Novel concepts for lipid identification from shotgun mass spectra using a customized query language

Dissertation

to receive the academic degree / zur Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat)

submitted to / vorgelegt an der
Dresden University of Technology
Department of computer Science /
Technische Universität Dresden
Fakultät Informatik

by / eingereicht von

Dipl.-Inf. Ronny Herzog
born / geboren am 12. September 1980 in Gera

Supervisor/ Betreuer:
Prof. Dr. Michael Schroeder, TU Dresden, Biotec und Fakultät für Informatik

Abgabe: 06.03.2012

I dedicate this work to my parents and to my daughter, Helena, who will change my life forever.

Acknowledgments

My greatest thanks go to my supervisor, Dr. Andrej Shevchenko. I thank him for his trust and motivation through all the years. A great help were his inspiring discussions and his clear, down-to-earth thinking which often put my thoughts from a nebulous realm into the reality.

I thank Michael Schroeder for being my supervisor. He made it possible for me to start my phd in his research group. I thank him for his support, the very interesting and helpful discussions and for putting the project some times in the right direction again. His clear thinking helped me often to get my thoughts into an understandable order.

I thank Dr. Dominik Schwudke who had the starting ideas for this project which was built upon his excellent former work. He helped me to get into the world of mass spectrometry, lipids and chemistry and was always the main resource of knowledge for this work. From his example I learned what it means to think analytical and work with great accuracy and awareness.

I also would like to thank Kai Schuhmann for his valuable discussions and input. He was one of the first users of LipidXplorer. He could add valuable ideas to the software and was a patient software tester. It was a pleasure to work with him. The same holds for Dr. Julio Sampaio as he was also one of the first software testers and a great partner for valuable discussions.

A special thank goes to Andrej Vasilj. Beside the very valuable discussions and the great fun I had, I thank him for proof reading my thesis.

Last but not least, I would like to thank every single present and former member of the Shevchenko laboratory: Dr. Magno Junqueria, Dr. Marc Gentzel, Dr. Christer Ejsing, Oksana Lavrynenko, Andrea Knaust, Henrik Thomas, Dr. Anna Shevchenko, Dr. Michael Groessl, Dr. Cyrus Papan, Dr. Susanne Sales, Dr. Maria Carvalho, Karem Gallardo, Dr. Christer Ejsing, Dr. Christian Klose, Dr. Marina Edelson-Averbukh, Dr. Natalie Wielsch and Virginia Maria Ferreira Resende. I had a joyful time, intriguing discussions and lots of fun.

A special thank is for Nicole, who made me a better person since I got to know her. She helped me a lot during my thesis by just being with me and sharing my bad days as well as my good days. But I thank her also very much for reading parts of my thesis and giving it a view from a non-expert. Thank you!

Abstract

Lipids are the main component of semipermeable cell membranes and linked to several important physiological processes. The research field of lipidomics aims at the quantitative molecular characterization of all lipid classes and species of cells, tissues or whole organisms. It was - and is - highly accelerated by mass spectrometry which is the method of choice for detailed structural and quantitative lipid analysis. The direct infusion of total lipid extracts into the mass spectrometer is called shotgun lipidomics and allows a fast spectra acquisition and direct quantification of many lipid classes in parallel but challenges the analysis by having a high density of information contained in the spectrum. Modern mass spectrometers improve the analysis of directly infused spectra with a high resolution, a high acquisition speed and a high accuracy. To tap the full potential of those features for an accurate lipid identification and quantification, it needs customized software, especially if biological experiments comprise numerous samples. However, the currently existing tools are narrowed only for certain mass spectrometers and only for a limited set of lipids. Software offering the identification of individual lipid species often relies on databases of MS/MS spectra. This is usually counterproductive since lipid fragmentation pathways strongly depend on both, the type of the mass spectrometer and the experiment settings.

The focal point of this thesis is the development of effective algorithms to identify lipids on any high and low resolution mass spectrometer in a high throughput manner. The main focus was laid on the effective storing and structuring of spectra data from numerous samples comprising biological experiments and a flexible and customizable lipid identification approach with which a high variety of lipid classes and species can be identified. Spectra were aligned and stored in a database called MastersScan which is the starting point for the high throughput lipid identification and quantification. The Molecular Fragmentation Query Language (MFQL) was developed to probe the MasterScan in order to identify lipid classes and lipid species. It makes use of the immanent combinatorial structure of lipids and furthermore supports their differing fragmentation pathways from spectra acquired with various different mass spectrometers. It supports different acquisition modes as well as their

differing features of resolution and accuracy. The correct quantification of lipids is supported by a comprehensive isotopic correction algorithm.

The thesis contains several experiments to validate and verify the underlying algorithms and the isotopic correction. A benchmark compares LipidXplorer with other lipidomics tools. It is shown that it outperforms the other in the conclusion of the benchmark.

Publications

1. A novel informatics concept for high throughput shotgun lipidomics based on the molecular fragmentation query language.
   Ronny Herzog, Dominik Schwudke, Kai Schuhmann, Julio Sampaio, Stefan R. Bornstein, Michael Schroeder and Andrej Shevchenko
   Genome Biology, 2011; 12(1): R8
2. LipidXplorer: a software for consensual cross-platform lipidomics
   Ronny Herzog, Kai Schuhmann, Dominik Schwudke, Julio Sampaio, Stefan R. Bornstein, Michael Schroeder, Andrej Shevchenko
   PloS One, accepted 2011, article in press
3. Shotgun lipidomics by tandem mass spectrometry under data-dependent acquisition control
   Dominik Schwudke, Gerhard Liebisch, **Ronny Herzog**, Gerd Schmitz and Andrej Shevchenko
   Methods in Enzymology, 2007; Vol. 433; 175-91
4. Top-Down lipidomics reveals ether lipid deficiency in blood plasma of hypertensive patients
   Jürgen Grässler, Dominik Schwudke, Peter Schwarz, **Ronny Herzog**, Andrej Shevchenko and Stefan R. Bornstein
   PLoS One, 2009; 4(7)
5. Shotgun lipidomics on high resolution mass spectrometers
   Dominik Schwudke, Kai Schuhmann, **Ronny Herzog**, Stefan R. Bornstein and Andrej Shevchenko
   CSH Perspectives in Biology, 2011´; Sep 1; 3(9)
6. Bottom-Up shotgun lipidomics by higher energy collisional dissociation on LTQ Orbitrap mass spectrometers
   Kai Schuhmann, **Ronny Herzog**, Dominik Schwudke, Wolfgang Metelmann-Strupat, Stefan R. Bornstein and Andrej Shevchenko
   Analytical Chemistry, 83(14), 2011
7. Shotgun lipidomics on a LTQ Orbitrap mass spectrometer by successive switching between acquisition polarity modes
   Kai Schuhmann, Reinaldo Almeida, Mark Baumert, **Ronny Herzog**, Stefan R. Bornstein

and Andrej Shevchenko
Journal of Mass Spectrometry, accepted Nov. 2011, article in press

Paper 1 describes the concepts and algorithms of the software LipidXplorer for lipid identification from mass spectra, which is the main work underlying this thesis. It contains furthermore a benchmark between LipidXplorer and other lipid identification tools and a discussion about its flexibility in lipid identification with the help of a practical example. My contribution to paper 1 are the concepts and implementations of the molecular fragmentation query language (comprised in Chapter 0), the MasterScan (comprised in Chapter 4), the underlying algorithms (comprised in Chapter 3), their verification (comprised in Chapter 6.1), the verification of the isotopic correction algorithms (comprised in Chapter 6.3), the concept and results of the benchmark between different lipid identification tools (comprised in Chapter 6.4) and the comparison of LipidXplorer running on different mass spectrometers (comprised in Chapter 6.7) and with different mass spectra resolution (comprised in Chapter 6.6).
Paper 2 describes a further development with which LipidXplorer is able to handle different types of spectral acquisitions. My contributions to paper 2 are the concept and implementation of the required algorithms and the comparison of lipid identification results between different acquisition types (comprised in Chapter 6.8).
Paper 3 and 5 describe the concept of shotgun lipidomics by tandem mass spectrometry under data-dependent acquisition. The identification of lipids is described by using the concepts and implementation of LipidXplorer.
Paper 4 uses LipidXplorer for the high throughput lipid analysis of a large data set comprising over 400 samples.
LipidXplorer was used for paper 6 where its lipid identification routines are used to identify fragmentation specific for higher energy collisional dissociation and for using background peaks for recalibration.
My contribution to paper 7 was a modification on LipidXplorer to read out specific information from mass spectra.

Table of contents

Table of figures

Figures Appendix:

Table of tables

1 Introduction

Lipids have been loosely defined as biological substances that are generally hydrophobic in nature and in many cases soluble in organic solvents (Smith, 2000). These chemical properties cover a broad range of molecules, such as fatty acids, phospholipids, sterols, sphingolipids, terpenes and others (Christie, 2003). Lipids are the main component of semipermeable cell membranes. The physical and chemical properties of the membranes directly affect cellular processes, making the role of the lipids dynamic rather than just a simple inert barrier (Shevchenko and Simons, 2010). Phospholipids, for example, are linked to important physiological processes such as bioenergetics, signal transduction across the cell membrane, cellular recognition and generation of signaling lipids (Ivanova, et al., 2007). Thus, understanding the lipid composition of cells may aid in the elucidation of cellular functions and may help to get insight towards the symptoms and treatment of various diseases.

The emerging scientific field of lipidomics aims at the quantitative molecular characterization of all lipid classes and species of cells, tissues or whole organisms. It faces a highly heterogeneous array of individual lipid species that collectively give rise to many tens of thousands of lipids that constitute each cell's lipidome (Han, et al., 2011). Many technologies (including mass spectrometry (MS), nuclear magnetic resonance spectroscopy, fluorescence spectroscopy, column chromatography, and microfluidic devices) have been used in lipidomics to identify, quantify and understand the structure and function of lipids in biological systems (Feng and Prestwich, 2005). Of those, mass spectrometry is by far the most used, because its technical characteristics go well with the identification of lipids and has furthermore a high sensitivity, an easy quantifiability and allows retrieving a high amount of information from molecules.

The progress of mass spectrometric driven lipidomics was highly accelerated by the breakthrough work of Karas and Hillenkamp (Karas, et al., 1985) and the noble-prize winning research of Fenn and Tanaka with the development of so-called soft-ionization techniques of electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI). These methods enable the ionization of molecules without unwanted fragmentation, increasing the accuracy, sensitivity and reproducibility of quantification from complex solutions. Another

important step was the appearance of two major mass spectrometry approaches which are nowadays routinely used for lipidomics investigations. One is the direct infusion of total lipid extracts into the mass spectrometer, so-called shotgun lipidomics, and the other is the coupling of a chromatographic separation device. The former's advantages are its high speed acquisition of spectra and simple lipid quantification while the latter greatly enhances the spectra's signal-to-noise ratio by fractionating the analyte and allows furthermore separation of lipid classes.

Depending on the architecture of the mass spectrometer, spectra are acquired with different acquisition methods. Greatly driven by the work of Han and Gross (Gross and Han, 2009; Han and Gross, 2003; Han and Gross, 2005; Han and Gross, 1994; Han and Gross, 2001; Han and Gross, 2003; Han and Gross, 2005; Han, et al., 2011; Han, et al., 2006), the so-called precursor-ion and neutral loss scanning (PIS and NLS) on triple quadrupole or triple quadrupole–linear ion trap (QTRAP) mass spectrometers became the method of choice for lipidomics over years. A more recently used acquisition method for lipidomics was the method of data-dependent acquisition (DDA). The DDA approach allows the generation of a comprehensive set of spectra from a single sample. Coupled with bioinformatics tools it was possible to simulate PIS and NLS on DDA acquired spectra and additionally introduce the new scanning method of Boolean scans which facilitates the detection of complex fragmenting lipid species (Schwudke, et al., 2007).

The introduction of the hybrid linear trap quadrupole (LTQ) Orbitrap mass spectrometer which incorporates high spectra resolution with a high acquisition speed accelerated the progress of shotgun lipidomics due to its advanced analytical possibilities. Many lipid species which were overlapping in low resolved spectra could now be distinguished clearly due to the high resolution capacity of the Orbitrap, making the use of PIS and NLS unnecessary and the acquisition speed very fast.

The work of this thesis carries forward the computer aided interpretation of spectra acquired with DDA applying a new concept and new algorithms to most efficiently use the information of mass spectra, especially by incorporation of high resolution measurements enabled by modern mass spectrometers like the LTQ Orbitrap. By doing so, it addresses current problems which are present in the

area of computational interpretation of shotgun lipidomics data sets:

1) Over the years numerous manufacturers of mass spectrometers developed their own, proprietary and non-interchangeable spectra data formats. This makes it difficult or even impossible to have a single tool cooperating with all the different mass spectrometers. This problem has been present in the field of mass spectrometry for quite some years and has led to the development of generic mass spectra file formats (Martens, et al., 2011; Pedrioli, et al., 2004). However, there is currently no lipidomics software supporting these open formats. Furthermore, different mass spectrometers have different characteristics which influence the resolution and the accuracy of the acquired spectra and therefore the lipid identification algorithm. However, most of the present tools are customized to a certain mass spectrometer and are not interchangeable. This is especially the case for different types of acquisitions, like precursor ion / neutral loss scan and data-dependent acquisition which have their own spectra format, respectively.
2) Biological experiments normally incorporate a number of different assays of an organism which may encompass different stages of development, mutants or species with differing attributes. Even in middle sized experiments this results in tens to hundreds of samples. These numerous samples result in thousands of mass spectra which routinely need to be pre-processed, structured and probed. Thus, there is a need for a feasible and smart batch processing routine.
3) The identification of molecular species from mass spectra requires the recognition of species specific signatures in the mass spectra. Most lipids are assembled in structural building blocks. Thus, they can be identified with customized and efficient algorithms which differ from identification procedures of other molecule classes. Nevertheless, lipids are structural diverse and produce various signatures. These specific signatures vary furthermore in dependency of the mass spectra resolution, accuracy, the chosen acquisition mode and other settings of the acquisition. Thus, their identification can only be based on a miscellaneous and flexible spectra interpretation routine. Another problem which is inherent in shotgun

lipidomics is that lipid species can be isobaric, i.e. they overlap with each other. This can lead easily to false positive and incomplete annotations. Current software falls short to handle this diversity. Most tools are specialized to certain mass spectrometer platforms and are not interchangeable.

4) An important issue is the validation of the identified lipid species. No matter which method was used, there is a probability that it produces false positive results. This holds especially in shotgun lipidomics, where spectra are richly filled with lipid specific information and lipids easily produce isobaric peaks. Unfortunately, no method exists currently for evaluating identified lipid species. One reason is that lipids have only a small number of specific mass spectra signals. Such, it is not possible to derive statistical significant probabilities of correct identified peaks. Therefore, it is of great importance to offer aid for the validation of the results.
5) A correct quantification of the identified species is necessary for any meaningful outcome of a biological experiment. The real amount of molecular species is distributed among its isotopes. This has to be considered for a correct quantification. The main problem which occurs concerning the quantification of lipids in shotgun lipidomics is the overlap of isotopes with other lipid species. It is therefore necessary to identify such overlaps and subtract them from the result which can be a complex task, especially for the isotopic subtraction of fragmented ions.

Even middle sized biological studies are not feasible anymore if there is no appropriate aid with automation. As it will be shown in this thesis, the existing software for shotgun lipidomics does not perform optimally and is still an underdeveloped area. Especially if it comes to experiments which need high throughput interpretation, no suitable software is available. Additionally to the problems mentioned before, current software tools usually target a selection of lipid classes that reflect the scientific interest of the developer’s team. It may not recognize uncommon lipid classes or species – for example, those comprising some unconventional fatty acid moieties – or may perform suboptimal.

1.1 The aim of the presented study

As stated above, there is an urgent need to develop new algorithms and techniques to overcome the limitations of the current software for interpretation and quantification of lipids from mass spectra in the field of shotgun lipidomics.
According to the points given in the introduction, the main conceptional highlights should encompass:

1) The support of spectra acquired on any mass spectrometer with any acquisition method.
2) The support of large scale lipidomics by a fast processing of numerous samples.
3) Customizable and flexible spectra interpretation routines, addressing lipids in their complex diversity along with complex shotgun mass spectra.
4) Transparency of the spectra handling and lipid identification to support verification of the result.
5) A comprehensive and accurate isotopic correction method for a correct quantification of lipid species.

The developed algorithms should be validated by several methods. This should be underlined by presenting examples of the usage of the software on different conditions and mass spectrometers.
The findings should be implemented in a software toolkit which should have an intuitive user interface so as to be usable by a broad audience of biologists and chemists working in the field of lipidomics. The overall correctness of the developed software should be validated by comparing it against other existing tools.

1.2 The main contributions of this thesis

The findings of (Schwudke, et al., 2006) and (Schwudke, et al., 2007) mark the starting point for the development of the software underlying this thesis, called LipidXplorer. The thesis introduces two new concepts: The first is the concept of the MasterScan, which is an experimental database incorporating all spectra acquired in a biological experiment. The second is the concept and development of a customized query language which enables probing the MasterScan for lipids. It is called molecular fragmentation query language (MFQL) and is the first query

language used for the identification and quantification of lipids. Both approaches together with the developed algorithms aim to solve all problems stated above.

1.3 The composition of the thesis

The first part of the thesis contains background information to build a basis for understanding the subject and can be found in Chapter 2. It is divided into two parts: the first part introduces basic biological and mass spectrometric concepts concerning lipids (Chapter 2.1 to Chapter 2.3) and the second part deals with available software tools for lipid identification with mass spectrometry and their pros and cons (Chapter 2.6 and Chapter 2.7).

The software LipidXplorer comprises a data base containing spectra of whole biological experiments. It is called MasterScan and is one of its main parts. Before the MasterScan can be created, single scans (presented in Chapter 2.3) must be averaged to consensus mass spectra. Then, the resulting spectra are aligned and can be stored. Several spectra alignment algorithms are presented in Chapter 3.1. The algorithms developed for LipidXplorer can be found in Chapter 3.2. Chapter 4 describes the implementation of the MasterScan.

The other main result of the work underlying this thesis is the development of MFQL. It is customized to probe the MasterScan and supports any lipid identification routine in an intuitive, transparent and user-friendly manner independently of the instrumentation platform. An introduction in its design can be found in Chapter 5.3 together with examples of its usage in Chapters 5.4 and 5.5. In Chapter 6.1 the implemented scan averaging and spectra alignment algorithms are validated in separate tests. Chapter 6.2 introduces the chemical entity isotopes and the necessity to do a so-called isotopic correction on the resulting lipid profile in order to have correctly quantified lipid species. The developed algorithms are validated with an experiment in Chapter 6.3. In Chapter 6.4 LipidXplorer is benchmarked against other lipidomics tools. This required the building of a lipid reference list from a complex lipid mixture. The processing speed is benchmarked in Chapter 6.5.

LipidXplorer is customizable to handle spectra acquired on different mass spectrometers with differing resolution, accuracy and different acquisition modes. To demonstrate LipidXplorer’s cross platform capabilities, it is shown working on

spectra differing in resolution in Chapter 6.6. In Chapter 6.7 LipidXplorer is used for spectra acquired on different mass spectrometers. Chapter 6.8 presents LipidXplorer's capability of using spectra acquired in different modes.

The usage of a query language for lipid identification allows a great flexibility. To depict this, in Chapter 6.9 it is used for the identification of a number of different lipid classes in parallel on a real sample.

Chapter 7 marks the end of the thesis with conclusions containing further examples in which LipidXplorer was successfully used for biological studies (Chapter 7.2). Its public availability (Chapter 6.10) is presented as well as an outlook on possible future directions of the project (Chapter 7.3).

2 Background

2.1 Lipids

Lipids constitute a broad group of naturally occurring molecules which include fats, waxes, sterols, fat-soluble vitamins, monoglycerides, diglycerides, phospholipids and others which have a low solubility in water but are soluble in nonpolar organic solvents (Harkewicz and Dennis, 2011). The lipids amphiphilic (Molecules possessing both hydrophilic and hydrophobic properties) nature allows them to form structures such as vesicles, liposomes or membranes in an aqueous environment. Lipids function mainly as energy storage, as structural components of cell membranes and as signaling molecules (Wenk, 2005). Especially because of the latter two functions, lipids are an important factor for a great number of cellular processes. This makes the study of lipids an important source for gathering insight into symptoms and treatment of various diseases. Recent research shows that lipids play a vital role in the brain, in Alzheimer disease (Sagin and Sozmen, 2008), in cancer (Denizot, et al., 2001), inflammation (Tselepis and John Chapman, 2002) or in the fertility of male primates (Roudebush, et al., 2005) and others.

The totality of lipids in cells, tissues or organisms is called the “lipidome”. Eukaryotic lipidomes comprise over a hundred lipid classes, each of which is represented by a large number of individual yet structurally related molecules. According to different estimates an eukaryotic lipidome might contain from 9,000 to 100,000 individual molecular lipid species in total (van Meer, 2005; Yetukuri, et al., 2008).

These great numbers originate from the structural diversity of lipids. The lipid repertoire can be comprehensively classified in eight categories (Fahy, et al., 2005):

- Fatty acyls
- Glycerolipids
- Glycerophospholipids
- Sphingolipids
- Sterol Lipids
- Prenol Lipids

- Saccharolipids
- Polyketides

each of which contain several lipid classes which on their part contain numerous lipid species.

The molecular structure of lipids is a combination of several structural building blocks. Glycerophospholipids, for example, have a common denominator: the glycerol molecule or glycerol backbone. Located at its sn-3 position is the polar head group which determines the lipid class. The other two hydroxyl groups are acylated with different long-chain fatty acids (or “aliphatic chains”; or “acyl chains”), which depending on their length and number of double bonds produce numerous lipid species. Glycerophospholipid classes could be for example Phosphatidylethanolamine (PE), Phosphatidylcholine (PC), Phosphatidylinositol (PI), Phosphatic acid (PA), Phosphatidylserine (PS) or Phosphatidylglycerol (PG) (see Figure 1).

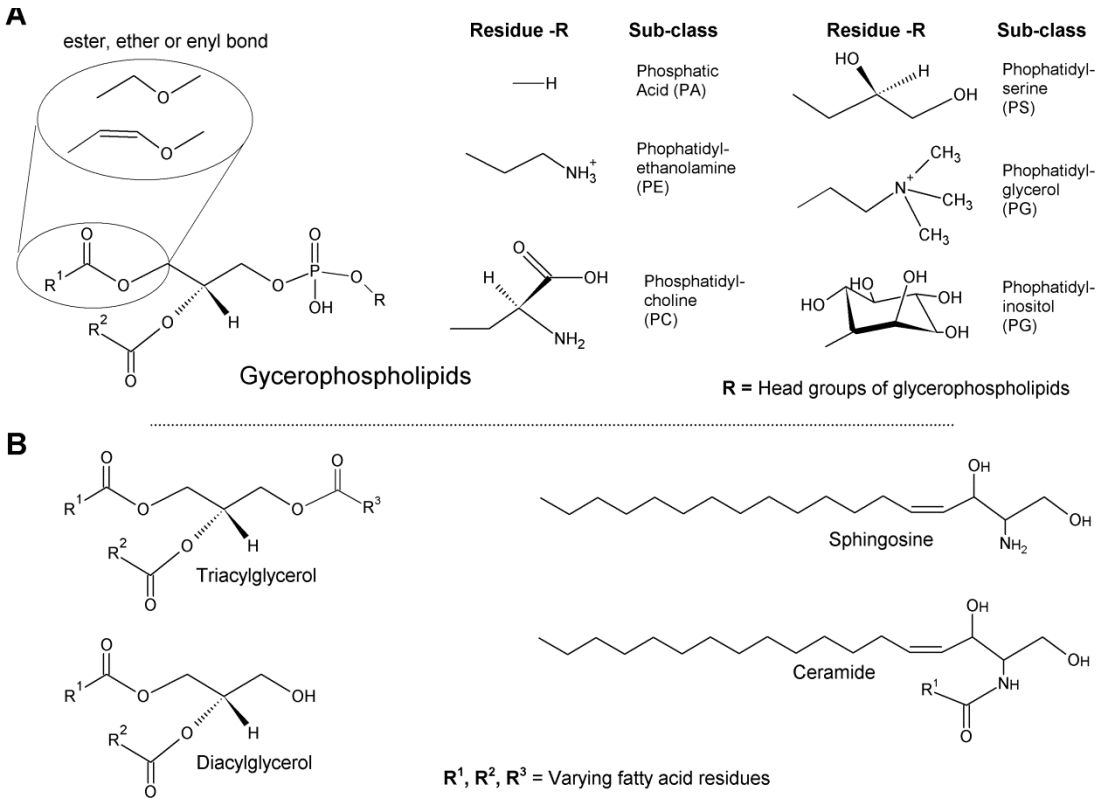


Figure 1: The complex diversity of lipid species shown on the example of Glycerophospholipids.

A: The structural composition of glycerophospholipids consists of different building blocks: the glycerol backbone with the fatty acyl chains (on the left) and the different lipid class specific head groups (on the right). Considering the presented head groups and assuming a diversity of fatty acids with a length of 12 to 22 carbons and zero to eight double bonds each, there exists theoretically 2880 different glycerophospholipids.

B: A selection of other major lipid classes is shown. This includes triacylglycerol (TAG), diacylglycerol (DAG), sphingosine and ceramide. (Source: according to (Watson, 2006))

The quantitative molecular characterization of the full lipidome (i.e. all lipids) of cells, tissues or whole organisms is called lipidomics and is an emerging scientific discipline (reviewed in (Dennis, 2009; Oresic, et al., 2008; van Meer, 2005; Wenk, 2005)).

Due to enormous compositional complexity and diversity of physicochemical properties of individual lipid molecules, lipidomic analyses rely heavily on mass spectrometry.

2.2 Fundamentals of Mass Spectrometry

At its very basics, a mass spectrometer separates and detects ions according to

their mass-to-charge ratio (*m/z*). Additional structural information can be provided by fragmenting intact ions with collision-induced dissociation (CID), which is called tandem MS or MS/MS. In the following, a short summary on common mass spectrometric instrumentation is given. A comprehensive survey can be found in (Gross, 2004).

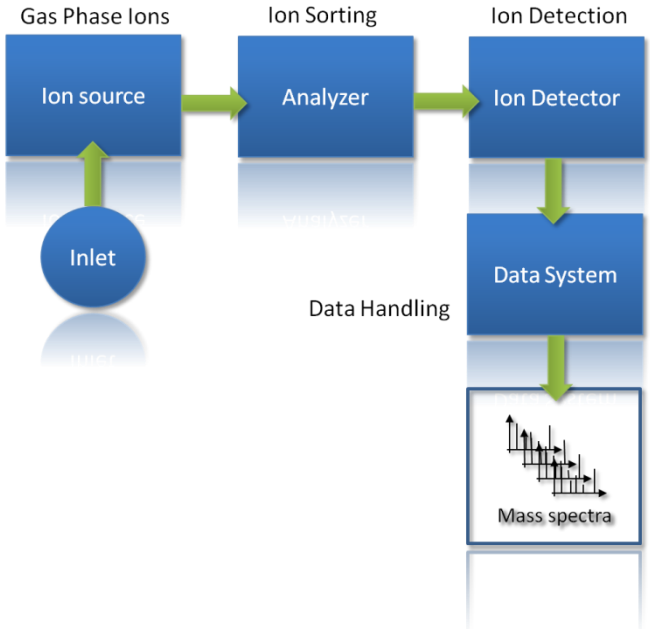


Figure 2: Schematic representation of the functional parts of a mass spectrometer.

The inlet transfers the sample into the vacuum of the mass spectrometer. In the Ion Source, neutral molecules are ionized and accelerated into the mass analyzer. The analyzer separates the ions by their mass to charge values (m/z) before the Ion Detector records their abundances. The signal is processed in the Data System which transforms it into a spectrum of m/z values (x-axis) vs. their individual intensities (y-axis) (Source: own design).

Typically, a mass spectrometer consists of three major parts: (i) an ion source, (ii) one or more mass analyzers that measure the m/z ratio of the ionized analytes and (iii) a detector that records the ion signal corresponding to each m/z value (Sampaio 2010) (see Figure 2).

2.2.1 The ionization of the analyte

Following the sample introduction, the analyte is ionized in the ion source, either operating at atmospheric or vacuum pressures. The generation of charged molecules is necessary to enable ion manipulation based on its mass-to-charge ratio. The oldest ionization method is that of electron impact (EI) and relies on the

analyte being in gas-phase. This restricted mass spectrometry to small and thermostable molecules due to the lack of proper techniques to “softly” ionize and transfer ionized molecules from the condensed phase to the gas phase without excessive fragmentation. In EI, a beam of electrons passes through the gas-phased sample knocking off other electrons if they collide with the neutral analyte. This produces positively charged ions in form of a radial cation (M+). However, the situation changed in the late 1980s and was crucial for the progress of modern mass spectrometric driven lipidomics. The groundbreaking work of Michael Karas and Franz Hillenkamp and the noble-price winning research of John B. Fenn and Koichi Tanaka with the development of so-called soft-ionization techniques of electrospray ionization (ESI) (Fenn, et al., 1989; Whitehouse, et al., 1985) and matrix-assisted laser desorption/ionization (MALDI) (Karas, et al., 1985; Karas and Hillenkamp, 1988) introduced two techniques for the routine and general accumulation of molecular ions of entire biomolecules. Those prevent molecules to fragment in the ion source and produce adduct ions (MH+) of intact molecules.

2.2.2 The analysis of the ions

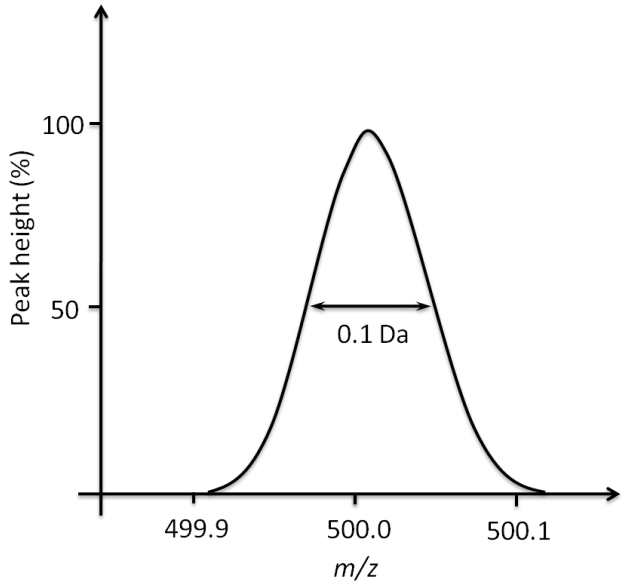
After ions are formed in the ion source, they are accelerated into the mass analyzer by an electric field. The mass analyzer separates them according to their mass-to-charge value and transports them further to the detector which records the abundance of the separated ions. The mass analyzers in a mass spectrometer can be of different types varying in physical principles and performance standards. The simplest and most frequently used mass analyzers are time-of-flight (TOF) analyzers (Chernushevich, et al., 2001), quadrupoles (Yost and Enke, 1978) and ion traps (IT) (Paul and Steinwedel, 1953). In a TOF analyzer, the ionized analyte is accelerated by an electric field of known strength (Schultz, 1946). A detector at a known distance records the time the particle needs to reach it. Since the velocity of the ion depends on its mass-to-charge ratio, one can calculate the m/z value back from the time it needs to reach the detector and the known experimental settings. Quadrupole analyzers consist of four circular rods, set parallel to each other and connected electrically. A radio frequency voltage is applied between one pair of rods and the other. Ions enter

the quadrupole and are attracted and repulsed by the periodically changing forces of the rods, because the sign of the electronic forces also change periodically. Only ions which have a stable trajectory under a given ratio of voltages pass through the quadrupoles while others collide with the rods. The radius of the trajectories depends on the m/z values of the ions. Therefore, only ions within a particular, very confined m/z range for a specific strength and frequency of applied voltages will reach the detector, whereas others will be deflected. Ion traps work by the same principles as quadrupole but by capturing the ions in a confined space like a vacuum system or tube. To prevent ions from escaping via either open end, they are trapped by electrodes of a slightly higher potential which are placed adjacent to the front and rear ends of the multipole. The principle is the same as for quadrupoles, the ions move on specific trajectories in an oscillating electric field. Such devices are known as linear ion traps (LIT). By changing the electric potential of the entrance and exit electrodes, ions of a particular *m/z* range are trapped and released from the LIT. The ejected ions are subsequently recorded by the detector. The detector records the charge induced when an ion hits a surface. It is typically built from an electron multiplier or a faraday cup which is an electrode where the ions deposit their charge and an electric current is produced. The current depends on the quantity of ions hitting the detector and is used then as the intensity value for a particular *m/z* value, so generating the mass spectrum. The data system at the end of this pipeline converts this data into a spectrum of m/z values versus abundance (see Figure 5 for an example spectrum).

2.2.3 Mass accuracy and mass resolution

Mass spectra are acquired as continuous spectra in which peaks have two common attributes: mass accuracy and mass resolution. Both values express the experimental errors or uncertainties of *m/z* values. Mass accuracy is the ability to measure or calibrate the instrument response against a known entity. It is usually expressed as a relative error in parts-per-million (ppm), but sometimes also as absolute error in Dalton (Da), or milli-mass unit (mmu), where 1 mmu equals 0.001 Da. It indicates the deviation of the instruments response from a known mass. The resolution (*R*) measures the ability of a mass spectrometer to

differentiate one ion from another. The resolution *R* is calculated as $R = \frac{m}{\Delta m}$, where Δm is determined with either of two common ways corresponding to IUPAC (International Union of Pure and Applied Chemistry): the peak valley definition and the peak width definition. The valley definition defines Δm as the closest spacing of two peaks of equal intensity with the valley (the lowest value of signal) between them less than a specified fraction of the peak height. Typical values are 10% or 50%. The peak width definition Δm is defined as the width of the peak measured at a specified fraction of the peak height, for example 0.5%, 5%, 10% or 50%. It is widely accepted to use the width at half-height which is called full-width half-height maximum (FWHM) for the degree of response of an instrument to a given ion (see Figure 3).



Accuracy:

True mass	= 500.0000
Measured mass	= 500.0020
Difference	= 0.002
Error	$= \frac{0.002}{500} \times 10^6 = 4\,ppm$

Resolution:

Mass	= 500.0000
Peak width (at 50%)	= 0.1
Resolution (FWHM)	$= \frac{500}{0.1} = 5000$

Figure 3: The definition and calculation accuracy and the resolution.

Shown is the determination of the mass accuracy and the full-width half-height maximum (FWHM) method for determining resolution for a mass spectrometer measured at a given ion. (Source: according to (Balogh, 2004))

2.2.4 High accuracy mass spectrometers

A great breakthrough in the field of mass spectrometry was the introduction of Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometers (Comisarow and Marshall, 1974). FT-ICR offered a so far unseen ultrahigh resolving power with $R \approx 10^5$ to $5 \cdot 10^6$ and highest mass accuracy with $\Delta m \approx 10^{-4}$ to 10^{-3} Da, attomol detection limits with soft ionization methods and a high mass range. The determination of *m/z* is based on the cyclotron frequency (i.e.

the movement) of the ion in a fixed magnetic field (Marshall, et al., 1998). The ions are trapped in a Penning trap – a magnetic field with electric trapping plates – where they are excited by an oscillating electric field perpendicular to the magnetic field. This causes the ions to circle in larger cyclotron radius. The frequency depends on their *m/z* value. Thus, the *m/z* value can be calculated back by Fourier transforming the recorded frequency. The drawback of FT-ICR mass spectrometers is that the high resolution and high accuracy come with high costs and space requirements for the super-conducting magnet.

The recently introduced Orbitrap mass analyzer (Makarov, et al., 2006; Makarov, et al., 2006) operates without a magnetic field and is so available at a much lower price and installation room requirement. However, it is nevertheless able to compete with the resolution and accuracy of a FT-ICR analyzer. As with FT-ICR, the recording of the ion oscillation and subsequent conversion into the frequency domain using Fourier transformation is used in the Orbitrap to obtain m/z values. But instead of a magnet an electrostatic field is used for trapping the ions. The electrostatic field is created in an Orbitrap consisting of an outer and inner coaxial electrode. The inner spindle shaped electrode confines the ions so that they both orbit around the central electrode and oscillate back and forth along the central electrode's long axis. The frequency of this oscillation depends on the m/z value of the ions and uses it to calculate the mass spectra.

To increase the power of modern mass spectrometers different types of mass analyzers and ion-guiding devices are constructed in single, so-called hybrid instruments. These enable a better support to various experimental setups and are furthermore often more affordable instruments. A popular example of a hybrid instrument is the hybrid quadrupole time-of-flight (Q-TOF) mass spectrometer that combines a quadrupole and a TOF analyzer which allows higher resolution and hastens the analysis. This makes it better suited for protein and proteome analysis (Aebersold and Mann, 2003).

2.2.5 The fragmentation of ions with tandem mass spectrometry

Tandem mass spectrometry (MS/MS) requires the utilization of multiple mass analyzers or of an ion-trap mass spectrometer. In an MS/MS experiment, the first analyzer is used to select a precursor ion that is typically accelerated to high

kinetic energy with an electrical potential to induce collisional heating and subsequent fragmentation with inert gases (this is often helium, nitrogen or argon) in the collision cell. This is called collision-induced dissociation (CID). The fragments of the precursor ion which carry an intrinsic charge are detected with the second mass analyzer. The resulting spectrum is called MS/MS spectrum or fragment spectrum (see Figure 5). Each fragment spectrum has a precursor or parent ion, which was prior selected for CID. The selection is within the isolation window which typically has a size of 1.0 Da but varies between different mass spectrometers. The narrower this window is the less neighbor ions are fragmented together with the chosen precursor, so the fragment spectrum is more specific. On the other hand, the narrower the isolation window is, the less sensitive is the fragment spectrum, since less precursor ions are selected.

2.2.6 Mass spectrometry in lipidomics

There exist two approaches for the mass spectrometry based identification and quantification of lipids within cells and tissues. The first approach includes a chromatographic separation before the injection into mass spectrometer, whereas the second involves the direct infusion of the total lipid extract into the mass spectrometer.

The first approach is based on separation of different lipid categories using chromatographic separation prior to the mass spectrometric analysis based on the differing affinity of the analyte towards the chosen mobile phase and stationary phase. The mobile phase can be liquid or gaseous while the stationary phase is usually a solid but may be an immobilized liquid. Commonly used chromatographic methods are high performance/pressure liquid chromatography (HPLC), gas chromatography (GC) and thin layer chromatography (TLC). A normal phase HPLC system for example separates lipids by means of the lipid class specific polar head group while a reverse phase HPLC system separates lipids by means of their fatty acids on the basis of their chain length, degree of unsaturation and substitution. Such, it is well suited to separate each of the lipid classes such as glycerophospholipids from glycerolipids (triacylglycerols (TAGs) or diacylglycerols (DAGs)). Due to the separation, different fractions of the sample are injected at different time stamps. This minimizes the ion suppression

and improves the sensitivity and therefore the ability to detect very low abundant lipid species. Nevertheless, it is to note that large concentrations of ions in the column eluate are necessary. The fractionation of the samples also limits the injection time for the individual lipid categories, thus decreasing the possible number of tandem mass spectrometric experiments and limits the analysis to a small number of lipids. This makes lipid quantification on a large scale problematic. The quantification of lipids also a complex task if chromatographic separation is used. Every single fraction of a chromatographic run has a different intensity and length in time while it eludes. The mass spectrometer knows the intensity at a certain time point, but not how many time points the fraction has and thus what the amount of the whole fraction is. Therefore, an accurate quantification is no possible. This problem can be solved by algorithms which identify fractions in the chromatographic run, calculate the amount of the fraction and relate it to the acquired mass spectrum. That this is a complex task to automate is mirrored in the number of different algorithms being developed for the processing of chromatographic runs (reviewed in (Castillo, et al., 2011)). Nonetheless, LC-MS / MS/MS platforms are broadly used in lipidomics (reviewed in (Han, et al., 2011)), especially for the targeted analysis of a small number of lipids (Liebisch, et al., 1999) or for applications of discovery and identification of novel lipids, particularly those which are present in low or very low abundance in a small scale (Kingsley and Marnett, 2009; Minkler and Hoppel, 2010; Tan, et al., 2006; Ziqiang, 2009).

The second approach is to directly infuse the total lipid extract into the mass spectrometer (see Figure 4 for a schematic comparison and Figure 6 for shotgun lipidomics pipeline), omitting a previous chromatographic separation, which is called shotgun mass spectrometry. It avoids difficulties from alterations in concentration, chromatographic anomalies, and ion-pairing alterations to improve the S/N ratio. It was originally described in 1994 by (Han and Gross, 1994). The numerous advantages of shotgun lipidomics include simplicity, efficiency, high sensitivity, ease of management and less expensive instrumental requirements which make it a popular choice for lipidomics. Probably the major advantage of shotgun lipidomics over a chromatographic pre-separation of lipids is that a mass spectrum of molecule ions of individual molecular species in a lipid class of

interest can be acquired at a constant concentration of the solution during direct infusion. This feature allows a simple, straight forward quantification and furthermore, virtually unlimited time to perform detailed tandem mass spectrometric experiments with multiple fragmentation strategies (Han, et al., 2011). However, the challenge of the direct infusion is the high amount of data contained in the spectrum because the sample is not distributed over the time (see Figure 4). Its whole content is contained in one spectrum which can lead to ion suppression and overlapping of peaks. Especially isotopes of molecular species tend to overlap with other molecular species which corrupts their real abundance or introduces unwanted peak shifts. That is why the resolution of a mass spectrometer is a key value for shotgun lipidomics. The higher a spectrum is resolved, the less negative side effects are produced by a direct infusion.

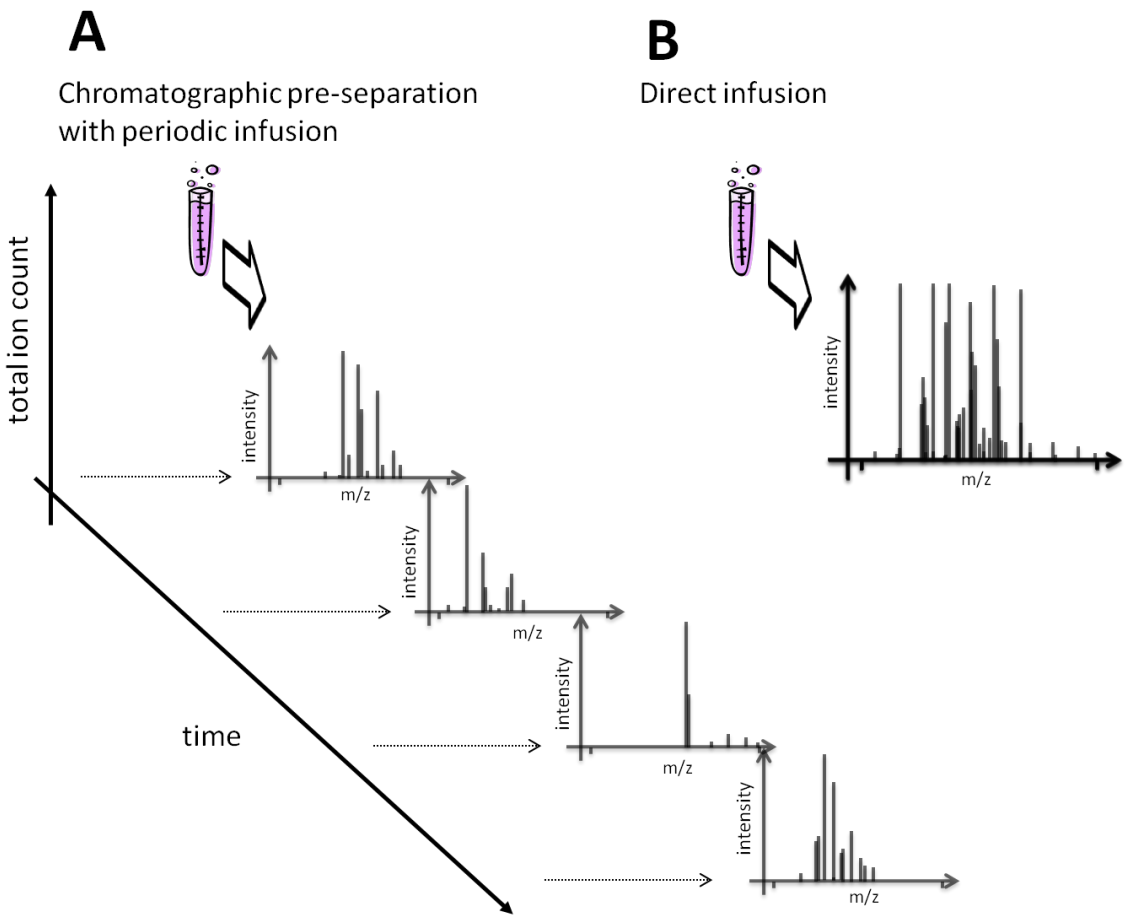


Figure 4: Comparison of chromatographic pre-separation and direct infusion.

The figure shows the major principles of mass spectrometric acquisition with a chromatographic pre-separation (A) and direct infusion (B). The chromatograph lets the analyte elute over time such that certain molecular categories elute at different time points. This is depicted with the profile along the time axis. At periodic time points, the analyte is infused into the mass spectrometer and spectra are acquired. In contrary, by direct infusion, the analyte is directly infused into the mass spectrometer (B). This results in a single spectrum containing all the molecular peaks of the analyte.

The state-of-the-art fragmentation methods for lipidomics were the detection of lipid species by precursor ion scan (PIS) and neutral loss scan (NLS) (Quehenberger, et al., 2010; Schmelzer, et al., 2007) (reviewed in (Pulfer and Murphy, 2003)) by triple quadrupole or triple quadrupole – linear ion trap (QTRAP) mass spectrometers (reviewed in (Blanksby and Mitchell, 2010; Gross and Han, 2011; Han and Gross, 2005)) where the instrument is set to detect one specific fragment originating from all precursor masses within a specified *m/z* range. In each analysis, the fragment mass (in case of precursor ion scanning) or the mass difference (in case of neutral loss scanning) is monitored and then the analysis is repeated for the next fragment mass / mass difference of interest.

Those scanning methods allow a simple identification of lipids. Usually lipids fragment into their polar head group, their naturally occurring fatty acyl chains (see Figure 5) or produce fragments as molecular adducts. The linear combinations of those lipid class and lipid species specific building blocks are used to reconstruct lipid species from PIS and NLS mass spectra (Brügger, et al., 1997). This made PIS and NLS the method of choice for lipid identification which was intensively promoted by the work of Han and Gross (Han and Gross, 2005; Han, et al., 2011), Murphy (Murphy, et al., 2001; Murphy and Axelsen, 2011), Hsu and Turk (Hsu and Turk, 2009) and Ejsing and Shevchenko (Ejsing, et al., 2006).

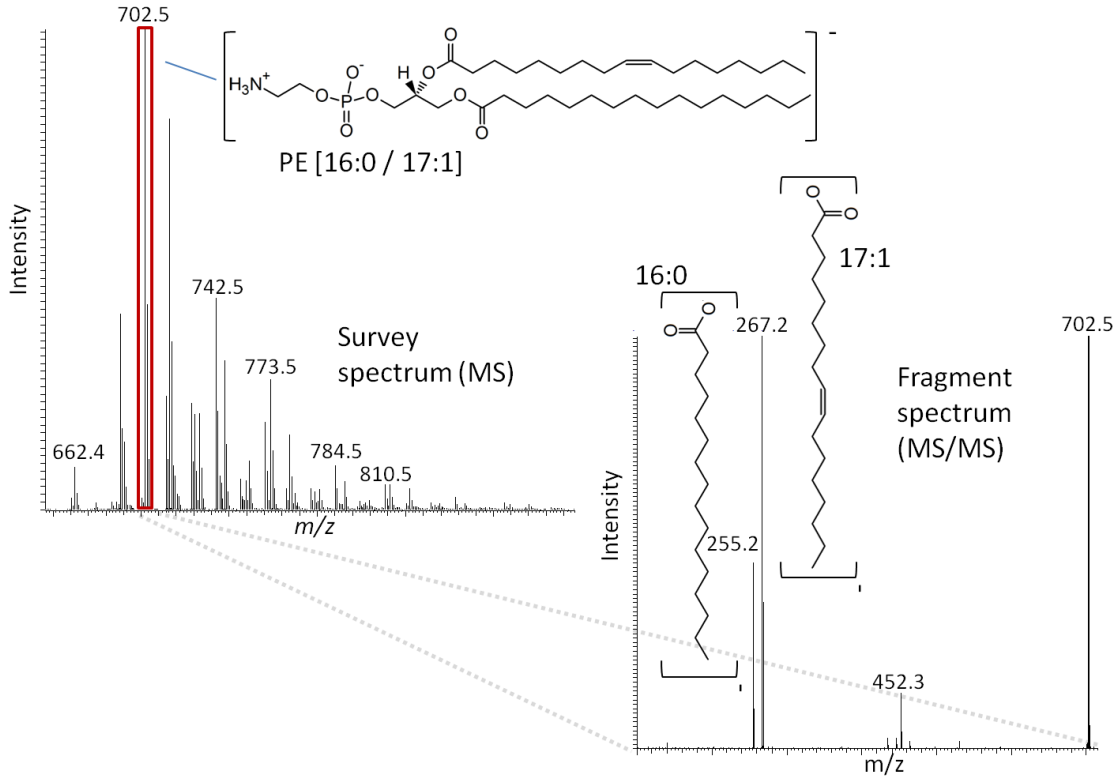


Figure 5: A tandem mass spectrum with annotation of a PE lipid species.

