LipidMS 3.0: an R-package and a web-based tool for LC-MS/MS data processing and lipid annotation

María Isabel Alcoriza-Balaguer [1], Juan Carlos García-Cañaveras [1,*], Francisco Javier Ripoll-Esteve [2], and Agustín Lahoz [1,3,*]

[1]Biomarkers and Precision Medicine Unit, Medical Research Institute-Hospital La Fe, Av. Fernando Abril Martorell 106, Valencia, 46026, Spain.

[2] Department of Informatics, Medical Research Institute-Hospital La Fe, Av. Fernando Abril Martorell 106, Valencia, 46026, Spain.

[3]Analytical Unit, Medical Research Institute-Hospital La Fe, Av. Fernando Abril Martorell 106, Valencia, 46026, Spain.

*email: juancarlos_garcia@iislafe.es, agustin.lahoz@uv.es

Abstract

Summary: LipidMS was initially envisioned to use fragmentation rules and data-independent acquisition (DIA) for lipid annotation. However, data-dependent acquisition (DDA) remains the most widespread acquisition mode for untargeted LC-MS/MS-based lipidomics. Here we present LipidMS 3.0, an R package that not only adds DDA and new lipid classes to its pipeline, but also the required functionalities to cover the whole data analysis workflow from pre-processing (i.e., peak-peaking, alignment and grouping) to lipid annotation. We applied the new workflow in the analysis of a serum dataset acquired in MS, DDA and DIA modes. Our results show that LipidMS 3.0 data pre-processing outperforms XCMS and complements those lipids annotated using MS-DIAL, one of the most widely used tools in lipidomics. To extend and facilitate LipidMS 3.0 usage among less experienced R-programming users the workflow has been also implemented as a web-based application.

Availability and Implementation: LipidMS R-package is freely available at https://CRAN.R-project.org/package=LipidMS and as a website at http://www.lipidms.com.

1 Introduction

Lipids are a heterogeneous group of non-polar molecules with important biological roles, such as structural components of cell membranes, energy storage or signaling molecules (Wenk, 2005). Recent advances made in mass spectrometry (MS)-based analyses of the lipidome, lipid annotation tools and *in silico*-generated spectra databases have enabled the characterized lipidome to continuously expand (Züllig *et al.*, 2020). Liquid chromatography (LC) coupled with high-resolution mass spectrometry is the most prominent analytical platform for untargeted lipidomics (Roca *et al.*, 2021). Currently, there are two main approaches to generate LC-MS/MS data: data-dependent acquisition (DDA), in which precursors from MS1 (full-scan) are selected and immediately fragmented (MS2) to generate a spectrum that can be directly queried against a spectra database (Koelmel, Kroeger, Gill, *et al.*, 2017); and data-independent acquisition (DIA), where no precursor from MS1 is isolated and all the ions are subsequently fragmented (Guo and Huan, 2020). A wide variety of tools have been developed for lipid annotation based on both DDA and DIA (Koelmel, Kroeger, Ulmer, *et al.*, 2017; Alcoriza-Balaguer *et al.*, 2019; Tsugawa *et al.*, 2015; Hartler *et al.*, 2017). Of them, the LipidMS R package (Alcoriza-Balaguer *et al.*, 2019) was envisioned to use fragmentation rules and DIA to annotate lipids with different structural elucidation levels. However, it required external software to process large datasets and missed the option to use DDA. Here we present LipidMS 3.0, a new R package release and web-based application (www.lipidms.com), that enables the use of DDA, adds new lipid classes to its repertoire, and includes new functionalities to cover the whole data processing workflow from pre-processing (i.e., peak-peaking, alignment and grouping) to lipid annotation (**Figure 1**). To test LipidMS, a dataset acquired in MS, DDA and DIA modes obtained upon the LC-MS/MS lipidomic analysis of human serum has been used. LipidMS v3.0 data pre-processing was compared to

XCMS 3.16 (Smith *et al.*, 2006) whereas lipid annotation was compared to MS-DIAL 4.80 (Tsugawa *et al.*, 2015), the current state of the art for lipidomics analysis.

