Advancing Lipidomic Bioinformatic Technologies: Visualization and Phospholipid Identification (VaLID) version 3.0

Graeme S.V. McDowell[1], Alexandre P. Blanchard[1], Daniel Figeys[1], Stephen Fai[2], and Steffany A.L. Bennett[1]

[1] Ottawa Institute of Systems Biology, Neural Regeneration Laboratory, Department of Biochemistry, Microbiology and Immunology, University of Ottawa, Ottawa, K1H 8M5, Canada.
{gmcdo092, apbla037, dfigeys, sbennet}@uottawa.ca
[2] Carleton Immersive Media Studio, Azrieli School of Architecture and Urbanism, Carleton University, Ontario, K1S 5B6, Canada.
sfai@cims.carleton.ca
http://neurolipidomics.com/resources.html

Abstract. There is a paucity of bioinformatic tools for spectral analysis capable of assigning and visualizing molecular identities from mass spectrometry-derived structural information. Predicting phospholipid lipid identities is a labour-intensive process given the extreme variability in structure based on permutations of only a few atomic 'building blocks'. Moreover, our ability to visualize all theoretically possible phospholipids present in lipidomic datasets is limited. To address this gap, we created the online lipidomic bioinformatic tool Visualization and Phospholipid Identification (VaLID). This work describes the expansion of this tool to include functional search capacity linking the VaLID 3.0 database of 1,324,224 theoretically possible phospholipids to PubMed and the Human Metabolome Database (HMDB) as well as the inclusion of lipid predictions using nomenclatures now useful for researchers employing a shotgun lipidomic approach. VaLID is freely available at http://neurolipidomics.com/resources.html.

Keywords: Phospholipids • Lipidomics • Database • Mass Spectrometry

1 Introduction

The emerging field of lipidomics seeks to understand how dynamic changes in membrane composition regulate cell function [1]. Neurolipidomics is the study of cellular, regional, and systemic lipid homeostasis in the central nervous system encompassing not only the identification and measurement of individual lipid isoforms but also the mRNA and protein expression profiles of metabolic enzymes and transporters, and the protein targets that affect downstream signalling [1]. Furthermore, a lipidomic analysis includes an unbiased assessment of lipid function ranging from the physico-

chemical basis of lipid behaviour to lipid-protein and lipid-lipid interactions, and the impact of dynamic lipid metabolism on cellular response to intrinsic and extrinsic stimuli [1]. Lipidomics, the systems-level analysis of lipids and their interacting moieties [2], depends upon our ability to answer two seemingly simple questions: How many lipid species are there? And what effect does lipid diversity have on cellular function? Recent advances in high performance liquid chromatography (LC), electrospray ionization (ESI), and matrix-assisted laser desorption ionization (MALDI) mass spectrometry (MS), coupled with new membrane separation and extraction methodologies, now provide us with the means to quantify lipid diversity [3-6]. For example, LC-ESI-MS enables researchers to determine not only phospholipid subclass, but also the number of carbons and the possible number of unsaturations in the fatty acid chains linked by acyl, akyl, or alkenyl linkages to the phospholipid backbone. These and other technological capacities have revived the idea of a highly dynamic membrane, and brought the focus of membrane biology back to their lipid constituents, with determinative roles of lipids in biological processes, such as potent second messenger molecules becoming more apparent [7]. Hundreds to thousands of lipid species can now be profiled in different subcellular membrane compartments [1]. Basic unitary conceptions are being challenged. Diacylglycerol, commonly conceived by cell biologists as a single lipid, is now recognized to be a family of over 50 structurally distinct species each controlling different cellular processes [8, 9].

Yet with this success come new challenges. The lipidomics field faces four formidable roadblocks as elegantly expounded by Niemela et al., [10] : ***(1) Data Processing and Lipid Identification*** – There is a paucity of bioinformatic tools for spectral analysis capable of assigning and visualizing molecular identities from MS-derived structural information; ***(2) Statistical Analysis*** – Lipidomic datasets involve "medium-scale" data ranging from tens to hundreds of lipids per family that, in turn, impact on other related medium-sized lipid networks (i.e., requiring analysis of thousands of "features"). These sizes are difficult to analyze using traditional statistical approaches that classically consider features of less than ten yet are not amenable to genomic/proteomic statistical methodologies where features exceed thousands; ***(3) Pathway Analysis*** – We lack accessible curated databases capable of predicting biochemical, signalling, and regulatory lipid pathways affected by changes in membrane composition; (***4) Modeling Tools*** – We have little to no capacity to rapidly model the impact of altering lipid composition on biological membrane properties and cellular signalling (i.e., lipid interactome pathway analysis).