The figure shows a survey (or MS) spectrum on the left and a fragment (or MS/MS) spectrum acquired with collision-induced dissociation of the ion with *m/z* 702.5 on right. The x-axis is the *m/z* value and the y-axis the abundance of the ions. The peak with the *m/z* value 702.5 is the ion of a Phosphatidylethanolamine (PE) [16:0/17:1] lipid species. The species is defined by [<number of carbon atoms>:<number of double bonds>] of its fatty acyl chains which is 16:0 and 17:1 in this case. Both fatty acyl chains can be found in the spectrum having the *m/z* of 255.2 (16:0) and *m/z* 267.2 (17:1). The spectra are acquired on an LTQ Orbitrap Velos instrument in negative mode with a resolution of 7500 for MS and MS/MS, respectively. (Source: own design)

In precursor ion scanning on triple quadrupole instruments only a single precursor or neutral loss scan can be performed at a time and the analysis must

be repeated in order to profile another lipid class. However, while it is very sensitive and sufficiently specific, it is still an inherently low throughput approach. Moreover, the specificity of ion fragment selection is with a low resolution, since triple quad instruments usually do not allow more than a range of 1-2 Da (Dalton) for selection of the fragment. In contrast to the state-of-the-art PIS or NLS is the method of data-dependent acquisition (DDA) (Schwudke, et al., 2006). Here, all detectable precursor ions (or, alternatively, all plausible precursors from a pre-defined inclusion list) are fragmented beginning with the most intense signals. The longer an acquisition is running, the more MS/MS spectra are produced, until ultimately, a comprehensive data set comprising all fragment ions from all ionizable lipid precursors is obtained. In combination with bioinformatics tools, data-dependent acquisition was successfully used to emulate precursor ion scans and neutral loss scans. This enabled not only the simultaneous acquisition of theoretically unlimited precursor ion scans, as it is possible only with a Q(q)TOF instrument (Ekroos, et al., 2002), but it also allows the simultaneous simulation of a theoretically unlimited number of neutral loss scans. Furthermore, it was possible to do so-called Boolean scans which involve the combination of two or more scans combined by logical operations allowing the identification of lipid species which have a more complex fragmentation pathway.

A great step towards the accurate identification of lipid molecules was done by the introduction of Fourier transform ion cyclotron resonance mass spectrometers (short: FTMS) (Fridriksson, et al., 1999; Ivanova, et al., 2001; Jones, et al., 2006; Leavell and Leary, 2006; O'Connor, et al., 2002). Because of their high resolution and mass accuracy lipid peaks can be distinguished easily from chemical noise and in some instances FTMS enables the unequivocal compositional assignment of the resolved isobaric molecular species without recourse to MS/MS experiments (Ishida, et al., 2004). Despite its recognized potential, it comprises technical limitations which impair the accuracy of the isotopic profiles, thus compromising the species quantification and hampered structural validation of lipid precursors (Jones, et al., 2005; Schwudke, et al., 2007). Furthermore, FTMS platforms are not suitable for high-throughput lipidomics since the acquisition of samples is time and cost intensive.

Recently, hybrid mass spectrometers which combine Fourier transform or

Orbitrap mass analyzers with a linear ion trap (Makarov, et al., 2006; Makarov, et al., 2006)) or with a quadrupole were introduced. These new machines addressed many of the previously mentioned concerns and furthermore accelerated the FTMS technique and enabled the acquisition of high accurate spectra without compromising acquisition speed. Since their introduction, LTQ Orbitrap instruments are frequently used in all fields dealing with the annotation of molecules by mass spectrometry. In (Schwudke, et al., 2007) it was shown that the increase in resolution and mass accuracy of LTQ Orbitrap instruments tremendously increases the number of accurate lipid identifications without utilizing tandem mass spectrometry, achieving tremendously short acquisition times and making high-throughput lipidomics possible.

Although it was developed for and applied on hybrid quadrupole time-of-flight (Q-TOF) mass spectrometers, the method of data-dependent acquisition experienced its full potential on high accuracy LTQ Orbitrap instruments. Precursor ion scans, neutral loss scans and Boolean scans could be now coupled with high accurate spectra. In addition to the “Bottom-Up” identification of lipids, where MS/MS spectra are examined first, the method of “Top-Down” lipidomics was founded (Schwudke, et al., 2007). Here, the high resolution MS spectrum is acquired first and yields a fundamental lipid annotation. To identify also the individual lipid species, optional tandem mass spectra are acquired for the selected lipids. Thus, tandem experiments are only done for selected lipids in contrast to a provisory fragmentation of all precursors. This makes this method is a more efficient approach. Even though, the annotation of lipids of the survey spectra (another name for the MS spectra) might be suitable in many cases. High accurate mass spectra enable to distinguish previously isobaric lipids species without doing MS/MS experiments. This allows a fast high throughput lipidomics method with which an acquisition and processing of numerous spectra is possible.

2.3 Basics of shotgun lipidomics, terms and definitions

The basic workflow of a shotgun lipidomics experiment is depicted in Figure 6.

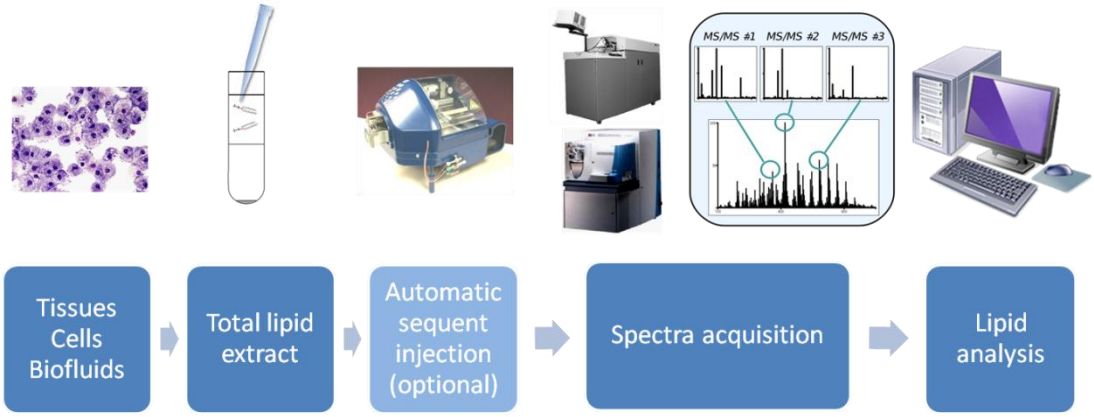


Figure 6: The workflow of shotgun lipidomics experiments.

Lipids are extracted from tissues, cells or biofluids. The total lipid extract is directly infused into the mass spectrometer. If a greater number of samples is present, the direct infusion of the extracts can be automated with a NanoMate® (Advion) which allows the automatic sequent injection. Spectra are acquired from the total lipid extracts and submitted to a subsequent computational analysis in search for lipids. The figure shows the basic setup as used in the laboratory where LipidXplorer was developed. The two mass spectrometers are an ABI Sciex QStar and a LTQ Orbitrap. (Source: own design)

In the case of biological experiments, spectra are normally acquired from different assays of an organism incorporating stages of development, mutants or species with differing attributes. The statistical significance of each assay is increased by the parallel generation of independent biological replicates. From those samples, lipids are extracted and directly infused in the mass spectrometer. This is optionally repeated with the same sample to acquire several technical replicates. Each provides a full set of MS and MS/MS spectra further termed as an acquisition (see Figure 7).

Because it produces a comprehensive data set comprising all fragment ions from all ionizable lipid precursors, the method of choice for the acquisition of spectra is data-dependent acquisition (DDA): first, a survey (MS) spectrum is acquired to determine *m/z* and abundances of precursor ions. Then, MS/MS spectra are acquired from several automatically selected precursors (the order is normally the decreasing abundance) and then the acquisition cycle (MS spectrum followed by a few MS/MS spectra) is repeated.

Survey spectra are acquired in the following way: within a certain period of time (for example, 30 s) a mass spectrometer repeatedly acquires individual spectra in much shorter intervals (for example, 1 s) that are termed as scans. To reduce the

signal-to-noise ratio all scan from a single acquisition are averaged to a consensus mass spectrum (see Figure 7).

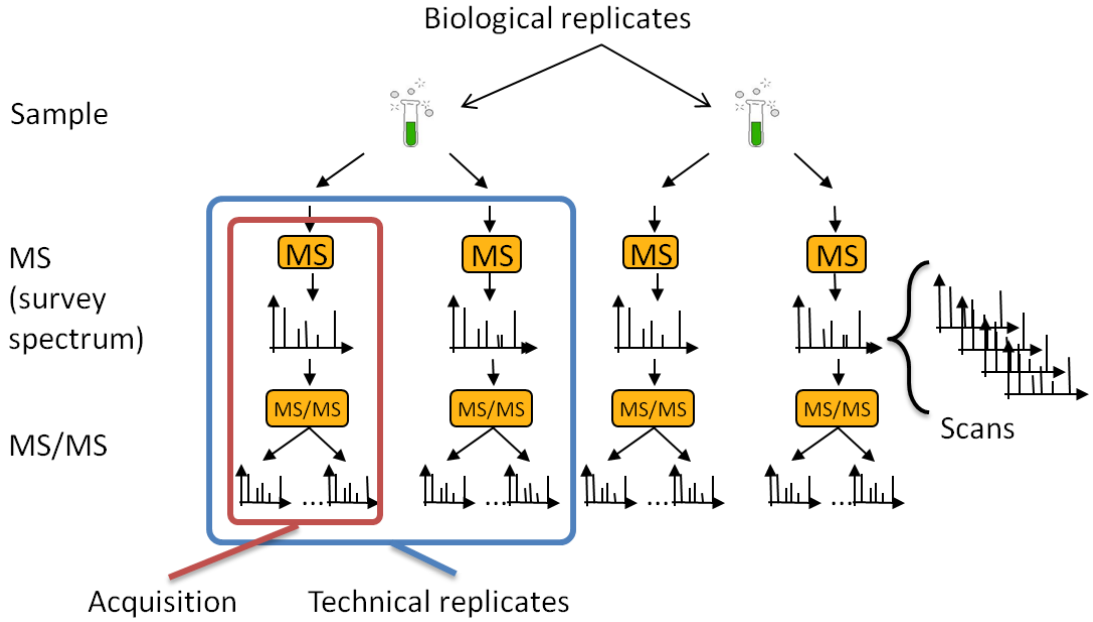


Figure 7: The shotgun lipidomics dataset.

Experiments are repeated in several independent *biological replicates* for each studied phenotype. Each biological replica is split into several samples from which lipids are extracted and extracts are independently analyzed by mass spectrometry. Mass spectra (MS) acquired from the total lipid extract surveys molecular ions of lipid precursors, which are subsequently fragmented in tandem mass spectrometric experiment yielding MS/MS spectra. Each spectrum is acquired in several *scans* that are subsequently averaged. A set of MS and MS/MS spectra is termed as an *acquisition* and several acquisitions are performed continuously hence creating a *technical replicate*. (Source: (Herzog, et al., 2011))

2.3.1 Mass spectra peak detection

A peak detection transforms the continuous spectrum, as it is produced by the mass spectrometer, into a peak list by calculating the centroids of peaks. The intensity of the peaks in the resulting peak list can be either the high of the original peak or its peak area. In the following the term mass spectrum is used for both continuous spectrum and peak list.

2.3.2 Peak tolerance

The term peak tolerance specifies the observed *m/z* shift of a monoisotopic peak with the theoretical *m/z* value of the lipids molecular formula. Since the accuracy and resolution have both an impact on the peak detection algorithm, the peak tolerance does not necessarily agree 100% to the mass spectrometers specific peak accuracy.

2.3.3 Peak occupancy

Having mass spectra from several samples, another attribute can be defined which is peak occupancy. It measures the frequency with which a particular peak was encountered in individual acquisitions within the full series of experiments. In contrast to accuracy and resolution, the peak occupancy depends on both, the instrument performance and the individual features of the analyzed samples. Thus, even multiple repetitive acquisitions do not fully compensate for the under-sampling of low abundant precursors, especially if they are detected with a poor signal-to-noise ratio. Since the method of data-dependent acquisition of MS/MS spectra is biased towards fragmenting more abundant precursors, low abundant precursors might not necessarily be fragmented in all acquisitions. Therefore, the peak occupancy attribute, helps to balance coverage and reproducibility of lipid peak detection.

2.4 Computational approaches for the automatic identification of molecules from shotgun mass spectra

Computational methods for the identification of molecules from mass spectra can be found in many research areas. In particularly the fields of proteomics and metabolomics – the large scale study of proteins and metabolites, respectively - developed a variety of by now established algorithms and sophisticated software toolkits.

The identification of metabolites can be separated in two basic approaches: the identification using a spectra database and the algorithmic interpretation of mass spectral peaks – called de-novo identification. The most common method is the comparison of experimental mass spectra with spectra of authentic standards. Many libraries of mass spectra exist (see review (Borland, et al., 2010)), for example the National Institute of Standards and Technology (NIST) database[1] or the Wiley registry of mass spectral data[2]. Each database provides a scoring function based on the similarity of the spectra which is measured with various distance measures (see (Gower and Legendr, 1986) for an overview). The

[1] http://chemdata.nist.gov/

[2] http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470520353.html

second approach is the de-novo identification and deduction of structure hypothesizes for compounds. The research in this area is in its beginnings (Neumann and Böcker, 2010). However, several approaches exist which use an intermediate step: They pre-process mass spectra to add more information before sending it to a metabolite database. In (Mohamed, et al., 2009), for example, the mass spectra were processed by a principle component analysis to reduce background ions. A further processing identified isotopes, ion adducts and fragments and generated a reduced list of *m/z* candidates. Then, the software MZSearcher simultaneously searches several lists of metabolomics databases (e.g. Kyoto Encyclopedia of Genes and Genomes (KEGG), Human Metabolome Data Base (HMDB), etc) to propose biomarker candidates. The software SIRIUS Starbust (Böcker and Rasche, 2008) goes a step further towards the de-novo identification of metabolites. It calculates the correct sum formula from MS and MS/MS data and proposes additionally a tree representation that is often close to the actual fragmentation tree of the compound. Another class of algorithms use the systematic bond disconnection algorithm (EPIC, elucidation of product ion connectivity) (Hill and Mortishire-Smith, 2005) which identifies the most likely explanatory fragments with a scoring function using a set of simple, generally applicable penalties. Other commercially available tools use databases of fragmentation rules based on gas phased election impact mass spectra were cleavage sites can be approximated and described with rules of possible fragmentation reactions[1,2]. However, the work of (Heinonen, et al., 2008; Horai, et al., 2009) shows that both tools could interpret only a part of experimental data. Being able to assign (sub-) structures to mass spectral peaks in reasonable time with a software tool, it is possible to screen comparatively large general-purpose molecular databases (e.g. PubChem, KEGG or ChemSpider) and calculate a score of the agreement between the spectrum and candidate compounds. This approach is used by (Hill, et al., 2008) and for the MetFrag suite (Wolf, et al., 2010).

Since proteins are large molecules and mass spectrometric driven proteomics is

[1] www.highchem.com/massfrontiermass-frontier.html

[2] http://www.acdlabs.com/products/adh/ms/ms_Frag/

most of the time coupled with chromatographic separation, it produces hundreds to tens of thousands (fragment-) ion spectra per hour of data acquisition. The complex computational challenge is to assign peptide sequences to these fragment spectra, infer the according proteins and determine their abundances from this high amount of data (Nesvizhskii, et al., 2007). Since the field of proteomics is quite older than lipidomics, a great variety of algorithms and software exists by now. The main source of information for MS/MS based proteomics is the fragment ion spectrum of a specific peptide ion. The first and central step for proteomics data processing is the correct assignment of such a fragment spectrum to a peptide sequence. A large number of computational approaches and software tools have been developed for an automatically assignment and can be classified in three groups: (i) Database searching, where theoretical or experimental peptide fragment spectra stored in databases are correlated with the acquired fragment spectra; (ii) de novo sequencing, where algorithms calculate the peptide sequence from the fragment spectra; (iii) hybrid approaches (Nesvizhskii, et al., 2007). Although written some time ago in 2007, the review of (Nesvizhskii, et al., 2007) lists 35 proprietary and freely available software tools associated to protein identification. Commonly accepted statistical methods offer the evaluation of the result (e.g. the false positive rate (FPR), the family-wise error rate (FWER) or the false discovery rate (FDR); see (Nesvizhskii, et al., 2007)). The quantification of proteins is as well a complex undertaking since the quantification is based on the numerous peptides coming from a protein. Three approaches have emerged: (i) quantification based on spectral counting, (ii) quantification via differential stable isotope labeling of peptides or proteins and (iii) label-free quantification based on the precursor ion signal intensities (see (Mueller, et al., 2008) for a review). More recent developed software aims at comprising all aspects of the whole process of protein identification, which includes spectra pre-processing (noise filtering, baseline subtraction, peak detection, the detection of isotopic distribution, peak overlaps, charge deconvolution, etc.) (see (Matthiesen, et al., 2010) for a review) in one toolkit. A well received and wide spread software toolkit is the Trans Proteomics Pipeline (Deutsch, et al., 2010; Keller, et al., 2005). It contains fully functional modules which can be controlled by a unified user interface. MaxQuant is a

single software tool including the all necessary algorithms and routines for the whole process of protein identification and quantification (Cox and Mann, 2008; Graumann, et al., 2011). The OpenMS proteomics pipeline (TOPP) (Kohlbacher, et al., 2007) is similar to TTP with additional functions for spectra processing. Moreover, it offers more control over the workflow of the data analysis in order to build customized applications or even implement own algorithms using existing data structures with a graphical user interface using a data flow paradigm (Reinert and Kohlbacher, 2009).

2.5 Processing mass spectra for lipid identification

Although lipids are a subset of metabolites, the algorithms for molecule identification from the field of metabolomics are not feasible for lipid identification. Metabolites have a far greater structural diversity which needs complex algorithms for their identification. In contrast, lipids have a more structured assembling which allows having less complicated algorithms to identify a greater variety of lipid species. Approaches utilizing a metabolome database have an inherent problem if parent ions overlap when metabolites are identified with the help of fragment spectra. Since the fragment spectra contain fragments of several parent ions the identification score can be easily corrupted. This hampers the identification with the shotgun method. For random test, the parent ion with m/z 660.4609 from a high accuracy *E.coli* acquisition with MS/MS experiments on an LTQ Orbitrap was input into MetFrag. Manual inspection and knowledge from the sample extraction determined the existence of three phosphatidylethanolamine species, namely PE [14:0/16:1], PE [14:1/16:0] and PE [12:0/18:1], under the same parent mass. MetFrag listed 44 compounds. One of the three lipid species was found ranked on the second place with a score of 0.8 and the other two with a lower score of 0.66 and 0.35 and ranks of the 4^{th} and 8^{th} position, respectively. The result is ambiguous. Even reducing the fragment spectra artificially to only three peaks which are specific for PE [14:1/16:0] resulted in the right species being ranked third, but now with a score of 0.97.

The algorithms used for the identification of proteins are even less applicable for lipid identification. Proteins are large molecules comprised of a single linear polymer chain of amino acids. The average protein has 466 amino acids and a

mass of 53 kilo Dalton (Lodish, et al., 2005). The amount of fragments produced is much higher and needs completely different approaches for identification and quantification. Lipids produce only a moderate number of specific fragments. Often these are class-specific fragments, immediately indicating the lipid class.

In contrast to metabolomics and proteomics, the field of lipidomics is a quite young research area. Thus, the number and variety of available software for lipid identification and quantification is quite smaller. With the advancing mass spectrometer technology and the increasing interest in lipidomics and therefore the growing numbers of samples to be examined, a software aided interpretation of mass spectra became necessary for lipidomics. This is reflected by the number of tools published in the last 5 years (Ejsing, et al., 2006; Hübner, et al., 2009; Leavell and Leary, 2006; Rockwood and Haimi, 2006; Schwudke, et al., 2006; Song, et al., 2007; Taguchi, et al., 2007; Yang, et al., 2009). But furthermore, recently developed computer-aided spectra interpretation algorithms increase the possibilities for lipid identification (Schwudke, et al., 2006). Especially the introduction of Orbitrap FT-MS in use with data dependent acquisition still lacks good software tools which take full advantage of these methods (Schwudke, et al., 2007).

2.6 Available software for lipid identification and quantification using shotgun lipidomics

Available software for identification of lipids from mass spectra can be separated into two groups: spreadsheet based software and stand-alone tools.

The first collection of tools does a simple comparison of *m/z* values from a database or given as a (MS Excel-) table with experimental *m/z* peak values. This approach is the most obvious, since it is a simple computational aid for the hand picking of MS peaks. These tools come as Excel plug-ins (Hübner, et al., 2009; Leavell and Leary, 2006; Rockwood and Haimi, 2006; Yang, et al., 2009) and include many more functions like isotopic correction, baseline correction or an automatic quantification to a given standard. The second group of software tools uses more complex lipid identification algorithms, especially by making use of tandem mass spectrometry. By using both, MS and MS/MS spectra, the calculation of the exact molecular species is possible. One characteristic is that

they directly import and process the mass spectra without using a step in-between like Microsoft Excel. This can be one spectrum at a time (LipidSearch) or a batch of spectra (LipidQA, LipidProfiler). Clearly, this increases the throughput of the lipid identification significantly, since one does not have to load each individual spectrum into spreadsheet software.

In the following, a selection of available lipid identification software is presented. The focus of the tools is their applicability to shotgun lipidomics. Some tools are written for liquid-chromatography mass spectrometry (LC-MS) (LIMSA and LipID), but can also be used for MS screening of shotgun spectra, if high resolution and high accurate spectra are available. A comprehensive list of available software for lipid analysis can be found on the web site of Oliver Fiehn[1].

2.6.1 Spreadsheet-based lipidomics tools

The following lipidomics tools are based on or are add-ins of the Microsoft Excel spreadsheet software, respectively.

2.6.1.1 LIMSA

The software called lipid mass spectrum analysis (LIMSA (Haimi, et al., 2006)) is a Microsoft Excel add-on with a dynamic lipid library that identifies and integrates the peaks in an imported spectrum, corrects the peak areas for overlap by isotopic peaks of other species and quantifies the identified species using included internal standards. It is instrument-independent because it processes text-format MS-spectra given as network common data format (NetCDF) files or as excel tables. Together with pasting the mass spectrum into the first two columns of a Microsoft Excel worksheet, the user gives a list of compounds with names and sum formulas as well as a list of isotope distributions for the lipid building atoms, which are both stored internally in the Microsoft Excel add-in. The compound list is selected or constructed from LIMSA's editable database, containing initially more than 3000 lipid species. The given peak list is then probed for peaks of the compound list. Unmatched compounds are removed from

[1] http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics/LipidAnalysis/

the compound list. The remaining species are corrected for overlapping isotopes in the manner described in (Rockwood and Haimi, 2006).

2.6.1.2 FAAT

The Fatty Acid Analysis Tool (FAAT) (Leavell and Leary, 2006) is a Microsoft Excel add-in consisting of two worksheets: the interface worksheet and the lipid library worksheet. The interface worksheet contains all necessary settings and data sources, the library worksheet is a user-defined list with [<lipid species name>, <exact mass>, <elemental composition>, <expected charge state>, <CH_2 heterogeneity>, <analysis polarity>] tuples. The targeted mass spectrometric platforms are high accuracy and high resolution instruments. The lipid identification is done by comparing the lipid species entries from the library worksheet with the content of the files given in the interface worksheet, involving a user-given mass error. A noteworthy feature of FAAT includes the possibility to align up to 5 spectra to a reference spectrum to assign isotopic shifts to the found lipid species which aids the assignment of elemental composition and structure (Converse, et al., 2003; Mougous, et al., 2004; Mougous, et al., 2006).

2.6.1.3 LipID

LipID is an easy-to-use program, which enables the assignment of single mass-to-charge values or lists of mass-to-charge values to various lipid species (Hübner, et al., 2009) in a similar manner as the above presented LIMSA. It does not account for MS/MS data and is therefore only suitable with LC-MS or high accuracy FT-MS data, because otherwise with a lower resolution, the number of possible matches for an *m/z* value would not allow an unequivocal identification. Differing to the other tools, it simplifies the lipid database handling by utilizing only the lipid class specific head group as input and automatically calculates species defined by their fatty acid compositions.

2.6.1.4 AMDMS-SL

The software called Automation of Multi-Dimensional Mass Spectrometry-based Shotgun Lipidomics (AMDMS-SL) (Yang, et al., 2009) is a product of Multi-Dimensional Mass Spectrometry-based Shotgun Lipidomics (MDMS-SL) platform

to analyze individual lipid species in shotgun lipidomics (Han and Gross, 2005; Han and Gross, 2001; Han and Gross, 2003; Han and Gross, 2005). It uses a build-in database to identify lipids either from MS spectra and/or precursor ion scan (PIS) and neutral loss scan (NLS) spectra. The spectra are again pasted into a Microsoft Excel Worksheet and using AMDMS-SL as an add-in. It features baseline correction and removal of background noise as well as an isotopic correction routine. The peak list is matched to the internal database. The examination of MS/MS data is limited to PIS and NLS scans. A remarkable feature of the software is that it includes a routine for quantification of the lipid species based on a given internal standard. For this it considers also the MS/MS fragment intensities. The authors claim to include a database of over 36,000 lipids which is unfortunately not included in the distribution of the software. The only way to identify lipid species of interest is to add the list of *m/z* values with associated lipid names of the lipids of interest into the first worksheet of the Excel Workbook. The software spots these *m/z* values within a given mass tolerance and calculates the quantities of the lipids to a given standard.

2.6.2 Stand-alone lipidomics tools

The following lipidomics tools do not depend on any other software.

2.6.2.1 LipidProfiler

LipidProfiler (now distributed as LipidView) is described in (Ejsing, et al., 2006) and is an advanced lipid identification software. It includes a database of several lipid classes given as *m/z* values of fragment masses. The identification is done by finding those masses in precursor ion scan and neutral loss scan spectra which can be acquired also in parallel as so-called multiple precursor ion scans (MPIS). It includes an advanced isotopic correction routine which features so-called interscan and intrascan isotopic correction. Also with LipidProfiler, the abundances of the lipid species can be automatically normalized to a given internal standard. To our current knowledge this is the only proprietary software for lipid analysis and is bound to ABI Sciex mass spectrometers.

2.6.2.2 LipidQA

LipidQA (Song, et al., 2007) is freely available and not dependent on Microsoft Excel. The software imports one or more mass spectra acquired with data dependent acquisition (DDA) in either a peak list format or in Thermo Fisher *.raw file format. It includes a lipid spectra database which was built automatically. For lipid identification it compares the spectra from the database with the given experimental spectra and calculates a score for the matching peaks. However, spectral comparison is arguably not feasible for lipidomics, especially for shotgun lipidomics, since isobaric species will mislead the scoring function.

2.6.2.3 LipidSearch

LipidSearch (Taguchi, et al., 2007) is a web service for lipid identification from mass spectra with numerous settings. Single spectra given as Thermo Fisher *.raw files or peak list files acquired in data dependent acquisition (DDA) or PIS/NLS mode are uploaded and the result returned as an html table. The settings include tolerance values for the peak identification as well as threshold values for noise reduction.

2.6.2.4 LipidInspector

The method of data dependent acquisition (DDA) is recognized as a high accuracy, high throughput acquisition method as compared to the popular PIS and NLS (Schwudke, et al., 2006). However, the resulting comprehensive data set needs an optimal interpretation algorithm. LipidInspector (Schwudke, et al., 2006) was developed for this purpose and simulates PIS and NLS on DDA acquired spectral data sets which are given in peak list formats. A small number of algorithms simulate several precursor ion scans and neutral scan which can be applied in parallel. The software was intentionally developed as proof of principle for lipid identification by simulation of PIS and NLS. Therefore, the software has only a limited number of features and is not extensible to other lipid classes. But it demonstrates well the advantages of using DDA acquired spectra for lipid identification. The acquisition needs to be done only once but the data set can be re-used any time with different PIS and NLS scans. No repetitive acquisitions are necessary which saves the amount of sample needed for the acquisition.

2.7 Limitations of the presented software tools

A critical point concerning all of the presented software tools is their use of a single generic lipid database (LipidQA, LipidSearch, LipidProfiler, LipidInspector, and LIMSA). It is often difficult and always impractical in a long term. Since lipidomics is a quite new scientific field, the database will not be complete. This thesis will show that even a recognized database like LipidMAPS is not complete even for common lipid species. Furthermore, different organisms contain different lipidomes. If a database is not limited to the scope of the organism, the probability to produce false positive identifications increases strongly.

Some tools (LipidQA, LipidSearch) rely on databases of MS/MS spectra and use a spectra matching approach. This is usually counterproductive because in shotgun lipidomic experiments the collision-induced dissociation (CID) of lipids yields mixed populations of fragment ions. The true matching score for each individual lipid of an isobaric precursor cannot be achieved, because the peaks of the other species might not fit in the individual matching spectrum and spoil its matching score. Since lipids normally have just a few fragments the impact can be quite strong.

Another limitation concerning all presented programs is their limited support of mass spectra file formats. LipidProfiler is bound to instruments of ABI Sciex; it is not compatible to other mass spectrometers file formats. LipidQA and LipidSearch support *.raw spectra from Thermo Scientific instruments and a simple generic peak list file format (*.pkl). LipidInspector supports another generic peak list format, which is *.dta. All the other tools, beside LipidProfiler, support only peak lists which are copy-pasted or loaded individually into an Excel spreadsheet. No tool supports the newly developed XML formats mzXML (Lin, et al., 2005), mzData (Orchard, et al., 2004) or mzML (Martens, et al., 2011), which are meant to be the standard mass spectra output formats to facilitate data sharing and analysis. Those xml formats do not only contain (mass, intensity) pairs, but also other relevant technical details about the spectra acquisition, like the precursor mass of an MS/MS spectrum, for example.

A typical lipidomics study might encompass 10 to 100 individual samples, from each of which 10 - 100 MS and 100 - 1000 MS/MS spectra are acquired. Thus, a typical shotgun experiment yields several hundreds of MS and MS/MS spectra

(see Figure 7) of which several are redundant coming from biological and technical replicates. Apparently, the biggest drawback of the Excel based software tools is that they are not able to perform in a high throughput manner. Copy-pasting the spectra into Excel is apparently not a valid option since it heavily challenges the users' time and patience. But also the tools which are not based on Excel have only limited batch processing modes. LipidQA allows the input of several spectra at once, but does not have any alignment routine, so that the results have to be compared by hand, which only slightly helps with high throughput lipidomics. Only LipidProfiler allows the import of several spectra and offers a subsequent alignment of the resulting lipid species.

All tools are optimized to a certain type of spectral acquisition, tandem mass spectrometry settings and accuracy/resolution of the used machine. All these factors influence the required interpretation algorithms. For example, the identification algorithms of LIMSA, FAAT and LipID only do a comparison of *m/z* values of the intact lipid ions, but do not support tandem mass spectra. Thus, those tools cannot identify individual lipid species. They are only able to identify non isobaric lipids without resolving the individual length and unsaturation of the fatty acids. However, isobaric species are quite common in lipidomics. Only a chromatographic pre-fractionation helps to overcome this situation for those tools.

To further compare the existing lipidomics tools, the requirements are down to a list of features which are presented in Table 1. It indicates the existence of the features with "Yes" and the absence with no entry in the table.

Table 1

Comparison of lipidomics tools by means of selected features

FEATURE[1]	LipID	LIMSA	FAAT	AMDMS-SL	LipidQA	Lipid Profiler	Lipid Search	Lipid Inspector
MS + MS/MS (support of DDA)						Yes	Yes	Yes
Database expandability	Yes	Yes				Yes		
Isotopic correction		Yes		Yes	Yes	Yes		
Cross - platform	Yes	Yes	Yes	Yes	Yes		Yes	Yes
Support of individual acquisition modes				Yes		Yes	Yes	
Spectra alignment			Yes					
Grouping						Yes		
Batch mode		Yes	Yes		Yes	Yes		Yes
Customization of identification algorithm								

[1]List of features: **MS + MS/MS,** lipid identification based on parallel introspection of MS and MS/MS spectra, required for identifying individual molecular species; **Database expandability,** users may expand references databases at will; **Isotopic correction,** overlapping isotopic clusters are detected and the intensities of corresponding monoisotopic peaks are adjusted; **Cross-platform**, can process spectra acquired on mass spectrometers from different vendors; Support of individual acquisition modes, supports different acquisition modes like data dependent acquisition (DDA) and precursor ion scan (PIS); **Spectra alignment,** supports alignment of multiple spectra within the series of experiments; **Grouping**, supports grouping of spectra of biological and technical replicates acquired from the same sample; **Batch mode,** supports processing of multiple spectra submitted as a batch; **Customization of identification algorithm,** supports the customization of the lipid identification algorithm to fit the mass spectrometers attributes.

Most probably, the limits of the presented software solutions result from the focus of the laboratory where they were developed. Every lab uses its own lipid identification strategy, its own type of mass spectrometer and is focused on the

subset of lipids which is of interest. As soon as the presented software tools are used for other machines, other spectral formats, other identification goals, they soon reach their limitations. Unfortunately this is hidden from the user in many cases. Most tools do not give sufficient insight into their lipid identification mechanism which makes it difficult or impossible to judge the result. Adding to this is the non-existence of benchmarks of lipid identification software.

3 From single scans to mass spectra

During shotgun analyses, spectra are acquired in the following way: within a certain period of time (for example 30 s) a mass spectrometer repeatedly acquires individual spectra in much shorter intervals (for example 1 s) that are termed as scans (see Figure 7). Although being successive acquisitions of the same sample, it is not fully reproducible and masses of identical precursors and fragments vary within certain ranges depending on the mass spectrometers accuracy and resolution. Hence, a subsequent averaging of all related scans into a single representative spectrum reduces the influence of those mass shifts. And this increases the mass accuracy and improves ion statistics of both the measured masses and abundances of corresponding peaks. This technique is commonly applied in proteomics (Frank, et al., 2008; Liu, et al., 2007).

In the following, LipidXplorer's scan averaging algorithm will be reported and compared to other spectrum averaging and alignment methods. The method of choice for LipidXplorer was a mixed binning algorithm.

3.1 Scan averaging and alignment algorithms

One way of averaging several spectra is to align them and calculate the average of the aligned peaks. Alignment, as it is defined here, is grouping together peaks of different spectra which are supposed to be of the same molecular origin. Even though, the alignment is also used further for structuring LipidXplorer's spectral database.

A spectrum can be modeled as an array $S = \{p_0, \ldots, p_n\}$ of tuples $p_i = (m_i, I_i)$ where m_i is the mass of the peak and I_i its intensity. The array is ordered by the peak mass. An aligned set of k spectra is therefore a set of t k-tuples $\hat{A} = \{A_1, A_2, \ldots, A_t\}$ of the form $A_{t'} = (p_1, p_2, \ldots, p_k)$ with $t = max(|S_1|, |S_2|, \ldots, |S_k|)$ where for every two entries $p_{j_1}, p_{j_2} \in A_i$ holds that either one or both entries are zero[1] or $|m_{j_1} - m_{j_2}| < T$ for all $A_i \in \hat{A}$. The tuple A_i represents one molecular

[1] If a signal with the m/z value m_i is not present in the in a sample i, an empty peak with $p_i = (m_i, 0)$ is inserted in the tuple $A_{t'}$ at position i.

species occurring in the different samples.

The simplest method of alignment is the alignment to a reference spectrum (Jeffries; Wong, et al., 2005; Wong, et al.). Here, all peaks are aligned to a reference peak list. Apparently the crucial step is to choose the right reference spectrum. Taking a random reference spectrum from the whole set of given spectra leads certainly to a biased alignment. But any other way besides having the mean spectrum as reference is actually a method of recalibration and the mean spectrum is what should actually be the result of the alignment.

In the following, several methods of spectra alignment which do not rely on a reference spectrum are presented. The methods utilized for alignment are dynamic programming (Frank, et al., 2008; Sauve and Speed, 2004), hierarchical clustering (Tibshirani, et al., 2004; Yu, et al., 2006) and so-called binning or bucketing approaches (Frank, et al., 2008; Geurts, et al., 2005; Morris, et al., 2005; Yasui, et al., 2003).

3.1.1 Dynamic programming

Dynamic programming is a method of breaking complex problems down into sub problems. For alignment of mass spectra this would mean to have a sequential decision making procedure in the sense that the correspondence between two peak sets is established in an accumulative way. Alignment with dynamic programming is extensively used to compare protein amino acid sequences (Higgins and Sharp, 1988; Needleman and Wunsch, 1970) with the aim to recognize related proteins or predict protein secondary structures. Thereby, the dynamic programming approach guarantees to maximize the overall similarity between two sequences. However, with mass spectra the situation is different. The point-wise correspondence between two data sets does not always exist neither is it wanted, in particular for spectra originating from different samples. If the experiment contains samples with different lipid composition and possibly also some blank runs, some data points are deliberately empty. There are no “gaps” like it is the case with the alignment of amino acid sequences. An alignment score can hardly be estimated. And moreover, dynamic programming is of polynomial complexity, especially the alignment of multiple spectra quickly leads to a computationally intractable problem (Lipman, et al., 1989; Robinson, et

al., 2007; Thompson, et al., 1994).

3.1.2 Hierarchical clustering

Agglomerative hierarchical clustering (Grabmeier and Rudolph, 2002) involves iterations during which two clusters that are at a minimum distance apart are combined together. This would serve exactly what is needed for mass spectra alignment, because peaks could be grouped together which are of the same molecular origin but slightly shifted in their *m/z* value due to limits of resolution and accuracy in the mass spectrometer. It is well known that the algorithm for hierarchical clustering of n peaks has a complexity of $O(n^3)$. Using priority queues, a complete linkage hierarchical clustering algorithm can achieve a runtime of $O(n^2 \log n)$ (Krznaric and Levcopoulos, 2002) but still this is quite time consuming and not feasible even for just a few mass spectra since they can easily contain thousands of peaks.

3.1.3 Binning alignment

Binning (or bucketing) alignment is probably the simplest alignment algorithm with $O(n)$ complexity. Here, all peaks of all given mass spectra originating from different samples are put in one ordered list before this list is “binned” with bins of a size matching the peak shifts originating from the spectra acquisition. Thus, peaks of the same molecular origin should fall into one bin. One can differentiate between “fixed binning” and “mixed binning” (Kazmi, et al., 2006) algorithms. While “fixed binning” uses a fixed bin size, the “mixed binning” approach relies on increasing the bin size with higher *m/z* values. It was observed that with certain mass spectrometers, the resolution changes with higher *m/z* values. This is especially the case with LTQ Orbitrap instruments (Makarov, et al., 2009). Since the bin size is highly connected to the mass spectrometers resolution, a mixed binning approach mirrors the characteristics of mass spectrometers more closely.
An important point of critique of the binning approach is the “boundary problem” where a peak that is marginally close to a fixed bin boundary falls outside of the bin although it belongs to the bins true replicate peaks or the other way round where two peaks of different origins fall into the same bin because of being close to the bin boundary. The techniques to avoid this effect is first to have a peak

dependent binning algorithm. This means to let every bin start with a peak which is opposite to discretizing the mass spectra in the bins sized slices first and collecting the peaks in the several bins later. Therefore, the probability that a true replicate peak falls outside its associated bin decreases. Second, the averaging algorithm should be repeated on the already averaged spectrum. This collects peaks which fell outside the bin size, although close to the bin border. The reason for that can be found in the assumed Gaussian distribution of peak shifts (James A, 1983; Weichuan, et al., 2005; Zubarev and Mann, 2007). Details can be found in Appendix 10.3. Using a changing bin size also reduces the boundary problem, since it is a more accurate simulation of the mass spectra's characteristics.

3.1.4 Heuristic alignment

In (Kazmi, et al., 2006) the boundary problem is addressed by adding a heuristic which breaks or merges adjacent bins with a complete linkage hierarchical alignment. The complexity of this algorithm is$O(nr^2 \log r)$ where n is the number of peaks and r the number of spectra to be aligned. The article shows a significant difference in the quality between a simple mixed binning algorithm and the approach of "heuristic alignment", while in their findings was no significant difference between the heuristic alignment and a complete linkage clustering.

The heuristic method proposed in (Kazmi, et al., 2006) seems to be a good compromise between speed and accuracy, yet a mixed binning algorithm was used for LipidXplorer. The reason is, as it will be shown in Chapter 6.5.1, that the mixed binning approach outperforms the heuristic method in speed about a manifold without a significant compromise in the mass accuracy as needed for lipid identification. Since LipidXplorer is meant to be high throughput lipidomics software, speed is an important factor.

LipidXplorer uses a customizable mixed binning algorithm where the degree of change of the bin size depends on the mass resolution gradient which depends on his part on the used mass spectrometer. For example: as presented in (Makarov, et al., 2009) there is a strong change in the resolution with increasing *m/z*, especially with the high mass accuracy LTQ Orbitrap instruments. LipidXplorer approximates this behavior with a linear *m/z* to resolution function defined by the degree of resolution change.

Like most spectra alignment algorithms it assumes that masses pertinent to the same peak are Gaussian distributed within individual spectra. The algorithm recognizes related peaks in each individual spectrum and averages their masses and intensities (see Appendix 10.3).

3.2 The implementation of the mixed binning algorithm for the scan averaging

Before spectra can be aligned and probed for lipids they have to be generated by averaging the individual scans (see Figure 7 for details) which will improve the signal-to-noise value. The mixed binning algorithm was used for scan averaging.
First, the algorithm considers all pertinent scans within the acquisition and combines all reported masses into a single peak list (see Figure 7). This list is then sorted by masses in ascending order and the averaging proceeds in steps, starting from the lowest detected mass. In every step, the algorithm considers the mass *m* and checks whether other masses fall into a bin of $[m, m+\frac{m}{R}]$ width, where $R(m)$ is the mass resolution at the mass m. $R(m)$ is assumed to change linearly within the full mass range; its slope (mass resolution gradient) and intercept (resolution at the lowest mass of the full mass range) are instrument-dependent features and have to be pre-calculated by the user from some reference spectra. This ensures that the alignment fits the instruments abilities. All masses within the bin are average weighted by peak intensities according to Equation 1:

$$m_{avg} = \frac{\sum_{m_{i\in B}} \frac{I(m_i)}{I_{\max}} m_i}{\sum_{m_{i\in B}} \frac{I(m_i)}{I_{\max}}} \qquad (1)$$

where $I(m_i)$ is the intensity of the peak having mass m_i; $I_{\max}$ is the intensity of the most abundant peak within the bin B and m_{avg} is the intensity weighted average mass.

The average mass is then stored as a single representative mass for this bin and the procedure is repeated for the next mass bin (see Figure 8). It was assumed that the variation of peak masses is normally distributed within the bin. This causes the previously discussed boundary problem and therefore the procedure is repeated several times (for details see Appendix 10.3). While applying the algorithm, it was found that three successive iterations are always sufficient for a complete separation of bins in a way that masses are collected correctly into their dedicated bins and no two adjacent bins are closer than the value of $\frac{m}{R(m)}$.

One known limitation of this algorithm is that abundant chemical noise might impact the binning accuracy. Therefore, a threshold value T is set by the user for a signal-to-noise ratio of peaks of his choice. Before the algorithm starts, all peaks below the threshold are sorted out of the peak lists. A commonly accepted estimate for calculating the limit of detection (LOD) is a signal-to-noise ratio of 3.0.

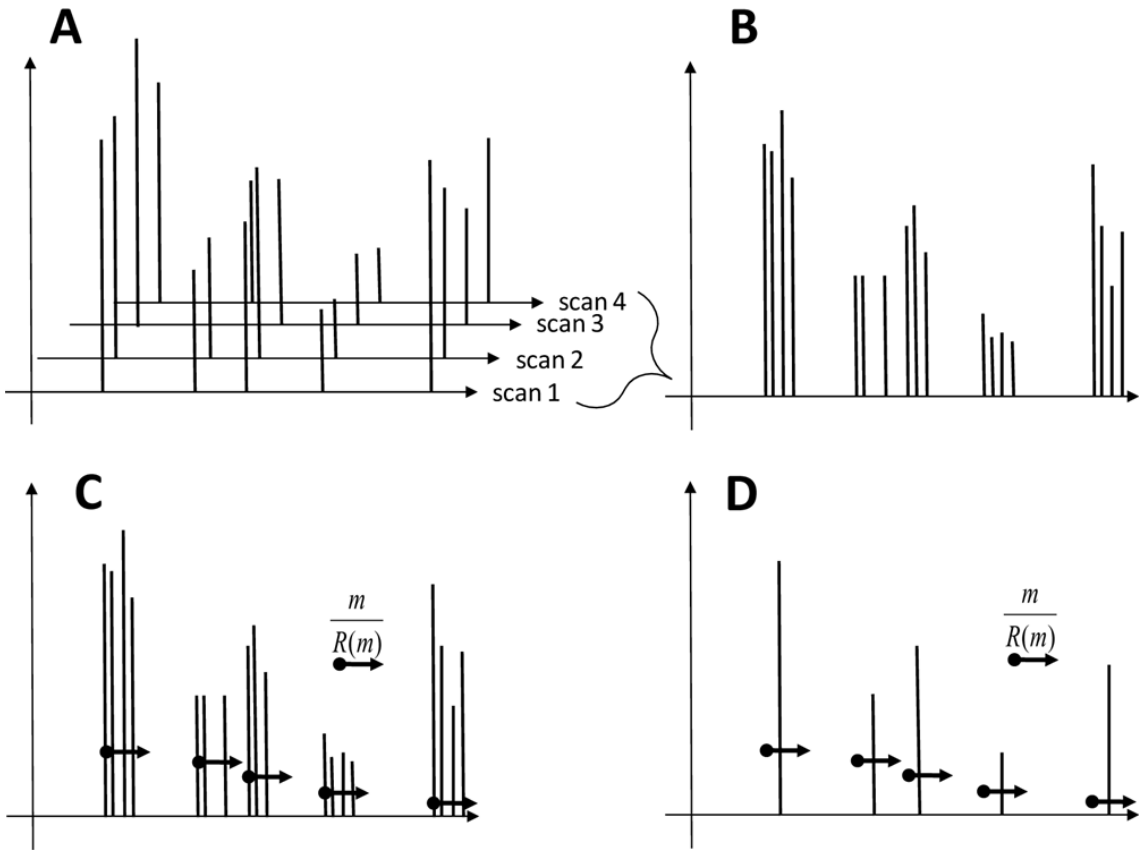


Figure 8: Scan averaging algorithm

Related individual scans (here as an example which only shows four scans) from a single acquisition or a precursor mass are imported (A); peaks are combined into a single peak list and sorted by means of the *m/z* value (B). The algorithm collects peaks which fall into bins with the size of [m; $m+\frac{m}{R(m)}$]. It starts from the lowest reported mass and continues with the next peak after a bin (C). The bold dots stand for the lowest mass of each bin, while the arrow length reflects the bin size $\frac{m}{R(m)}$. Within each bin, masses are weight averaged by peak intensities (see Equation 1) and stored. The procedure (steps C and D) is repeated two more times on the binned spectrum (not shown). In this way, a single representative average spectrum (D) is produced from several individual scans (A). (Source: (Herzog, et al., 2011))

4 The MasterScan: a database of shotgun spectra

As mentioned in the introduction, a common biological experiment comprises tens to hundreds of samples of which every single spectrum should be probed for lipids. This needs a good organization of the spectra to have a feasible batch processing routine and an aid for the interpretation of the result. To achieve this, the database MasterScan was developed for which the spectra are aligned by means of the peak's *m/z* value before they are stored. Aligned peaks are represented by their average *m/z* value. Due to the alignment, only peak intensities assigned to the averaged masses, rather than masses of individual peaks, are stored in the MasterScan. This decreases the redundancy, improves ion statistics, saves memory and allows the processing of several spectra at once. The alignment utilizes the same algorithms as the scan averaging because at its core the problems are similar.

For alignment of *m/z* values from the different spectra, a modified version of the mixed binning alignment algorithm was used. The difference is that peak abundances are not averaged but saved in a dictionary. The aligned peaks consist of the average *m/z* value of all binned peaks coupled with a dictionary containing the intensities for each occurrence in a sample:

(*m/z* for peak n, {"sample 1" : abundance of peak n in sample 1}, …, {"sample m" : abundance of peak n in sample m}Smith, 2000)

The MS/MS spectra are associated to their precursor masses from the survey spectrum. Since the isolation window is greater than the mass spectrometers resolution, fragment spectra are associated to several precursor masses. If two isolation windows overlap, the survey masses are distributed to the closest precursor masses which are given in the header of the MS/MS spectra.

