMetA's workflow includes all the functionalities needed, for data processing, group-based comparisons and graphical outputs, which facilitates the interpretation of results (Figure 35).

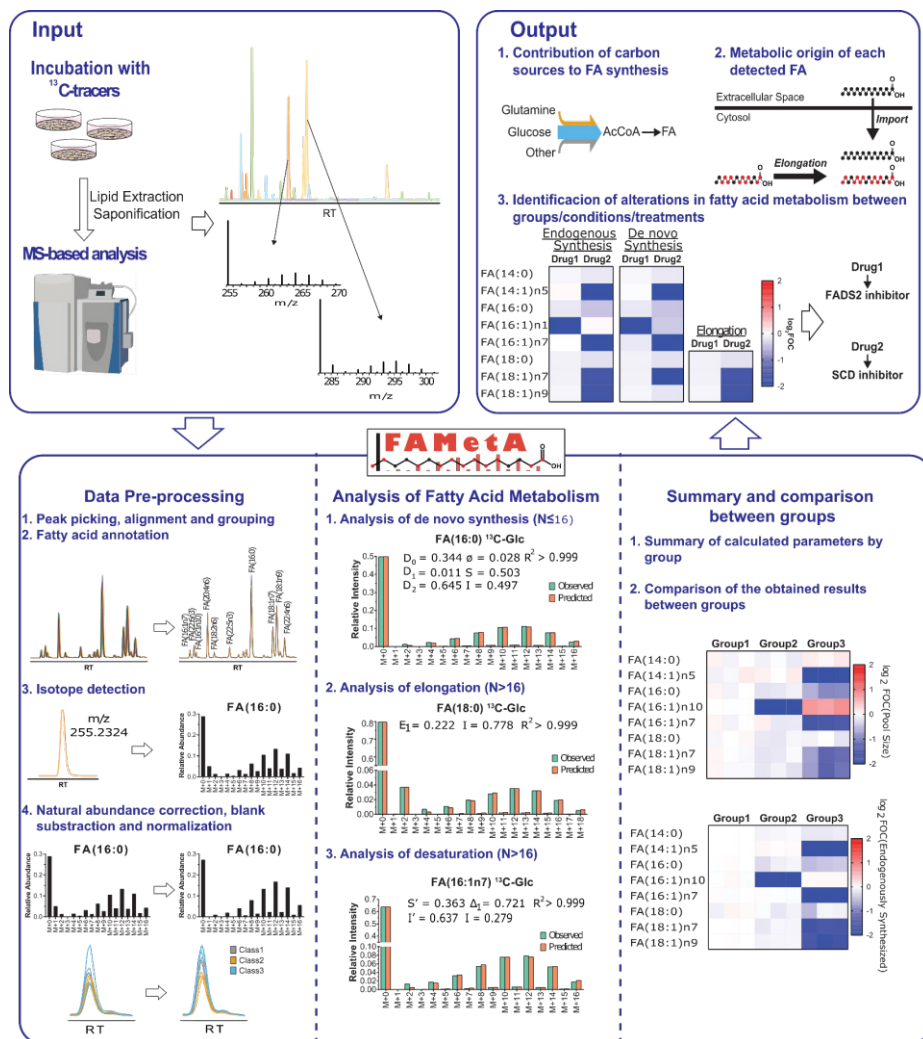


Figure 35. FAMeTA overview. FAMeTA is an R package and a web-based platform for the estimation of FA metabolism based on mass isotopologue data, generated after incubation with suitable ^{13}C -tracers and based on LC-MS or GC-MS analysis of fatty acid extracts. The main steps within FAMeTA workflow include data processing, estimation of metabolism parameters for each sample and fatty acid based on the obtained mass isotopologue distributions and finally the combination of those individual results to provide a global view of fatty acid metabolism network for each condition of interest and the comparison between them. The most relevant biologically relevant outputs that can be obtained, depending on the experimental design include the fractional contribution of each tested carbon source, the detailed description of the metabolic origin of each detected fatty acid and the elucidation of alterations in fatty acid metabolism between conditions of interest.

2. Features and implementation

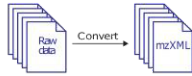
2.1. FAMetA workflow

The FAMetA workflow (Figures 35-37) starts by loading the raw MS data files in the mzXML format, which can be obtained with any MS file converter, such as msConvert from ProteoWizard¹⁶⁹, and a csv file containing the required metadata (sample name, acquisition mode, sample group, or class, and any additional information like external measures for normalization) (Figure 36, steps 1-2). First, data processing can be performed in the R environment or through the web-based application using our suggested workflow, which combines functions from FAMetA and our previously developed R-package LipidMS^{150,151} (Figure 36, steps 2-5). LipidMS is called for the first processing step, which runs peak-picking, alignment and grouping through the functions *batchdataProcessing*, *alignmsbatch* and *groupmsbatch* (Figure 36, step 2). Then FAMetA is called, and *annotateFA* and *curateFAannotations* functions are used to identify unique FA isomers. Automatic FA annotations can be exported to a csv file and modified by removing rows of unwanted FA, modifying the initial and end retention times, or adding new rows with missing compounds. Unique compound names with nomenclature “FA(16:1)n7”, where n7 (omega-7) indicates the last double-bond position, are required to differentiate FA isomers. For any unknown positions, letters x, y and z are could be used (i.e., FA(16:1)nx). Internal standards for later normalization can be also added in a new row at this point by indicating IS in the compound name column (Figure 36, step 3). Once all the FA of interest have been correctly identified, FA isotopes can be extracted using the *searchFAisotopes* function (Figure 36, step 4). Finally, data can be corrected and normalized using the *dataCorrection* function, which

OPTION A

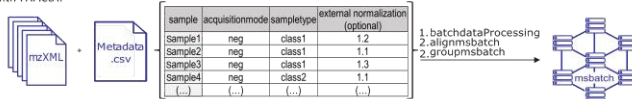
1. Data conversion (MSConvert)

Convert raw data files to a format compatible with the pre-processing software. Any file converter such as MSConvert from Proteowizard may be used. In the case of LipidMS (R-package with an output directly compatible with FAMeTA), mzXML format must be used.



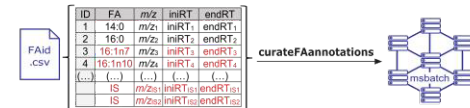
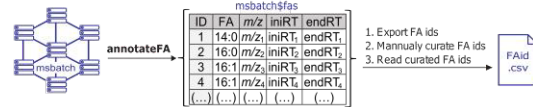
2. Data pre-processing (LipidMS)

We propose LipidMS R-package to perform peak-picking, alignment and grouping. It will return an msbatch object compatible with FAMeTA.



3. Fatty acid annotation and manual curation

Annotate FA based on *m/z*. Annotations have to be manually curated to identify isomers. During manual curation missing FA or internal standards can be added and parameters modified.



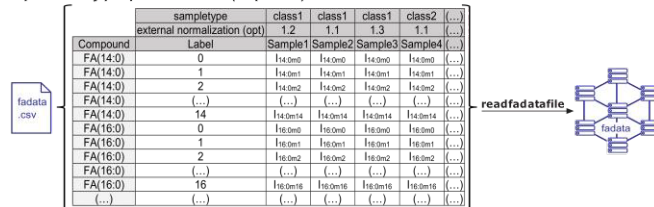
4. Isotope detection

Search for ¹³C isotopes based on *m/z* and peak shape correlation.



OPTION B

Import already pre-processed data (steps 1 - 4)



5. Natural abundance correction and normalization

Correct for natural abundance of ¹³C isotopes using the Accucor algorithm and, if available, IS-based normalization, blank subtraction and external factor normalization (e.g. protein, cell number...)

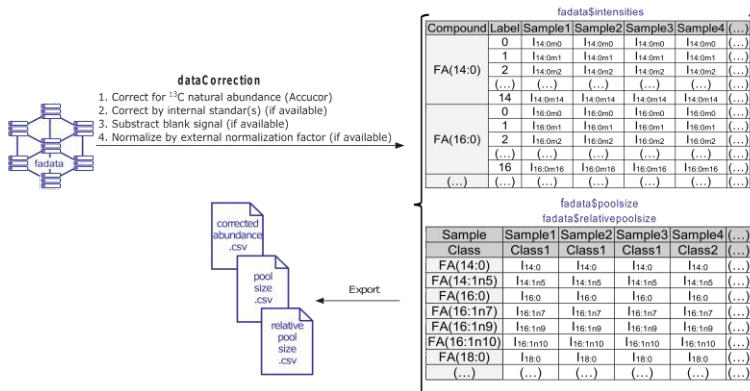


Figure 36. Detailed workflow for data pre-processing. Data pre-processing can be performed with a combination of our developed in-house R package LipidMS and FAMetA (Option A), or using any other suitable pre-processing tool (Option B). For data pre-processing with LipidMS, raw data files must firstly be converted into mzXML. LipidMS uses raw data files in the mzXML format and a csv metadata file as input to cover peak-peaking, alignment, grouping and peak filling. Output is a *msbatch* object that can be directly used by FAMetA to perform other pre-processing steps, including FA annotation and isotope detection. Output is a *fadata* object that can be used to conduct the final pre-processing step, which is natural abundance correction and normalisation. *Italics* depict the functions that can be executed in LipidMS or FAMetA.

runs four different steps (all of which are optional): data correction for natural ^{13}C abundance using the *Accucor* algorithm¹⁷⁰; data normalization with internal standards; blank subtraction; and external normalization (Figure 36, step 5). Alternatively, the external data processed by other available software/tools can be loaded at this point of the workflow or before the data correction and normalization steps. Then, the actual FA metabolism analysis can be performed by sequentially running the *synthesisAnalysis*, *elongationAnalysis* and *desaturationAnalysis* functions (Figure 37, steps 1-3). The first two functions model isotopologue distributions by non-linear regression with many initial values^{171,172} to ensure that the best fits are found. By default, a maximum of 1,000 iterations for synthesis and 10,000 for elongation are performed for each set of initial values to fit the isotopologue distributions (*maxiter* parameter) or until the model has converged 100 times (*maxconvergence* parameter). If no results are obtained or parameters come close to the limits of the confidence intervals, these parameters can be increased to improve the results. The third function employs the previous results to estimate the desaturation values. Finally, the summarized results tables and heatmaps are obtained using the *summarizeResults* function to export and explore the results (Figure 37, step 4).

1. Analysis of de novo synthesis (carbon number ≤16)

Estimate fraction of newly synthesized FA and contribution of the ¹³C tracer to acetate pool using a quasi-multinomial distribution.

fadata synthesis results

fadata → synthesisAnalysis → fadata

| Compound | Sample | Class | D ₁ | D ₂ | D ₃ | Φ | S | I | E | Δ | I |
|----------|---------|--------|-----------------|-----------------|-----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| FA(14:0) | Sample1 | Class1 | D ₁₁ | D ₁₂ | D ₁₃ | Φ ₁ | S ₁ | I ₁ | E ₁ | Δ ₁ | I ₁ |
| | Sample2 | Class1 | D ₂₁ | D ₂₂ | D ₂₃ | Φ ₂ | S ₂ | I ₂ | E ₂ | Δ ₂ | I ₂ |
| | Sample3 | Class1 | D ₃₁ | D ₃₂ | D ₃₃ | Φ ₃ | S ₃ | I ₃ | E ₃ | Δ ₃ | I ₃ |
| | Sample4 | Class2 | D ₄₁ | D ₄₂ | D ₄₃ | Φ ₄ | S ₄ | I ₄ | E ₄ | Δ ₄ | I ₄ |

fadata synthesis predicted values

| Compound | Label | Sample1 | Sample2 | Sample3 | Sample4 | (...) |
|----------|-------|---------|---------|---------|---------|-------|
| FA(14:0) | 0 | Isotop | Isotop | Isotop | Isotop | (...) |
| | 1 | Isotop | Isotop | Isotop | Isotop | (...) |
| | 2 | Isotop | Isotop | Isotop | Isotop | (...) |
| | (...) | (...) | (...) | (...) | (...) | (...) |

fadata synthesis plots \$FA (xxx)

Export → pred values DNS .csv, results DNS .csv, results DNS .pdf

2. Analysis of elongation

Estimate fraction of elongated FA.

fadata elongation results

fadata → elongationAnalysis → fadata

| Compound | Sample | Class | S ₁ | E ₁ | E ₂ | E ₃ | E ₄ | I |
|----------|---------|--------|-----------------|-----------------|-----------------|-----------------|-----------------|----------------|
| FA(18:0) | Sample1 | Class1 | S ₁₁ | E ₁₁ | E ₁₂ | E ₁₃ | E ₁₄ | I ₁ |
| | Sample2 | Class1 | S ₂₁ | E ₂₁ | E ₂₂ | E ₂₃ | E ₂₄ | I ₂ |
| | Sample3 | Class1 | S ₃₁ | E ₃₁ | E ₃₂ | E ₃₃ | E ₃₄ | I ₃ |
| | Sample4 | Class2 | S ₄₁ | E ₄₁ | E ₄₂ | E ₄₃ | E ₄₄ | I ₄ |

fadata elongation predicted values

| Compound | Label | Sample1 | Sample2 | Sample3 | Sample4 | (...) |
|----------|-------|---------|---------|---------|---------|-------|
| FA(18:0) | 0 | Isotop | Isotop | Isotop | Isotop | (...) |
| | 1 | Isotop | Isotop | Isotop | Isotop | (...) |
| | 2 | Isotop | Isotop | Isotop | Isotop | (...) |
| | (...) | (...) | (...) | (...) | (...) | (...) |

fadata elongation plots \$FA (xxx)

Export → pred values E .csv, results E .csv, results E .pdf

3. Analysis of desaturation

Estimate the fraction of desaturated FA based on previous results for synthesis and elongation analysis.

fadata desaturation results

fadata → desaturationAnalysis → fadata

| Compound | Sample | Class | Δ | I |
|------------|---------|--------|----------------|----------------|
| FA(16:1n7) | Sample1 | Class1 | Δ ₁ | I ₁ |
| | Sample2 | Class1 | Δ ₂ | I ₂ |
| | Sample3 | Class1 | Δ ₃ | I ₃ |
| | Sample4 | Class2 | Δ ₄ | I ₄ |

Export → results Δ .csv

4. Summary and graphical output

fadata results summary

| Class | D ₁ | | D ₂ | | D ₃ | | Φ | | S | | I | | E | | Δ | | I | |
|--------------|-----------------|------------------|-----------------|------------------|-----------------|------------------|----------------|------------------|----------------|------------------|----------------|------------------|----------------|------------------|----------------|------------------|-------|-------|
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| Class1 (ref) | D ₁₁ | D _{1SD} | D ₁₁ | D _{1SD} | D ₁₁ | D _{1SD} | Φ ₁ | Φ _{1SD} | S ₁ | S _{1SD} | I ₁ | I _{1SD} | E ₁ | E _{1SD} | I ₁ | I _{1SD} | (...) | (...) |
| Class2 | D ₂₁ | D _{2SD} | D ₂₁ | D _{2SD} | D ₂₁ | D _{2SD} | Φ ₂ | Φ _{2SD} | S ₂ | S _{2SD} | I ₂ | I _{2SD} | E ₂ | E _{2SD} | I ₂ | I _{2SD} | (...) | (...) |
| Class3 | D ₃₁ | D _{3SD} | D ₃₁ | D _{3SD} | D ₃₁ | D _{3SD} | Φ ₃ | Φ _{3SD} | S ₃ | S _{3SD} | I ₃ | I _{3SD} | E ₃ | E _{3SD} | I ₃ | I _{3SD} | (...) | (...) |

fadata heatmap \$relativepoolsize\$log2FOC

fadata heatmap \$synthesized\$log2FOC

fadata results results

| Compound | Sample | Class | D ₁ | D ₂ | D ₃ | Φ | S | I | E | Δ | I |
|----------|---------|--------|-----------------|-----------------|-----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| FA(14:0) | Sample1 | Class1 | D ₁₁ | D ₁₂ | D ₁₃ | Φ ₁ | S ₁ | I ₁ | E ₁ | Δ ₁ | I ₁ |
| | Sample2 | Class1 | D ₂₁ | D ₂₂ | D ₂₃ | Φ ₂ | S ₂ | I ₂ | E ₂ | Δ ₂ | I ₂ |
| | Sample3 | Class1 | D ₃₁ | D ₃₂ | D ₃₃ | Φ ₃ | S ₃ | I ₃ | E ₃ | Δ ₃ | I ₃ |
| | Sample4 | Class2 | D ₄₁ | D ₄₂ | D ₄₃ | Φ ₄ | S ₄ | I ₄ | E ₄ | Δ ₄ | I ₄ |

Export → results .csv, summary .csv, summary .pdf

fadata heatmap \$poolsize\$raw
fadata heatmap \$poolsize\$zscore
fadata heatmap \$poolsize\$log2FOC
adadata heatmap \$relativepoolsize\$raw
fadata heatmap \$relativepoolsize\$zscore
fadata heatmap \$synthesized\$raw

Figure 37. Detailed FAMetA workflow and output. Starting with the *fadata* object generated during data processing (Figure 36), FAMetA sequentially performs the analysis of DNS (*synthesisAnalysis* function), elongation (*elongationAnalysis* function) and desaturation (*desaturationAnalysis* function). The results for each step can be exported or a summary of all the calculated parameters and a group-based comparison can be obtained by executing the function *summarizeResults*.

2.2. Implementation of the quasi-multinomial distribution

MID of FA usually show an overdispersion which is not properly modelled by multinomial distributions. This overdispersion can be attributed to different factors such as cellular heterogeneity, time-dependent variations that result from changes in nutrient availability or differences between the various intracellular FA pools (e.g., differences between lipid classes or between FA/lipids located in different organelles). All these factors contribute to a narrowing or widening of the expected distributions if FA synthesis would fit exactly to a multinomial distribution (Figure 38). To address this issue, we decided to implement quasi-multinomial modelling instead of the formerly used multinomial modelling^{126-128,144-147}, which provides a ϕ parameter that accounts for data overdispersion. As shown in Figure 38, where multinomial and quasi-multinomial distributions have been used to fit different experimental FA MID obtained from literature^{129,155,167}, the later improves data fitting. The residuals obtained for quasi-multinomial distributions are smaller than those obtained for multinomial distributions, what has been confirmed by a log-likelihood ratio test and right-tailed chi-square distribution. Despite the better fit provided by quasi-multinomial adjustment, no significant differences in the calculated values for the DNL parameters have been observed.

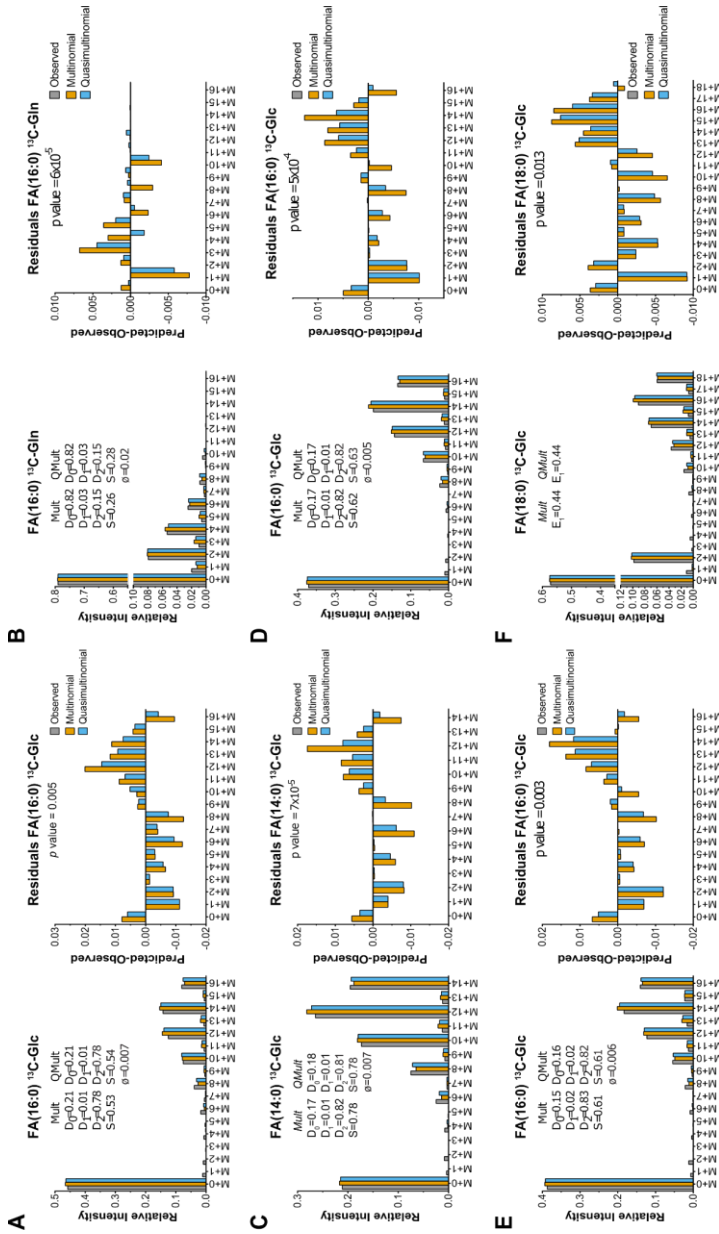


Figure 38. Fitting experimental mass-isotopologue FA data to multinomial and quasi-multinomial distributions. A-B), FA(16:0) in the A549 cells upon incubation with A) U- ^{13}C -glucose or B) U- ^{13}C -glutamine, data obtained from ref.¹⁶⁷ C-D), FA(14:0) and FA(16:0) in the H1.299 cells upon incubation with U- ^{13}C -glucose, data obtained from ref.²⁹ E-F) FA(16:0) and FA(18:0) in the MCF7 cells upon incubation with U- ^{13}C -glucose, data obtained from ref.¹⁵⁵ For each dataset, the experimental data, the fitting done using the FAMetA algorithm with multinomial or quasi-multinomial distributions, and the residuals are shown. The reported *p-values* correspond to the comparisons between multinomial and quasi-multinomial fitting using a log-likelihood ratio test and right-tailed chi-square distribution.

2.3. Estimation of DNL parameters

We considered FA(16:0) as the final product of the DNL. Therefore, FAMetA can estimate the DNL parameters for FA up to 16C. For these species, I and S represent the fraction of the FA pool that is imported and synthesized, respectively, and sum 1:

$$I_{16:0} + S_{16:0} = 1 \quad (\text{Equation 9})$$

For the DNL analysis, FA isotopologue distributions (previously corrected for the natural abundance of the ^{13}C isotopes) are modelled with the following sum of the weighted quasi-multinomial distributions adapted from ref.¹⁷³:

$$P(m = 0) = I + S * (1 + N * \Phi) * \frac{D_0}{1 + N * \Phi} * \left(\frac{D_0 + N * \Phi}{1 + N * \Phi} \right)^{N-1} \quad (\text{Equation 10})$$

$$P(m) = \sum_{j=1}^k P(X_0 = x_{0,j}, X_1 = x_{1,j}, X_2 = x_{2,j}) ; \text{ for } 1 \leq m \leq M \quad (\text{Equation 11})$$

, where:

$$\begin{aligned} P(X_0 = x_{0,j}, X_1 = x_{1,j}, X_2 = x_{2,j}) &= S * \frac{N!}{x_{0,j}! x_{1,j}! x_{2,j}!} * (1 + N * \Phi) * \frac{D_0}{1 + N * \Phi} \\ &* \left(\frac{D_0 + x_{0,j} * \Phi}{1 + N * \Phi} \right)^{x_{0,j}-1} * \frac{D_1}{1 + N * \Phi} * \left(\frac{D_1 + x_{1,j} * \Phi}{1 + N * \Phi} \right)^{x_{1,j}-1} \\ &* \frac{D_2}{1 + N * \Phi} * \left(\frac{D_2 + x_{2,j} * \Phi}{1 + N * \Phi} \right)^{x_{2,j}-1} \end{aligned} \quad (\text{Equation 12})$$

, given that:

$$x_{i,j} = 0, 1, \dots, N$$

$$\sum_{i=1}^2 x_{i,j} = x_{0,j} + x_{1,j} + x_{2,j} = N$$

(Equation 13)

$$\sum_{i=1}^2 i * x_{i,j} = 0 * x_{0,j} + 1 * x_{1,j} + 2 * x_{2,j} = m$$

(Equation 14)

$$0 \leq \phi \leq \frac{1 - \max(D_0, D_1, D_2)}{N}$$

(Equation 15)

M is the total number of carbons in the FA molecule and N equals $M/2$. This represents the number of acetyl-CoA molecules used for the synthesis of an FA of length M . m is the number of ^{13}C atoms incorporated into the FA molecule. D_0 , D_1 and D_2 represent the fraction of acetyl-CoA with 0, 1 or 2 atoms of ^{13}C , respectively, and sum 1. x_0 , x_1 and x_2 represent the number of acetyl-CoA units with 0, 1 or 2 ^{13}C atoms that provide an M -carbon FA with an m label. For a given pair of N and m values, up to k combinations of the x_0 , x_1 and x_2 values fulfil Equations 13 and 14. ϕ accounts for overdispersion and can be set at 0 to reduce the quasi-multinomial distributions to multinomial distributions. The *in silico* validation (described below) of the above-described equations demonstrates an overestimation of ϕ and an underestimation of S and D_2 for values of $D_2 \geq 0.75$ (Additional Figure S29, Appendix 2). In these situations, the upper limit of ϕ is set at $0.5 * (1 - \max(D_0, D_1, D_2)) / N$. Note that overdispersion parameter ϕ modifies D_0 , D_1 and D_2 for each synthesis step, which allows distribution to widen. Based on this model, non-linear regression¹⁷¹ with many sets of plausible initial values (adapted from ref.¹⁷²) is used to fit the observed isotopologue distributions of FA up

to 16C, and to estimate parameters D_1 , D_2 , ϕ and S . When analyzing multiple samples per group, S and D_2 values can be checked to ensure homogeneity within each group. If not, we can assume D_2 should remain within a narrow range for a given condition and thus fix D_2 by the mean of the rest of the samples in the group for the outlier sample and repeat the analysis to improve the calculation of the S value. To improve the analysis results, distributions of FA up to 16C are firstly fitted, and the estimated parameters D_1 , D_2 and ϕ are then used to model longer FA.

2.4. Estimation of elongation parameters

The main product of the DNL of FA is FA(16:0)⁷⁷. Therefore, the main route for elongation starts at 16C and then adds units of two carbons in each elongation step. Elongation from FA(14:0) is a minor route¹²⁹ and is omitted for simplicity. For the FA ranging from 18 to 26 carbons, the following equations are considered:

$$I_{18:0} + E_1(I_{16:0} + S_{16:0}) = I_{18:0} + E_1 = 1$$

(Equation 16)

$$I_{20:0} + E_2 * (I_{18:0} + E_1 * (I_{16:0} + S_{16:0})) = I_{20:0} + E_2 = 1$$

(Equation 17)

$$I_{22:0} + E_3 * (I_{20:0} + E_2 * (I_{18:0} + E_1 * (I_{16:0} + S_{16:0}))) = I_{22:0} + E_3 = 1$$

(Equation 18)

$$\begin{aligned}
 I_{24:0} + E_4 * (I_{22:0} + E_3 * (I_{20:0} + E_2 * (I_{18:0} + E_1 * (I_{16:0} + S_{16:0})))) &= I_{24:0} + E_4 \\
 &= 1
 \end{aligned}$$

(Equation 19)

$$\begin{aligned}
 I_{26:0} + E_5 * (I_{24:0} + E_4 * (I_{22:0} + E_3 * (I_{20:0} + E_2 * (I_{18:0} + E_1 * (I_{16:0} + S_{16:0})))) &= I_{26:0} + E_5 \\
 &= 1
 \end{aligned}$$

(Equation 20)

For the elongation analysis of endogenous FA, isotopologue distributions are modelled using Equations 10-12 for synthesis up to FA(16:0), followed by single and independent elongation steps ($E_1, E_2 \dots, E_n$). As each step is independent and involves the addition of a unique acetyl-CoA molecule, overdispersion cannot be considered and each step is modelled by using multinomial distributions. The probability of incorporating 0, 1 or 2 ^{13}C atoms into the FA to be elongated equals $E_i D_0$, $E_i D_1$, and $E_i D_2$, respectively. For FA longer than 16C, only synthesis and elongation terms are estimated ($S, E_1, E_2 \dots, E_n$), while the rest (D_0, D_1, D_2 and ϕ) are inherited from the results obtained for the FA(16:0). In case no results are available for FA(16:0), FAMetA uses FA(14:0), mean of all FA of 16C (FA(16:X)) or mean of all FA of 14C (FA(14:X)) in this order of priority. For FA(18:0), FA isotopologue distributions (previously corrected for natural ^{13}C isotopes abundance) are modelled with the following equations:

$$P_{18:0}(m = 0) = I_{18:0} + E_1 * D_0 * P_{16:0}(m = 0)$$

(Equation 21)

$$P_{18:0}(m = 1) = E_1 * D_0 * P_{16:0}(m = 1) + E_1 * D_1 * P_{16:0}(m = 0)$$

(Equation 22)

$$P_{18:0}(m) = E_1 * D_0 * P_{16:0}(m = m) + E_1 * D_1 * P_{16:0}(m = m - 1) + E_1 * D_2 * P_{16:0}(m = m - 2)$$

$$\text{for } 2 \leq m \leq M; P_{16:0}(m > 16) = 0$$

(Equation 23)

Analogous equations can be obtained for FA with $M > 18$ by adding elongation terms to previously existing distributions. The *in silico* validation of the above-described equations demonstrates that elongation terms can only be accurately determined when the contribution (D_2) of the ^{13}C -tracer is greater than 0.05. In addition, based on this validation data and to ensure that reliable results are obtained, by default, FAMetA only estimates elongation parameters for those samples whose D_2 parameter has been estimated to be greater than 0.1. For n6 and n3 series (Figure 33), elongation is usually expected from FA(18:2)n6 and FA(18:3)n3. Thus, synthesis ($S_{16:0}$) and the first elongation step (E_1) are set at 0. If isotopologue M+2 is observed given the degradation of FA(18:2)n6 or FA(18:3)n3, followed by one elongation step, then E_1 is estimated. However, the endogenously synthesized fraction remains at 0. Once again, non-linear regression¹⁷¹ with multiple initial values¹⁷² is used to fit the observed isotopologue distributions of the elongated FA.

2.5. Estimation of desaturation

After estimating the synthesis and elongation parameters, these results can be used for the indirect calculation of the FA fraction that comes from desaturation in the unsaturated FA. For a given unsaturated FA (e.g., FA(18:1)n9), we can conceptually consider a one-step elongation-desaturation reaction (in this example, directly from

FA(16:0) to FA(18:1)n9), or a two-step elongation followed by a desaturation process (in this example FA(16:0) is elongated to FA(18:0) and then desaturated to FA(18:1)n9) (Figure 34). By means of FAMetA, we can directly estimate both E_i and E_i' from the isotopologue distributions of FA(18:0) and FA(18:1)n9, respectively. From alternative paths, the relative import and endogenous synthesis pathways of FA(18:1)n9 can be written as:

$$I_{18:1n9}' + E_1' * (S_{16:0} + I_{16:0}) = 1 \quad (\text{Equation 24})$$

$$I_{18:1n9} + \Delta * E_1 * (S_{16:0} + I_{16:0}) + \Delta * I_{18:0} = 1 \quad (\text{Equation 25})$$

By combining both equations, we can define that:

$$I_{18:1n9}' = I_{18:0} * \Delta + I_{18:1n9} \quad (\text{Equation 26})$$

and thus, calculate desaturation parameter Δ as:

$$\Delta = \frac{E_1'}{E_1} \quad (\text{Equation 27})$$

If both E_i' and E_i are below the confidence interval, which for desaturation is set to 0.05 by default, parameter Δ is not calculated, and E_i' remains as the endogenously synthesized fraction. If the stationary state is not reached, values > 1 can be obtained for the desaturation parameter that is, in this case, replaced with 1.

This same approach can be used for all the known desaturation steps provided when the precursor and product FA isomers are correctly and uniquely identified, and the stationary state is reached.

For the FA synthesized from desaturation activities, Δ is considered the fraction from endogenous synthesis. So the imported fraction is calculated as $1-\Delta$. With unknown isomers or missing precursors, S' or E' is returned for the DNS of FA until 16C or the elongation of longer FA, respectively. The range of reactions included in FAMetA are described in Figure 33^{129,174-176}. Additional reactions (desaturations) can be included for unknown/additional FA by modifying *desaturationdb* in the FAMetA R package.

2.6. Model assumptions

In order to interpret the results correctly, the model assumptions made by FAMetA should be considered:

- 1) The acetyl-CoA pool contributing to lipogenesis has a uniform labelling pattern.
- 2) The lipogenic acetyl-CoA pool reaches isotopic steady state quickly compared with the total labelling time.
- 3) For FA of 16C or longer, the final product of FASN (i.e., DNL) is FA(16:0).
- 4) For the FA belonging to the n3 and n6 series, S parameter is set to 0.
- 5) At any given time point $I = \text{import} + \text{pre-existing FA}$, and only when the pre-existing FA have been completely replaced (i.e., the actual steady state has been achieved) $I = \text{import}$.
- 6) There is a single FA pool.

