LCMS: An R package for automated semitargeted analysis in lipidomics

Caroline Peltier, Glenda Vasku, Marine Crépin, Stephanie Cabaret, Olivier Berdeaux

Received: 13 April 2023 | Revised: 19 October 2023 | Accepted: 13 December 2023

DOI: 10.1002/ejlt.202300077

European Journal of Lipid Science and Technology

SHORT COMMUNICATION

LCMS: An R package for automated semitargeted analysis in lipidomics

Caroline Peltier[1,2] | **Glenda Vasku**[1,2] | **Marine Crépin**[1,2] | **Stephanie Cabaret**[1,2] | **Olivier Berdeaux**[1,2]

[1]Centre des Sciences du Goût et de l'Alimentation, CNRS, INRAE, Institut Agro, University of Bourgogne, Dijon, France

[2]CNRS, INRAE, PROBE Research Infrastructure, ChemoSens Facility, Dijon, France

Correspondence
Caroline Peltier, Centre des Sciences du Goût et de l'Alimentation, CNRS, INRAE, Institut Agro, University of Bourgogne, Dijon, France.
Email: caroline.peltier@inrae.fr

Abstract

While nontargeted analysis aims to profile and report the relative distributions of a wide range of molecules from different lipid classes/subclasses, its major challenge is the annotation and identification of the molecules. Semitargeted analysis circumvents the problem by establishing a (potentially large) list of molecules to be targeted in the samples that are identified before the analysis. This approach is particularly adapted for lipid analysis to help with the automation of lipid annotation and identification. However, the manual extraction of peaks for many molecules and many samples is time consuming. Consequently, an automation of these extractions is deeply required. This paper presents a free R package for the automation of semitargeted analysis for lipid analysis. From raw files collected with LC-MS device and a list of molecules to target (containing their class), it automatically returns Excel files containing the intensities for each targeted molecule and each sample. This package allows a fast computation of the intensities. Furthermore, it guarantees the reproducibility of the results and is freely available and user-friendly.

Practical Applications: With the help of the R package presented in this paper, the use of semitargeted lipidomics as an alternative to untargeted analysis should be investigated by more labs. Work on the comparisons between the approaches could be conducted. While untargeted methods are mostly used, they require long pretreatments and identification of molecules of interest. On the contrary, in semitargeted analysis, once the integration table and retention time are obtained, the results are fast and directly interpretable. An idea for lipidomics would be to use untargeted lipidomics to compute the integration table and retention table, then use semitargeted analysis for a fast computation of well identified molecules.

KEYWORDS
lipidomics, mass spectrometry, R package, semitargeted

Abbreviations: LC, MS liquid chromatography mass spectrometry; PC, phosphatidylcholine; TIC, total ion chromatogram.

1 | INTRODUCTION

Lipids are complex and heterogeneous molecular entities, playing an intricate and key role in understanding biological activities and disease processes. Lipidomics aim to identify and quantitatively define the lipid classes/subclasses, including their molecular species.[1] According to Checa et al.,[2] two major analytical strategies exist in lipidomics: targeted and nontargeted analyses. In a targeted lipidomic approach, a limited number of predefined lipid-specific signals are used to establish precisely and accurately relative abundances of the expected endogenous lipids. Therefore, targeted analysis offers high specificity and accurate quantification. Besides, nontargeted lipidomic approaches are used for a global analysis of all measurable lipids present in a sample and must be coupled to chemometric methods to extract valuable information of relevant signals, which are subsequently identified (by database searching and/or comparison to analytical standards). Different software solutions for lipid identifications and data processing (such as Workflow4 Metabolomics, MZmine, MS-dial, Lipidsearch software, etc.) exist to run such approaches. One of the major limitations in untargeted approaches is the identification of the molecules.

Semitargeted approach[3] is an alternative that falls in the middle between untargeted and targeted ones aiming to identify and relatively quantify numerous metabolites. While it focuses on known lipid classes or specific groups of lipids (for example, defined in a database), it also allows for the detection and exploration of unexpected or novel lipid species within those classes. It provides a balance between specificity and the potential for discovery.