The representative masses of the aligned peaks, their intensities in individual MS spectra and aligned MS/MS spectra associated with corresponding precursor masses represent the content of the MasterScan (see Figure 9). This makes the MasterScan a comprehensive database containing all shotgun lipidomics information of all samples produced in the full series of biological experiments.

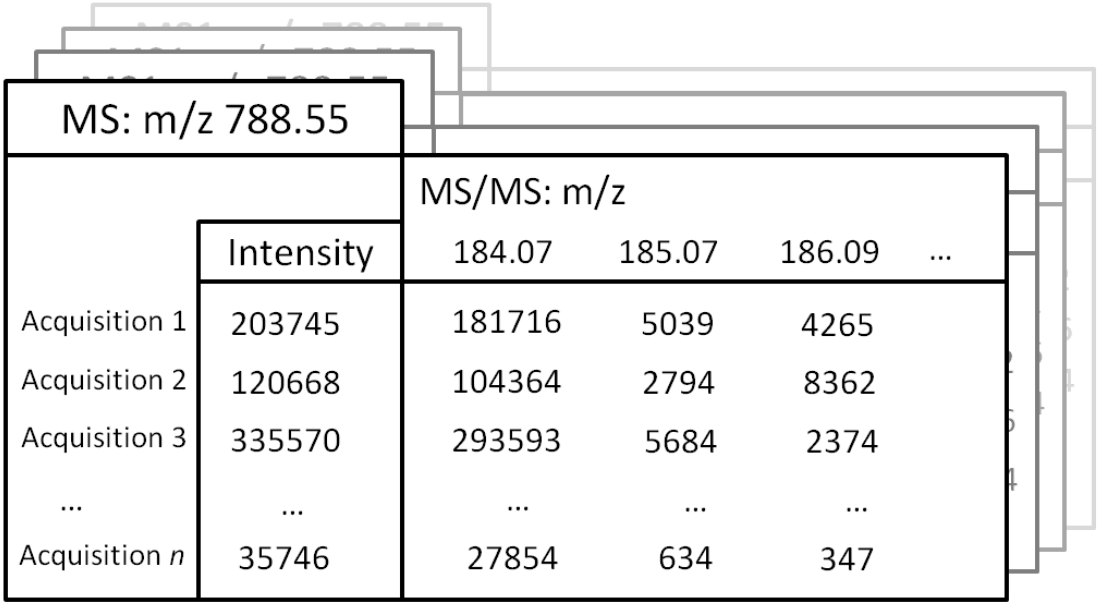


Figure 9: Organization of a MasterScan file.

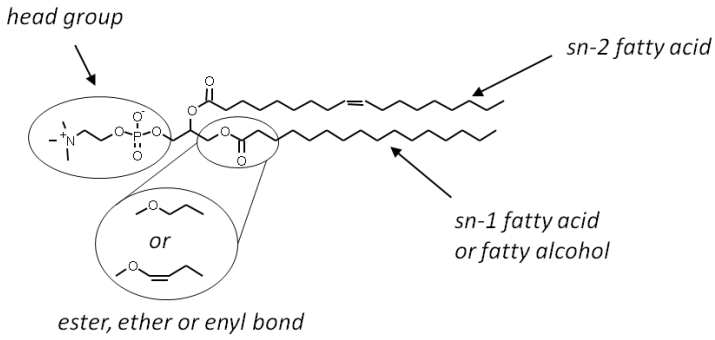
LipidXplorer imports and aligns MS and MS/MS spectra into a flat file database called MasterScan. It is shown here as a file cabinet addressed at the top-level by the precursor masses of the MS spectrum, while their intensities are assigned to individual acquisitions. In this example the lipid precursor with *m/z* 788.55 was observed in all acquisitions with the intensity (in arbitrary units) of 203745 in the Acquisition 1; 120668 in the Acquisition 2; … until 35746 in the Acquisition *n*. This precursor *m/z* 788.55 was fragmented in each acquisition. Masses of fragments were aligned and substituted by the averaged representative masses, while the intensities of corresponding peaks in each individual acquisition were stored. For example, fragment with *m/z* 184.07 was of the intensity of 181716 in the Acquisition 1; 104364 in the Acquisition 2; …, till 27854 in the Acquisition *n*. (Source: (Herzog, et al., 2011)).

5 The molecular fragmentation query language (MFQL)

5.1 The challenges of lipid identification

Compared to Metabolites, lipids do not have such an enormous structural diversity. Most lipid classes consist of structural building blocks which allow customized spectra interpretation routines. This characteristic was used when the molecular fragmentation query language was developed.

Nevertheless, within their molecular category, lipids have diverse complex structures. The class of Phosphatidylcholine (PC) lipid species for example, comprises a glycerol structure to which its specific head group and two fatty acid moieties are attached. The fatty acids differ by the number of carbon atoms and unsaturations and different bonds at the sn-1 position. Every of those configurations is an individual lipid species whose *m/z* value might easily overlap with other lipid species. For an accurate determination of most common glycerophospho lipid classes the examination of fragment spectra is therefore necessary (see Figure 10).



All lipids of PC class: 'C[30..48] H[30..200] N[1] O[7..8] P[1]'

All PC (esters): 'C[30..48] H[30..200] N[1] O[8] P[1]'

PC 36:1 : 'C[41] H[86] N[1] O[8] P[1]'

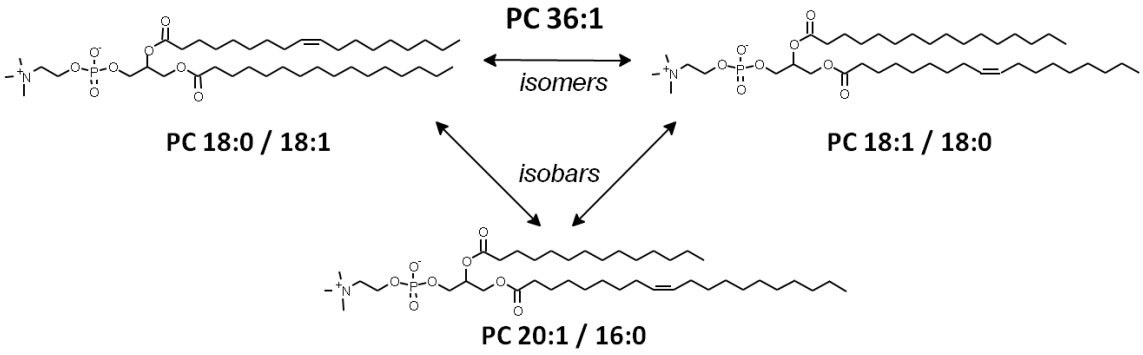


Figure 10: Structural complexity of lipid species and sum composition constraints.

Phosphatidylcholine (PC) should act as a representative example: PC molecules consist of a posphorylcholine head group attached to the glycerol backbone at the *sn-3* position, while fatty acid moieties occupy *sn-1* and *sn-2* positions (alternatively, a fatty alcohol moiety could be attached at the *sn-1* position). Fatty acid moieties differ by the number of carbon atoms and double bonds, but also by the relative location at the glycerol backbone, so that isomeric structures having exactly the same fatty acid moieties are possible. Every configuration of the fatty acid moieties and the kind of bond on the sn-1 represents an individual lipid species. Additionally, lipid species can occur as isometric or isobaric species. Most generic constraints ("All lipids of PC class" or "All PC esters") encompass sum compositions of species with all naturally occurring fatty acids. However, because of the fatty acid variability, some species of other lipid classes (such as phosphatidylethanolamine (PE)) might meet the same constraint. Therefore, for most common glycerophospholipid classes the characterization of individual molecular species could not solely rely on their intact masses, irrespective of how accurately they were measured. MS/MS experiments that produce structure-specific ions contribute more specific constraints, such as the number of carbons and double bonds in individual moieties, characteristic head group fragment, characteristic loss of a fatty acid moiety, among others. These constraints can be bundled by Boolean operations. (Source: (Herzog, et al., 2011))

5.2 A query language for the MasterScan

A query language is a specified language for requesting information from a database. The information is organized by using a certain paradigm. For example, the paradigm of relational structured databases (Codd, 1970) matches

data by common characteristics found within the dataset. In the same sense mass spectra were structured and organized into the MasterScan database to be able to probe it for patterns of information which lead to the identification of complex molecules. From current knowledge, this is the first time a query language is used for identification of lipids.

The probably most widespread query language is the ANSI standardized structured query language[1] (SQL). Although there are plenty powerful implementations, it is more beneficial to develop a new query language. The reason is that SQL queries are not intuitive when it comes to the identification of lipids from mass spectra. Keywords and terms are general and designed for any content a database can have. This makes them very technical and not intuitive. SQL queries on the MasterScan could easily become tens of lines of code with several nested loops, if it should identify lipid species. If, on the other hand, the query language is tailored to a specific problem, terms and keywords can be limited to the specific needs and the language customized to the target audience. This makes the query language far more efficient and increases the readability of the code. It could be furthermore tailored to be usable by anyone, without having a prior knowledge of programming.

This led to the development of a new query language, called Molecular Fragmentation Query Language (MFQL). MFQL allows the formalization of available or assumed knowledge of lipid fragmentation pathways into queries that are custom-made for probing the MasterScan database. In the following the design principles will be introduced and examples of queries presented.

5.3 The design of MFQL

The design of MFQL follows the design of the Structured Query Language (SQL), but has significant differences in order to narrow its scope to lipid identification in mass spectra. Furthermore, the target audience for the software should be nature scientists and not informatics. Another important design goal was to allow also non-programmers to have an easy start with MFQL, since this is the target audience for LipidXplorer. Therefore, the chosen keywords should reflect very

[1] http://en.Wikipedia.org/Wiki/SQL

well the vocabulary of how one would describe the identification of lipids from mass spectra. These are for example ‘IDENTIFY’, ‘SUCHTHAT’, ‘REPORT’ etc. The Backus-Naur Normal Form (BNF) of the language can be found in Appendix 10.5.

5.3.1 The supported data types

MFQL supports basic data types like floating point numbers and strings. Additionally it has two specific data types: chemical sum compositions and so-called sc-constraints:

1) chemical sum compositions: are composed of chemical elements and numbers which state the abundance of the elements. Examples for chemical sum compositions are: C36 H68 O10 P1 or C35 H67 O8 N1 P1. To mark them as sum compositions they are put into single quotation marks.
2) sc-constraints: With sc-constraints whole sets of chemical sum compositions can be specified. Ranges for the chemical elements in a formula are given as well as a range of allowed double bonds (Double Bond Range – DBR). With sc-constraints the user can address all species of a molecular class. As example would be: ‘C[33..49] H[50..100] O[8] N[1] P[1] db(2.5,9.5)’ for all PE species with variable length of its acyl chains and variable number of double bounds.

With those data types, precursors and fragments are addressed directly in an MFQL query.

5.3.2 The four sections of an MFQL query

An MFQL query is structured into four sections for the definition of variables, the specification of the scope, optional constraints and the format of the output:

DEFINE: here, sum compositions, sc-constraints, masses or groups of masses are defined, initialized and associates to user given variable names.

IDENTIFY: determines where and how the DEFINE content is applied. It usually encompasses searches for precursor and/or fragment ions in MS and MS/MS

spectra.

SUCHTHAT: defines optional constraints that are formulated as mathematical expressions and inequalities, numerical values, peak attributes, sum compositions and functions. Several individual constraints can be bundled by logical operations and applied together.

REPORT: establishes the content of the output. The output is always a table where the columns are defined in this section. Every column gets an arbitrary name and a content which can be the content of the defined variables or a formatted string.

With one single MFQL query all detectable species of a lipid class which share common fragmentation pathways can be annotated at once. The MFQL concept takes full advantage of the completeness of shotgun lipidomics datasets that contain all fragment ions produced from all plausible precursors. In this way, MFQL supports the parallel application of any shotgun lipidomic approach, such as for example top-down screening, multiple precursor and neutral loss scanning, multiple reaction monitoring among others.
MFQL will be further elucidated in the following with the help of two practical examples. The first scenario is the identification of phosphatidylcholine (PC) lipid species in positive ion mode acquired spectra (see Figure 11). The second scenario shows the identification of phosphatidylethanolamine (PE) from spectra acquired in negative ion mode (see Figure 12).

5.4 First example: The identification of phosphatidylcholine lipid species with MFQL in positive ion mode

Phosphatidylcholine is a major component of cell membranes. Its structure is built up from a functional head group attached by two fatty acyl moeties and is typical for many lipid classes. In MS/MS experiments, molecular cations of PC species split off their head group and produce the specific phosphorylcholine fragment having the sum composition of ‘C5 H15 O4 N1 P1’ and *m/z* value 184.07. The identification of PC species proceeds by identifying this fragment ion in MS/MS

spectra together with the accurately determined mass of the intact precursor in the MS spectrum. The complete MFQL query together with the chemical structure of a PC ion can be found in Figure 11.

The first step is the assignment of a query name:

```
QUERYNAME = Phosphatidylcholine;
```

Next is the definition of variables which describe the characteristics of the lipid species. In this case, the query should find the singly charged PC head group fragment in the MS/MS spectra and therefore:

```
DEFINE headPC = 'C5 H15 O4 N1 P1' WITH CHG = +1;
```

In a shotgun experiment not all fragmented peaks will originate from PC. For higher search specificity all precursors are defined who are expected to produce the headPC fragment in MS/MS spectra. This is done with an obligatory sc-constraint for MS peaks which defines all possible sum compositions for PC ions when including different chain lengths of the fatty acids and variable number of double bonds. The latter is expressed as a "double bond range" DBR:

```
DEFINE prPC = 'C[30..48] H[30..200] N[1] O[8] P[1]' WITH CHG =
+1, DBR = (1.5, 7.5);
```

The following 'IDENTIFY' section specifies that 'prPC' precursors should be identified in positive MS spectra and the 'headPC' fragment in positive MS/MS spectra. With this 'IDENTIFY' structure LipidXplorer can be commanded to find neutral losses by defining variables with a charge of zero. This makes them automatically a neutral loss. All statements under 'IDENTIFY' are connected with 'AND' to request that 'headPC' is searched only in MS/MS spectra of 'prPC'.

```
IDENTIFY
   prPC IN MS1+ AND
   headPC IN MS2+
```

The search space is further narrowed down by applying optional project-specific compositional constraints formulated in the 'SUCHTHAT' section. For example, it is generally assumed that mammals do not produce fatty acid moeties having odd number of carbon atoms. Therefore, the constraint is to consider only lipids with even-numbered fatty acid moieties.

```
SUCHTHAT
    isEven(prPC.chemsc[C]);
```

In this case the function ‘isEven()’ requests that candidate PC precursors should contain an even number of carbon atoms in sum. Since the head group of PC and the glycerol backbone contain 5 and 3 carbon atoms, respectively, this implies that a lipid could not comprise fatty acid moieties with odd and even number of carbon atoms at the same time. Cases of lipids comprising two odd fatty acids unfortunately give also a positive result with ‘isEven()’. But since PC do not fragment into their single fatty acids in positive mode, their individual length cannot be retrieved.

```
REPORT
    MASS = prPC.mass;
    NAME = "PC [%d:%d]" % "((prPC.chemsc - headPC.chemsc)[C] - 3,
prPC.chemsc[db] - 1.5)";
    CHEMSC = prPC.chemsc;
    ERROR = "%dppm" % "(prPC.errppm)";
    INTENS = prPC.intensity;
    FRAGINTENS = headPC.intensity;;
```

The last section REPORT defines which kind of information and in which format it should be reported as result. The result is always written in a table and stored in *.csv file format. The content and set by the user in the REPORT section. It includes annotation of recognized lipid species, the abundances of characteristic ions for subsequent quantification and additional information pertinent to the analysis, such as masses, mass differences (errors) etc. (a whole list of possible attributes can be found in Appendix 10.6). In this example LipidXplorer was advised to report five columns:

- NAME – the lipid species name
- MASS – the *m/z* value of the ion
- CHEMSC – the chemical sum composition of the lipid
- ERROR – the difference between the calculated and the measured mass
- INTENS – the abundance of the intact ion reported for each individual acquisition
- FRAGINTENS – the abundance of the signature fragment ion

The NAME of the species is made by string formatting: There have to be two

strings separated by a '%' where the left is the text to be printed containing optional placeholders for decimal numbers ('%d'), floating point numbers ('%f') or strings ('%s') and the right string contains a list with attributes which are filled in the placeholders. In this example two placeholders '%d' of the lipid class name "PC [%d:%d]" are filled with the number of carbon atoms and the number of double bonds. The number of carbon atoms is calculated by subtracting 3 carbons for the glycerol backbone from the difference of the sum composition for 'headPC' and the precursor 'prPC'.

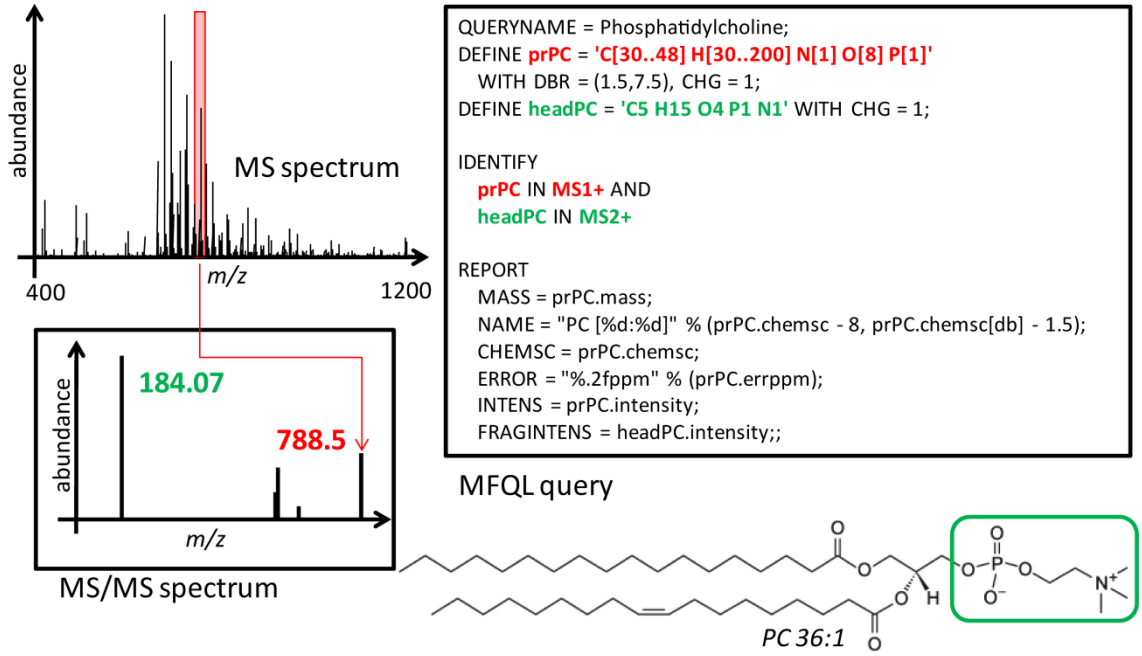


Figure 11: MFQL identification of phosphatidylcholine (PC).

The chemical structure of PC 36:1 is shown on the bottom of the figure. Upon their collisional fragmentation, molecular cations of PC produce the specific head group fragment with *m/z* 184.07 and sum composition 'C5 H15 O4 P1 N1'. The MS spectrum on the upper left was acquired by direct infusion of a total lipid extract into the mass spectrometer (inset). All detectable peaks were subjected to MS/MS. The spectrum acquired from the precursor *m/z* 788.5 (designated by arrow) is presented below. The precursor ion was isolated within 1 Da mass range and therefore several isobaric lipid precursors were co-isolated for MS/MS and produced abundant fragment ions unrelated to PC. These ions were disregarded by this MFQL query and did not affect PC identification. On the upper right is the MFQL query for identifying PC species, details are provided in the text. The peaks originating from the PC ion are colored in red for the precursor mass and green for the PC head group. (Source: own design)

5.5 Second example: The identification of phosphatidylethanolamine (PE) lipid species with MFQL in negative ion mode

PE is another major component of biological membranes and its structure typical for numerous lipid classes. PE consists of a glycerol backbone to which the phosphoethanolamine head group and two fatty acid moieties are attached *via*

phosphoether and ester bonds, respectively.

In contrast to the fragmentation of PC in positive mode (see Chapter 5.4), PE produces two acyl anions of its fatty acid moieties fragmented upon its collision-induced dissociation if the mass spectrometer is switched to negative mode. With MFQL these fatty acids can be accommodated to determine the lipid species:

```
QUERYNAME = Phosphatidylethanolamine;
DEFINE PR = 'C[33..49] H[50..100] O[8] N[1] P[1]' WITH DBR =
(2.5,9.5), CHG = -1;
DEFINE FA1 ='C[12..22] H[20..50] O[2]' WITH DBR = (1.5,7.5),
CHG = -1;
DEFINE FA2 ='C[12..22] H[20..50] O[2]' WITH DBR = (1.5,7.5),
CHG = -1;
```

In the DEFINE section the precursor plus the two fatty acid moieties are specified. In both cases sc-constraints were used to query for all possible fatty acid combinations which vary in number of carbon atoms, hydrogen atoms and number of double bonds.

```
IDENTIFY
    PR IN MS1- AND
    FA1 IN MS2- AND
    FA2 IN MS2-
```

The IDENTIFY section now delegates LipidXplorer to find the intact ion (‘PR’) together with two fragments fitting into the sc-constraints of ‘FA1’ and ‘FA2’.

```
SUCHTHAT
     FA1.chemsc + FA2.chemsc + 'C5 H11 O4 N1 P1' ==
PR.chemsc
```

This example shows the flexibility of the SUCHTHAT section: Mathematical equations are used to assert that the sum of both fatty acid moieties plus the lipid specific head group (‘C5 H11 O4 N1 P1’) fits the sum composition of the precursor ion. Without this statement all fatty acyl chains which are in the fragment spectrum but originate from other, isobaric lipid species would be wrongly associated to PE. In this way it is ensured that only the right fatty acids are associated with the lipid.

```
REPORT
```

```
    MASS = PR.mass;
    CHEMSC = PR.chemsc;
    ERROR = "%2.2fppm" % (PR.errppm);
    NAME = "PE [%d:%d]" % (PR.chemsc[C] - 5, PR.chemsc[db] -
2.5);
    SPECIE = "PE [%d:%d / %d:%d]" % (FA1.chemsc[C],
FA1.chemsc[db] - 1.5, FA2.chemsc[C], FA2.chemsc[db] - 1.5);
    INTENS = PR.intensity;

    FASINTENS = sumIntensity(FA1.intensity, FA2.intensity);;
```

The REPORT section is similar to the PC query example. Additionally there is a SPECIES column reporting the lipid name and the length of the individual fatty acid moities as well as the number of double bonds. The FASINTENS column uses the function ‘sumIntensity()’ to sum the abundances of both given variables. This is useful for the quantification of fatty acids. In particularly it considers if both fatty acids are different or the same, because if they are the same, they must not be summed. If a lipid has two times the same fatty acid (e.g. a PE[16:0/16:0]) both produce a single peak in the mass spectrum. If they would be summed, their intensity would be doubled which would be wrong.

A screenshot with the result of a PE lipid annotation from an E.coli sample can be found in Appendix 10.10.

The distribution of LipidXplorer is supported by its own Wiki page[1] which contains whole libraries of lipid queries. Thus, users can share and download queries for lipids of interest and use them without writing it from the scratch. Details can be found in Chapter 6.10.

[1] https://wiki.mpi-cbg.de/wiki/lipidx/index.php/Main_Page

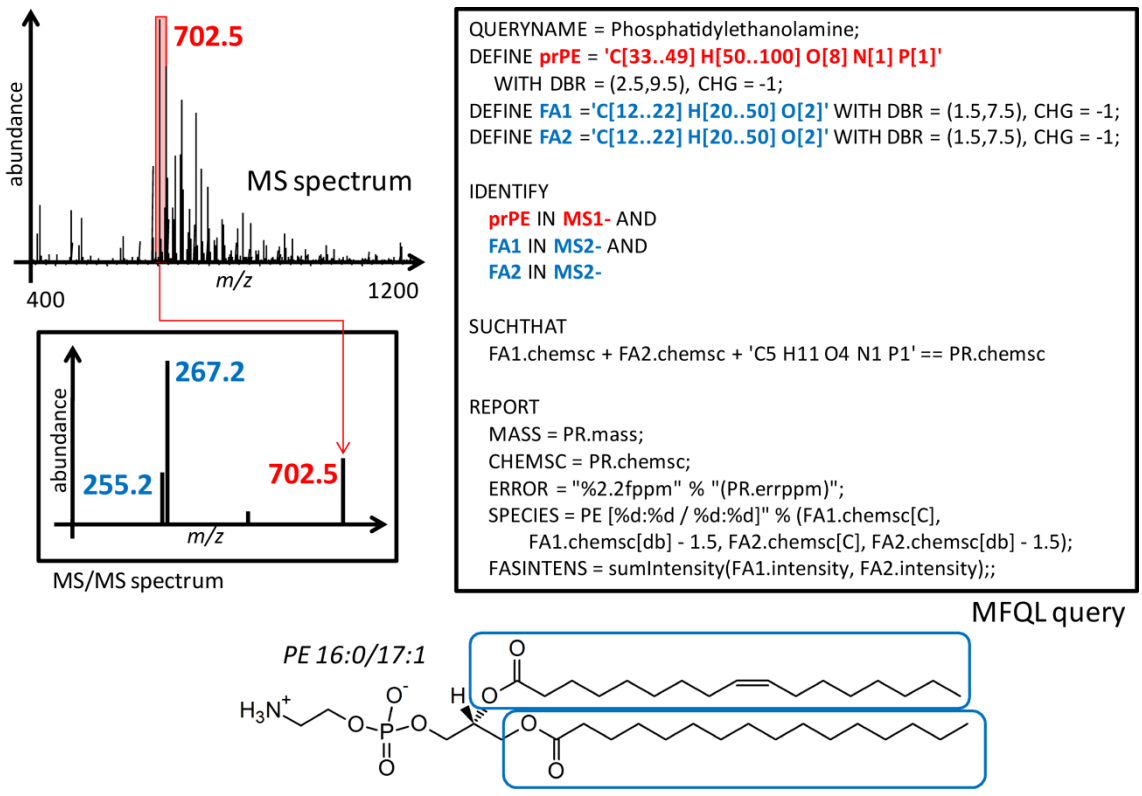


Figure 12: MFQL Identification of phosphatidylethanolamine (PE) species.

PE consists of a glycerol backbone to which the phosphoethanolamine head group and two fatty acid moieties are attached *via* phosphoether and ester bonds, respectively. The chemical structure of PE 16:0/17:1 is shown at the lower panel with the two fatty acid moieties (16:0 and 17:1) highlighted. Its molecular anion with *m/z* 702.5 was detected in the survey MS spectrum and then MS/MS spectrum was acquired. The latter was dominated by abundant acyl anion fragments (*m/z* 255.2 and 267.2) originating from corresponding fatty acid moieties. To identify PE molecules, MFQL first defines the obligatory sum composition constraints (sc-constraints) for their intact masses (prPE) and characteristic fatty acid fragments (FA1 and FA2). No specifically expected masses, but only compositional constraints are defined in the query. In the 'IDENTIFY' section the query requests that, for a given precursor peak (in MS1-) and a pair of fragments (both found in the MS2- spectrum acquired from this precursor) all three above DEFINED compositional constraints have to simultaneously met - "SUCHTHAT" masses of both fatty acids and of the phosphoethanolamine head group complement the mass of intact precursor. The 'REPORT' section describes the format of data output. (Source: (Herzog, et al., article in press))

6 The implementation of the MasterScan and MFQL into the software LipidXplorer

The concepts of the MasterScan and MFQL were implemented into the software LipidXplorer. It is a fully functional software tool kit equipped with an intuitive graphical user interface (GUI) (see Appendix 10.1 for screenshots) guiding the user through its functions.

LipidXplorer is organized in several functional modules (see Figure 13). It starts importing raw mass spectra by averaging individual scans into representative MS and MS/MS spectra. These spectra are further aligned by means of *m/z* of precursor and fragment ions, respectively, and then MS/MS spectra are associated with the corresponding precursor masses and stored in the MasterScan database. Spectral importing routines are instrument dependent and consider common peak attributes: mass resolution and its change over the full range of *m/z*; minimum peak intensity thresholds specified separately for MS and MS/MS spectra; width of precursor isolation window in MS/MS experiments and the polarity mode. LipidXplorer corrects observed masses by linear approximation of the mass shift calculated from a few reference masses effectively enabling a simple offline recalibration of spectral data (if any are detectable in the spectrum). It also pre-filters spectra by user-defined peak intensity and occupation thresholds that are also specified separately for MS and MS/MS modes.

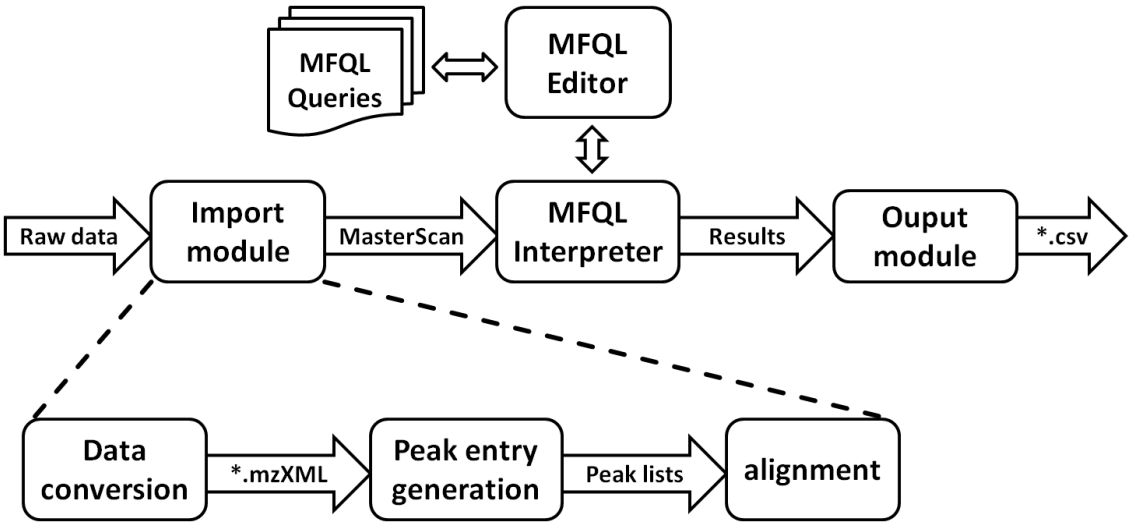


Figure 13: Architecture of LipidXplorer.

Boxes represent functional modules and arrows represent the data flow between the modules. The import module converts technical replicates (collections of successively acquired MS and MS/MS spectra) into a flat file database *MasterScan* (.sc). Then the interpretation module probes the *MasterScan* with interpretation queries written in MFQL. Finally, the output module exports the findings in a user-defined format. All LipidXplorer settings (irrespectively to what particular module they apply to) are controlled via a single graphical user interface. (Source: (Herzog, et al., 2011))

6.1 Validation of the algorithms of LipidXplorer

The scan averaging and alignment algorithms which were presented in this work and implemented into LipidXplorer were validated in several scenarios. The scan averaging algorithm was validated by benchmarking it against another proprietary averaging algorithm. The spectral alignment algorithm was validated by benchmarking it against artificially generated spectra and furthermore, it was validated with the use of real life samples. The results are presented in the following.

6.1.1 Validation of the scan averaging algorithm

The performance of the scan averaging algorithm was compared with the scan averaging of the proprietary software Xcalibur – a dedicated tool for processing spectra acquired on Thermo Fisher Scientific mass spectrometers and a de facto standard in the field of processing high-resolution spectra. 325 samples of lipid extracts were acquired on a LTQ Orbitrap mass spectrometer with the mass resolution of 100,000. Each acquisition consisted of 19 scans, which were independently averaged by Xcalibur and LipidXplorer. Then, each pair of

averaged spectra of the same acquisition was aligned by means of peak masses. Two masses m_1 and m_2 were considered identical if $|m_2 - m_1| < \frac{m_1}{R(m_1)}$, where mass resolution R =100,000. In the aligned spectra peaks were selected which exceed the intensity threshold of 1%, 0.5% and 0.1% of the base peak intensity to test if the algorithm performance was affected by chemical noise. It is usually assumed that the dynamic range (the ratio of intensities of the most abundant to the least abundant signal) in Orbitrap spectra does not typically exceed 1000-fold (Makarov, et al., 2006) and therefore the peak intensity threshold of 0.1% corresponds to peaks that are at the edge of reliable detection. It is to note that the LipidXplorer averaging algorithm performed well on peaks selected at the lowest threshold: only 7% of peaks mismatched, while the mass differences between the aligned peaks were, on average within 0.3 ppm and their intensities differed by less than 3%. The Spearman rank correlation factors (SRCF) were calculated using the intensities of aligned peaks. The average Spearman rank correlation factors are presented in Table 2. In conclusion, the mixed binning algorithm implemented in LipidXplorer performed equally well as the related algorithm in Xcalibur. See Additional file 1 for the experiment data.

Table 2
Comparison of scan averaging algorithms in Xcalibur and LipidXplorer.

Intensity threshold	1%	0.5%	0.1%
No. of peaks	158.40 ± 23.57	237.62 ± 37.36	736.22 ± 128.71
Mass difference, ppm	0.06 ± 0.09	0.08 ± 0.09	0.30 ± 0.09
Intensity difference, %	0.61 ± 0.87	0.72 ± 0.86	3.00 ± 1.24
Spearman rank correlation	0.99 ± 0.02	0.98 ± 0.02	0.94 ± 0.03
Mismatched masses, %	1.45 ±1.44	2.37 ± 1.57	7.06 ± 2.36

All values are average ± std. dev.

6.1.2 Validation of the spectra alignment algorithm

The spectra alignment algorithm of LipidXplorer aligns spectra by means of the

m/z values within the mass spectrometers resolution, while in contrast to the scan averaging algorithm the peak abundances are preserved and stored in a dictionary. The ideal validation test would encompass a collection of real-life spectra with fully known content, i.e. all ions in the spectra should be known exactly. This is hardly feasible, because the exact content of even commercially available lipid extracts is not known 100%. Thus, even if one would do a hand annotation of the spectra, which would take a great amount of time, it would still be biased. Therefore, the algorithm was validated in two separate tests. In the first test, the peak abundances were effectively disregarded, yet the correct masses were exactly known and the dataset composition was controlled. The second test relied on a compendium of real-life spectra of total lipid extracts having typical distribution and variability of abundances of genuine lipid peaks, along with a large number of background peaks and chemical noise. However, the exact composition of lipid species in each sample was not known.

6.1.2.1 Validation using artificial spectra

An experiment was designed in which several spectra were computationally generated from a template spectrum and aligned in a MasterScan. The *m/z* values of the peaks in the MasterScan were then correlated to the *m/z* values of the peaks in the original template spectrum. The template spectrum was designed such that the distance between the two adjacent peaks with the masses m_1 and m_2 was $\frac{m_1}{R(m_1)}$, where R = 500. Within a mass range of 500 to 945, which covers most of lipid precursors, the template contained 319 peaks that were spaced, on average, by the distance of 1.4 Da. From this template 256 spectra were generated in which masses of peaks were randomly selected from Gaussian distributions having the centroid m and $\sigma = \frac{2m}{R(m)}$, where $R = 100,000$ and m is the corresponding mass from the template spectrum. Under the selected resolution and spacing, the peaks in the simulated spectra did not overlap.

Conventionally, LipidXplorer successively repeats spectra binning three times as

discussed in Chapter 3.2 and Appendix 10.3. However, for this test, LipidXplorer was configured such that peaks were binned one, two and three times, respectively. After importing the spectra, it was anticipated that *i)* all 319 peaks of the template spectrum should be present in the MasterScan and *ii)* occupations of individual peaks through all 256 spectra should mirror Gaussian distribution, if peaks were only binned once. Therefore, it was expected to find 319 peaks with the average occupation of 0.68, since this is the number of peaks falling into the rage of [m- σ , m+σ] of the distribution, which equals the bin size of $\frac{m}{R(m)}$.

And Indeed, it was found that after one–step binning 319 peaks were correctly aligned and had an average occupation of 0.65 (Table 3). The average mass difference between the template and aligned peaks was 0.9 mDa. As expected, repeating the binning a second and a third time substantially improved the binning accuracy (Appendix 10.3 and Additional file 2).

Table 3
Computational validation of the peak alignment algorithm.

No. of binning cycles	Avg. peak occupation	Avg. mass difference, ppm
1	0.65 ± 0.05	1.3 ± 0.8
2	0.87 ± 0.08	1.6 ± 0.7
3	0.97 ± 0.04	0.4 ± 0.4

6.1.2.2 Validation of spectra alignment using real life samples

The previous test assumed that in the aligned spectra no unrelated peaks fall into the same mass bin. However, this is hardly possible in real-life shotgun spectra. Therefore, there was the need to test if the alignment accuracy was affected by the complexity of analyzed lipid mixtures and by chemical noise. To this end, lipid species identified by LipidXplorer were compared in the series of individual spectra and in the spectra aligned within the MasterScan.

Using 128 MS spectra of total lipid extracts of different human blood plasma samples (Graessler, et al., 2009) a MasterScan file was compiled, in which individual spectra were mass-aligned by the alignment algorithm of LipidXplorer.

In parallel, each of these 128 spectra was submitted individually to LipidXplorer and lipid species were identified under the same settings. Then the spectra were aligned by means of the identified species (not by peak masses, as it is the case for the MasterScan). Noteworthy is that in both tests, the intensities of peaks in the individual spectra were preserved. The Pearson Correlation Factor (PCF) between the intensities of peaks of the same lipid species in the same acquisitions was computed, either determined in the raw “as submitted” spectrum (lipids were identified in individual spectra), or aligned within the MasterScan file (lipids were identified by probing the MasterScan). It was anticipated that the accurate alignment of multiple spectra would increase the mass accuracy of each individual peak and improve peak identifications. A total of 218 lipid species was recognized by both methods. Of those, three and six species were not identified in the MasterScan and in the individually processed spectra, respectively. When the Pearson correlation factors of the intensity vectors of the identified lipids from both methods were calculated (see Figure 14) it was found that the correlation factors of 15 lipid species out of the total of 218 fell below of 0.8. Their case-by-case inspection showed that isotopic clusters of three species in individual spectra were altered by background or spray instability. The remaining 12 lipid species were very low abundant and their peak intensities were below 0.1% of the intensities of base peaks in corresponding spectra. This led to the conclusion that, while building a MasterScan, mass-alignment of peaks was, in general, correct. The full test dataset is available in Additional file 3.

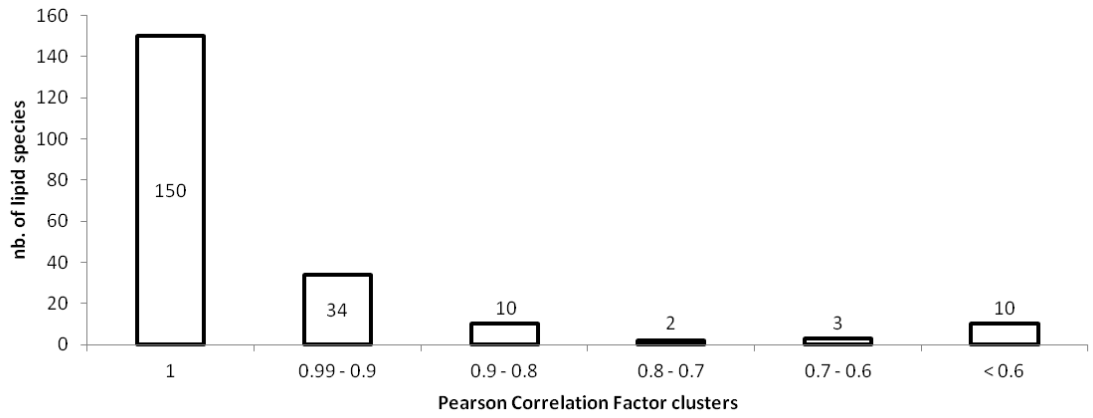


Figure 14: Pearson correlation factors (PCF) of peaks abundances in the MasterScan and individual spectra.

In total, the dataset consisted of 128 high resolution MS spectra of total lipid extracts in which 219 peaks of individual lipid species were recognized. The exact number of peaks assigned to lipid species is provided for each PCF bin. The average PCF calculated for the entire dataset had a value of 0.94 (Source: (Herzog, et al., 2011))

6.2 Isotopic correction of identified lipid species

Isotopes are variants of atoms (see Figure 15) of a particular chemical element, which have the same number of protons and electrons, but differing numbers of neutrons. The number of protons and neutrons in the nucleus is not the same for two isotopes of any element. For example, ^{12}C, ^{13}C and ^{14}C are three stable isotopes of the element carbon (C) with *m/z* values 12.000, 13.0033 and 14.0032. Every isotope occurs with a certain probability. A ^{13}C, for example, has a probability of 0.0107, i.e. every hundredth carbon is a ^{13}C. Only a small number of isotopes have a significant probability. From the common lipid elements – C, H, O, N and P – the probability of their common isotopes is listed in Table 4.

The isotopic distribution is the collection of all isotopes of a chemical

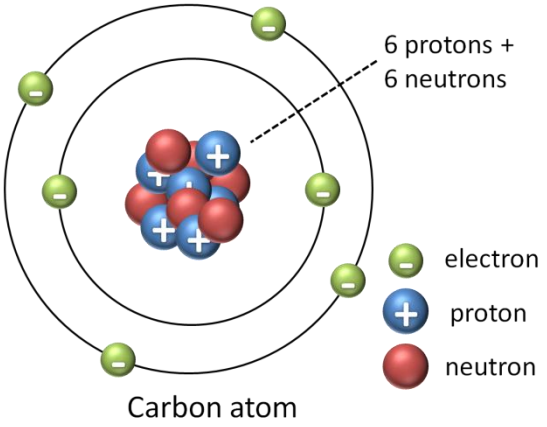


Figure 15: The components of an atom shown with the help of a carbon atom.

The figure shows a carbon atom. It consists of a nucleus of six neutrons and six protons surrounded by an electron cloud consisting of six electrons. Protons are positively charged and electrons negatively, while neutrons are neutral. (Source: own design)

sum formula. The theoretical isotopic distribution of formulas of the main elements occurring in lipids, $C x_1 H x_2 O x_3 P x_4 N x_5$, contains all combinations of the stable isotopes. I.e.: all combinations I_C x I_H x I_O x I_P x I_N where I_C = (Hughey, et al., 2001), I_O = {^{16}O, ^{17}O, ^{18}O}, I_P = {^{31}P}, I_N = {^{14}N, ^{15}N} and I_H = {^{1}H, D (Deuterium)}. Thus, there are $|I_C| \cdot |I_O| \cdot |I_P| \cdot |I_N| \cdot |I_H|$ different isotopes having different *m/z* values.

Given the isotopes probabilities, the isotopic distribution can be estimated with the binomial distribution (Rockwood and Haimi, 2006). The monoisotopic mass is defined as the sum of the masses of the atoms in a molecule using the unbound, ground-state, rest mass of the principal (most abundant) isotope for each element instead of the isotopic average mass[1]. In the following all isotopes different to the monoisotopic peak will be called “isotope”.

Table 4
Probability of common isotopes of the most abundant elements in lipids

Isotope	**Probability**[1]
^{13}C	0.0107
^{2}H/D (Deuterium)	0.000155
^{17}O	0.00038
^{18}O	0.00205
^{15}N	0.00364

[1]numbers taken from: (Böhlke, et al., 2005)

The isotope ^{13}C has the most significant impact on the isotopic distribution of lipids. Only ^{18}O can have a slight influence for lipids if it is present in a very high abundance and its sum composition can contain a greater number of ^{18}O.

The isotopic distribution of elements leads to two effects which need to be considered by any lipid identification software in order to get a valid quantification of molecules. The first effect concerns the abundance of the monoisotopic peak and the second the overlap of isotopes with other lipid species.

[1] http://en.wikipedia.org/wiki/Monoisotopic_mass

6.2.1 The effect of the isotopic distribution on the monoisotopic peak (Type I)

The first effect concerns the abundance of the lipid species. The abundance of a molecule is distributed across its isotopic cluster, which varies depending on the number of elements with a high probability of generating isotopes. As stated before, for lipids these are ^{13}C and ^{18}O. For example, the phosphatidylcholine (PC) species with fatty acyls 16:0 and 18:1 has 42 carbons. This yields a monoisotopic ion at *m/z* 760.586 that represents 61.7 % of the ions created when counting all stable isotopes. The remaining 38.3% appear in the ^{13}C and ^{18}O containing ions. Since the influence of this effect depends on the number of ^{13}C and ^{18}O in the molecule, the monoisotopic peak of a lipid with lower mass has a higher abundance than the monoisotopic peak of a lipid with a higher *m/z* value although the real abundance, which is obtained by counting all isotopes, is the same. The difference can be up to 12% within a lipid class (see Figure 16). The method of correcting this effect was previously called isotopic correction Type I (Han and Gross, 2003).

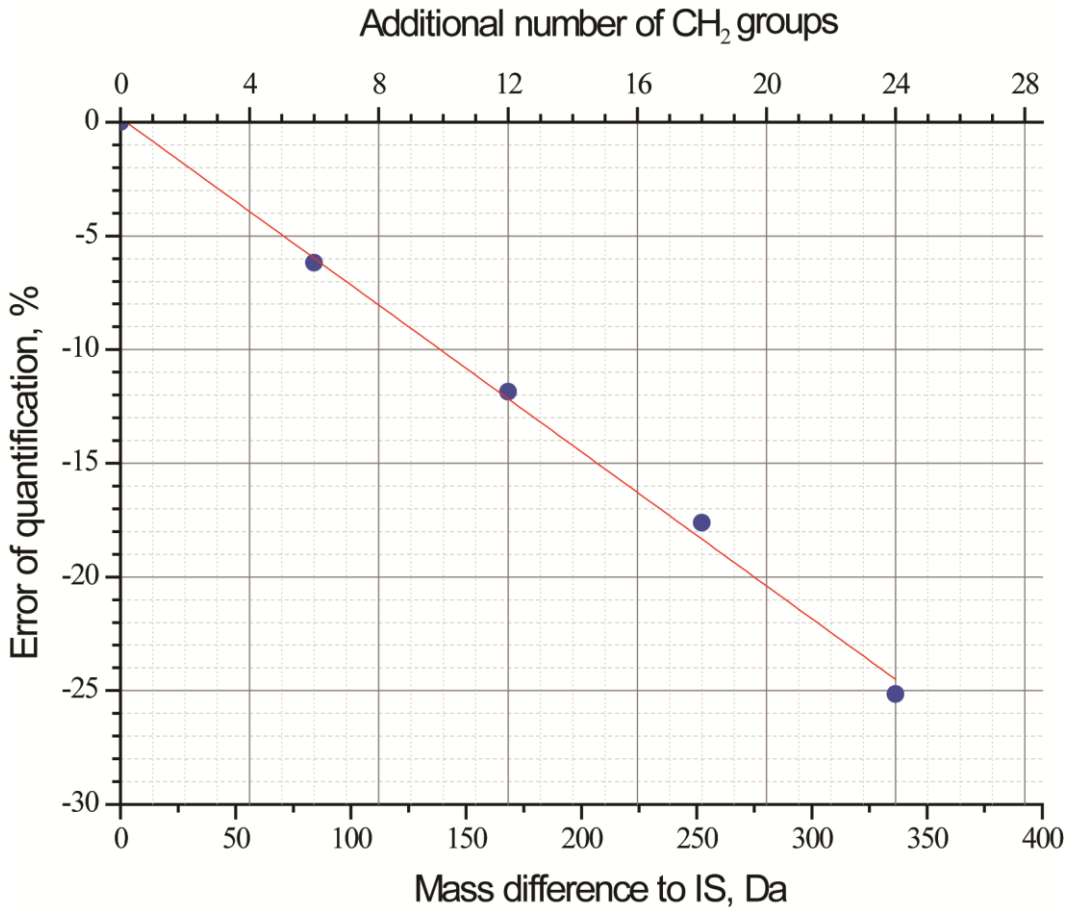


Figure 16: The quantification error if the Type I isotopic correction is not applied.