2.7. Data requirements for FA modelling

Before performing the FA metabolism analysis, the user should check that the FA of interest have been labelled enough to obtain isotopologue distributions of good quality (avoid missing isotopologues) that guarantee that the calculated parameters fall within the ranges that allow their accurate estimation. When curating FA annotations, FA names must follow the nomenclature FA(C:d)ns, where C is the total number of carbon, d is the number of double bounds and ns refers to the omega series, which indicates the position of the last double bound starting from the end of the chain. Duplicated identities are not allowed and the series must belong either to known series (i.e. 3, 5, 6, 7, 7a (i.e. second double bond introduced by FADS2 at 16C), 7b (i.e. second double bond introduced by FADS2 at 18C), 9, 10, 12, 13) or to unknown series where the letters x, y and z are used for nomenclature.

2.8. Implementation

FAMetA has been developed in an R programming environment¹⁷⁷ and is available via CRAN (<https://CRAN.R-project.org/package=FAMetA>). The source code and development version are also available at <https://github.com/maialba3/FAMetA>. In addition, a web-based implementation of FAMetA has been built using the Shiny R-package¹⁴⁹, which is accessible at www.fameta.es. Example data files, scripts and tutorials for the R package and the web application can be found at <http://www.fameta.es> via the “Resources” tab.

2.8.1. R package

FAMetA functions can be divided into three groups: i) functions aimed to data pre-processing, which are imported from LipidMS; ii) functions devoted to fatty acids annotation and manual curation; and iii) functions designed to perform the metabolic analysis. Two first groups of functions work with *msbatch* objects from LipidMS, while the last uses *fadata* lists containing, at least, samples metadata, FA identities, abundances for all expected isotopologues and internal standard intensities for normalization if available. After the data correction step, MID and pool size (i.e., total sum of isotopologues intensities for each FA) are also added. MID are used for the subsequent estimation of DNL, elongation and desaturation analysis.

2.8.2. Web-based tool

In order to provide a user-friendly GUI interface that covers all the required steps for FA analysis, FAMetA has also been implemented as a web-based tool using Shiny¹⁴⁹, which can be accessed through <http://www.fameta.es>. After accessing the tool, the following tabs will take users through the FAMetA workflow. In this case, each tab is devoted to run one step in the FAMetA workflow, so that users will run them sequentially and they will receive an email for each tab:

- **Data pre-processing.** First step in FAMetA workflow consists of data pre-processing using the LipidMS R package. At this tab (Figure 39), mzXML files and a metadata csv file are required. Metadata file must have at least three columns: sample (mzXML file names), acquisitionmode (MS) and samplotype (QC, group1, group2, etc.). Once all files have

been uploaded, pre-processing parameters must be tuned. After this first step has been performed, users will receive an email containing the FA annotation results.

- **Manual curation.** Automatic FA annotations can be modified by editing the csv file received by email: removing rows of unwanted FA, modifying the initial and end retention times, or adding new rows with missing compounds. The internal standards for later normalization can also be added at this point to a new row by indicating IS in the compound name column. Once FA annotations have been curated, ^{13}C isotopologues for each FA will be searched and MID will be sent by email (Figure 40).
- **Metabolic analysis.** At this tab (Figure 41), a csv file containing metadata and MID must be provided (csv file received after the previous step). Then, data correction is required which will run four different steps (all of them are optional): data correction for natural ^{13}C abundance using the *Accucor* algorithm¹⁷⁰, data normalization with internal standards, blank subtraction and external normalization. Finally, the actual FA metabolism analysis can be performed.

Extra documentation, examples and links to the source code, which is available in github or CRAN, can be found at www.fameta.es by clicking on the “Resources” tab.

Figure 39. Data pre-processing tab of the FAMetA web tool. Users can upload the mzXML files and tune the processing parameters employed by LipidMS to run the first step of FAMetA’s workflow.

Figure 40. Manual curation tab of the FAMetA web tool. Users can upload the revised FA annotations to obtain the subsequent mass isotopologue distributions.

The image shows the 'Metabolic analysis' tab of the FAMetA web tool. The interface is split into two main panels. The left panel, titled 'FAMetA: Fatty Acid Metabolic Analysis', contains a 'Job Name' field with the value 'Job_2023-02-16'. Below it is an 'Import FA data (csv file)' section with a 'Browse...' button and the text 'No file selected'. There are also input fields for 'sep' (column delimiter, '-'), 'dec' (decimal character, '-'), and an 'Email (to send your results):' field. The 'Data correction' section includes 'correct13C' (a dropdown menu set to 'TRUE'), 'resolution' (a text input field with '140000'), 'purity13C' (a text input field with '0.99'), and 'blankgroup' (a text input field with 'blank'). The right panel, also titled 'FAMetA', contains 'Synthesis parameters' and 'Elongation parameters'. The 'Synthesis parameters' section includes 'R2Thr' (0.95), 'maxiter' (1000), 'maxconvergence' (100), 'startpoints' (5), and 'propagateD' (a dropdown menu set to 'TRUE'). The 'Elongation parameters' section includes a checked 'Run elongation' checkbox, 'R2Thr' (0.95), and 'maxiter' (10000).

Figure 41. Metabolic analysis tab of the FAMetA web tool. Once FA MID have been correctly obtained, they are modelled to estimate the FA metabolic parameters.

3. FAMetA performance evaluation

3.1. *In silico* validation

To validate the FAMetA algorithm and its implementation, *in silico* MID were generated. To simulate experimental distributions, multiple values covering the expected range for each parameter were used. For each theoretical isotopologue distribution, 10 realizations of Gaussian noise were simulated at four noise levels: 0%, 2%, 5%, or 10% relative standard deviation (RSD). The generated data was used to calculate the RSD and relative error of each modelled synthesis parameter for the following FA, which comprise an example of all the reactions included in FAMetA: FA(16:0) (Additional Figure S29, Appendix 2), FA (18:0) (Additional Figure S30, Appendix 2), FA(20:0) (Additional Figure S31, Appendix 2), FA(22:0) (Additional Figure S32, Appendix 2), FA(24:0) (Additional Figure S33, Appendix 2), FA(16:1)n7 (Additional Figure S34A-B, Appendix 2), and FA(18:1)n9 (Additional Figure S34C-D, Appendix 2). FAMetA accurately determined the complete set of FA synthesis parameters (relative error < 15%, RSD < 15%) whenever the fractional contribution of the tracer (D_2) and the parameters to be calculated for a given FA (i.e., S , E_1 , E_2 , E_3 , and E_4) fell within the 0.05 - 0.9 range. This ensures its applicability in an actual biological scenario.

3.2. Biological validation

Once FAMetA algorithms were validated using *in silico* distributions, FAMetA performance was evaluated using actual experimental data. To this end, a variety of *in vitro* and *in vivo* experimental settings were used. Firstly, mouse CD8⁺ T-cells were

incubated for 72h with different uniformly ^{13}C labelled tracers (U- ^{13}C -glucose, U- ^{13}C -glutamine, U- ^{13}C -lactate or U- ^{13}C -acetate) and in the presence or absence of well-known inhibitors of FA metabolism enzymes (i.e., FASN (GSK2194069, FASNi)¹⁵⁷, SCD1 (A93572, SCD1i)^{159,160} and FADS2 (SC26196, FADS2i)¹⁵⁸) to evaluate FAMetA in a controlled disturbance scenario. Then, we analyzed previously published data generated using *in vivo* experimental models (incorporation of U- ^{13}C -fructose into saponified circulating FA in wild-type and intestine-specific ketohexokinase (KHK-C) knockout mice after drinking normal water for 8 weeks, or 5% or 10% sucrose water)¹⁷⁸ to validate FAMetA in a real *in vivo* scenario.

3.2.1. FAMetA enables the analysis of FA metabolism *in vitro*

For the mouse CD8⁺ T-cells, total lipids were extracted from cell pellets and saponified to release FA, which were subsequently analyzed by LC-MS. Twenty-seven known FA were detected in the samples, including a variety of saturated, monounsaturated and polyunsaturated FA within the range from 14 to 24 carbons. FAMetA accurately modelled the obtained MID for all of them and valuable biological information about nutrient preferences and metabolic origin of each particular FA was obtained. First, we evaluated the contribution of different ^{13}C -tracers to the FA synthesis under standard culture conditions (i.e., RPMI media and normoxia). In this scenario, glucose is the preferred carbon source ($D \approx 0.7$) in the active mouse CD8⁺ T-cells (Figure 42A), with a minor contribution of glutamine ($D \approx 0.08$) (Figure 42B). If present in media, lactate and

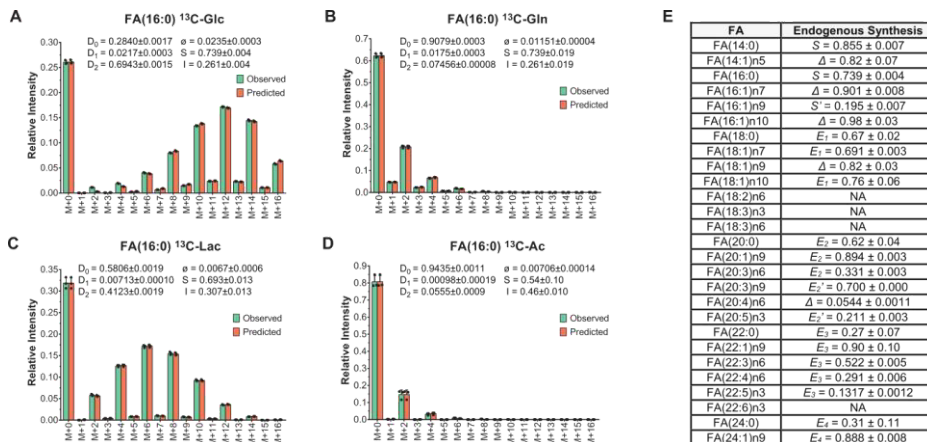


Figure 42. Biological validation of FAMetA in active mouse CD8⁺ T-cells incubated with different U-¹³C-tracers. Estimation of the FA metabolism parameters in the active mouse CD8⁺ T-cells incubated for 72 h with various U-¹³C-tracers. A-D) Estimation of the sources and the DNL parameters for FA(16:0) upon incubation with A) U-¹³C-glucose, B) U-¹³C-glutamine, C) U-¹³C-lactate or D) U-¹³C-acetate. E) Summary of the endogenously synthesized fraction for the 27 known FA detected in the active mouse CD8⁺ T-cells upon incubation with U-¹³C-glucose.

glucose are metabolically exchangeable at the lactate dehydrogenase (LDH) level. Thus, lactate feeds the pyruvate pool and FA synthesis (Figure 42C). When supplemented in media, acetate feeds the acetyl-CoA pool and contributes to FA synthesis (Figure 42D). These results are consistent with previously published data obtained by different approaches^{163,164,167}. Although most of the identified FA are present in culture media, endogenous synthesis is the preferential route for saturated and monounsaturated FA, whereas polyunsaturated FA preferentially come from exogenous sources (Figure 42E).

Then, FA metabolism was evaluated upon incubation with U-¹³C-glucose and treatment with different known inhibitors. Treatment with FASNi and SCD1i slightly decreases cell proliferation, but FADS2i does not (Figure 43A). Changes in the relative pool size of the detected FA appear (Figure 43B); e.g., SCD1i lowers the intracellular

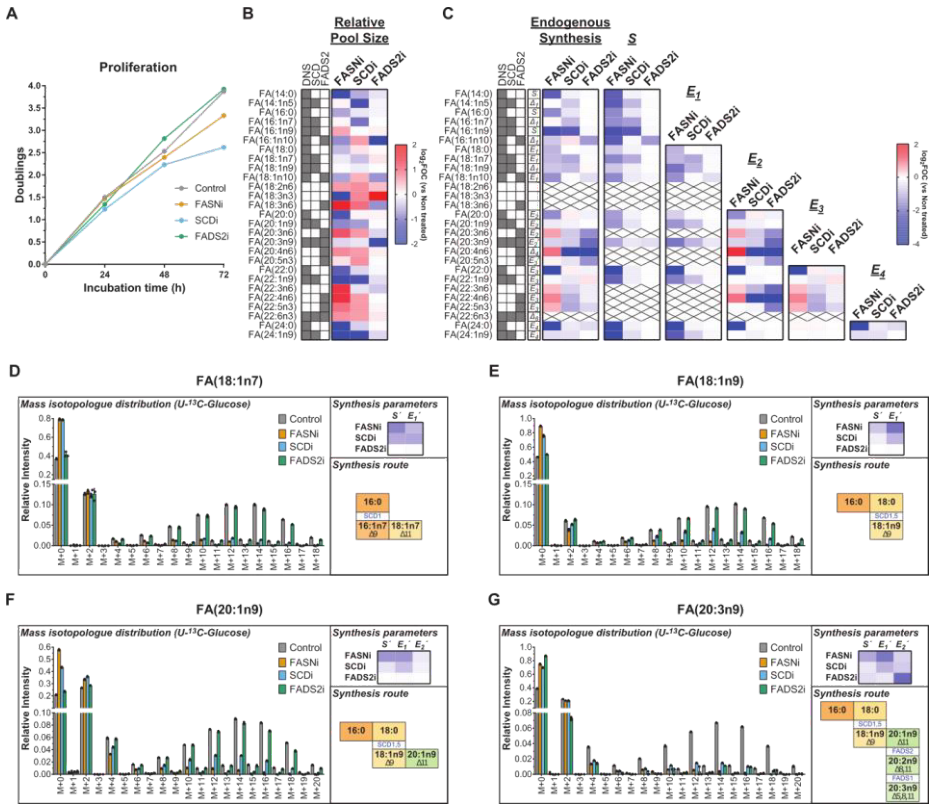


Figure 43. Biological validation of FAMeTA in active mouse CD8⁺T-cells incubated with U-¹³C-glucose and different inhibitors of the FA metabolism. A-G) Analysis of alterations in FA biosynthesis in the active mouse CD8⁺ T-cells incubated for 72 h with U-¹³C-glucose induced by FASN inhibitor GSK2194069, SCD inhibitor A93572 and FADS2 inhibitor SC26196. A) Mean proliferation of the active mouse CD8⁺ T cells during the 72-hour incubation period. B) Heatmap showing for each identified FA the mean value of the log₂ fold-of-change (vs. untreated) in the relative pool size. C) Heatmap showing the mean value of the log₂ fold-of-change (vs. untreated) for each identified FA in the following parameters: endogenously synthesized fraction, calculated S , E_1 , E_2 , E_3 and E_4 . For each FA, the parameter reported for the endogenous synthesis is indicated. D-G) Mass isotopologue distribution, the mean value of the log₂ fold-of-change (vs. untreated) in the synthesis parameters and synthesis route for D) FA(18:1)n7, E) FA(18:1)n9, F) FA(20:1)n9 and G) FA(20:3)n9. In all cases n=3. Individual points are shown for the mass isotopologue distributions, and the mean values are reported elsewhere. The shadowed cells in B and C indicate the activities (DNS, SCD or FADS2) involved in the synthesis of a particular FA. On the heatmaps, crosses indicate missing or NA values. In D-G, the horizontal transitions in the synthesis route description denote elongations (enzymes not indicated), and vertical transitions denote desaturations (enzymes indicated).

levels of the n5, n7, and n9 series FA, and increases the relative abundance of FADS2 products (e.g., sapienic acid, FA(16:1)n10), while FADS2i considerably diminishes sapienic acid abundance, which is consistent with previous reports on the complementary and compensatory roles of SCD1 and FADS2¹¹⁹ (Figure 43B). When analyzing endogenous synthesis, the changes reveal which enzymes are involved in the synthesis of each identified FA. FASNi decreases the endogenous synthesis of all the FA that come from FA(16:0), and SCD1i and FADS2i decreases the endogenous synthesis of all the FA that these enzymes are involved in (e.g. n9 series FA for SCD1i, n10 series FA for FADS2i) (Figure 43C). When focusing on each calculated synthesis parameter, identifying the step in which each enzyme acts and mapping synthesis routes are straightforward. For example, for FA(18:1)n7 and FA(18:1)n9, SCDi differentially affects synthesis parameters. In FA(18:1)n9, where SCD acts at the 18-carbon level, the most prominent decrease is in calculated E_1 (i.e., $E_1' = E_1 * \Delta$), in FA(18:1)n7, where SCD acts at the 16-carbon level, both calculated S (i.e., $S' = S * \Delta$), and E_1 decreases upon treatment with SCDi (Figure 43D-E). The SCDi inhibition pattern observed in FA(18:1)n9 is mirrored in FA(20:1)n9 and FA(20:3)n9 (Figure 43C-G). In addition, FADS2i decreases the calculated E_2 (i.e., $E_2' = E_2 * \Delta$) for FA(20:3)n9, which is indicative of FADS2 introducing a double bond at the 20-carbon level (Figure 43G). Thus, FAMetA allows the identification of both changes in general patterns and particular synthesis parameters induced by FA metabolism inhibitors.

3.2.2. FAMetA enables the analysis of the FA metabolism *in vivo*

To test whether FAMetA could handle with *in vivo* data, we analyzed previously published data on the incorporation of U-¹³C-fructose into saponified circulating FA in wild-type and intestine-specific ketohexokinase (KHK-C) knockout mice after drinking normal water for 8 weeks, or 5% or 10% sucrose water¹⁷⁸. *In vivo* generated data is characterized by low synthesis of FA, slight contribution of the ¹³C-tracer to the FA biosynthesis and a high proportion of odd-labelled isotopologues. Despite this, the estimated parameters fall within the high-confidence ranges established using the *in silico* validation. Overall, FAMetA has proven to properly fit *in vivo* data (Figure 44).

The observed general trend suggests increased DNL, elongation and desaturation upon sucrose treatment, with a more pronounced effect on the KHK-C knockout mice (Figure 44A-E). Neither sucrose nor KHK-C ablation influences the fractional contribution of fructose to DNL (Figure 44F-G), but exposure to drinking fructose significantly alters DNL (S), elongation (E_1 and E_2) and desaturation (Figure 44H-K). The *post hoc* comparisons reveal significantly heightened FA(16:0) synthesis when drinking more fructose, but only in the KHK-C knockout group (Figure 44H), as well as augmented desaturation when drinking more fructose in both the wild-type and KHK-C knockout mice (Figure 44F). These results agree with and extend those reported by the authors of the study, which reported increased total ¹³C-labelled carbons in saponified circulating palmitate that accounts for the cumulative effect of DNL, the contribution of fructose to DNL and the palmitate concentration¹⁷⁸.

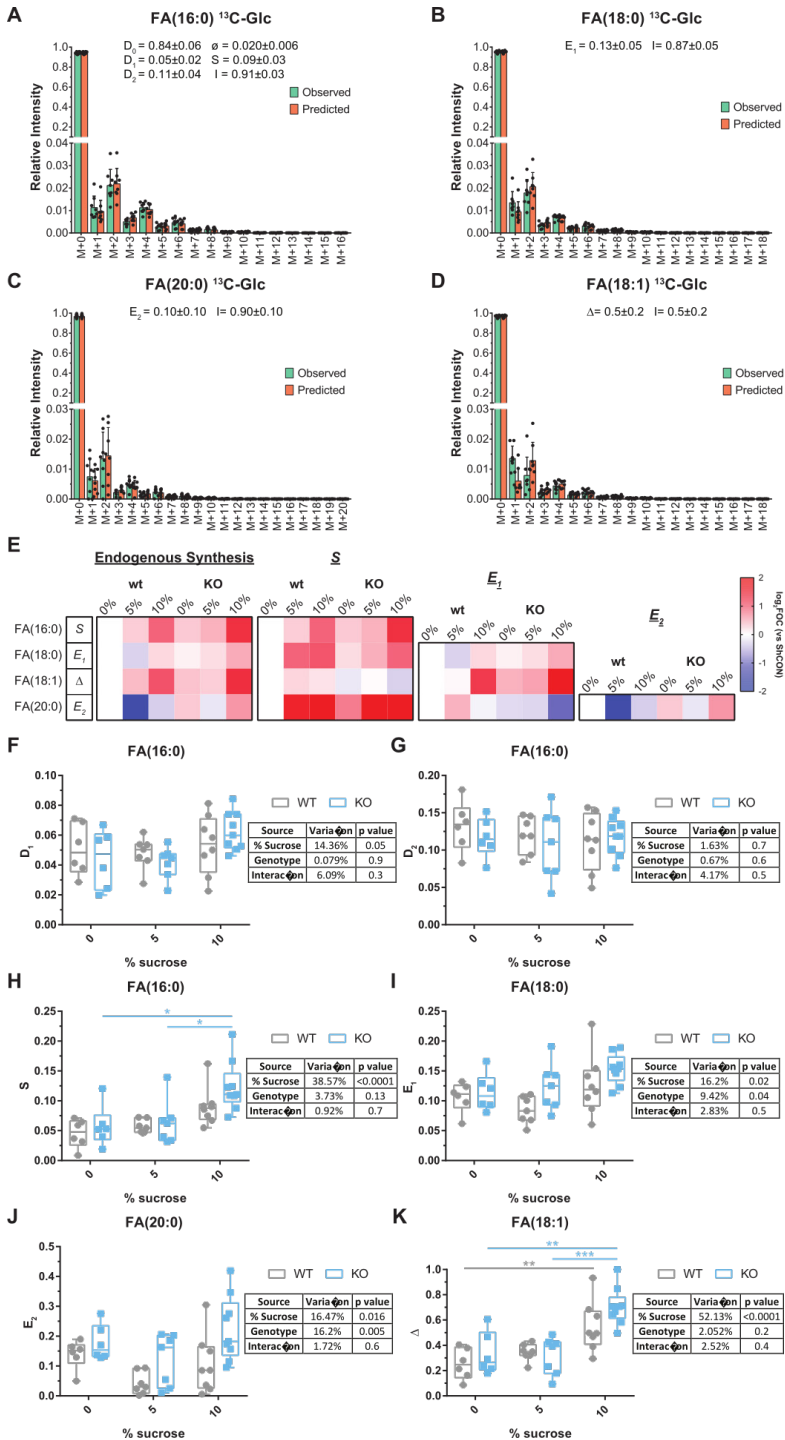


Figure 44. Biological validation of FAMetA in WT and KHK-C mice after drinking normal or 5% or 10% sucrose water, data obtained from ref.¹⁷⁸. A-D) FAMetA was used to fit all the reported experimental MID for the WT 10% sucrose group. E) Heatmap showing the log₂ fold of change (vs. WT 0% sucrose) for each reported FA in the following parameters: endogenously synthesised fraction, calculated S, E1 and E2. For each FA, the parameter reported for the endogenous synthesis is indicated. The shadowed cells indicate the activities (DNS, elongation (ELOVL) or SCD1-mediated desaturation (SCD)) involved in the synthesis of a particular FA. F-K) The calculated FA synthesis parameters obtained with FAMetA. The tables summarise the result of the two-way ANOVA performed for each calculated parameter. Paired differences are calculated by a post hoc Tukey test. The *p*-values obtained for the reported significant differences: S parameter for FA(16:0), KO 10% vs. KO 0%, *p*-value=0.014, KO 10% vs. KO 5%, *p*-value=0.03; Δ parameter for FA(18:1), WT 10% vs. WT 0%, *p*-value=0.006, KO 10% vs. KO 0%, *p*-value=0.0012, KO 10% vs. KO 5%, *p*-value=0.0004 (n=6,6,7,7,8,8).

3.3. Comparison between FAMetA and other available tools

Once the capabilities of FAMetA were proven, we decided to compare the functionalities implemented in FAMetA with those from previously available approaches and tools (i.e., ISA, ConvISA¹⁵⁵, Kamphorst *et al.*¹⁵⁶ and FASA¹²⁹). Table 10 shows the functions implemented by FAMetA and previous methods. Compared to previous tools, it allows the characterization of a broader FA biosynthesis network as it includes DNL, elongation and desaturation in a single tool. In addition, FAMetA offers the possibility of running all the required steps from data pre-processing to analysis of FA metabolism and graphical representation, and thanks to the web-based application, this can be performed in a user-friendly environment. Moreover, regarding the FAMetA results and parameters, it improves the estimation of the elongation steps enabling an easier interpretation of the estimated parameters, and the implementation of quasi-multinomial fitting that includes de parameter Φ , accounts for data overdispersion.

Table 10. Comparison of features implemented within the main available tools for the analysis of FA metabolism. ✓ means that the feature is covered, ~ means that it is covered but with some limitations (detailed in Comments) and X, not covered.

| | ISA | ConvISA ¹⁵⁵ | Kamphorst ¹⁵⁶ | FASA ¹²⁹ | FAMetA |
|---|---|---|--|--|-------------------------|
| De novo lipogenesis | ✓ | ✓ | ✓ | ✓ | ✓ |
| Contribution of labeled nutrient to lipogenic AcetylCoA pool | ✓ | ✓ | ✓ | ✓ | ✓ |
| Elongation | X | ~ | ✓ | ✓ | ✓ |
| Desaturation | X | X | ~ | X | ✓ |
| Data pre-processing | X | X | X | X | ✓ |
| Graphical output | X | X | X | X | ✓ |
| Implementation | Matlab | Matlab script | Matlab script | Matlab toolbox | R-package Web-based app |
| Comments | The actual algorithm is not released as script or equivalent, but has to be implemented by users or used within a metabolic flux tool | Elongation calculated only for FA(18:0) | Steady state must be achieved as M+0 = import. Desaturation based on total labeling and exemplified only for FA(18:1)n9. | Elongation described as de novo lipogenesis up to the total number of carbons plus multiple import-elongation terms. | |

As FASA is the unique existing tool available to model FA elongation up to 26C, the performance of FAMetA to estimate elongation parameters was then compared with FASA. The comparison between FAMetA and FASA was performed using a dataset published by FASA developers, which contained information of twelve FA determined in the H1229 cells incubated with U-¹³C-glucose and/or U-¹³C-glutamine

either with or without down-regulation of the SREBP cleavage activating protein (SCAP) labelled as shControl and shSCAP, respectively¹²⁹.

We firstly compared them in computing speed terms. The processing time with Intel Xeon E5-1620 CPU (3.5GHz) with 32GB RAM in Windows is ~170min for FASA (Matlab R2022a) and ~ 12min for the FAMetA R package (R v4.1.1, RStudio v1.4, FAMetA v0.1.3). The same analysis on the FAMetA webserver takes ~30min. The FAMetA algorithm calculates the fractional contribution of the carbon source (D_0 , D_1 and D_2) and overdispersion parameter (ϕ) based on the distribution of FA(16:0). These values are then employed to fit the remaining FA. The same strategy was employed for FASA by firstly fitting FA(16:0) and then the remaining FA by setting the D_0 , D_1 and D_2 values.

Then, we compared them in terms of FA metabolic modelling. FAMetA and FASA present differences in the way they calculate the FA biosynthesis parameters. While FAMetA calculates import, DNL, elongation and desaturation, FASA does not calculate desaturation. In addition, FAMetA and FASA calculate elongation by different approaches that makes a difference in terms of interpretation of the results. FAMetA provides the direct estimation of each step in a specific FA synthesis pathway (Figure 34). For example, the FA(20:0) sources are described as $I_{20:0} + E_2 = I_{20:0} + E_2 * (I_{18:0} + E_1 * (I_{16:0} + S_{16:0}))$, where each parameter (S , E_1 , E_2) directly represents a single synthesis route step, and E_2 is the direct estimation of the fraction of FA(20:0) that results from the elongation of the total FA(18:0) pool. Conversely in FASA, FA(20:0) sources are described as $S + IE_2 + IE_1 + I$, where S (elongated from FA(16:0)) actually represents $S_{16:0} * E_1 * E_2$; IE_2 (elongated from the imported FA(16:0)), $I_{16:0} * E_1 * E_2$; IE_1 (elongated from

FA(18:0)), $I_{18:0} * E_2$, and I represents the fraction of the directly imported FA(20:0)¹²⁹. Using FASA, the authors of the original study conclude that SCAP down-regulation decreases both DNL and elongation¹²⁹ (Figure 45). Although they do not report the detailed results of each synthesis parameter for every reported FA, we analyzed the dataset using FASA to find that several parameters change for each FA, and it is difficult to ascertain clear patterns to provide a more detailed conclusion than that proposed by the authors. Using FAMetA, we identify that SCAP down-regulation decreases the synthesis of monounsaturated n7 (i.e., FA(16:1)n7 and FA(18:1)n7) and n9 (i.e., FA(18:1)n9, FA(20:1)n9, and FA(22:1)n9) FA (Figure 45B). When focusing on particular synthesis parameters, the calculated S (i.e., $S' = S * \Delta$) is the most altered parameter for the n7 series, which is consistent with SCD1 introducing the double bond at the 16-carbon level. The calculated E_i (i.e., $E_i' = E_i * \Delta$) is the most altered parameter for the n9 series, and is consistent with SCD1 introducing the double bond at the 18-carbon level (Figure 45B). Our refined analysis, which includes the calculation of desaturation and easy-to-interpret direct estimations of each elongation step, identifies that the main decrease occurs in the endogenous synthesis of the SCD1-derived n7 and n9 series of FA. This indicates diminished SCD1 activity as the main metabolic change induced after SCAP silencing.

Thus, we conclude that compared to FASA, FAMetA provides a more comprehensive characterization of the FA biosynthetic network, a better and more intuitive description of each synthesis parameter, and a more complete workflow that goes from data processing to group-based comparisons and graphical representation. It is also more efficient from a computing perspective.

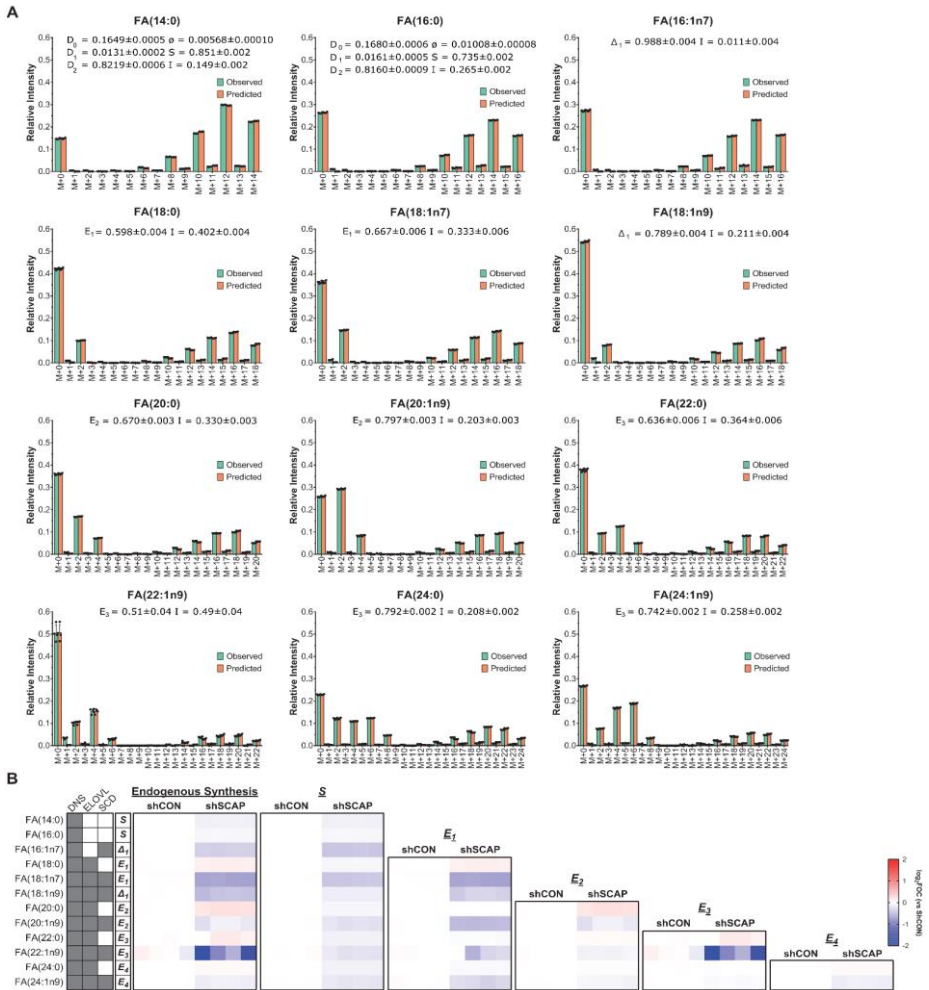


Figure 45. Analysis of the influence of the down-regulation of SCAP on the FA metabolism in the H1299 cells; data obtained from ref.¹²⁹. A) FAMeTA was used to fit all the reported experimental mass-isotopologue distributions for the control condition (shCON). B) Heatmap showing the log₂ fold of change (vs. shCON) for each reported FA in the following parameters: endogenously synthesized fraction, calculated S , E_1 , E_2 , E_3 and E_4 . For each FA, the parameter reported for endogenous synthesis is indicated. The shadowed cells indicate the activities (DNS, elongation (ELOVL), or SCD1-mediated desaturation (SCD)) involved in the synthesis of a particular FA. n=4.

4. FAMetA enables the identification of unknown FA in biological samples

Finally, to test whether FAMetA was able to help in the identification of unknown FA, we decided to carry out the unbiased analysis of total FA in the non-small cell lung cancer (NSCLC) A549 cell line, which revealed high FA diversity (62 species), including several FA (33 species) that do not match with the used FA standards (Figure 46A-B). Here, we hypothesize that the information provided by the retention time of each FA combined with the FAMetA analysis of the MS-data generated using U-¹³C-glucose and well-characterized inhibitors (i.e., FASNi, SCDi, and FADS2i) would serve as a valuable strategy to identify unknown and unexpected FA by the reconstruction of their metabolic synthesis route. All the detected unknown FA incorporated ¹³C from U-¹³C-glucose, which confirms their endogenous metabolic origin. In all the cases the information provided by the inhibition profile and the retention time allowed us to propose identities for them all (Figure 46C, Figure 47 and Additional Figure S34, Appendix 2).