These experiments are considered semitargeted because, while the list of metabolites is defined, the hypothesis may not target only one specific metabolite[3] but rather a large number of them. However, doing this task manually is time consuming and monotonous. In addition, existing software that perform such tasks are time-consuming and present challenges, especially when large batches of samples are being analyzed. Therefore, the objective of this paper is to present an R package dedicated to automated semitargeted analysis for lipidomic analysis.

2 | MATERIALS AND METHODS

2.1 | Semitargeted LC-MS analysis protocol

A semitargeted procedure consists in different steps. First, lipids of the samples are extracted then recaptured in a relevant solvent mixture. Adding a standard for each lipid class to be studied is recommended. Then, the sequence should be executed according to a dedicated experimental design including blanks, regular quality controls (that are a pool constituted of the same quantity of all samples, also called QC) and samples.

In parallel to this step, a few samples (preferentially QC) are analyzed for the identification of lipids to be targeted. Such approach is based on solid identifications/annotations of each lipid through the use of tandem mass spectrometry (MS/MS or MS^2), a combination of manual spectral elucidation or usage of existing software designated to perform lipid identifications (e.g., W4M, Lipid Maps, Lipidsearch software). This aims to produce a basis of lipid species that will be called "integrationTable" in the rest of the paper. This preliminary step requires the expertise of lipidomists and needs to be continuously updated.

Finally, a retention time window is required for each lipid class. It is obtained with the extracted chromatograms of specific species. We recommend utilizing internal standards that might be used to point the respective RT window limitations for lipid classes. Internal standards need to be very close to the class that is being analyzed and preferably marked (e.g., deuterated). If not, other similar nondeuterated standards are acceptable, only if they have similar chemical characteristics with the lipid class. These retention times will be stored in "classTable" in the rest of the paper. Equipped with the raw data, integrationTable and classTable, the user is ready for analyzing data with the LCMS R package.

2.2 | An R package for automatizing semi targeted LC-MS

R (R Core Team, 2020) is a free software dedicated to statistical analysis that allows packages to be created. The package developed for semitargeted LC-MS was called LCMS. It uses other packages such as MSnbase for reading .mzXML files,[4] ggplot2 for the graphical outputs, and openxlsx for working with Excel files.

2.3 | Inputs of the package

The pipeline of lipid analysis is represented on Figure 1 and is based on three inputs: the raw data, the integrationTable, and the classTable.

The raw data (1 in Figure 1) should be mzXML files. MSConvert software can be used to transform. raw as .mzXML files.

The integration Table (2 in Figure 1) is an Excel file containing the targeted lipid base organized in several columns: "class" for the class of the lipid species, "name" for the name of the species, "compo" for the composition of the lipid, and "mz" for the ratio *m/z*. When standards are present, an additional column "std" should be added, containing "yes" if the molecule is a standard, "no" otherwise. This allows standards not to be included in statistics about the samples.

The class time table, or classTable (3 in Figure 1), is an Excel file containing three columns: "class" for the class to be considered, "rtmin" for the minimal limit of the retention time window to be considered for this class, "rtmax" for the maximal limit of the retention time window. Times must be entered in seconds (and not in minutes).

Further parameters can also be entered such as ppm resolution required to find a match that need to be chosen on the resolution of the mass spectrometer or display options of the resulting Excel file.