With higher *m/z* values the relative intensity of a monoisotopic peak decreases if compared to an internal standard. Thus, although the same amount of lipid was spiked into the mass spectrometer, the observed peak intensities decrease, the quantification error increases with higher *m/z*. This figure shows how strong this error increases in steps of CH_2 groups, i.e. with longer fatty acyl chains. The lipid species lactocyl ceramide (LacCer) [d18:1/12:0] differs to its class sibling LacCer [d18:1/24:0] about 12 CH_2 units, thus the quantity of LacCer [d18:1/24:0] is underestimated about approximately 12%. (Source: courtesy by Kai Schuhmann)

6.2.2 The effect of the isotopic distribution on the other lipid species (Type II)

The second effect is the overlap of isotopes on neighboring lipid species (see Figure 17) which occurs especially in low resolution mass spectra. It is also an inherent problem of shotgun lipidomics. Without a previous chromatographic separation, all lipid species from a sample are contained in a single mass spectrum which leads to a high probability for overlapping isotopes. Isotopic overlaps can occur in survey spectra as well as in MS/MS spectra. The lipid phosphatidylethanolamine (PE) [18:2/18:0] for example has the *m/z* value 742.5

and its second isotope the *m/z* value 744.6 in negative ion mode. The neighboring PE [18:1/18:0] has an *m/z* value of 744.6. Having a resolution of 7,500 which makes a tolerance window of 0.1 Da, both peaks overlap and the second isotope of PE [18:2/18:0] increases the intensity of the peak in *m/z* 744.6. The effect is limited in high resolution acquisitions. With acquisitions having a resolution ~100,000 there are no isotopic overlaps in *m/z* ranges <500 while above *m/z* 500 there isotopes overlap spontaneously. This depends on the abundance of the isotope and the overlapped monoistopic peak, because the peak profile is Gaussian-like making the peaks wider on the base. Therefore, it can easily happen that they are merged by the peak detection algorithm into one peak.

Depending on the peak abundance, the isotopic overlap can have a tremendous effect on the quantification. Figure 19 shows a practical example where the isotopic overlap produces a twofold change in the abundance on the overlapped lipid species.

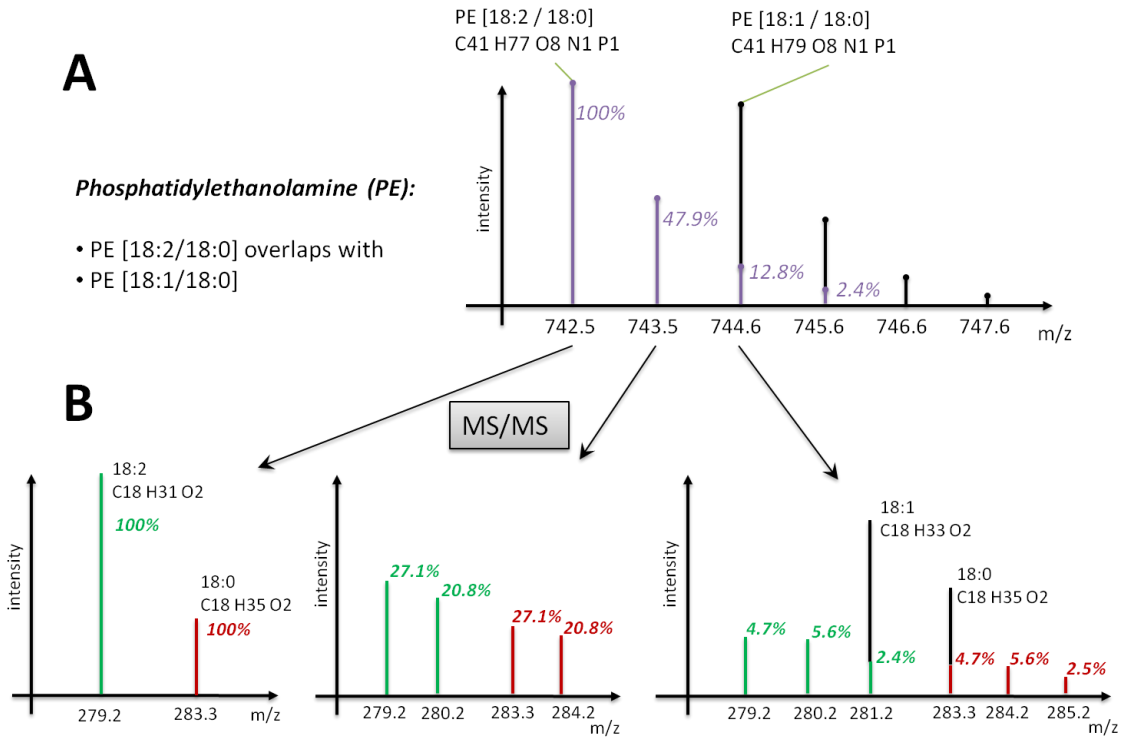


Figure 17: Isotopic overlapping shown on an example of two PE species.

Shown is a theoretical isotopic distribution of two lipid species. (A): The second isotope of Phosphatidylethanolamine (PE) [18:2/18:0] overlaps with the monoisotopic peak of PE [18:1/18:0] and corrupts the peaks intensity. Every isotope is labeled with its abundance. (B): The fragments also contain isotopes and they can also overlap with monoisotopic fragments of neighboring species. The fragmentation produces all possible distributions of isotopes contained in the parent ion. The fatty acid 18:2 of the second of PE [18:2 / 18:0] for example, produces fragments having no isotope (*m/z* 279.2), one isotope (*m/z* 280.2) and two isotopes (*m/z* 281.2), respectively. The last fragment overlaps with the fatty acid 18:1 of PE [18:1 / 18:0]. The isotopic correction algorithm corrects the abundances by subtracting the abundances of the isotopes from the monoisotopic peaks in the MS and MS/MS spectrum. The MS/MS correction algorithm is explained in more detail below in Chapter 6.2.4. (Source: own design)

6.2.3 Type I isotopic correction

The LipidXplorer type I correction virtually sums up the abundance of all isotopes by calculating the isotopic distribution of lipid species due to its chemical sum composition and divides the inverse of the ratio of the monoisotopic peak to the isotopes by the monoisotopic intensity. For example: let I_m be the intensity of the monoisotopic peak. By calculating the isotopic distribution of I_m, it becomes known that I_m contributes $p\%$ to the isotopic cluster. To get its original abundance, I_m is multiplied by $\frac{1}{p/100}$.

The Type I isotopic correction is done after the Type II correction, because if a

monoisopic peak is overlapped with the isotope of another lipid, the result of the Type I correction would be wrong.

6.2.4 Type II isotopic correction

The correction of MS isotopes is done by calculating the isotopic distribution from the sum composition of the identified lipids by using the algorithm described in (Palmblad, et al., 2001) and successively deconvolving the neighboring peaks. LipidXplorer starts with the lipid having the lowest *m/z* value. The calculated isotopic distribution contains the abundances of the isotopes according to the abundance of the monoisotopic peak. After the calculation, LipidXplorer checks if there are neighboring lipid peaks which overlap with one of the isotopes. The overlap is recognized, if the peak is within the tolerance given by the resolution. If yes, it subtracts the intensity of the isotope from the lipid. After processing all isotopes it continues with the same routine on the next lipid with higher *m/z* value until the end of the spectrum. The result is a de-isotoped MS spectrum. Important to note is that this algorithm works more accurately the more lipids are identified, implying that the more queries were run on the MasterScan, the more accurate is the quantization.

The isotopic correction of the MS/MS spectra is more complex, since the fragments of the isotopic peaks contain all possible isotope combinations (as depicted in Figure 17). The main difference to the MS isotopic correction is that the distribution is now 2-dimensional – MS and MS/MS peaks have to be taken into account together. Another step in complexity is that fragments of one isotopic distribution have to be considered through all MS/MS spectra at once. I.e. to calculate the isotopic distribution of the fragment isotopes, the intensity of the monoisotopic fragment is needed which is located in another MS/MS spectrum. This is called interscan isotopic correction (see (Ejsing, et al., 2006)).

The values for the probability of the fragment isotopes are calculated by multiplying the probabilities of the fragment isotopes and the probabilities that the isotope occurs in the fragment's neutral loss. For example, the probabilities for the fragments of the first isotope are distributed over the fragment containing the isotope (termed F_1N_0) and the fragment which does not contain the isotope

(termed F_0N_1). The probabilities in F_1N_0 are calculated by multiplying the probability of the fragment containing the isotope (F_1) and the probability of the neutral loss containing no isotope (N_0). The same is done for F_0N_1. In this way LipidXplorer calculates the isotopic distribution up to the 4^{th} isotope. Insight in this data is given by the LipidXplorer GUI if a sum composition and a fragment sum composition is being input (see Figure A 5).

6.3 Validation of isotopic correction

6.3.1 Validation of Type I isotopic correction

For the validation of the Type I isotopic correction, it was compared with an isotopic correction done by hand. The experiment was done on an LTQ Orbitrap where five triacylglycerol (TAG) lipid species were injected having the same molar ratio and their *m/z* values were spaced such that there was no isotopic overlap. The observed isotopes of the TAG species were summed and added to the monoisotopic peak by hand, respectively. The result was compared to LipidXplorer's isotopic correction algorithm and showed a strong similarity. To further show the impact of the Type I correction, LipidXplorer identified the TAG species with the isotopic correction switched off. The result is shown in Figure 18.

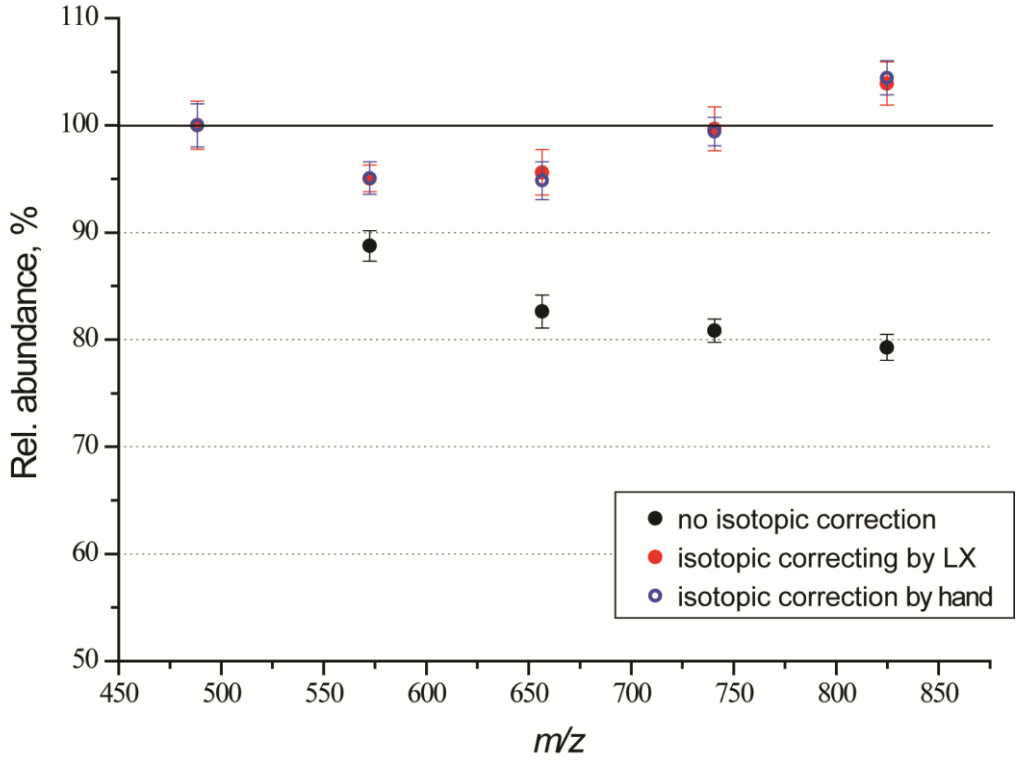


Figure 18: Validation of the Type I isotopic correction by comparison with a correction by hand.

Mass spectra acquired on an LTQ Orbitrap instrument were processed with LipidXplorer without (black dots) and with (red dots) Type I isotopic correction on triacylglycerol (TAG) lipid species. Correspondingly, the abundances of TAG species with higher *m/z* are underestimated. Furthermore, the TAG species were processed by hand (blue dots) by summing up the intensities of all isotopes. A strong correlation between the hand processing and LipidXplorer can be seen. (Source: courtesy by Kai Schuhmann)

6.3.2 Verification of Type II isotopic correction

To test the algorithm, a mixture of four phosphatic acid (PA) standards with a molar ratio 1:9:1:1 was injected into a LTQ Orbitrap Velos mass spectrometer and MS and MS/MS spectra acquired. The two standards PA [18:0/18:2] and PA [18:1/18:1] have the same exact masses and a ratio of 1:9. Therefore, a ratio of 10:1:1 of precursor ion intensities in the MS spectrum was expected. For the quantification of the individual lipid species in MS/MS spectra, the intensities of the fatty acid moieties were summed up and a ratio of 1:9:1:1 was expected (see Figure 19).

The measured molar ratios agreed with the expected. This experiment underscored the notion that the isotopic correction is absolutely required to determine the content of relatively low abundant species. Even at the moderate

dynamic range of 1:9, the abundance of PA 18:0/18:1 would have been drastically overestimated in both MS and MS/MS measurements. A screenshot from the output of LipidXplorer for both modes can be found in Appendix 10.9 and the experimental data in Additional file 4.

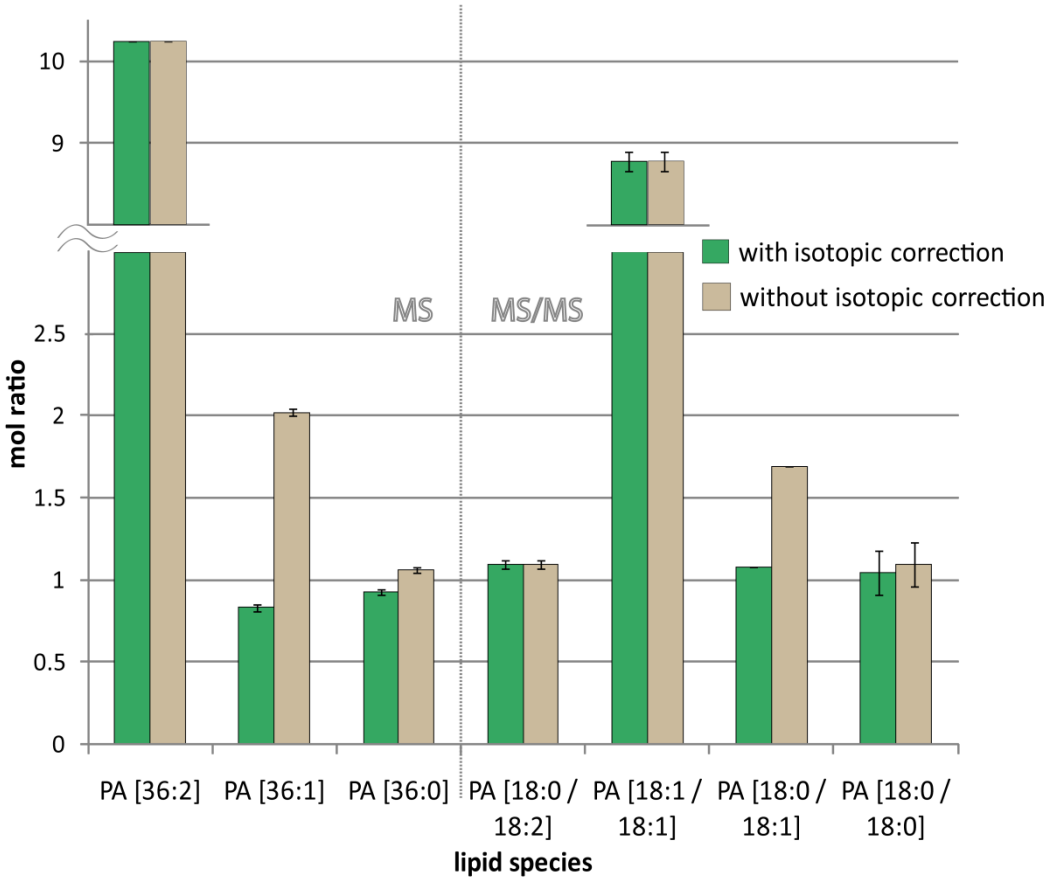


Figure 19: Validation of the isotopic correction algorithm using PA mixture.

Molar ratios of PA standards were determined with (green) and without (grey) isotopic correction of abundances of peaks within partially overlapping isotopic clusters. The molar ratios in the MS spectrum were determined from the abundances of precursor peaks; in the MS/MS spectra as a sum of the abundances of acyl anions of the fatty acids moieties. (Source: (Herzog, et al., 2011))

6.4 Benchmarking the performance of LipidXplorer's lipid identification

The current literature on lipid identification software does not offer benchmarks for lipid identification software. From the knowledge gained by the work underlying this thesis, there is currently no article which tries to verify the published algorithms or compare determined lipid profiles with other software tools. The reason might be that there are no statistical estimates of species

identification confidence which would support lipid identification with a false discovery rate or similar as well as a lack of test-datasets with 100% known content. Therefore, it is difficult to compare different software tools with each other since there is no basic data set on which their results must agree on.
For this reason, a lipid reference list was developed from a test dataset acquired from E.coli total lipid extracts with which the rate of false positive identifications was estimated. In a second step, LipidXplorer's performance of lipid identification was compared with other software tools that support shotgun lipidomics experiments by interpreting peak lists produced from MS and MS/MS spectra.

6.4.1 The generation of a lipid reference list from a complex lipid mixture

The composition of any complex lipid extract of a real sample might not be known exactly and it is therefore difficult to judge if any particular identification is a false positive. On the other hand, an extract of a real sample mirrors the complexity of a shotgun acquisition containing chemical noise, background peaks and isotopic overlaps, the entire challenges shotgun lipidomics software has to deal with. The idea was to identify lipids by hand under very stringent mass spectrometric settings to have a 100% true positive hits reference list. To refine this list, the hand annotated result was intersected with the result from annotating lipids with a third party mass spectrometer using its own software tool LipidProfiler (Ejsing, et al., 2006). Any lipid identification software has to be able to identify the lipids from this list.
The first step was to produce a dataset of MS and MS/MS spectra by analyzing a commercially available total lipid extract of E.coli on an Orbitrap LTQ XL mass spectrometer using data-dependent acquisition in negative mode. It is known that, due to collision-induced dissociation, molecular anions of glycerophospholipids produce abundant acyl anions of their fatty acid moieties that enable an unequivocal identification of individual molecular species (Ekroos, et al., 2002). The glycerophospholipidome of wild type *E.coli* comprises bulk quantities of PE and PG and minor amounts of PA, which should be identifiable with any available software. Also, *E.coli* does not produce lipids with polyunsaturated fatty acid (PUFA) moieties (Geiger, et al., 2010; Kikuchi, et al.,

2000; Oursel, et al., 2007). Thus, species of other glycerophospholipid classes (such as, Phosphatidylinositol (PI) and Phosphatidylserine (PS)) or any species containing PUFA, if identified by the software, will likely represent false positives. Cardiolipins, another major component of the *E.coli* lipidome, could be detected as both singly and doubly charged molecular anions, which might lead to inconsistent interpretations of both MS and MS/MS spectra by different software. The identification of cardiolipins was therefore deliberately omitted from the benchmarking protocol.

As stated above, the lipid composition of the standard *E.coli* extract was first obtained by determining a list of species by manual interpretation of spectra acquired at a LTQ Orbitrap XL machine with, high mass resolution of 100,000 and 15,000 (FWHM, *m/z* 400) in MS and MS/MS modes, respectively, which allowed the imposition of stringent constraints for matching both, precursor and fragment peaks. In this way, 38 lipid species of PE, PG and PA classes were identified.

Next, the same extract was independently analyzed by the method of multiple precursor ion scanning (MPIS) on a quadrupole time-of-flight mass spectrometer (Ejsing, et al., 2006). The interpretation of the MPIS dataset by LipidProfiler software confirmed 36 species representing 95% of the species identified manually. The intersection of species identified by manual interpretation of high resolution spectra and by MPIS / LipidProfiler was taken as the reference list.

Within this list, 78%, i.e. only a fraction of lipids from the reference list were present in the LIPIDMAPS database (Table 5), which is recognized as one of the most complete databases. It should be underscored that, while compiling the reference list, it was aimed at providing the most conservative minimalistic estimate of the lipid composition, *i.e.* only species that *must be* identifiable by any further software tests and any other lipid identification software were included. This does not imply that PE and PGs species other that in the reference list are necessarily false positives.

6.4.2 Testing lipid identification tools against the reference list

In summary, the software benchmarking procedure relied upon the following

rationale: the rate of false negative identifications was estimated by comparing the software output to the reference list and the rate of false positive identifications was estimated by forcing the software to identify species from lipid classes that are not produced by *E.coli*. For the latter test, only the lipid classes whose precursors readily produce molecular anions and whose masses might overlap with precursors of genuine *E.coli* lipids (PE, PG, PA) in low resolution mass spectra were considered. Although LipidXplorer could have restricted the search space by sc-constraints and thus, reduce the expected rate of false positives, for having a real comparison with the other tested programs it was set to report hits with fatty acid moieties having up to 22 carbon atoms and up to 6 double bonds.

A separate dataset was acquired in 8 technical replicates from the same *E.coli* extract under the low mass resolution of 800 for both MS and MS/MS modes that is common for triple quadrupole or ion trap instruments. This dataset was independently processed by LipidXplorer, LipidQA and LipidSearch (Table 4). LipidQA and LipidSearch could only process each replicate independently. Therefore, their output was aligned by means of the reported lipid species. Species identified in less than four (out of the total of 8) replicates were discarded. The same criterion was applied using the occupation threshold of 50% with LipidXplorer.

LipidXplorer produced a total of 53 identifications, which included 36 (100%) species from the reference list plus another 17 species (see Appendix 10.7 for corresponding MFQL queries). According to the above convention, one species was declared as false positive. Both LipidQA and LipidSearch reported fewer species agreeing with the reference list and more false positives (Table 5 and Appendix 10.11). A full list of species identified by all software tools is presented in Additional file 5.

Based on these findings it was concluded that LipidXplorer outperformed the currently available software in interpreting shotgun lipidomics datasets.

Table 5
Benchmarking LipidXplorer identification performance using *E.coli* lipidome

Lipid class	Reference List	LipidMaps[4]	LipidQA[3]	LipidSearch	LipidXplorer
True positives					
PA[1]	0	0	0/1	0/1	0/0
PE	21	18	12/14	14/21	21/27
PG	15	10	8/13	9/17	15/25
Compliance[2], %			56	64	100
False positives					
PS			2	0	0
PI			0	0	0
PUFA species[5]			7	2	1
Total			9	2	1

[1]PA is a very minor (<0.01 mol%) component of the E.coli lipidome.
[2]Compliance is a ratio of the total number of identified species that belong to the reference list to the total number of identified species. It is calculated for species of all three lipid classes together.
[3]The number of identified species is presented as: number of the species that belong to the reference list / total number of identified species. The numbers are presented separately for each class.
[4]The lipid species database is at www.lipidmaps.org
[5]Putative lipids with polyunsaturated fatty acids (PUFA) were searched within for PA, PE and PG lipid classes. As PUFA here fatty acids having more than 2 double bonds were assumed to account even for the rarest instances when a moiety might contain one double bond and one cyclopropane ring.

6.5 Benchmarking the speed of LipidXplorer

The following chapter presents some raw estimations on how much time for spectra import and lipid identification is needed by LipidXplorer. Nevertheless, it depends very much on the kind of spectra which are imported and the kind of MFQL queries used for lipid identification. There is a great difference in the time needed for importing and identification if the spectra contain MS/MS experiments, for example. The number of fragment variables which are used in the IDENTIFY section of a query has also a great influence on the time needed for the lipid

identification. The more variables are used, the slower will be the identification. This is even a polynomial dependency.

First, a comparison of different alignment algorithms, as described in Chapter 3.1.4 is presented. The mixed binning alignment is compared against the heuristic alignment approach as published by (Kazmi, et al., 2006).

6.5.1 Test of LipidXplorer's mixed binning algorithm against the heuristic hierarchical alignment approach

As presented previously (see Chapter 3.1.4), the heuristic alignment of (Kazmi, et al., 2006) seems to be a good middle way between the very slow but very accurate hierarchical clustering and the very fast but in-accurate mixed binning algorithm. In practice, the hierarchical clustering is not necessarily the better choice because the gained advantage of accuracy has no significant influence on the lipid identification ability but has to be bought dearly with time. Both approaches were benchmarked against each other and the result is presented in the following.

Running both aligning methods on a data set and comparing peak by peak afterwards would give no significant result since one cannot know if peaks are aligned right or wrong. Thus, the reference list of lipids from Chapter 6.4.1 was used containing lipids which have to be present in the spectra as a benchmark for correct aligned peaks. If the alignment algorithms would differ substantially, it would be reflected in the lipid profile comparison. The heuristic alignment algorithm of (Kazmi, et al., 2006) was implemented as an optional averaging and alignment algorithm for MS and MS/MS spectra in the import routine of LipidXplorer. Eight replicates of an *E.coli* sample with low and high resolution in negative ion mode were imported, averaged and aligned, respectively. Lipids were identified with the MFQL queries for PE and PG as used in Chapter 6.4. The resulting lipid profiles of both methods were compared with each other. The number of identified lipids was 41 for the heuristic algorithm to 42 for the mixed binning algorithm. The one lipid which did not match (PE [17:1/18:0]) was of low abundance and is not listed in the lipid reference list. However, the mixed binning algorithm is a "greedy" algorithm and leads sometimes to a slightly wider

interpretation of the bin size. Otherwise, the profile revealed no differences in the quality and quantity of the identified lipid species (see Table 6). Such, it could be concluded that the mixed binning algorithm has the same accuracy as the heuristic hierarchical alignment algorithm.

Table 6
The differences of the averaged abundances of each individual lipid species from both methods were summed and set in ratio to the totality of all lipid species

	Difference between the abundances in MS	Difference between the abundances in MS/MS
Low resolution (R = 7500 for MS and MS/MS)	0.006 ‰	0.3 ‰
High resolution (R = 100000 for MS and 15000 for MS/MS)	0.0 %	0.02 %

In Table 7 the time needed to import the spectra with the mixed binning algorithm and the heuristic algorithm is listed. Apparently, the mixed binning algorithm is on average at least five times faster than the heuristic algorithm. The big difference of the needed time for the acquisition with the linear ion trap (E.coli LIT/LIT) can be explained with the quite bigger amount of peaks falling into one bin. The mixed binning algorithm is not affected by this, since it just collects all peaks which are in one bin. The heuristic alignment on the other hand needs to do a hierarchical alignment much more often, which is much more time intensive.

Table 7
The import time of spectra acquired with different resolution to compare the heuristic with the mixed binning algorithm

Sample[2]	Nb. of samples	Time with mixed binning algorithm (in sec)[1]	Time with heuristic alignment (in sec)[1]	Mixed binning algorithm is x times faster than heuristic algorithm
E.coli 100,000/15,000	8 samples	0:18	1:26	4.7
E.coli 100,000/LIT	4 samples	0:28	2:33	5.5
E.coli 30,000/LIT	4 samples	0:10	1:13	7.3
E.coli 7,500/LIT	6 samples	0:09	0:51	5.7
E.coli LIT/LIT	8 samples	0:14	18:56	81.1

[1]On an Intel Core 2 Duo CPU (T9300; 2.50 GHz) computer under Windows 7.
[2]Samples are the same as used for the benchmarking of different lipid identification tools (see Chapter 6.4). The entry states the sample name and the resolution for MS and MS/MS mode at which the samples where acquired. The resolution of the acquisition with the linear ion trap (LIT) is unit resolution, i.e. a resolution of one Dalton.

It could be shown that the heuristic alignment does not outperform the mixed binning algorithm. The quality of the result is the same with both algorithms with the difference that the heuristic alignment algorithm needs much more time to perform and might not be feasible for large projects, especially because the hierarchical alignment has a polynomial complexity.

6.5.2 Benchmarking the speed of LipidXplorer

LipidXplorer was used in (Raa, et al., 2009) to determine all lipids which were affected by the uptake and intracellular transport of Shiga toxin in HEp-2 cells. 32 samples given in *.mzXML format were imported into LipidXplorer each consisting of 55 MS and 110 MS/MS scans. It took 59 sec on an Intel Core 2 Duo CPU (T9300; 2.50 GHz) computer operating under Windows Vista. The total size of the *.mzXML files was 45MB, whereas the size of the produced MasterScan file was only 3.35MB. The identification of species of 6 lipid classes (PC, PC-O, PE, PE-O, SM and TAG) required 59 seconds with LipidXplorer.

LipidXplorer was used in (Graessler, et al., 2009) and revealed an ether lipid deficiency in blood plasma of hypertensive patients. A subset of the used blood plasma samples was used to test how the processing speed of LipidXplorer

depends on the spectra dataset size. Therefore, mzXML files totaling 168 MB that comprised 248 acquisitions in MS only mode were imported, each containing about 2400 peaks. Building the MasterScan file took 13 min at the same desktop PC as used above and required 0.7 GB of RAM. Subsequent screening of the 29.1 MB MasterScan file with 16 MFQL queries only required 6.5 sec. It is to note that a MasterScan is only built once from all spectra acquired in the project. Further interpretation of the dataset, including repetitive screening for other lipid class(es) or using alternative signature ions, does not imply changing the MasterScan.

6.6 LipidXplorer supports spectral interpretation of mass spectra acquired with different resolution

The mass resolution and mass accuracy of detected peaks are determined by the used mass spectrometer. In the following it will be shown that LipidXplorer consistently and accurately interprets spectra acquired at different mass resolution and accuracy.

Several independent shotgun analyzes of an *E.coli* total lipid extract were performed on a LTQ Orbitrap XL mass spectrometer under different target mass resolution settings (described as experiments *i*) – *v*) in Materials and Methods 8.2 and 8.4) (Figure 20) and interpreted the datasets with LipidXplorer. Within the series of successive MS experiments the mass accuracy of the Orbitrap analyzer was dependent only on the target resolution R and, therefore, for matching the masses of lipid species, it was assumed that the tolerance at the mass $m = \frac{m}{R(m)}$. The interest for this experiment was in the number of false positive assignments of detected peaks to PE-*O* species that are not produced in *E.coli*, but closely resemble the structure and often have the masses isobaric with abundant PE species. The difference in exact masses of isobaric PE and PE-*O* species is 0.0364 Da. Thus, their peaks can be distinguished in high resolution spectra (Schwudke, et al., 2007; Schwudke, et al., 2007). Since the same sample was analyzed each time and the same precursor and fragment masses were expected, the experiment provided a consistent dataset for benchmarking the

LipidXplorer performance in interpreting spectra acquired on low- and high-resolution instruments.

Diacyl (PE) and alkylacyl (PE-*O*) lipids were distinguished by assigning the correct sum compositions to peaks observed at the mass resolution of 30,000. The related queries can be found in Appendix 10.7.1 and Appendix 10.7.2. The number of false assignments to PE-*O* dropped from 33 at the MS resolution of 7,500 to 10 at the MS resolution of 30,000 that, as expected, distinguished peaks with the mass offset of approx. 0.030 Da. The increase of mass resolution in the MS spectra up to 100,000 further decreased the number of false positives, yet did not eliminate them completely. When the MS/MS mass resolution was also increased to 15,000 and customized to match fragment masses with an accuracy of better than 0.005 Da, the number of false positive assignments dropped to zero (see Figure 20). Thus, it could be demonstrated that LipidXplorer takes full advantage of the high mass resolution and mass accuracy of a hybrid tandem mass spectrometer. But the experiment also showed that averaging and alignment of related peaks in multiple experiments does not compensate for the limited identification specificity of low resolution machines (see Additional file 6).

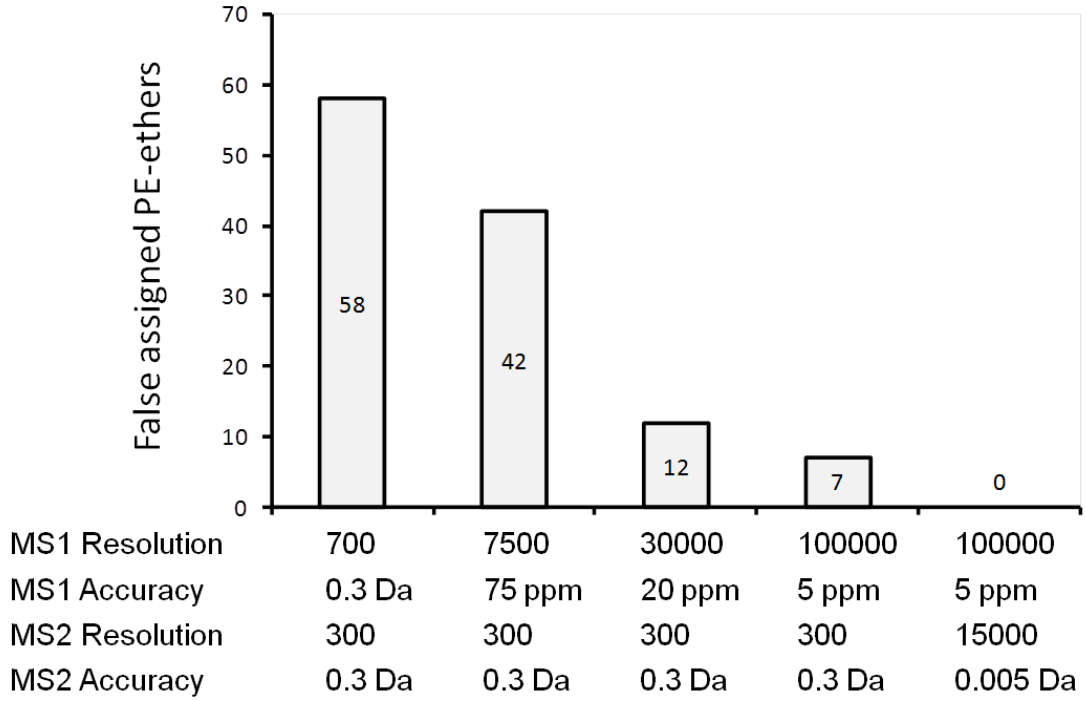


Figure 20: LipidXplorer accurately interprets both high and low resolution mass spectra.

The number of PE-*O* species falsely assigned by LipidXplorer software in shotgun analysis of a total *E. coli* lipid extract under different target mass resolutions. (Source: (Herzog, et al., 2011))

6.7 LipidXplorer supports consistent cross-platform identification of lipids.

By its design and operational principles, LipidXplorer is not tethered to any particular mass spectrometry platform. The program imports shotgun spectra as instrument-independent peak lists or mzXML files. Converters from proprietary formats to mzXML are available for all major instrument platforms. When building a MasterScan, LipidXplorer considers the few generic features of MS and MS/MS spectra, such as mass resolution and mass accuracy, while MFQL allows adapting lipid identification routines to machine-dependent molecular fragmentation pathways. This implies that even if spectra were acquired at different machines and acquisition types (MS or MS/MS), their LipidXplorer interpretation should result in quantitatively consistent profiles. To substantiate this, LipidXplorer's cross-platform performance was validated in two steps. First, it will be demonstrated that lipid quantification by LipidXplorer incorporates established and independent analytical methods, which rely on different

instruments, operation modes and software to ensure that LipidXplorer interpretations are correct. Second, LipidXplorer will be employed for interpreting shotgun datasets of MS and MS/MS spectra acquired at different instruments and demonstrate that it produced quantitatively concordant molecular species profiles.

To this end, a total lipid extract of *E.coli* was analyzed at the LTQ Orbitrap Velos by MS and data-dependent MS/MS. Then, the same extract was analyzed on a quadrupole time-of-flight mass spectrometer QSTAR Pulsar *i* by data-dependent acquisition and also by the method of multiple precursor ion scanning (MPIS), which is a unique feature of QSTAR machines (Ejsing, et al., 2006; Ekroos, et al., 2002). The dataset of MPIS spectra was processed by LipidProfiler software (see Chapter 2.6.2.1). For a better consistency, the mass resolution of the Orbitrap was set at 7,500 such that it was close to the mass resolution of the QSTAR. MS and MS/MS spectra were imported into the MasterScan databases as *.mzXML files and the same MFQL queries for PE and PG lipid species as used in Chapter 6.4 (see Appendix 10.7.1 and 10.7.3, respectively) were applied. 24 major species (15 from PE and 9 from PG lipid classes) were identified and quantified with good signal-to-noise ratios and an occupation threshold setting of 1.0. This was important for a consistent comparison of independent experiments. MS quantification relied on the intensities of intact molecular anions of corresponding species, while for MS/MS quantification the MFQL queries reported the intensities of acyl anion fragments of corresponding fatty acid moieties of each fragmented lipid precursor (Ejsing, et al., 2006; Schwudke, et al., 2006). It was observed that the relative abundances of species quantified in MS and MS/MS spectra acquired at the LTQ Obitrap and QSTAR instruments highly correlated - including the lipid profile acquired by MPIS and analyzed with another software (LipidProfiler) (see Figure 21 and Table 8). Furthermore, the relative abundances of individual species determined by LipidXplorer in MS and MS/MS spectra acquired at different machines and different modes were correlated (for example, Orbitrap MS *vs* QSTAR MS/MS or Orbitrap MS/ MS *vs* QSTAR MS) and compared to profiles acquired at the same machine in different modes (Orbitrap MS *vs* Orbitrap MS/MS or QSTAR MS *vs* QSTAR MS/MS), see Appendix 10.12 and

Additional file 7. In all independent comparisons (Figure 21 and Appendix 10.12) a good correlation of the individual lipid species' relative quantities was observed. Importantly, the slopes of the scatter plots were all close to the value of 1.0 indicating that LipidXplorer introduced no instrument-dependent or method-dependent systematic bias.

In conclusion, this test showed that LipidXplorer processes spectra acquired at different mass spectrometers and by different (MS and MS/MS) methods in consistent and quantitative manner.

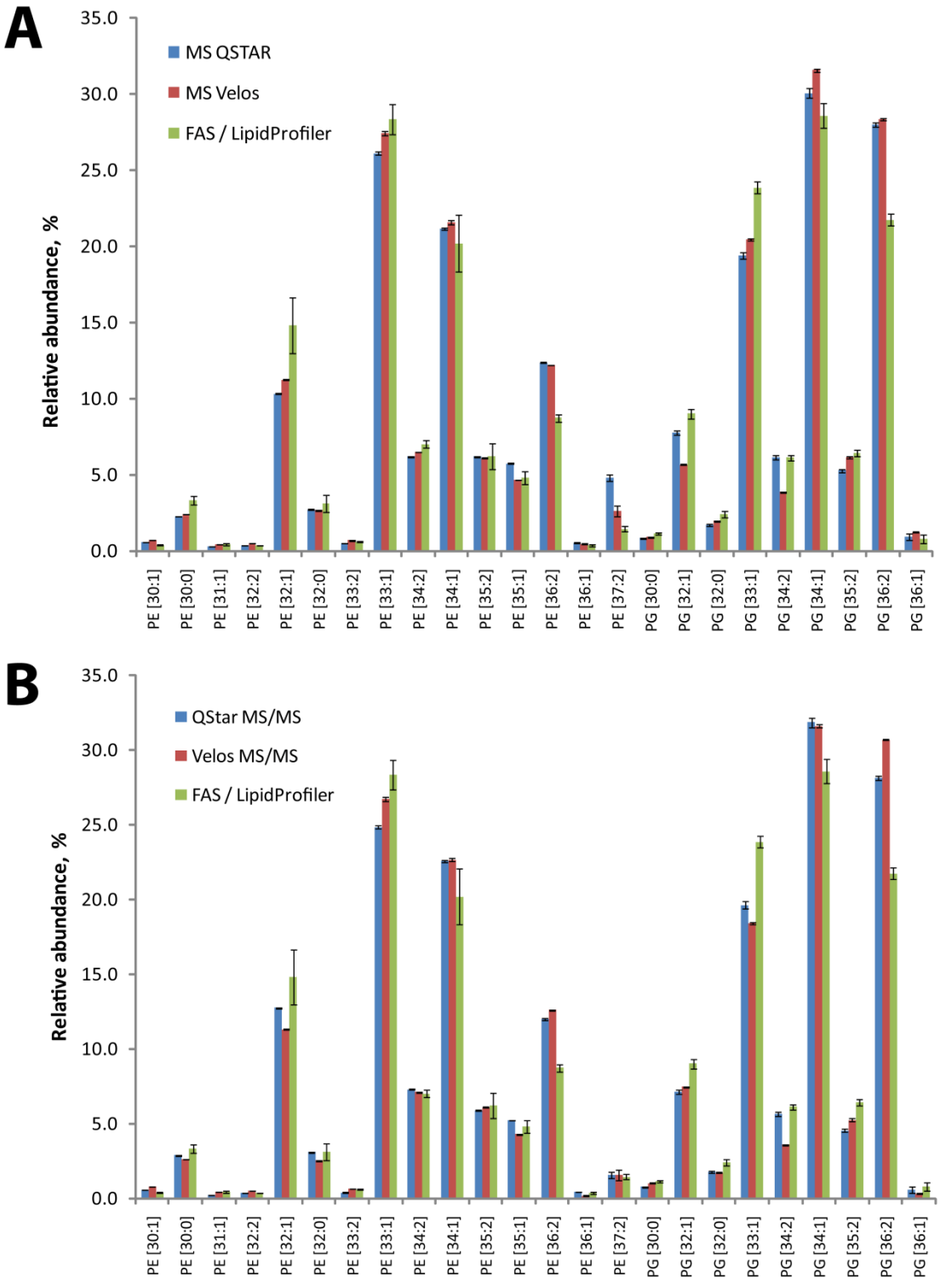


Figure 21:. LipidXplorer supports the interpretation of spectra acquired at different mass spectrometers.

Comparison of the relative abundances of 24 major PE and PG lipid species identified in a total *E.coli* extract in MS (panel A) and data-dependent MS/MS modes at the LTQ Orbitrap Velos (bars in red) and QSTAR Pulsar *i* (bars in blue) mass spectrometers, while spectra were interpreted by LipidXplorer. The same extract was analyzed by multiple precursor ion scanning (MPIS) at the

QSTAR Pulsar *i* and LipidProfiler software (bars in green). The abundances of the individual fatty acid moieties were summed for (B) to have a better comparison with the profile of the precursor abundances (A). Species abundances were normalized to the total abundance of the lipid class; error bars (SD) calculated on the basis of 6 experiments. Correlation coefficients and slopes of scatter plots for each pair-wise comparison are presented in Table 8. Original data can be found in Additional file 7. (Source: (Herzog, et al., 2011))

Table 8
Cross-platform correlation of relative abundances of *E.coli* lipids[1]

Mode	Statistical estimates[2]	Orbitrap *vs* QSTAR[3]		Orbitrap *vs* MPIS QSTAR[4]		QSTAR *vs* MPIS QSTAR[5]	
		PE	PG	PE	PG	PE	PG
MS	Correlation coefficient R^2	0.99	0.99	0.97	0.94	0.95	0.94
	Slope	0.95	1.14	1.0	0.85	1.03	0.89
MS/MS	Correlation coefficient R^2	0.99	0.99	0.97	0.94	0.95	0.94
	Slope	0.96	0.96	1.00	0.85	1.03	0.89

[1] Relative abundances of PE and PG species are presented in Figure 21.
[2] The values of correlation coefficients R^2 and slopes were calculated from corresponding scatter plots.
[3] Species were quantified by LipidXplorer using MS and MS/MS spectra acquired independently at the Orbitrap and QSTAR, respectively. MS/MS experiments were performed in data-dependent acquisition mode.
[4] Species were quantified by LipidXplorer using, respectively, MS and MS/MS spectra acquired at the Orbitrap machine. The relative abundances of individual species were correlated against correspondent relative abundances independently determined by multiple precursor ion scanning (MPIS) at the QSTAR and LipidProfiler software.
[5] Species were quantified by LipidXplorer using, respectively, MS and MS/MS spectra acquired at the QSTAR machine in data-dependent acquisition mode. Relative abundances of individual species were correlated with the ones determined by MPIS and LipidProfiler.

6.8 LipidXplorer supports different acquisition types

Since LipidXplorer was originally developed for mass spectra acquired with data-dependent acquisition, the spectra datasets were structured similarly: they consisted of survey MS spectra and full MS/MS spectra acquired either from

peaks detected in survey spectra, or from peaks whose masses matched the masses from a pre-compiled inclusion list. This factor contributes to LipidXplorer's interpretation consistency irrespectively of the instrument platform. Although data-dependent acquisition is a powerful approach (Schuhmann, et al., 2011; Schwudke, et al., 2007; Schwudke, et al., 2006) it is most efficient on rapid high mass resolution tandem instruments such as hybrid quadrupole time-of-flight (Chernushevich, et al., 2001) or LTQ Orbitrap (Makarov, et al., 2006; Scigelova and Makarov, 2006) mass spectrometers. However, a large body of lipidomics work is performed by triple quadrupole or triple quadrupole - linear ion trap (QTRAP) mass spectrometers (reviewed in (Blanksby and Mitchell, 2010; Gross and Han, 2011; Han and Gross, 2005)) using precursor- or neutral loss scanning (Quehenberger, et al., 2010; Schmelzer, et al., 2007), reviewed in (Pulfer and Murphy, 2003). In this way, no full MS/MS spectra are acquired: the instrument is set to detect one specific fragment originating from all precursors masses within a specified *m/z* range. In each analysis, a fragment mass (in case of precursor ion scanning) or mass difference (in case of neutral loss scanning) is monitored and then the analysis is repeated for the next fragment mass / mass difference of interest. This analysis produces a differently structured dataset to which the generic interpretation with MFQL is not applicable (see Figure 22 (B)).

In the following it will be demonstrated how low mass resolution precursor ion - and neutral loss scan spectra can be interpreted by LipidXplorer and provide evidence that these interpretations are consistent with alternative analyses by data-dependent-driven acquisition of full MS/MS spectra.

A precursor ion scan spectrum is acquired as follows: in centroided mode it plots the abundance of one specific fragment for each mass step (spaced by a certain small increment, usually 0.1 to 0.2 Da) within a specified *m/z* range (see Figure 22). Usually, several precursor ion spectra are acquired in one experiment. In order to have them being prepared for the interpretation with MFQL, LipidXplorer transforms them into the spectra format as it is with data dependent acquired spectra: [precursor mass]: [$frag_1$, abundance], [$frag_2$, abundance], …, [$frag_n$, abundance], i.e. from the format in panel B to the format of panel A in Figure 22. The first step in the transformation algorithm is the alignment of all spectra to

have the precursor masses associated to all their fragment masses (as shown in Figure 22 (B)). Here, a customized version of LipidXplorer's spectra alignment algorithm was used (see Chapter 3.2) which is adjustable to the mass resolution and accuracy of the used mass spectrometer. Effectively, this procedure creates a "virtual" MS/MS spectrum for each precursor mass observed in individual precursor ion spectra: the only difference with a conventional shotgun dataset is that no "survey" MS spectra were acquired. This is emulated by storing the precursor *m/z* values from the PIS spectra. Therefore, upon building a MasterScan from transposed spectra, lipid identification proceeds with MFQL queries in the same way as it would be done for data dependent acquired spectra.

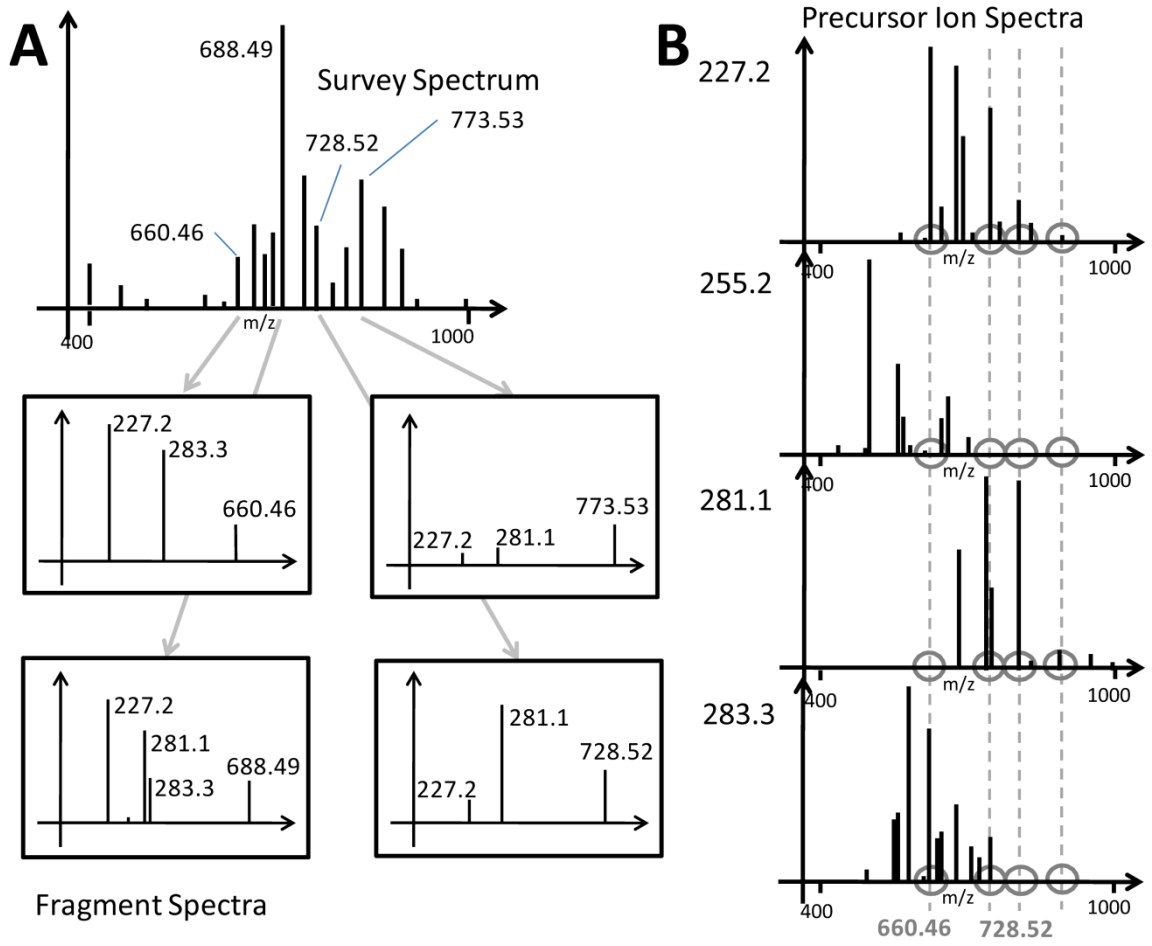


Figure 22: Data-dependent acquisition (DDA)-driven MS/MS and Precursor Ion Scan (PIS) Spectra.