For example, five FA(18:2) (18:2n6, nv, nx, ny, nz) are detected in the NSCLC cell line A549 (Figure 46B-C). Based on their retention time and expected n-series, v, x, y and z should be > 6 (Figure 46B). FA(18:2)nz is identified as FA(18:2)n10 because SCDi does not affect any synthesis parameter and FADS2 decreases the calculated E_i (i.e., $E_i' = E_i * \Delta$) (Figure 47). For FA(18:2)nv and FA(18:2)nx, SCD1i decreases the calculated S more than E_i , but the opposite occurs for FA(18:2)ny. Thus FA(18:2)nv,nx and FA(18:2)ny are respectively identified as FA(18:2)n7 and FA(18:2)n9 (Figure 47A-C).

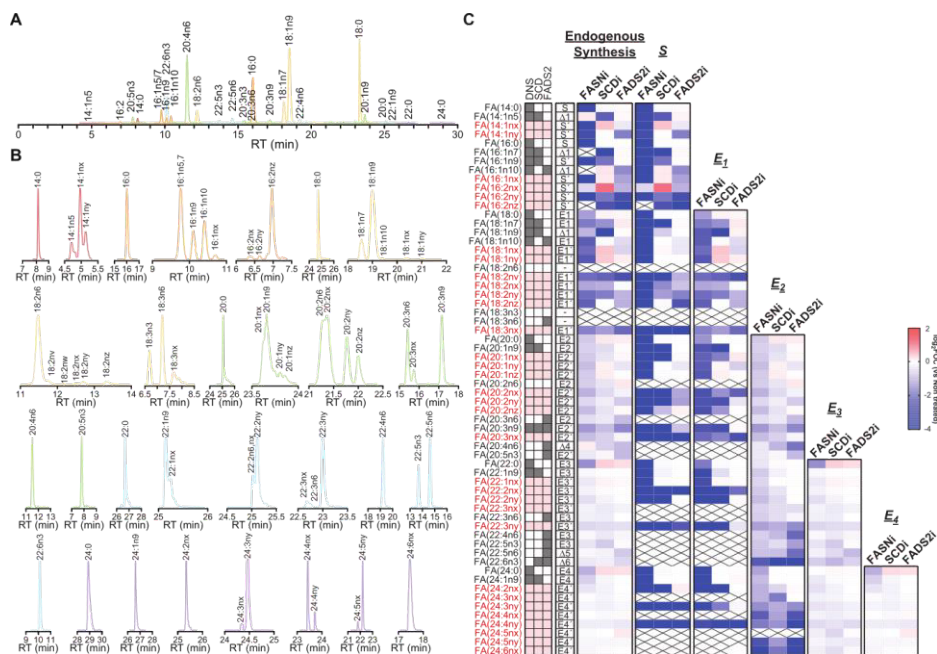


Figure 46. Analysis of the FA diversity in the human NSCLC cell line A549 incubated for 72 h with U-¹³C-glucose induced by the use of different FA metabolism inhibitors (FASNi, SCDi and FADS2i). A-B) Chromatographic separation of the saponified FA from the A549 cells in culture. A) Combined chromatogram showing all the detected FA. B) Individual chromatograms for each detected FA. C) Heatmap showing the mean value of the log₂ fold-of-change (vs. untreated) for each detected FA in the following parameters: endogenously synthesized fraction, calculated S , E_1 , E_2 , E_3 and E_4 . For each FA, the parameter reported for the endogenous synthesis is indicated. The shadowed cells indicate the activities (DNS, SCD or FADS2) involved in the synthesis of a particular FA. Red denotes the FA whose synthesis route is unknown. On the heatmap, crosses indicate missing or NA values.

Based on the FADS2i inhibition profile, we conclude that FADS2 introduces the second double bond at the 18-carbon level for FA(18:2)nv because FADS2i decreases the calculated E_i more than the calculated S , and at the 16-carbon level for FA(18:2)nx because FADS2i decreases the calculated S . Therefore, the four unknown FA(18:2) are identified as FA(18:2)n7(Δ 6,11), FA(18:2)n7(Δ 8,11), FA(18:2)n9(Δ 6,9) and FA(18:2)n10(Δ 5,8), respectively (Figure 47A-D).

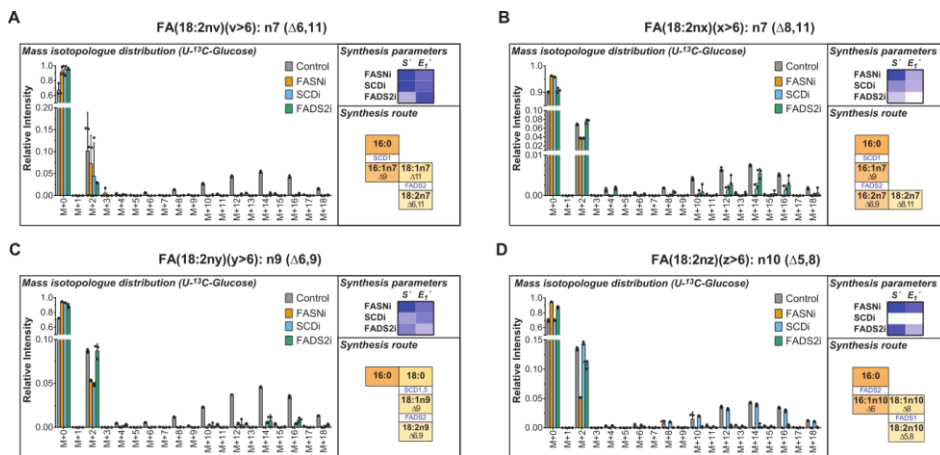


Figure 47. Elucidation of the synthesis route of unidentified FA species by combining FAMetA and FA metabolism inhibitors. A-D) The mass-isotopologue distribution, the mean value of the log₂ fold-of-change (vs. untreated) in the synthesis parameters, and the proposed synthesis route for A) FA(18:2nv), B) FA(18:2nx), C) FA(18:2ny) and D) FA(18:2nz), whose identities do not match any standard employed for the method development. In all cases n=3. Individual points are shown for the mass isotopologue distributions. The mean values are reported elsewhere. In the synthesis route description, horizontal transitions denote elongations (enzymes not indicated) and vertical transitions depict desaturations (enzymes indicated).

Based on this strategy, we built a decision tree that guides the identification of each double bond position based on the inhibition profile (Figure 48). Following this strategy, we identified a total of 33 unknown FA (Additional Figure S35). Of them, 11 FA were confirmed with commercially available standards (Figure 49), and 9 of them did not match with any previously described FA. Therefore, FAMetA combined with our proposed strategy disclose a more comprehensive FA biosynthetic landscape of A4594 cells, including the description of 11 novel FA that belong to already described n-series (Figure 50).

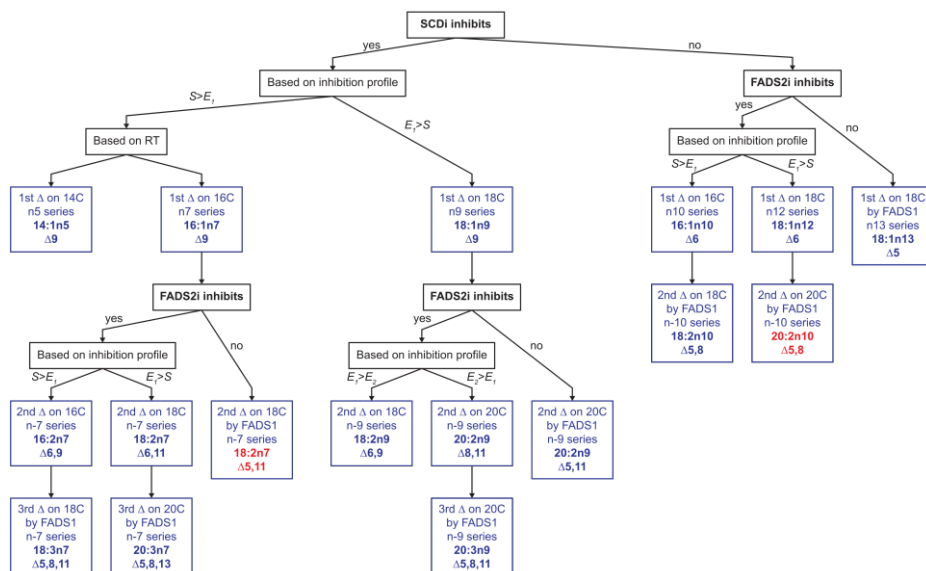


Figure 48. The algorithm employed to identify unknown FA by the reconstruction of their biosynthesis route. The depicted algorithm is applied to identify the double bond positions for FA based on the inhibition profile obtained upon incubation with U-¹³C-glucose, either with or without SCDi or FADS2i. The algorithm applies to FA whose origin can be tracked to FA(14:0)/FA(16:0). The previous assumptions must be met: 1) the FA incorporates labelling and intensity suffices to obtain values for all/most expected isotopomers; 2) FASNi decreases parameter S or distribution is consistent with the origin being FA(14:0)/FA(16:0). Based on the chromatographic profile, we expect the FA to elute by increasing n-series (i.e., RT(n5 series) ≤ RT(n7 series) ≤ RT(n9 series), etc.). The algorithm allows to identify the initial FA for the FA synthesis routes described in Figure 33; thus the actual position of the double bonds has to be extrapolated for the FA of a different carbon length to that indicated in the algorithm. In red, FA for which we can anticipate the identification and synthesis route based on the described strategy, but were not detected or unambiguously assigned experimentally in the A549 cells because FADS1 inhibitors were lacking.

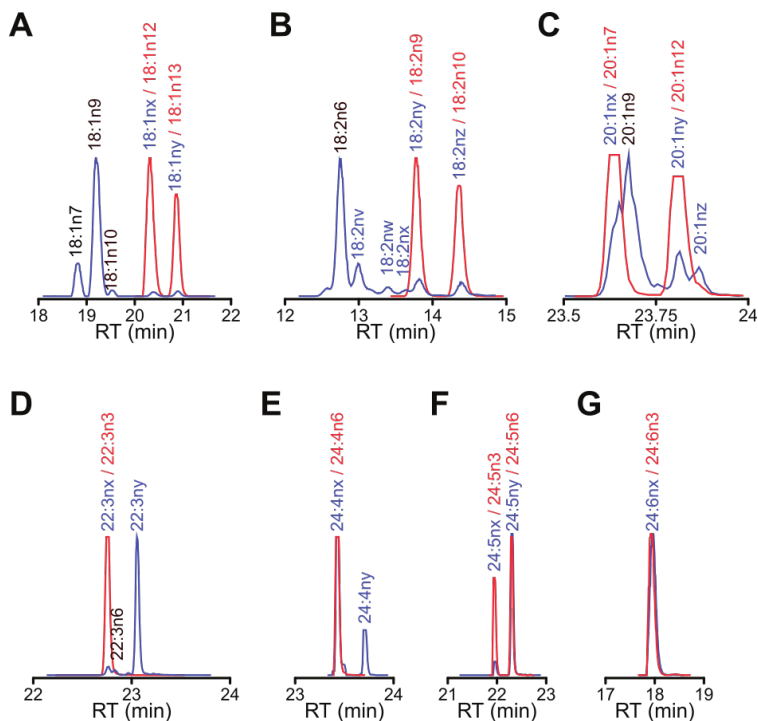


Figure 49. Confirmation of the identity of 11 unknown FA in the A549 cells with chemical standards. Chromatographic separation of A) FA 18:1, B) 18:2, C) 20:1, D) 22:3, E) 24:4, F) 24:5 and G) 24:6. In blue, the saponified FA from the A549 cells in culture. In red, chemical standards. Text in black, the FA that initially matched the chemical standards used to develop the method; in blue, a notation of the unknown FA detected in the A549 cells; in red, the chemical standards used to confirm the identity of the selected unknown FA

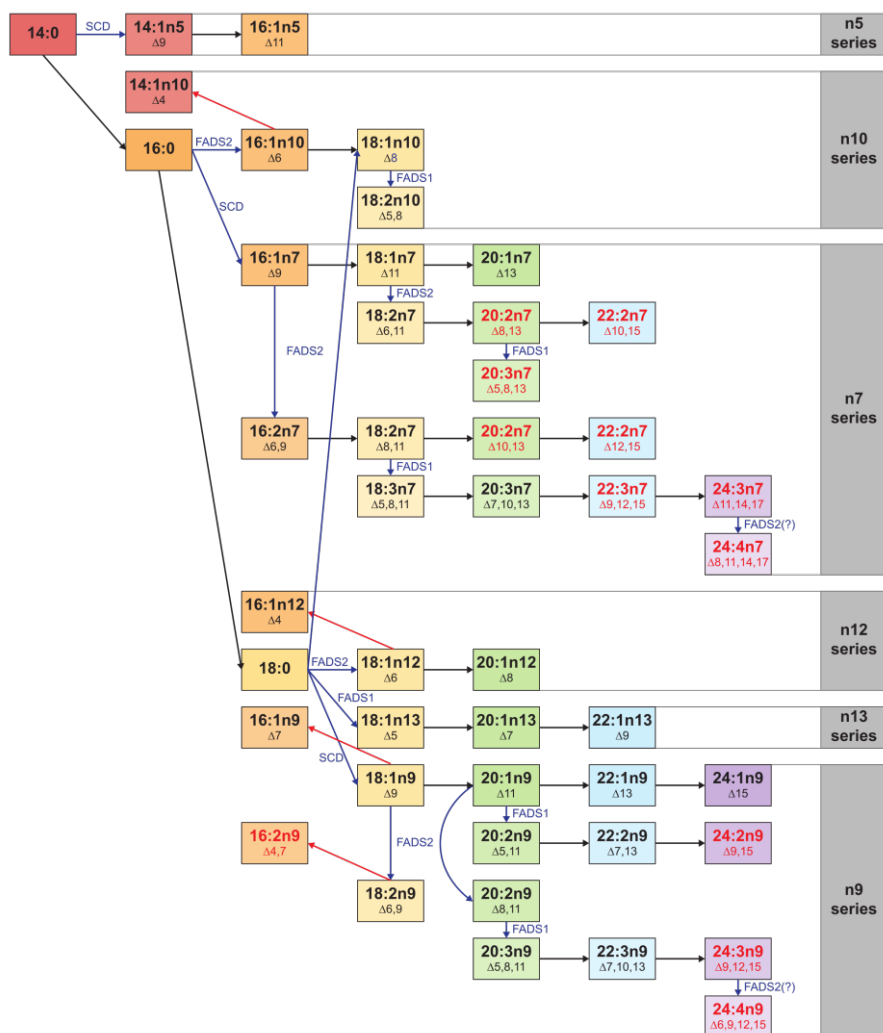


Figure 50. FA biosynthesis routes in the NSCLC cell line A549. Summary of the FA metabolism network in the A549 cells for those FA that come from DNL. Black arrows denote elongations, blue arrows desaturations (the responsible enzyme is indicated) and a red arrow degradation. Red depicts the FA that have not been previously described.

5. Future improvements of FAMetA

Future developments of mass isotopologue data analysis tools, including FAMetA, should address some unresolved issues like using labelled-FA as nutrients, distinguishing the uptake of exogenous FA and the lipolysis of stored lipids, estimating the synthesis rate of the FA that result from the degradation of a longer FA (e.g. FA(16:1)n9, where $S' = S * E_i * \text{degradation}$), or the resolution of the FA metabolism properties of particular lipid classes of interest or organelles. Additionally, the FAMetA algorithm is exclusively designed to fit the data from ^{13}C -based tracers for even-chain FA. Thus, future efforts should focus on implementing calculations based on ^2H -tracers, such as $^2\text{H}_2\text{O}$, which contributes to FA synthesis via direct H_2O incorporation, and also via $\text{NADPH}^{150,151}$, and to expand the reactions to cover odd-chain FA, in which not only the lipogenic Acetyl-CoA has to be estimated, but also the lipogenic Propionyl-CoA pool¹⁷⁹.

Conclusions

Based on the work presented here, the following conclusions can be drawn for Chapter 1, focused on LipidMS:

- 1) LipidMS was developed, an R-package aimed to lipid identification in untargeted LC-MS lipidomics.
- 2) LipidMS covers the whole workflow required to process untargeted LC-MS raw data from peak-picking to lipid identification.
- 3) Fragmentation rule-based identification implemented in LipidMS reduces false positive identifications and increases the level of structural elucidation compared to currently used tools (e.g., MS-DIAL).
- 4) LipidMS allows the simultaneous processing of data acquired using full scan, DIA and DDA modes, although only DIA and DDA data are used for lipid annotation.
- 5) The analysis of additivated vs non-additivated serum proves that LipidMS data processing workflow is valid for untargeted LC-MS analysis, as the results are comparable with those obtained using XCMS and MS-DIAL.
- 6) LipidMS, in its R version, is highly customizable allowing the modification of FA chains or sphingoid bases used to build the databases and even, the fragmentation rules employed for lipid identification.
- 7) Besides the R package, LipidMS can be used via web through www.lipidms.es so that no knowledge of R programming is required.

On the other hand, the following conclusions can be drawn for Chapter 2, focused on FAMetA:

- 1) FAMetA was developed, an R-package aimed to the analysis of the fatty acids metabolism based on mass isotopologue data.
- 2) FAMetA allows the analysis of the FA biosynthetic network, including the contribution of ^{13}C -tracers to the synthesis of FA, *de novo* synthesis, elongation and desaturation for FA up to 26C by the combined used of ^{13}C -tracers and LC-MS.
- 3) The quasi-multinomial distribution employed by FAMetA that allows modelling the data overdispersion observed in the FA isotopologue distributions through the Φ parameter.
- 4) FAMetA improves the assessment and interpretation of the elongation terms compared to FASA by individually estimating each elongation step instead of a sum of its multiple steps.
- 5) FAMetA includes an indirect estimation of FA desaturation based on the synthesis parameters of the precursor and the product FA of each reaction for the whole FA network.
- 6) The combined use of $\text{U-}^{13}\text{C}$ -glucose, well-known inhibitors of FA metabolism enzymes and data analysis with FAMetA allows the comprehensive characterization of the FA biosynthetic network, including the identification of 11 previously unknown FA.
- 7) Besides the R package, FAMetA can be used via web through www.fameta.es so that no knowledge of R programming is required.

Resumen en castellano

El desarrollo de la bioinformática y de las tecnologías analíticas han permitido la irrupción de las aproximaciones ómicas en la ciencia. Estas plataformas de perfilado molecular masivo tienen como objetivo la determinación del conjunto de biomoléculas (genes, proteínas, metabolitos, etc.) que forman parte de un sistema biológico. Entre ellas, la metabolómica pretende caracterizar el conjunto de metabolitos, moléculas de pequeño tamaño que actúan como precursores, intermediarios o productos finales del metabolismo. Los niveles de los metabolitos vienen determinados por todos aquellos procesos bioquímicos encargados de su producción, consumo y eliminación y, por tanto, son un reflejo directo del estado fisiológico del sistema biológico en estudio. La gran diversidad de propiedades físico-químicas de los metabolitos, que determinan en gran medida que técnicas analíticas deben utilizarse para su caracterización, han favorecido la aparición de subdisciplinas dentro de la metabolómica centradas en el análisis de un grupo concreto de metabolitos con características compartidas. Los lípidos son un subgrupo numeroso y heterogéneo de metabolitos que se caracterizan por su naturaleza hidrofóbica/anfifílica y que tienen una gran importancia biológica como intermediarios o productos de rutas de señalización, componentes estructurales de las membranas celulares y fuentes de energía. El análisis holístico de estos lípidos ha supuesto que la lipidómica se establezca como una subdisciplina de la metabolómica con entidad y características propias. El metabolismo de los lípidos juega un papel central en los sistemas biológicos y su estudio puede contribuir a la comprensión de los mecanismos que subyacen a diferentes condiciones patológicas. En los últimos años se han identificado alteraciones en los perfiles lipídicos generales y en especies lipídicas particulares en enfermedades de alta prevalencia como el cáncer, el hígado graso no alcohólico, la diabetes, las cardiopatías y las enfermedades neurológicas. Actualmente, se están

dirigiendo grandes esfuerzos para conocer no solo los mecanismos relacionados con los lípidos que subyacen a las enfermedades, sino también para encontrar nuevos biomarcadores que permitan predecir el diagnóstico, el pronóstico o la respuesta al tratamiento. Sin embargo, la mayoría de los biomarcadores lipídicos propuestos no están validados o no son útiles como biomarcadores clínicos debido a la falta de especificidad o sensibilidad de estas moléculas. Además, la interpretación biológica de las alteraciones del metabolismo de los lípidos es limitada porque aún se desconocen las funciones específicas de la mayoría de las especies de lípidos. En la mayoría de los casos, solo se utilizan los niveles globales de las clases de lípidos y los ácidos grasos libres totales para la interpretación de los resultados, pasando por alto la composición de las cadenas de ácidos grasos de los lípidos complejos. Por lo tanto, aún se requieren avances en métodos analíticos y herramientas bioinformáticas que mejoren el análisis del lipidoma para comprender completamente el metabolismo de los lípidos y sus implicaciones en cada enfermedad.

Actualmente, la espectrometría de masas acoplada a cromatografía líquida (LC-MS) es la técnica analítica más empleada para el análisis del metaboloma y del lipidoma. En LC-MS, los metabolitos se separan en primer lugar por cromatografía líquida para, a continuación, ser ionizados y detectados por espectrometría de masas. El resultado final es un conjunto de datos crudos caracterizados por tres variables, tiempo de retención (RT), relación masa-carga (m/z) e intensidad, que deben ser procesados para extraer las señales asociadas a los diferentes metabolitos presentes en las muestras. En función del objetivo de un análisis metabolómico llevado a cabo por LC-MS, se distinguen dos tipos de aproximaciones: metabolómica dirigida o *targeted*, cuyo objetivo es la cuantificación de un pequeño conjunto metabolitos bien caracterizados, y metabolómica no dirigida o *untargeted*, cuyo objetivo consiste en

conseguir la mayor cobertura posible del metaboloma. Las aproximaciones *targeted* se realizan con espectrómetros de masas de baja resolución, como puede ser un triple cuadrupolo (TQ, por sus siglas en inglés) y para cada metabolito de interés se deben definir a priori las características a emplear en su detección, esto es su ion molecular (precursores o *parent ions*) y los fragmentos característicos que se generan tras la fragmentación de los mismos en la celda de colisión (fragmentos o *daughter ions*). Estos equipos suelen trabajar en modo *multiple reaction monitoring* (MRM) en el que múltiples metabolitos de interés se detectan en base a las características mencionadas. En las aproximaciones no dirigidas, al no disponer de un conjunto predefinido de metabolitos de interés, los datos deben ser procesados con el objetivo de extraer las señales de la mayor cantidad posible metabolitos que, *a priori*, son desconocidos. La identificación de los metabolitos se realiza tanto en base a la masa exacta del ion molecular detectado como en base a su estructura, dilucidada gracias a la fragmentación del ion molecular. Por tanto, el análisis *untargeted* se suele realizar con equipos de alta resolución que además posean la capacidad de fragmentar los iones generados. En la mayoría de los casos los equipos disponen de un cuadrupolo que permite filtrar los iones de interés de forma previa a su fragmentación en la celda de colisión y posterior análisis. En función de si existe o no un filtrado previo de los iones en el cuadrupolo antes de ser introducidos en la celda de colisión, podemos distinguir entre adquisición dependiente de datos (DDA), en la que se seleccionan un número determinado de iones que son seleccionados en el cuadrupolo y posteriormente fragmentados o adquisición independiente de datos (DIA), en la que todos los iones que coeluyen en un momento determinado son introducidos en la celda de colisión. En el caso de los datos adquiridos en DDA existe una conexión directa entre los fragmentos generados y el precursor, mientras que en el caso de DIA se deben utilizar técnicas de análisis de datos para poder establecer la conexión/correlación entre los

precursores y sus correspondientes fragmentos. Los equipos más habituales para el análisis metabolómico *untargeted* son el cuadrupolo-tiempo de vuelo (QToF, por sus siglas en inglés) y el cuadrupolo-orbitrap.

A pesar del gran interés que ha despertado la lipidómica en los últimos años, la gran heterogeneidad, el tamaño del lipidoma y la falta de estándares comerciales han dificultado la correcta identificación de los lípidos en análisis por LC-MS no dirigida, lo que sigue suponiendo el principal cuello de botella en el avance del estudio del lipidoma. Además, como ya se ha mencionado, la interpretación biológica de los resultados es limitada debido a que las funciones específicas de la mayoría de las especies de lípidos son aún desconocidas. Por este motivo, el objetivo general planteado en esta tesis fue el desarrollo de nuevos métodos y herramientas bioinformáticas que faciliten la caracterización del lipidoma y el estudio del metabolismo de lípidos, particularmente ácidos grasos. Para ello se propusieron dos objetivos principales:

- 1) Desarrollo de una herramienta que mejore la anotación de lípidos en los análisis por LC-MS no dirigida. Esta herramienta debe cubrir todos los pasos necesarios para el procesamiento de los datos e implementar la anotación de lípidos basada en reglas de fragmentación para datos DDA y DIA.
- 2) Desarrollo de un método que permita el estudio del conjunto de reacciones implicadas en la biosíntesis de ácidos grasos basado en el uso combinado de LC-MS y trazadores de ^{13}C .

Esta tesis se divide en dos capítulos en los que se explican con detalle cada una de las dos herramientas desarrolladas a lo largo de esta tesis, LipidMS (Capítulo 1), un paquete de R para el procesamiento de datos de LC-MS no dirigida y la anotación de lípidos, y FAMetA (Capítulo 2), una herramienta basada en distribuciones de isotopólogos para el análisis

exhaustivo del metabolismo de los ácidos grasos, ambas con el objetivo de mejorar el análisis del lipidoma basado en espectrometría de masas.

Por un lado, LipidMS fue desarrollado con el objetivo específico de mejorar la identificación de lípidos en LC-MS mediante el uso de reglas de fragmentación. Como ya se ha mencionado, el tamaño, la complejidad y la heterogeneidad del lipidoma junto con la falta de estándares lipídicos disponibles, hacen de la anotación de lípidos uno de los pasos más limitantes y costosos del procesamiento de datos en los estudios lipídicos por LC-MS. La identificación precisa de cualquier metabolito en LC-MS, requiere la comprobación del RT, m/z y espectro MS/MS con un estándar disponible comercialmente. En el caso de los lípidos, debido a la enorme variedad de especies lipídicas y al reducido número de estándares disponibles, esta estrategia no puede aplicarse en la mayoría de los casos. En este sentido, la definición de patrones de fragmentación para diferentes clases de lípidos ha permitido la construcción *in silico* de librerías de espectros MS/MS que se utilizan para la anotación de lípidos mediante el uso de algoritmos de *spectral matching*. Sin embargo, esta estrategia presenta múltiples limitaciones. En primer lugar, un único valor de m/z para un precursor no es suficiente para identificar el ion molecular debido a la gran cantidad de solapamientos entre especies isoméricas e isobáricas, por lo que una correcta anotación de isótopos y aductos es de suma importancia en lipidómica no dirigida. Además, aunque la información del MS/MS puede ayudar a distinguir algunos de estos solapamientos, no es suficiente en muchos casos en los que se obtienen fragmentos comunes entre diferentes clases de lípidos o entre diferentes especies de una misma clase. Por otra parte, si el espectro MS/MS contiene un número reducido de fragmentos con intensidades elevadas, los cálculos de similitud entre espectros pueden estar sesgados dando lugar a resultados iguales o muy similares para diferentes especies isobáricas e isoméricas. Esto es muy frecuente en los lípidos, donde los fragmentos específicos de clase, que

sólo informan sobre la subclase de un lípido (por ejemplo, los fragmentos de la cabeza polar), o los fragmentos correspondientes a las cadenas de ácidos grasos que sólo informan sobre la composición de las cadenas, pero no sobre la clase o subclase de la especie lipídica de interés, son comunes a un gran número de especies. Por otro lado, cuando los compuestos isobáricos o isoméricos coeluyen durante la separación cromatográfica, lo que también es común debido a la naturaleza estructural de los lípidos a modo de bloques, se obtienen espectros MS/MS complejos tanto para los datos adquiridos en DDA como en DIA, lo que dificulta las anotaciones de lípidos. Como alternativa, la identificación de lípidos basada en reglas de fragmentación y en la presencia o ausencia de los fragmentos esperados para cada clase de lípido se ha implementado en un número reducido de herramientas bioinformáticas. En el momento en que se empezó esta tesis doctoral, solo unas pocas herramientas como LDA⁶⁸ o LipidMatch⁶⁹, estaban basados en reglas de fragmentación, y la mayoría, únicamente trabajaban con datos adquiridos en DDA. Por otro lado, MS-DIAL¹³³ permitía trabajar con datos adquiridos en DIA, pero la anotación de lípidos estaba basada en *spectral matching*. En versiones posteriores MS-DIAL incorporó la anotación basada en reglas de fragmentación a través de LipidMatch^{69,136}. En este contexto, LipidMS fue diseñado inicialmente con el objetivo de anotar lípidos en muestras individuales utilizando datos adquiridos en DIA y anotaciones basadas en reglas de fragmentación, aunque más tarde fue ampliado a DDA, ya que es el modo de adquisición más comúnmente utilizado. Por otro lado, LipidMS dependía inicialmente del uso de herramientas externas de procesamiento para analizar secuencias de múltiples muestras. Para superar esta limitación, las nuevas versiones del paquete han incorporado las funcionalidades necesarias para cubrir todo el flujo de trabajo en el procesamiento de los datos: extracción de picos, alineación, agrupación e integración de picos. Una vez generada la matriz con todas las señales detectadas en el *dataset*, LipidMS inicia la

identificación de lípidos en aquellas muestras adquiridas en DIA o DDA utilizando la información tanto de MS¹ como de MS². Con respecto a otras herramientas disponibles, LipidMS incorpora dos estrategias que ayudan a maximizar el número de asignaciones correctas y a minimizar las incorrectas. Por un lado, el conjunto de reglas de fragmentación ha sido definido de tal forma que se prioriza el uso de fragmentos específicos de clase bien caracterizados en lugar de fragmentos más intensos, pero menos específicos, como son las cadenas de ácidos grasos (que pueden ser comunes a gran cantidad de clases de lípidos). Por otro lado, los lípidos suelen ionizar en forma de múltiples aductos (p.ej. [M+H]⁺, [M+Na]⁺ y [M+NH₄]⁺, en el caso de ESI+). En muchas ocasiones los aductos de una especie lipídica concreta pueden ser confundidos con otra especie, por tanto, una correcta asignación de todos los aductos detectados para un lípido concreto de forma previa al análisis de los fragmentos generados contribuye a dar mayor robustez a las identificaciones generadas y a minimizar el número de anotaciones incorrectas. La última versión de LipidMS incluye las reglas de fragmentación predefinidas para 28 clases de lípidos y permite customizar tanto las reglas de fragmentación como los *building blocks* utilizados para generar las librerías necesarias para la identificación. En función de los fragmentos encontrados, cada especie identificada puede anotarse con diferentes niveles de elucidación estructural: a nivel de clase, cuando solo se han encontrado fragmentos característicos de la clase o subclase de lípido, lo que confirma el tipo de lípido y la composición total de carbonos y dobles enlaces pero no la composición de las cadenas; a nivel de composición de las cadenas de ácidos grasos, cuando además de los fragmentos de clase se han encontrado fragmentos específicos de estas cadenas; y a nivel de posición de las mismas, cuando las intensidades relativas de los fragmentos correspondientes a las cadenas permiten dilucidar la posición de cada uno de las ácidos grasos dentro de la estructura del lípido complejo. LipidMS

fue evaluado mediante el análisis de un suero humano comercial aditivado y no aditivado con un total de 68 estándares lipídicos y comparado con dos de los softwares más comúnmente empleados en el procesamiento de datos de metabolómica y lipidómica no dirigida: XCMS¹³⁴ y MS-DIAL¹³⁶. En primer lugar, la comparación con XCMS demuestra que los algoritmos de procesamiento implementados en la última versión de LipidMS funcionan correctamente ya que los resultados obtenidos con ambos softwares son similares. Por otro lado, la comparación con MS-DIAL demuestra que LipidMS reduce el número de identificaciones incorrectas y mejora el nivel de elucidación estructural de las especies identificadas pese a que MS-DIAL es capaz de anotar un número mucho mayor de especies, por lo que LipidMS y MS-DIAL podrían utilizarse de manera complementaria. También es importante subrayar que LipidMS soporta el procesamiento simultáneo de las siguientes combinaciones de modos de adquisición MS: todas las muestras adquiridas en DIA; todas las muestras adquiridas en DDA; combinación de muestras DIA y DDA; combinación de *full scan* y DIA; combinación de *full scan* y DDA; y combinación de *full scan*, DDA y DIA, lo que permite integrar con mayor facilidad y de manera automática los resultados de las anotaciones obtenidas en DIA y DDA con el resto de los datos. Futuras mejoras de LipidMS deberían incluir la ampliación de las clases de lípidos y de las cadenas de ácidos grasos y bases esfingoides utilizadas para ofrecer una mejor cobertura del lipidoma, la estandarización de LipidMS para hacerlo compatible con otros paquetes de R, o la posibilidad de analizar datos de lípidos marcados con trazadores isotópicos.