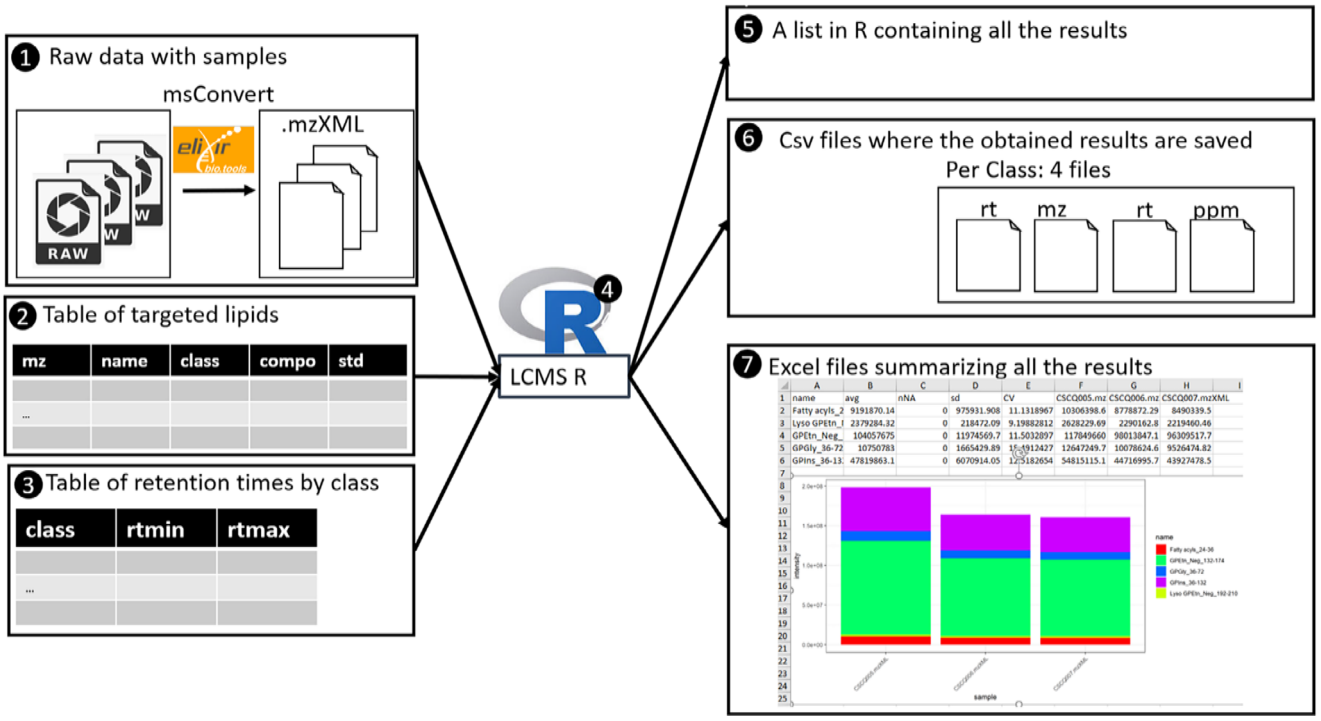


FIGURE 1 Inputs (1, 2, 3) and outputs (5, 6, 7) of the LCMS R package. (1) Represents the raw data stored as .mzXML files, (2) is the table of targeted lipids (example of integrationTable), (3) is a the table of retention times (example of classTable).

2.4 | Details of the algorithm and its parameters

The use of LCMS package is based on three steps: (i) doing all calculations with the *calculateIntensity* function that returns an R object (5 in Figure 1), (ii) writing the resulting csv files containing the results with the function *getCsvOfIntensity* (intermediate results, 6 in Figure 1), and (iii) writing an Excel file synthetizing all the results with some statistics, specific color code and graphs (7 in Figure 1) with the function *getExcelOfIntensity*.

2.4.1 | Calculation of intensities

Each .mzXML file contains all centroided peaks with their *m/z*, retention time, and values. Additionally, each molecule of the *integrationTable* has a specific class and theoretical *m/z* (denoted as m_{th}). The corresponding retention times are referenced as intervals in *classTable* (denoted as rt_{min} and rt_{max}). For one given centroided peak, its intensity is calculated as the equivalent of area under the curve (AUC) of the chromatographic peak, which corresponds to its value multiplied by the time resolution. The time resolution is estimated as the average difference between two successive retention times.

Next, given a *ppm* parameter (part per million, default to 10), the algorithm simply consists of summing all the AUC of the centroided peaks whose *m/z* values fall in the interval $[m_{th} - ppm \times 10^{-6} \times m_{th}; m_{th} + ppm \times 10^{-6} \times m_{th}]$ and whose retention time are in the range $[rt_{min}, rt_{max}]$.

The parameter minimalNumberOfPoints (default set to 0) allows for the calculation of the intensity only for a minimal number of centroid peaks in the chromatogram (to avoid potential anomalies).