There is pressing need for bioinformatic interest in lipidomics. We simply do not have the necessary tools to mine our new rich lipid compositional datasets for functional significance. Lipidomics has yet to benefit from the same development of analytical tools available to proteomics and genomics through concerted bioinformatic efforts. The goal of our work is to develop new tools to address the first and third challenges (Lipid Identification and Pathway Analysis). As our technology improves, more precise lipid identification becomes possible [11]. A major issue lies in the identification of isobaric lipids in complex matrices – species with identical mass to charge (m/z) ratios. An example of this property can be seen in Figure 1. Genomics and proteomics can capitalize on sequence-based signatures; lipids lack easily defina

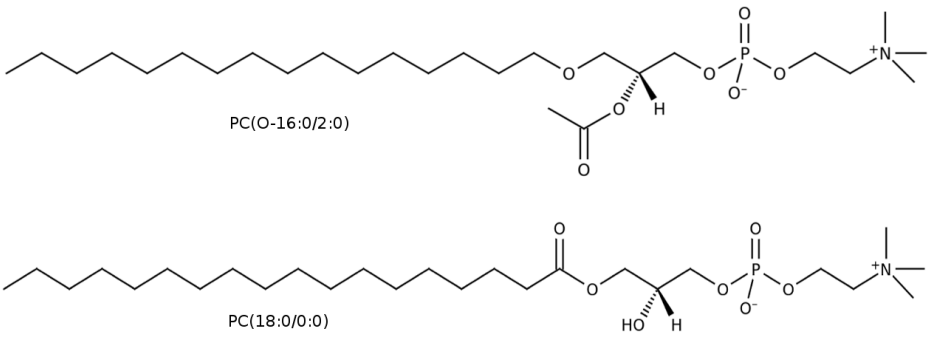

Fig. 1. Isobaric lipids. Both these lipids have the same mass to charge ratio of 523.3638 in $[M+H^+]$ mode where M is mass. Both lipids share a phosphocholine polar headgroup linked to the glycerophospholipid backbone by a phosphor-diester bond. However, the top lipid, a platelet activating factor alkylacylglycerophosphocholine, is defined by a different structure in both carbon chain identity and linkage to the glycerophospholipid backbone, activating different receptors and signalling pathways than the bottom lipid, a lyso-phosphatidylcholine.

ble molecular fingerprints. Identities must be established from structural determinants. Here, bioinformatic tools are urgently needed. For example, the LIPID metabolites and pathways strategy (LIPID MAPS) consortium[1], and LipidBank[2] have created tools that aid in the identification of lipids via MS, and have led the field in standardizing ontologies, notation, and protocols [11, 12]. In the case of LipidMAPS, their lipid databases were created from lipids curated from both literature sources and data generated by their own core laboratories and those of their partners each using specific cell systems [13]. While these databases and the associated search tools are invaluable, not all lipids are represented. Our group, working with neural membranes, has found that many species in our spectra were not present in the LipidMAPS structural databases. Thus, to participate in advancing lipidomic bioinformatics, we created a database and online prediction engine which was designed to be entirely comprehensive – Visualization and Phospholipid Identification (VaLID)[3] [14]. VaLID is web-based application linking a user-friendly search engine to an exhaustive database which contains all theoretically possible phospholipid species combined with different drawing and visualization features. In its first and second iteration, it was designed for use by researchers employing LC-ESI-MS technologies to explore lipid diversity. Here, we describe its expansion to include functional search capacity linking VaLID 3.0 databases to both PubMed[4] and the Human Metabolome Database (HMDB)[5] as

1 www.lipidmaps.org
2 http://lipidbank.jp
3 http://neurolipidomics.com/resources.html
4 http://www.ncbi.nlm.nih.gov/pubmed/
5 http://www.hmdb.ca

well as the expansion of VaLID's prediction engine to include lipid prediction using nomenclatures now useful for researchers employing a shotgun lipidomic approach.