Por otro lado, FAMetA surgió como respuesta al segundo objetivo de esta tesis, que consistía en desarrollar una herramienta que facilite el estudio del metabolismo de los ácidos grasos. El uso de trazadores de ¹³C y detección basada en MS es el método de referencia para el análisis del metabolismo de los ácidos grasos. Este método se basa en la incorporación

sucesiva de unidades de dos carbonos marcadas con el isótopo estable del carbono ^{13}C , a través del acetyl-CoA, hacia los ácidos grasos durante las reacciones de síntesis y elongación y el posterior análisis de las distribuciones de isotopólogos obtenidas. Gracias a la diferencia de masa entre las especies preexistentes o las sintetizadas a través de fuentes no marcadas con respecto a las generadas a partir de la fuente que contiene ^{13}C , se puede realizar un análisis del metabolismo basado en la distribución de isotopólogos (especies de una misma molécula que difieren únicamente en su masa como consecuencia de la incorporación de ^{13}C en lugar del ^{12}C , que es la especie mayoritaria de forma natural). A pesar de que se han desarrollado varios algoritmos y herramientas para extraer información sobre el metabolismo de los ácidos grasos mediante la modelización de estas distribuciones de isotopólogos, estas siguen proporcionando una información limitada y difícil de interpretar^{129,153-155,166,167}. La mayoría de estos métodos únicamente proporcionan información sobre la lipogénesis *de novo* para los ácidos grasos de hasta 16 o 18 carbonos o no reflejan los pasos biológicos reales de los procesos de elongación. Además, la desaturación no se tiene en cuenta para la red completa de ácidos grasos. Con el fin de superar estas limitaciones, desarrollamos FAMetA, una herramienta que utiliza las distribuciones de isotopólogos de los ácidos grasos obtenidas por la incorporación de acetyl-CoA marcado con ^{13}C para estimar cada uno de los pasos de la mayoría de las reacciones biosintéticas implicadas en el metabolismo de los ácidos grasos: lipogénesis *de novo* (*S*), elongación (*E*), desaturación (Δ) e importación (*I*). Además, FAMetA permite estimar la contribución relativa del trazador empleado al *pool* de acetyl-CoA (D_0 , D_1 y D_2 , haciendo referencia a si contiene 0, 1 o 2 átomos de ^{13}C respectivamente). Para ácidos grasos de hasta 16 carbonos, la síntesis *de novo* se ha modelizado tradicionalmente utilizando distribuciones multinomiales que permiten la estimación de los parámetros *I*, *S* y D_0 , D_1 , D_2 . En FAMetA hemos reemplazado las distribuciones multinomiales por

distribuciones quasi-multinomiales que son capaces de modelizar y cuantificar (mediante el parámetro Φ) la sobredispersión habitualmente observada de forma experimental en las distribuciones obtenidas. Para los ácidos grasos de más de 16 carbonos, además de los parámetros S e I , también se estiman hasta cinco términos de elongación (E_n , haciendo referencia $n=1$ al primer paso de elongación para ácidos grasos de 18 carbonos y $n=5$ el último paso para ácidos grasos de 26 carbonos) que representan cada uno de los pasos de elongación individuales de un precursor con X átomos de carbonos, a un producto de longitud $X+2$. En comparación con herramientas anteriores, la forma en que FAMetA calcula las elongaciones, refleja mejor cómo se elongan los ácidos grasos dentro de las células, lo que permite una interpretación biológica directa de los parámetros de elongación estimados. Además, FAMetA incorpora la estimación indirecta de la desaturación para la red metabólica de los ácidos grasos mediante una estrategia que utiliza los parámetros de síntesis estimados para el precursor y el producto de la reacción de desaturación en lugar del marcaje total. Por último, el flujo de trabajo de FAMetA incluye todas las funcionalidades necesarias para el procesamiento de datos, las comparaciones por grupos y los resultados gráficos, lo que facilita la interpretación de los resultados. Para testar la validez de los algoritmos implementados en FAMetA, en primer lugar, se simuló un conjunto de distribuciones de isotopólogos a partir de valores conocidos de los diferentes parámetros calculados por FAMetA, y se comprobó que FAMetA es capaz de determinar con precisión el conjunto completo de parámetros de la síntesis de ácidos grasos (error relativo $< 15\%$, RSD $< 15\%$ para todos los parámetros) siempre que la contribución relativa del trazador (D_2) y los parámetros a calcular para un determinado ácido graso, es decir, S , E_1 , E_2 , E_3 y E_4 , se encuentren dentro del intervalo 0.05-0.9, lo que garantiza su aplicabilidad en un escenario biológico real. A continuación, FAMetA fue evaluado en diferentes escenarios biológicos tanto *in vivo* como *in vitro*,

con y sin la presencia de inhibidores conocidos de reacciones específicas del metabolismo de los ácidos grasos, comprobando que FAMetA permite determinar los parámetros asociados a estas reacciones la red metabólica completa y, además, en un escenario de uso de inhibidores, FAMetA es capaz de detectar los cambios específicos inducidos en el metabolismo. Además, comparado con FASA¹²⁹, la única herramienta que hasta el momento incluía el análisis de ácidos grasos elongados más allá de 18 carbonos, FAMetA proporciona una caracterización más completa de la red biosintética de los ácidos grasos, una descripción mejor y más intuitiva de cada uno de los parámetros de síntesis y un flujo de trabajo más completo que va desde el procesamiento de datos hasta las comparaciones basadas en grupos y la representación gráfica. Por último, el uso de inhibidores específicos combinado con el análisis de FAMetA, nos ha permitido estudiar en profundidad la red metabólica de biosíntesis de ácidos grasos en células A549, identificando 33 ácidos grasos *a priori* desconocidos, 11 de los cuales pudieron ser confirmados con estándares comerciales. Además, 12 de ellos no han sido previamente descritos en mamíferos, aunque pertenecen a series n/omega ya descritas. Futuras versiones de FAMetA deberían incorporar el análisis de otro tipo de trazadores a parte de los de ¹³C, permitir el uso de ácidos grasos marcados como trazadores, ampliar la red de reacciones para incluir los ácidos grasos de cadena impar y abordar la degradación. En resumen, en comparación con herramientas anteriores, FAMetA ofrece: i) la caracterización de una red biosintética de ácidos grasos más amplia ya que incluye en una única herramienta el análisis de síntesis *de novo*, elongación y desaturación; ii) la posibilidad de ejecutar los pasos necesarios desde el procesamiento de datos hasta el análisis del metabolismo de los ácidos grasos y la representación gráfica en una única herramienta; iii) un entorno de fácil manejo gracias a su implementación como un paquete de R y una versión web con interfaz gráfica; iv) mejor ajuste a los datos experimentales gracias a la

implementación de un ajuste quasi-multinomial que incluye el parámetro Φ para tener en cuenta la sobredispersión de los datos; v) mejor modelado de las reacciones de elongación, lo que permite una interpretación más sencilla de los parámetros estimados; y vi) parámetros y representaciones gráficas fáciles de interpretar que permiten obtener conclusiones biológicas significativas.

Para ambas herramientas, tanto LipidMS como FAMetA, se han implementado versiones web utilizando el paquete Shiny con el objetivo de hacer su uso más accesible para la mayoría de usuarios ya que ofrece una interfaz gráfica intuitiva que no requiere conocimientos de R ni instalación previa. LipidMS se encuentra disponible en <http://www.lipidms.es/> y FAMetA en <https://www.fameta.es/>.

En base al trabajo desarrollado en esta tesis, las siguientes conclusiones pueden ser extraídas para el Capítulo 1, centrado en LipidMS:

- 1) Se ha desarrollado LipidMS, un paquete de R dirigido a la identificación de lípidos en lipidómica no dirigida basada en LC-MS.
- 2) LipidMS cubre todo el flujo de trabajo necesario para procesar datos de LC-MS no dirigida, desde la extracción de picos hasta la identificación de lípidos.
- 3) La identificación basada en reglas de fragmentación implementada en LipidMS reduce el número de falsos positivos en la identificación de lípidos y mejora el nivel de elucidación estructural en comparación con las herramientas utilizadas actualmente como, por ejemplo, MS-DIAL.
- 4) LipidMS permite el procesamiento simultáneo de datos adquiridos en full *scan*, DIA y DDA, aunque para la anotación únicamente se utilizan aquellos adquiridos en DIA y DDA.

- 5) El análisis de suero aditivado frente a no aditivado demuestra que el procesamiento de datos implementado en LipidMS es válido para el análisis por LC-MS no dirigida, ya que los resultados son comparables a los obtenidos mediante XCMS y MS-DIAL.
- 6) LipidMS, en su versión de paquete de R, es altamente personalizable permitiendo la modificación de las cadenas de ácidos grasos o de las bases esfingoides utilizadas para construir las bases de datos y las reglas de fragmentación empleadas para la identificación de lípidos.
- 7) Además del paquete de R, LipidMS puede ser utilizado a través de su herramienta web (www.lipidms.es) sin necesidad de tener conocimientos de programación en R.

Por otro lado, las siguientes conclusiones pueden ser extraídas del Capítulo 2, dedicado a FAMetA:

- 1) Se ha desarrollado FAMetA, un paquete de R dirigido al análisis del metabolismo de ácidos grasos basado en distribuciones de isotopólogos.
- 2) FAMetA permite analizar la red biosintética de los ácidos grasos incluyendo la contribución de los trazadores de ^{13}C a la síntesis de ácidos grasos, síntesis *de novo*, elongación y desaturación para ácidos grasos de hasta 26C mediante el uso combinado de trazadores ^{13}C y LC-MS.
- 3) La distribución quasi-multinomial empleada en FAMetA, que incluye el parámetro Φ , permite modelizar la sobredispersión observada en las distribuciones de isotopólogos de los ácidos grasos.
- 4) FAMetA mejora la estimación y la interpretación de los términos de elongación comparado con FASA, al estimar individualmente cada paso de elongación en lugar de una suma de múltiples pasos.

- 5) FAMetA incluye la estimación indirecta de la desaturación basada en los parámetros de síntesis del precursor y el producto de cada reacción para la red completa de ácidos grasos.
- 6) El uso combinado de glucosa completamente marcada con ^{13}C e inhibidores de enzimas del metabolismo de ácidos grasos junto con el análisis llevado a cabo por FAMetA permite la caracterización exhaustiva de la red biosintética de los ácidos grasos, incluyendo la identificación de 11 ácidos grasos a desconocidos previamente.
- 7) Además del paquete de R, FAMetA puede utilizarse a través de su herramienta web (www.fameta.es) sin necesidad de tener conocimientos previos de programación en R.

References

1. Kuehnbaum, N. L. & Britz-Mckibbin, P. New advances in separation science for metabolomics: Resolving chemical diversity in a post-genomic era. *Chemical Reviews* vol. 113 2437–2468 (2013).
2. Dettmer, K., Aronov, P. A. & Hammock, B. D. Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews* vol. 26 51–78 (2007).
3. Begou, O., Gika, H. G., Wilson, I. D. & Theodoridis, G. Hyphenated MS-based targeted approaches in metabolomics. *Analyst* vol. 142 3079–3100 (2017).
4. Jacob, M., Lopata, A. L., Dasouki, M. & Abdel Rahman, A. M. Metabolomics toward personalized medicine. *Mass Spectrometry Reviews* vol. 38 221–238 (2019).
5. Zamboni, N., Saghatelian, A. & Patti, G. J. Defining the Metabolome: Size, Flux, and Regulation. *Molecular Cell* vol. 58 699–706 (2015).
6. López-López, Á., López-González, Á., Barker-Tejeda, T. C. & Barbas, C. A review of validated biomarkers obtained through metabolomics. *Expert Review of Molecular Diagnostics* vol. 18 557–575 (2018).
7. Johnson, C. H., Ivanisevic, J. & Siuzdak, G. Metabolomics: Beyond biomarkers and towards mechanisms. *Nature Reviews Molecular Cell Biology* vol. 17 451–459 (2016).
8. Roca, M., Alcoriza, M. I., Garcia-Cañaveras, J. C. & Lahoz, A. Reviewing the metabolome coverage provided by LC-MS: Focus on sample preparation and chromatography-A tutorial. *Analytical Chimica Acta* 1147, 38–55 (2021).
9. Alonso, A., Marsal, S. & Julià, A. Analytical methods in untargeted metabolomics: State of the art in 2015. *Frontiers in Bioengineering and Biotechnology* vol. 3 (2015).

10. Emwas, A. H. M. The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research. *Methods in Molecular Biology* 1277, 161–193 (2015).
11. Chen, S. et al. Pseudotargeted metabolomics method and its application in serum biomarker discovery for hepatocellular carcinoma based on ultra high-performance liquid chromatography/triple quadrupole mass spectrometry. *Analytical Chemistry* 85, 8326–8333 (2013).
12. García-Cañaveras, J. C., Donato, M. T., Castell, J. v. & Lahoz, A. Targeted profiling of circulating and hepatic bile acids in human, mouse, and rat using a UPLC-MRM-MS-validated method. *Journal of Lipid Research* 53, 2231–2241 (2012).
13. Schrimpe-Rutledge, A. C., Codreanu, S. G., Sherrod, S. D. & McLean, J. A. Untargeted Metabolomics Strategies—Challenges and Emerging Directions. *Journal of the American Society Mass Spectrometry* 27, 1897–1905 (2016).
14. Dunn, W. B., Wilson, I. D., Nicholls, A. W. & Broadhurst, D. The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. *Bioanalysis* 4, 2249–2264 (2012).
15. Bijlsma, S. et al. Large-scale human metabolomics studies: A strategy for data (pre-) processing and validation. *Analytical Chemistry* 78, 567–574 (2006).
16. Begou, O., Gika, H. G., Theodoridis, G. A. & Wilson, I. D. Quality control and validation issues in LC-MS metabolomics. in *Methods in Molecular Biology* vol. 1738 15–26 (2018).
17. León, Z., García-Cañaveras, J. C., Donato, M. T. & Lahoz, A. Mammalian cell metabolomics: Experimental design and sample preparation. *Electrophoresis* vol. 34 2762–2775 (2013).

18. Vuckovic, D. Current trends and challenges in sample preparation for global metabolomics using liquid chromatography-mass spectrometry. *Analytical and Bioanalytical Chemistry* vol. 403 1523–1548 (2012).
19. Lu, W. et al. Metabolite measurement: Pitfalls to avoid and practices to follow. *Annual Review of Biochemistry* vol. 86 277–304 (2017).
20. Wang, Y., Liu, S., Hu, Y., Li, P. & Wan, J. B. Current state of the art of mass spectrometry-based metabolomics studies - a review focusing on wide coverage, high throughput and easy identification. *RSC Advances* vol. 5 78728–78737 (2015).
21. Knolhoff, A. M., Kneapler, C. N. & Croley, T. R. Optimized chemical coverage and data quality for non-targeted screening applications using liquid chromatography/high-resolution mass spectrometry. *Analytical Chimica Acta* 1066, 93–101 (2019).
22. Wishart, D. S. et al. HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Res* 46, D608–D617 (2018).
23. Petras, D. et al. GNPS Dashboard: collaborative exploration of mass spectrometry data in the web browser. *Nature Methods* vol. 19 134–136 (2022).
24. Horai, H. et al. MassBank: A public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry* 45, 703–714 (2010).
25. Domingo-Almenara, X. et al. The METLIN small molecule dataset for machine learning-based retention time prediction. *Nature Communications* 10 (2019).
26. LIPIDMAPS database. <https://www.lipidmaps.org/>.
27. mzCloud database. <https://www.mzcloud.org/>.
28. Koelmel, J. P. et al. Expanding Lipidome Coverage Using LC-MS/MS Data-Dependent Acquisition with Automated Exclusion

- List Generation. *Journal of the American Society Mass Spectrometry* 28, 908–917 (2017).
29. Guo, J. & Huan, T. Comparison of Full-Scan, Data-Dependent, and Data-Independent Acquisition Modes in Liquid Chromatography-Mass Spectrometry Based Untargeted Metabolomics. *Analytical Chemistry* 92, 8072–8080 (2020).
 30. Stolt, R. et al. Second-order peak detection for multicomponent high-resolution LC/MS data. *Analytical Chemistry* 78, 975–983 (2006).
 31. Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry* 78, 779–787 (2006).
 32. Tautenhahn, R., Bottcher, C. & Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* 9 (2008).
 33. Wang, S. Y., Kuo, C. H. & Tseng, Y. J. Ion trace detection algorithm to extract pure ion chromatograms to improve untargeted peak detection quality for liquid chromatography/time-of-flight mass spectrometry-based metabolomics data. *Analytical Chemistry* 87, 3048–3055 (2015).
 34. Martin Loos. *enviPick: Peak Picking for High Resolution Mass Spectrometry Data*. <https://github.com/blosloos/enviPick>.
 35. Danielsson, R., Bylund, D. & Markides, K. E. Matched filtering with background suppression for improved quality of base peak chromatograms and mass spectra in liquid chromatography-mass spectrometry. *Analytica Chimica Acta* vol. 454 (2002).
 36. Du, P., Kibbe, W. A. & Lin, S. M. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics* 22, 2059–2065 (2006).

37. Smith, R., Ventura, D. & Prince, J. T. LC-MS alignment in theory and practice: A comprehensive algorithmic review. *Brief Bioinform* 16, 104–117 (2013).
38. Vest Nielsen, N.-P., Carstensen, J. M. & Smedsgaard, J. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A* vol. 805 www.imm.dtu.dk (1998).
39. Christin, C. et al. Optimized time alignment algorithm for LC-MS data: Correlation optimized warping using component detection algorithm-selected mass chromatograms. *Analytical Chemistry* 80, 7012–7021 (2008).
40. Kassidas, A., Macgregor, J. F. & Taylor, P. A. Synchronization of Batch Trajectories Using Dynamic Time Warping.
41. Prince, J. T. & Marcotte, E. M. Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. *Analytical Chemistry* 78, 6140–6152 (2006).
42. Katajamaa, M., Miettinen, J. & Orešič, M. MZmine: Toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* 22, 634–636 (2006).
43. Pluskal, T., Castillo, S., Villar-Briones, A. & Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* 11, (2010).
44. Sumner, L. W. et al. Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* 3, 211–221 (2007).
45. Kind, T. et al. Identification of small molecules using accurate mass MS/MS search. *Mass Spectrometry Reviews* vol. 37 513–532 (2018).

46. Massbank of North America (MoNA).
<https://mona.fiehnlab.ucdavis.edu/>.
47. Misra, B. B. Data normalization strategies in metabolomics: Current challenges, approaches, and tools. *European Journal of Mass Spectrometry* 26, 165-174 (2020).
48. Hendriks, M. M. W. B. et al. Data-processing strategies for metabolomics studies. *TrAC - Trends in Analytical Chemistry* vol. 30 1685-1698 (2011).
49. Cuevas-Delgado, P., Dudzik, D., Miguel, V., Lamas, S. & Barbas, C. Data-dependent normalization strategies for untargeted metabolomics—a case study. *Analytical BioAnalytical Chemistry* 412, 6391-6405 (2020).
50. Brunius, C., Shi, L. & Landberg, R. Large-scale untargeted LC-MS metabolomics data correction using between-batch feature alignment and cluster-based within-batch signal intensity drift correction. *Metabolomics* 12 (2016).
51. Schiffman, C. et al. Filtering procedures for untargeted lc-ms metabolomics data. *BMC Bioinformatics* 20 (2019).
52. Stanstrup, J. et al. *The MetaRbolomics book*.
53. Bartel, J., Krumsiek, J. & Theis, F. J. Statistical methods for the analysis of high-throughput metabolomics data. *Computational and Structural Biotechnology Journal* vol. 4 e201301009 (2013).
54. Gromski, P. S. et al. A comparative investigation of modern feature selection and classification approaches for the analysis of mass spectrometry data. *Analytical Chimica Acta* 829, 1-8 (2014).
55. Degenhardt, F., Seifert, S. & Szymczak, S. Evaluation of variable selection methods for random forests and omics data sets. *Brief Bioinform* 20, 492-503 (2019).

56. Acharjee, A. Comparison of Regularized Regression Methods for ~Omics Data. *Journal of Postgenomics Drug & Biomarker Development* 03 (2012).
57. Hervás, D., Prats-Montalbán, J. M., García-Cañaveras, J. C., Lahoz, A. & Ferrer, A. Sparse N-way partial least squares by L1-penalization. *Chemometrics and Intelligent Laboratory Systems* 185, 85-91 (2019).
58. Xia, J. & Wishart, D. S. MSEA: A web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Research* vol. 38 (2010).
59. Xia, J., Wishart, D. S. & Valencia, A. MetPA: A web-based metabolomics tool for pathway analysis and visualization. in *Bioinformatics* vol. 27 2342-2344 (Oxford University Press, 2011).
60. Li, S. et al. Predicting Network Activity from High Throughput Metabolomics. *PLoS Computational Biology* 9 (2013).
61. Chong, J. et al. MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis. *Nucleic Acids Research* 46, W486-W494 (2018).
62. Djoumbou-Feunang, Y. et al. BioTransformer: A comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *Journal of Cheminformatics* 11 (2019).
63. Brunk, E. et al. Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nature Biotechnology* 36, 272-281 (2018).
64. Romero, P. et al. Open Access Computational prediction of human metabolic pathways from the complete human genome. vol. 6 (2004).

65. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* vol. 28 (2000).
66. Heller, S. R., McNaught, A., Pletnev, I., Stein, S. & Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J Cheminform* 7 (2015).
67. Kind, T. et al. LipidBlast in silico tandem mass spectrometry database for lipid identification. *Nat Methods* 10, 755–758 (2013).
68. Hartler, J. et al. Deciphering lipid structures based on platform-independent decision rules. *Nat Methods* 14, 1171–1174 (2017).
69. Koelmel, J. P. et al. LipidMatch: An automated workflow for rule-based lipid identification using untargeted high-resolution tandem mass spectrometry data. *BMC Bioinformatics* 18 (2017).
70. Züllig, T., Trötz Müller, M. & Köfeler, H. C. Lipidomics from sample preparation to data analysis: a primer. *Analytical and Bioanalytical Chemistry* vol. 412 2191–2209 (2020).
71. Wenk MR. The emerging field of lipidomics. *Nat Rev Drug Discov* 4, 594–610 (2005).
72. Fahy, E. et al. A comprehensive classification system for lipids. *Journal of Lipid Research* 46, 839–861 (2005).
73. Fahy, E. et al. Update of the LIPID MAPS comprehensive classification system for lipids. *Journal of Lipid Research* vol. 50 (2009).
74. Quehenberger, O. et al. Lipidomics reveals a remarkable diversity of lipids in human plasma. *Journal of Lipid Research* 51, 3299–3305 (2010).
75. Blanco, A. & Blanco, G. Chapter 5 - Lipids. *Medical Biochemistry* 99–119 (Academic Press, 2017).

76. Pietrocola, F., Galluzzi, L., Bravo-San Pedro, J. M., Madeo, F. & Kroemer, G. Acetyl coenzyme A: A central metabolite and second messenger. *Cell Metabolism* vol. 21 805–821 (2015).
77. Batchuluun, B., Pinkosky, S. L. & Steinberg, G. R. Lipogenesis inhibitors: therapeutic opportunities and challenges. *Nature Reviews Drug Discovery* vol. 21 283–305 (2022).
78. Broadfield, L. A., Pane, A. A., Talebi, A., Swinnen, J. v. & Fendt, S. M. Lipid metabolism in cancer: New perspectives and emerging mechanisms. *Developmental Cell* vol. 56 1363–1393 (2021).
79. Bittman, R. Glycerolipids: Chemistry. in *Encyclopedia of Biophysics* (ed. Roberts, G. C. K.) 907–914 (Springer Berlin Heidelberg, 2013).
80. Gyamfi, D., Awuah, E. O. & Owusu, S. Lipid metabolism: An overview. *The Molecular Nutrition of Fats* 17–32 (Elsevier, 2018).
81. Diagne Joseite Fauvel, A., Record, M., Chap, H. & Douste-blazy, L. Studies on ether phospholipids ii. Comparative composition of various tissues from human, rat and guinea pig. *Biochimica et Biophysica Acta* vol. 793 (1984).
82. van der Veen, J. N. et al. The critical role of phosphatidylcholine and phosphatidylethanolamine metabolism in health and disease. *Biochimica et Biophysica Acta - Biomembranes* vol. 1859 1558–1572 (2017).
83. Henneberry, A. L., Wright, M. M. & McMaster, C. R. The major sites of cellular phospholipid synthesis and molecular determinants of fatty acid and lipid head group specificity. *Mol Biol Cell* 13, 3148–3161 (2002).
84. Wright, M. M. & McMaster, C. R. PC and PE Synthesis: Mixed Micellar Analysis of the Cholinephosphotransferase and Ethanolaminephosphotransferase Activities of Human

- Choline/Ethanolamine Phosphotransferase 1 (CEPT1). *Lipids* 37, 663–672 (2002).
85. Vance, J. E. Phosphatidylserine and phosphatidylethanolamine in mammalian cells: Two metabolically related aminophospholipids. *Journal of Lipid Research* vol. 49 1377–1387 (2008).
 86. Balla, T. Phosphoinositides: Tiny Lipids With Giant Impact on Cell Regulation. *Physiol Rev* 93, 1019–1137 (2013).
 87. Fujita, M. & Kinoshita, T. GPI-anchor remodeling: Potential functions of GPI-anchors in intracellular trafficking and membrane dynamics. *Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids* vol. 1821 1050–1058 (2012).
 88. Leventis, P. A. & Grinstein, S. The distribution and function of phosphatidylserine in cellular membranes. *Annual Review of Biophysics* vol. 39 407–427 (2010).
 89. Thakur, R., Naik, A., Panda, A. & Raghu, P. Regulation of membrane turnover by phosphatidic acid: Cellular functions and disease implications. *Frontiers in Cell and Developmental Biology* vol. 7 (2019).
 90. Dudek, J. Role of cardiolipin in mitochondrial signaling pathways. *Frontiers in Cell and Developmental Biology* vol. 5 (2017).
 91. Lykidis, A., Jackson, P. D., Rock, C. O. & Jackowski, S. The Role of CDP-Diacylglycerol Synthetase and Phosphatidylinositol Synthase Activity Levels in the Regulation of Cellular Phosphatidylinositol. *Journal of Biological Chemistry* (1997).
 92. Stuhne-Sekalec, L., Chudzik, J. & Stanacev, N. Z. Participation of the microsomal CDP-diglycerides in the mitochondrial biosynthesis of phosphatidylglycerol. *Biochemistry and Cell Biology* 64 (1986).

93. Houtkooper, R. H. et al. Identification and characterization of human cardiolipin synthase. *FEBS Lett* 580, 3059–3064 (2006).
94. Stone, S. J. & Vance, J. E. Phosphatidylserine synthase-1 and -2 are localized to mitochondria-associated membranes. *Journal of Biological Chemistry* 275, 34534–34540 (2000).
95. Carreira, A. C. et al. Mammalian sphingoid bases: Biophysical, physiological and pathological properties. *Progress in Lipid Research* vol. 75 (2019).
96. Futerman, A. H. Sphingolipids. in *Biochemistry of Lipids, Lipoproteins and Membranes: Sixth Edition* 297–326 (Elsevier Inc., 2016).
97. Pralhada Rao, R. et al. Sphingolipid Metabolic Pathway: An Overview of Major Roles Played in Human Diseases. *Journal of Lipids* 2013, 1–12 (2013).
98. Chen, yang & Cao, Y. The sphingomyelin synthase family: proteins, diseases and inhibitors. *Biol Chem* (2017)
99. Mouritsen OG & Zuckermann MJ. What's So Special About Cholesterol? *Lipids* 39 (2004).
100. Piper, D. E. et al. The high-resolution crystal structure of human LCAT1. *Journal of Lipid Research* 56, 1711–1719 (2015).
101. Hilvo, M. et al. Novel theranostic opportunities offered by characterization of altered membrane lipid metabolism in breast cancer progression. *Cancer Research* 71, 3236–3245 (2011).
102. Patterson, A. D. et al. Aberrant lipid metabolism in hepatocellular carcinoma revealed by plasma metabolomics and lipid profiling. *Cancer Res* 71, 6590–6600 (2011).
103. Puri, P. et al. A lipidomic analysis of nonalcoholic fatty liver disease. *Hepatology* 46, 1081–1090 (2007).

104. García-Cañaveras, J. C. et al. A lipidomic cell-based assay for studying drug-induced phospholipidosis and steatosis. *Electrophoresis* 38, 2331-2340 (2017).
105. Rhee, E. P. et al. Lipid profiling identifies a triacylglycerol signature of insulin resistance and improves diabetes prediction in humans. *Journal of Clinical Investigation* 121, 1402-1411 (2011).
106. Meikle, P. J. et al. Plasma lipidomic analysis of stable and unstable coronary artery disease. *Arterioscler Thromb Vasc Biol* 31, 2723-2732 (2011).
107. Han, X. et al. Metabolomics in early Alzheimer's disease: Identification of altered plasma sphingolipidome using shotgun lipidomics. *PLoS One* 6, (2011).
108. Yang, L. et al. Recent advances in lipidomics for disease research. *Journal of Separation Science* vol. 39 38-50 (2016).
109. Butler, L. M. et al. Lipids and cancer: Emerging roles in pathogenesis, diagnosis and therapeutic intervention. *Advanced Drug Delivery Reviews* vol. 159 245-293 (2020).
110. Hao, Y. et al. Investigation of lipid metabolism dysregulation and the effects on immune microenvironments in pan-cancer using multiple omics data. *BMC Bioinformatics* 20 (2019).
111. Dória, M. L. et al. Fatty acid and phospholipid biosynthetic pathways are regulated throughout mammary epithelial cell differentiation and correlate to breast cancer survival. *FASEB Journal* 28, 4247-4264 (2014).
112. García-Cañaveras, J. C. & Lahoz, A. Tumor microenvironment-derived metabolites: A guide to find new metabolic therapeutic targets and biomarkers. *Cancers* vol. 13 (2021).

113. Padanad, M. S. et al. Fatty Acid Oxidation Mediated by Acyl-CoA Synthetase Long Chain 3 Is Required for Mutant KRAS Lung Tumorigenesis. *Cell Reports* 16, 1614–1628 (2016).
114. Pascual, G. et al. Targeting metastasis-initiating cells through the fatty acid receptor CD36. *Nature* 541, 41–45 (2017).
115. Capece, D. & Franzoso, G. Rewired lipid metabolism as an actionable vulnerability of aggressive colorectal carcinoma. *Molecular and Cellular Oncology* vol. 9 (2022).
116. Camarda, R. et al. Inhibition of fatty acid oxidation as a therapy for MYC-overexpressing triple-negative breast cancer. *Nature Medicine* 22, 427–432 (2016).
117. Wang, Y. Y. et al. Mammary adipocytes stimulate breast cancer invasion through metabolic remodeling of tumor cells. *Journal of Clinical Investigation* 2 (2017).
118. Nieman, K. M. et al. Adipocytes promote ovarian cancer metastasis and provide energy for rapid tumor growth. *Nature Medicine* 17, 1498–1503 (2011).
119. Vriens, K. et al. Evidence for an alternative fatty acid desaturation pathway increasing cancer plasticity. *Nature* 566, 403–406 (2019).
120. Jang, C., Chen, L. & Rabinowitz, J. D. Metabolomics and Isotope Tracing. *Cell* vol. 173 822–837 (2018).
121. Fernández-García, J., Altea-Manzano, P., Pranzini, E. & Fendt, S. M. Stable Isotopes for Tracing Mammalian-Cell Metabolism In Vivo. *Trends in Biochemical Sciences* vol. 45 185–201 (2020).
122. Wilkinson, D. J. Historical and contemporary stable isotope tracer approaches to studying mammalian protein metabolism. *Mass Spectrometry Reviews* vol. 37 57–80 (2018).
123. Schlame, M., Xu, Y., Erdjument-Bromage, H., Neubert, T. A. & Ren, M. Lipidome-wide ¹³C flux analysis: A novel tool to

- estimate the turnover of lipids in organisms and cultures. *Journal of Lipid Research* 61, 95-104 (2020).
124. Boumann, H. A. et al. The two biosynthetic routes leading to phosphatidylcholine in yeast produce different sets of molecular species. Evidence for lipid remodeling. *Biochemistry* 42, 3054-3059 (2003).
125. Skotland, T. et al. Determining the Turnover of Glycosphingolipid Species by Stable-Isotope Tracer Lipidomics. *Journal of Molecular Biology* 428, 4856-4866 (2016).
126. Tumanov, S., Bulusu, V. & Kamphorst, J. J. Analysis of Fatty Acid Metabolism Using Stable Isotope Tracers and Mass Spectrometry. *Methods in Enzymology* vol. 561 197-217 (Academic Press Inc., 2015).
127. Kamphorst, J. J., Fan, J., Lu, W., White, E. & Rabinowitz, J. D. Liquid chromatography-high resolution mass spectrometry analysis of fatty acid metabolism. *Analytical Chemistry* 83, 9114-9122 (2011).
128. Zheng, J. et al. Stable isotope labeling combined with liquid chromatography-tandem mass spectrometry for comprehensive analysis of short-chain fatty acids. *Analytical Chimica Acta* 1070, 51-59 (2019).
129. Argus, J. P. et al. Development and Application of FASA, a Model for Quantifying Fatty Acid Metabolism Using Stable Isotope Labeling. *Cell Rep* 25, 2919-2934.e8 (2018).
130. Tredwell, G. D. & Keun, H. C. ConvISA: A simple, convoluted method for isotopomer spectral analysis of fatty acids and cholesterol. *Metabolic Engineering* 32, 125-132 (2015).
131. Han, X., Yang, K. & Gross, R. W. Multi-dimensional mass spectrometry-based shotgun lipidomics and novel strategies for

- lipidomic analyses. *Mass Spectrometry Reviews* vol. 31 134–178 (2012).
132. Köfeler, H. C. et al. Recommendations for good practice in ms-based lipidomics. *Journal of Lipid Research* vol. 62 38 (2021).
133. Tsugawa, H. et al. MS-DIAL: Data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature Methods* 12, 523–526 (2015).
134. Tautenhahn, R., Patti, G. J., Rinehart, D. & Siuzdak, G. XCMS online: A web-based platform to process untargeted metabolomic data. *Analytical Chemistry* 84, 5035–5039 (2012).
135. Koelmel, J. P. et al. LipidMatch: An automated workflow for rule-based lipid identification using untargeted high-resolution tandem mass spectrometry data. *BMC Bioinformatics* 18, (2017).
136. Tsugawa, H. et al. A lipidome atlas in MS-DIAL 4. *Nature Biotechnology* 38, 1159–1163 (2020).
137. Kyle, J. E. et al. LIQUID: An open source software for identifying lipids in LC-MS/MS-based lipidomics data. *Bioinformatics* 33, 1744–1746 (2017).
138. Ni, Z., Angelidou, G., Lange, M., Hoffmann, R. & Fedorova, M. LipidHunter Identifies Phospholipids by High-Throughput Processing of LC-MS and Shotgun Lipidomics Datasets. *Analytical Chemistry* 89, 8800–8807 (2017).
139. Ni, Z., Angelidou, G., Hoffmann, R. & Fedorova, M. LPptiger software for lipidome-specific prediction and identification of oxidized phospholipids from LC-MS datasets. *Science Reports* 7 (2017).
140. Goracci, L. et al. Lipostar, a comprehensive platform-neutral cheminformatics tool for lipidomics. *Analytical Chemistry* 89, 6257–6264 (2017).