The scheme explains how data-dependent acquisition of full MS/MS spectra (DDA-driven MS/MS) and precursor ion spectra are interrelated. In DDA mode (A) a tandem mass spectrometer first acquires a survey spectrum presenting masses of intact lipids and then fragments all detectable precursors (here the *m/z* values 660.46; 688.49; 728.52 and 773.53 as an example) to acquire full MS/MS spectra (with an *m/z* range from the lowest expected fragment to the *m/z* of the intact precursor). Hence, a dataset of MS/MS spectra comprising all fragment ions generated from all detectable precursors is produced. In the MS/MS spectra (panel A) *m/z* of characteristic acyl anion fragments (*m/z* 227.2; 281.1; 283.3) produced from fatty acid moieties in molecular anions of glycerophospholipids were designated. When operating in precursor ion scanning (PIS) mode, (B) the mass spectrometer registers the intensity of one pre-selected fragment (in this example, the same fatty acid acyl anion) produced from all masses within the *m/z* range of the precursors. In this mode, only precursors producing this specific fragment will produce a peak, while others not. Usually, on most common triple quadrupole mass spectrometers PIS spectra are acquired successively for a large number of fragments (like, all acyl anions of major fatty acids). In the first step of the transformation procedure these spectra are aligned to reveal what fragments out of the expected pool were produced from a particular precursor (dotted line). For example, the lipid with *m/z* 660.46 produced acyl anions with *m/z* 227.2 and 283.3 that correspond to 14:0 and 18:0 fatty acids. At the same time, fragments other than from the pre-selected set, will remain undetected. The scheme exemplifies that DDA-driven MS/MS and PIS produce complementary structural evidence, although they originate from two completely different modes of spectra acquisition. (Source: (Herzog, et al., article in press; Herzog, et al., 2012))

To validate the cross-platform interpretation capabilities of LipidXplorer,

commercial samples of total lipid extracts from *E.coli* were analyzed by shotgun experiments performed in different analytical modes at different mass spectrometers. Upon fragmentation, PE and PG produce abundant acyl anion fragments of their fatty acid moieties that, together with their precursor masses, unequivocally identify the molecular species (Ejsing, et al., 2006; Ejsing, et al., 2009; Ekroos, et al., 2002; Ståhlman, et al., 2009) (see also Figure 12). First, MS/MS spectra from over 120 plausible PE and PG precursors were acquired on a quadrupole time-of-flight instrument QSTAR Pulsar *i* and LTQ Orbitrap Velos using data-dependent acquisition in negative ion mode. Precursor ions were selected with unit mass resolution so that only monoisotopic peaks of plausible precursors were fragmented. Collision energies were optimized as described in (Ejsing, et al., 2006; Schuhmann, et al., 2011) and either ramped with precursor masses (QSTAR) or applied as a "normalized" collision energy (LTQ Orbitrap). For better consistency, spectra were acquired with approximately the same mass resolution of R=7500 for both LTQ Orbitrap and QSTAR. Both experiments were performed in four technical replicates, processed and individual species quantified with the MFQL queries for PE and PG presented in Chapter 6.4.

In parallel, the same extract was infused into the triple quadrupole mass spectrometer Thermo Scientific TSQ Vantage and 72 precursor ion scan spectra were acquired for masses of all acyl anion fragments detected in the experiment above. Precursors were selected within 1.0 Da isolation window, consistently with the experiment settings applied on the QSTAR and Orbitrap. The MasterScan was composed from the aligned and transformed spectra and interpreted by the same MFQL queries identifying lipid species of PE and PG classes as for the LTQ Orbitrap and QSTAR experiments. Quantitative profiles obtained in three independent experiments on hybrid tandem machines QSTAR and LTQ Obitrap and on a triple quadrupole Vantage instrument, were in good agreement had a high correlation as shown in Figure 23 and

Table 9 below and Additional file 8.

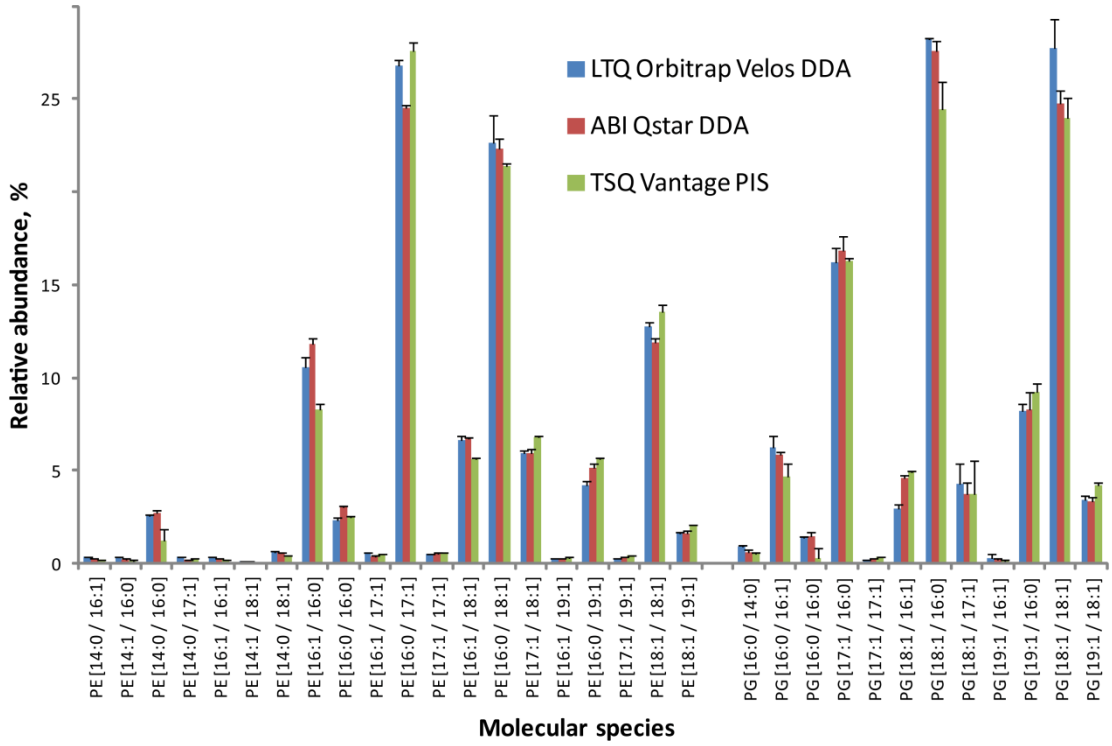


Figure 23: Comparison of the lipid profiles obtained by three independent analytical methods.

Total lipid extract from E.coli was analyzed on the QSTAR and LTQ Orbitrap Velos mass spectrometers in DDA mode and on the TSQ Vantage triple quadrupole mass spectrometer by precursor ion scanning for acyl anion fragments. The same MFQL queries were employed to identify and quantify lipids of PE and PG classes. Relative abundances of individual species were normalized to the total abundance of all species of each class. Error bars represent standard deviations (SD, n=3 for experiments on the TSQ Vantage and n=4 on the QSTAR and LTQ Orbitrap mass spectrometers). Relative abundances determined on LTQ Orbitrap and QSTAR correlated with R^2 and slope of 0.99 and 0.94, respectively; on LTQ Orbitrap and TSQ Vantage: R^2 = 0.98 and slope 0.93; QSTAR and TSQ Vantage R^2 = 0.98 and slope 0.98. (see Appendix 10.12) (Source: (Herzog, et al., 2012))

Table 9
A pair wise correlation of the lipid profiles of *E.coli* acquired on different platforms and modes[1].

Statistical estimates[2]	Velos vs QSTAR[3]		QSTAR vs Vantage[4]		Velos vs Vantage[5]	
	PE	PG	PE	PG	PE	PG
Correlation coefficient R^2	0.99	0.99	0.97	0.99	0.98	0.98
Slope	0.94	0.94	1.04	0.91	0.99	0.86

[1] Relative abundances of PE and PG species are presented in Figure 23.
[2] The values of correlation coefficients R^2 and slopes were calculated from corresponding scatter plots.
[3] Species were quantified by LipidXplorer using MS and MS/MS spectra acquired independently at the Orbitrap Velos and QSTAR, respectively. MS/MS experiments were performed in data-dependent acquisition mode on Orbitrap Velos and QSTAR.
[4] Species were quantified by LipidXplorer using, respectively, MS and MS/MS spectra acquired at the QSTAR machine acquired in data-dependent mode. The relative abundances of individual species were correlated against correspondent relative abundances determined by precursor ion scanning at the Vantage.
[5] Species were quantified by LipidXplorer using, respectively, MS and MS/MS spectra acquired at the Orbitrap Velos machine in data-dependent acquisition mode. Relative abundances of individual species were correlated against correspondent relative abundances determined by precursor ion scanning at the Vantage.

Using the same *E.coli* lipid extract, it was further tested if LipidXplorer could consistently interpret neutral loss scan spectra. Upon collisional fragmentation in positive ion mode, molecular cations of PE and ammonium adducts of PG undergo facile neutral losses of their head groups (Δ *m/z* 141.02 and Δ *m/z* 189.04, respectively), which are conventionally used for their shotgun profiling (Hsu and Turk, 2000; Schwudke, et al., 2006). Shotgun neutral loss experiments on a Vantage triple quadrupole mass spectrometer were performed under basic instrument settings described above, however spectra were acquired in positive mode under CE= 22 eV. Spectra were acquired in three replicas, processed using MFQL queries accounting for the head group neutral losses (see Appendix 10.7.7 and 10.7.8, respectively) and normalized abundances of species compared (Figure 24). For each lipid class, we observed good correlation between the profiles, suggesting that LipidXplorer consistently interpreted both precursor ion scanning and neutral loss scanning spectra.

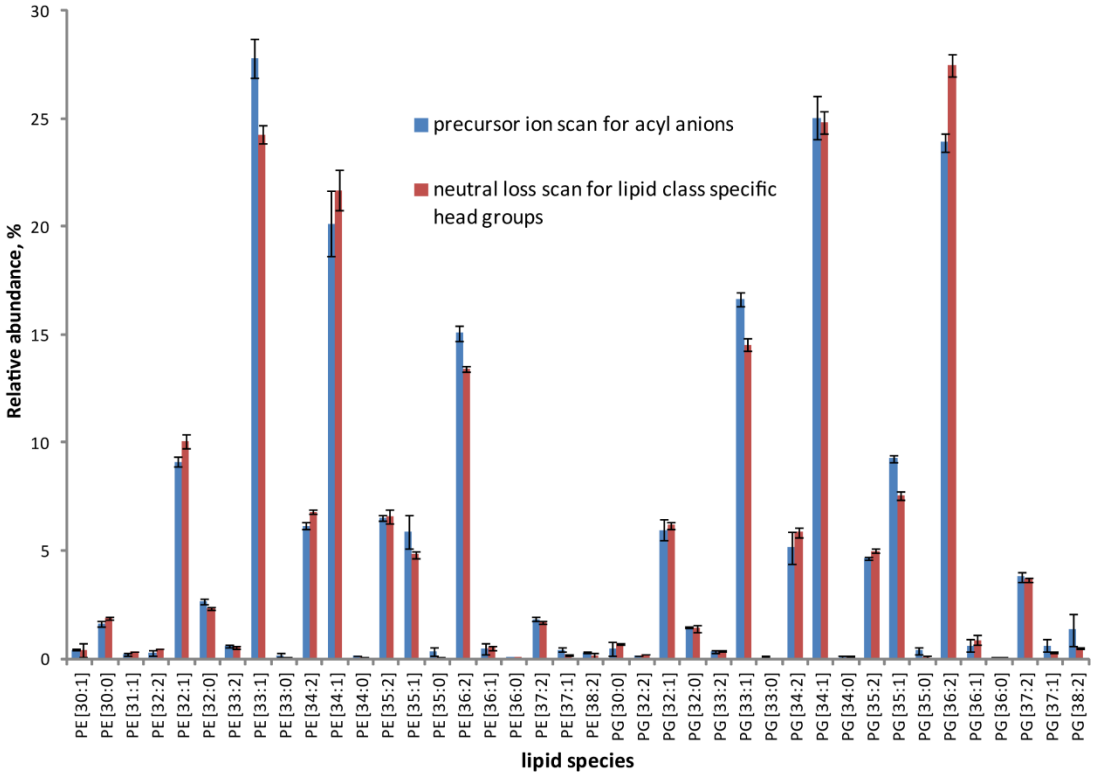


Figure 24: Comparison of the lipid profiles obtained by precursor ion scanning and neutral loss scanning.

Total lipid extract from *E.coli* was analyzed on the TSQ Vantage triple quadrupole mass spectrometer by precursor ion scanning for acyl anion fragments (profiles are the same as in Figure 23) and lipid-class specific positive ion mode neutral loss scanning for the loss of head group (Δ *m/z* 141.02 and Δ *m/z* 189.04 for $[M+H]^+$ molecular ions of PE and ammonium adducts $[M+NH_4]^+$ of PG, respectively). Relative abundances of individual species were normalized to the total abundance of all species of each class. Error bars represent standard deviations (SD, n=3 for experiments on the TSQ Vantage). Relative abundances of species determined on TSQ Vantage by precursor ion scanning and neutral loss scanning correlated with R^2 = 0.98 and slope of 0.94 for PE and R^2 = 0.98 and slope 1.03 for PG (see Additional file 9). (Source: (Herzog, et al., 2012))

6.9 MFQL exploits the diversity of lipid fragmentation pathways.

Lipid identification relies on specific “signature” ions detectable in MS and/or MS/MS mode that, not necessarily unequivocally, distinguish the molecular species from molecules of other lipid classes or of the same class. MFQL allows recognizing many of these ions and /or their combinations simultaneously in each MS/MS spectrum, i.e. the expression capacity of MFQL is accurate enough to identify lipid species of a whole class without interfering with other queries. This is demonstrated in the following: accurate and coherent lipid assignments are employed in parallel to recognize individual species of multiple lipid classes in

total lipid extracts.

A dataset of MS and MS/MS spectra was acquired in 6 technical replicates from a commercially available bovine heart total lipid extract on a LTQ Orbitrap XL mass spectrometer in negative ion mode. Using LipidXplorer software, a MasterScan database was compiled and probed with MFQL queries composed for 19 lipid classes and 188 lipids of 15 classes were identified (see

Table 10). MFQL queries are provided in Appendix 10.8 and the full list of identified species in Additional file 10.

The interpretation of shotgun datasets by LipidXplorer takes advantage of the independent use of several signature ions for each lipid class. If detected at high mass resolution, precursor ions of intact lipids are signature ions themselves. Some lipid classes, such as triacylglycerol (TAG), diacylglycerol (DAG) and cardiolipin (CL) have a unique composition of N, O and P atoms and can be unequivocally identified solely by their intact masses with no recourse to MS/MS as presented in (Schwudke, et al., 2007).

Otherwise, species identification relies on signature ions in MS/MS spectra, such as acyl anions of fatty acid moieties, products of neutral losses of fatty acid moieties, head group fragments *etc*. As an example, it will be demonstrated in the following how multiple signature ions are used to identify molecular species of structurally related phosphatidycholine (PC) and phosphatidycholine-ether (PC-*O*) lipids (see Figure 10). The analysis was performed in negative ion mode in which both PC and PC-*O* were detected as molecular adducts with acetate anions (see Figure 25). Species of both classes have the phosphorylcholine head group attached to the glycerol backbone at the *sn-3* position and the fatty acid moiety at the *sn-2* position (see Figure 10). However, at *sn-1* position ester (PC) species have another fatty acid moiety, whereas ether (PC-*O*) species have a fatty alcohol moiety. To identify PC species, four signature ions could be considered (see Table 10 and Figure 25): the intact precursor ion; a fragment ion produced by a neutral loss of 74 Da that is specific for the head group fragment; and two acyl anions of fatty acid moieties (see Figure 25A). PC-*O* species can be identified (and distinguished from PC) also by four signature ions (see Table 10). Compared to PC, the accurate masses of the intact precursor ion and of the fragment of 74 Da neutral loss should meet different sc-constraints. The third signature ion is the fragment of the neutral loss of the fatty acid moiety and the fourth is the acyl anion of the fatty acid moiety itself (see Figure 22B). At the same time, masses of the fatty acid and fatty alcohol moieties should be complementing the intact precursor mass. The high mass resolution of the Orbitrap mass analyzer allows us to distinguish peaks of intact isobaric PC and

PC-*O* as well as of sphingomyline (SM) (also present in the total extract) and the first isotopic peaks of PC. In MS/MS spectra the peaks of neutral loss products from co-selected PC, PC-*O* and SM precursors could be clearly distinguished.

It is to note that signature ions could be recognized by MFQL queries even if fragments originating from accidentally co-fragmented precursors are also present. A lipid identification based on spectra comparison could easily give a low score in this case (as it is the case in (Song, et al., 2007)), because it would match too few peaks from the spectrum. Users are also fully flexible to choose the signature ions and sc-constraints for species identification and alter MFQL queries accordingly, while the species profiles produced by alternative interepretations remain quantitatively consistent (see Figure 22C).

Probing the MasterScan with correspondent MFQL queries effectively emulated several lipid class specific and lipid species specific precursor ion and neutral loss scans (Ejsing, et al., 2006; Ekroos, et al., 2003; Han and Gross, 2001; Schwudke, et al., 2006) (see Figure 25). Signature ions might be associated with any structural feature of a lipid molecule and the power of MFQL concept is that any of those can be recognized and used for the identification and quantification of individual species. Therefore, the combination of MFQL-assisted interpretation and the organization of shotgun lipidomics datasets in a MasterScan database enables cross-platform, accurate and comprehensive lipidomics analysis of complex biological samples.

Table 10
Multifaceted identification of bovine brain lipid species by LipidXplorer.

Lipid class[1]	Number of identified species	Number of signature ions	FA[2]	FAO[2]	HG[2]	NL[2]	MS[4]
PC	13	4	X X[3]			X	X
PC-O	17	4	X	X		X	X
LPC	4	3	X			X	X
Cer	3	2					X X
CL	10	1					X
LCL	2	2					X X
DAG	13	1					X
PE	22	3	X X				X
PE-O	35	3	X	X			X
LPE	4	2	X				X
PG	10	3	X X				X
PI	13	4	X X		X		X
PS	10	4	X X			X	X
SM	7	2				X	X
TAG	25	1					X

[1]MFQL queries identifying species of these lipid classes are presented in Appendix 10.8
[2]FA – acyl anions of fatty acid moieties; FAO – product of the neutral loss of a fatty acid moiety from sn-2 position of the glycerol backbone; HG – fragment of the head group specific for all species of the same lipid class; NL – neutral loss of a fragment specific for all species of the same class; MS – precursor mass
[3]Two X symbols indicate that two signature ions of the same type are observed – such as, for example, two acyl anions of both fatty acid moieties
[4]Two X symbols indicate that precursor species of this class are detected in two molecular forms (like, deprotonated ion and acetate adduct), or as doubly- and singly- charged ions.

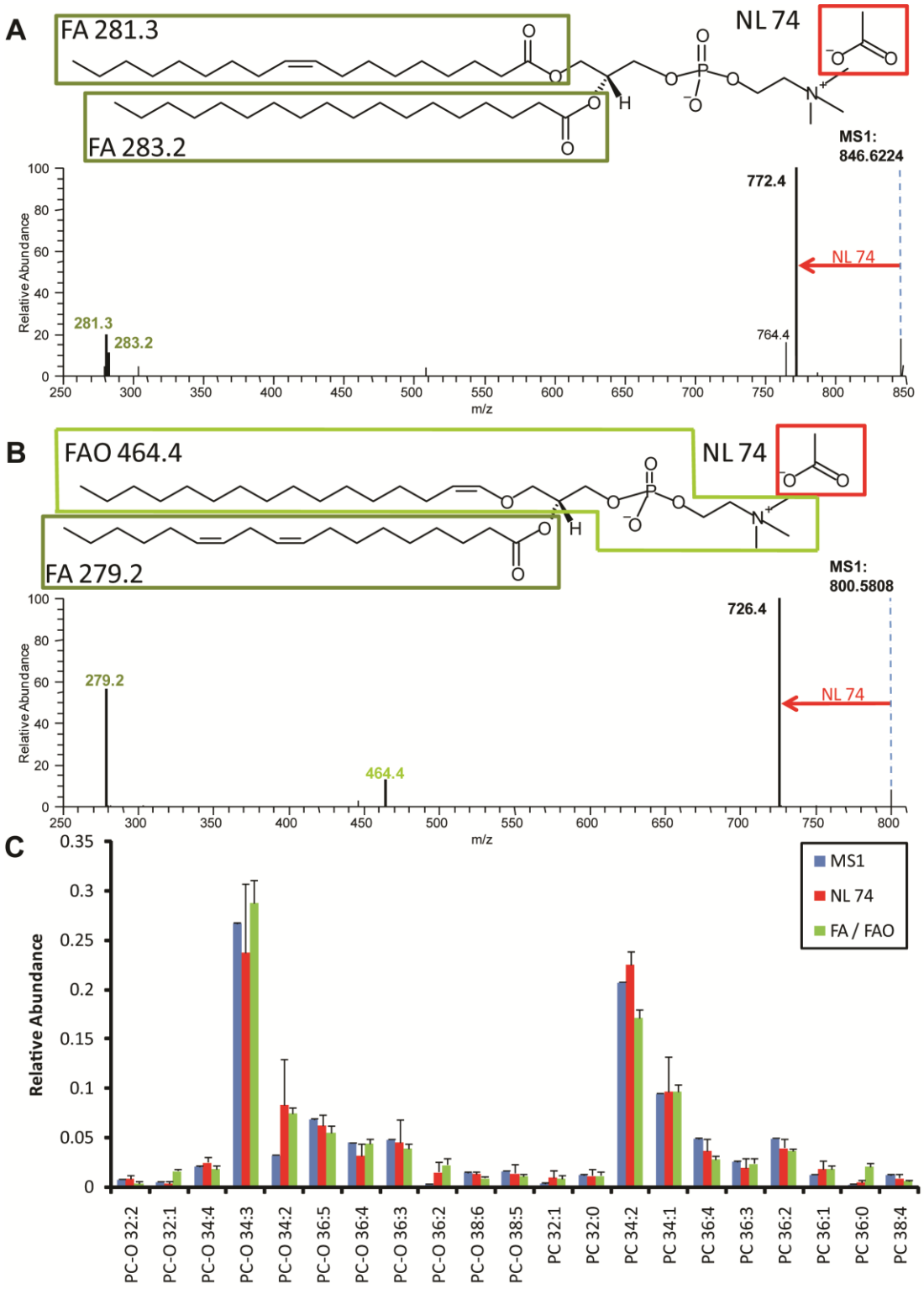


Figure 25: Identification of PC and PC-*O* by MFQL queries relying on complementary signature ions.

A) MS/MS spectrum of the precursor ion of the acetate adduct of PC 36:1 (*m/z* 846.6224), in which four signature ions are recognized: the intact ion (MS); the fragment of the neutral loss of the acetate and methyl group (*Δ m/z* = 74.0 (NL); the acyl anions of the two fatty acid moieties

(FA 281.3 and FA 283.2), both boxed in the chemical structure above.

B) MS/MS spectrum of the acetate adduct of PC-*O* 34:3 (*m/z* 800.5808). Signature ions are the same as in panel A, except *m/z* 464.4 representing the fragment produced by the neutral loss of the *sn-2* fatty acid moiety.

C) Quantitative profiles of PC and PC-*O* species reported from abundances of different signature ions. MS: precursor ions in MS spectra; NL 74: neutral loss *Δ m/z* 74 in MS/MS spectra; FA /FAO: acyl anions of fatty acid moieties and (for PC-*O*) neutral loss of *sn-2* fatty acid moiety. The relative abundance of species was normalized to the total abundance of species within each (PC or PC-*O*) class. (Source: (Herzog, et al., 2011))

6.10 LipidXplorer is available from its own wiki page

When software is released to the public, it has to be maintained. Otherwise it will die after some time. This is especially true for open source software. A lot of smart scientific open source software dies shortly of its publication because there is no support and feedback. The maintenance must not only include regular updates and bug fixes of the software it also should be a forum where users can discuss about problems and ideas and are informed about updates. In this way, the software stays in the focus of the users and the developers yields itself to further developments.

The LipidXplorer project is supported with a wiki[1] site. It includes the full documentation on LipidXplorer, a lipid identification tutorial, a section with frequently asked questions (FAQ), the library of MFQL queries and sample spectra datasets (see Figure 26).

The documentation handles all aspects of the functionality of the software and contains a description of MFQL. Furthermore, a hands-on tutorial is included providing a step-by-step introduction using a provided set of samples and MFQL scripts. It is comprised with a tutorial on MFQL.

All MFQL queries which are used for the presented real-life applications (see Chapter 7.2) are listed in the Wiki together with the citation of the associated articles. This includes the samples and queries used in this thesis. Since it is a Wiki every reader can change its content. Thus, it is a platform on which MFQL queries can be shared. In the long run it could serve as a platform for a growing shotgun lipidomics community. The support for LipidXplorer is completed by a discussion forum based on Google Groups, which is already used by a number of

[1] https://wiki.mpi-cbg.de/Wiki/lipidx/index.php/Main_Page

people. Currently queries for the following lipids are available on the wiki:

- queries for acquisitions in negative mode:
  - Phosphatidylethanolamine (PE)
  - Lyso-Phosphatidylethanolamine (LPE)
  - Phosphatidylglycerol (PG)
  - Lyso- Phosphatidylglycerol (LPG)
  - Phosphatidylinositol (PI)
  - Lyso- Phosphatidylinositol (LPS)
  - Phosphatidylserine (PS)
  - Phosphatic Acid (PA)
  - Lyso-Phosphatic Acid (LPA)
  - Phosphatidylethanolamine ether (PE-O)
  - Phosphatidylcholine (PC)
  - Lyso-Phosphatidylcholine (LPC)
  - Phosphatidylcholine ether (PC-O)
  - Sphingomylien (SM)
  - Ceramide (Cer)
  - Diacylglycerol (DAG)
  - Triacylglycerol (TAG)
  - Cardiolipin (CL)
  - Maradolipids (MLP)
- queries for acquisitions in positive mode:
  - Phosphatidylethanolamine (PE)
  - Lyso-Phosphatidylethanolamine (LPE)
  - Phosphatidylinositol (PI)
  - Phosphatidylserine (PS)
  - Phosphatidylethanolamine ether (PE-O)
  - Phosphatidylcholine (PC)
  - Lyso-Phosphatidylcholine (LPC)
  - Phosphatidylcholine ether (PC-O)
  - Sphingomylien (SM)
  - Ceramide (Cer)
  - Glucolsylceramide
  - Diacylglycerol (DAG)
  - Triacylglycerol (TAG)

Since a single query contains the theoretical definition of roughly 60 to 600 lipid species all the 32 queries cover around 1200 to 12000 lipid species.

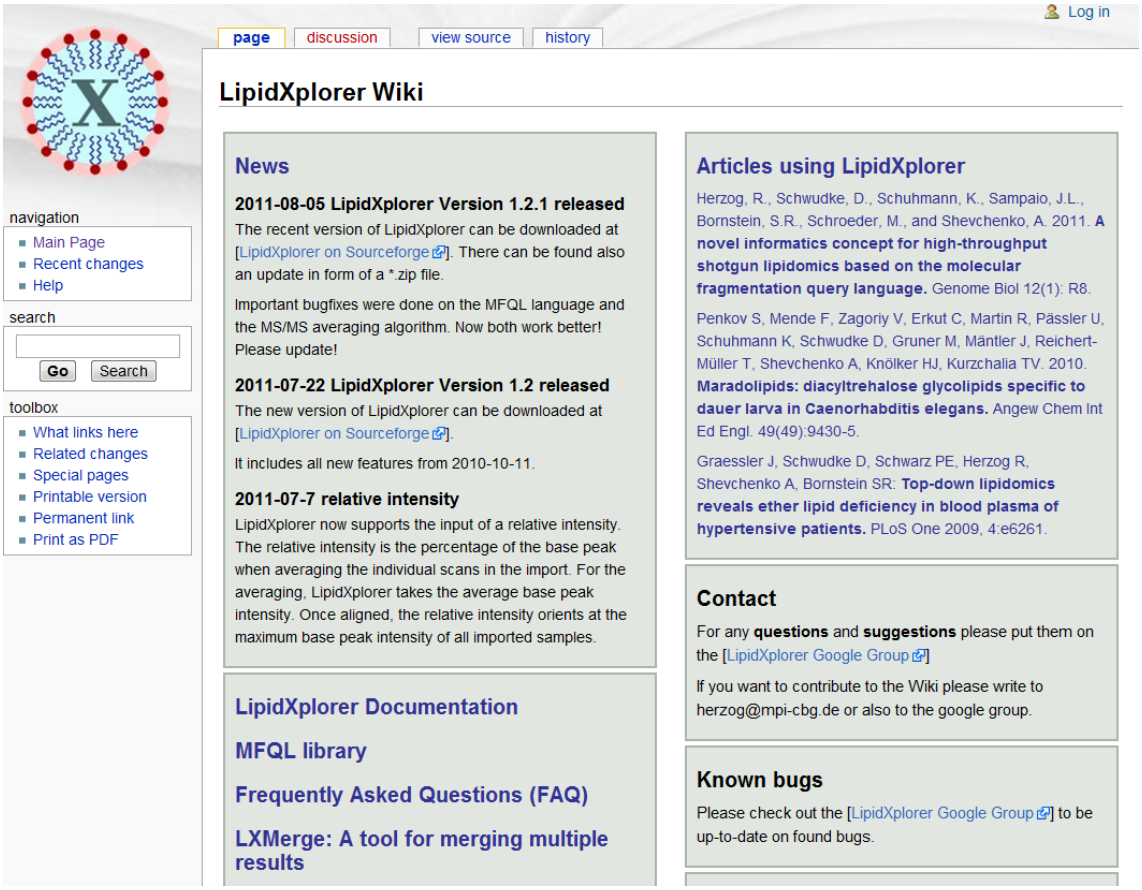


Figure 26: A screenshot of the LipidXplorer Wiki.

The Wiki should be the main resource of everything related to LipidXplorer. It features a News section, all the documentations and tutorials, the MFQL library with queries for numerous lipid classes in acquired in different conditions, a “FAQ” (Frequently Asked Questions), links to the LipidXplorer download site and all articles in which LipidXplorer was utilized (https://wiki.mpi-cbg.de/wiki/lipidx/index.php/Main_Page).

7 Conclusion

This thesis presented solutions for the problem of computer aided lipid identification of mass spectra. Algorithms for spectra averaging and spectra alignment to import mass spectra into a relational flat file data base were developed. The MasterScan allows probing of structural information from numerous mass spectra of entire biological experiments, acquired at any machine in any acquisition mode. If data-dependent acquisition is used, the MasterScan contains the comprehensive set of spectra. Without doing subsequent acquisitions the MasterScan can be probed for other lipid species over and over again. With MFQL a specified query language for lipid analysis was developed. It allows probing for lipid fragmentation pathways and yields a de-novo interpretation of mass spectra to identify lipid classes and lipid species. The presented algorithms and MFQL were implemented into a full featured software kit called LipidXplorer.

The aims from the beginning of the thesis were:

1) The support of spectra acquired on any mass spectrometer with any acquisition method.
2) The support of large scale lipidomics by a fast processing of numerous samples.
3) Customizable and flexible spectra interpretation routines, allow addressing lipids in their complex diversity along with complex shotgun mass spectra.
4) Transparency of the spectra handling and lipid identification to support verification of the result.
5) A comprehensive and accurate isotopic correction method for a correct quantification of lipid species.

The aims were met in the following ways:

1) The support of any mass spectrometer is enabled by a customizable spectra import routine which encompasses spectra variables like accuracy, resolution, signal-to-noise ratio or spectra recalibration. It also enables the import and use of spectra from different mass spectra architectures and acquisitions. With the support of transposing precursor

ion / neutral loss scan spectra, the main shotgun lipidomics acquisition methods are covered.

2) A fast processing of several samples is achieved by a subsequent spectra alignment. Once aligned, lipids in all samples can be detected in one run. Furthermore, it gives an analytical view on the mass spectra peaks by observing the occupation of peaks or having a direct view on the up- and down-regulation of lipid species. If data-dependent acquisition is used, the MasterScan contains the comprehensive set of spectra. Without doing subsequent acquisition the MasterScan can be probed for other lipid species.
3) To get the utmost flexibility and customizability, a custom made query language for probing the aligned spectra for lipid species was developed. This is the first time a query language is used for lipid identification from mass spectra. It allows an utmost flexibility of data interpretation and identification of lipids with their various fragmentation behaviors which vary across the chosen mass spectrometers and acquisition settings.
4) With the customizable import settings and the lipid identification routines defined in MFQL, a high transparency is achieved. All steps, from the mass spectra import through the alignment up to the lipid identification can be controlled and understood by the user. Thus, the user is able to validate the result based on the identification procedure.
5) To have a correct quantification of lipid species, LipidXplorer implements two isotopic correction algorithms. The Type I correction calculates the real amount of a lipid species by comprising the its whole isotopic cluster. The Type II isotopic correction algorithm is especially important for shotgun lipidomics because of the high density of shotgun spectra. Here, the isotopes from identified lipid species which overlap with other identified lipid species are subtracted from the spectrum. The correctness was shown with two experiments (see Chapter 6.3).

Furthermore, it was shown that several queries can be run in parallel without interfering each other and moreover decipher MS and MS/MS spectra of isobaric lipid species. The correctness of LipidXplorer's import algorithms was verified and it was shown with a benchmark that the concepts presented in this thesis

outperform other lipid identification approaches. The benchmark furthermore pointed out that there is lack of completeness in modern lipid data bases like LipidMaps, although it ought to be one of the largest existing lipid databases (Fahy, et al., 2007; Schmelzer, et al., 2007). But it did not contain the lipids which were identified by a stringent hand annotation. The following table below is the same table as presented in the introduction (see Table 1) but with an additional column for LipidXplorer.

Table 11
Comparison of the presented software tools by means of selected features with LipidXplorer

FEATURE[1]	LipID	LIMSA	FAAT	AMDMS-SL	LipidQA	Lipid Profiler	Lipid Search	Lipid Inspector	LipidXplorer
MS + MS/MS (support of DDA)						Yes	Yes	Yes	Yes
Database expandability	Yes	Yes				Yes			Yes
Isotopic correction		Yes		Yes	Yes	Yes			Yes
Cross - platform	Yes	Yes	Yes	Yes	Yes		Yes	Yes	Yes
Support of individual acquisition modes				Yes		Yes	Yes		Yes
Spectra alignment			Yes						Yes
Grouping						Yes			Yes
Batch mode		Yes	Yes		Yes	Yes		Yes	Yes
Customization of identification algorithm									Yes

[1]List of features: **MS + MS/MS,** lipid identification based on parallel introspection of MS and MS/MS spectra, required for identifying individual molecular species; **Database expandability,** users may expand references databases at will; **Isotopic correction,** overlapping isotopic clusters are detected and the intensities of corresponding monoisotopic peaks are adjusted; **Cross-platform**, can process spectra acquired on mass spectrometers from different vendors; Support of individual acquisition modes, supports different acquisition modes like data dependent acquisition (DDA) and precursor ion scan (PIS); **Spectra alignment,** supports alignment of multiple spectra within the series of experiments; **Grouping**, supports grouping of spectra of biological and technical replicas acquired from the same sample; **Batch mode,** supports processing of multiple spectra submitted as a batch; **Customization of identification algorithm,** supports the customization of the lipid identification algorithm to fit the mass spectrometers attributes.

LipidXplorer includes all the features of the current software tools for lipid analysis in one software kit. It is prepared for both scenarios: a high throughput

analysis and an in-depth analytical analysis of mass spectra. The flexibility of MFQL allows the detection of unconventional lipids.
It includes a graphical user interface for an easy access to its rich settings and allows especially non-programmers to have a quick start with the software.

7.1 Discussion concerning the individual aims

In the following, the individual aims stated at the beginning of the thesis are discussed individually.

7.1.1 Discussion about aim 1 – the support of spectra acquired on any mass spectrometer with any acquisition mode

LipidXplorer supports any mass spectrometer by using generic data formats. The most preferred one is the *.mzXML format. Most raw spectra formats can be converted into *.mzXML using third party converters. The problem is that the converters always depend on the vendors libraries which have to provide functions for reading out the actual spectra data. It is not always the case that these functions are provided. For ABI Sciex mass spectrometers, for example, LipidXplorer uses the mzWiff[1] converter to automatically convert ABI Sciex *.wiff files to *.mzXML. Currently, the conversion is supported for spectra acquired in data-dependent mode, but not for spectra acquired in precursor ion/ neutral loss scanning mode. Unfortunately, PIS and NLS are the preferred modes for most ABI Sciex instruments. For raw files from Thermo Scientific mass spectrometers LipidXplorer uses the ReAdW[2] converter and can automatically convert *.raw files to *.mzXML. This works for data-dependent acquired spectra as well as for precursor ion/ neutral loss scan spectra (acquired with the TSQ Vantage, for example).
Standard data formats for mass spectra like *.mzXML (Lin, et al., 2005; Pedrioli, et al., 2004), *.mzData or the recent developed *.mzML (Martens, et al., 2011) get more and more popular, because numerous software tools for processing mass spectra are available nowadays. The project msconvert, a tool from the

[1] http://tools.proteomecenter.org/wiki/index.php?title=Software:mzWiff

[2] http://tools.proteomecenter.org/wiki/index.php?title=Software:ReAdW

ProteoWizard[1] library, tries to encompass the spectra formats of the most used mass spectrometers. Newer versions of LipidXplorer will support this converter and enable the tool to automatically convert the different formats. But also msconvert depends on the support from the vendors software and therewith its limitations.

7.1.2 Discussion about aim 2 – the support of large scale lipidomics

LipidXplorer supports large scale lipidomics by aligning and storing the individual spectra of a biological experiment into the MasterScan. One outcome of the alignment is a single spectrum containing the averaged peaks and the associated intensities as a list. Thus, there is effectively only one spectrum to be probed, no matter how many samples were imported. The back draw of this routine is that although the lipid identification is very efficient, the alignment itself might need quite an amount of time and memory, depending which alignment algorithm was used and how many samples should be imported. As presented in Chapter 6.5.2, the import of 168 MB of spectral data, for example, needed 0.7 GB of memory. The question is, if this heavy consumption of resources mainly results from the use of Python as programming language or from the algorithms itself. Since effort has been put on having a resource-beneficial implementation of the algorithms, it is probably the use of Python which causes this. Since Python is an interpreted scripting language it is naturally slower than a low level compiler, like for example C/C++, because it has a large overhead for the internal data handling. It is known, for example, that the function call overhead is relatively high in Python (see Python speed performance tips[2]). To address this problem and make LipidXplorer more effective, routines which need a lot of resources could be re-written in C/C++ and integrated in the software. This is supported by Python which offers a simple way to write Python modules in C[3]. However, there are also other ways to address this problem. If the computer resources are exhausted because of too many samples to be imported, an alternative would be to not generate a MasterScan but to process each sample individually. This will not be

[1] http://proteowizard.sourceforge.net/

[2] http://wiki.python.org/moin/PythonSpeed/PerformanceTips

[3] http://docs.python.org/c-api/

faster and one would not benefit from the averaging of the individual *m/z* values, but since there is no alignment algorithm needed, the resources requirements are low. This is technically possible by running LipidXplorer from a batch script for each sample individually and merges the results. The tool LXMerge does the merging and is available at the LipidXplorer wiki[1]. As the feedback from users suggests, the processing speed and hardware requirements are an important factor and is further discussed in Chapter 7.3.1.

7.1.3 Discussion about aim 3 – customizable and flexible spectra interpretation

Using a query language for lipid identification from mass spectra offers a great flexibility. Individual fragmentation signatures can be addressed as well as specific parameters from the mass spectrometer platform or the kind of probed sample. To benefit from this, it requires a basic knowledge of MFQL and knowledge about lipid fragmentation pathways. Thus, the user has to make an effort to learn MFQL. It might need some time at the beginning, especially if one does not have any experience with programming languages. On the other hand, if one starts to learn MFQL, one learns about lipid fragmentation pathways in mass spectra in parallel. Everything which is to be known of how lipids are identified in mass spectra is encoded in MFQL. Without reading numerous articles, the user can gather knowledge from MFQL scripts. Many examples can be downloaded from the LipidXplorer wiki, used as templates and can be customized to the specific demands of the utilized mass spectrometer, acquisition mode and biological organism. However, it is also possible to use LipidXplorer without writing MFQL queries. The LipidXplorer wiki offers a library of queries which covers major lipid classes. Queries just need to be downloaded, pasted into LipidXplorer and used for lipid identification. The current list of lipid classes is listed in Chapter 6.10. Whenever available, the MFQL scripts are posted together with citations of the articles which made use of them.

[1] https://wiki.mpi-cbg.de/wiki/lipidx/index.php/LipidXMergingTool

7.1.4 Discussion about aim 4 – transparency of the spectra handling and lipid identification

The transparency of LipidXplorer aids the validation of the results. Since every step of the identification process can be understood, the correctness of the result can be deduced. However, this manual inspection of the acquired spectra is time consuming and biased by the knowledge of the user. It is therefore not applicable to large scale lipidomics efforts. But on the other hand there no statistical framework currently exists which is able to either estimate a global false discovery rate or compute the local probability that a particular assignment is correct. Such a framework would be of a great help. To give an example: The field of proteomics, deals with the identification and quantification of proteins and peptides from mass spectra. Here, scoring functions are common (Eriksson, et al., 2000; Gras, et al., 1999; Perkins, et al., 1999; Sadygov and Yates, 2003). The accuracy of an identified species is estimated with a score calculated from the mass spectral peaks of the numerous peptides or amino acid which are produced by digesting or fragmenting proteins, respectively. In opposite, lipids do not produce so many fragments. The number of lipid specific fragments is in the most cases one or two. This is not sufficient to get a significant statistical conclusion. Here it might need several different variables which are associated to an identified mass spectral peak like the mass error, the resolution, the accuracy or the signal-to-noise ratio, respectively. This is a field of valuable further research.

7.1.5 Discussion about aim 5 – the correct quantification of lipid species

The isotopic correction is based on the natural occurrence of isotopes. It simply calculates the isotopic distribution from a lipids sum composition and subtracts the isotopes from the spectrum. It is important that LipidXplorer's algorithm starts with the lowest *m/z* value and continues step-wise with the next higher *m/z* value, because in this way isotopes are always calculated from peaks which were already corrected.

A characteristic of the algorithm is that isotopes are only calculated for peaks which have been assigned with a sum composition. This means furthermore that the more queries are used in parallel for probing the spectra, the more sum compositions will be assigned to peaks and therefore the more accurate the

isotopic correction will work. This makes the quality of the isotopic correction depending on the number of queries used for probing the spectra. It could be that there are lipid species in the spectrum that produce isotopic overlaps but are not in the focus of the research, i.e. are not probed with a query. One approach to prevent this dependency would be to estimate the average isotopic distribution from a given *m/z* value. This would result in a basic deconvolution of lipid containing mass spectra. However, the great diversity of lipid species might make it a complex task to find a proper averaging method which yields accurate isotopic distributions. Here, more research should be beneficial.

7.2 Real-life applications of LipidXplorer

LipidXplorer has been extensively tested in real-life applications and already contributed to interesting biological results in the works of: (Carvalho, et al., 2010; Gerl, et al., 2012; Graessler, et al., 2009; Klose, et al., 2010; Penkov, et al., 2010; Raa, et al., 2009; Reich, et al., 2010; Saito, et al., 2009).

The work of (Carvalho, et al., 2010) shows that *Drosophila melanogaster* requires bulk membrane sterol and steroid hormones in order to complete adult development. LipidXplorer was used for the identification of membrane lipid species, in particular ceramide, ceramide-PE and hexocyl-ceramide. For this work the top-down lipid identification paradigm was used for the first time. High accuracy MS spectra were acquired with and LTQ Orbitrap mass spectrometer from 17 samples and selected lipids were further injected into an ABI Qstar instrument for fragmentation and subsequent quantification. Moreover, due to the flexibility of lipid identification with MFQL, LipidXplorer could identify lipid species with unusual long chain bases.

The aim of the study in (Graessler, et al., 2009) was to elucidate if hypertension specifically affects the blood plasma lipidome independently and differently from the effects induced by obesity and insulin resistance. LipidXplorer was used for the large scale screening of lipids from the plasma lipidome of 19 men with hypertension and 51 normotensive male controls. The analysis encompassed 95 lipid species of 10 major lipid classes from a set of 151 high resolution mass spectra. The major finding of this study revealed that the overall content of ether

lipids decreased in the blood plasma of hypertensive individuals.

(Penkov, et al., 2010) discovered a lipid class, maradolipids, which are the first diacyltrahaloses found to be produced in animal organisms. The nematode *Caenorhabditis elegans* synthesizes maradolipids for the highly stress-resistant dauer larvae which might be important for their resistance to extreme environmental stress. The structure of maradolipid species was first elucidated with nuclear magnetic resonance (NMR) (Bloch, 1946; Purcell, et al., 1946) technology. The found structure and the fragmentation pathway was stored with MFQL and could be used for further mass spectrometric experiments.

In (Raa, et al., 2009) the influence of the glycosphingolipid synthesis for the uptake and intracellular transport of Shiga toxin in HEp-2 cells was investigated. 135 lipid species from the major lipid classes of PC, PE and sphingomyelins (SM) could be identified and major changes in the lipidome for HE-2 cells in different stages could be observed when inhibiting the glycerophospholipid synthesis. Furthermore, the identification of ceramide, hexocylceramide and globotriaosylceramide species with LipidXplorer could contribute to the overall result of this study.

For the studies of (Saito, et al., 2009) and (Reich, et al., 2010), LipidXplorer played also a major role for the identification of lipid species of major lipid classes. In both studies, also species of the non-polar lipid class of triacylglycerols could be identified and contributed to the result.

Last but not least, the concepts and algorithms of this work will be used for a start-up company which is forming right now. The goal of the company is to set up a high throughput lipid identification and quantification pipeline for large number of samples. It includes the storing, acquisition, identification and interpretation of samples. The first service targets at an interpretation of ca. 100,000 samples from clinical studies.

7.3 Future directions

Although LipidXplorer fulfills the requirements of a modern shotgun lipidomics platform and – as it could be shown is this thesis - is ahead of similar programs in its field, it still has space for improvements. This concerns mainly the processing speed which can be slow under certain circumstances. The slowdown becomes

noticeable if a great number of samples should be imported and several queries probing for fragments are used, respectively. This can result in a processing time of several hours. Proposals for addressing this issue are stated in the following chapter.
In the chapters after this, approaches are presented which could expand the field in which LipidXplorer can be applied.

7.3.1 Upgrade speed

Python is an interpreted scripting language and therefore naturally slower than compiled languages. It could be criticized that Python was used for programming LipidXplorer in a first run instead of a low-level programming language like C/C++. However, as an interpreted language it is a good choice for rapid prototype development. Python supports all the features of a modern scripting language (Ousterhout, 1998) which is in particular its fast development cycle and clear structure of code. While its development, LipidXplorer benefits from that because parts of it could be re-written easily after some logical errors were detected or ideas changed about how certain problems should have been approached. What makes Python in addition so attractive as a Bioinformatics programming language is its support of scientific calculations[1] and Bioinformatics[2,3] with a great number of modules.
As already discussed in Chapter 7.1.2 the alignment of samples is a resource demanding process. For example, if a great number of samples should be imported (> 200), the import can take several hours even by using the fast mixed bucketing algorithm. The same holds for the memory which is used heavily for the alignment algorithm. Re-writing resource demanding parts of the code in a low-level language would be a solution, because Python supports the import of modules written in C. Besides that, another possibility would be to parallelize the import algorithms because nowadays, every modern Computer accommodates several processors. The scan averaging process could be optimized by running each sample in a separate thread. Thus, different samples would be processed in

[1] http://numpy.scipy.org

[2] http://biopython.org/wiki/Main_Page

[3] http://www.awaretek.com/tutorials.html#bio

parallel. At its best the speed would be increased by the number of CPUs. However, not every algorithm can be split up into parallelizable sub-problems. The alignment algorithm cannot be split in the same way, since the data from all samples are already contained in one list. However, this list could be split up into different parts, processed in several threads and re-combined again. This is possible because peaks which should be aligned are always in a close neighborhood, such that there is no need to consider the whole peak list at once as it would be the case for hierarchical clustering, for example. The memory usage would also be decreased if processes were parallelized, since they were split into smaller sub-problems.