2 Program Description

VaLID is composed of three parts: (1) a group of comprehensive phospholipid databases, (2) multiple search engines enabling lipid prediction, and now basic functional annotation, and (3) various drawing options for lipids within the database. The program is constantly being updated with new bioinformatic features in response to community feedback. We describe here the addition of linking the search engine and database with capacity to query PubMed and the HMDB for known functions and physicochemical properties of target species. To aid researchers employing shotgun lipidomic MS methodologies, we also describe addition of additional lipid nomenclatures to lipid prediction and functional query, defining species by headgroup and total number of carbons, as well as the total number of unsaturations in prediction search.

2.1 Database

The backbone for VaLID is its phospholipid database. Here, we expanded these databases to make the program amenable not only to lipidomic researchers employing LC-ESI-MS technologies but also shotgun lipidomics. To facilitate this transition, the original database was split into separate units, one per phospholipid subclass. The databases were created to be comprehensive (i.e., to contain all the possible lipids within their specific phospholipid subclasses) with one restriction. Species are limited to the realm of biological possibility as they contain only phospholipids with carbon chains ranging in length from 0 to 30 carbons at the *sn*-1 and *sn*-2 positions with combinations of up to six unsaturations in the cis position. The databases also contain species with three different chain linkage options, alkyl, acyl, and alkenyl, representing an ether, ester and vinyl ether bonds, respectively. The average and exact masses for every combination of chain *sn*-1 and *sn*-2 chain lengths, with each of the appropriate linkages are calculated, and added to the exact or average mass [15] of a defining headgroup. Here, the exact mass represents the isotopic mass of the most prevalent isotope of the atom of interest, whereas the average mass is the weighted average of the isotopic masses, given multiple isotopic masses.

The first version of VaLID released in early 2013 contained eight phospholipid subclasses, glycerophosphates (PA), glyceropyrophosphates (PPA), glycerophosphocholines (PC), glycerophosphoethanolamines (PE), glycerophosphoglycerols (PG), glycerophosphoglycerophosphates (PGP), glycerophosphoserines (PS) and cytidine 5'-diphosphate 1,2-diacyl-*sn*-glycerols (CDP-DG). This represented approximately 736,000 unique lipid species [14]. In late 2013, it was coded to include four more phospholipid groups: glycerophosphoinositols (PI), glycerophosphoinositol monophosphates (PIP), glycerophosphoinositol bisphosphates (PIP_2) and glycerophosphoinositiol trisphosphates (PIP_3) and updated to be capable of visualizing a total 1,324,224 theoretically possible phospholipids predicted from any user-inputted m/z

value and MS condition [16]. Now, we describe additional search features and linkage to functional annotation databases.