141. Hutchins, P. D., Russell, J. D. & Coon, J. J. LipiDex: An Integrated Software Package for High-Confidence Lipid Identification. *Cell Systems* 6, 621-625 (2018).
142. Fahy, E. et al. LipidFinder on LIPID MAPS: Peak filtering, MS searching and statistical analysis for lipidomics. *Bioinformatics* 35, 685-687 (2019).
143. Alvarez-Jarreta, J. et al. LipidFinder 2.0: Advanced informatics pipeline for lipidomics discovery applications. *Bioinformatics* 37, 1478-1479 (2021).
144. Sarafian, M. H. et al. Objective set of criteria for optimization of sample preparation procedures for ultra-high throughput untargeted blood plasma lipid profiling by ultra performance liquid chromatography-mass spectrometry. *Analytical Chemistry* 86, 5766-5774 (2014).
145. Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R. & Neumann, S. CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical Chemistry* 84, 283-289 (2012).
146. Alcoriza-Balaguer, M. I. et al. LipidMS: An R Package for Lipid Annotation in Untargeted Liquid Chromatography-Data Independent Acquisition-Mass Spectrometry Lipidomics. *Anal Chem* 91, 836-845 (2019).
147. Alcoriza-Balaguer, M. I., García-Cañaveras, J. C., Ripoll-Esteve, F. J., Roca, M. & Lahoz, A. LipidMS 3.0: an R-package and a web-based tool for LC-MS/MS data processing and lipid annotation. *Bioinformatics* 38, 4826-4828 (2022).
148. Li, H., Cai, Y., Guo, Y., Chen, F. & Zhu, Z.-J. MetDIA: Targeted Metabolite Extraction of Multiplexed MS/MS Spectra Generated

- by Data-Independent Acquisition. *Analytical Chemistry* 88, 8757-5764 (2016).
149. Chang, W. et al. Shiny: Web Application Framework (<https://CRAN.R-project.org/package=shiny>). (2021).
150. Fu, X. et al. Measurement of lipogenic flux by deuterium resolved mass spectrometry. *Nature Communications* 12, (2021).
151. Zhang, Z., Chen, L., Liu, L., Su, X. & Rabinowitz, J. D. Chemical Basis for Deuterium Labeling of Fat and NADPH. *Journal of American Chemistry Society* 139, 14368-14371 (2017).
152. Antoniewicz, M. R. A guide to ¹³C metabolic flux analysis for the cancer biologist. *Experimental and Molecular Medicine* vol. 50 (2018).
153. Kelleher, J. K. & Nickol, G. B. Isotopomer Spectral Analysis: Utilizing Nonlinear Models in Isotopic Flux Studies. in *Methods in Enzymology* vol. 561 303-330 (Academic Press Inc., 2015).
154. Kelleher, J. K., Kelleher, W. G. & Masterson, T. M. Model equations for condensation biosynthesis using stable isotopes and radioisotopes. (1992).
155. Tredwell, G. D. & Keun, H. C. ConvISA: A simple, convoluted method for isotopomer spectral analysis of fatty acids and cholesterol. *Metabolic Engineering* 32, 125-132 (2015).
156. Kamphorst, J. J. et al. Hypoxic and Ras-transformed cells support growth by scavenging unsaturated fatty acids from lysophospholipids. *Proceedings of the National Academy of Sciences U S A* 110, 8882-8887 (2013).
157. Hardwicke, M. A. et al. A human fatty acid synthase inhibitor binds β -ketoacyl reductase in the keto-substrate site. *Nature Chemical Biology* 10, 774-779 (2014).

158. Obukowicz, M. G. et al. Identification and Characterization of a Novel 6/5 Fatty Acid Desaturase Inhibitor As a Potential Anti-Inflammatory Agent. *Biochemical pharmacology* vol. 55 (1998).
159. Xin, Z. et al. Discovery of piperidine-aryl urea-based stearyl-CoA desaturase 1 inhibitors. *Bioorganic and Medicinal Chemistry Letters* 18, 4298–4302 (2008).
160. von Roemeling, C. A. et al. Stearyl-CoA desaturase 1 is a novel molecular therapeutic target for clear cell renal cell carcinoma. *Clinical Cancer Research* 19, 2368–2380 (2013).
161. Ghergurovich, J. M. et al. A small molecule G6PD inhibitor reveals immune dependence on pentose phosphate pathway. *Nature Chemical Biology* 16, 731–739 (2020).
162. García-Cañaveras, J. C. et al. SHMT inhibition is effective and synergizes with methotrexate in T-cell acute lymphoblastic leukemia. *Leukemia* 35, 377–388 (2021).
163. García-Cañaveras, J. C. et al. SHMT inhibition is effective and synergizes with methotrexate in T-cell acute lymphoblastic leukemia. (2021).
164. Qiu, J. et al. Acetate Promotes T Cell Effector Function during Glucose Restriction. *Cell Rep* 27, 2063-2074.e5 (2019).
165. Kamphorst, J. J., Chung, M. K., Fan, J. & Rabinowitz, J. D. Quantitative analysis of acetyl-CoA production in hypoxic cancer cells reveals substantial contribution from acetate. *Cancer Metabolism* 2 (2014).
166. Antoniewicz, M. R., Kelleher, J. K. & Stephanopoulos, G. Elementary Metabolite Units (EMU): a novel framework for modeling isotopic distributions. *Metab Eng* (2007) doi:10.1016/j.ymben.2006.09.001.

167. Metallo, C. M. et al. Reductive glutamine metabolism by IDH1 mediates lipogenesis under hypoxia. *Nature* 481, 380–384 (2012).
168. Buescher, J. M. et al. A roadmap for interpreting ¹³C metabolite labeling patterns from cells. *Current Opinion in Biotechnology* vol. 34 189–201 (2015).
169. Chambers, M. C. et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology* vol. 30 (2012).
170. Su, X., Lu, W. & Rabinowitz, J. D. Metabolite Spectral Accuracy on Orbitraps. *Analytical Chemistry* (2017).
171. Elzhov, T., Mullen, K., Spiess, A.-N. & Bolker, B. minpack.lm: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK, Plus Support for Bounds (<https://CRAN.R-project.org/package=minpack.lm>). (2022).
172. Padfield, D., Yvon-Durocher, G., Buckling, A., Jennings, S. & Yvon-Durocher, G. Rapid evolution of metabolic traits explains thermal adaptation in phytoplankton. *Ecology Letters* 19, 133–142 (2016).
173. Consul, P. C. & Ettl, S. P. Some Discrete Multinomial Probability Models with Predetermined Strategy. *Biometrical Journal* (1977).
174. Guillou, H., Zadavec, D., Martin, P. G. P. & Jacobsson, A. The key roles of elongases and desaturases in mammalian fatty acid metabolism: Insights from transgenic mice. *Progress in Lipid Research* vol. 49 186–199 (2010).
175. Purdy, J. G., Shenk, T. & Rabinowitz, J. D. Fatty acid elongase 7 catalyzes lipidome remodeling essential for human cytomegalovirus replication. *Cell Rep* 10, 1375–1385 (2015).
176. Deák, F., Anderson, R. E., Fessler, J. L. & Sherry, D. M. Novel Cellular Functions of Very Long Chain-Fatty Acids: Insight From

ELOVL4 Mutations. *Frontiers in Cellular Neuroscience* vol. 13 (2019).

177. R Core Team. R: A language and environment for statistical computing. (2022).
178. Jang, C. et al. The small intestine shields the liver from fructose-induced steatosis. *Nature Metabolism* 2, 586-593 (2020).
179. Crown, S. B., Marze, N. & Antoniewicz, M. R. Catabolism of branched chain amino acids contributes significantly to synthesis of odd-chain and even-chain fatty acids in 3T3-L1 adipocytes. *PLoS One* 10, (2015).

Appendix 1: List of Publications

All the results included in this thesis have been published in the following articles:

- Roca M, **Alcoriza MI**, Garcia-Cañaveras JC, Lahoz A. Reviewing the metabolome coverage provided by LC-MS: Focus on sample preparation and chromatography-A tutorial. *Anal Chim Acta*. 2021. 1147:38-55. doi: 10.1016/j.aca.2020.12.025.

Introduction and analysis of the human metabolome coverage by untargeted LC-MS based on a survey in metabolic and spectral databases cover part of the introduction of this thesis.

- **Alcoriza-Balaguer MI**, García-Cañaveras JC, López A, Conde I, Juan O, Carretero J, Lahoz A. LipidMS: An R Package for Lipid Annotation in Untargeted Liquid Chromatography-Data Independent Acquisition-Mass Spectrometry Lipidomics. *Analytical Chemistry*. 2019. 91(1):836-845. doi: 10.1021/acs.analchem.8b03409.
- **Alcoriza-Balaguer MI**, García-Cañaveras JC, Ripoll-Esteve FJ, Roca M, Lahoz A. LipidMS 3.0: an R-package and a web-based tool for LC-MS/MS data processing and lipid annotation. *Bioinformatics*. 2022. 38(20):4826-4828. doi: 10.1093/bioinformatics/btac581.

These two articles cover the results presented in the Chapter 1 of this thesis.

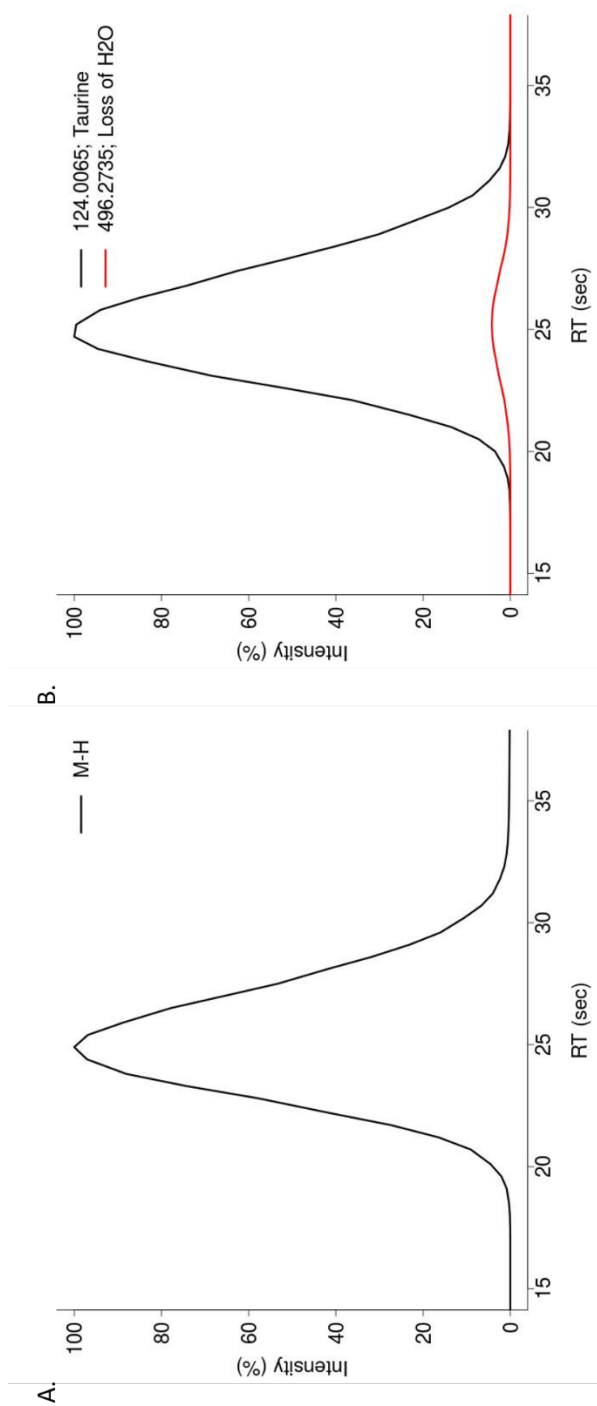
- Alcoriza-Balaguer MI, García-Cañaveras JC, Benet M, Juan-Vidal O, Lahoz A. FAMetA: a mass isotopologue-based tool for the comprehensive analysis of fatty acid metabolism. *Briefings in Bioinformatics*. 2023. 1-14. doi: doi.org/10.1093/bib/bbad064

This article covers the results presented in the Chapter 2 of this thesis.

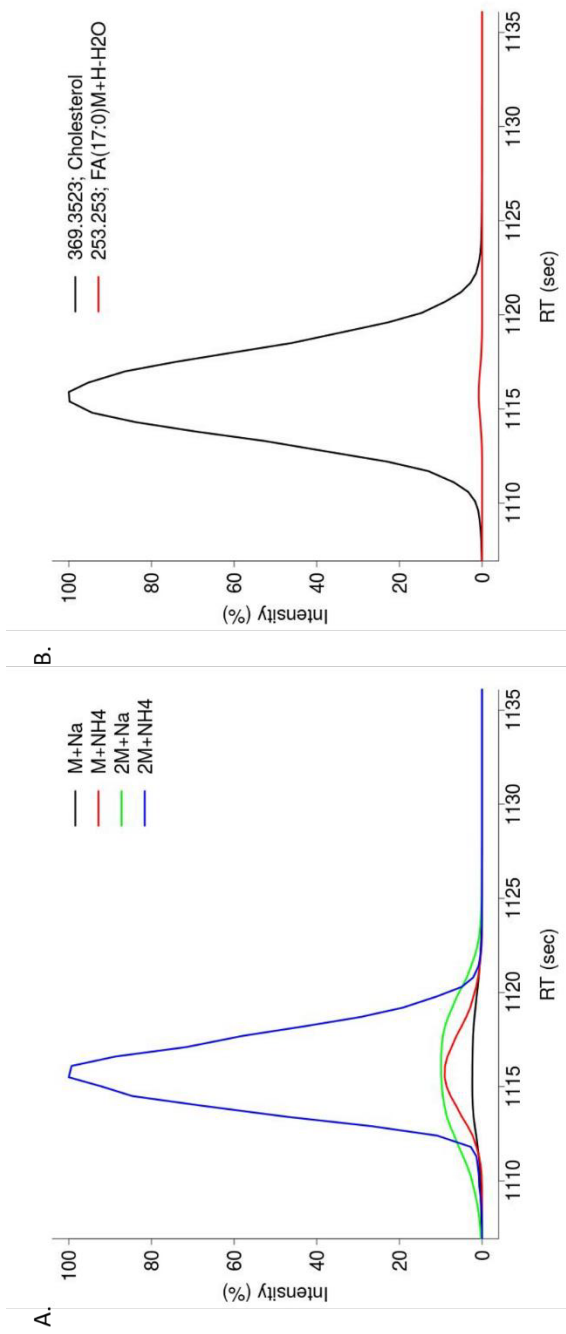
In addition, throughout the development of this thesis, I have also contributed in the following publications:

- Palanca-Ballester, C., Hervas, D., Villalba, M. *et al.* Translation of a tissue epigenetic signature to circulating free DNA suggests *BCAT1* as a potential noninvasive diagnostic biomarker for lung cancer. *Clin Epigenet* 14, 116 (2022). <https://doi.org/10.1186/s13148-022-01334-3>
- Ibáñez-Martínez E, López-Nogueroles M, Alcoriza-Balaguer MI, Pérez I, Roca-Marugán M, Pemán-García J, Lahoz-Rodríguez A, Solé-Jover A. Non-Invasive Infections Diagnosis in Lung Transplant Recipients in Exhaled Breath Condensate. *J. Heart Lung Transplant*. 2021 April;40(4). Doi; 10.1016/j.healun.2021.01.1894.
- Ballester, M., Sentandreu, E., Luongo, G., Santamaria R., Bolonio M., Alcoriza-Balaguer MI., Palomino-Schätzlein M., Pineda-Lucena A., Castell J., Lahoz A., Bort R. Glutamine/glutamate metabolism rewiring in reprogrammed human hepatocyte-like cells. *Sci Rep* 9, 17978 (2019). <https://doi.org/10.1038/s41598-019-54357-x>.

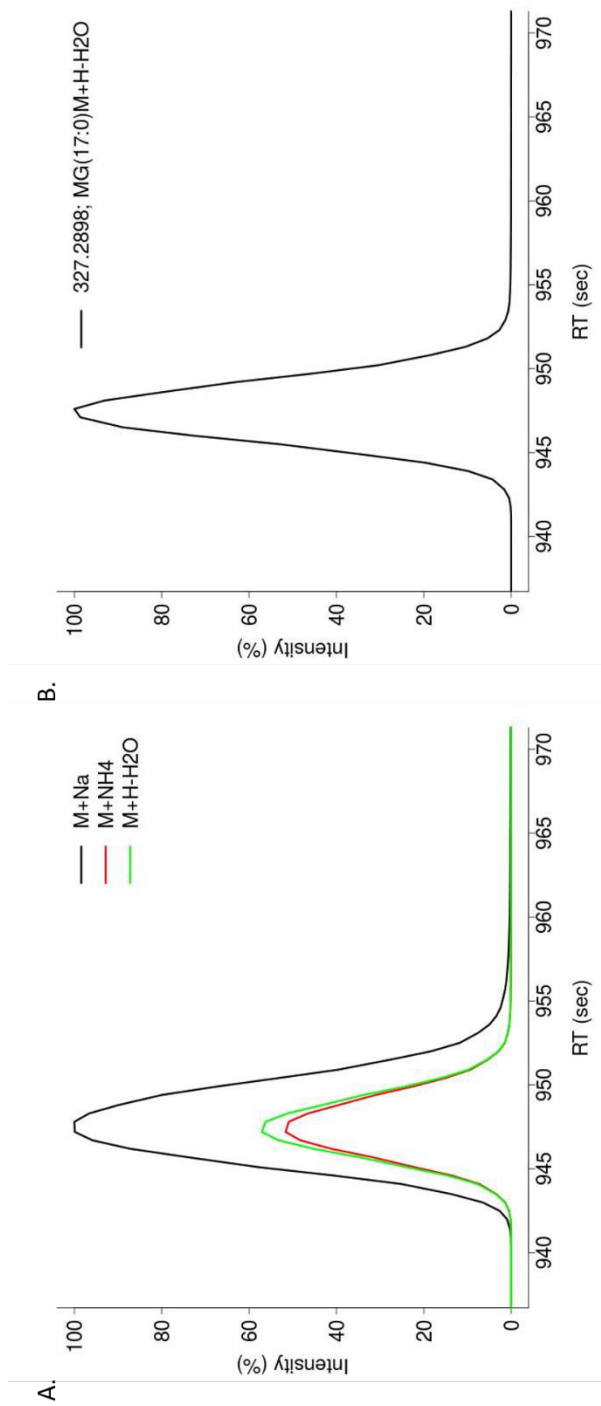
Appendix 2: Additional figures and tables



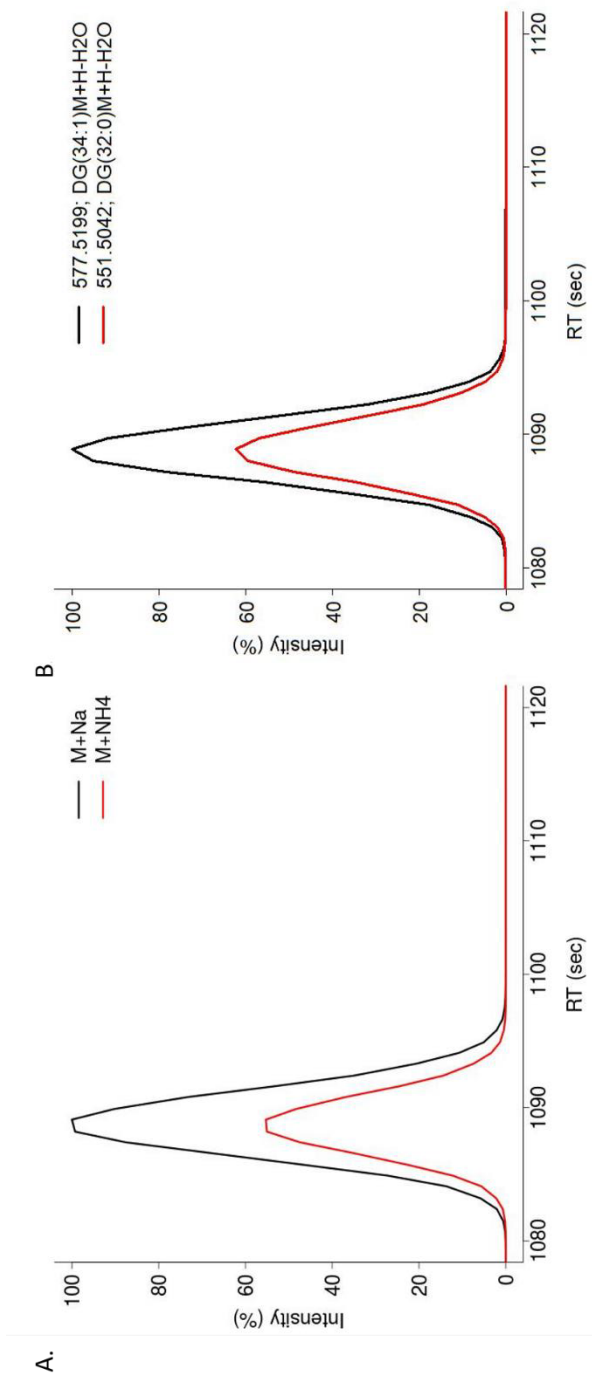
Additional Figure S1. TCA fragmentation pattern in ESI-. A) Chromatographic profiles of the precursor ions in MS¹. B) Chromatographic profiles of informative fragments in MS² using DIA. Coelution between precursor and product fragments can be observed (A-B).



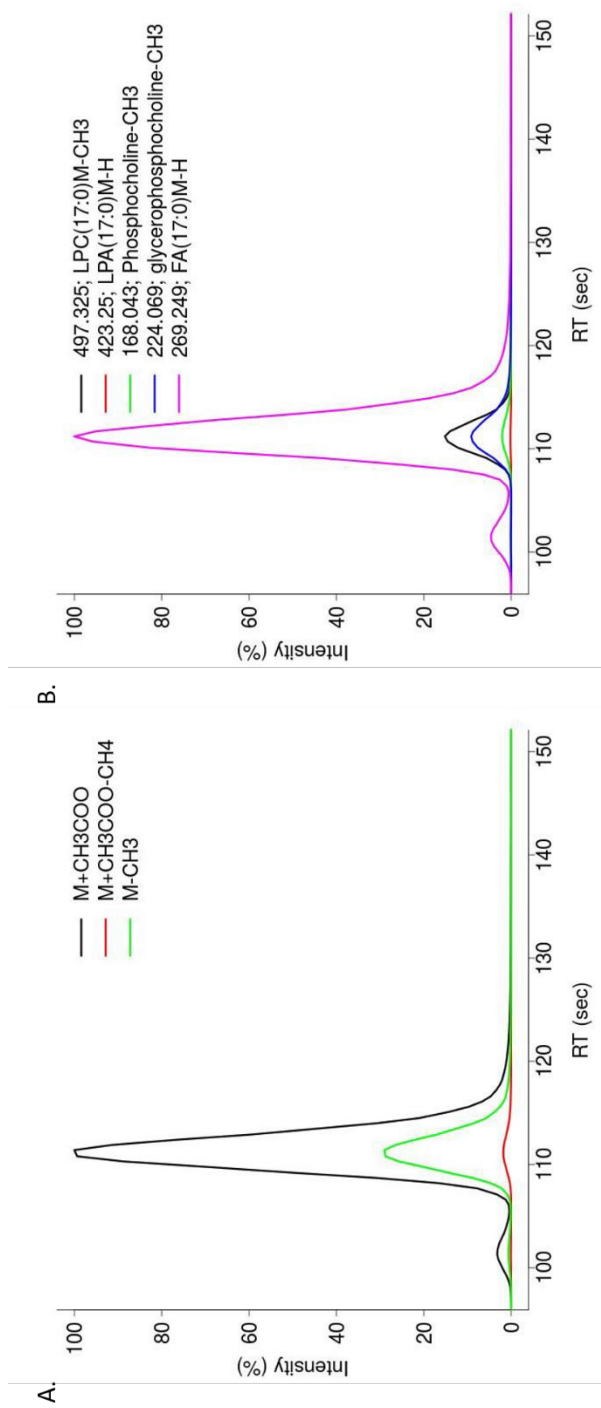
Additional Figure S2. CE(17:0) fragmentation pattern in ESI+. A) Chromatographic profiles of the precursor ions in MS¹. B) Chromatographic profiles of informative fragments in MS² using DIA. Coelution between precursor and product fragments can be observed (A-B).



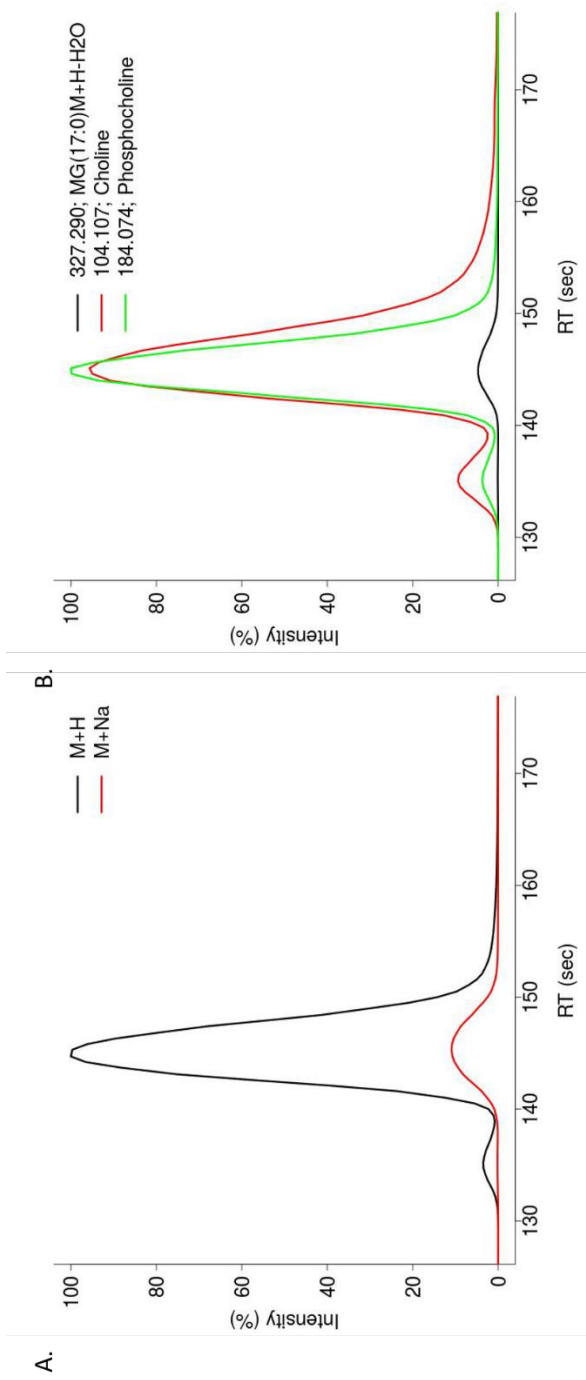
Additional Figure S3. DG(17:0/17:0) fragmentation pattern in ESI+. A) Chromatographic profiles of the precursor ions in MS¹. B) Chromatographic profiles of informative fragments in MS² using DIA. Coelution between precursor and product fragments can be observed (A-B).



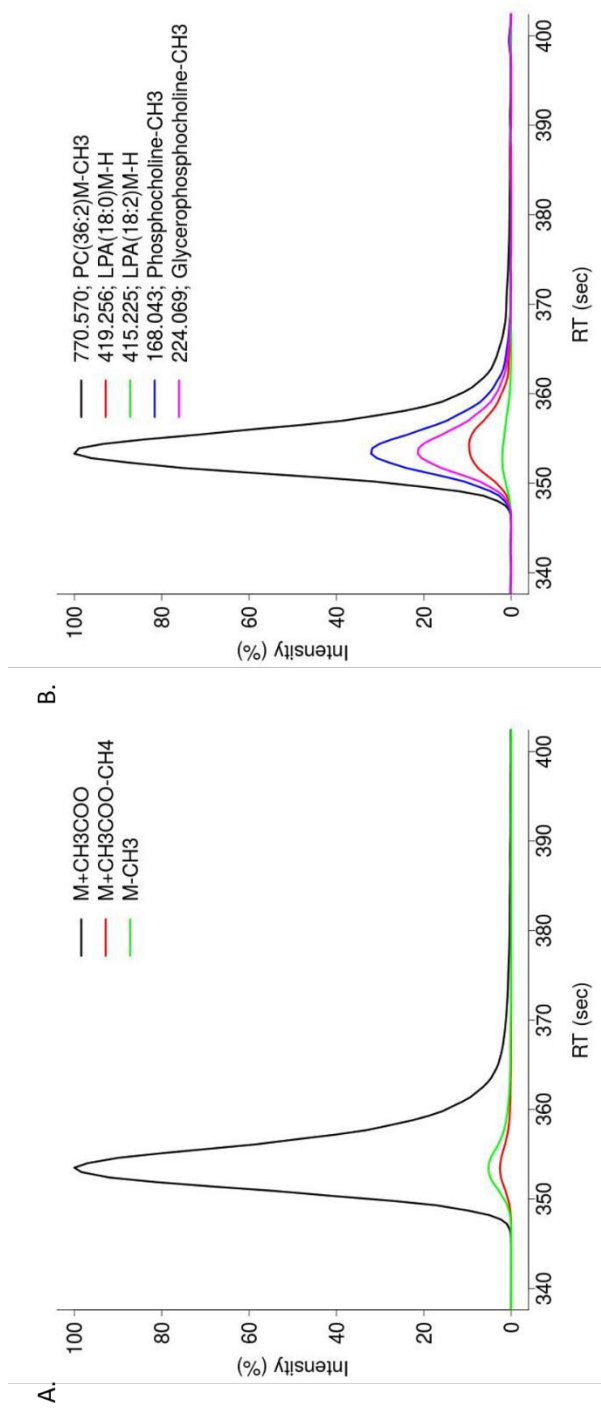
Additional Figure S4. TG(18:0/16:0/18:1) fragmentation pattern in ESI+. A) Chromatographic profiles of the precursor ions in MS¹. B) Chromatographic profiles of informative fragments in MS² using DIA. Coelution between precursor and product fragments can be observed (A-B).