A bottleneck of the lipid identification routine is the required generation of possible fragment combinations for a precursor mass. In order to detect the right lipid, LipidXplorer needs to calculate all possible combinations of calculated sum compositions of each individual fragment of a query. The possibilities increase exponential. For example, TAG species are built up from three fatty acids, i.e. there are up to $(3n)^3$ possible variations with n being the number of fragments. If there are m isobaric species, i.e. different TAG species having the same precursor *m/z* value n has the following value: $1 < n < 3m$. The problem with the number of possible variations increases with low resolution mass spectra, where naturally more possibilities for fragment sum compositions are found.

7.3.2 Subsequent data interpretation and analysis

The output format of LipidXplorer is a comma separated file (*.csv) containing columns defined in the MFQL queries. Most commonly, it contains the lipid's name and its intensity values. At this point, LipidXplorer does not provide any function for normalizing the intensities to a given standard or generation of statistical data. Intentionally, no such functions were implemented since there are plenty of different ways to do standardization and subsequent data analysis. The idea for the design of LipidXplorer was to have a clear data interface for both input and output. The development should be strictly focused on the functionality of lipid identification and spectra preprocessing. Nevertheless, post processing the results and producing graphics would give a good overview over the identified lipid species and aid with their interpretation. This could include bar diagrams

showing the intensities of whole lipid classes, the up- and down-regulation of major lipid species, comparisons between samples with a principle component analysis, etc. Technically, this could be a report in *.pdf format which includes this collection of diagrams and charts. The only limitation is that the report variables of each MFQL query would have to be the same, because otherwise the software would not know which column contains the same kind of data.

7.3.3 Incorporation of LC-MS

Although shotgun lipidomics is the most used technique for mass spectrometry of lipidomics, it might reach its limits if low sensitive peaks should be identified. This is due to the high density of relevant peaks in shotgun spectra. Often there are isotopic peaks overlapping with monoisotopic peaks making a correct quantification difficult. Furthermore, the ion suppression is high and especially small peaks might disappear. In opposite, a chromatographic pre-separation enhances the sensitivity because different lipid categories elute at different retention times. In liquid chromatography coupled with a mass spectrometer (LC-MS/MS-MS), the analyte is infused in the mass spectrometer during its elution from the LC column. Spectra are acquired periodically during this time with a frequency of one minute, for example. Thus, the MS analysis is done only on a small fraction of the whole sample and has every spectrum now marked with a retention time stamp. This information can be used as an additional lipid specific characteristic and incorporated into LipidXplorer. With MFQL, LipidXplorer could easily take full advantage of LC-MS/MS-MS features. There would be no restriction on lipid identification with MFQL, the same principles as for shotgun lipidomics hold here. MFQL has only to be expanded with commands addressing the retention time. Algorithms for (LC-) peak detection, feature recognition and quantification could easily be added as module in front of LipidXplorer’s import and work parallel with the import of shotgun lipidomics data.

7.3.4 Applications of MFQL in other areas

The approach of using a query language for the identification of molecular species from mass spectra is not limited to lipids. It can be used with all kind of molecules which have a small number of specific fragments and follow specific

fragmentation pathways.

An area in which LipidXplorer could be used is the field of metabolomics which is about the identification of metabolites. Metabolites are a superset of lipids and have a greater structural diversity and not as many repeating structures like most of the major lipid classes. Without computational interpretation, metabolites are identified by means of specific organic neutral losses (McLafferty and Turecek, 1993) manually. The computational identification of metabolites is mostly based on comparison with spectra databases or comparison with an intermediated step which elucidates possible sum compositions or fragments (see Chapter 2.4).

However, it was also shown in (Dudley, et al., 2010; Lim, et al., 2007) that some metabolite classes follow certain rules while fragmentation. This makes them good candidates to be also identified with MFQL. All algorithms and functions can be used as they are for metabolites. With some additional functions MFQL might cover even a greater range of metabolites, since it is customized to lipids as it is now.

Another possible application of LipidXplorer is the identification of metabolites in shotgun spectra. The algorithms and tools presented in Chapter 2.4 elucidate fragmentation rules and sum compositions for metabolites from mass spectra. In (Rasche, et al., 2010; Wolf, et al., 2010), for example, fragmentation patterns were deduced in silico with the use of basic knowledge about bond dissociation. To obtain correct results, it needs a clean spectrum where no precursor mass overlaps with another compound. But a once elucidated fragmentation could be easily stored as an MFQL query and used for metabolite identification in shotgun spectra. Then, this molecular species can be identified in a high throughput manner from complex mixtures.

8 Materials and Methods

8.1 Annotation of lipid species

Individual molecular species are annotated as follows: <lipid class > [<no. of carbon atoms in the first fatty acid or fatty alcohol moiety >:<no. of double bonds in the first fatty acid or fatty alcohol moiety >/<no. of carbon atoms in the second fatty acid moiety >:<no. of double bonds in the second fatty acid moiety >]. For example, PC [18:0/18:1] stands for a phosphatidylcholine comprising the moieties stearic (18:0) and oleic (18:1) fatty acids. If the exact composition of fatty acid or fatty alcohol moieties is not known, the species are annotated as: <lipid class > <no. of carbon atoms in both moieties >:<no. of double bonds in both moieties >. In this way, PC [36:1] stands for a PC species having 36 carbon atoms and one double bond in both fatty acid moieties.

8.2 Mass spectrometry experiments in general

Mass spectrometry experiments were performed on a linear trap quadrupole (LTQ) Orbitrap XL hybrid, LTQ Orbitrap Velos, TSQ Vantage mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) and, were specified, on a modified QSTAR Pulsar *i* quadrupole time-of-flight mass spectrometer (MDS Sciex, Concord, Ontario, Canada), all equipped with a robotic nanoflow ion source TriVersa (Advion BioSciences, Ithaca, NY, USA). If not specified otherwise, data-dependent acquisition was performed as described in (Schwudke, et al., 2006). A data-dependent acquisition cycle consisted of one MS spectrum followed by MS/MS spectra acquired from ten most abundant precursor ions, whose masses were subsequently excluded from further MS/MS experiments. MS/MS spectra were acquired on a LTQ Orbitrap using pulsed Q collision-induced dissociation (PQD) under the normalized collision energy of 21%. Fragment ions were detected at the linear ion trap (IT) or Orbitrap analyzers, as indicated separately for each experiment. The linear ion trap was operated at the low (unit) mass resolution *R*, while the mass resolution of the Orbitrap was set for each experiment separately using the target resolution parameter specified as Full Width Half Maximum (FWHM) of the peak at *m/z* 400. Where specified, LTQ Orbitrap MS/MS spectra were acquired by the method of higher energy collision-induced dissociation (HCD). Precursor ions were isolated by the linear ion trap at

the unit resolution, fragmented in the HCD cell under the normalized collision energy of 45% and fragment ions detected by the Obitrap analyzer at a mass resolution of 7,500. Shotgun lipidomics experiments on triple stage quadrupole (TSQ) Vantage were performed under basic instrument settings.

MPIS scans were acquired on a quadrupole time-of-flight mass spectrometer QSTAR Pulsar *i* (AB Sciex, Toronto, Ontario, Canada) and interpreted by LipidProfiler software as described in (Ejsing, et al., 2006). Data-dependent MS/MS experiments on a QSTAR Pulsar *i* were performed as described in (Schwudke, et al., 2006).

8.3 Implementation of LipidXplorer software

LipidXplorer was programmed in Python 2.6. It imports spectra in *.mzXML (Pedrioli, et al., 2004) or peak lists in the *.dta/*.csv format. Free converters to *.mzXML are available here (http://tools.proteomecenter.org/wiki/index.php?title=Software:TPP or (http://proteowizard.sourceforge.net/). LipidXplorer automatically converts *.raw or *.wiff files into *.mzXML using, respectively, ReAdW or mzWiff programs.

LipidXplorer organizes mass spectra in a database-like format termed MasterScan (*.sc). The MasterScan is saved using Python's PICKLE function (http://docs.python.org/library/pickle.html) for Python object serialization.

The MFQL interpreter is written using PLY (Python Lex-Yacc) (http://www.dabeaz.com/ply/), a lexer/parser generator based on Lex and Yacc. A collection of MFQL scripts is included in the distributed version of LipidXplorer and supports quantitative profiling of 19 major lipid classes. The routine for calculating sum compositions is an exhaustive search algorithm written in C and imported into Python.

The algorithm for calculating isotopic distributions was developed by Dr Magnus Palmblad (University of Reading, UK) (Palmblad, et al., 2001) and converted to Python by Dr Brian H Clowers using the NUMPY module (http://numpy.scipy.org/).

LipidXplorer is available under general public license (GPL) at (https://sourceforge.net/projects/lipidxplorer/files/). Full documentation on LipidXplorer, including the installation guidelines, a lipid identification tutorial and

a library of MFQL scripts are provided at (https://wiki.mpi-cbg.de/wiki/lipidx/index.php/Main_Page). A sample dataset of shotgun mass spectra is also available for testing local installations of the software.

8.4 LipidXplorer benchmarking (Chapter 6.4): the dataset

E. coli total lipid extract was purchased from Avanti Polar Lipids (Alabaster, AL, USA) and analyzed on the LTQ Orbitrap XL instrument in negative ion mode. A solution of the total lipid concentration of 2.5 µg/ml in 7.5 mM ammonium acetate in choloroform/methanol/2-propanol (1/2/4, v/v/v) was infused into the mass spectrometer by TriVersa robotic ion source using a chip with the diameter of spraying nozzles of 4.1 µm. To produce the spectra dataset, the extract was analyzed in several independent experiments: experiment I, eight acquisitions under the unit mass resolution (*R*) settings using ion trap (IT) to acquire both MS and MS/MS spectra; experiment II, six acquisitions with *R* = 7,500 for MS spectra (Orbitrap) and unit resolution for MS/MS spectra (IT); experiment III, four acquisitions with *R* = 30,000 for MS spectra (Orbitrap) and unit resolution for MS/MS spectra (IT); experiment IV, four acquisitions with *R* = 100,000 for MS spectra (Orbitrap) and unit resolution for MS/MS spectra (IT); experiment V, seven acquisitions with *R* = 100,000 for MS spectra (Orbitrap) and *R* = 15,000 for MS/MS spectra (Orbitrap).

In the experiments I to IV, each acquisition produced approximately 33 MS and 330 MS/MS spectra; in the experiment V, 10 MS and 100 MS/MS spectra were acquired. To reduce undersampling, in the experiment V, acquisition of MS/MS spectra was navigated by the inclusion list compiled from 40 masses of plausible PE, PG and PA precursors. A list of molecular lipid species was produced by manual interpretation of spectra acquired in the experiment V with requested mass tolerance of better than 3 ppm for precursors and 5 ppm for specific fragment ions. Only lipid species identified in at least four out of seven replicated analyses were included.

Spectra acquired in each of the experiments I to IV were further processed by LipidXplorer to produce corresponding MasterScan files. The dataset from the experiment I for comparative benchmarking of LipidXplorer against LipidQA and LipidSearch programs was used. Since LipidQA and LipidSearch do not align the

spectra from replicated analyses, each acquisition was processed independently and then a non-redundant list of all identified lipid species was compiled.

8.5 LipidXplorer benchmarking (Chapter 6.4): the procedure

Eight acquisitions containing complete sets of MS and MS/MS spectra were submitted as *.raw files. The output was aligned by reported lipid species (only with LipidQA and LipidSearch). Individual lipid species were considered as positively identified if they were recognized in four or more replicated analyses. In all tests the programs were prompted to identify species of PE, PI, PS, PG and PA classes. Mass tolerance was set at 0.3 Da in MS and MS/MS modes; fatty acid moieties were assumed to comprise 12 to 22 carbon atoms and 0 to 6 double bonds.

Settings specific for each tested program were as follows.

LipidXplorer: 'MS threshold' was set to 100 and 'MS/MS threshold' to 5 counts per peak area; 'Resolution gradient' was set to 1; other common spectra import settings were as in Additional file 6 (setting: 'FAS_LTQ').

LipidQA (spectra were imported as *.raw files): 'MS error' and the 'MS/MS error' were both set to 0.3 Da; 'Finnigan Filter', on; 'Quantification', off; 'Mode selection', Neg. Mode; 'If MS2 spectra were centroided', checked. Only species with a score above 0.5 were accepted. The current version of LipidQA is available at (http://msr.dom.wustl.edu/Personnel/Staff_Scientist_Song_Haowei.htm).

Lipid Search version 2.0 beta: 'SearchType' was set to 'MS2,MS3'; 'ExpType' to 'Infusion'; 'Precursor tol' to '0.3 Da'; 'Product peak tol' to 0.3 Da; 'Intensity threshold' to 0.01; 'Threshold type' to Relative; 'M-score Threshold' to 10.0. The current version of LipidSearch is available at (http://lipidsearch.jp/lipidsearch/lipidsearch.do).

LipidProfiler v.1.0.97: the software was used for creating a reference list of lipids in the *E. coli* extract and utilized a separate dataset acquired on a QSTAR Pulsar *i* mass spectrometer by the MPIS method. Intensity threshold was set to 0.2%; all lipid species reported as 'confirmed results' in at least four independent acquisitions.

8.6 Validation of isotopic correction (Chapter 6.3)

In two independent replicates a mixture of PA standards consisting of PA [18:0/18:2], PA [18:1/18:1], PA [18:0/18:1] and PA [18:0/18:0] (all from Avanti Polar Lipids) with the molar ratio of 1:9:1:1 was analyzed on a LTQ Orbitrap Velos. Spectra were acquired under data-dependent acquisition control in negative mode using the linear ion trap analyzer under a target resolution of 800 for both MS and MS/MS. Precursors were fragmented using collision-induced dissociation. To process the dataset, mass tolerance was set to 300 ppm for MS and 500 ppm for MS/MS, spectra; occupation threshold was set to 0.5.

8.7 Validation of the spectra alignment algorithm (Chapter 6.1.2)

A dataset of 128 MS spectra of human blood plasma extracts acquired on a LTQ Orbitrap XL mass spectrometer was used. Spectra were imported into a MasterScan file assuming a mass resolution of 127,500 (FWHM, at *m/z* 400), a mass accuracy of 4 ppm, and an occupation threshold of 0.5. Post-acquisition adjustment of peak masses was achieved using two reference masses of lipid standards spiked into the samples prior to extraction (Graessler, et al., 2009). Lipids of 11 major classes (PC, PC-*O*, PE, PE-*O*, LPC, LPE, SM, DAG, TAG, Chol and CholEst) were identified by their accurate masses with no recourse to MS/MS.

8.8 LipidXplorer supports consistent cross-platform identification of lipids (Chapter 6.7)

A total lipid extract of *E. coli* was analyzed by multiple precursor ion scanning (Ejsing, et al., 2006) and by data-dependent acquisition (Schwudke, et al., 2006) on a QSTAR Pulsar *i* mass spectrometer. The same extract was analyzed by data-dependent HCD at the LTQ Orbitrap Velos mass spectrometer. Each analysis was performed in four replicates. Datasets of shotgun MS and MS/MS spectra were imported into MasterScan files built separately for each mass spectrometer and lipid species identified by MFQL queries (see Additional file 7 for the import settings and Appendix 10.7 for the queries). Lipid species were quantified in MS mode by using the intensities of their molecular ions. For MS/MS quantification, MFQL queries recognized and reported the sum of abundances of

acyl anion fragments for each individual precursor. Relative quantities of individual lipids were calculated by normalizing to the total abundance of all species of the same lipid class. Parameters of linear correlation of lipid species profiles obtained by different methods (correlation coefficient R^2 and slope) were computed by Microsoft Excel.

8.9 LipidXplorer supports different acquisition types (Chapter 6.8)

A total lipid extract of *E. coli* was analyzed by on the quadrupole time-of-flight instrument QSTAR Pulsar I and the LTQ Orbitrap Velos using data-dependent acquisition in negative ion mode. Precursor ions were selected with unit mass resolution such that only monoisotopic peaks of plausible precursors were fragmented. Collision energies were optimized as described in (Ejsing, et al., 2006; Schuhmann, et al., 2011) and either ramped with precursor masses (QSTAR) or applied as a “normalized” collision energy (Schuhmann, et al., 2011) (LTQ Orbitrap). For better consistency, spectra were acquired with approximately the same mass resolution of R=7500 for both LTQ Orbitrap and QSTAR. Both experiments were performed in four technical replicates, processed and individual species quantified with the MFQL queries for PE and PG used for Chapter 6.4 (see Appendix 10.7). The same extract was infused into the TSQ Vantage. Precursor ion scans for common acyl anion fragments were repeatedly performed in negative ion mode for 40 minutes resulting in a total of 72 scans. Precursors were selected within a 1.0 Da isolation window, consistently with the experiment settings applied on the QSTAR and Orbitrap. The MasterScan was composed from the aligned and transformed spectra and interpreted by the same MFQL queries identifying lipid species of PE and PG classes as for the LTQ Orbitrap and QSTAR experiments. Quantification was done according to Chapter 8.8.

The neutral loss scans on the TSQ Vantage were performed with the same instrument settings as stated above, however spectra were acquired in positive mode under CE= 22 eV.

8.10 MFQL exploits the diversity of lipid fragmentation pathways - Analysis of bovine heart total lipid extract (Chapter 6.9)

A total lipid extract of bovine heart (Avanti Polar Lipids) was analyzed in six technical replicates on a LTQ-Orbitrap XL mass spectrometer using a target resolution of 100,000 for MS spectra (Orbitrap) and unit resolution for MS/MS (IT) in negative ion mode. Six replicates were acquired, each consisting of 31 MS and 310 MS/MS spectra.

8.11 Publicly accessible depository of spectra

Mass spectra used for benchmarking and validating of LipidXplorer are available in original formats (*.raw for LTQ Orbitrap and *.wiff for QSTAR Pulsar *i*) at the LipidXplorer Wiki page (https://wiki.mpi-cbg.de/wiki/lipidx/index.php/Main_Page).

9 Abbreviations

Names of lipid classes

Cer	ceramides
Chol	cholesterol
CholEst	cholesterol ester
CL	cardiolipin
DAG	diacylglycerol
EPIC	elucidation of product ion connectivity
FPR	false positive rate
FWER	family-wise error rate
FDR	false discovery rate
HMDB	human metabolome database
IUPAC	international union of pure and applied chemistry
KEGG	Kyoto encyclopedia of genes and genomes
LCL	triacyl-lysocardiolipin
LPA	lyso-phosphatic acid
LPC	lyso-phosphatidylcholines
LPE	lyso-phosphatidylethanolamines
LPG	lyso- phosphatidylglycerol
LPI	Lyso-phosphatidylinositol
MLP	maradolipid
NIST	national institute of standards and technology
PA	phosphatidic acid
PC	phosphatidylcholines
PC-*O*	1-alkyl-2-acylglycerophosphocholines
PE	phosphatidylethanolamines
PE-*O*	1-alkyl-2-acylglycerophosphoethanolamines
PG	phosphatidylglycerols
PI	phosphatidylinositols
PS	phosphatidylserines
SM	sphingomyelins
TAG	triacylglycerols

Others

BNF	Backus-Naur Normal Form
CID	collision-induced dissociation
CPU	central processing unit
Da	Dalton
DDA	data-dependent acquisition
EI	electron impact
ESI	electrospray ionization
eV	electron volt
FA	Fatty Acid - acyl anions of fatty acid moieties
FAO	Fatty Acid ether – alkyl anions of fatty acid moieties
FT-ICR	Fourier transform ion cyclotron resonance
FTMS	Fourier transform mass spectrometer
FWHM	full width at half maximum
GLC	gas liquid chromatography
GUI	graphical user interface
HCD	higher energy collision-induced dissociation
HG	head group
HPLC	high performance/pressure liquid chromatography
IT	ion trap
IUPAC	International union of pure and applied chemistry
LC	liquid chromatography
LC-MS	liquid-chromatography mass spectrometry
LIT	linear ion traps
LOD	limit of detection
LTQ	linear trap quadrupole
m/z	mass-to-charge ratio
MALDI	matrix-assisted laser desorption/ionization
MFQL	molecular fragmentation query language
mM	millimolar; 1 mM = 10^{-3} mol/l
mmu	milli mass unit; 1 mmu = 0.001 Da = 1 mDa (millidalton)
MPIS	multiple precursorion scanning
MS	mass spectrometry / mass spectrum
MS/MS	tandem mass spectrometry / fragment spectrum
NLS	neutral loss scan
PCF	Pearson correlation factor

PIS	precursor ion scan
PQD	pulsed Q collision-induced dissociation
PUFA	polyunsaturated fatty acid
ppm	parts-per-million; 1 ppm = 0.0001 %
Q-TOF	quadrupole time-of-flight
QTRAP	quadrupole ion trap
sc	sum composition
SQL	structured query language
SRCF	Spearman rank correlation factors
TLC	thin layer chromatography
TOF	time-of-flight
TSQ	triple stage quadrupole
µg	microgram; 1 µg = 10^{-6} g

10 Appendix

10.1 The Graphical User Interface (GUI)

The GUI consists of four panels which guide the user through the importing and identification process of LipidXplorer:

a. In ‘Import Source’ the user selects the folder with his experiments. He also specifies the importing format and optionally the grouping of the source spectra.
b. With ‘Import Settings’ all settings concerning the kind of the used mass spectrometer and acquisition options are set.
c. In the ‘Run’ panel the user chooses the MFQL queries he wants to use for lipid identification and starts the identification process.
d. ‘MS Tools’ is outside the lipid identification pipeline but provides some helpful functions calculate sum compositions from masses and the other way round as well as the isotopic distribution of precursors and fragment masses.

Additionally, LipidXplorer has an own editor implemented, which uses syntax highlighting of MFQL keywords for a better readability.

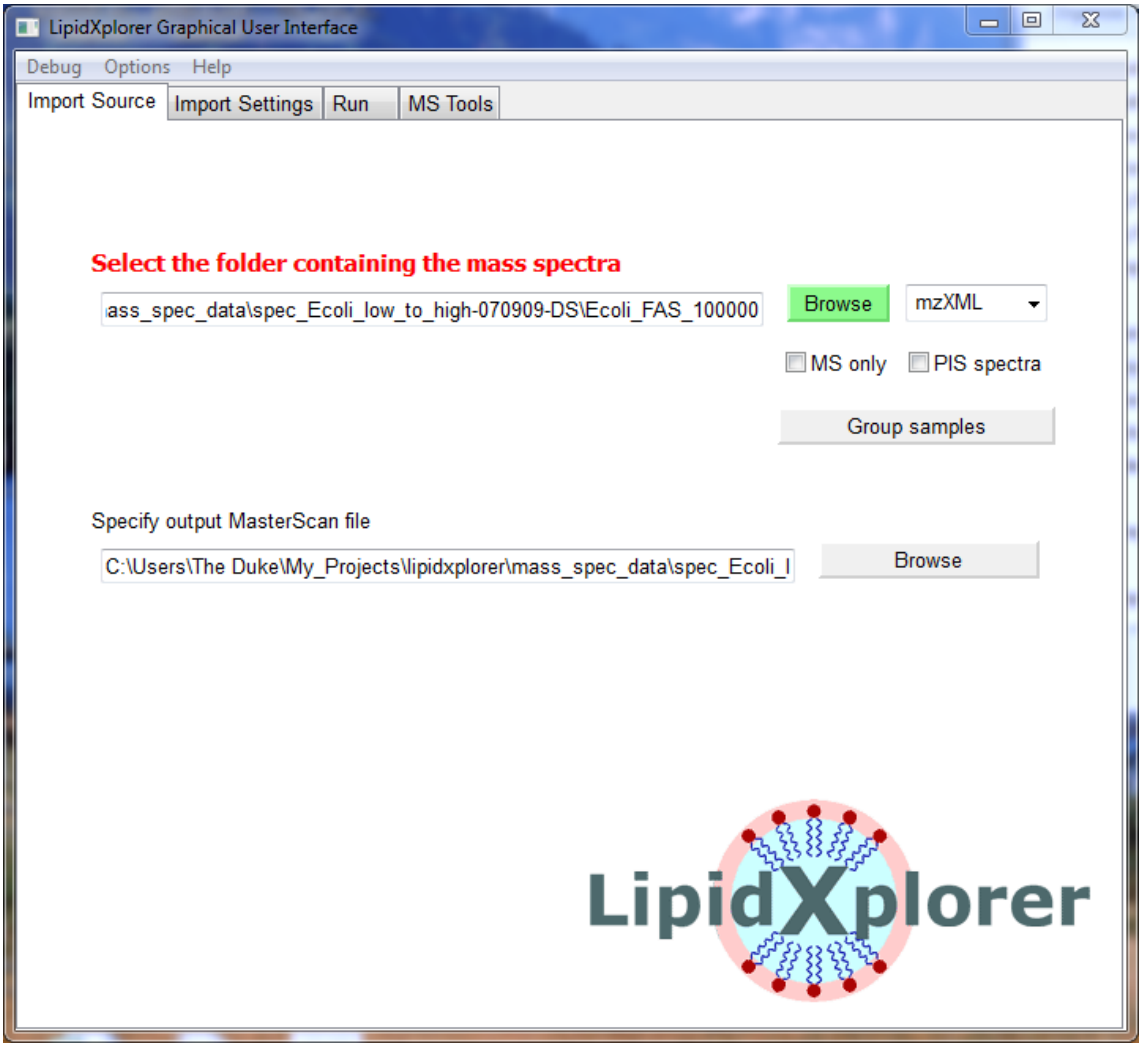


Figure A 1: **Screenshot of the Import source panel of LipidXplorer.**

Here the folder containing the mass spectra is specified, either by drag'n'drop or the "Browse" button. The format of the spectra has to be specified as well. By checking "PIS spectra", LipidXplorer imports precursor ion scan spectra which were acquired on triple quad instruments. Spectra replicates can be grouped with "Group samples" to enhance the ion statistics.

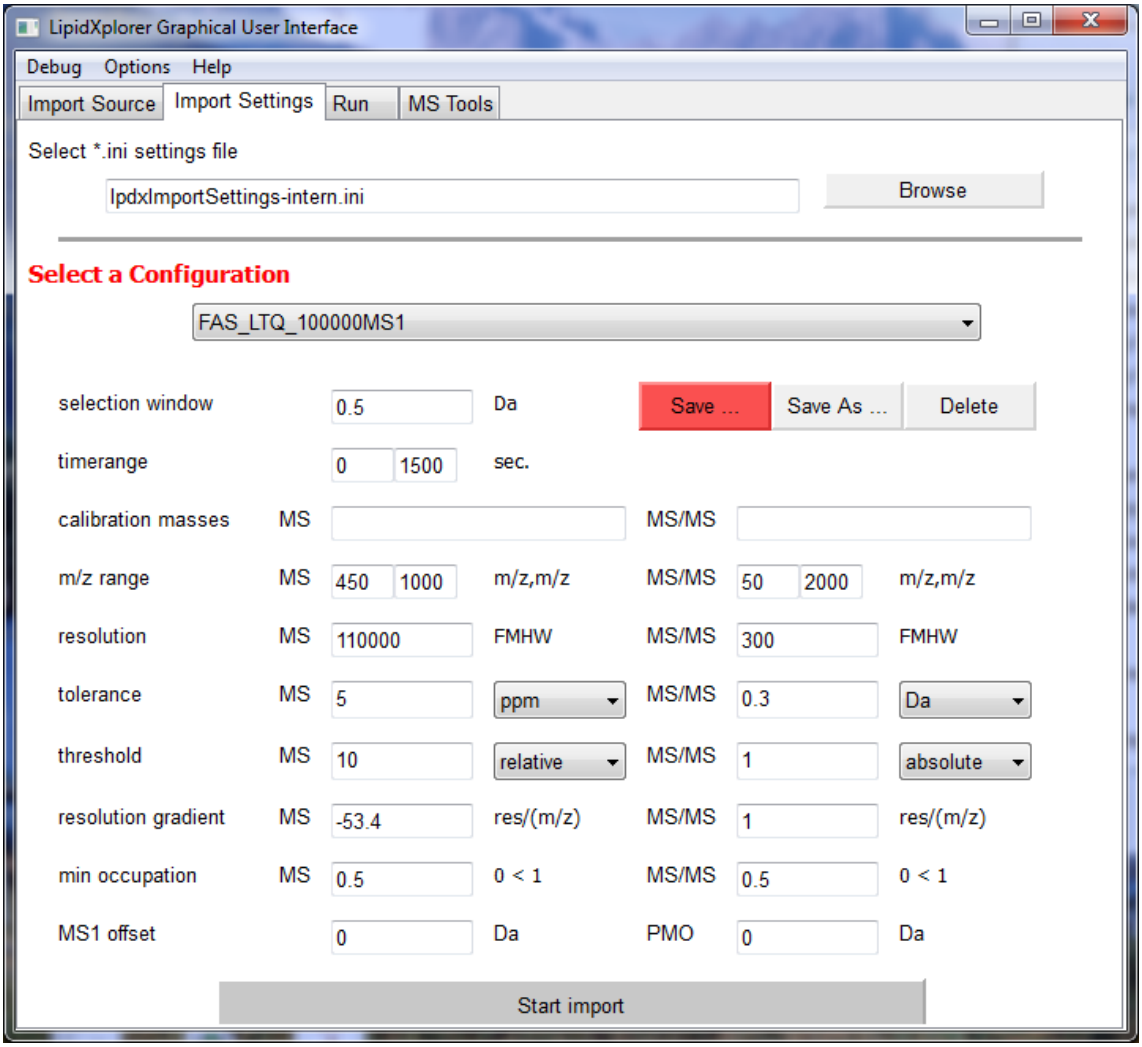


Figure A 2: Screenshot of the Import Settings panel.

In this panel the settings for the used mass spectrometer are specified. The settings can be stored in a configuration which itself is stored in an *.ini settings file. In this way, different configuration can be comfortely managed.

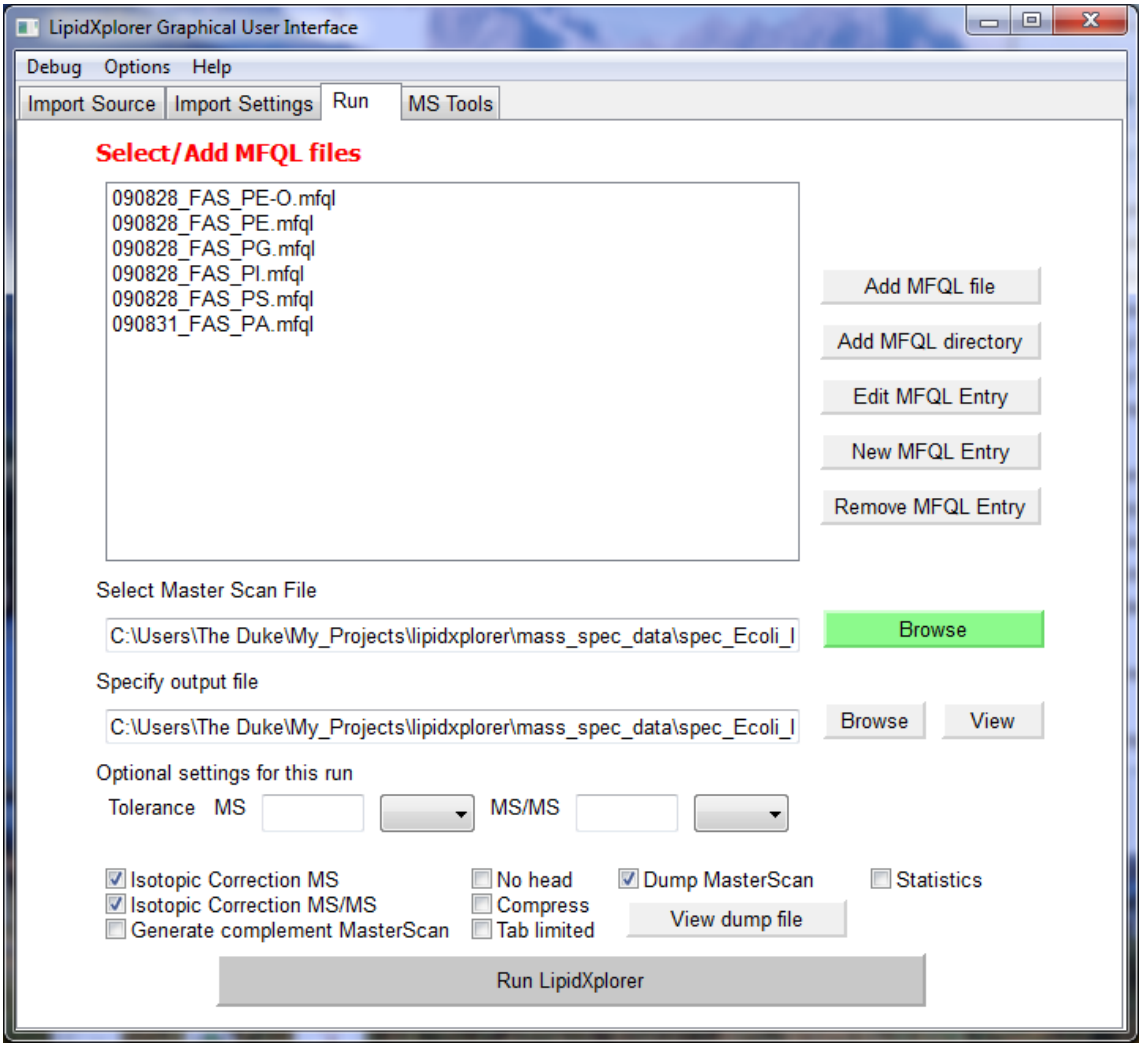


Figure A 3: Screenshot of the Run panel.

Here, the MFQL files which should be used for probing the MasterScan are specified. They also can be drag'n'dropped into the "Select/Add MFQL files" window. Furthermore, the user can "dump" the MasterScan to a *.csv file which can be read by any spread sheet software. This gives the user insight in how LipidXplorer manages the spectral data and executes the lipid identification.

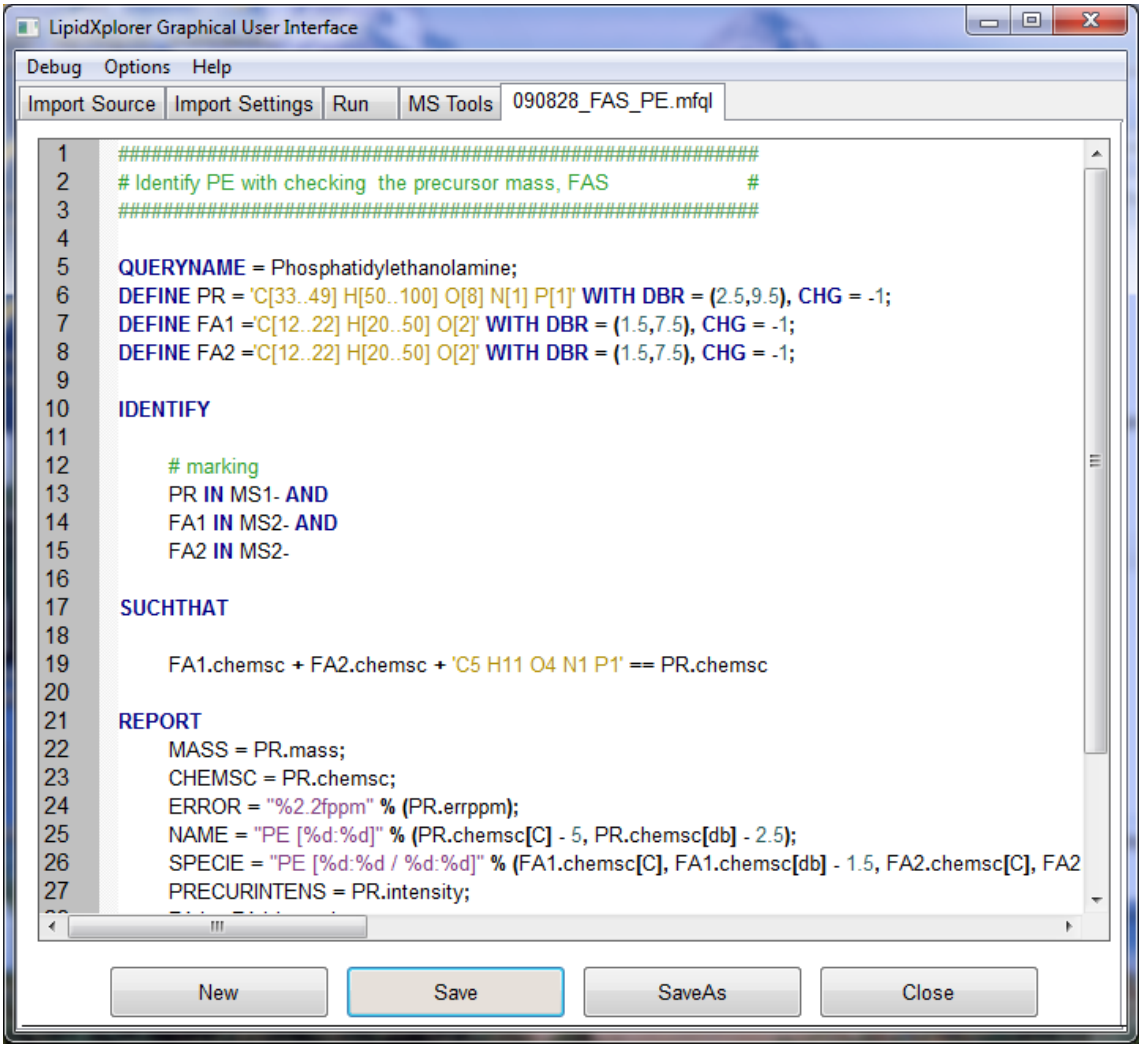


Figure A 4: Screenshot of the MFQL editor.

LipidXplorer provides an editor for MFQL files. It features MFQL-syntax highlighting and the standard editor features.

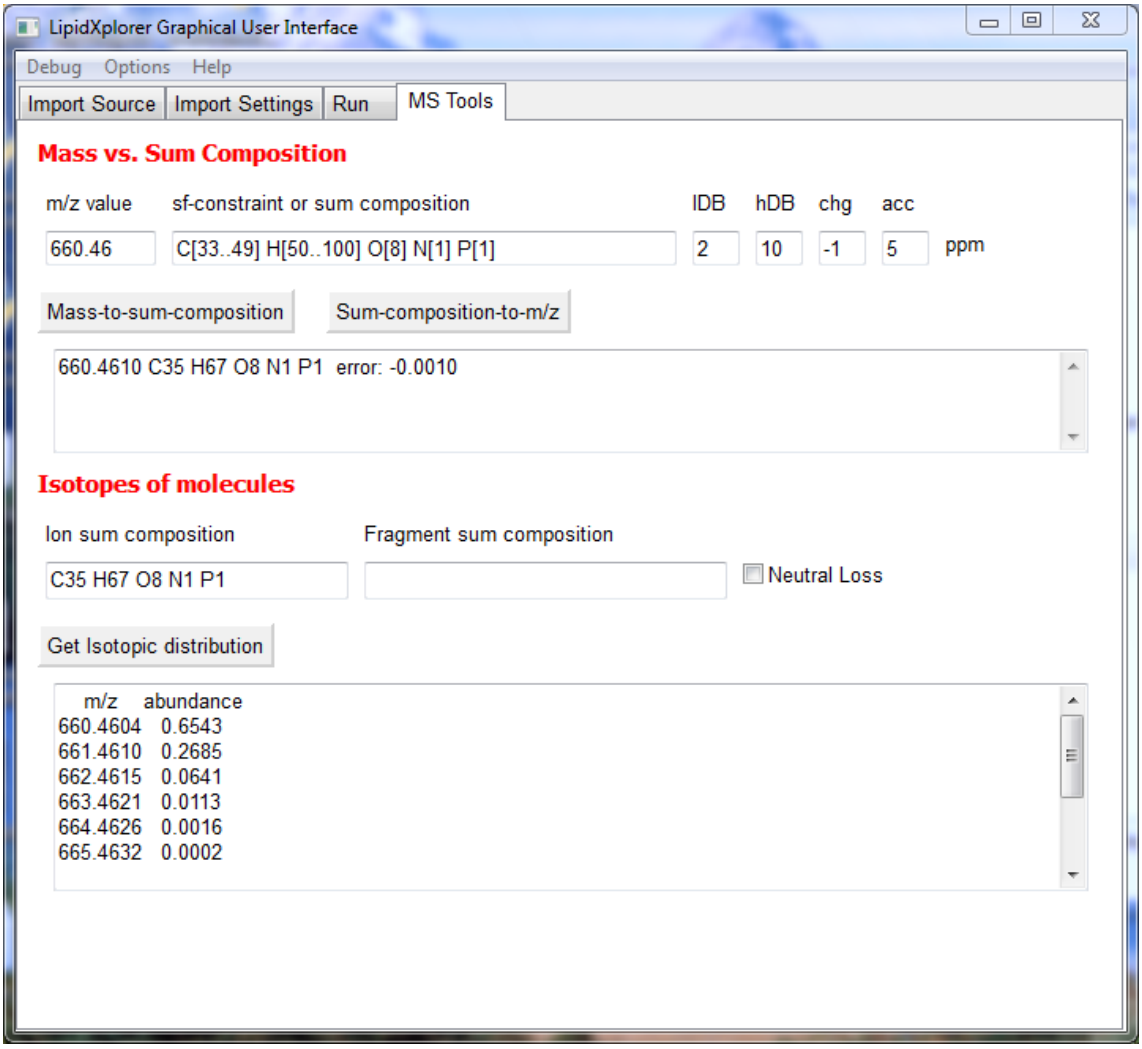


Figure A 5: Screenshot of the MS-Tools panel.

With the upper part, one can calculate sum compositions from a given m/z value and a given sc-constraint or vice versa from a sum composition to its m/z value for a given charge. The lower part shows the isotopic distribution of optionally a precursor sum composition or of a fragment sum composition.

10.2 Detailed description of scan averaging algorithm

The notion of a spectrum will be introduced: A mass spectrum $S = \{p_0, \dots, p_n\}$ is a set of peaks $p_i = (m_i, I_i, L_i)$, where m_i is the mass of the peak, I_i its intensity and L_i the initially empty set of intensities for $i \in \{0, \dots, n\}$.

The input to the algorithm is:

- The single scan survey spectra given as peak lists S_i for $i = 1, \dots, n$
- The resolution $R(m)$ is assumed to change linearly within the full mass range; its slope (the mass resolution gradient) and intercept (the resolution at the lowest mass of the full mass range) are instrument-dependent features pre-calculated by the user from some reference spectra.
- The bin size of given mass $b(m) = {}^{m}/_{R(m)}$
- A intensity threshold value T.
- An empty set S_{result}

First step is to put all peaks of all given spectra together in a set $\tilde{S} = \bigcup_{i=1}^{n} S_i$. The peaks $p_1, \dots, p_{|\tilde{S}|} \in \tilde{S}$ are sorted increasingly according to their m/z value. The algorithm begins with the smallest mass $p_1 \in \tilde{S}$:

1. Repeat:
    a. Collect all peaks, whose m/z values are not greater than $m_i + b(m_i)$ in a set $B = \{p_i, p_{i+1}, \dots, p_{i+k}\}$ for $k \in \mathbb{N}$
    b. If there is at least one peak $p_i \in B$ with $I_i \geq T$ continue with c), otherwise go to d)
    c. Calculate the intensity weighted average $m_{avg} = \frac{\sum_{j=i}^{i+k} m_j I_j}{\sum_{j=i}^{i+k} I_j}$ of the masses, calculate the average intensity as $I_{scan} = \frac{\sum_{j=i}^{i+k} I_j}{k}$ and have $L_i = L_i \cup \{I_i, \dots, I_{i+k}\}$. Store the result in the new spectrum $S_{new} = S_{new} \cup p_i$ with $p_i = (m_{avg}, I_{scan}, L_i)$
    d. Go to the subsequent peak of the greatest peak in B, i.e. $i := i + k + 1$ and continue the algorithm with a) till the last entry in $\tilde{S}$ is reached. If all peaks of S are processed then $\tilde{S} := S_{new}$
2. Calculate the resulting spectrum S_{result} by calculating for every $p_i \in \tilde{S}$: $S_{result} = S_{result} \cup p_{result}$ where $p_{result} = (m_i, I_{scan})$ with $I_{scan} = \frac{\sum_{j \in L_i} j}{n}$

3. Reduce the spectrum by peaks which are below an intensity threshold T_{stn} which reflects the decreasing signal-to-noise ratio when averaging several spectra. This value depends on the user given intensity threshold T: $T_{stn} = \frac{T}{\sqrt{n}}$. For this reason, the threshold is lower with a greater number of averaged spectra. The resulting spectrum is therefore:

   For every $p_i = (m_i, I_i) \in S_{result}$ do:

   if $I_i \geq T_{stn}$ then $S_{endresult} = S_{endresult} \cup p_i$

10.3 Work scheme of binning process in scan averaging

The distribution of peaks, as they are produced from the mass spectrometer, is assumed to be Gaussian. This leads to a threefold repetition of the scan averaging algorithm (explained in more detail in Figure A 6). The required number of repetitions comes from experiences with the scan averaging algorithm. Always after three repetitions, the algorithm did not find any peak within the binning window.

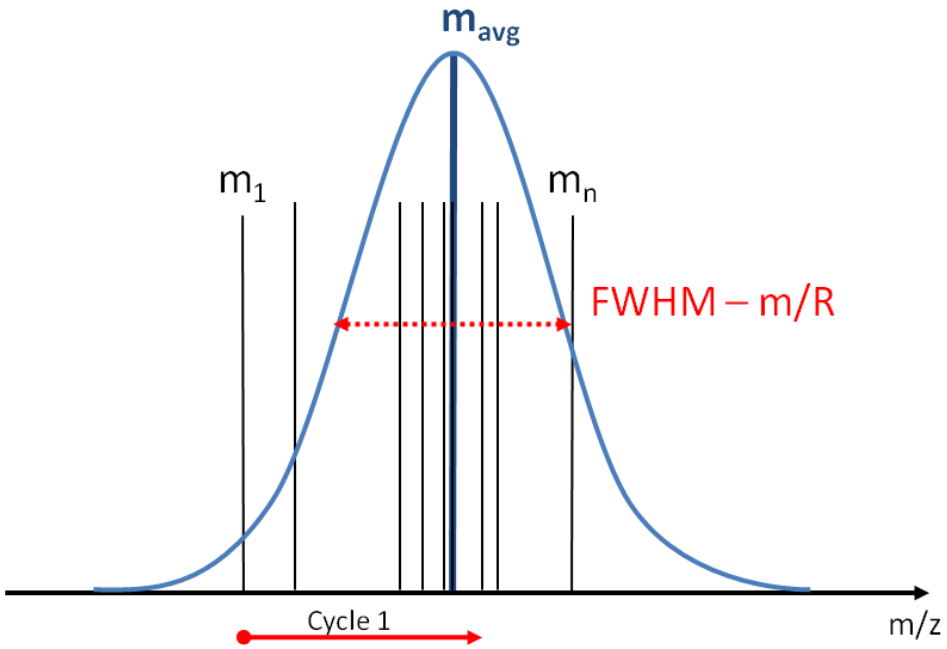


Figure A 6: The peak distribution in mass spectra.

In a collapsed peak list the centroid masses $\{m_1, \dots, m_n\}$Smith, 2000 of all scans n fall within a cluster. For the binning process it is assumed that all $m_1, \dots, m_n$ are normal distributed around the "true mass" m_{avg}. This distribution is sampled by its FWHM which can be determined from $R(m)$. In the first circle, the smallest mass m_1 is applied to determine the start of the bin. Because of the determination of the higher bin border as $m_1 + R(m)$ the first determined averaged mass of the first cycle (Cycle 1) bin must not reflect the complete distribution measured masses of that cluster and more cycles have to follow to bin the peaks correctly. (Source: (Herzog, et al., 2011))

10.4 Detailed description of spectra alignment algorithm

The notion of a spectrum is introduced: A mass spectrum $S = \{p_0, \dots, p_n\}$ is a set of peaks $p_i = (m_i, I_i, L_i)$, where m_i is the mass of the peak, I_i its intensity and L_i the initially empty set of intensities for $i \in \{0, \dots, n\}$.