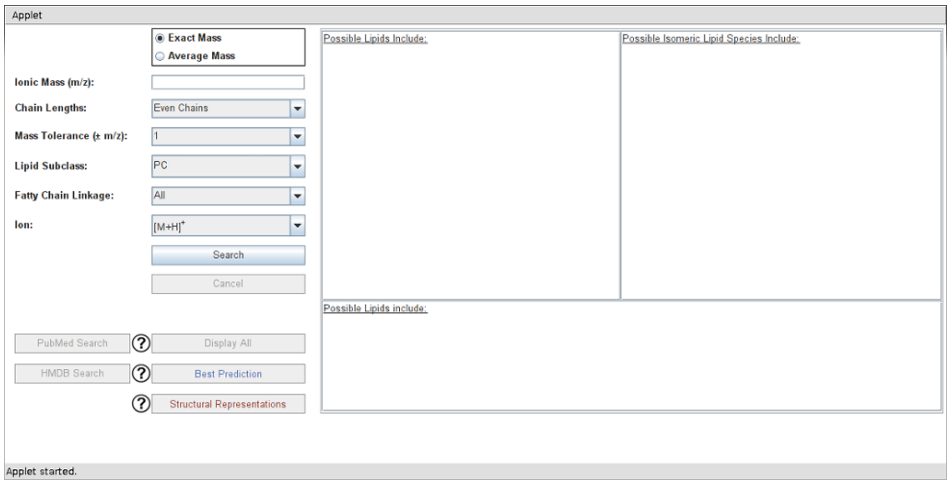

Fig. 2. Graphical Interface of VaLID 3.0

2.2 Search Function

VaLID's graphical interface (Figure 2) allows users to access the databases and return only entries tailored to their particular predetermined specifications (Figure 3). Lipid nomenclature adheres to that developed by the LipidMAPS consortium [12, 17]. Like VaLID 1.0 and 2.0, VaLID 3.0 has multiple selectable fields enabling users to customize their searches based on their particular MS methodologies and prediction requirements: exact or average mass, the ionic mass, even or odd chains, mass tolerance, lipid subclass, fatty chain linkage, and selected ion mode. Users first choose results returned for average or exact mass. The m/z ratio of lipid to be predicted is inputted in the field labelled ionic mass. The mass tolerance represents a range, from 0.0001 to 2 m/z, above and below the desired mass, to accommodate the limit of sensitivity of the user's mass spectrometer. Searches can be restricted to phospholipids with an even number, odd number, or both, carbon chain length at *sn*-1 and *sn*-2 positions. Particular lipid subclasses are defined by target phospholipid headgroup and various combinations of headgroup modification options (Figure 4). Users can specify the type and combination of fatty chain linkages that they wish to search (i.e., an ester (acyl), ether (alkyl) bond at one or both chains, a vinyl ether (alkenyl) at the *sn*-1 position, or all of these options). The MS ion mode employed by the user is inputted, enabling VaLID to predict lipid identity based on appropriate mass. We introduce in VaLID 3.0 a feature to search lipid function and properties within both PubMed and the HMDB (Figure 3). A third box for lipid results, using nomenclature defining headgroup and total number of carbons and possible total number of unsaurations in both chains encompassing all of the lipids returned in the top two boxes, is now returned for users wishing to predict species using these terms (Figure 3).

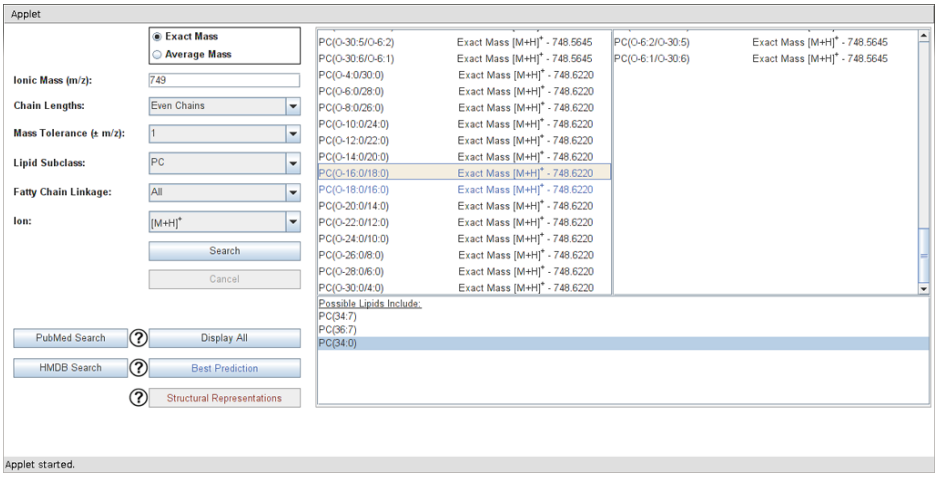

Fig. 3. An example of a search for phosphocholines with mass of 749 within a tolerance of ±1 m/z, selecting for all linkages, but restricted to even chains, from a spectra collected in $[M+H^+]$ ion mode. Possible isomeric lipids (where *sn*-1 and *sn*-2 chains are in alternate positions) theoretically possible are shown in the box on the right hand side. New to version 3.0 are the results labelled "Possible Lipids Include", containing a list of the lipid subclasses, as well as the total number of carbons in both chains, and the number of unsaturations that represent the sum total of all potential species returned in the top right and left boxes. The PubMed and HMDB search buttons are also visible and enabled once a target species is selected in one of the three result boxes.

