Bioinformatics, 33(11), 2017, 1744–1746
doi: 10.1093/bioinformatics/btx046
Advance Access Publication Date: 31 January 2017
Applications Note

OXFORD

Data and text mining

LIQUID: an-open source software for identifying lipids in LC-MS/MS-based lipidomics data

Jennifer E. Kyle*,†, Kevin L. Crowell†, Cameron P. Casey, Grant M. Fujimoto, Sangtae Kim, Sydney E. Dautel, Richard D. Smith, Samuel H. Payne and Thomas O. Metz*

Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Jonathan Wren

Received on June 9, 2016; revised on January 20, 2017; editorial decision on January 23, 2017; accepted on January 27, 2017

Abstract

Summary: We introduce an open-source software, LIQUID, for semi-automated processing and visualization of LC-MS/MS-based lipidomics data. LIQUID provides users with the capability to process high throughput data and contains a customizable target library and scoring model per project needs. The graphical user interface provides visualization of multiple lines of spectral evidence for each lipid identification, allowing rapid examination of data for making confident identifications of lipid molecular species. LIQUID was compared to other freely available software commonly used to identify lipids and other small molecules (e.g. CFM-ID, MetFrag, GNPS, LipidBlast and MS-DIAL), and was found to have a faster processing time to arrive at a higher number of validated lipid identifications.

Availability and Implementation: LIQUID is available at http://github.com/PNNL-Comp-Mass-Spec/LIQUID.

Contact: jennifer.kyle@pnnl.gov or thomas.metz@pnnl.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A key bottleneck in LC-MS/MS-based lipidomics studies is the ability to process spectral data in a high throughput manner and simultaneously provide confident identification and accurate quantification of detected lipid species (Cajka and Fiehn, 2014) as incorrect identifications can lead to misleading biological interpretations. Because the lipidomics community has not yet developed a robust approach for estimating the false discovery rate associated with lipid molecular species identification based on MS/MS spectra, identifications reported by current software tools should be manually verified by the user to ensure accuracy. As larger scale studies are increasingly desired (Chan *et al.*, 2012), analysts need improved software for lipid identification. Many tools are not designed for high-volume verification of identifications. In particular, few programs use accessory evidence such as isotopic profiles or MS level extracted ion chromatograms. Here we introduce an open-source lipid identification software, LIQUID (Lipid Quantification and Identification), where: (i) the scoring model is trainable, (ii) the search database is customizable, (iii) the experimental lines of evidence used to make confident identifications of lipids are easily accessible, (iv) single target and (v) fragment pattern search are available to enable tracking of similar and repeating patterns of MS/MS spectra corresponding to unidentified lipids.

2 Implementation

LIQUID was developed using C# .Net version 4.5.1 and Windows Presentation Foundation (WPF) and is available at http://github.com/PNNL-Comp-Mass-Spec/LIQUID. Lipid species included in the LIQUID reference database were seeded from LipidMaps (Sud *et al.*, 2007), and the classification, naming and nomenclature follow the conventions used therein. To allow for the expansion of lipids not