2 Features and implementation

2.1 Data preprocessing

LipidMS was initially designed to annotate lipids in single samples files and, to this end, external software was required for batch data preprocessing. In its new version, LipidMS 3.0 includes a battery of data preprocessing functions to analyze untargeted LC-MS and LC-MS/MS datasets:

- ***Peak-picking***. The first step in LipidMS preprocessing consists in extracting a peak list for each sample of the dataset using the enviPick algorithm (https://CRAN.R-project.org/package=enviPick). This function also includes annotation for the ^{13}C isotopes based on mass difference, relative intensity and correlation between peaks.
- ***Peak alignment***. Once peak lists have been extracted, recurrent peaks from different samples are grouped based on mass-to-charge ratio (*m/z*) and retention time (RT) clustering (**Figures S1 and S2**) and then, LOESS regression is applied to correct RT drifts among samples.
- ***Peak grouping***. After samples alignment, the peaks from the different samples that correspond to the same feature are grouped based on *m/z* and RT by using the same clustering algorithms as in peak alignment (**Figures S1 and S2**), but by employing more restrictive parameters for m/z and RT tolerance. The result is a feature matrix for the dataset.
- ***Filling peaks***. Once all feature peaks have been defined, peak areas are extracted again by a target approach that avoids missing peaks and improves area estimations.

Further details of these preprocessing functions can be found in the **Supplementary Information**.

2.3 Data-dependent acquisition

Despite the important advancements made for DIA-based annotation, DDA remains the most widespread data acquisition mode in LC-MS/MS (Guo and Huan, 2020). For this reason, LipidMS 3.0 now incorporates both data acquisition methods. For DDA, the same previously established fragmentation rules are applied (Alcoriza-Balaguer *et al.*, 2019), but now fragments are searched directly in an MS/MS spectra linked with a specific precursor from MS1. For each precursor ion, the MS/MS spectra used for annotation are the closest to the precursor RT, which improves differentiation between isomeric lipid species.

2.3 New lipid classes

LipidMS initially covered 24 classes, for which fragmentation rules were optimized based on experimental and *in silico*-based spectra (Alcoriza-Balaguer *et al.*, 2019). Since it was first released, new fragmentation rules for six lipid classes (acylceramides, ceramides phosphate, plasmanyl and plasmenyl phosphocholines and phosphoethanolamines) have been obtained based on the experimental spectra of available commercial lipids. Hence, these new lipid classes are now added to the LipidMS 3.0 repertoire (**Supplementary Tables S1-S3** and **Supplementary Figures S3-S14**).

2.4 LipidMS workflow

LipidMS 3.0 uses raw data files in the mzXML format, which can be obtained employing any MS file converter, such as msConvert from ProteoWizard (Chambers *et al.*, 2012), and a csv metadata file, which contains sample names, acquisition mode (MS, DDA or DIA) and sample type (blank, quality control, group 1, etc.) as input (**Figure 1**). The recommended workflow to follow is to acquire samples in MS mode and pooled QC samples in MS mode, which is used for data

normalization, and in DDA or DIA, which is utilized to perform lipid annotation. Data pre-processing is performed with the algorithms and parameters optimized for common lipidomics approaches in which the elution of multiple isomers in a narrow RT window is common (Alcoriza-Balaguer *et al.*, 2019). Lipid identification is based on fragmentation rules using DIA (based on co-elution) or DDA (clean spectra) data, and finally the potential identities are matched to the whole dataset. LipidMS provides two outputs: a table that summarizes the intensity and identity of all the detected lipids across samples; and plots depicting information that supports the proposed lipid identities, as well as the achieved level of confidence for each identification (Alcoriza-Balaguer *et al.*, 2019).

2.5 Implementation

LipidMS 3.0 has been developed as an R package 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. While the R environment presents several advantages, such as highly flexible and customizable workflows that allow the management of large datasets, it requires R programming skills to make the most of its usage. To access to a broader range of users, LipidMS 3.0 has been also implemented as a web-based application with a user-friendly GUI interface (**Figure S15**), which is accessible at http://www.lipidms.com. Example data files, scripts and tutorials for the R package and the web application can be found at http://www.lipidms.com via the "Resources" tab.

2.6 Performance Evaluation

To evaluate the performance of LipidMS 3.0, a serum dataset acquired using MS, DIA and DDA modes and obtained upon untargeted LC-MS lipidomic analysis was used. LipidMS 3.0 workflow was compared to a combination of data pre-processing with XCMS 3.16 (Smith *et al.*, 2006) and

lipid annotation with LipidMS 3.0 and to MS-DIAL 4.80 workflow (Tsugawa *et al*., 2015). The automated annotation of LipidMS 3.0 workflow outperformed the combination of XCMS and LipidMS 3.0 annotation, likely due to the pre-processing algorithms implemented in LipidMS, which achieve a better extraction of the multiple isomeric species typically found in lipidomic studies (**Supplementary Figure S16-S17**). Compared to MS-DIAL, LipidMS 3.0 retrieved a lower number of lipid annotations but provided a higher level of structural information. This increased confidence can be attributed to the strong constrains applied in LipidMS, which is based on fragmentation rules, compared to MS-DIAL, which is based on spectral matching (**Supplementary Results** and **Supplementary Figures S16-S18**).