In contrast to other software solutions, no Gaussian approximation or chromatogram smoothing are employed, which can result in slight differences between results.

Consequently, the main parameters to be used at this stage are the path to the repository where are stored the .mzXML files: *ppm* parameter and *minimalNumberOfPoints* parameter.

The function returns an R object allowing R users to easily use their results for their favorite statistical analyses (PCA, PLS, univariate models, CoDa).

2.4.2 | Results of the package

Steps (ii) and (iii) allow the results to be written on the computer.

For each class, the function *getCsvOfIntensity* creates four csv files whose lines are molecules and columns are samples: (i) "mzo" files containing the average of *m/z* observed, (ii) "ppm" files containing the ppm observed, (iii) "int" containing the absolute obtained intensities, (iv) "pct" containing the obtained relative intensities. An additional csv file named "param" contains the used parameters.

The only parameters required for this function are a result of *calculateIntensity* and the path of the repository where to save the results. Then, these .csv files are used to produce a single user-friendly Excel file containing all the results with the following organization: (i) a "class"

tab for each lipid class, (ii) a "total" tab for the sum on all lipid species, and (iii) a parameter tab containing all the parameters used for the analysis (useful to reproduce exactly the analysis). Regarding the "class" tabs, for each lipid species, the intensity of each sample was reported and the average, standard deviations, and coefficient of variation (CV) were calculated.

Several parameters are available for this function: *output*, *CVthresold*, *timeSd*, and *includeStd*.

The parameter *output* specifies if the .xlsx is built with intensities ("int") or relative intensities ("pct").

A color code was implemented such as the cells containing intensities higher than the average + *timeSd* × standard deviation was highlighted in blue (and green for intensities lower than this quantity).

Cells corresponding to CV were also highlighted in red when higher than the *CVthreshold* value.

For each class, the composition of each sample is also displayed in a barplot where different species are displayed with different colors.

Finally, when standards were added in the samples, the *includeStd* parameter (TRUE or FALSE) specifies if the standards are displayed (or not) in the distributions.

Regarding the "total" tab, intensities were also summed by class then summarized with the same statistics (average, standard deviation, and CV).

A "totalPct" tab is also added, corresponding to the percentages of intensities of the different classes.

2.5 | Illustration on an example

The use of the package was illustrated on the data of three quality controls of biological tissues (retinas) obtained after liquid chromatography coupled with an Orbitrap high-resolution mass spectrometer (Orbitrap Fusion Tribrid mass spectrometer, Thermo Scientific, USA) with Heated Electrospray Ionization H-ESI in centroid mode. The .raw files were extracted using the software Xcalibur 4.1.

The protocol is fully described in Vasku et al.[5] These results are included in the package and can be reproduced by any user. We chose to present results for negative ionization mode that contains less classes and lipid species than positive ionization mode and consequently results in smaller files. Of course, the package can be similarly used for positive mode, without constraints.

Regarding the obtention of the integration table, approximately 15 lipid classes were assessed, and about 500 lipid species from these classes/subclasses were annotated. The identification/annotation process was focused to MS/MS spectra in negative and positive ionization mode of pooled retina QC. MS/MS was used to distinguish parent molecular ions and their fragmented molecular ions, resulting in the annotation of the molecule. This was accomplished by employing the software Lipidsearch (Thermo Scientific, USA), which can automatically annotate lipid species from raw files. However, this approach requires further examination, which was achieved by manual spectrum elucidation of each species. In this purpose, manual spectral elucidation was performed by examining the spectrum of each molecule ion to determine its structure. For instance, *PC (18:0/16:0)* is a phosphatidylcholine (PC) species comprised of two saturated fatty acids in positions *sn*-1 and *sn*-2. This species was annotated by examining the MS/MS fragmented molecule in negative and positive modes. The negative MS/MS fragmentation of the molecular ion $[M\text{-}H]^-$ reveals the *m/z* of each fatty acid that composes it (*m/z* 283 for 18:0 and *m/z* 255 for 16:0). While the positive MS/MS fragmentation of molecular ion $[M+H]^+$ shows a major fragment *m/z* 184 corresponding to the polar head (choline), indicating that the species is a PC.[6] Numerous lipid species were manually annotated in this manner, which indicated a level 3 of lipid metabolite identification, as has been suggested by the LIPID MAPS initiative.