2.3 Drawing Feature

Every lipid within the database can be drawn via VaLID's visualization feature with curated high-resolutions species identified in neural tissue by our group further presented in a series of rigid models produced using Maya® nParticles as we have described [14]. For example, PC(18:0/18:1) – that is, a glycerophosphocholine with a fatty chain of 18 carbons in length which is fully saturated, in one position, and a fatty chain, 18 carbons long with one unsaturation in the other – has many structural possibilities. Additionally, the predicted species could have *sn*-1 and *sn*-2 chains reversed. The unsaturation in one of the chains could be at any carbon along the chain. The increase in the number of unsaturations and chain length greatly increases the number

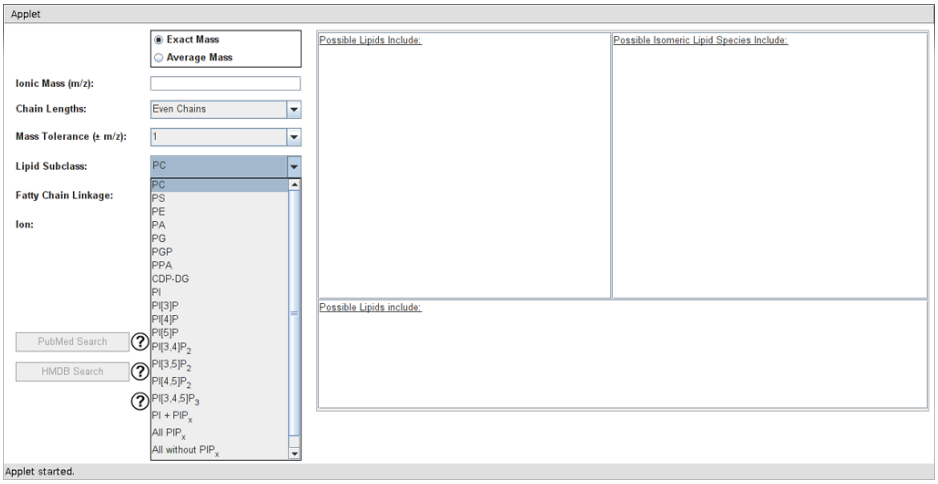

Fig. 4. The phospholipid subclasses searchable in VaLID 3.0.

of possible structures for a given lipid, representing a bioinformatic challenge to phospholipid representation. As such, we created a drawing algorithm that calculates, then draws, the atom location and bond connectivity in a 2D Cartesian plane. The results are displayed using ChemAxon's MarvinView software (Figure 5). The structures were created to match the drawing specifications laid out by the LipidMAPS consortium [12]. Using the features of MarvinView, the user can download any of the structures individually, and save them as a molecular model (.mol) file, which can be opened in multiple chemical drawing tools, such as Marvin or ChemDraw, or saved as an image, such as a PNG file. As described previously [14, 16], there are multiple drawing options available for users within VaLID's drawing component. If a lipid within the "Possible Lipids Include" or the "Possible Isomeric Lipids Include" boxes is selected, and the "Display All" button is pressed, every combination of unsaturations, as described above, will be calculated and drawn. Lipids can also appear in blue or red text. Should a lipid appear in blue font, there is a structure that can be considered more common, or "more likely" to occur. If one of these lipids is highlighted, and the "Best Prediction" button is pressed, the common structure(s) of each individual chain is drawn, as opposed to every combination. For example, if the lipid selected contains "20:4", the *Best Prediction* would return the structure of arachidonic acid, with corresponding positions of double bonds. These fatty chains, with common structures, were identified based on their relative abundance in mammalian cells [16]. Lipids that are returned in red font are lipids which have been curated by the CIHR Training Program in Neurodegenerative Lipidomics (CTPNL) in neural tissue and we provide high-resolution 3D models (VaLID view models) for download by highlighting the lipid and pressing the "Structural Representation" button. These models are derived using Maya® nParticles, converted into smooth polygonal meshes directed to the original x, y, and z coordinates, imported as points in space, recapitulating the original molecular structure developed using MarvinView as described above in an abstracted, organic, form. Resulting VaLID view models are available for download

as rigid polygons. They are also available on request fitted with a rig of movable joints between atoms to facilitate membrane reconstruction and modeling.