3 Conclusions

Lipid annotation remains a bottleneck in lipidomic data analyses. To overcome it, LipidMS 3.0 offers an integral workflow that allows high-confidence lipid identification and quantification using DDA and DIA data and fragmentation rules. Overall, we demonstrate the significant improvements achieved by the new LipidMS release and how its use complements the information provided by existing tools. We believe that LipidMS 3.0 is a valuable addition to existing tools and has the potential to become a key resource to annotate complex lipids. LipidMS 3.0 is an open-access and platform-independent software that can be executed locally with an R environment (https://CRAN.R-project.org/package=LipidMS) or online by http://www.lipidms.com with no installation requirements.

Funding

M.A.A.-B. is supported by a PFIS contract from the Carlos III Health Institute of the Spanish Ministry of Economy and Competitiveness [FI18/00224]. J.C.G.-C. is supported by a grant from the Conselleria de Sanidad Universal y Salud Pública, Generalitat Valenciana, as part of Plan GenT, Generació Talent [DEI-01/20-C]. A.L. is supported by the European Regional Development Fund (FEDER) and the Carlos III Health Institute of the Spanish Ministry of Economy and Competitiveness [PI20/00580]. Part of the equipment used in this work was co-funded by the Generalitat Valenciana and European Regional Development Fund (FEDER) funds (PO FEDER of Comunitat Valenciana 2014-2020).

References

Alcoriza-Balaguer,M.I. *et al.* (2019) LipidMS: An R Package for Lipid Annotation in Untargeted Liquid Chromatography-Data Independent Acquisition-Mass Spectrometry Lipidomics. *Anal. Chem.*, **91**, 836–845.

Chambers,M.C. *et al.* (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.*, **30**, 918–920.

Guo,J. and Huan,T. (2020) Comparison of Full-Scan, Data-Dependent, and Data-Independent Acquisition Modes in Liquid Chromatography–Mass Spectrometry Based Untargeted Metabolomics. *Anal. Chem.*, **92**, 8072–8080.

Hartler,J. *et al.* (2017) Deciphering lipid structures based on platform-independent decision rules. *Nat. Methods*, **14**, 1171–1174.

Koelmel,J.P., Kroeger,N.M., Gill,E.L., *et al.* (2017) Expanding Lipidome Coverage Using LC-

MS/MS Data-Dependent Acquisition with Automated Exclusion List Generation. *J. Am. Soc. Mass Spectrom.*, **28**, 908–917.

Koelmel,J.P., Kroeger,N.M., Ulmer,C.Z., *et al.* (2017) LipidMatch: An automated workflow for rule-based lipid identification using untargeted high-resolution tandem mass spectrometry data. *BMC Bioinformatics*, **18**, 1–11.

Roca,M. *et al.* (2021) Reviewing the metabolome coverage provided by LC-MS: Focus on sample preparation and chromatography-A tutorial. *Anal. Chim. Acta*, **1147**, 38–55.

Smith,C.A. *et al.* (2006) XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.*, **78**, 779–787.

Tsugawa,H. *et al.* (2015) MS-DIAL: Data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Methods*, **12**, 523–526.

Wenk,M.R. (2005) The emerging field of lipidomics. *Nat. Rev. Drug Discov.*, **4**, 594–610.

Züllig,T. *et al.* (2020) Lipidomics from sample preparation to data analysis: a primer. *Anal. Bioanal. Chem.*, **412**, 2191–2209.

Figure 1. LipidMS 3.0 workflow. LipidMS is an open-access and platform-independent software for lipid identification from LC-MS/MS-based lipidomics that can be executed locally with an R environment (https://CRAN.R-project.org/package=LipidMS) or online (www.lipidms.com). Briefly, the LipidMS workflow comprises the following steps: 1) raw data files in the mzXML format and a csv metadata file (sample, acquisition mode, which can be MS, DDA or DIA, and sample type) are used as input; 2) data preprocessing is executed, including peak-peaking, alignment, grouping and peak filling; 3) lipids are annotated based on the established fragmentation rules for those samples acquired in DIA or DDA; 4) two main outputs are returned: a data matrix containing 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.