The resulting lipid species, named as "integration table" along with the accurate *m/z*, were then employed for data processing in R.

The class table was obtained on the extracted chromatograms of standards with Freestyle software (Thermofisher, USA).

Regarding the parameters of R functions, the Excel file was computed with intensity with ppm = 10, a thresholdCV = 10, and a timeSd = 1. These values were chosen for illustrating the color code in the example, but we recommend to use thresholdCV = 30 and timeSd = 3 as default values (see Supplementary Data 1).

2.6 | Stability, validity, and availability of the R package

In order to validate the method, intensities were manually extracted from Freestyle software (Thermofisher, USA) then automatically with LCMS R package in order to compare them. Details of this study are available in Supplementary Data 2 and in the package "./tests/test_compareAll.r." It shows correlations higher than 0.99 between the log intensities was calculated on 354 species for a sample.

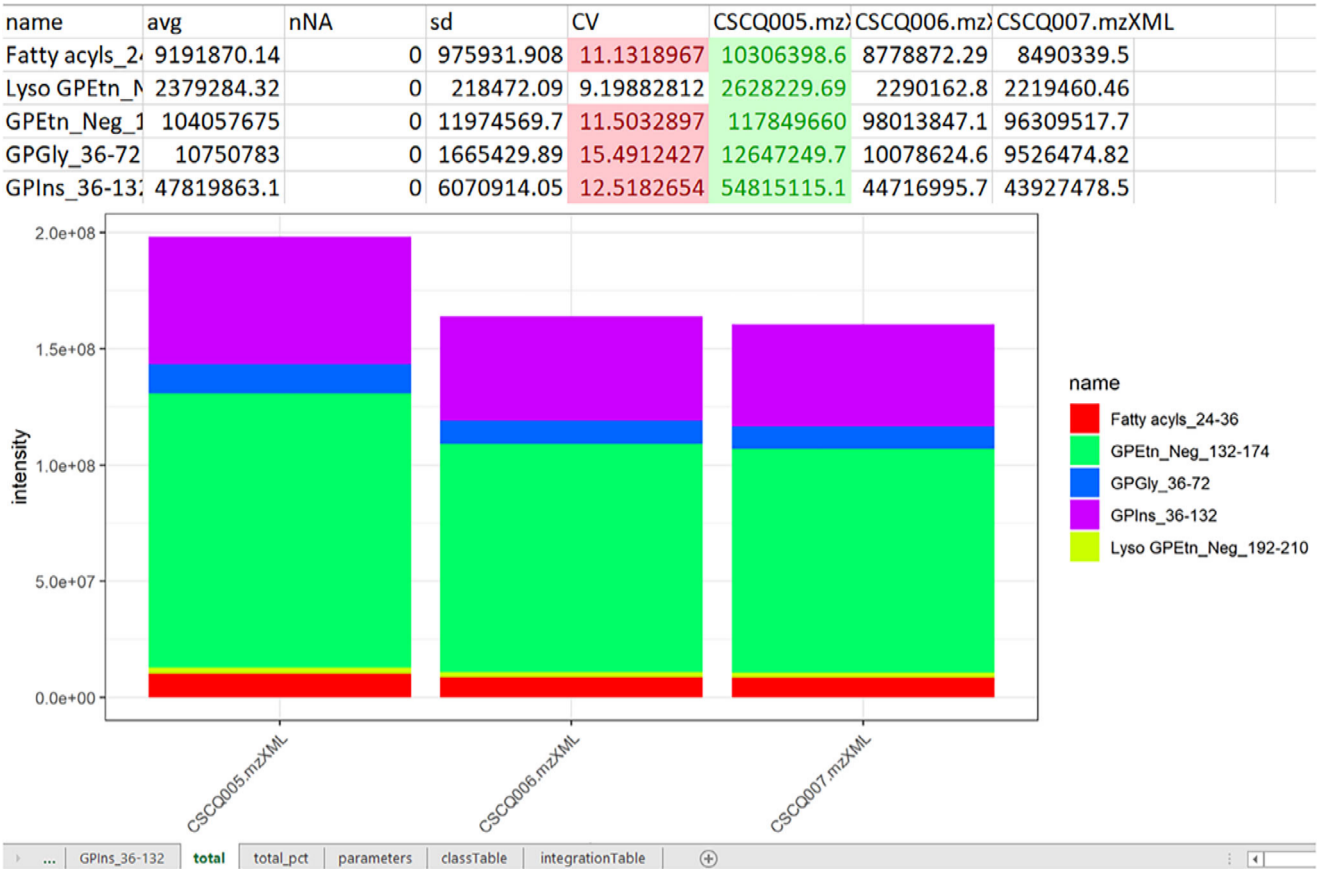
In order to assess the stability of the packages, automatic tests were also included in the package guaranteeing that any update does not modify the results. The R package LCMS presented in this paper is freely available on www.github.com/ChemoSens/LCMS.