The input to the algorithm is:

- All spectra S_i for $i = 1, \dots, n$. The spectra were averaged with LipidXplorers averaging algorithm.
- The resolution $R(m)$ is assumed to change linearly within the full mass range; its slope (the mass resolution gradient) and intercept (the resolution at the lowest mass of the full mass range) are instrument-dependent features pre-calculated by the user from some reference spectra.
- The bin size of given mass $b(m) = {}^{m}/_{R(m)}$
- An empty set S_{result}

First step is to put all peaks of all given spectra together in a set $\tilde{S} = \bigcup_{i=1}^{n} S_i$. The peaks $p_1, \dots, p_{|\tilde{S}|} \in \tilde{S}$ are sorted increasingly according to their m/z value. The algorithm begins with the smallest mass $p_1 \in \tilde{S}$:

1) Repeat:

    a. Collect all peaks, whose m/z values are not greater than $m_i + b(m_i)$ in a set $B = \{p_i, p_{i+1}, \dots, p_{i+k}\}$ for $k \in \mathbb{N}$

    b. Calculate the average $m_{avg} = \frac{\sum_{j=i}^{i+k} m_j}{k}$ of the masses and have $L_i = L_i \cup \{I_i, \dots, I_{i+k}\}$. Store the result in the new spectrum $S_{new} = S_{new} \cup p_i$ with $p_i = (m_{avg}, I_{scan}, L_i)$

    c. Go to the subsequent peak of the greatest peak in B, i.e. $i := i + k + 1$ and continue the algorithm with a) till the last entry in $\tilde{S}$ is reached. If all peaks of S are processed then $\tilde{S} := S_{new}$

4. The resulting spectrum $\tilde{S}$ contains all information of all spectra which were given for the import. The intensity values for a peak $p_i = (m_i, I_i, L_i)$ in L_i are from peaks of the corresponding sample, i.e. one peak p_i is a representative of all peaks with the same m/z value (“same” according to the resolution) through all samples.

10.5 The Backus-Naur-Form (BNF) of the Molecular Fragmentation Query Language (MFQL)

In the following is a description of the MFQL syntax with a BNF.

10.5.1 MFQL tokens

The tokens are given as regular expressions:

1) Identifier in LipidXplorer begin with a Letter and can contain numbers:

   ```
   Id -> [A-z][0-9A-z]*
   ```

2) A chemical sum composition can contain the elements carbon (C), hydrogen (H), nitrogen (N), natrium (Na), oxygen (O), phosphorus (P), deuterium (D), sulfur (S) and ^{13}C (Ci):

   ```
   sumComposition -> '([CHNOPDS][ia]?[0-999])+'
   ```

3) A sc-constraint is a set of sum compositions. It is defined like a sum composition with variable numbers of elements:

   ```
   scConstraint -> '([CHNOPDS][ia]?\[[0-999]\]..\[[0-999]\])+'
   ```

4) A value is the token for a (floating point-) number:

   ```
   value -> [+-]?[1-9]*(\.[0-9]+)?
   ```

5) The token ‘stringWithPlaceholders’ is a string containing C-typical placeholders for inserting values given as attributes. The placholders can be the following:

   a) *%d* for a decimal value

   b) *%m.nf* for a floating point number, where *m* is the number of digits on the left side of the decimal point and *n* the number of digits on the right side of the decimal point.

   c) *%s* for a string value

6) The following keywords are included in MFQL:

   a) QUERYNAME, DEFINE, IDENTIFY, SUCHTHAT, REPORT, AND, OR, NOT, IN, MS

10.5.2 The MFQL BNF diagram

The Backus-Naur Form is a notation technique for context-free grammars, often used to describe syntax of programming languages. A BNF specification is a set

of derivation rules, written as

<symbol> :: __expression__

where <symbol> is a so-called “nonterminal” and the ‘__expression__’ consists of one or more sequences of symbols which can be separated with a ‘|’ to indicated a choice. The right side of a rule is a possible substitution for the <symbol> on the left.

Below is the complete BNF of MFQL:

```
<start> ::  QUERYNAME = id ";" <definitions>, IDENTIFY <identify>

<definitions> ::  <definition> ";" | <definitions> ";"

<definition> ::  DEFINE id = <content>

<content> :: (sumComposition | scConstraint | value)

<identify> ::   <identification> (SUCHTHAT <suchthat> | REPORT
<report>)

<identification> ::  <scan> | <scan> AND <identification>

<scan> ::  id IN MS(1|2)(+|-)

<suchthat> ::  <conditions> REPORT <report>

<conditions> :: (NOT <condition> | <condition>) | (AND|OR) (NOT
<conditions> | <condition>)

<condition> ::  <equation> | (<|<=|>|>=|==) <condition>)

<equation> ::  <term> | ((+|-|*|/) <equation>)

<term> ::  <variable> | <function> | value | id

<variable> ::  id | (id"["id"]") | (id"."id) | (id"."id"["id"]")

<function> ::  id"("<attributes>")"

<attributes> ::  <term> | "," <attributes>

<report> ::  <mask> ";" | <report> ";"

<mask> ::  id = <equation> | <term> | <string>

<string> ::  stringWithPlaceholders "%" "(" <attributes> ")"
```

10.6 List of peak attributes

- *mass* - the m/z value of the peak
- *chemsc* - the sum composition of the peak.
- *frsc* - the sum composition of the fragment. If the peak is a fragment, it is the same as chemsc, if it is a neutral loss, it returns the sum composition of the fragment.
- *nlsc* - the sum composition of the neutral loss. If the peak is a neutral loss, it is the same as chemsc, if it is a fragment, it returns the sum composition of the neutral loss.
- *frmass* - the mass of the fragment. If the peak is a fragment, it is the same as mass, if it is a neutral loss, it returns the mass of the fragment.
- *nlmass* - the mass of the neutral loss. If the peak is a neutral loss, it is the same as mass, if it is a fragment, it returns the mass of the neutral loss.
- *errppm* - is the difference between the exact mass and the measured mass in ppm
- *errda* - the error in Da
- *errres* - := $\frac{m/z}{errda}$
- *intensity* - is the list of all intensities. Specific samples can be addressed with "[" and "]". For example: precursor.intensity[wildType*] return the intensities of all samples whose name start with "wildType"

10.7 MFQL queries used for E.coli

10.7.1 MFQL query for Phosphatidylethanolamine (PE) in negative ion mode

```
QUERYNAME = Phosphatidylethanolamine;
DEFINE PR = 'C[33..49] H[50..100] O[8] N[1] P[1]' WITH DBR = (2.5,9.5), CHG = -1;
DEFINE FA1 ='C[12..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
DEFINE FA2 ='C[12..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
IDENTIFY
       PR IN MS1- AND
       FA1 in MS2- AND
       FA2 in MS2-
SUCHTHAT
       FA1.chemsc + FA2.chemsc + 'C5 H11 O4 N1 P1' == PR.chemsc
REPORT
       MASS = PR.mass;
       CHEMSC = PR.chemsc;
ERROR = "%2.2fppm" % "(PR.errppm)";
       NAME = "PE [%d:%d]" % "((PR.chemsc)[C] - 5, (PR.chemsc)[db] - 2.5)";
       SPECIE = "PE [%d:%d / %d:%d]" % "(FA1.chemsc[C], FA1.chemsc[db] - 1.5,
FA2.chemsc[C], FA2.chemsc[db] - 1.5)";
       PRECURINTENS = PR.intensity;
       FAS = sumIntensity(FA1, FA2);;
```

10.7.2 MFQL query for Phosphatidylethanolamine ether (PE-O) in negative ion mode

```
QUERYNAME = Phosphatidylethanolamineether;
DEFINE PR = 'C[33..49] H[50..100] O[7] N[1] P[1]' WITH DBR = (1.5,8.5), CHG = -1;
DEFINE FA1 ='C[12..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
DEFINE FAO ='C[17..27] H[20..80] O[6] N[1] P[1]' WITH DBR = (0.5,6.5), CHG = -1;
IDENTIFY
       PR IN MS1-  AND
       FA1 in MS2-  AND
       FAO in MS2-
SUCHTHAT
       FA1.chemsc + FAO.chemsc == PR.chemsc + 'O1 H1'
REPORT
       MASS = "%4.4f" % "(PR.mass)";
       CHEMSC = PR.chemsc;
       ERROR = "%2.2fppm" % "(PR.errppm)";
       NAME = "PE-O [%d:%d]" % "((PR.chemsc)[C] - 5, (PR.chemsc)[db] - 1.5)";
       SPECIE = "PE-O [%d:%d / %d:%d]" % "(FAO.chemsc[C] - 5, FAO.chemsc[db] - 0.5,
FA1.chemsc[C], FA1.chemsc[db] - 1.5)";
       PRECURINTENS = PR.intensity;
       FAS = FA1.intensity + FAO.intensity;;
```

10.7.3 MFQL query for Phosphatidylglycerol (PG) in negative ion mode

```
QUERYNAME = Phosphatidylglycerol;
DEFINE PR = 'C[34..50] H[30..120] O[10] P[1]' WITH DBR = (2.5,9.5), CHG = -1;
DEFINE FA1 ='C[12..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
DEFINE FA2 ='C[12..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
IDENTIFY
       PR IN MS1- AND
       FA1 in MS2-  AND
       FA2 in MS2-
SUCHTHAT
       FA1.chemsc + FA2.chemsc + 'C6 H12 P1 O6' == PR.chemsc
REPORT
       MASS = "%4.4f" % "(PR.mass)";
       CHEMSC = PR.chemsc;
       ERROR = "%2.2fppm" % "(PR.errppm)";
       NAME = "PG [%d:%d]" % "((PR.chemsc)[C] - 6, (PR.chemsc)[db] - 2.5)";
       SPECIE = "PG [%d:%d / %d:%d]" % "(FA1.chemsc[C], FA1.chemsc[db] - 1.5,
FA2.chemsc[C], FA2.chemsc[db] - 1.5)";
       PRECURINTENS = PR.intensity;
       FAS = sumIntensity(FA1, FA2);;
```

10.7.4 MFQL query for Phosphatidylinositol (PI) in negative ion mode

```
QUERYNAME = Phosphatidylinositol;
DEFINE PR = 'C[37..53] H[30..140] O[13] P[1]' WITH DBR = (3.5,10.5), CHG = -1;
DEFINE headPI = 'C[6] H[10] O[8] P[1]' WITH DBR = (1.5,4.5), CHG = -1;
DEFINE FA1 ='C[14..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
DEFINE FA2 ='C[14..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
IDENTIFY
```

```
        PR IN MS1- AND
        FA1 in MS2- AND
        FA2 in MS2- AND
        headPI in MS2-
SUCHTHAT
        FA1.chemsc + FA2.chemsc + headPI.chemsc + 'C3 H6 O1' == PR.chemsc
REPORT
        MASS = "%4.4f" % "(PR.mass)";
        CHEMSC = PR.chemsc;
        ERROR = "%2.2fppm" % "(PR.errppm)";
        NAME = "PI [%d:%d]" % "((PR.chemsc)[C] - 9, (PR.chemsc)[db] - 3.5)";
        SPECIE = "PI [%d:%d / %d:%d]" % "(FA1.chemsc[C], FA1.chemsc[db] - 1.5,
FA2.chemsc[C], FA2.chemsc[db] - 1.5)";
        PRECURINTENS = PR.intensity;
        FAS = FA1.intensity + FA2.intensity;;
```

10.7.5 MFQL query for Phosphatidylserine (PS) in negative ion mode

```
QUERYNAME = Phosphatidylserine;
DEFINE PR = 'C[34..50] H[30..120] O[10] N[1] P[1]' WITH DBR = (3.5,10.5), CHG = -1;
DEFINE headPS = 'C[3] H[5] O[2] N[1]' WITH DBR = (-0.5,6.5), CHG = 0;
DEFINE FA1 ='C[12..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
DEFINE FA2 ='C[12..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
IDENTIFY
        PR IN MS1- AND
        FA1 in MS2- AND
        FA2 in MS2- AND
        headPS in MS2-
SUCHTHAT
        FA1.chemsc + FA2.chemsc + headPS.nlsc + 'C3 H6 P1 O4' == PR.chemsc
REPORT
        MASS = "%4.4f" % "(PR.mass)";
        CHEMSC = PR.chemsc;
        ERROR = "%2.2fppm" % "(PR.errppm)";
        NAME = "PS [%d:%d]" % "((PR.chemsc)[C] - 6, (PR.chemsc)[db] - 3.5)";
        SPECIE = "PS [%d:%d / %d:%d]" % "(FA1.chemsc[C], FA1.chemsc[db] - 1.5,
FA2.chemsc[C], FA2.chemsc[db] - 1.5)";
        PRECURINTENS = PR.intensity;
        FAS = sumIntensity(FA1, FA2);;
```

10.7.6 MFQL query for Phosphatic acid (PA) in negative ion mode

```
QUERYNAME = PhosphatidicAcid;
DEFINE PR = 'C[31..47] H[30..120] O[8] P[1]' WITH DBR = (2.5,9.5), CHG = -1;
DEFINE FA1 ='C[12..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
DEFINE FA2 ='C[12..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
IDENTIFY
        PR IN MS1- AND
        FA1 in MS2- AND
        FA2 in MS2-
SUCHTHAT
        FA1.chemsc + FA2.chemsc + 'C3 H6 P1 O4' == PR.chemsc
REPORT
        MASS = "%4.4f" % "(PR.mass)";
        CHEMSC = PR.chemsc;
```

```
        ERROR = "%2.2fppm" % "(PR.errppm)";
        NAME = "PA [%d:%d]" % "((PR.chemsc)[C] - 3, (PR.chemsc)[db] - 2.5)";
        SPECIE = "PA [%d:%d / %d:%d]" % "(FA1.chemsc[C], FA1.chemsc[db] - 1.5,
FA2.chemsc[C], FA2.chemsc[db] - 1.5)";
        PRECURINTENS = PR.intensity;
        FAS = sumIntensity(FA1, FA2);;
```

10.7.7 MFQL query for Phosphatidylethanolamine (PE) in positive ion mode acquired with neutral loss scanning

Although it is acquired as neutral loss scan, the head group of PE is specified as fragment and not as neutral loss in the query. This is because the neutral loss scan spectra are stored in *.mzXML having *m/z* 141.02 as header.

```
QUERYNAME = Phosphatidylethanolamine;
DEFINE prPE = 'C[33..47] H[30..200] N[1] O[8] P[1]' WITH DBR = (0.5,7.5), CHG = 1;
DEFINE headPE = 'C2 H8 O4 N1 P1' WITH CHG = 1;
IDENTIFY
        prPE IN MS1+ AND
        headPE in MS2+
REPORT
        MASS = prPE.mass;
        NAME = "PE [%d:%d]" % "((prPE.chemsc)[C] - 5, (prPE.chemsc)[db] - 1.5)";
        CHEMSC = prPE.chemsc;
        ERROR = "%2.2fppm" % "(prPE.errppm)";
        INTENS = prPE.intensity;
        FRAGINTENS = headPE.intensity
```

10.7.8 MFQL query for Phosphatidylglycerol (PG) in positive ion mode acquired with neutral loss scanning

Although it is acquired as neutral loss scan, the head group of PG is specified as fragment and not as neutral loss in the query. This is because the neutral loss scan spectra are stored in *.mzXML having *m/z* 189.04 as header.

```
QUERYNAME = Phosphatidylglycerol;
DEFINE prPG = 'C[33..47] H[30..200] N[1] O[10] P[1]' WITH DBR = (0.5,7.5), CHG = 1;
DEFINE headPG = 'C3 H12 O6 P1 N1' WITH CHG = 1;
IDENTIFY
        prPG IN MS1+ AND
        headPG in MS2+
REPORT
        MASS = prPG.mass;
        NAME = "PG [%d:%d]" % "((prPG.chemsc)[C] - 6, (prPG.chemsc)[db] - 0.5)";
        CHEMSC = prPG.chemsc;
        ERROR = "%2.2fppm" % "(prPG.errppm)";
        ERRORMSMS = "%2.2fppm" % (headPG.errppm);
        INTENS = prPG.intensity;
        FRAGINTENS = headPG.intensity;;
```

10.8 MFQL queries used for Bovine Heart

10.8.1 MFQL queries for ceramide

```
QUERYNAME = ceramides;
DEFINE PR = 'C[32..44] H[30..100] N[1] O[3]' WITH DBR = (1.5,3.5), CHG = -1;
DEFINE PRA = 'C[34..42] H[30..100] N[1] O[5]' WITH DBR = (1.5,3.5), CHG = -1;
IDENTIFY
       PRA IN MS1- OR
        PR IN MS1-
REPORT
       MASS = "%4.4f" % "(PRA.mass)";
       CHEMSC = PRA.chemsc;
       ERROR = "%2.2fppm" % "(PRA.errppm)";
       NAME = "Cer [%d:%d]" % "((PRA.chemsc)[C]-2, (PRA.chemsc)[db] - 1.5)";
       PRECURINTENS = PRA.intensity;;
```

10.8.2 MFQL queries for cardiolipin (CL)

```
QUERYNAME = Cardiolipin;
DEFINE PR = 'C[65..87] H[90..180] O[17] P[2]' WITH DBR = (5,16), CHG = -2;
IDENTIFY
       PR IN MS1- WITH TOLERANCE = 5 ppm
SUCHTHAT
       isOdd(PR.chemsc[C])
REPORT
       MASS = "%4.4f" % "(PR.mass)";
       CHEMSC = PR.chemsc;
       ERROR = "%2.2fppm" % "(PR.errppm)";
       NAME = "CL [%d:%d]" % "((PR.chemsc)[C] - 9, (PR.chemsc)[db] - 5)";
       PRECURINTENS = PR.intensity;;
```

10.8.3 MFQL queries for diacylglycerol (DAG)

```
QUERYNAME = Diacylglycerol;
DEFINE PR = 'C[33..49] H[35..110] O[7]' WITH DBR = (2.5,9.5), CHG = -1;
IDENTIFY
       PR IN MS1-
SUCHTHAT
       isOdd(PR.chemsc[C])
REPORT
       MASS = "%4.4f" % "(PR.mass)";
       CHEMSC = PR.chemsc;
       ERROR = "%2.2fppm" % "(PR.errppm)";
       NAME = "DAG [%d:%d]" % "((PR.chemsc)[C] - 5, (PR.chemsc)[db] - 2.5)";
       PRECURINTENS = PR.intensity;;
```

10.8.4 MFQL queries for lyso phosphatic acid (LPA)

```
QUERYNAME = lysoPhosphatidicAcid;
DEFINE PR = 'C[17..25] H[30..70] O[7] P[1]' WITH DBR = (1.5,7.5), CHG = -1;
DEFINE FA1 ='C[14..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
IDENTIFY
       PR IN MS1- AND
       FA1 IN MS2-
```

```
SUCHTHAT
        isEven(FA1.chemsc[C]) AND
        FA1.chemsc + 'C3 H7 O5 P1' == PR.chemsc
REPORT
        MASS = "%4.4f" % "(PR.mass)";
        CHEMSC = PR.chemsc;
        ERROR = "%2.2fppm" % "(PR.errppm)";
        NAME = "LPA [%d:%d]" % "((PR.chemsc)[C] - 3, (PR.chemsc)[db] - 1.5)";
        PRECURINTENS = PR.intensity;
        FAS = FA1.intensity;;
```

10.8.5 MFQL queries for lyso phosphatidylcholine (LPC)

```
QUERYNAME = lysoPhosphatidylcholine;
DEFINE PR = 'C[24..32] H[30..80] O[9] N[1] P[1]' WITH DBR = (1.5,7.5), CHG = -1;
DEFINE headPC = 'C[3] H[6] O[2]' WITH CHG = 0;
DEFINE FA1 ='C[14..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
IDENTIFY
        PR IN MS1- AND
        FA1 in MS2- AND
        headPC in MS2- WITH TOLERANCE = 0.5 Da
SUCHTHAT
        isEven(PR.chemsc[C]) AND
        isEven(FA1.chemsc[C]) AND
        FA1.chemsc + headPC.nlsc + 'C7 H16 P1 O5 N1' == PR.chemsc
REPORT
        MASS = "%4.4f" % "(PR.mass)";
        CHEMSC = PR.chemsc;
        ERROR = "%2.2fppm" % "(PR.errppm)";
        NAME = "LPC [%d:%d]" % "((PR.chemsc)[C] - 10, (PR.chemsc)[db] - 1.5)";
        PRECURINTENS = PR.intensity;
        NLSPIS = headPC.intensity;
        FAS = FA1.intensity;;
```

10.8.6 MFQL queries for lyso phosphatidylethanolamine (LPE)

```
QUERYNAME = lysoPhosphatidylethanolamine;
DEFINE PR = 'C[19..27] H[30..70] O[7] N[1] P[1]' WITH DBR = (1.5,7.5), CHG = -1;
DEFINE FA1 ='C[14..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
IDENTIFY
        PR IN MS1- AND
        FA1 in MS2-
SUCHTHAT
        isOdd(PR.chemsc[C]) AND
        isEven(FA1.chemsc[C]) AND
        FA1.chemsc + 'C5 H12 O5 N1 P1' == PR.chemsc
REPORT
        MASS = "%4.4f" % "(PR.mass)";
        CHEMSC = PR.chemsc;
        ERROR = "%2.2fppm" % "(PR.errppm)";
        NAME = "LPE [%d:%d]" % "((PR.chemsc)[C] - 5, (PR.chemsc)[db] - 1.5)";
        PRECURINTENS = PR.intensity;
        FAS = FA1.intensity;;
```

10.8.7 MFQL queries for lyso phosphatidylglycerol (LPG)

```
QUERYNAME = lysoPhosphatidylglycerol;
DEFINE PR = 'C[20..28] H[30..70] O[9] P[1]' WITH DBR = (1.5,7.5), CHG = -1;
DEFINE FA1 ='C[14..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
IDENTIFY
       PR IN MS1- AND
       FA1 IN MS2-
SUCHTHAT
       isEven(FA1.chemsc[C]) AND
       FA1.chemsc + 'C6 H13 O7 P1' == PR.chemsc
REPORT
       MASS = "%4.4f" % "(PR.mass)";
       CHEMSC = PR.chemsc;
       ERROR = "%2.2fppm" % "(PR.errppm)";
       NAME = "LPG [%d:%d]" % "((PR.chemsc)[C] - 6, (PR.chemsc)[db] - 1.5)";
       PRECURINTENS = PR.intensity;
       FAS = FA1.intensity;;
```

10.8.8 MFQL queries for lyso phosphatidylinositol (LPI)

```
QUERYNAME = LysoPhosphatidylinositol;
DEFINE PR = 'C[37..53] H[30..140] O[12] P[1]' WITH DBR = (2.5,9.5), CHG = -1;
DEFINE headPI = 'C[6] H[10] O[8] P[1]' WITH DBR = (1.5,4.5), CHG = -1;
DEFINE FA1 ='C[14..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
IDENTIFY
       PR IN MS1- AND
       FA1 in MS2- AND
       headPI in MS2-
SUCHTHAT
       isOdd(PR.chemsc[C]) AND
       isEven(FA1.chemsc[C]) AND
       FA1.chemsc +  headPI.chemsc + 'C3 H7 O2' == PR.chemsc
REPORT
       MASS = "%4.4f" % "(PR.mass)";
       CHEMSC = PR.chemsc;
       ERROR = "%2.2fppm" % "(PR.errppm)";
       NAME = "LPI [%d:%d]" % "((PR.chemsc)[C] - 9, (PR.chemsc)[db] - 2.5)";
       PRECURINTENS = PR.intensity;
       NLSPIS = headPI.intensity;
       FAS = FA1.intensity;;
```

10.8.9 MFQL queries for lyso cardiolipin (LCL)

```
QUERYNAME = lysoCardiolipin;
DEFINE PR = 'C[51..69] H[90..180] O[16] P[2]' WITH DBR = (4,16), CHG = -2;
DEFINE PRA = 'C[51..69] H[90..180] O[16] P[2]' WITH DBR = (3.5,15.5), CHG = -1;
IDENTIFY lysoCardiolipin WHERE
       PR IN MS1- WITH TOLERANCE = 5 ppm OR
       PRA IN MS1- WITH TOLERANCE = 5 ppm
REPORT
       MASS = "%4.4f" % "(PR.mass)";
       CHEMSC = PR.chemsc;
        ERROR = "%2.2fppm" % "(PR.errppm)";
       NAME = "LysCL [%d:%d]" % "((PR.chemsc)[C] - 9, (PR.chemsc)[db] - 4)";
        PRECURINTENS = PR.intensity;;
```

10.8.10 MFQL queries for phosphatic acid (PA)

```
QUERYNAME = PhosphatidicAcid;
DEFINE PR = 'C[31..47] H[30..120] O[8] P[1]' WITH DBR = (2.5,8.5), CHG = -1;
DEFINE FA1 ='C[14..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
DEFINE FA2 ='C[14..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
IDENTIFY
       PR IN MS1- AND
       FA1 in MS2- AND
       FA2 in MS2-
SUCHTHAT
       isOdd(PR.chemsc[C]) AND
       isEven(FA1.chemsc[C]) AND
       isEven(FA2.chemsc[C]) AND
       FA1.chemsc + FA2.chemsc + 'C3 H6 P1 O4' == PR.chemsc
REPORT
       MASS = "%4.4f" % "(PR.mass)";
       CHEMSC = PR.chemsc;
       ERROR = "%2.2fppm" % "(PR.errppm)";
       NAME = "PA [%d:%d]" % "((PR.chemsc)[C] - 3, (PR.chemsc)[db] - 2.5)";
       SPECIE = "PA [%d:%d / %d:%d]" % "(FA1.chemsc[C], FA1.chemsc[db] - 1.5,
FA2.chemsc[C], FA2.chemsc[db] - 1.5)";
       PRECURINTENS = PR.intensity;
       FAS = sumIntensity(FA1, FA2);;
```

10.8.11 MFQL queries for phosphatidylcholine (PC)

```
QUERYNAME = Phosphatidylcholine;
DEFINE PR = 'C[38..54] H[30..130] O[10] N[1] P[1]' WITH DBR = (2.5,9.5), CHG = -1;
DEFINE headPC = 'C[3] H[6] O[2]' WITH CHG = 0;
DEFINE FA1 ='C[14..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
DEFINE FA2 ='C[14..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
IDENTIFY
       PR IN MS1- AND
       FA1 in MS2- AND
       FA2 in MS2- AND
       headPC in MS2- WITH TOLERANCE = 0.5 Da
SUCHTHAT
       isEven(PR.chemsc[C]) AND
       isEven(FA1.chemsc[C]) AND
       isEven(FA2.chemsc[C]) AND
       FA1.chemsc + FA2.chemsc + headPC.nlsc + 'C7 H15 P1 O4 N1' == PR.chemsc
REPORT
       MASS = "%4.4f" % "(PR.mass)";
       CHEMSC = PR.chemsc;
       ERROR = "%2.2fppm" % "(PR.errppm)";
       NAME = "PC [%d:%d]" % "((PR.chemsc)[C] - 10, (PR.chemsc)[db] - 2.5)";
       SPECIE = "PC [%d:%d / %d:%d]" % "(FA1.chemsc[C], FA1.chemsc[db] - 1.5,
FA2.chemsc[C], FA2.chemsc[db] - 1.5)";
       PRECURINTENS = PR.intensity;
       NLSPIS = headPC.intensity;
       FAS = sumIntensity(FA1, FA2);;
```

10.8.12 MFQL queries for phosphatidylcholine ether (PC-O)

```
QUERYNAME = Phosphatidylcholineether;
DEFINE PR = 'C[38..54] H[30..120] O[9] N[1] P[1]' WITH DBR = (1.5,8.5), CHG = -1;
```

```
DEFINE headPC = 'C[3] H[6] O[2]' WITH CHG = 0;
DEFINE FA1 ='C[14..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
DEFINE FAO ='C[21..29] H[20..80] O[6] N[1] P[1]' WITH DBR = (0.5,6.5), CHG = -1;
IDENTIFY
       PR IN MS1- AND
       FA1 in MS2- AND
       FAO in MS2- AND
       headPC in MS2- WITH TOLERANCE = 0.5 Da
SUCHTHAT
       isEven(PR.chemsc[C]) AND
       isEven(FA1.chemsc[C]) AND
       isOdd(FAO.chemsc[C]) AND
       FA1.chemsc + FAO.chemsc + headPC.nlsc == PR.chemsc + 'O1 H1'
REPORT
       MASS = "%4.4f" % "(PR.mass)";
       CHEMSC = PR.chemsc;
        ERROR = "%2.2fppm" % "(PR.errppm)";
       NAME = "PC-O [%d:%d]" % "((PR.chemsc)[C] - 10, (PR.chemsc)[db] - 1.5)";
       SPECIE = "PC-O [%d:%d / %d:%d]" % "(FAO.chemsc[C] - 7, FAO.chemsc[db] - 0.5,
FA1.chemsc[C], FA1.chemsc[db] - 1.5)";
       PRECURINTENS = PR.intensity;
       NLSPIS = headPC.intensity;
       FAS = sumIntensity(FA1, FAO);;
```

10.8.13 MFQL queries for phosphatidylethanolamine (PE)

```
QUERYNAME = Phosphatidylethanolamine;
DEFINE PR = 'C[33..49] H[50..100] O[8] N[1] P[1]' WITH DBR = (2.5,9.5), CHG = -1;
DEFINE FA1 ='C[14..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
DEFINE FA2 ='C[14..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
IDENTIFY
       PR IN MS1- AND
       FA1 in MS2- AND
       FA2 in MS2-
SUCHTHAT
       isOdd(PR.chemsc[C]) AND
       isEven(FA1.chemsc[C]) AND
       isEven(FA2.chemsc[C]) AND
       FA1.chemsc + FA2.chemsc + 'C5 H11 O4 N1 P1' == PR.chemsc
REPORT
       MASS = "%4.4f" % "(PR.mass)";
       CHEMSC = PR.chemsc;
       ERROR = "%2.2fppm" % "(PR.errppm)";
       NAME = "PE [%d:%d]" % "((PR.chemsc)[C] - 5, (PR.chemsc)[db] - 2.5)";
       SPECIE = "PE [%d:%d / %d:%d]" % "(FA1.chemsc[C], FA1.chemsc[db] - 1.5,
FA2.chemsc[C], FA2.chemsc[db] - 1.5)";
       PRECURINTENS = PR.intensity;
       FAS = sumIntensity(FA1, FA2);;
```

10.8.14 MFQL query for phosphatidylethanolamine ether (PE-O)

```
QUERYNAME = Phosphatidylethanolamineether;
DEFINE PR = 'C[33..49] H[50..100] O[7] N[1] P[1]' WITH DBR = (1.5,8.5), CHG = -1;
DEFINE FA1 ='C[14..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
DEFINE FAO ='C[19..27] H[20..80] O[6] N[1] P[1]' WITH DBR = (0.5,6.5), CHG = -1;
IDENTIFY
       PR IN MS1- WITH TOLERANCE = 2.5 ppm AND
```

```
       FA1 in MS2- WITH TOLERANCE = 0.3 Da AND
       FA0 in MS2- WITH TOLERANCE = 0.3 Da
SUCHTHAT
       isOdd(PR.chemsc[C]) AND
       isEven(FA1.chemsc[C]) AND
       isOdd(FA0.chemsc[C]) AND
       FA1.chemsc + FA0.chemsc == PR.chemsc + 'O1 H1'
REPORT
       MASS = "%4.4f" % "(PR.mass)";
       CHEMSC = PR.chemsc;
       ERROR = "%2.2fppm" % "(PR.errppm)";
       NAME = "PE-O [%d:%d]" % "((PR.chemsc)[C] - 5, (PR.chemsc)[db] - 1.5)";
       SPECIE = "PE-O [%d:%d / %d:%d]" % "(FA0.chemsc[C] - 5, FA0.chemsc[db] - 0.5,
FA1.chemsc[C], FA1.chemsc[db] - 1.5)";
       PRECURINTENS = PR.intensity;
       FAS = FA1.intensity + FA0.intensity;;
```

10.8.15 MFQL query for phosphatidylglycerol (PG)

```
QUERYNAME = Phosphatidylglycerol;
DEFINE PR = 'C[34..50] H[30..120] O[10] P[1]' WITH DBR = (2.5,8.5), CHG = -1;
DEFINE FA1 ='C[14..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
DEFINE FA2 ='C[14..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
IDENTIFY
       PR IN MS1-  AND
       FA1 in MS2- AND
       FA2 in MS2-
SUCHTHAT
       isEven(PR.chemsc[C]) AND
       isEven(FA1.chemsc[C]) AND
       isEven(FA2.chemsc[C]) AND
       FA1.chemsc + FA2.chemsc + 'C6 H12 P1 O6' == PR.chemsc
REPORT
       MASS = "%4.4f" % "(PR.mass)";
       CHEMSC = PR.chemsc;
       ERROR = "%2.2fppm" % "(PR.errppm)";
       NAME = "PG [%d:%d]" % "((PR.chemsc)[C] - 6, (PR.chemsc)[db] - 2.5)";
       SPECIE = "PG [%d:%d / %d:%d]" % "(FA1.chemsc[C], FA1.chemsc[db] - 1.5,
FA2.chemsc[C], FA2.chemsc[db] - 1.5)";
       PRECURINTENS = PR.intensity;
       FAS = sumIntensity(FA1, FA2);;
```

10.8.16 MFQL query for phosphatidylinositol (PI)

```
QUERYNAME = Phosphatidylinositol;
DEFINE PR = 'C[37..53] H[30..140] O[13] P[1]' WITH DBR = (3.5,10.5), CHG = -1;
DEFINE headPI = 'C[6] H[10] O[8] P[1]' WITH DBR = (1.5,4.5), CHG = -1;
DEFINE FA1 ='C[14..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
DEFINE FA2 ='C[14..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
IDENTIFY
       PR IN MS1- AND
       FA1 in MS2- AND
       FA2 in MS2- AND
       headPI in MS2-
SUCHTHAT
       isOdd(PR.chemsc[C]) AND
       isEven(FA1.chemsc[C]) AND
```

```
        isEven(FA2.chemsc[C]) AND
        FA1.chemsc + FA2.chemsc + headPI.chemsc + 'C3 H6 O1' == PR.chemsc
REPORT
        MASS = "%4.4f" % "(PR.mass)";
        CHEMSC = PR.chemsc;
        ERROR = "%2.2fppm" % "(PR.errppm)";
        NAME = "PI [%d:%d]" % "((PR.chemsc)[C] - 9, (PR.chemsc)[db] - 3.5)";
        SPECIE = "PI [%d:%d / %d:%d]" % "(FA1.chemsc[C], FA1.chemsc[db] - 1.5,
FA2.chemsc[C], FA2.chemsc[db] - 1.5)";
        PRECURINTENS = PR.intensity;
        NLSPIS = headPI.intensity;
        FAS = FA1.intensity + FA2.intensity;;
```

10.8.17 MFQL query for phosphatidylserine (PS)

```
QUERYNAME = Phosphatidylserine;
DEFINE PR = 'C[34..50] H[30..120] O[10] N[1] P[1]' WITH DBR = (3.5,10.5), CHG = -1;
DEFINE headPS = 'C[3] H[5] O[2] N[1]' WITH DBR = (-0.5,6.5), CHG = 0;
DEFINE FA1 ='C[14..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
DEFINE FA2 ='C[14..22] H[20..50] O[2]' WITH DBR = (1.5,7.5), CHG = -1;
IDENTIFY
        PR IN MS1- AND
        FA1 in MS2- AND
        FA2 in MS2- AND
        headPS in MS2- WITH TOLERANCE = 0.5 Da
SUCHTHAT
        isEven(PR.chemsc[C]) AND
        FA1.chemsc + FA2.chemsc + headPS.nlsc + 'C3 H6 P1 O4' == PR.chemsc
REPORT
        MASS = "%4.4f" % "(PR.mass)";
        CHEMSC = PR.chemsc;
        ERROR = "%2.2fppm" % "(PR.errppm)";
        NAME = "PS [%d:%d]" % "((PR.chemsc)[C] - 6, (PR.chemsc)[db] - 3.5)";
        SPECIE = "PS [%d:%d / %d:%d]" % "(FA1.chemsc[C], FA1.chemsc[db] - 1.5,
FA2.chemsc[C], FA2.chemsc[db] - 1.5)";
        PRECURINTENS = PR.intensity;
        NLSPIS = headPS.intensity;
        FAS = sumIntensity(FA1, FA2);;
```

10.8.18 MFQL query for sphingomyelin (SM)

```
QUERYNAME = Sphingomyelin;
DEFINE PR = 'C[39..49] H[30..130] O[8] N[2] P[1]' WITH DBR = (1.5,3.5), CHG = -1;
DEFINE headSMPC = 'C[3] H[6] O[2]' WITH CHG = 0;
IDENTIFY
        PR IN MS1- AND
        headSMPC in MS2- WITH TOLERANCE = 0.5 Da
SUCHTHAT
        isOdd(PR.chemsc[C])
REPORT
        MASS = "%4.4f" % "(PR.mass)";
        CHEMSC = PR.chemsc;
        ERROR = "%2.2fppm" % "(PR.errppm)";
        NAME = "SM [%d:%d]" % "((PR.chemsc)[C] - 7, (PR.chemsc)[db] - 1.5)";
        PRECURINTENS = PR.intensity;
        NLSPIS = headSMPC.intensity;;
```

10.8.19 MFQL query for triacylglycerol (TAG)

```
QUERYNAME = Triacylglycerol;
DEFINE PR = 'C[47..65] H[50..160] O[8]' WITH DBR = (3.5,10.5), CHG = -1;
IDENTIFY
       PR IN MS1- WITH TOLERANCE = 2.5 ppm
SUCHTHAT
        isOdd(PR.chemsc[C])
REPORT
       MASS = "%4.4f" % "(PR.mass)";
       CHEMSC = PR.chemsc;
        ERROR = "%2.2fppm" % "(PR.errppm)";
       NAME = "TAG [%d:%d]" % "((PR.chemsc)[C] - 5, (PR.chemsc)[db] - 3.5)";
       PRECURINTENS = PR.intensity;;
```

10.9 LipidXplorer's output of the identification of PA standards (Chapter 6.3)

5	With isotopic correction								
6	MASS	CHEMSC	ERROR	NAME	SPECIE	PRECURINTENS:sample1	PRECURINTENS:sample2	FAS:sample1	FAS:sample2
7	699.496	C39 H72 O8 P1	-1.42ppm	PA [36:2]	PA [18:0 / 18:2]	1569864.898	1601589.136	2.81E+04	2.96E+04
8	699.496	C39 H72 O8 P1	-1.42ppm	PA [36:2]	PA [18:1 / 18:1]	1569864.898	1601589.136	2.25E+05	2.38E+05
9	701.51	C39 H74 O8 P1	-3.83ppm	PA [36:1]	PA [18:0 / 18:1]	129600.0553	127682.0372	2.72E+04	2.97E+04
10	703.5259	C39 H76 O8 P1	-3.46ppm	PA [36:0]	PA [18:0 / 18:0]	140227.2752	147075.0099	2.41E+04	3.13E+04
11									
12									
13	Without isotopic correction								
14	MASS	CHEMSC	ERROR	NAME	SPECIE	PRECURINTENS:sample1	PRECURINTENS:sample2	FAS:sample1	FAS:sample2
15	699.496	C39 H72 O8 P1	-1.42ppm	PA [36:2]	PA [18:0 / 18:2]	1569864.898	1601589.136	28102.86475	29561.46338
16	699.496	C39 H72 O8 P1	-1.42ppm	PA [36:2]	PA [18:1 / 18:1]	1569864.898	1601589.136	224660.4375	238489.8906
17	701.51	C39 H74 O8 P1	-3.83ppm	PA [36:1]	PA [18:0 / 18:1]	311526.9722	313285.3806	42783.23242	46321.69922
18	703.5259	C39 H76 O8 P1	-3.46ppm	PA [36:0]	PA [18:0 / 18:0]	160311.7301	167038.917	25320.72266	32619.28906

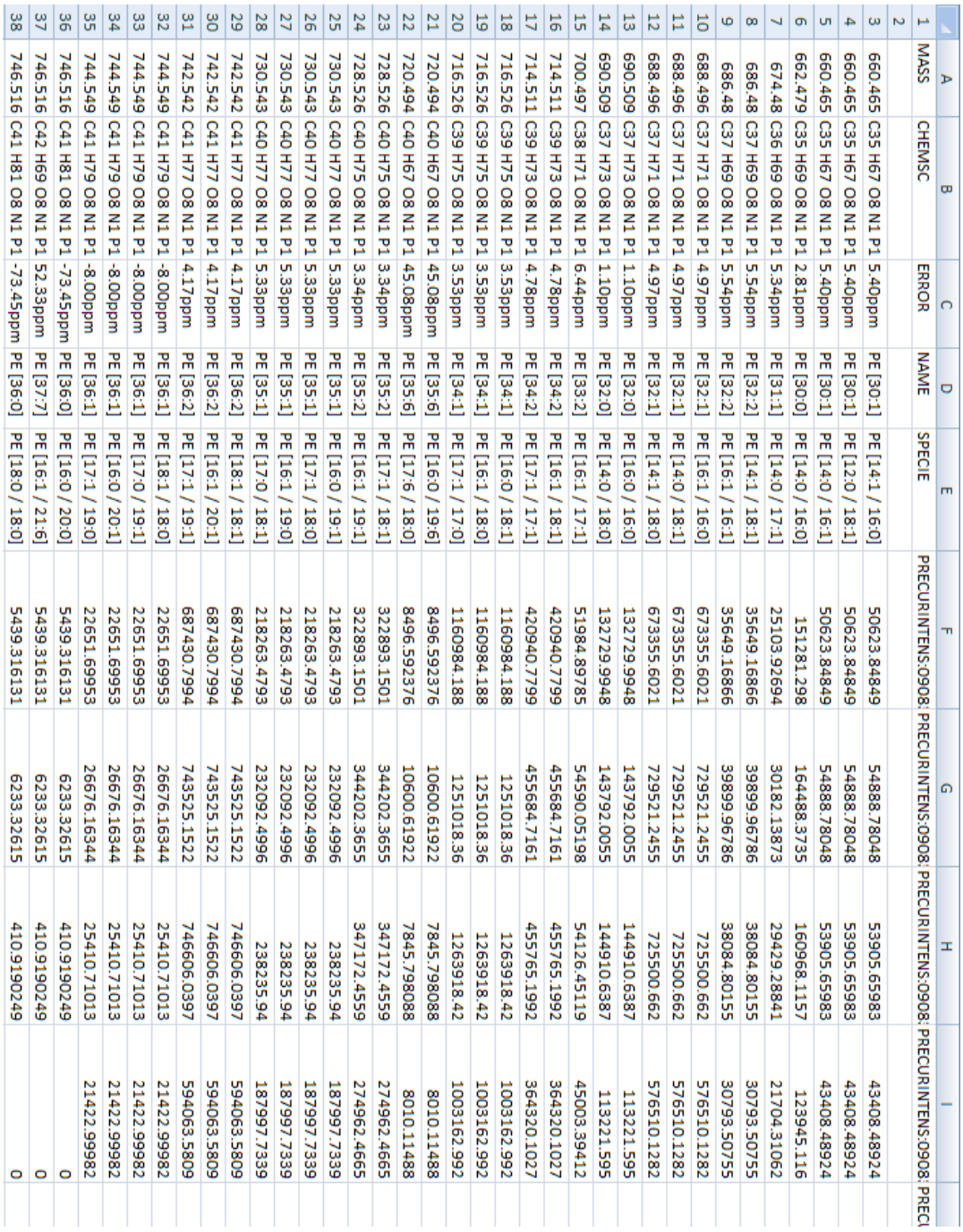
Figure A 7: The LipidXplorer output of the identification of PA standards.

The used MFQL queries can be found in Appendix <…>. The rows specified in the query are: MASS for the m/z value of the identified peak, CHEMSC for the sum composition of the identified lipid, ERROR for the distance of the theoretical m/z value to the experimental, NAME for the lipid name, SPECIE for the lipid species, PRECURINTENS contains the abundance of the intact ions and includes all examined samples (‘sample1’ and ‘sampel2’), FAS contains the abundance of the

sum of the identified peak of the acyl anions and includes also all examined samples.

10.10 The output of LipidXplorer

In the following a screenshot of the output of a LipidXplorer run in Microsoft Excel is shown. The spectra data set was acquired from E.coli on a LTQ Orbitrap XL with low resolution settings as in Materials and Methods 8.4 the experiment II. For a better view on the data, the picture is turned about 90 degrees.



	A	B	C	D	E	F	G	H	I
1	MASS	CHEMSC	ERROR	NAME	SPECIE	PRECURINTENS:0908:	PRECURINTENS:0908:	PRECURINTENS:0908:	PRECURINTENS:0908:	PRECI
2										
3	660.465	C35 H67 O8 N1 P1	5.40ppm	PE [30:1]	PE [14:1 / 16:0]	50623.84849	54888.78048	53905.65983	43408.48924	
4	660.465	C35 H67 O8 N1 P1	5.40ppm	PE [30:1]	PE [12:0 / 18:1]	50623.84849	54888.78048	53905.65983	43408.48924	
5	660.465	C35 H67 O8 N1 P1	5.40ppm	PE [30:1]	PE [14:0 / 16:1]	50623.84849	54888.78048	53905.65983	43408.48924	
6	662.479	C35 H69 O8 N1 P1	2.81ppm	PE [30:0]	PE [14:0 / 16:0]	151281.298	164488.3735	160968.1157	123945.116	
7	674.48	C36 H69 O8 N1 P1	5.34ppm	PE [31:1]	PE [14:0 / 17:1]	25103.92694	30182.13873	29429.28841	21704.31062	
8	686.48	C37 H69 O8 N1 P1	5.54ppm	PE [32:2]	PE [14:1 / 18:1]	35649.16866	39899.96786	38084.80155	30793.50755	
9	686.48	C37 H69 O8 N1 P1	5.54ppm	PE [32:2]	PE [16:1 / 16:1]	35649.16866	39899.96786	38084.80155	30793.50755	
10	688.496	C37 H71 O8 N1 P1	4.97ppm	PE [32:1]	PE [16:1 / 16:0]	673355.6021	729521.2455	725500.662	576510.1282	
11	688.496	C37 H71 O8 N1 P1	4.97ppm	PE [32:1]	PE [14:0 / 18:1]	673355.6021	729521.2455	725500.662	576510.1282	
12	688.496	C37 H71 O8 N1 P1	4.97ppm	PE [32:1]	PE [14:1 / 18:0]	673355.6021	729521.2455	725500.662	576510.1282	
13	690.509	C37 H73 O8 N1 P1	1.10ppm	PE [32:0]	PE [16:0 / 16:0]	132729.9948	143792.0055	144910.6387	113221.595	
14	690.509	C37 H73 O8 N1 P1	1.10ppm	PE [32:0]	PE [14:0 / 18:0]	132729.9948	143792.0055	144910.6387	113221.595	
15	700.497	C38 H71 O8 N1 P1	6.44ppm	PE [33:2]	PE [16:1 / 17:1]	51984.89785	54590.05198	54126.45119	45003.39412	
16	714.511	C39 H73 O8 N1 P1	4.78ppm	PE [34:2]	PE [16:1 / 18:1]	420940.7799	455684.7161	455765.1992	364320.1027	
17	714.511	C39 H73 O8 N1 P1	4.78ppm	PE [34:2]	PE [17:1 / 17:1]	420940.7799	455684.7161	455765.1992	364320.1027	
18	716.526	C39 H75 O8 N1 P1	3.53ppm	PE [34:1]	PE [16:0 / 18:1]	1160984.188	1251018.36	1263918.42	1003162.992	
19	716.526	C39 H75 O8 N1 P1	3.53ppm	PE [34:1]	PE [16:1 / 18:0]	1160984.188	1251018.36	1263918.42	1003162.992	
20	716.526	C39 H75 O8 N1 P1	3.53ppm	PE [34:1]	PE [17:1 / 17:0]	1160984.188	1251018.36	1263918.42	1003162.992	
21	720.494	C40 H67 O8 N1 P1	45.08ppm	PE [35:6]	PE [16:0 / 19:6]	8496.592376	10600.61922	7845.798088	8010.11488	
22	720.494	C40 H67 O8 N1 P1	45.08ppm	PE [35:6]	PE [17:6 / 18:0]	8496.592376	10600.61922	7845.798088	8010.11488	
23	728.526	C40 H75 O8 N1 P1	3.34ppm	PE [35:2]	PE [17:1 / 18:1]	322893.1501	344202.3655	347172.4559	274962.4665	
24	728.526	C40 H75 O8 N1 P1	3.34ppm	PE [35:2]	PE [16:1 / 19:1]	322893.1501	344202.3655	347172.4559	274962.4665	
25	730.543	C40 H77 O8 N1 P1	5.33ppm	PE [35:1]	PE [16:0 / 19:1]	218263.4793	232092.4996	238235.94	187997.7339	
26	730.543	C40 H77 O8 N1 P1	5.33ppm	PE [35:1]	PE [17:1 / 18:0]	218263.4793	232092.4996	238235.94	187997.7339	
27	730.543	C40 H77 O8 N1 P1	5.33ppm	PE [35:1]	PE [16:1 / 19:0]	218263.4793	232092.4996	238235.94	187997.7339	
28	730.543	C40 H77 O8 N1 P1	5.33ppm	PE [35:1]	PE [17:0 / 18:1]	218263.4793	232092.4996	238235.94	187997.7339	
29	742.542	C41 H77 O8 N1 P1	4.17ppm	PE [36:2]	PE [18:1 / 18:1]	687430.7994	743525.1522	746606.0397	594063.5809	
30	742.542	C41 H77 O8 N1 P1	4.17ppm	PE [36:2]	PE [16:1 / 20:1]	687430.7994	743525.1522	746606.0397	594063.5809	
31	742.542	C41 H77 O8 N1 P1	4.17ppm	PE [36:2]	PE [17:1 / 19:1]	687430.7994	743525.1522	746606.0397	594063.5809	
32	744.549	C41 H79 O8 N1 P1	-8.00ppm	PE [36:1]	PE [18:1 / 18:0]	22651.69953	26676.16344	25410.71013	21422.99982	
33	744.549	C41 H79 O8 N1 P1	-8.00ppm	PE [36:1]	PE [17:0 / 19:1]	22651.69953	26676.16344	25410.71013	21422.99982	
34	744.549	C41 H79 O8 N1 P1	-8.00ppm	PE [36:1]	PE [16:0 / 20:1]	22651.69953	26676.16344	25410.71013	21422.99982	
35	744.549	C41 H79 O8 N1 P1	-8.00ppm	PE [36:1]	PE [17:1 / 19:0]	22651.69953	26676.16344	25410.71013	21422.99982	
36	746.516	C41 H81 O8 N1 P1	-73.45ppm	PE [36:0]	PE [16:0 / 20:0]	5439.316131	6233.32615	410.9190249	0	
37	746.516	C42 H69 O8 N1 P1	52.33ppm	PE [37:7]	PE [16:1 / 21:6]	5439.316131	6233.32615	410.9190249	0	
38	746.516	C41 H81 O8 N1 P1	-73.45ppm	PE [36:0]	PE [18:0 / 18:0]	5439.316131	6233.32615	410.9190249	0	

Figure A 8: A screenshot of a LipidXplorer result.