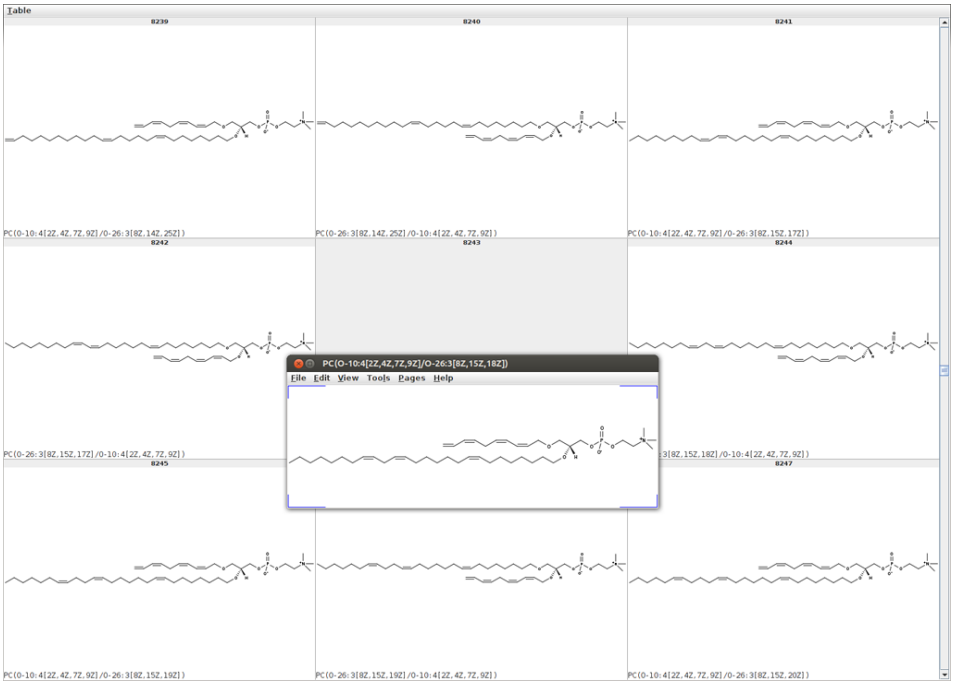

Fig. 5. When the Display All, or Best Prediction buttons are pressed, the structures of the selected lipid are calculated. When the calculations for the structures are completed, a new window pops up containing ChemAxon's MarvinView software, containing a table of all the structures. Each one of these structures can be viewed independently, where they can be saved as a molecular (.mol) file, or as an image.

3 Conclusion

The program VaLID was created initially to aid lipid researchers in predicting lipid identity from LC-ESI-MS spectra. To our knowledge, it is the first lipidomic MS prediction tool that contains all the chemically possible phospholipids up to 30 carbons in both chains, covering twelve subclasses. We describe here the addition in version 3.0 of features making VaLID's prediction capacity useful for researchers employing shotgun lipidomic approaches as well as connection of VaLID's databases to the HMDB and PubMed databases. This additions allow researchers to search existing literature for published information (where available) on 1,324,224 phospholipids. VaLID is made freely available as part of the CIHR Training Program in Neurodegenerative Lipidomics at http://neurolipidomics.com/resources.html and its data-

bases and engines are hosted by the Carleton Immersive Media Studios[6]. Version 3.0 was released at the 2014 IWBBIO 2014 (2nd International Work-Conference on Bioinformatics and Biomedical Engineering) meeting on April 7 2014.

4 Acknowledgements

This resource was funded by the Canadian Institutes of Health Research (CIHR) MOP 89999 to DF and SALB and a Strategic Training Initiative in Health Research (STIHR) CIHR/ Training Program in Neurodegenerative Lipidomics (CTPNL) and Institute of Aging TGF 96121 to DF, SF, and SALB. APB received a FRSQ and CTPNL graduate scholarship; GSVM received a CTPNL graduate scholarship.