A R version higher than R.4.0.2 is required for this package. Compilation and tests were conducted with computers with the following characteristics: ≥16 Go of RAM, Intel Core i7, and Windows 10 Pro, 64 bits.

3 | RESULTS AND DISCUSSION

3.1 | Example of results

Figure 2 shows a part of the Excel file ("total" tab) obtained for the example dataset (included in the package). The Excel file is available in Supplementary Data 3. The "total" tab is similar to all other tabs: each line corresponds to a class (species in "class" tab). First, some statistics are given: for FattyAcyls, the average intensity



name	avg	nNA	sd	CV	CSCQ005.mz)	CSCQ006.mz)	CSCQ007.mzXML
Fatty acyls_2	9191870.14	0	975931.908	11.1318967	10306398.6	8778872.29	8490339.5
Lyso GPEtn_N	2379284.32	0	218472.09	9.19882812	2628229.69	2290162.8	2219460.46
GPEtn_Neg_1	104057675	0	11974569.7	11.5032897	117849660	98013847.1	96309517.7
GPGly_36-72	10750783	0	1665429.89	15.4912427	12647249.7	10078624.6	9526474.82
GPIns_36-13	47819863.1	0	6070914.05	12.5182654	54815115.1	44716995.7	43927478.5

FIGURE 2 Screenshot of the "total" tab of the resulting Excel file (with parameters ppm = 10, thresholdCV = 10, timeSd = 1). For each class, it includes the intensities obtained for each sample (CSCQ005, CSCQ006, CSCQ007) and statistics such as the average (avg), number of missing values (nNA), standard deviation (sd), and coefficient of variation (cv). The barplot indicates the intensity of each class for each sample.

on samples was around 9.19 million, the standard deviations was 0.97 million and for the file CSCQ005, the measured intensity was higher than 10 million. As this quantity was higher than the average + timeSd × standard deviations, this intensity was highlighted in green. The coefficient of variation (CV) that were higher than 10 were also correctly highlighted in red. The barplot represents the distribution of each class. All samples had similar distributions and were highly concentrated in GPEtn_Neg (in green, Figure 2). The number after the class corresponds to the retention times used during calculation.

3.2 | Discussion around the input parameters

The proposed semitargeted approach in lipidomics requires a preestablished database, in which the correct *m/z* values are given for each molecular ion. These *m/z* are calculated prior through (i) Lipidsearch software and (ii) manual spectral elucidation of QCs. In addition to the *m/z* information, it is important to add the retention time information for each class as the minimum and maximum retention time (RT). Such practice reduces the annotation errors for each molecular ion. It must be stated that the annotation process in lipidomics is challenging, and nowadays there is not yet a standardized approach that utilizes a general lipid database that can be used on almost every sample (human or animal) in order to identify lipid species with high credibility. At this point, a lipid database is required to be established prior to data processing, preferably in QC samples that represent theoretically all analyzed samples; therefore, the major limit of semitargeted approach is the establishment of the integration table and the molecular classes/subclasses/species. Another way to obtain integrationTable would be to run untargeted analysis on a few samples, obtain the molecules to be targeted, then use this table for semitargeted analysis.