The query for phosphatidylethanolamine (PE) lipid species was used on an *E.coli* sample acquired in low resolution negative ion mode on a LTQ Orbitrap. The query can be found in Appendix 10.7.1. The output includes the columns for the lipid's *m/z* value ('MASS'), its chemical sum composition ('CHEMSC'), the error to the exact *m/z* value in parts-per-million ppm ('ERROR'), the lipid name ('NAME'), the lipid species ('SPECIES') and the list of abundances for each sample ('PRECURINT').

10.11 The reference list for the LipidXplorer benchmark (Chapter 6.4)

In the following the hand annotated reference list is presented containing all lipid species which have to be found by any lipid identification software in *E.coli*. The list was intersected with the results of LipidProfiler as specified in Chapter 6.4.

Table 12
The lipid reference list compared to the identifications of the tested tools (see Chapter 6.4)

				Matching lipid species			
	Phosphatidylethanolamine			LipidMaps	LipidSearch	LipidQA	LipidXplorer
	m/z	lipid name	lipid species				
1	660.46	PE 30:1	PE 14:0 / 16:1	x	x		x
2	660.46	PE 30:1	PE 16:0 / 14:1	x			x
3	662.48	PE 30:0	PE 14:0 / 16:0	x	x	x	x
4	674.48	PE 31:1	PE 14:0 / 17:1	x	x	x	x
5	686.48	PE 32:2	PE 14:1 / 18:1	x			x
6	686.48	PE 32:2	PE 16:1 / 16:1	x	x	x	x
7	688.49	PE 32:1	PE 14:0 / 18:1	x			x
8	688.49	PE 32:1	PE 16:0 / 16:1	x	x	x	x
9	690.51	PE 32:0	PE 16:0 / 16:0	x	x	x	x
10	700.49	PE 33:2	PE 16:1 /17:1	x	x	x	x
11	702.51	PE 33:1	PE 16:0 / 17:1	x	x	x	x
12	714.51	PE 34:2	PE 16:1 / 18:1	x	x	x	x
13	714.51	PE 34:2	PE 17:1 / 17:1	x			x
14	716.52	PE 34:1	PE 16:0 / 18:1	x	x	x	x
15	716.52	PE 34:1	PE 16:1 / 18:0	x			x
16	728.52	PE 35:2	PE 16:1 / 19:1				x
17	728.52	PE 35:2	PE 17:1 / 18:1	x	x	x	x
18	730.54	PE 35:1	PE 16:0 / 19:1		x	x	x
19	742.54	PE 36:2	PE 18:1 /18:1	x	x	x	x
20	756.56	PE 37:2	PE 18:1 / 19:1		x	x	x
21	770.5	PE 38:2	PE 18:1 / 20:1	x			x
	Phosphatidylglycerol						
	m/z	lipid name	lipid species				
1	693.47	PG 30:0	PG 14:0 / 16:0	x	x	x	x
2	719.48	PG 32:1	PG 16:0 / 16:1	x	x	x	x
3	719.48	PG 32:1	PG 14:0 / 18:1	x			x
4	733.5	PG 33:1	PG 16:0 / 17:1	x	x	x	x
5	745.5	PG 34:2	PG 16:1 / 18:1	x	x	x	x
6	745.5	PG 34:2	PG 17:1 / 17:1	x			x
7	747.52	PG 34:1	PG 16:1 / 18:0	x			x
8	747.52	PG 34:1	PG 16:0 / 18:1	x	x	x	x
9	759.52	PG 35:2	PG 17:1 / 18:1	x	x	x	x
10	759.52	PG 35:2	PG 16:1 / 19:1	x			x
11	773.54	PG 36:2	PG 17:1 / 19:1				x
12	773.54	PG 36:2	PG 18:1 / 18:1	x	x	x	x
13	787.5	PG 37:2	PG 18:1 / 19:1		x	x	x
14	801.5	PG 38:2	PG 18:1 / 20:1		x	x	x

15	801.5	PG 38:2	PG 19:1 / 19:1	x	x	x

10.12 Comparison of lipid profiles from spectra acquired on different mass spectrometers

In Chapter 6.7 *E.coli* extracts were acquired at different mass spectrometers (LTQ Orbitrap Velos and ABI QStar) and analyzed with the same queries for PE and PG to show that LipidXplorer returns the same lipid profile among different platforms. Additionally to Table 8 and Figure 21, a pair-wise correlation of the used platforms is shown in the following. Thereby, MS and MS/MS profiles are correlated within and between the used platforms. This results in the following correlations:

- Orbitrap MS *vs* Orbitrap MS/MS
- QSTAR MS *vs* QSTAR MS/MS
- Orbitrap MS *vs* QSTAR MS/MS
- QSTAR MS *vs* Orbitrap MS/ MS
- Orbitrap MS/ MS *vs* QSTAR MS/MS

10.12.1 Orbitrap Velos MS vs Orbitrap Velos MS/MS

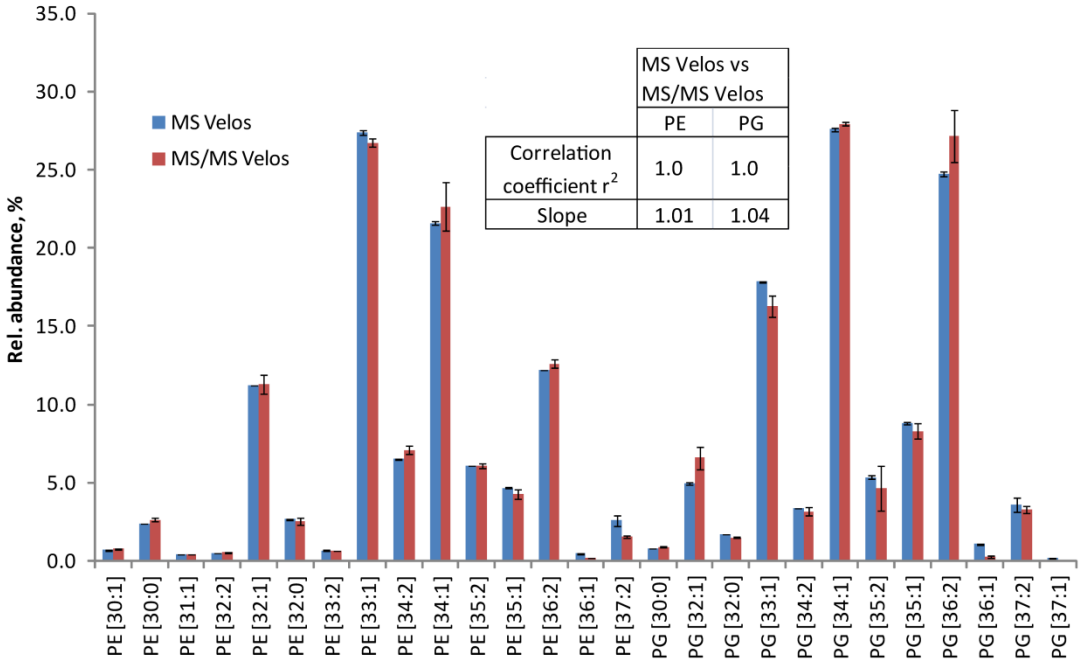


Figure A 9: Correlation of the lipid profiles of the MS and MS/MS acquisitions at the LTQ Orbitrap Velos.

Additionally the slope and the R^2 correlation coefficient between the lipid profiles of PG and PE, respectively are given in a table. (Source: (Herzog, et al., 2012))

10.12.2 QSTAR MS vs QSTAR MS/MS

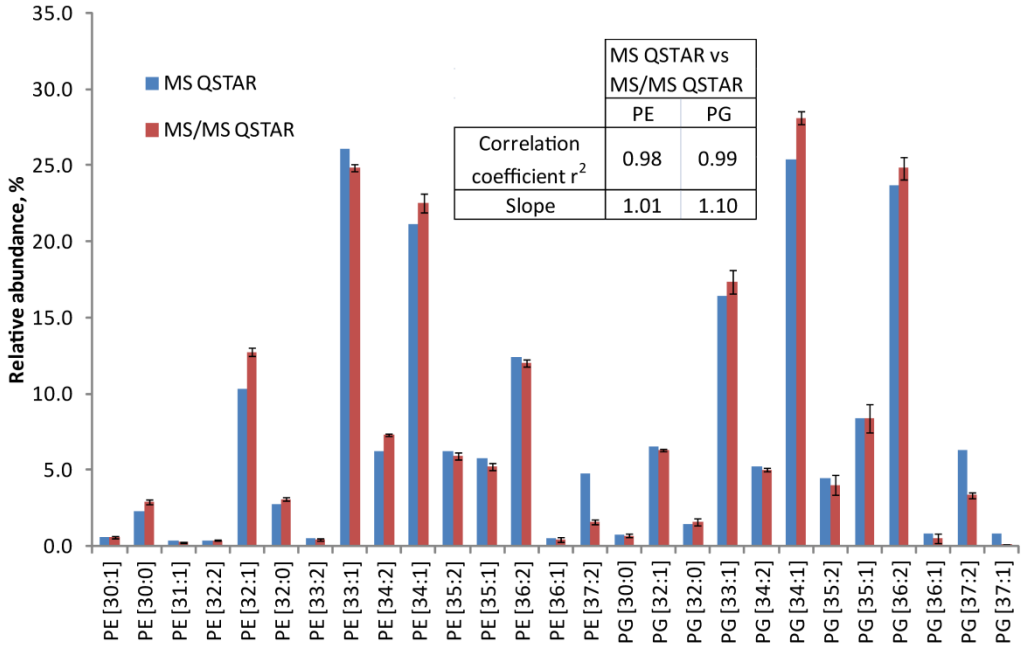


Figure A 10: Correlation of the lipid profiles of the MS and MS/MS acquisitions at the ABI QSTAR.

Additionally the slope and the R^2 correlation coefficient between the lipid profiles of PG and PE, respectively are given in a table. (Source:(Herzog, et al., 2012))

10.12.3 Orbitrap Velos MS vs QSTAR MS/MS

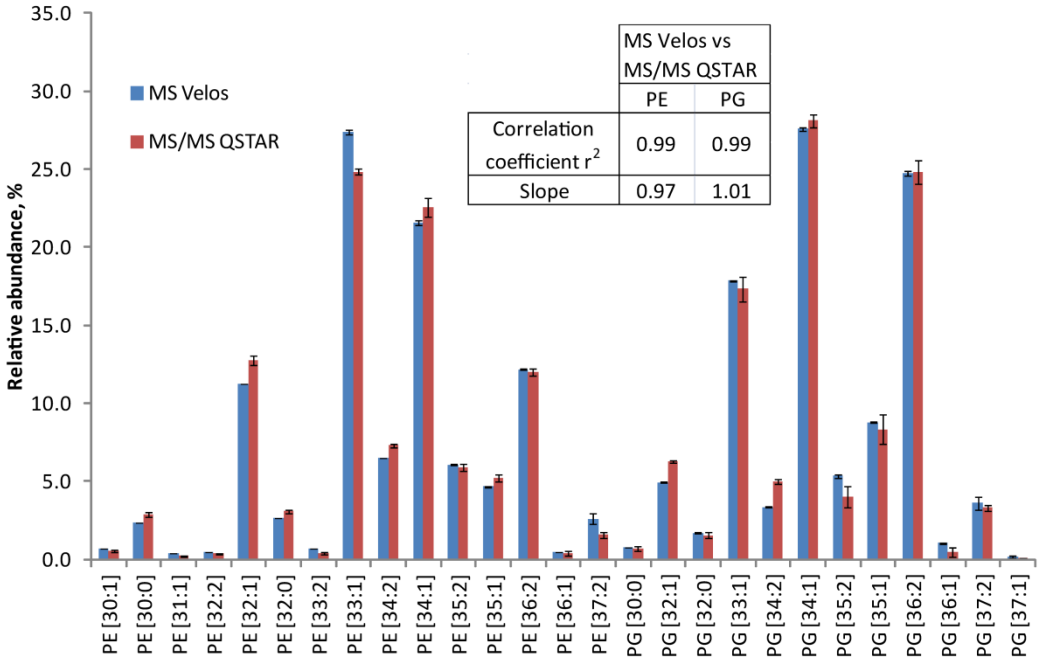


Figure A 11: Correlation between the lipid profile acquired at the Orbitrap Velos in MS and at the QSTAR in MS/MS mode.

Additionally the slope and the R^2 correlation coefficient between the lipid profiles of PG and PE, respectively are given in a table. (Source:(Herzog, et al., 2012))

10.12.4 QSTAR MS vs Orbitrap Velos MS/MS

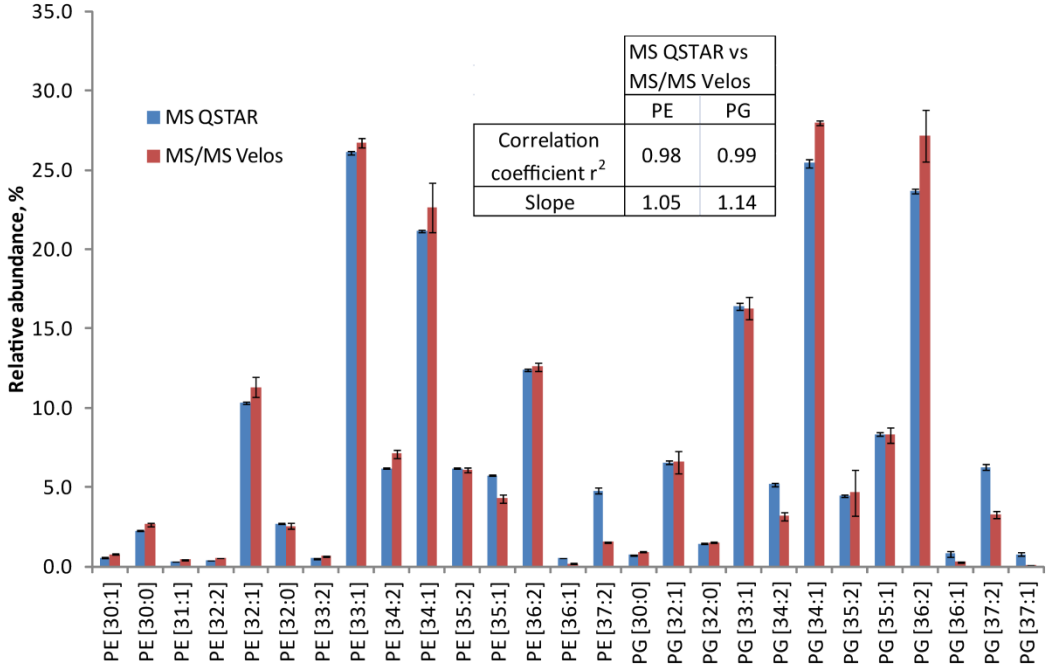


Figure A 12: Correlation between the lipid profiles acquired at the Orbitrap Velos in MS/MS and at the QSTAR in MS mode.

Additionally the slope and the R^2 correlation coefficient between the lipid profiles of PG and PE, respectively are given in a table. (Source: (Herzog, et al., article in press; Herzog, et al., 2012))

10.12.5 QSTAR MS/MS vs Orbitrap Velos MS/MS

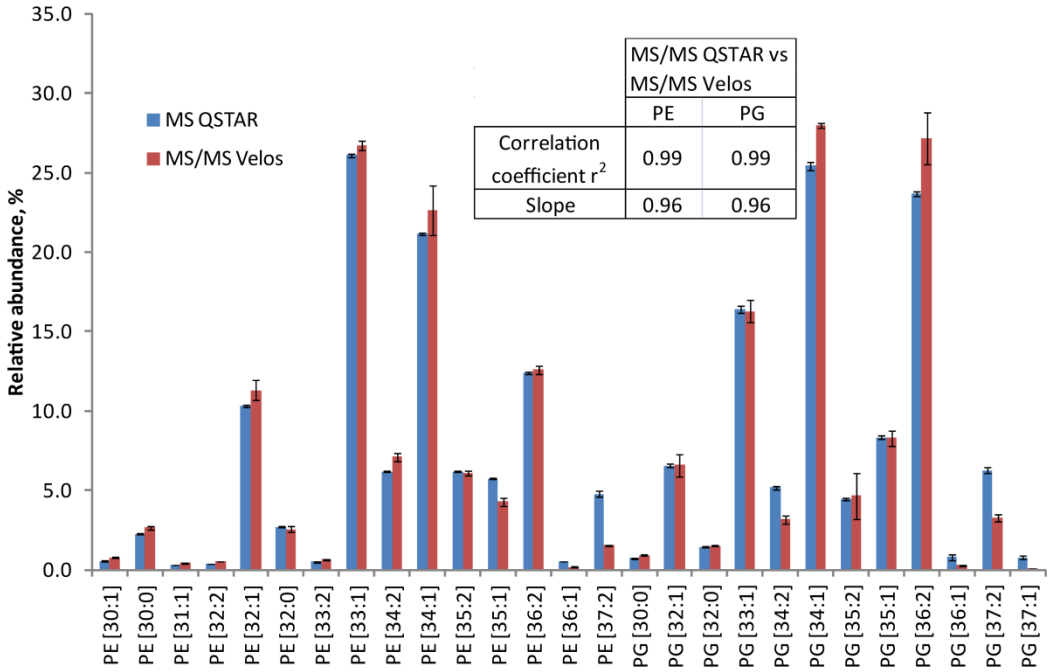


Figure A 13: Correlation between the lipid profiles acquired at the QSTAR in MS and at the Orbitrap Velos in MS/MS mode.

Additionally the slope and the R^2 correlation coefficient between the lipid profiles of PG and PE, respectively are given in a table. (Source: (Herzog, et al., 2012))

11 Additional files

Nb.	Description	Link to the content of the CD-ROM
1	Comparison of 325 spectra averaged by Xcalibur and LipidXplorer	Additional file 1
2	List of averaged masses for validating the averaging algorithm	Additional file 2
3	List of identified lipids and PCF for validating the alignment algorithm	Additional file 3
4	List of identified PA species for the validation of LipidXplorer's isotopic correction algorithm	Additional file 4
5	Lipid species identified in shotgun spectra of *E.coli* extracts by different software	Additional file 5
6	Lipid species identified by LipidXplorer in *E.coli* extract from spectra acquired at different mass resolution settings	Additional file 6
7	Correlation of relative abundances of *E.coli* lipid species determined in MS and MS/MS modes at the independent experiments at the QSTAR and Orbitrap mass spectrometers	Additional file 7
8	Correlation of relative abundances of *E.coli* lipid species determined in MS and MS/MS modes at the independent experiments at the QSTAR Orbitrap and Vantage mass spectrometers in different acquisition modes.	Additional file 8
9	Correlation of relative abundances of *E.coli* lipid species determined in positive and negative precursor ion and neutral loss scan on the TSQ Vantage, respectively.	Additional file 9
10	Lipid species identified by LipidXplorer in the bovine heart extract	Additional file 10

Bibliography

Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics, *Nature*, **422**, 198-207.

Balogh, M.P. (2004) Debating Resolution and Mass

Accuracy in Mass Spectrometry, *Spectroscopy*, **19**.

Blanksby, S. and Mitchell, T. (2010) Advances in Mass Spectrometry for Lipidomics, *Annual Review of Analytical Chemistry*, **3**, 433-465.

Bloch, F. (1946) Nuclear Induction, *Physical Review*, **70**, 460-474.

Böcker, S. and Rasche, F. (2008) Towards de novo identification of metabolites by analyzing tandem mass spectra, *Bioinformatics*, **24**, i49-i55.

Böhlke, J.K.*, et al.* (2005) Isotopic Compositions of the Elements, 2001, *Journal of Physical and Chemical Reference Data*, **34**, 57-67.

Borland, L., Brickhouse, M. and Fountain, A.W. (2010) Review of chemical signature databases, *Anal Bioanal Chem*, **397**, 1019-1028.

Brügger, B.*, et al.* (1997) Quantitative analysis of biological membrane lipids at the low picomole level by nano-electrospray ionization tandem mass spectrometry, *Proceedings of the National Academy of Sciences*, **94**, 2339-2344.

Carvalho, M.*, et al.* (2010) Survival strategies of a sterol auxotroph, *Development*, **137**, 3675-3685.

Castillo, S.*, et al.* (2011) Algorithms and tools for the preprocessing of LC–MS metabolomics data, *Chemometrics and Intelligent Laboratory Systems*, **108**, 23-32.

Chernushevich, I.V., Loboda, A.V. and Thomson, B.A. (2001) An introduction to quadrupole–time-of-flight mass spectrometry, *Journal of Mass Spectrometry*, **36**, 849-865.

Christie, W.W. (2003) *Lipid Analysis*. Oily Press, Bridgewater, UK.

Codd, E.F. (1970) A relational model of data for large shared data banks, *Commun. ACM*, **13**, 377-387.

Comisarow, M.B. and Marshall, A.G. (1974) Fourier transform ion cyclotron resonance spectroscopy, *Chemical Physics Letters*, **25**, 282-283.

Converse, S.E.*, et al.* (2003) MmpL8 is required for sulfolipid-1 biosynthesis and Mycobacterium tuberculosis virulence, *Proceedings of the National Academy of Sciences*, **100**, 6121-6126.

Cox, J. and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification, *Nat Biotech*, **26**, 1367-1372.

Denizot, Y.*, et al.* (2001) Is there a role of platelet-activating factor in human lung cancer?, *Lung Cancer*, **33**, 195-202.

Dennis, E.A. (2009) Lipidomics joins the omics evolution, *Proc Natl Acad Sci U S A*, **106**, 2089-2090.

Deutsch, E.W.*, et al.* (2010) A guided tour of the Trans-Proteomic Pipeline, *PROTEOMICS*, **10**, 1150-1159.

Dudley, E.*, et al.* (2010) Targeted metabolomics and mass spectrometry, *Advances in protein chemistry and structural biology*, **80**.

Ejsing, C.S.*, et al.* (2006) Automated Identification and Quantification of Glycerophospholipid Molecular Species by Multiple Precursor Ion Scanning, *Analytical Chemistry*, **78**.

Ejsing, C.S.*, et al.* (2009) Global analysis of the yeast lipidome by quantitative shotgun mass spectrometry, *Proc Natl Acad Sci U S A*, **106**, 2136-2141.

Ekroos, K.*, et al.* (2002) Quantitative profiling of phospholipids by multiple precursor ion scanning on a hybrid quadrupole time-of-flight mass spectrometer, *Anal Chem*, **74**, 941-949.

Ekroos, K.*, et al.* (2003) Charting molecular composition of phosphatidylcholines by fatty acid scanning and ion trap MS3 fragmentation, *J Lipid Res*, **44**, 2181-2192.

Eriksson, J., Chait, B.T. and Fenyö, D. (2000) A Statistical Basis for Testing the Significance of Mass Spectrometric Protein Identification Results, *Analytical Chemistry*, **72**, 999-1005.

Fahy, E.*, et al.* (2007) Bioinformatics for Lipidomics. In, *Lipidomics and Bioactive Lipids: Mass‐Spectrometry‐Based Lipid Analysis*. pp. 247-273.

Fahy, E.*, et al.* (2005) A comprehensive classification system for lipids, *Journal of Lipid Research*, **46**, 839-862.

Feng, L. and Prestwich, G.D. (2005) *Functional lipidomics*. Dekker-CRC, New York.

Fenn, J.B.*, et al.* (1989) Electrospray ionization for mass spectrometry of large biomolecules, *Science*, **246**, 64-71.

Frank, A.*, et al.* (2008) Clustering Millions of Tandem Mass Spectra, *Journal of Proteome Research*, **7**, 113-122.

Frank, A.*, et al.* (2008) Interpreting Top-Down Mass Spectra Using Spectral Alignment, *Anal Chem*.

Fridriksson, E.K.*, et al.* (1999) Quantitative Analysis of Phospholipids in Functionally Important Membrane Domains from RBL-2H3 Mast Cells Using Tandem High-Resolution Mass Spectrometry†, *Biochemistry*, **38**, 8056-8063.

Geiger, O.*, et al.* (2010) Amino acid-containing membrane lipids in bacteria, *Progress in Lipid Research*, **49**, 46-60.

Gerl, M.J.*, et al.* (2012) Quantitative analysis of the lipidomes of the influenza virus envelope and MDCK cell apical membrane, *The Journal of Cell Biology*.

Geurts, P.*, et al.* (2005) Proteomic mass spectra classification using decision tree based ensemble methods, *Bioinformatics*, **21**, 3138-3145.

Gower, J.C. and Legendr, P. (1986) Metric and Euclidean properties of dissimilarity coefficients, *J Classif*, **3**, 5-48.

Grabmeier, J. and Rudolph, A. (2002) Techniques of Cluster Algorithms in Data Mining, *Data Mining and Knowledge Discovery*, **6**, 303-360.

Graessler, J.*, et al.* (2009) Top-Down Lipidomics Reveals Ether Lipid Deficiency in Blood Plasma of Hypertensive Patients, *PLoS One*, **4**.

Gras, R.*, et al.* (1999) Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection, *Electrophoresis*, **20**, 3535-3550.

Graumann, J.*, et al.* (2011) A framework for intelligent data acquisition and real-time database searching for shotgun proteomics, *Molecular & Cellular Proteomics*.

Gross, J. (2004) *Mass Spectrometry: A textbook*. Springer, Berlin.

Gross, R. and Han, X. (2009) Shotgun lipidomics of neutral lipids as an enabling technology for elucidation of lipid-related diseases, *Am J Physiol Endocrinol Metab*, **297**, E297-303.

Gross, Richard W. and Han, X. (2011) Lipidomics at the Interface of Structure and Function in Systems Biology, *Chemistry & Biology*, **18**, 284-291.

Haimi, P.*, et al.* (2006) Software tools for analysis of mass spectrometric lipidome data, *Anal Chem*, **78**, 8324-8331.

Han, X. and Gross, R. (2003) Global analyses of cellular lipidomes directly from crude extracts of biological samples by ESI mass spectrometry: a bridge to lipidomics, *J. Lipid Res.*, **44**, 1071-1079.

Han, X. and Gross, R. (2005) Shotgun lipidomics: multidimensional MS analysis of cellular lipidomes, *Expert Rev Proteomics*, **2**, 253-264.

Han, X. and Gross, R.W. (1994) Electrospray ionization mass spectroscopic analysis of human erythrocyte plasma membrane phospholipids, *Proceedings of the National Academy of Sciences*, **91**, 10635-10639.

Han, X. and Gross, R.W. (2001) Quantitative analysis and molecular species fingerprinting of triacylglyceride molecular species directly from lipid extracts of biological samples by electrospray ionization tandem mass spectrometry, *Anal Biochem*, **295**, 88-100.

Han, X. and Gross, R.W. (2003) Global analyses of cellular lipidomes directly from crude extracts of biological samples by ESI mass spectrometry: a bridge to lipidomics, *J Lipid Res*, **44**, 1071-1079.

Han, X. and Gross, R.W. (2005) Shotgun lipidomics: electrospray ionization mass spectrometric analysis and quantitation of cellular lipidomes directly from crude extracts of biological samples, *Mass Spectrom Rev*, **24**, 367-412.

Han, X., Yang, K. and Gross, R.W. (2011) Multi-dimensional mass spectrometry-based shotgun lipidomics and novel strategies for lipidomic analyses, *Mass Spectrometry Reviews*, n/a-n/a.

Han, X.*, et al.* (2006) Factors influencing the electrospray intrasource separation and selective ionization of glycerophospholipids, *J Am Soc Mass Spectrom*, **17**, 264-274.

Harkewicz, R. and Dennis, E.A. (2011) Applications of Mass Spectrometry to Lipids and Membranes, *Annual Review of Biochemistry*, **80**, 301-325.

Heinonen, M.*, et al.* (2008) FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data, *Rapid Communications in Mass Spectrometry*, **22**, 3043-3052.

Herzog, R.*, et al.* (article in press) LipidXplorer: a software for consensual cross-platform lipidomics, *PLoS One*.

Herzog, R.*, et al.* (2012) LipidXplorer: A Software for Consensual Cross-Platform Lipidomics, *PLoS One*, **7**, e29851.

Herzog, R.*, et al.* (2011) A novel informatics concept for high-throughput shotgun lipidomics based on the molecular fragmentation query language, *Genome Biology*, **12**, R8.

Higgins, D.G. and Sharp, P.M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer, *Gene*, **73**, 237-244.

Hill, A.W. and Mortishire-Smith, R.J. (2005) Automated assignment of high-resolution collisionally activated dissociation mass spectra using a systematic bond disconnection approach, *Rapid Communications in Mass Spectrometry*, **19**, 3111-3118.

Hill, D.W.*, et al.* (2008) Mass Spectral Metabonomics beyond Elemental Formula: Chemical Database Querying by Matching Experimental with Computational Fragmentation Spectra, *Analytical Chemistry*, **80**, 5574-5582.

Horai, H.*, et al.* (2009) Traceable Analysis of Multiple-Stage Mass Spectra through Precursor-Product Annotations, *Proceedings of GCB' 2009*, 173-178.

Hsu, F.-F. and Turk, J. (2000) Charge-driven fragmentation processes in diacyl glycerophosphatidic acids upon low-energy collisional activation. A mechanistic proposal, *Journal of The American Society for Mass Spectrometry*, **11**, 797-803.

Hsu, F.-F. and Turk, J. (2009) Electrospray ionization with low-energy collisionally activated dissociation tandem mass spectrometry of glycerophospholipids: Mechanisms of fragmentation and structural characterization, *Journal of Chromatography B*, **877**, 2673-2695.

Hübner, G., Crone, C. and Lindner, B. (2009) lipID - a software tool for automated assignment of lipids in mass spectra, *Journal of Mass Spectrometry*, **44**, 1676-1683.

Hughey, C.*, et al.* (2001) Kendrick Mass Defect Spectrum: A Compact Visual Analysis for Ultrahigh-Resolution Broadband Mass Spectra, *Analytical Chemistry*, **73**, 4676-4681.

Ishida, M.*, et al.* (2004) High-resolution analysis by nano-electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry for the identification of molecular species of phospholipids and their oxidized metabolites, *Rapid Communications in Mass Spectrometry*, **18**, 2486-2494.

Ivanova, P.T.*, et al.* (2001) Electrospray ionization mass spectrometry analysis of changes in phospholipids in RBL-2H3 mastocytoma cells during degranulation, *Proceedings of the National Academy of Sciences*, **98**, 7152-7157.

Ivanova, P.T.*, et al.* (2007) Glycerophospholipid Identification and Quantitation by Electrospray Ionization Mass Spectrometry. In Brown, H.A. (ed), *Methods in Enzymology*. Academic Press, pp. 21-57.

James A, Y. (1983) A general approach to calculating isotopic distributions for mass spectrometry, *International Journal of Mass Spectrometry and Ion Physics*, **52**, 337-349.

Jeffries, N. Algorithms for alignment of mass spectrometry proteomic data, *Bioinformatics*, **21**, 3066-3073.

Jones, J.J., Batoy, S.M.A.B. and Wilkins, C.L. (2005) A comprehensive and comparative analysis for MALDI FTMS lipid and phospholipid profiles from biological samples, *Computational Biology and Chemistry*, **29**, 294-302.

Jones, J.J.*, et al.* (2006) Characterizing the Phospholipid Profiles in Mammalian Tissues by MALDI FTMS, *Analytical Chemistry*, **78**, 3062-3071.

Karas, M., Bachmann, D. and Hillenkamp, F. (1985) Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules, *Analytical Chemistry*, **57**, 2935-2939.

Karas, M. and Hillenkamp, F. (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons, *Analytical Chemistry*, **60**, 2299-2301.

Kazmi, S.*, et al.* (2006) Alignment of high resolution mass spectra: development of a heuristic approach for metabolomics, *Metabolomics*, **2**, 75-83.

Keller, A.*, et al.* (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats, *Mol Syst Biol*, **1**.

Kikuchi, S., Shibuya, I. and Matsumoto, K. (2000) Viability of an Escherichia coli pgsANull Mutant Lacking Detectable Phosphatidylglycerol and Cardiolipin, *Journal of Bacteriology*, **182**, 371-376.

Kingsley, P.J. and Marnett, L.J. (2009) Analysis of endocannabinoids, their congeners and COX-2 metabolites, *Journal of Chromatography B*, **877**, 2746-2754.

Klose, C.*, et al.* (2010) Yeast Lipids Can Phase-separate into Micrometer-scale Membrane Domains, *Journal of Biological Chemistry*, **285**, 30224-30232.

Kohlbacher, O.*, et al.* (2007) TOPP—the OpenMS proteomics pipeline, *Bioinformatics*, **23**, e191-e197.

Krznaric, D. and Levcopoulos, C. (2002) Optimal algorithms for complete linkage clustering in d dimensions, *Theoretical Computer Science*, **286**, 139-149.

Leavell, M. and Leary, J. (2006) Fatty Acid Analysis Tool (FAAT): An FT-ICR MS Lipid Analysis Algorithm, *Analytical Chemistry*, **78**, 5497-5503.

Liebisch, G.*, et al.* (1999) Quantitative measurement of different ceramide species from crude cellular extracts by electrospray ionization tandem mass spectrometry (ESI-MS/MS), *Journal of Lipid Research*, **40**, 1539-1546.

Lim, H.-K.*, et al.* (2007) Metabolite identification by data-dependent accurate mass spectrometric analysis at resolving power of 60 000 in external calibration mode using an LTQ/Orbitrap, *Rapid Communications in Mass Spectrometry*, **21**, 1821-1832.

Lin, S.M.*, et al.* (2005) What is mzXML good for?, *Expert Review of Proteomics*, **2**, 839-845.

Lipman, D.J., Altschul, S.F. and Kececioglu, J.D. (1989) A tool for multiple sequence alignment, *Proceedings of the National Academy of Sciences of the United States of America*, **86**, 4412-4415.

Liu, J.*, et al.* (2007) Methods for peptide identification by spectral comparison, *Proteome Sci*, **5**, 3.

Lodish, H.*, et al.* (2005) *Molecular Cell Biology*. WH Freeman and Company, New York.

Makarov, A.*, et al.* (2006) Performance Evaluation of a Hybrid Linear Ion Trap/Orbitrap Mass Spectrometer, *Analytical Chemistry*, **78**, 2113-2120.

Makarov, A., Denisov, E. and Lange, O. (2009) Performance evaluation of a high-field orbitrap mass analyzer, *Journal of The American Society for Mass Spectrometry*, **20**, 1391-1396.

Makarov, A.*, et al.* (2006) Dynamic range of mass accuracy in LTQ Orbitrap hybrid mass spectrometer, *J Am Soc Mass Spectrom*, **17**, 977-982.

Makarov, A.*, et al.* (2006) Dynamic range of mass accuracy in LTQ orbitrap hybrid mass spectrometer, *Journal of The American Society for Mass Spectrometry*, **17**, 977-982.

Marshall, A.G., Hendrickson, C.L. and Jackson, G.S. (1998) Fourier transform ion cyclotron resonance mass spectrometry: A primer, *Mass Spectrometry Reviews*, **17**, 1-35.

Martens, L.*, et al.* (2011) mzML—a Community Standard for Mass Spectrometry Data, *Molecular & Cellular Proteomics*, **10**.

Matthiesen, R.*, et al.* (2010) Discussion on common data analysis strategies used in MS-based proteomics, *Proteomics*, **11**, 604-619.

McLafferty, F.W. and Turecek, F. (1993) *Interpretation of mass spectra*. University Science Books, Sausalito.

Minkler, P.E. and Hoppel, C.L. (2010) Separation and characterization of cardiolipin molecular species by reverse-phase ion pair high-performance liquid chromatography-mass spectrometry, *Journal of Lipid Research*, **51**, 856-865.

Mohamed, R.*, et al.* (2009) Comprehensive analytical strategy for biomarker identification based on liquid chromatography coupled to mass spectrometry and new candidate confirmation tools, *Anal Chem*, **81**, 7677-7694.

Morris, J.*, et al.* (2005) Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum, *Bioinformatics*, **21**, 1764-1775.

Mougous, J.D.*, et al.* (2004) Identification, function and structure of the mycobacterial sulfotransferase that initiates sulfolipid-1 biosynthesis, *Nat Struct Mol Biol*, **11**, 721-729.

Mougous, J.D.*, et al.* (2006) A sulfated metabolite produced by stf3 negatively regulates the virulence of Mycobacteriumtuberculosis, *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 4258-4263.

Mueller, L.N.*, et al.* (2008) An Assessment of Software Solutions for the Analysis of Mass Spectrometry Based Quantitative Proteomics Data, *Journal of Proteome Research*, **7**, 51-61.

Murphy, R., Fiedler, J. and Hevko, J. (2001) Analysis of Nonvolatile Lipids by Mass Spectrometry, *Chemical Reviews*, **101**, 479-526.

Murphy, R.C. and Axelsen, P.H. (2011) Mass spectrometric analysis of long-chain lipids, *Mass Spectrometry Reviews*, **30**, 579-599.

Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins, *Journal of Molecular Biology*, **48**, 443-453.

Nesvizhskii, A.I., Vitek, O. and Aebersold, R. (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry, *Nat Meth*, **4**, 787-797.

Neumann, S. and Böcker, S. (2010) Computational mass spectrometry for metabolomics: Identification of metabolites and small molecules, *Anal Bioanal Chem*, 2779-2788.

O'Connor, P., Mirgorodskaya, E. and Costello, C. (2002) High pressure matrix-assisted laser desorption/ionization Fourier transform mass spectrometry for minimization of ganglioside fragmentation, *Journal of The American Society for Mass Spectrometry*, **13**, 402-407.

Orchard, S.*, et al.* (2004) Advances in the development of common interchange standards for proteomic data, *PROTEOMICS*, **4**, 2363-2365.

Oresic, M., Hanninen, V.A. and Vidal-Puig, A. (2008) Lipidomics: a new window to biomedical frontiers, *Trends Biotechnol*, **26**, 647-652.

Oursel, D.*, et al.* (2007) Lipid composition of membranes of Escherichia coli by liquid chromatography/tandem mass spectrometry using negative electrospray ionization, *Rapid Communications in Mass Spectrometry*, **21**, 1721-1728.

Ousterhout, J.K. (1998) Scripting: higher level programming for the 21st Century, *IEEE Xplore*, **31**, 23-30.

Palmblad, M., Buijs, J. and Håkansson, P. (2001) Automatic analysis of hydrogen/deuterium exchange mass spectra of peptides and proteins using calculations of isotopic distributions, *Journal of The American Society for Mass Spectrometry*, **12**, 1153-1162.

Paul, W. and Steinwedel, H. (1953) Ein neues Massenspektrometer ohne Magnetfeld, *Zeitschrift Naturforschung Teil A*, **8**, 448.

Pedrioli, P.*, et al.* (2004) A common open representation of mass spectrometry data and its application to proteomics research, *Nature Biotechnology*, **22**, 1459-1466.

Penkov, S.*, et al.* (2010) Maradolipids: Diacyltrehalose Glycolipids Specific to Dauer Larva in Caenorhabditis elegans, *Angewandte Chemie International Edition*, **49**, 9430-9435.

Perkins, D.N.*, et al.* (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data, *Electrophoresis*, **20**, 3551-3567.

Pulfer, M. and Murphy, R. (2003) Electrospray mass spectrometry of phospholipids, *Mass Spectrom Rev*, **22**, 332-364.

Purcell, E.M., Torrey, H.C. and Pound, R.V. (1946) Resonance absorption by nuclear magnetic moments in a solid, *Physical Review*, **69**, 37.

Quehenberger, O.*, et al.* (2010) Lipidomics reveals a remarkable diversity of lipids in human plasma, *Journal of Lipid Research*, **51**, 3299-3305.

Raa, H.*, et al.* (2009) Glycosphingolipid Requirements for Endosome-to-Golgi Transport of Shiga Toxin, *Traffic*, **10**, 868-882.

Rasche, F.*, et al.* (2010) Computing Fragmentation Trees from Tandem Mass Spectrometry Data, *Analytical Chemistry*, **83**, 1243-1251.

Reich, A.*, et al.* (2010) Lipidome of narrow-band ultraviolet B irradiated keratinocytes shows apoptotic hallmarks, *Experimental Dermatology*, **19**, e103-e110.

Reinert, K. and Kohlbacher, O. (2009) OpenMS and TOPP: Open Source Software for LC-MS Data Analysis. In., pp. 201-211.

Robinson, M.*, et al.* (2007) A dynamic programming approach for the alignment of signal peaks in multiple gas chromatography-mass spectrometry experiments, *BMC Bioinformatics*, **8**, 419.

Rockwood, A. and Haimi, P. (2006) Efficient calculation of accurate masses of isotopic peaks, *Journal of The American Society for Mass Spectrometry*, **17**, 415-419.

Roudebush, W.E.*, et al.* (2005) The significance of platelet-activating factor and fertility in the male primate: a review, *Journal of Medical Primatology*, **34**, 20-24.

Sadygov, R.G. and Yates, J.R. (2003) A Hypergeometric Probability Model for Protein Identification and Validation Using Tandem Mass Spectral Data and Protein Sequence Databases, *Analytical Chemistry*, **75**, 3792-3798.

Sagin, F.G. and Sozmen, E.Y. (2008) Lipids as key players in Alzheimer disease - Alterations in metabolism and genetics, *Curr. Alzheimer Res.*, **5**, 4-14.

Saito, K.*, et al.* (2009) Ablation of cholesterol biosynthesis in neural stem cells increases their VEGF expression and angiogenesis but causes neuron apoptosis, *Proceedings of the National Academy of Sciences*, **106**, 8350-8355.

Sauve, A.C. and Speed, T.P. (2004) Normalization, baseline correction and alignment of high-throughput mass spectrometry data. , *Proceedings of the Genomic Signal Processing and Statistics*, **2004**.

Schmelzer, K.*, et al.* (2007) The Lipid Maps Initiative in Lipidomics. In Brown, H.A. (ed), *Methods in Enzymology*. Academic Press, pp. 171-183.

Schuhmann, K.*, et al.* (2011) Bottom-Up Shotgun Lipidomics by Higher Energy Collisional Dissociation on LTQ Orbitrap Mass Spectrometers, *Analytical Chemistry*, **83**, 5480-5487.

Schultz, H.L. (1946) System for High Speed Counting of Nuclear Particles, *Proceedings of the American Physical Society*, **69**, 674-674.

Schwudke, D.*, et al.* (2007) Top-Down Lipidomic Screens by Multivariate Analysis of High-Resolution Survey Mass Spectra, *Anal Chem*.

Schwudke, D.*, et al.* (2007) Shotgun lipidomics by tandem mass spectrometry under data-dependent acquisition control, *Methods Enzymol*, **433**, 175-191.

Schwudke, D.*, et al.* (2006) Lipid Profiling by Multiple Precursor and Neutral Loss Scanning Driven by the Data-Dependent Acquisition, *Analytical Chemistry*, **78**, 585-595.

Scigelova, M. and Makarov, A. (2006) Orbitrap Mass Analyzer – Overview and Applications in Proteomics, *PROTEOMICS*, **6**, 16-21.

Shevchenko, A. and Simons, K. (2010) Lipidomics: coming to grips with lipid diversity, *Nature Reviews Molecular Cell Biology*, **11**, 593-598.

Smith, A. (2000) *Oxford Dictionary of Biochemistry and Molecular Biology.* . Oxford University Press, Oxford, UK.

Song, H.*, et al.* (2007) Algorithm for processing raw mass spectrometric data to identify and quantitate complex lipid molecular species in mixtures by data-dependent scanning and fragment ion database searching, *J Am Soc Mass Spectrom*, **18**, 1848-1858.

Ståhlman, M.*, et al.* (2009) High-throughput shotgun lipidomics by quadrupole time-of-flight mass spectrometry, *Journal of Chromatography B*, **877**, 2664-2672.

Taguchi, R., Nishijima, M. and Shimizu, T. (2007) Basic Analytical Systems for Lipidomics by Mass Spectrometry in Japan. In, *Lipidomics and Bioactive Lipids: Mass‐Spectrometry‐Based Lipid Analysis*. pp. 185-211.

Tan, B.*, et al.* (2006) Targeted lipidomics: Discovery of new fatty acyl amides, *The AAPS Journal*, **8**, E461-E465.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic acids research*, **22**, 4673-4680.

Tibshirani, R.*, et al.* (2004) Sample classification from protein mass spectrometry, by 'peak probability contrasts', *Bioinformatics*, **20**, 3034-3044.

Tselepis, A.D. and John Chapman, M. (2002) Inflammation, bioactive lipids and atherosclerosis: potential roles of a lipoprotein-associated phospholipase A2, platelet activating factor-acetylhydrolase, *Atherosclerosis Supplements*, **3**, 57-68.

van Meer, G. (2005) Cellular lipidomics, *Embo J*, **24**, 3159-3165.

Watson, A.D. (2006) Thematic review series: Systems Biology Approaches to Metabolic and Cardiovascular Disorders. Lipidomics: a global approach to lipid analysis in biological systems, *Journal of Lipid Research*, **47**, 2101-2111.

Weichuan, Y., Xiaoye, L. and Hongyu, Z. (2005) Aligning peaks across multiple mass spectrometry data sets using a scale-space based approach. *Computational Systems Bioinformatics Conference, 2005. Workshops and Poster Abstracts. IEEE*. pp. 126-127.

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

Whitehouse, C.M.*, et al.* (1985) Electrospray interface for liquid chromatographs and mass spectrometers, *Anal Chem*, **57**, 675-679.

Wiki LipidXplorer Wiki.

Wolf, S.*, et al.* (2010) In silico fragmentation for computer assisted identification of metabolite mass spectra, *BMC Bioinformatics*, **11**, 148.

Wong, J., Cagney, G. and Cartwright, H. (2005) SpecAlign--processing and alignment of mass spectra datasets, *Bioinformatics*, **21**, 2088-2090.

Wong, J.W.H., Cagney, G. and Cartwright, H.M. SpecAlign—processing and alignment of mass spectra datasets, *Bioinformatics*, **21**, 2088-2090.

Yang, K.*, et al.* (2009) Automated Lipid Identification and Quantification by Multidimensional Mass Spectrometry-Based Shotgun Lipidomics, *Analytical Chemistry*, **81**, 4356-4368.

Yasui, Y.*, et al.* (2003) An Automated Peak Identification/Calibration Procedure for High-Dimensional Protein Measures From Mass Spectrometers, *J Biomed Biotechnol*, **2003**, 242-248.

Yetukuri, L.*, et al.* (2008) Informatics and computational strategies for the study of lipids, *Mol Biosyst*, **4**, 121-127.

Yost, R.A. and Enke, C.G. (1978) Selected ion fragmentation with a tandem quadrupole mass spectrometer, *Journal of the American Chemical Society*, **100**, 2274-2275.

Yu, W.*, et al.* (2006) Detecting and aligning peaks in mass spectrometry data with applications to MALDI, *Computational Biology and Chemistry*, **30**, 27-38.

Ziqiang, G. (2009) Discovering novel brain lipids by liquid chromatography/tandem mass spectrometry, *Journal of Chromatography B*, **877**, 2814-2821.

Zubarev, R. and Mann, M. (2007) On the Proper Use of Mass Accuracy in Proteomics, *Molecular & Cellular Proteomics*, **6**, 377-381.

Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts habe ich Unterstützungsleistungen von folgenden Personen erhalten: Prof. Dr. Michael Schröder, Dr. Andrej Shevchenko und Dr. Dominik Schwudke.

Weitere Personen waren an der geistigen Herstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich nicht die Hilfe eines kommerziellen Promotionsberaters in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt und ist auch noch nicht veröffentlicht worden.

Ich bestätige, dass ich die geltende Promotionsordnung der Fakultät Informatik der Technischen Universität Dresden anerkenne.

Dresden, 02.03.2012 Ronny Herzog