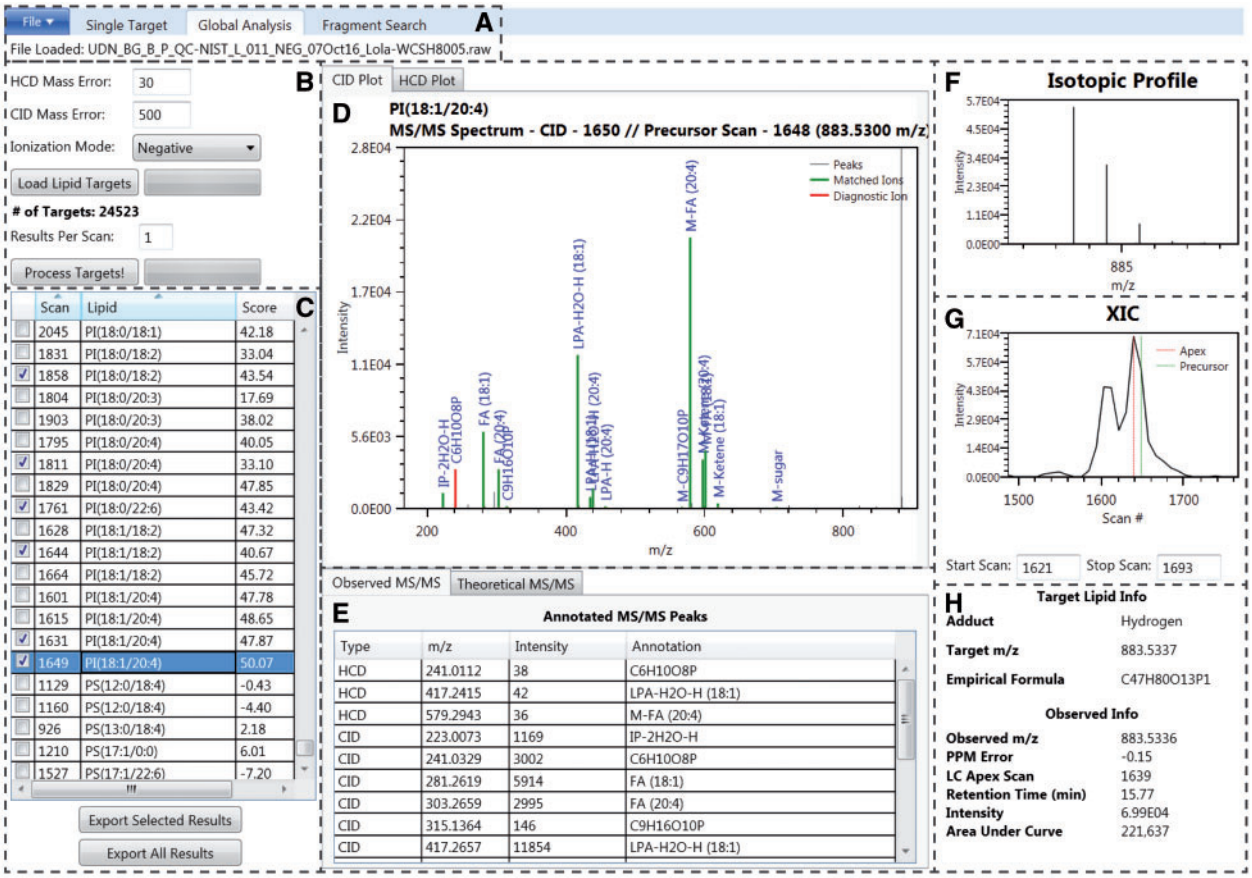

Fig. 1. LIQUID interface. **(A)** File input. **(B)** Search parameters. **(C)** Table of output. Checked boxes denote validated identifications. The highlighted species, PI(18:1/20:4), is selected as an example for the remaining displayed features. **(D)** MS/MS spectra with annotated diagnostic (red) and fragment ions (green). **(E)** List of observed MS/MS fragments, including their structural annotations. The theoretical MS/MS fragments for the associated lipid species are also available. **(F)** Isotopic profile of the lipid species selected for MS/MS analysis and generated using data from MS scans. **(G)** Extracted ion chromatogram (XIC) for the precursor ion. The peak apex is highlighted in red while the nearest MS level scan is highlighted green. Peak start and stop scan numbers can be input for determination of peak area, which is shown in (H). **(H)** Target lipid information

currently indexed by LipidMaps (e.g. non-mammalian and newly discovered classes and species), information needed to identify new lipids can be added to the database following the formats outlined by LipidMaps. The current version of the LIQUID database has added an additional >1500 lipid species beyond those contained in LipidMaps.

LIQUID scores lipid spectra based on expected fragment peak intensities learned from training data, following the approach used previously to model peptide spectra (Payne *et al.*, 2008). For any given lipid class, each fragment is characterized by its observed intensity, yielding a probability table. The final score of a lipid/spectrum match is the sum of log likelihood scores for each fragment. Retraining fragmentation probabilities and adding new species to the reference database are both described in the Supplementary Information.

3 Results

The current LIQUID reference database contains over 21,300 unique lipid targets across 6 lipid categories, 26 lipid classes and 66 subclasses (Supplementary Tables S1–S2), including yeast and plant lipids. The reference target list can be tailored based on which lipids are expected to be detected in the sample. For example, lipid subclasses found in photosynthetic membranes (e.g. sulfoquinovosyldiacylglycerol) could be excluded in a mammalian cell line study. The ability to customize the lipid species and classes in the target list allows for informed sample processing based upon biological knowledge. Single target analysis and fragment pattern search (i.e. to enable tracking of similar and repeating patterns of MS/MS spectra corresponding to unidentified lipids) are also available (Supplementary Figs S1 and S2).

LIQUID enables high throughput lipidomics data analysis and semi-quantitative identification of lipid species from LC-MS/MS data. We demonstrate these capabilities with global analysis of data from NIST SRM 1950 human plasma (Phinney *et al.*, 2013) resulting in 297 confident lipid identifications across 18 lipid subclasses (see Supplementary Information and Tables S8–S9). Processing the uploaded data file (Fig. 1A) using the lipid target list (Fig. 1B) requires approximately 1 min per LC-MS/MS file that typically contains over 10 000 MS/MS spectra. LIQUID accepts .raw (Thermo Scientific) and mzML files. Principle features of the software that allow users to make confident identifications include (i) annotated MS/MS spectra that highlight fragment peaks, such as the diagnostic ion (highlighted red) and associated chain ions (highlighted green) (Fig. 1D and E) that match to those found in the theoretical database (Fig. 1E), (ii) the observed isotopic profile of the precursor ion (Fig. 1F), (iii) the shape of the extracted ion chromatogram (XIC) of

the precursor ion revealing the locations of both the peak apex and MS level scan closest to the corresponding MS/MS spectra (Fig. 1G), (iv) observed mass measurement error from predicted and (v) retention time (Fig. 1H).