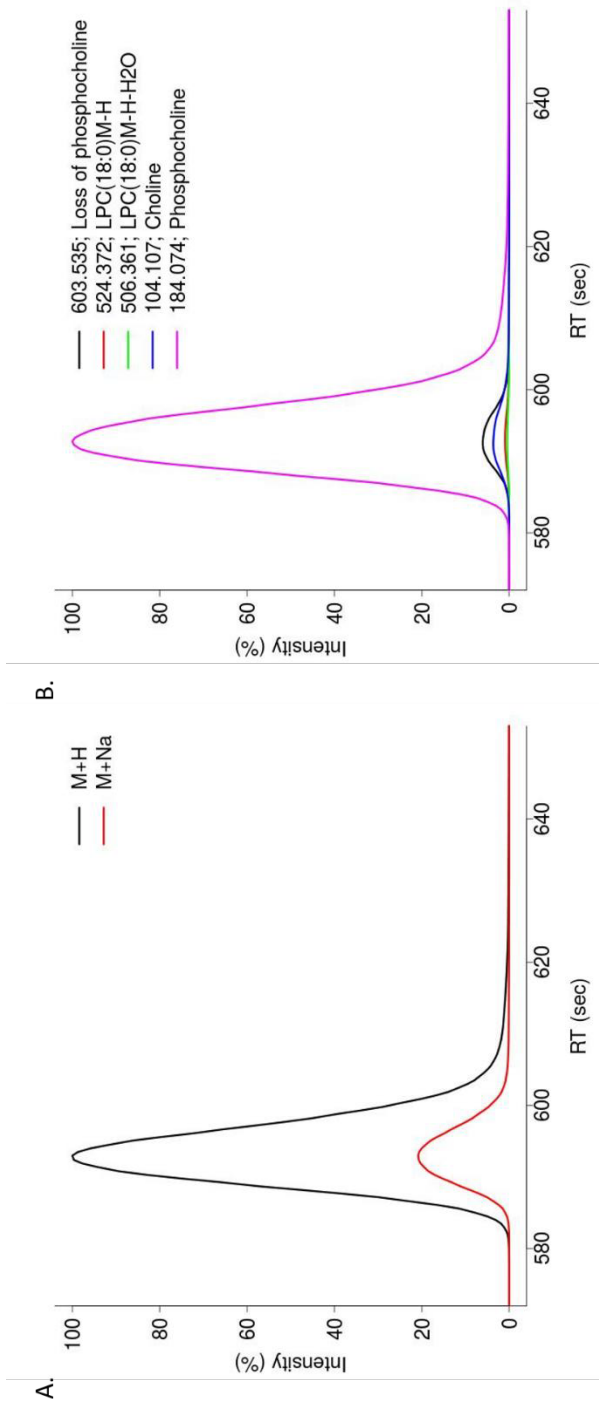
Additional Figure S5. LPC(17:0) fragmentation pattern in ESI-. A) Chromatographic profiles of the precursor ions in MS¹. B) Chromatographic profiles of informative fragments in MS² using DIA. Coelution between precursor and product fragments can be observed (A-B).



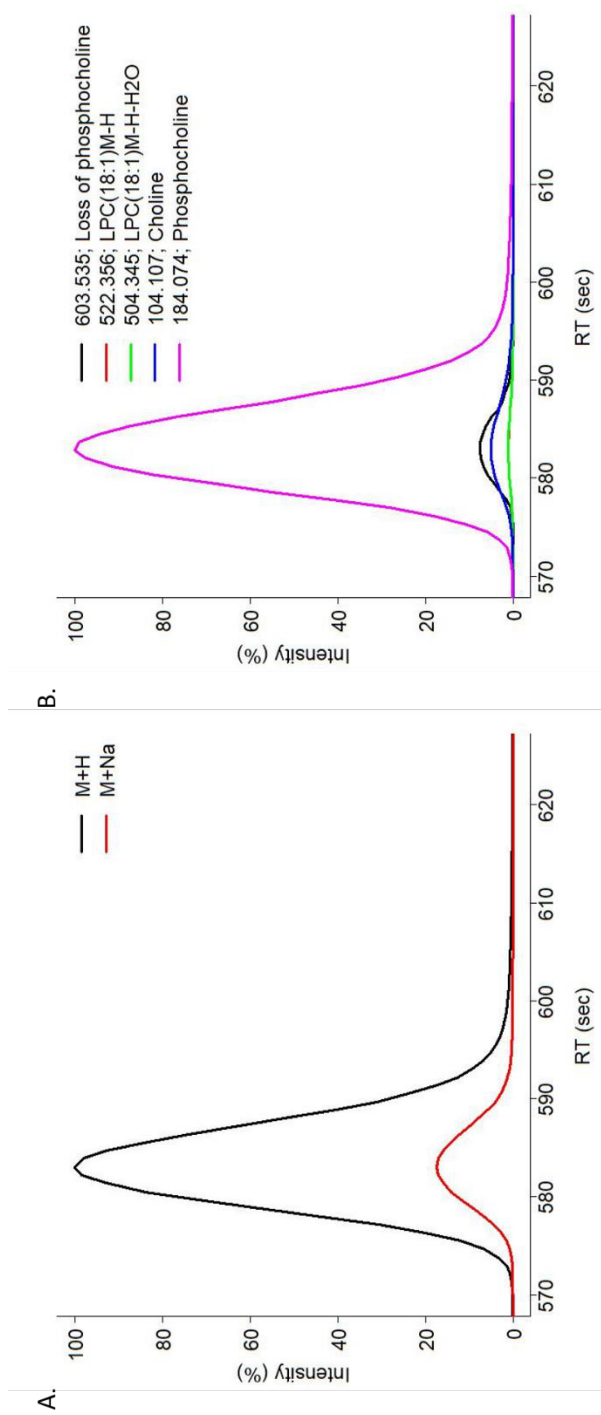
Additional Figure S6. LPC(17:0) fragmentation pattern in ESI+. A) Chromatographic profiles of the precursor ions in MS¹. B) Chromatographic profiles of informative fragments in MS² using DIA. Coelution between precursor and product fragments can be observed (A-B).



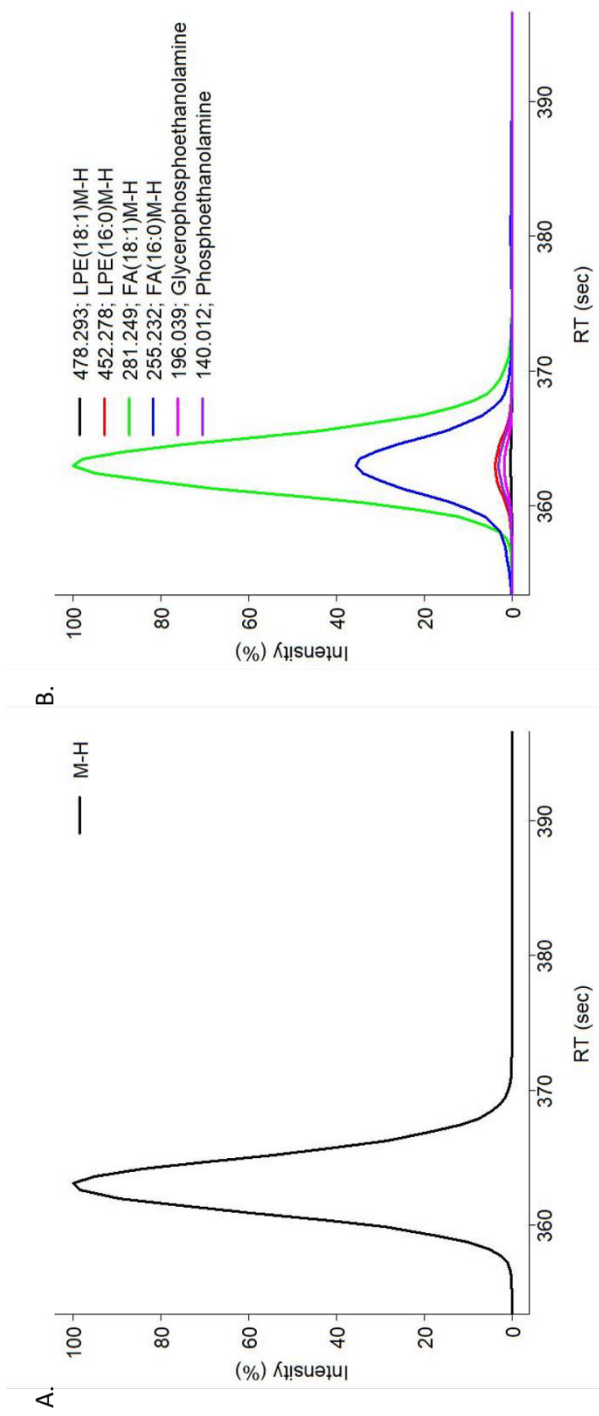
Additional Figure S7. PC(18:0/18:2) fragmentation pattern in ESI- MS². A) Chromatographic profiles of the precursor ions in MS¹. B) Chromatographic profiles of informative fragments in MS² using DIA. Coelution between precursor and product fragments can be observed (A-B).



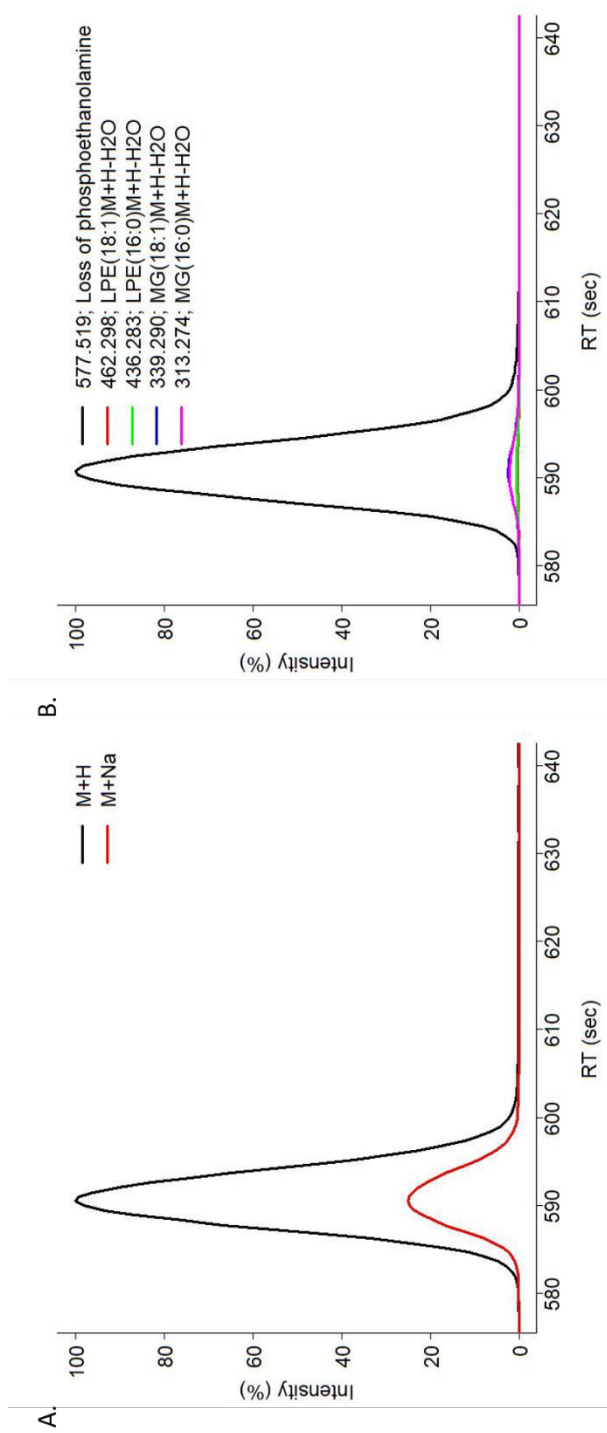
Additional Figure S8. PC(18:0/18:2) fragmentation pattern in ESI+. A) Chromatographic profiles of the precursor ions in MS¹. B) Chromatographic profiles of informative fragments in MS² using DIA. Coelution between precursor and product fragments can be observed (A-B).



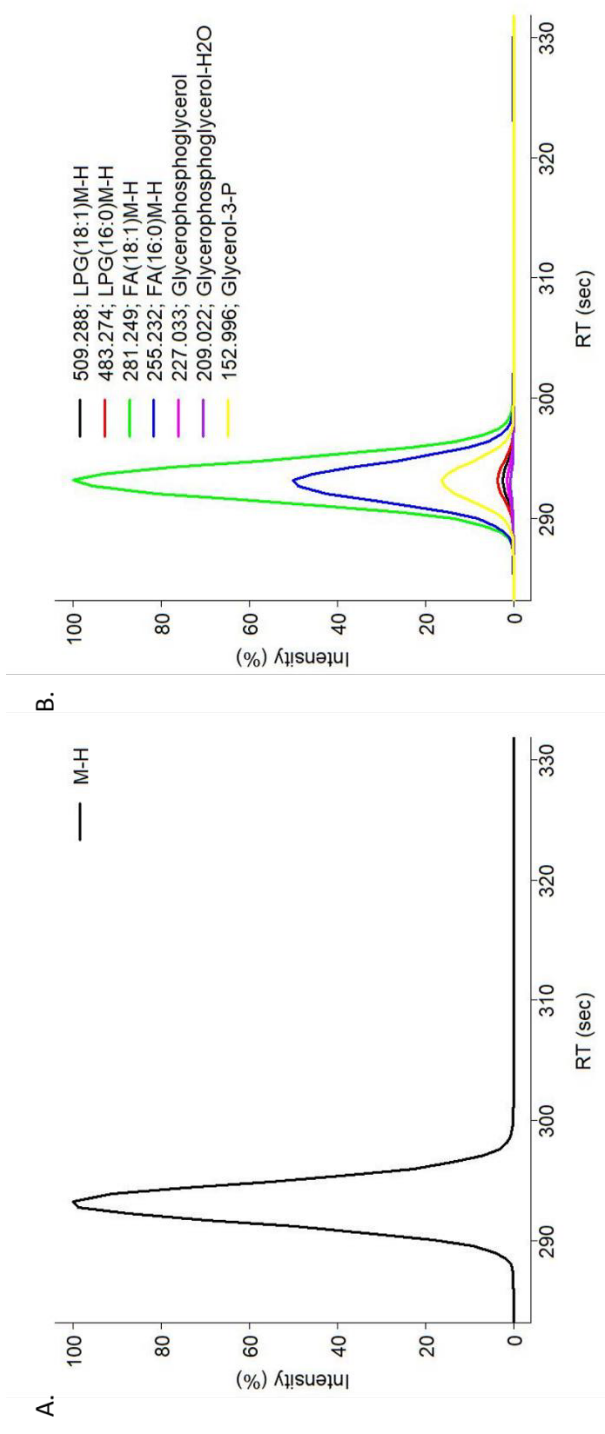
Additional Figure S9. PC(18:1/18:1) fragmentation pattern in ESI+. A) Chromatographic profiles of the precursor ions in MS¹. B) Chromatographic profiles of informative fragments in MS² using DIA. Coelution between precursor and product fragments can be observed (A-B).



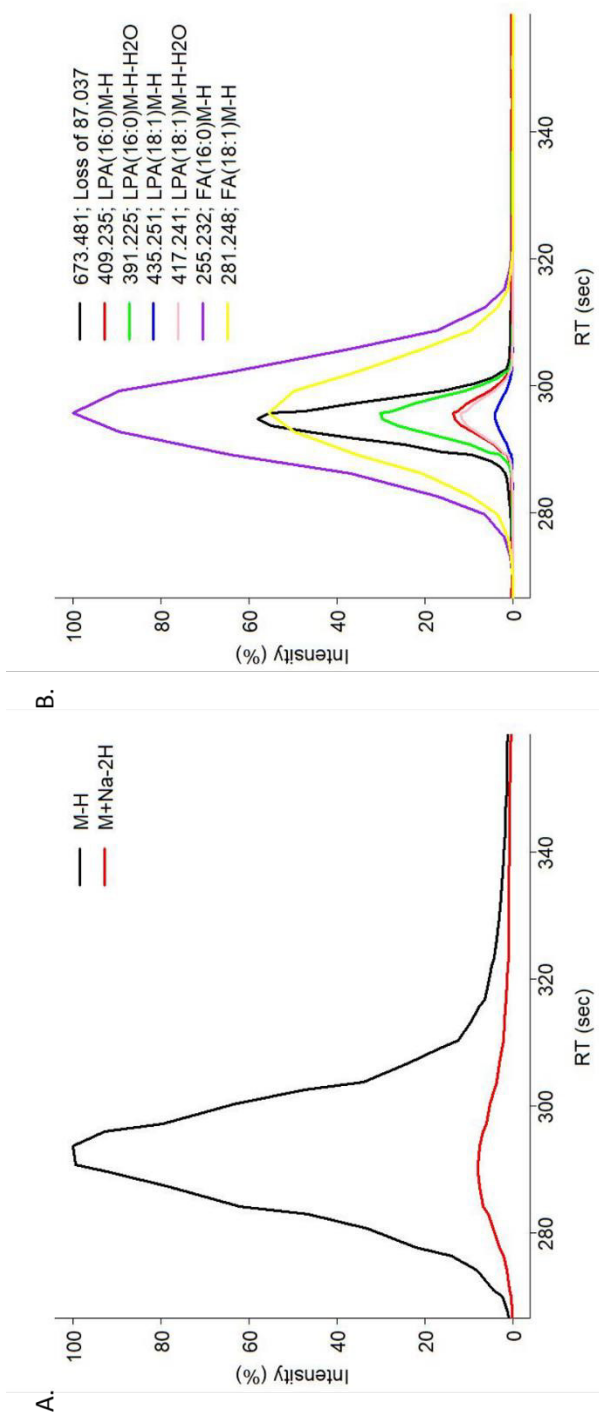
Additional Figure S10. PE(16:0/18:1) fragmentation pattern in ESI-. A) Chromatographic profiles of the precursor ions in MS¹. **B)** Chromatographic profiles of informative fragments in MS² using DIA. Coelution between precursor and product fragments can be observed (A-B).



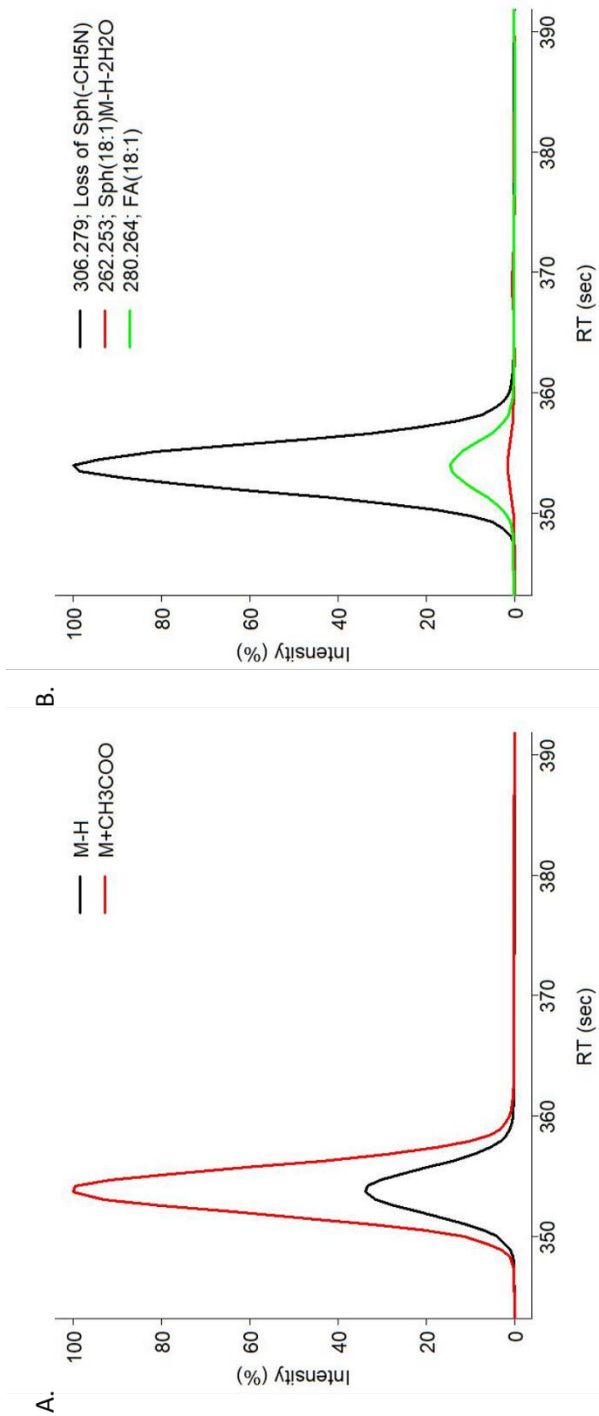
Additional Figure S11. PE(16:0/18:1) fragmentation pattern in ESI+. A) Chromatographic profiles of the precursor ions in MS¹. B) Chromatographic profiles of informative fragments in MS² using DIA. Coelution between precursor and product fragments can be observed (A-B).



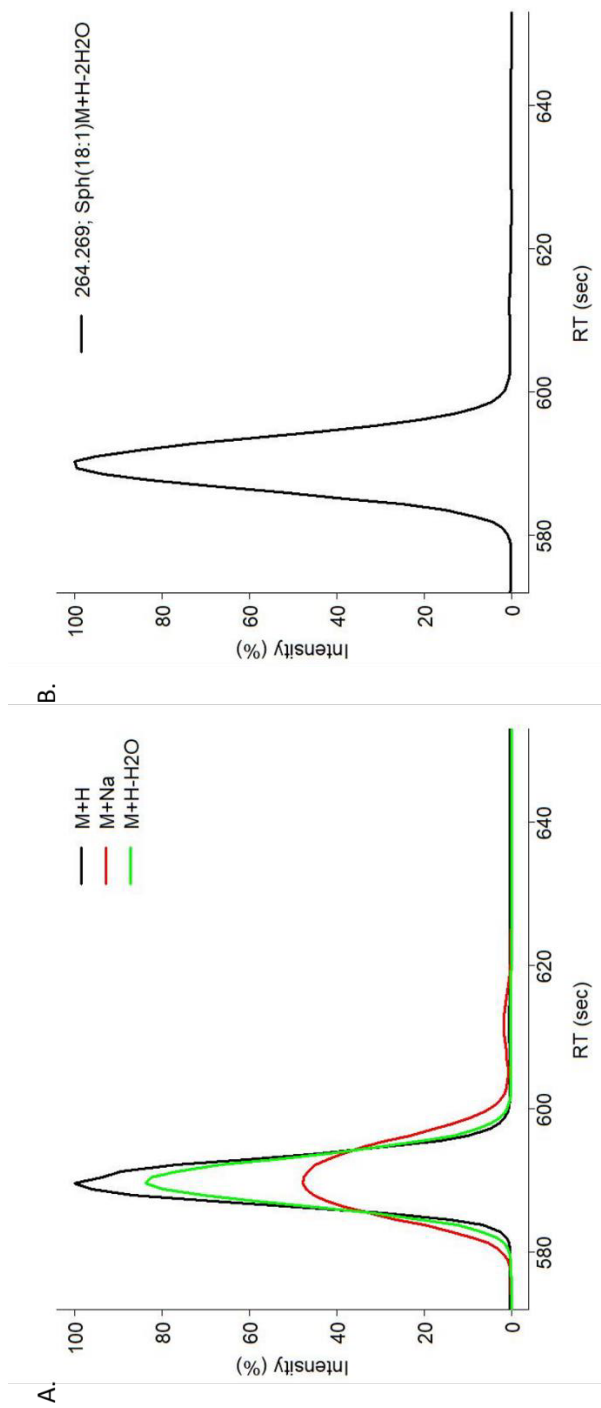
Additional Figure S12. PG(16:0/18:1) fragmentation pattern in ESI. A) Chromatographic profiles of the precursor ions in MS¹. B) Chromatographic profiles of informative fragments in MS² using DIA. Coelution between precursor and product fragments can be observed (A-B).



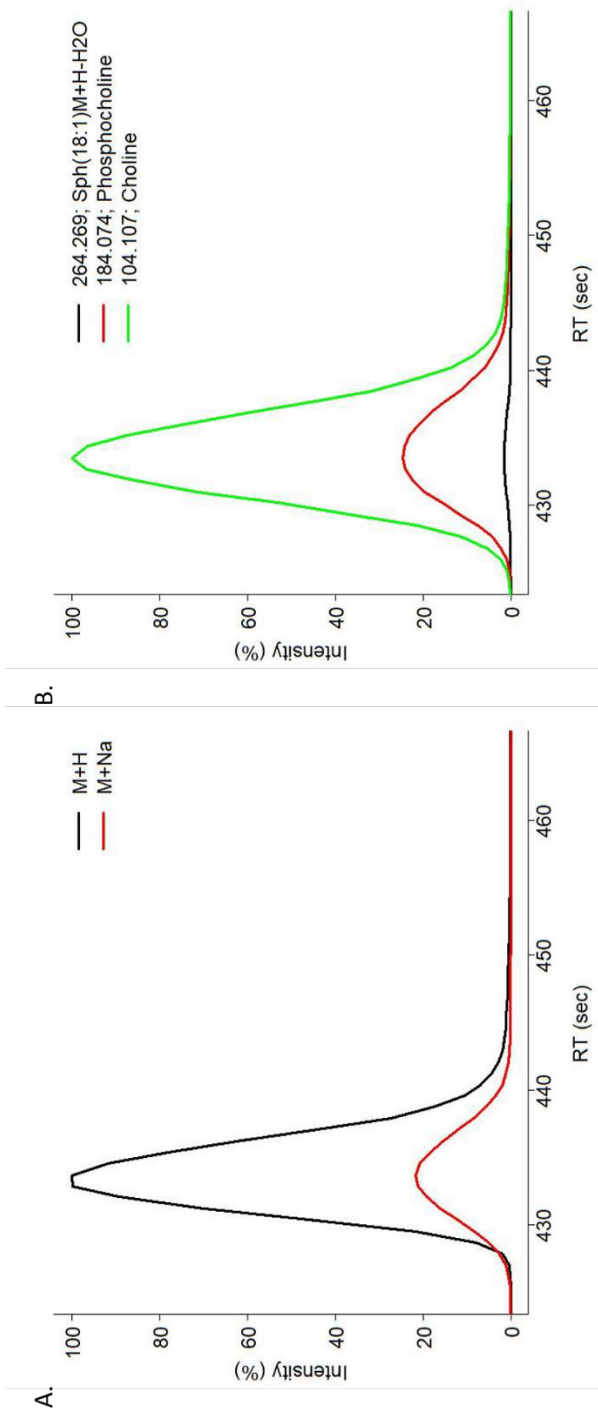
Additional Figure S13. PS(16:0/18:1) fragmentation pattern in ESI-. A) Chromatographic profiles of the precursor ions in MS¹. **B)** Chromatographic profiles of informative fragments in MS² using DIA. Coelution between precursor and product fragments can be observed (A-B).



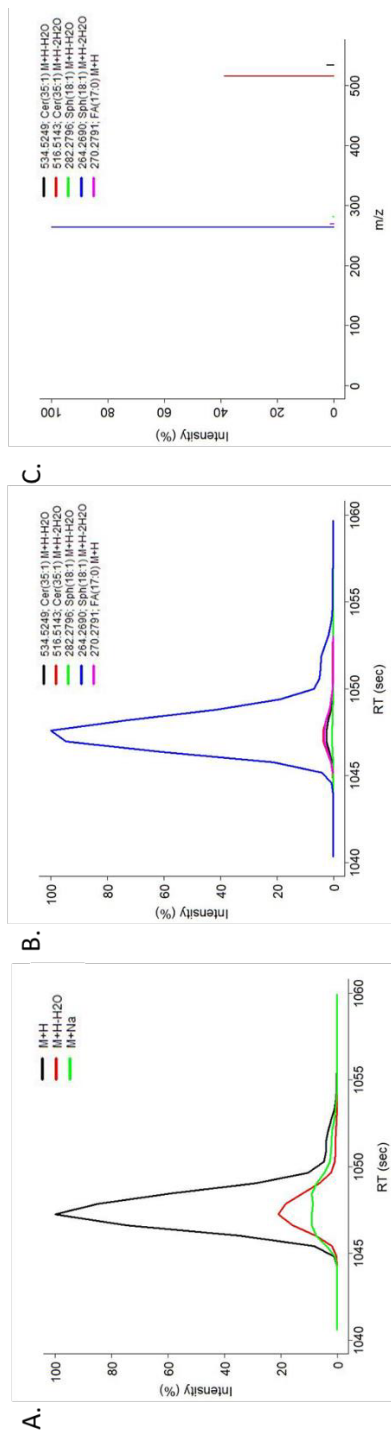
Additional Figure S14. Cer(d18:1/18:1) fragmentation pattern in ESI. A) Chromatographic profiles of the precursor ions in MS¹. B) Chromatographic profiles of informative fragments in MS² using DIA. Coelution between precursor and product fragments can be observed (A-B).



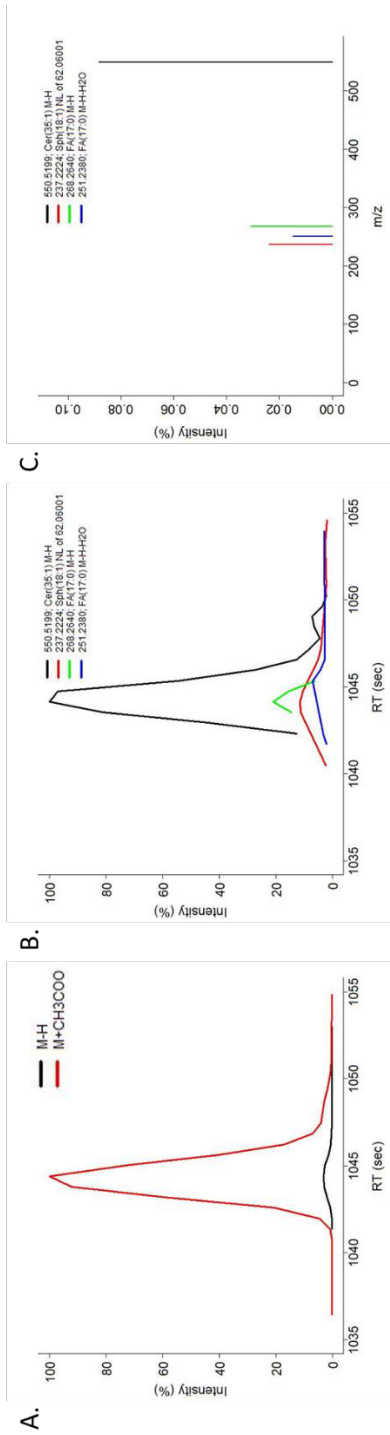
Additional Figure S15. Cer(d18:1/18:1) fragmentation pattern in ESI+. A) Chromatographic profiles of the precursor ions in MS¹. B) Chromatographic profiles of informative fragments in MS² using DIA. Coelution between precursor and product fragments can be observed (A-B).



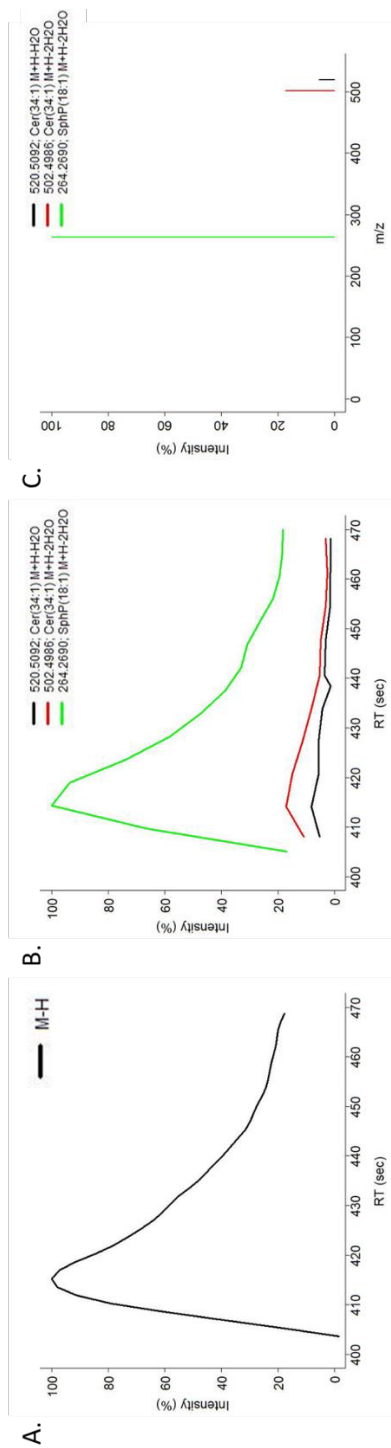
Additional Figure S16. SM(d18:1/16:0) fragmentation pattern in ESI+. A) Chromatographic profiles of the precursor ions in MS¹. B) Chromatographic profiles of informative fragments in MS² using DIA. Coelution between precursor and product fragments can be observed (A-B).



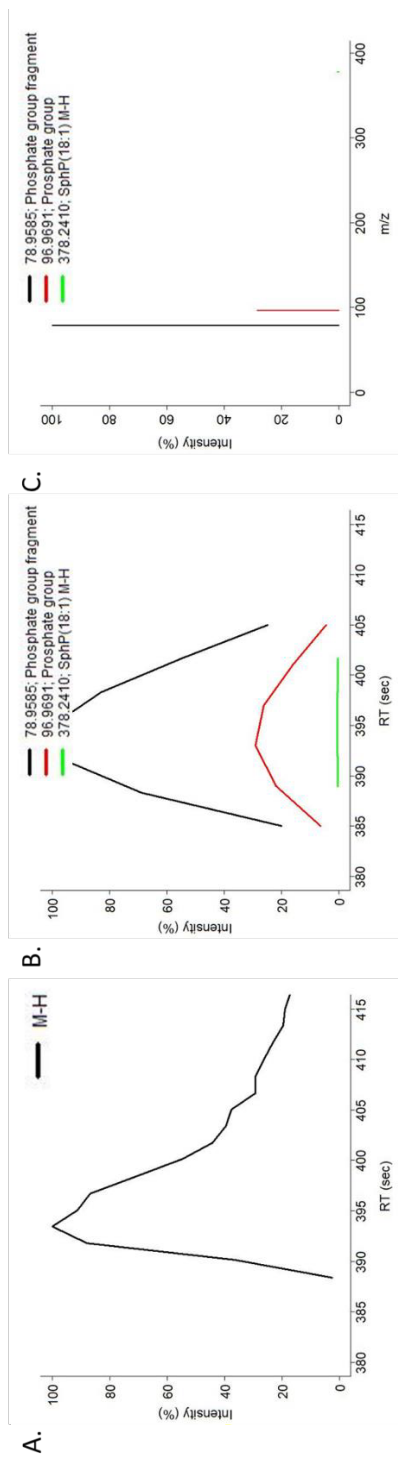
Additional Figure S17. AcylCer(18:1;d18:1/17:0) fragmentation profile for ESI+. A) Chromatographic profiles of the precursor ions in MS¹. B) Chromatographic profiles of informative fragments in MS² using DIA. The coelution between the precursor and product fragments is observed (A-B). C) Fragmentation spectra for MS² using DDA for precursor ion 816.7808 (M+H)⁺. Only the specific fragments used for annotation are shown.