Regarding the class table, the chromatogram should be accurately analyzed in order to find out retention times adapted to all samples and all lipid species in each class. It implies that the chromatograms are well aligned. Further works could be done for including the alignment of chromatograms.

4 | CONCLUSIONS

The proposed semitargeted approach in lipidomics is an automated data processing pipeline that can be adapted to different samples emerging from mass spectrometry data. Although nowadays there is not yet a standardized option for lipidomics, efforts have been made

to opt for a more accurate analysis. However, untargeted approaches in lipidomics remain a challenge due to many problematics such as internal standards, a wide variety of lipid classes that require different analytical techniques and lastly, long manual annotation processes. Semitargeted analysis represents a compelling alternative to untargeted analysis. It is based on the constitution of a database of prior identified molecular ions. Using this database, the LCMS R package presented in this paper automates the extraction of intensities of numerous samples, in order to avoid the time-consuming process of manual annotation of each single sample. Therefore, this package could be useful for any lipidomist interested in exploring semitargeted analysis. Future efforts in developing a protocol for establishing the database of lipid species or in comparing untargeted and semitargeted approaches would be of significant interest and utility in the field of lipidomics analysis.

AUTHOR CONTRIBUTIONS

Caroline Peltier: Conceptualization (equal); data curation (equal); methodology (lead); software (lead); validation (equal); writing—original draft (equal); writing—review and editing (equal). **Olivier Berdeaux**: Funding acquisition (equal); methodology (equal); supervision (equal); validation (equal); writing—review and editing (equal). **Glenda Vasku**: Conceptualization (equal); data curation (equal); methodology (equal); software (supporting); validation (equal); writing—original draft (supporting); writing—review and editing (equal). **Marine Crépin**: Data curation (equal); methodology (equal); software (equal); validation (equal); writing—review and editing (equal). **Stephanie Cabaret**: Data curation (equal); methodology (equal); software (equal); validation (equal); writing—review and editing (equal).

DATA AVAILABILITY STATEMENT

The package is freely available on github with the related data.

ACKNOWLEDGEMENTS

All the people that have participated to this paper are also authors of this document.

CONFLICT OF INTEREST STATEMENT

The authors have declared no conflict of interest.

ORCID

Caroline Peltier https://orcid.org/0000-0002-7926-955X

REFERENCES

1. Ramani Venkata, A., & Ramesh, M. (2021). A concise review on lipidomics analysis in biological samples. *ADMET and DMPK*, *9*(1), 1–22. https://doi.org/10.5599/admet.913
2. Checa, A., Bedia, C., & Jaumot, J. (2015). Lipidomic data analysis: Tutorial, practical guidelines and applications. *Analytica Chimica Acta*, *885*, 1–16. https://doi.org/10.1016/j.aca.2015.02.068
3. Liu, X., & Locasale, J. W. (2017). Metabolomics: A primer. *Trends in Biochemical Sciences*, *42*(4), 274–284. https://doi.org/10.1016/j.tibs.2017.01.004
4. Gatto, L., & Lilley, K. (2012). MSnbase—An R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics*, *28*, 288–289.
5. Vasku, G., Peltier, C., He, Z., Thuret, G., Gain, P., Gabrielle, P.-H., Acar, N., & Berdeaux, O. (2023). Comprehensive mass spectrometry lipidomics of human biofluids and ocular tissues. *Journal of Lipid Research*, *64*(3), 100343. https://doi.org/10.1016/j.jlr.2023.100343
6. Berdeaux, O., Juaneda, P., Martine, L., Cabaret, S., Bretillon, L., & Acar, N. (2010). Identification and quantification of phosphatidylcholines containing very-long-chain polyunsaturated fatty acid in bovine and human retina using liquid chromatography/tandem mass spectrometry. *Journal of Chromatography A*, *1217*(49), 7738–7748. https://doi.org/10.1016/j.chroma.2010.10.039

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Peltier, C., Vasku, G., Crépin, M., Cabaret, S., & Berdeaux, O. (2024). LCMS: An R package for automated semitargeted analysis in lipidomics. *European Journal of Lipid Science & Technology*, *126*, e2300077. https://doi.org/10.1002/ejlt.202300077