[6] http://www.cims.carleton.ca

5 References

1. Bennett, S.A.L., Valenzuela, N., Xu, H., Franko, B., Fai, S., Figeys, D.: Using neurolipidomics to identify phospholipid mediators of synaptic (dys)function in Alzheimer's Disease. Frontiers in physiology 4, 168 (2013)
2. Wenk, M.R.: The Emerging Field of Lipidomics. Nature Reviews Drug Discovery 4, (2005)
3. Piomelli, D., Astarita, G., Rapaka, R.: A neuroscientist's guide to lipidomics. Nat Rev Neurosci 8, 743-754 (2007)
4. Brown, H.A., Murphy, R.C.: Working towards an exegesis for lipids in biology. Nat Chem Biol 5, 602-606 (2009)
5. Bou Khalil, M., Hou, W., Zhou, H., Elisma, F., Swayne, L.A., Blanchard, A.P., Yao, Z., Bennett, S.A.L., Figeys, D.: Lipidomics era: Accomplishments and challenges. Mass Spectrom Rev 29, 877-929 (2010)
6. Xu, H., Valenzuela, N., Fai, S., Figeys, D., Bennett, S.A.L.: Targeted lipidomics - advances in profiling lysophosphocholine and platelet-activating factor second messengers. The FEBS journal 280, 5652-5667 (2013)
7. Zehethofer Nicole, Pinto, D.M.: Recent developments in tandem mass spectrometry for lipidomic analysis. Analytica Chimica Acta 627, (2008)
8. Deacon, E.M., Pettitt, T.R., Webb, P., Cross, T., Chahal, H., Wakelam, M.J., Lord, J.M.: Generation of diacylglycerol molecular species through the cell cycle: a role for 1-stearoyl, 2-arachidonyl glycerol in the activation of nuclear protein kinase C-betaII at G2/M. J Cell Sci 115, 983-989 (2002)
9. Callender, H.L., Forrester, J.S., Ivanova, P., Preininger, A., Milne, S., Brown, H.A.: Quantification of diacylglycerol species from cellular extracts by electrospray ionization mass spectrometry using a linear regression algorithm. Anal Chem 79, 263-272 (2007)
10. Niemela, P.S., Castillo, S., Sysi-Aho, M., Oresic, M.: Bioinformatics and computational methods for lipidomics. J Chromatogr B Analyt Technol Biomed Life Sci 877, 2855-2862 (2009)
11. Andrej, S., Kai, S.: Lipidomics: coming to grips with lipid diversity. Nature Reviews 11, 6 (2010)
12. Fahy, E., Subramarniam, S., Brown, H.A., Glass, C.K., Merrill Jr, A.H., Murphy, R.C., Raetz, C.R.H., Russell, D.W., Seyama, Y., Shaw, W., Shimizu, T., Spener, F., van Meer, G., VanNieuwenhze, M.S., White, S.H., Witztum, J.L., Dennis, E.A.: A comprehensive classification system for lipids. Journal of Lipid Research 46, (2005)
13. Sud, M., Fahy, E., Cotter, D., Brown, A., Dennis, E.A., Glass, C.K., Merrill Jr, a.H., Murphy, R.C., R.H., R.C., Russell, D.W., Subramaniam, S.: LMSD: LIPID MAPS structure database. Nucleic Acids Research 35, (2006)
14. Blanchard, A.P., McDowell, G.S., Valenzuela, N., Xu, H., Gelbard, S., Bertrand, M., Slater, G.W., Figeys, D., Fai, S., Bennett, S.A.L.: Visualization and Phospholipid Identification (VaLID): online integrated search engine capable of identifying and visualizing glycerophospholipids with given mass. Bioinformatics 29, 284-285 (2013)

15. De Laeter, J.R., Böhlke, J.K., De Bièvre, P., H., H., H.S., P., K.J.R., R., P.D.P, T.: Atomic Weights of the Elements: Review 2000. Pure Applied Chemistry 75, 683-800 (2003)
16. McDowell, G.S.V., P Blanchard, A., Taylor, G.P., Figeys, D., Fai, S., Bennett, S.A.L.: Predicting glycerophosphoinositol identities in lipidomic datasets using VaLID (Visualization and Phospholipid Identification) – an online bioinformatic search engine. submitted (2013)
17. Fahy, E., Sud, M., Cotter, D., Subramaniam, S.: LIPID MAPS online tools for lipid research. Nucleic Acids Research 35, W606-W612 (2007)