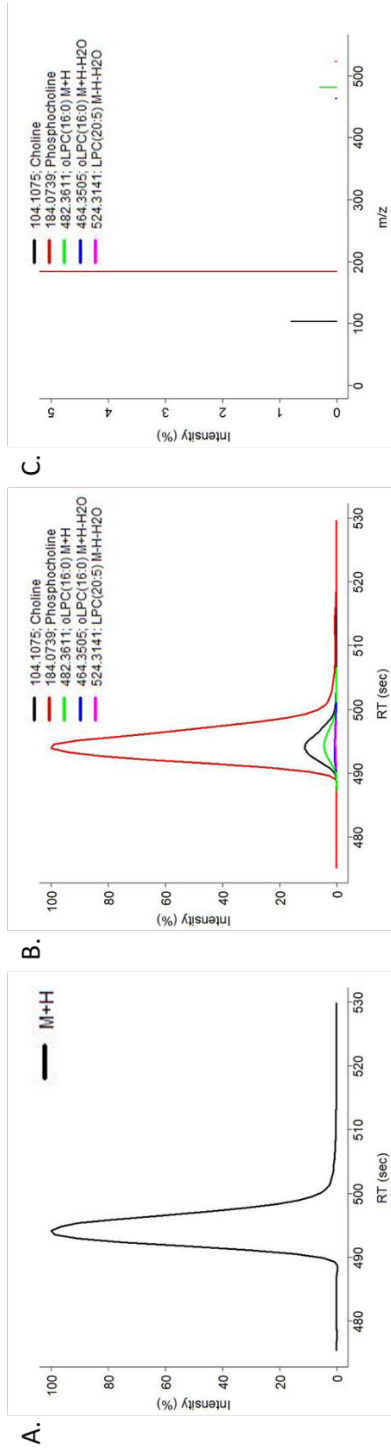
Additional Figure S18. The AcyCer(18:1;dl18:1/17:0) fragmentation profile for ESI- A) Chromatographic profiles of the precursor ions in MS². B) Chromatographic profiles of the informative fragments in MS² using DIA. Coelution between the precursor and product fragments is observed (A-B). C) Fragmentation spectra for MS² using DDA for precursor ion 814.7652 ([M-H]). Only the specific fragments used for annotation are shown.



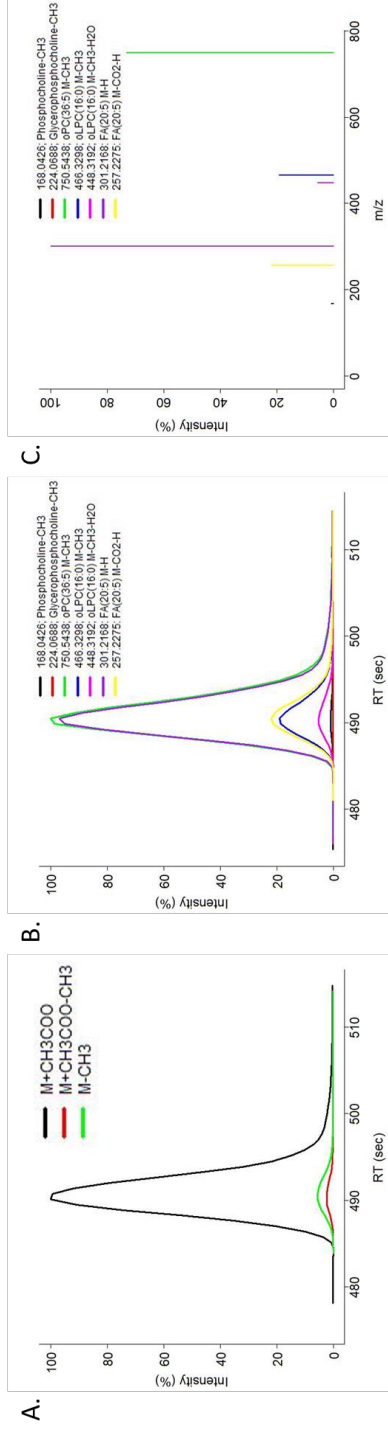
Additional Figure S19. The CerP(d18:1/16:0) fragmentation profile for ESI+. A) Chromatographic profiles of the precursor ions in MS¹. B) Chromatographic profiles of the informative fragments in MS² using DIA. Coelution between the precursor and product fragments is observed (A-B). C) Fragmentation spectra for MS² using DDA for precursor ion 618.4863 ([M+H]⁺). Only the specific fragments used for annotation are shown.



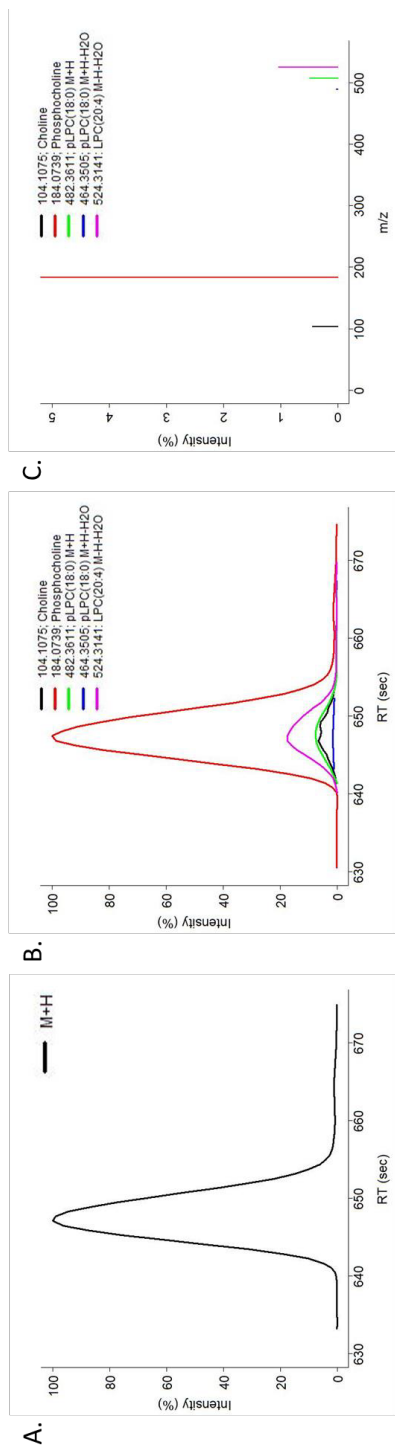
Additional Figure S20. The CerP(d18:1/16:0) fragmentation profile for ESI-. A) Chromatographic profiles of the precursor ions in MS². **B)** Chromatographic profiles of the informative fragments in MS² using DIA. **Coelution between the precursor and product fragments is observed (A-B).** **C)** Fragmentation spectra for MS² using DDA for precursor ion 616.4717 ([M-H]). Only the specific fragments used for annotation are shown.



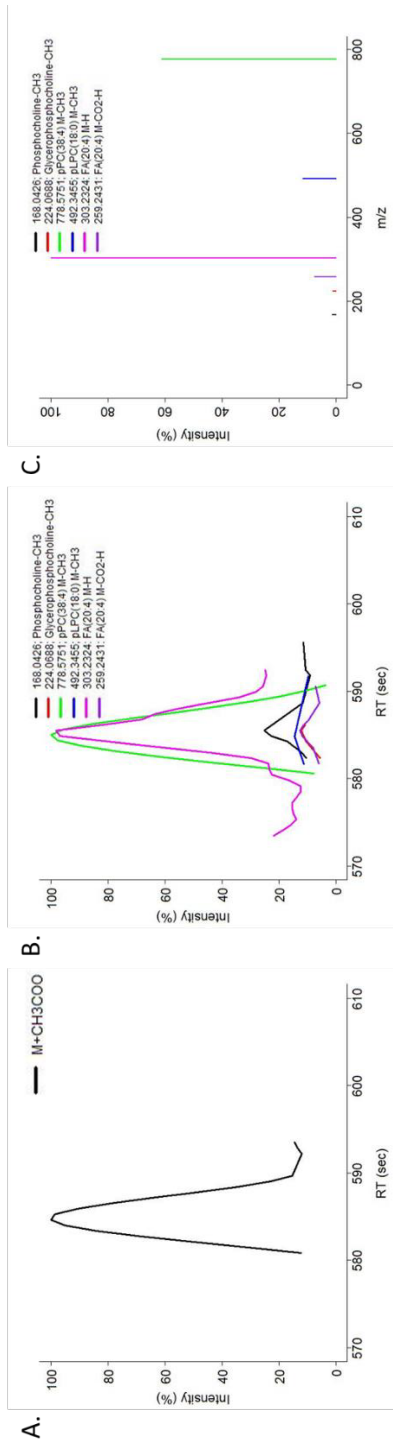
Additional Figure S21. The PC(O-16:0/20:5) fragmentation profile for ESI+. A) Chromatographic profiles of the precursor ions in MS⁺. B) Chromatographic profiles of the informative fragments in MS² using DIA. Coelution between the precursor and product fragments is observed (A-B). C) Fragmentation spectra for MS² using DDA for precursor ion 766.5751 ([M+H]⁺). Only the specific fragments used for annotation are shown. Fragment 184 represents 100% relative intensity.



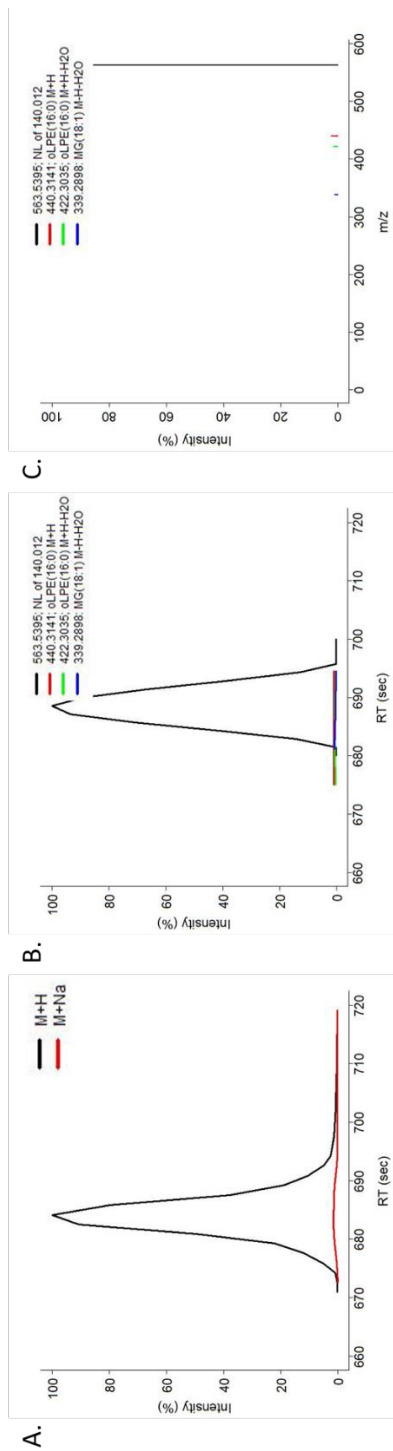
Additional Figure S22. The PC(O-16:0/20:5) fragmentation profile for ESI-A) Chromatographic profiles of the precursor ions in MS⁺. B) Chromatographic profiles of the informative fragments in MS⁻ using DIA. Coelution between the precursor and product fragments is observed (A-B). C) Fragmentation spectra for MS⁻ using DDA for precursor ion 824.5812 ([M+CH₃COO]⁻). Only the specific fragments used for annotation are shown.



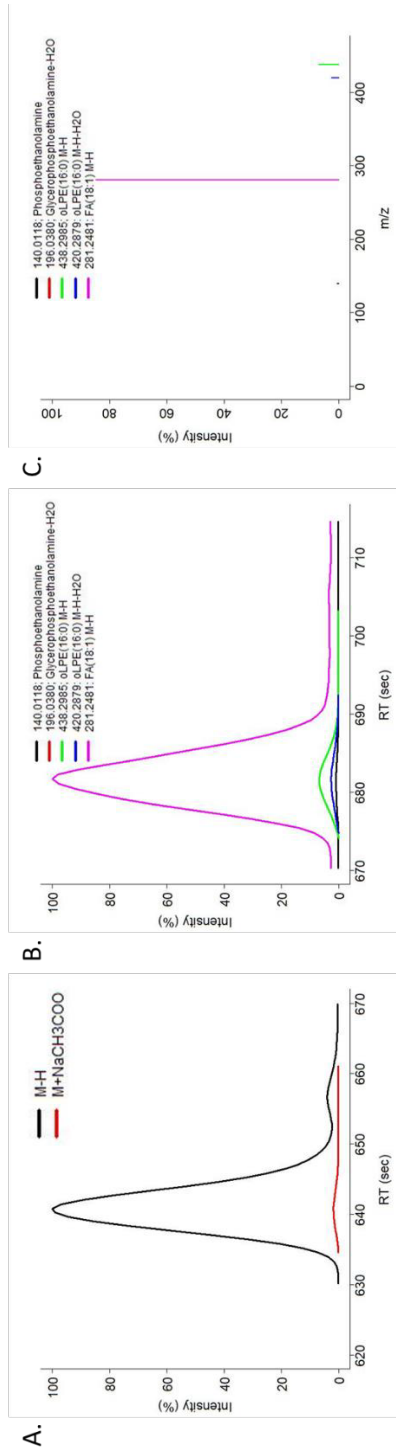
Additional Figure S23. The PC(P-18:0/20:4) fragmentation profile for ESI^+ . A) Chromatographic profiles of the precursor ions in MS^+ . B) Chromatographic profiles of the informative fragments in MS^+ using DIA. Coelution between the precursor and product fragments is observed (A-B). C) Fragmentation spectra for MS^- using DDA for precursor ion 794.6064 (M+H). Only the specific fragments used for annotation are shown. Fragment 184 represents 100% relative intensity.



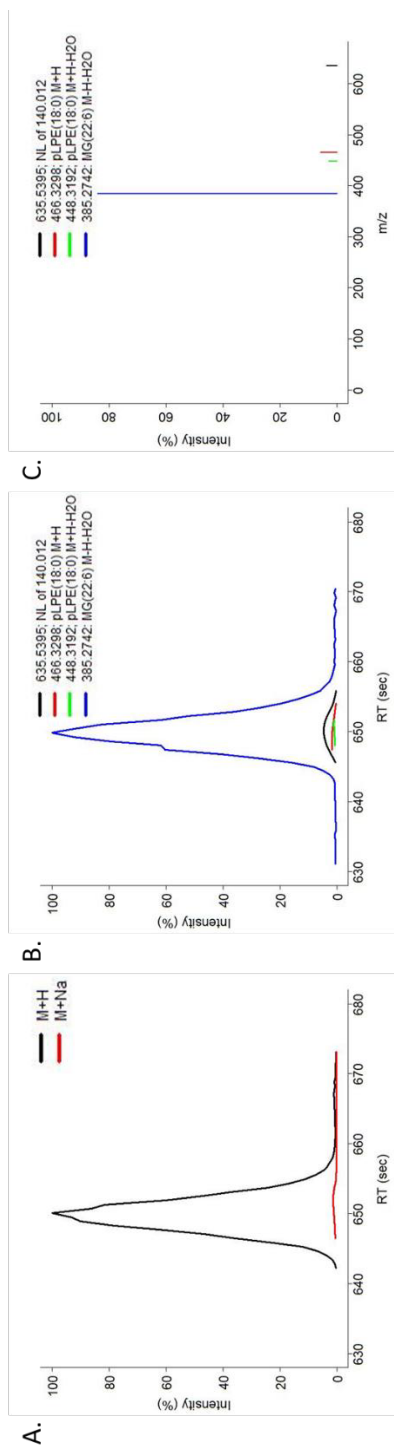
Additional Figure S24. The PCP-18:0/20:4 fragmentation profile for ESI. A) Chromatographic profiles of the precursor ions in MS¹. B) Chromatographic profiles of the informative fragments in MS² using DIA. Coelution between the precursor and product fragments is observed (A-B). C) Fragmentation spectra for MS² using DDA for precursor ion 852.6125 ([M+CH₃COO]⁺). Only the specific fragments used for annotation are shown.



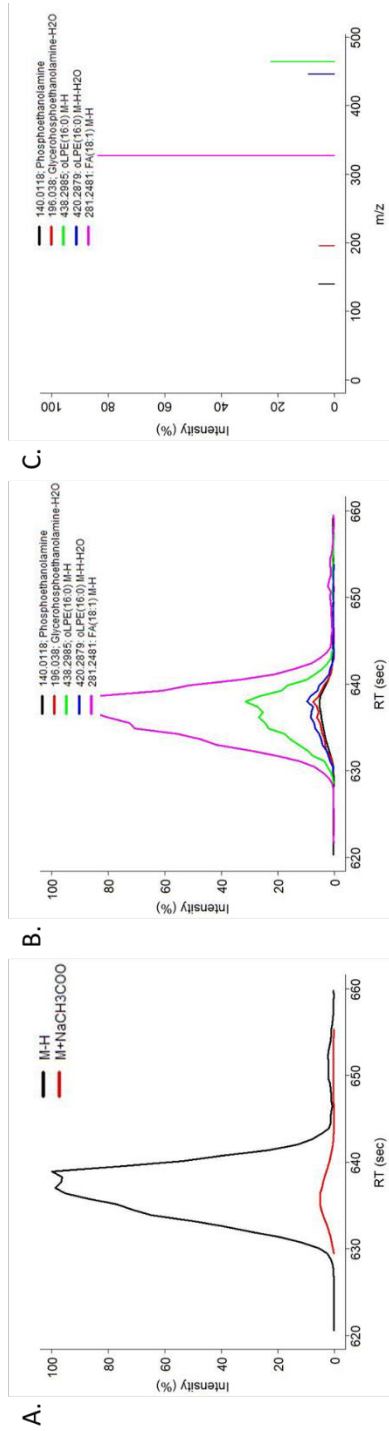
Additional Figure S25. The PE(O-16:0/18:1) fragmentation profile for ESI⁺. A) Chromatographic profiles of the precursor ions in MS¹. B) Chromatographic profiles of the informative fragments in MS² using DIA. Coelution between the precursor and product fragments is observed (A-B). C) Fragmentation spectra for MS² using DDA for precursor ion 704.5594 ([M+H]⁺). Only the specific fragments used for annotation are shown.



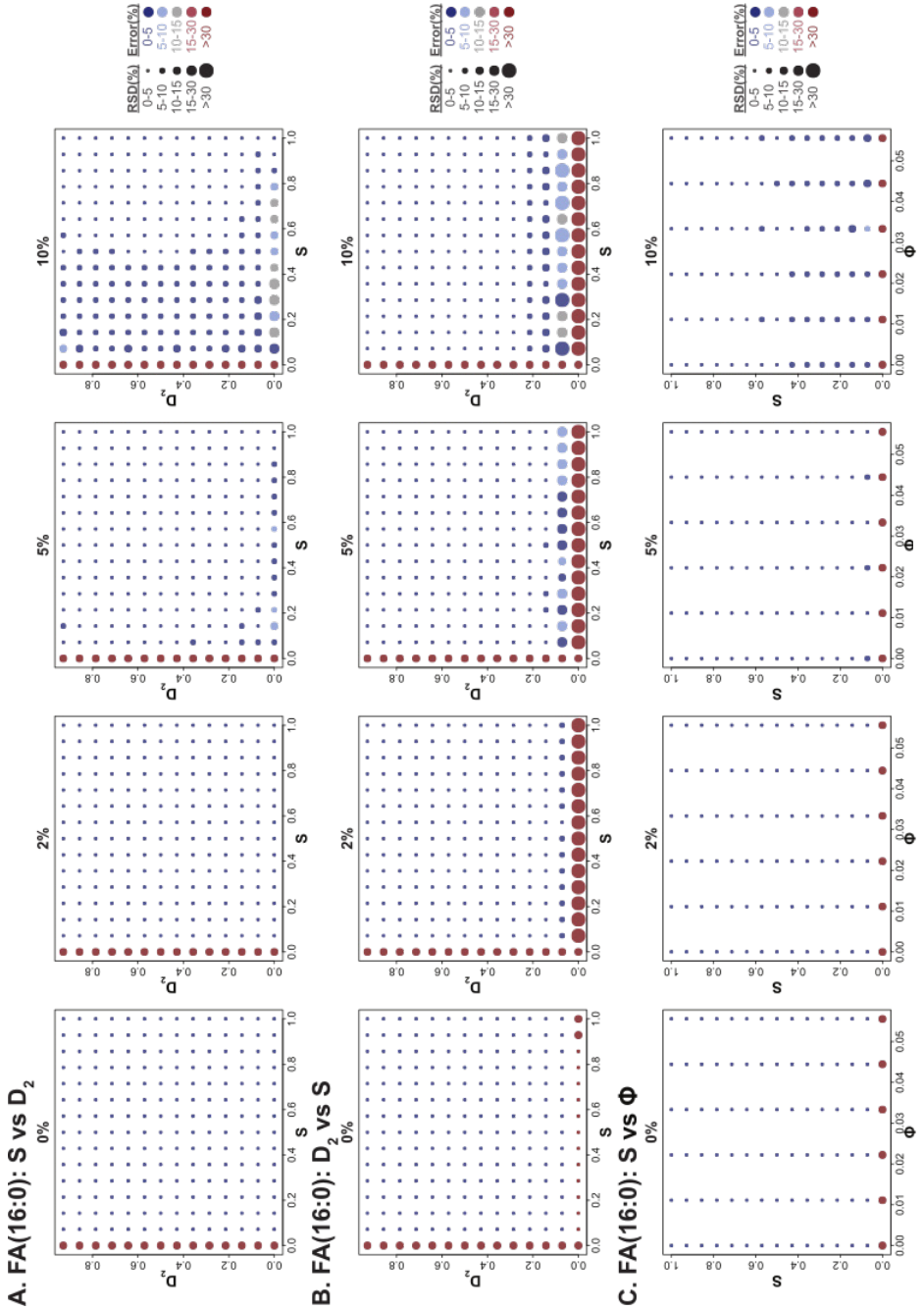
Additional Figure S26. The PE(O-16:0/18:1) fragmentation profile for ESI. A) Chromatographic profiles of the precursor ions in MS¹. B) Chromatographic profiles of the informative fragments in MS² using DIA. Coelution between the precursor and product fragments is observed (A-B). C) Fragmentation spectra for MS² using DDA for precursor ion 784.5469 (M+NaCH₃COO⁺). Only the specific fragments used for annotation are shown.



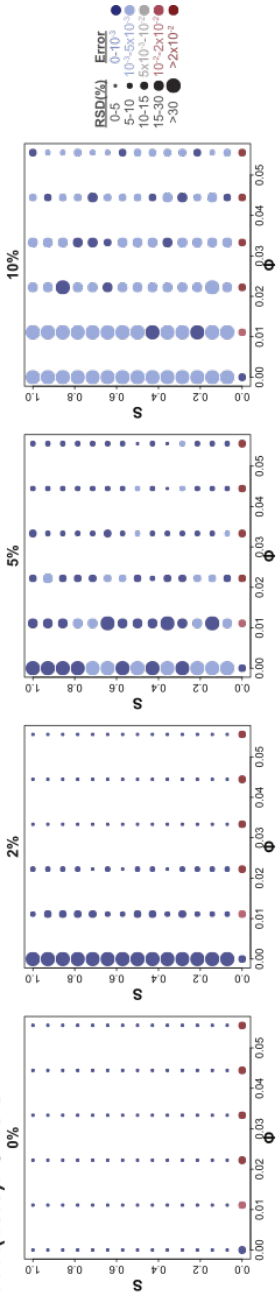
Additional Figure S27. The PE(P-18:0/22:6) fragmentation profile for ESI+. A) Chromatographic profiles of the precursor ions in MS¹. B) Chromatographic profiles of the informative fragments in MS² using DIA. Coelution between the precursor and product fragments is observed (A-B). C) Fragmentation spectra for MS² using DDA for precursor ion 776.5594 ([M+H]⁺). Only the specific fragments used for annotation are shown.



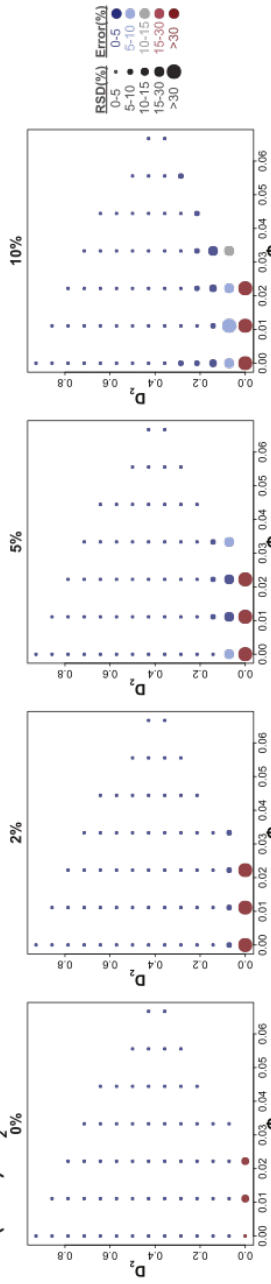
Additional Figure S28. The PEP-18:0/22:6) fragmentation profile for ESI-. A) Chromatographic profiles of the precursor ions in MS¹. B) Chromatographic profiles of the informative fragments in MS² using DIA. Co-elution between the precursor and product fragments is observed (A-B). C) Fragmentation spectra for MS² using DDA for precursor ion 774.5438 (M-H). Only the specific fragments used for annotation are shown.



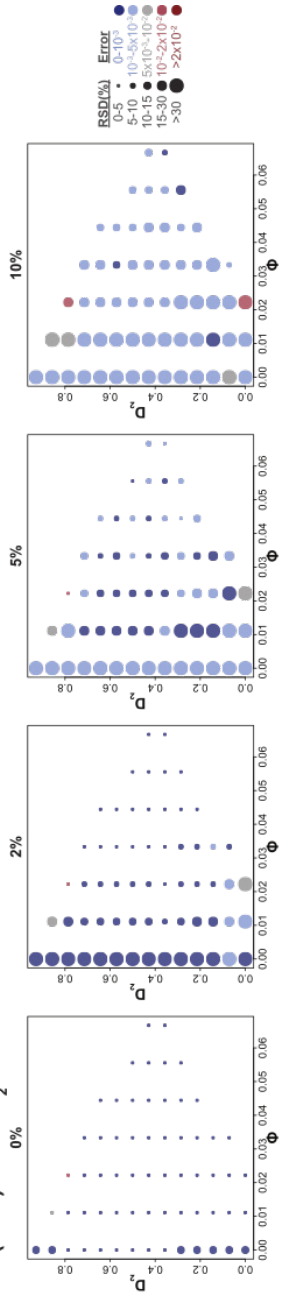
D. FA(16:0): Φ vs S



E. FA(16:0): D_2 vs Φ

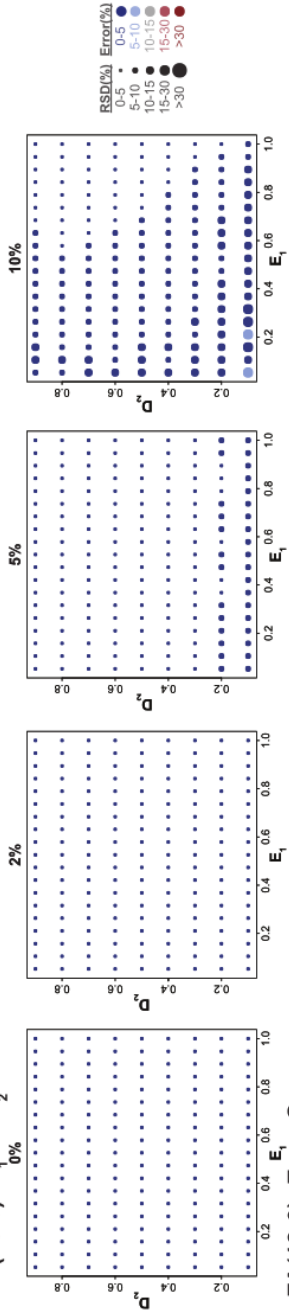


F. FA(16:0): Φ vs D_2

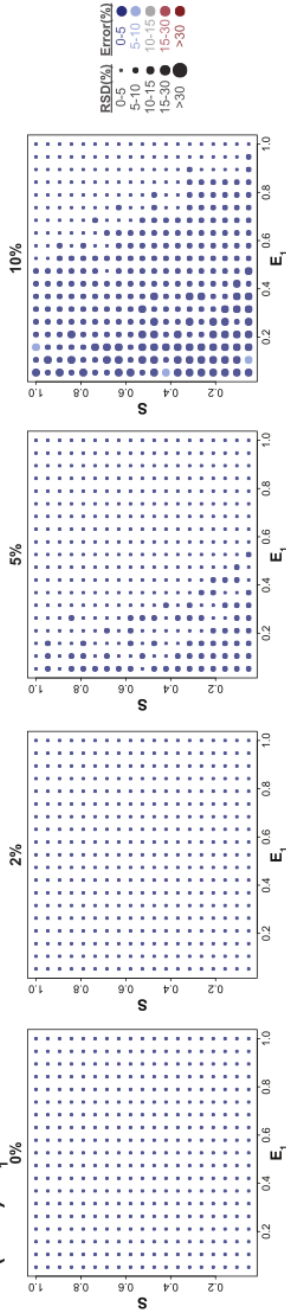


Additional Figure S29. *In silico* validation of the estimation of the *de novo* synthesis of FA(16:0). To evaluate FAMetA's ability to estimate the DNL analysis parameters, realistic values for D_1 (5 values from 0 to 0.2), D_2 (15 values from 0 to 1), ϕ (10 points from 0 to 0.1) and S (15 values from 0 to 1) were combined to simulate 3,945 theoretical FA(16:0) distributions, to which the 0%, 2%, 5% and 10% noise levels were added to obtain 10 different noised distributions for each set of parameters. A, Evaluation of S as a function of S and D_2 , B, Evaluation of D_2 as a function of S and D_2 , C, Evaluation of S as a function of S and ϕ ($D_2=0.5$). D, Evaluation of ϕ as a function of S and ϕ ($D_2=0.5$). E, Evaluation of D_2 as a function of D_2 and ϕ ($S=0.5$). F, Evaluation of ϕ as a function of D_2 and ϕ ($S=0.5$). G, Evaluation of D_1 as a function of D_1 and S , H, Evaluation of D_1 as a function of D_1 and D_2 , I, Evaluation of D_1 as a function of D_1 and ϕ .