LIQUID also enables users to specify the number of identification results per MS/MS scan (Fig. 1B), which allows the identification of co-eluting species (Supplementary Fig. S3). Many lipid species are well separated using reversed-phase LC; however, challenges in separating particular isobaric lipid species (e.g. triacylglycerides) and between lipid subclasses (e.g. diacylglycerophosphocholines and diacylglycerophosphoethanolamines) makes confident identifications difficult. In addition to isobaric and isomeric species within a single MS peak, the XIC visualization feature assists in structural isomer identification that are LC separated (Supplementary Fig. S4). Although structural details of isomers cannot always be assigned, their presence should be documented as they can be altered in perturbed systems (Kyle *et al.*, 2016) and have different biological functions. Additional features offered by LIQUID are described in the Supplementary Information.

Confidently identified lipids are exported in an output file containing the lipid common name and Lipid Maps nomenclature, exact and observed *m/z*, retention time, instrument scan numbers corresponding to the LC peak apices and MS/MS scan events, LC peak apex intensity and associated database identifiers [i.e. Lipid MAPS (Sud *et al.*, 2007), PubChem (Kim *et al.*, 2016), InChi (Heller *et al.*, 2015), HMDB (Wishart *et al.*, 2009)]. Once exported, output files can be re-loaded at a later date if re-analysis of the data is required.

LIQUID was compared to other freely avalible software commonly used to identify lipids and other small molecules [e.g. CFM-ID (Allen *et al.*, 2014), MetFrag (Ruttkies *et al.*, 2016), GNPS (Wang *et al.*, 2016), LipidBlast (Kind *et al.*, 2013) and MS-DIAL (Tsugawa *et al.*, 2015)] using data from LC-MS/MS-based lipidomics analysis of both the NIST SRM 1950 human plasma sample and a mixed standard comprised of 19 authentic lipid molecular species (see Supplementary Information and Tables S7–S28). LIQUID was found to confidently identify more lipid species than any of the other softwares and with overall faster combined processing and validation time.

All the tested software programs have strengths and features that make them unique. One of LIQUID's clear strengths is the larger number of fragment ions used to make a confident identification versus other softwares. For example, GNPS and MS-DIAL will base some lipid class identifications on only the diagnostic ion, leaving the chain composition undetermined and therefore only allowing the class but not lipid molecular species to be identified (see Supplementary Information and Tables S18–S19, S22–S23 for examples). Additionally, each matched fragment in LIQUID is clearly annotated with the molecular formula and/or chain in which it is derived. These fragment ions are also highlighted with color allowing for their rapid examination to determine if all fragments (e.g. diagnostic head group of phospholipids and both fatty acyl chains) are identified.

Funding

This work was supported by the National Institute of Allergy and Infectious Diseases [U19AI106772], and the Department of Energy (DOE), Office of Biological and Environmental Research (OBER), Genomic Science Program, under the Pacific Northwest National Laboratory (PNNL) Pan-omics project. Lipidomics data were generated in the Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the U.S. DOE and located at PNNL in Richland, Washington. PNNL is a multi-program national laboratory operated by Battelle for the DOE under Contract DE-AC05-76RLO 1830.

Conflict of Interest: none declared.

References

Allen,F. *et al.* (2014) CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Res.*, **42**, W94–W99.

Cajka,T. and Fiehn,O. (2014) Comprehensive analysis of lipids in biological systems by liquid chromatography-mass spectrometry. *Trends Anal. Chem. TRAC*, **61**, 192–206.

Chan,R.B. *et al.* (2012) Comparative lipidomic analysis of mouse and human brain with Alzheimer disease. *J. Biol. Chem.*, **287**, 2678–2688.

Heller,S.R. *et al.* (2015) InChI, the IUPAC international chemical identifier. *J. Cheminf.*, **7**, 23.

Kim,S. *et al.* (2016) PubChem substance and compound databases. *Nucleic Acids Res.*, **44**, D1202–D1213.

Kind,T. *et al.* (2013) LipidBlast in silico tandem mass spectrometry database for lipid identification. *Nat. Methods*, **10**, 755–758.

Kyle,J.E. *et al.* (2016) Uncovering biologically significant lipid isomers with liquid chromatography, ion mobility spectrometry and mass spectrometry. *Analyst*, **141**, 1649–1659.

Payne,S.H. *et al.* (2008) Phosphorylation-specific MS/MS scoring for rapid and accurate phosphoproteome analysis. *J. Proteome Res.*, **7**, 3373–3381.

Phinney,K.W. *et al.* (2013) Development of a standard reference material for metabolomics research. *Anal. Chem.*, **85**, 11732–11738.

Ruttkies,C. *et al.* (2016) MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J. Cheminf.*, **8**, 3.

Sud,M. *et al.* (2007) LMSD: LIPID MAPS structure database. *Nucleic Acids Res.*, **35**, D527–5D532.

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

Wang,M. *et al.* (2016) Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnol.*, **34**, 828–837.

Wishart,D.S. *et al.* (2009) HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.*, **37**, D603–D610.