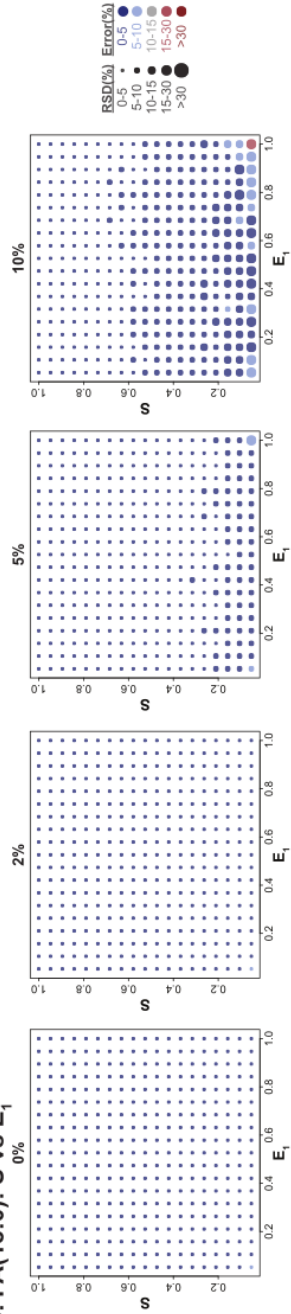
A. FA(18:0): E_1 vs D_2



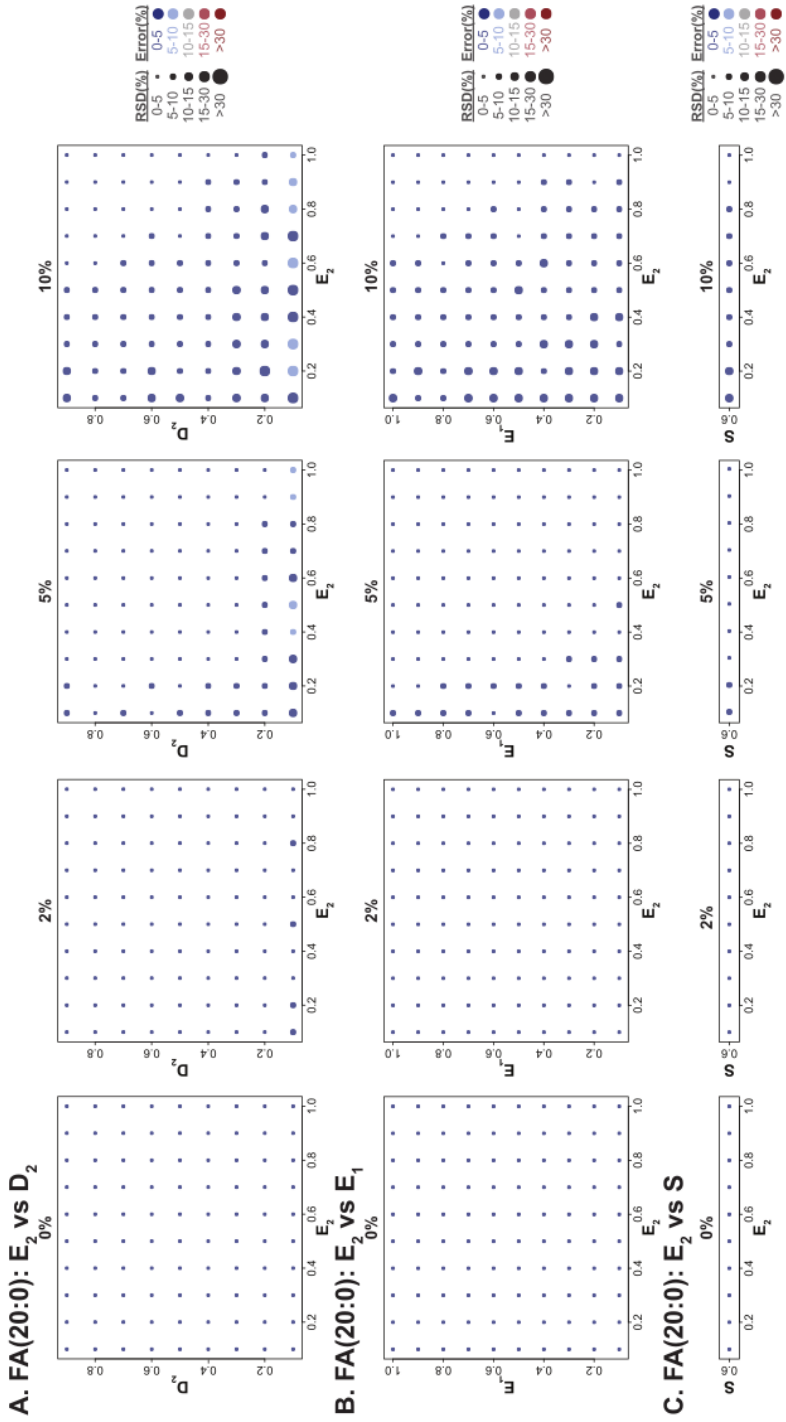
B. FA(18:0): E_1 vs S

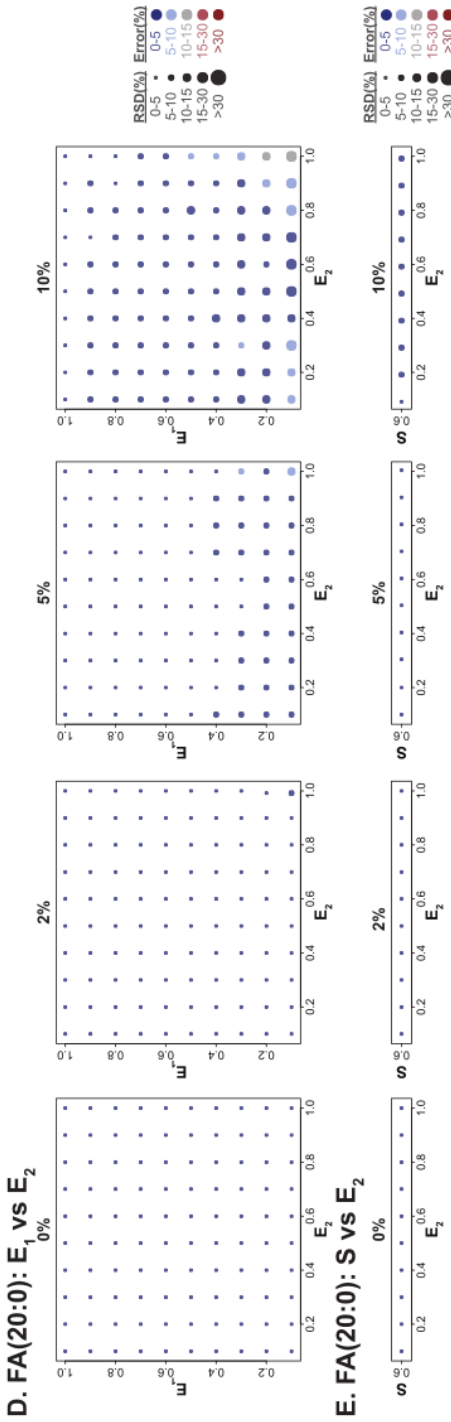


C. FA(18:0): S vs E_1

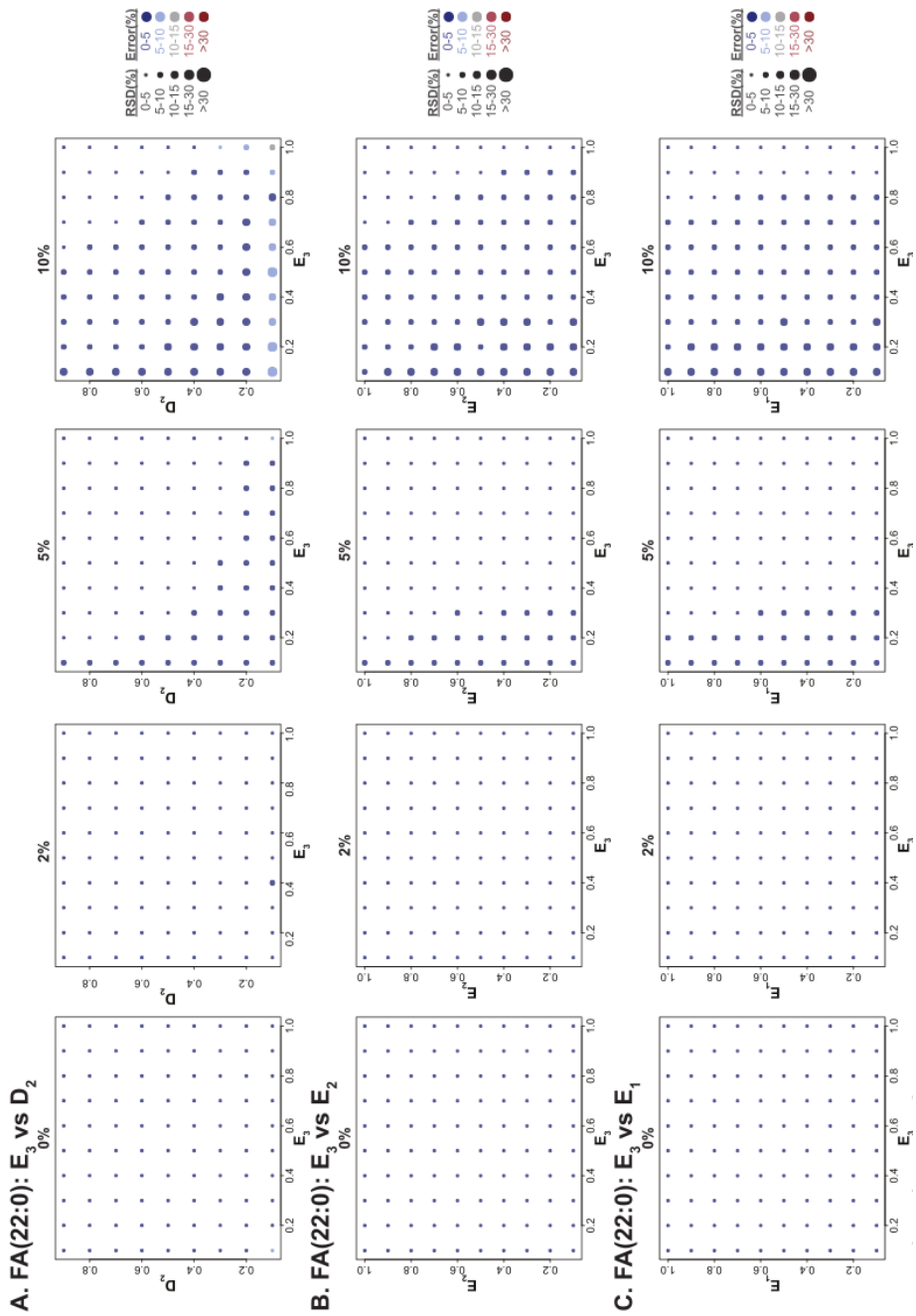


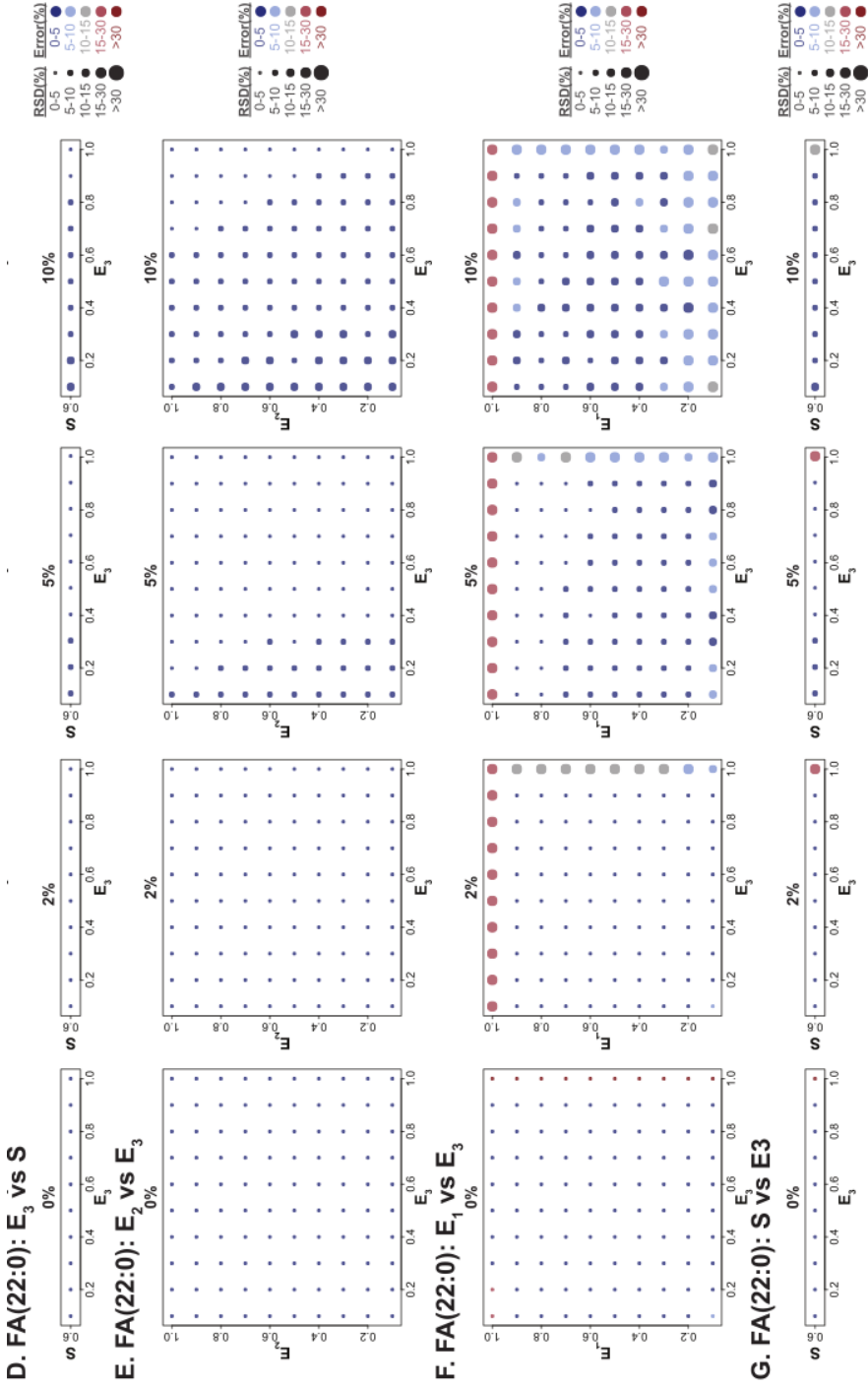
Additional Figure S30. *In silico* validation of the estimation of the *de novo* synthesis of FA(18:0). To evaluate FAMetA's ability to estimate parameters of elongation, the following values were set to simulate the mass-isotopologue data: D_1 and ϕ were set at 0.05, and 0.01, respectively, D_2 varies from 0.1 to 0.9, and E_i and S from 0.05 to 1. The 0%, 25, 5% and 10% noise levels were added to obtain 10 different noised distributions for each set of parameters. A-B, Evaluation of E_i and D_2 (A), and E_i and S (B). C, Evaluation of S as a function of E_i and S .



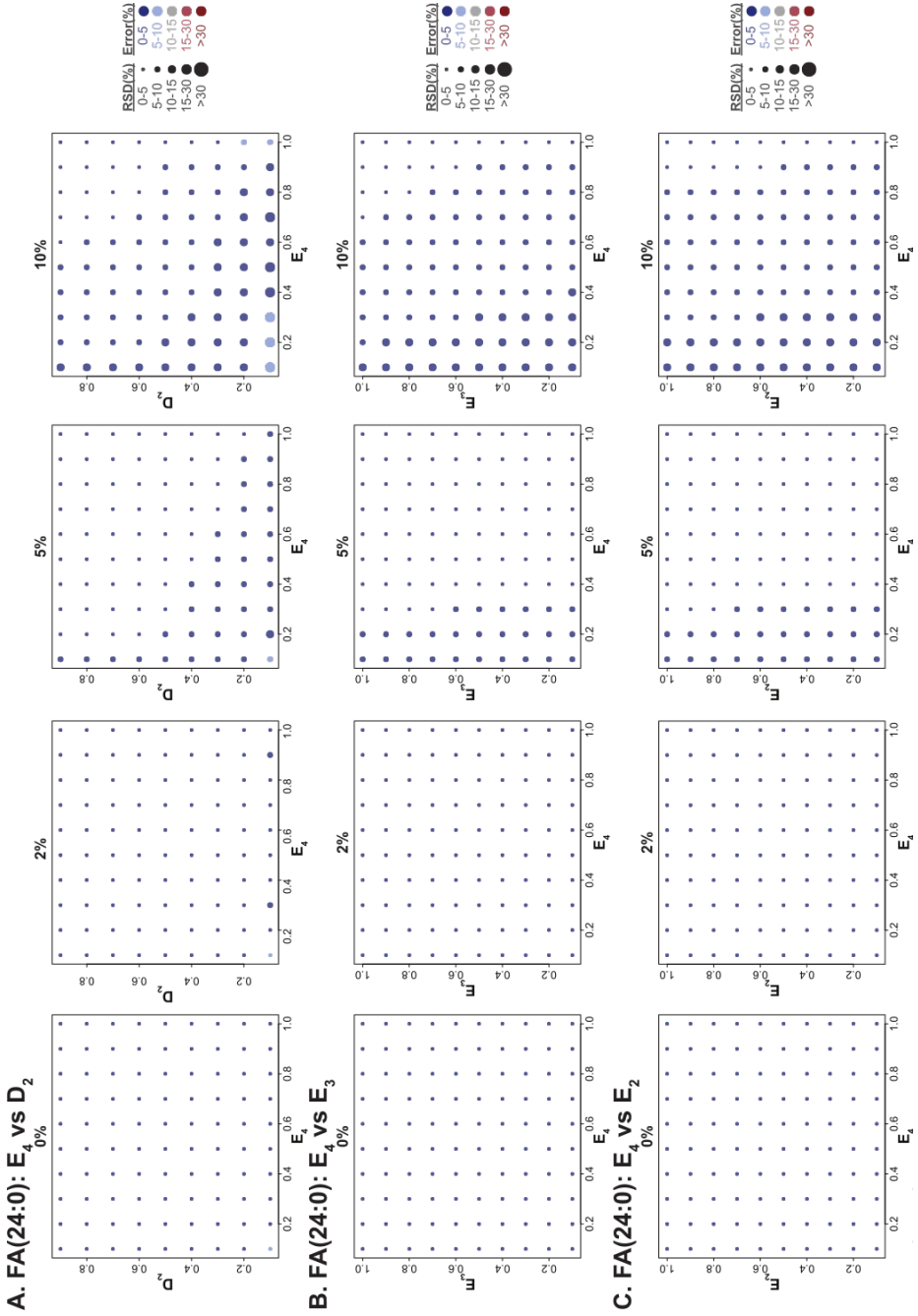


Additional Figure S31. *In silico* validation of the estimation of the *de novo* synthesis of FA(20:0). To evaluate FAMetA's ability to estimate parameters of elongation, the following values were set to simulate the mass-isotope data: S , D_1 and ϕ were set at 0.6, 0.05 and 0.01, respectively, D_2 varies from 0.1 to 0.9, and E_n from 0.1 to 1. The 0%, 2%, 5% and 10% noise levels were added to obtain 10 different noised distributions for each set of parameters. A-C, Evaluation of E_2 as a function of E_1 and D_2 (A), E_2 and E_1 (B), and E_2 and S (C). D, Evaluation of E_1 as a function of E_2 and E_1 . E, Evaluation of S as a function of E_2 and S .

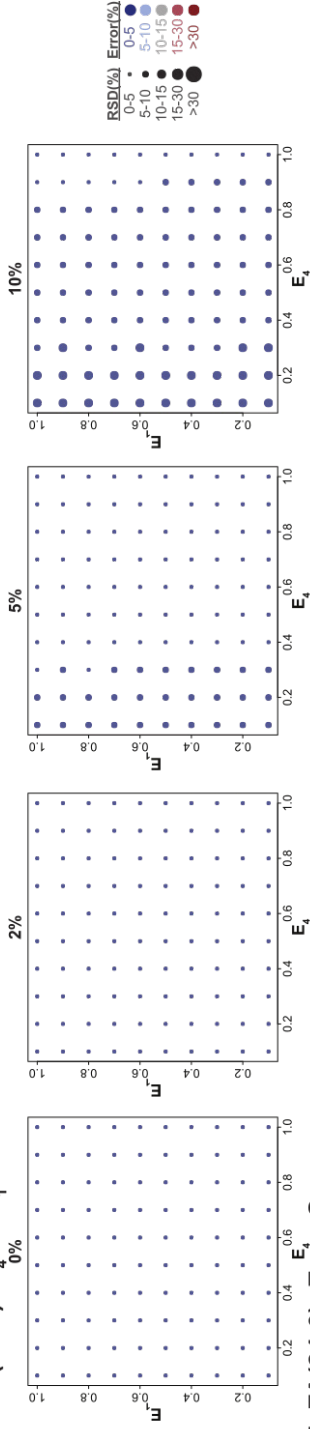




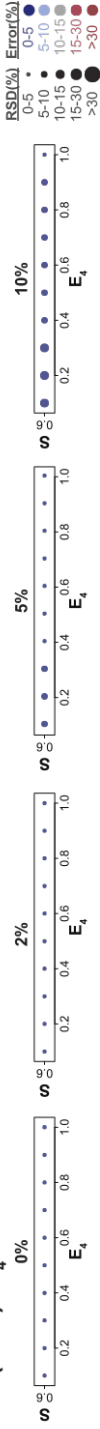
Additional Figure S32. *In silico* validation of the estimation of the *de novo* synthesis of FA(22:0). To evaluate FAMetA's ability to estimate parameters of elongation, the following values were set to simulate the mass-isotopologue data: S , D_1 and ϕ were set at 0.6, 0.05 and 0.01, respectively, D_2 varies from 0.1 to 0.9, and E_n from 0.1 to 1. The 0%, 2%, 5% and 10% noise levels were added to obtain 10 different noised distributions for each set of parameters. A-D, Evaluation of E_3 as a function of E_5 and D_2 (A), E_5 and E_2 (B), E_5 and E_1 (C), and E_5 and S (D). E, Evaluation of E_2 as a function of E_3 and E_2 . F, Evaluation of E_1 as a function of E_3 and E_1 . G, Evaluation of S as a function of E_3 and S .



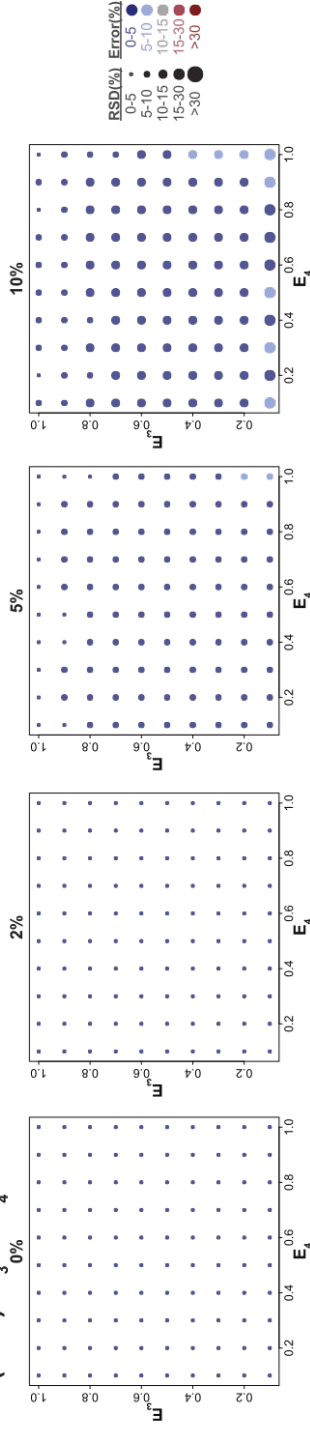
D. FA(24:0): E_4 vs E_1



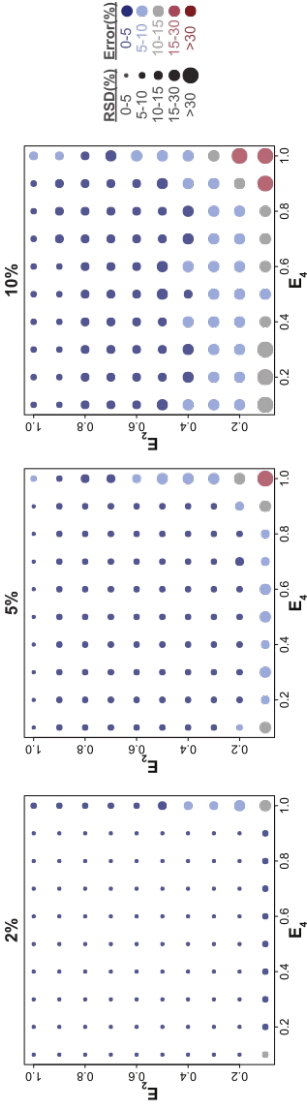
E. FA(24:0): E_4 vs S



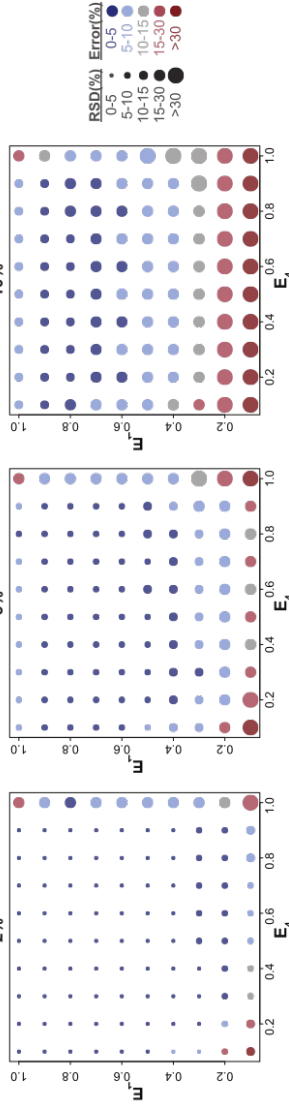
F. FA(24:0): E_3 vs E_4



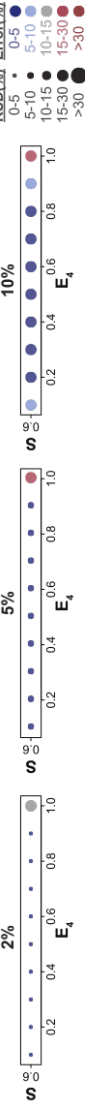
G. FA(24:0): E_2 vs E_4
0%



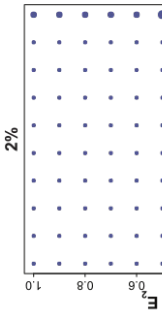
H. FA(24:0): E_1 vs E_4
0%



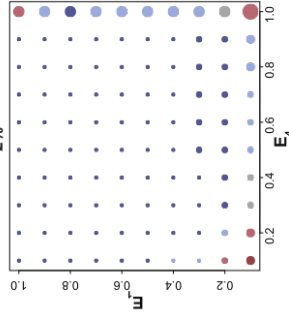
I. FA(24:0): S vs E_4
0%



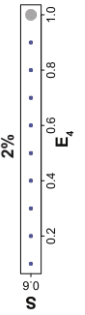
2%



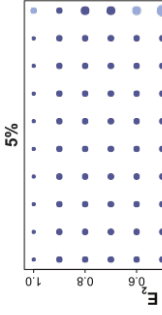
2%



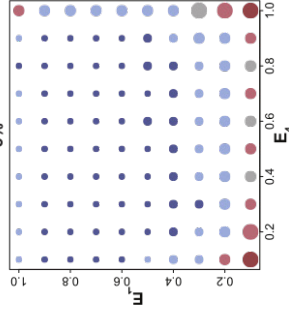
2%



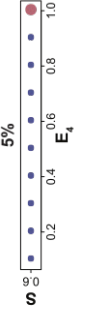
5%



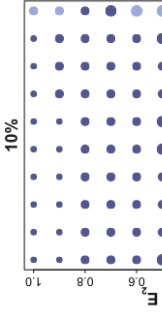
5%



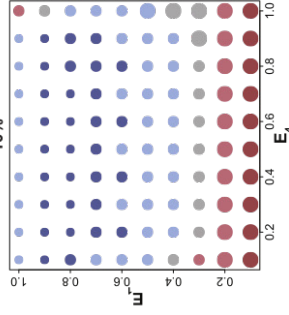
5%



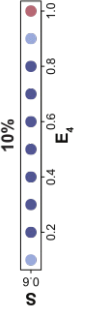
10%



10%



10%

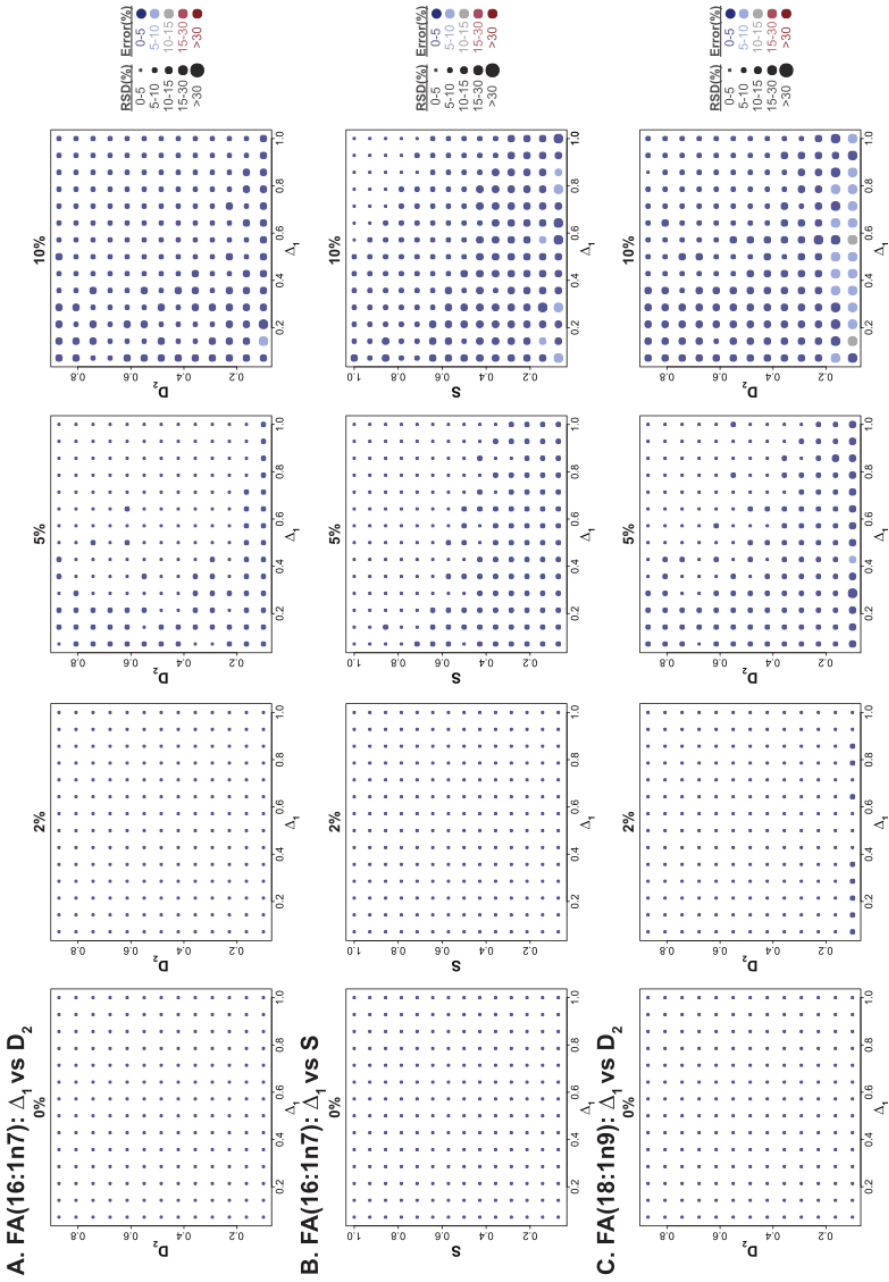


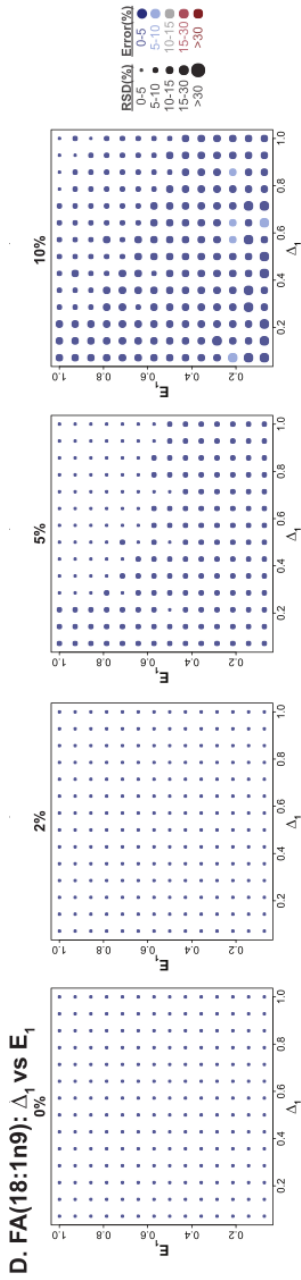
RSD(%) Error(%)
 0-5 ● 0-5
 5-10 ● 5-10
 10-15 ● 10-15
 15-30 ● 15-30
 >30 ● >30

RSD(%) Error(%)
 0-5 ● 0-5
 5-10 ● 5-10
 10-15 ● 10-15
 15-30 ● 15-30
 >30 ● >30

RSD(%) Error(%)
 0-5 ● 0-5
 5-10 ● 5-10
 10-15 ● 10-15
 15-30 ● 15-30
 >30 ● >30

Additional Figure S33. *In silico* validation of the estimation of the *de novo* synthesis of FA(24:0). To evaluate FAMetA's ability to estimate parameters of elongation, the following values were set to simulate the mass-isotopologue data: S , D_1 and ϕ were set at 0.6, 0.05 and 0.01, respectively, D_2 varies from 0.1 to 0.9, and E_n from 0.1 to 1. The 0%, 2%, 5% and 10% noise levels were added to obtain 10 different noised distributions for each set of parameters. A-E, Evaluation of E_4 as a function of E_1 , E_2 (A), E_3 and E_5 (B), E_4 and E_2 (C), E_4 and E_1 (D), and E_3 and S (E). F, Evaluation of E_3 as a function of E_4 and E_5 . G, Evaluation of E_2 as a function of E_4 and E_5 . H, Evaluation of E_1 as a function of E_4 and E_5 . I, Evaluation of S as a function of E_4 and S .





Additional Figure S34. *In silico* validation of the estimation of FA(16:1n7) and FA(18:1n9). To evaluate FAMetA's ability to estimate parameters of desaturation, the following values were set to simulate the mass-isotopologue data: D_i and ϕ were set at 0.05 and 0.01, respectively, D_2 varies from 0.1 to 0.9, Δ_1 varies from 0.1 to 1; for FA(16:1n7), S varies from 0.1 to 1; for FA(18:1n9) S is set at 0.6 and E_i varies from 0.1 to 1. The 0%, 2%, 5% and 10% noise levels were added to obtain 10 different noised distributions for each set of parameters. A-B, Evaluation of Δ_1 for FA(16:1n7) as a function of D_2 and Δ_1 , (A) or S and Δ_1 , (C). C-D, Evaluation of Δ_1 for FA(18:1n9) as a function of D_2 and Δ_1 , (C) or E_i and Δ_1 , (D).

