Greazy: Open-Source Software for Automated Phospholipid MS/MS Identification

By

Michael Kochen

Thesis
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE
in

Biomedical Informatics
August, 2015
Nashville, Tennessee

Approved:

David L. Tabb

Bing Zhang

John A. McLean

TABLE OF CONTENTS

Page

LIST OF TABLES iii

LIST OF FIGURES iv

Chapter

1. Introduction 1

Lipids in Cellular Function and Disease 1
Current Lipidomics Tools 3
Challenges in Lipidomics 5
Research Goals 6

2. Greazy: Open-Source Software for Automated Phospholipid MS/MS Identification 7

Introduction 7
Methods 8
Software 8
Data Sets 17
Results and Discussion 18
Characterization of the Search Space 19
Validation against the NIST 2014 Small Molecule MS/MS Library 22
Comparison of Greazy/LipidLama to LipidSearch 27
Analysis of Biological Replicates with Greazy and LipidLama 32
Conclusions and Future Work 35

Appendix

A. Fragmentation Models 36

REFERENCES 44

LIST OF TABLES

Table Page

1. Glycerophospholipids with Metal Adducts in Positive Ion Mode 10

2. Lipids Present in all Four Biological Replicates of SARS Infected Calu-3 2B4 Cells 34

A1. Glycerophospholipid Fragments with Nonmetal Adducts in Positive Ion Mode 36

A2. Glycerophospholipids in Negative Ion Mode 37

A3. Phosphatidylinositides in Negative Ion Mode 38

A4. Phosphosphingolipid Fragments with Nonmetal Adducts in Positive Ion Mode 39

A5. Phosphosphingolipid Fragments with Metal Adducts in Positive Ion Mode.....39

A6. Phosphosphingolipid Fragments in Negative Ion Mode.....40

A7. Cardiolipin Fragments in Positive Ion Mode 40

A8. Cardiolipin Fragments in Negative Ion Mode – Deprotonated.....41

A9. Cardiolipin Fragments in Negative Ion Mode – Doubly Deprotonated.....42

A10. Cardiolipin Fragments in Negative Ion Mode – Doubly Deprotonated plus Alkali Metal...43

LIST OF FIGURES

Figure Page

1. Schematic of the Greazy/LipidLama Workflow 7

2. Mixture Modeling Approach to FDR estimation 14

3. Graphical Comparison of Experimental and Theoretical Spectra 16

4. Schematic of Search Space Construction 19

5a. Effect of Fatty Acid Chain Length on Search Space Size 21

5b. Effect of Chain Length and Double Bond Content on Search Space Size 21

6. NIST Library Lipids Identified by Greazy 23

7a. Hypergeometric vs Combined Score: Negative Mode 24

7b. Hypergeometric vs Combined Score: Positive Mode 24

8. NIST Library Spectra Identified by Greazy 25

9. Score vs Collision Energy: Negative Mode 26

10. Score vs Collision Energy: Positive Mode 26

11. Score vs Spectrum Peak Count 27

12a. Greazy/LipidLama vs LipidSearch: Spectral Identifications 30

12b. Greazy/LipidLama vs LipidSearch: Lipid Identifications 30

13a. Greazy/LipidLama vs LipidSearch: Negative Mode Spectral Identifications 31

13b. Greazy/LipidLama vs LipidSearch: Positive Mode Spectral Identifications 31

14a. Venn Diagram of Four SARS infected Calu-3 2B4 Cell Replicates 33

14b. Lipid Class Count per Replicate 33

CHAPTER 1

Introduction

Lipidomics is the analysis of lipids from biological systems (cells, organelles, organs, etc.) at the broadest possible scale. This includes not only the identification and quantification of the full complement of lipids in the system but also the interactions of lipids with other cellular components, the pathways of lipid metabolism, and the many biological roles that lipids play in cellular function and disease[1]. In order to analyze the entirety of a lipidome, high-throughput technologies are employed which produce overwhelming amounts of data requiring sophisticated software for interpretation. Such tools have been used to great effect in genomics and proteomics. Unfortunately the software tools available for the analysis of lipidomic data sets lag those in other -omics fields. Much of this difficulty stems from the structural complexity and variability of lipids, as well as a relative lack of available datasets for comparative analysis. These and other problems must be addressed for any high-throughput data analysis tools to be viable for large-scale lipid analysis.

Lipids in Cellular Function and Disease

There are three major biological roles played by lipids: energy storage, signaling, and as the principle structural element of cellular membranes[1]. In energy storage lipids make up the long term energy reserves for all animal species. Triacylglycerols and sterol esters, because of their greatly reduced fatty acid components, are the main lipids found here[2]. In cellular signaling lipids often act as secondary messengers which relay signals received at the cell surface to

targets within the cell. Many lipids do this by acting as indirect messengers in which a portion of the lipid, a head group for example, is cleaved off to carry the signal to the target molecule[3]. Glycerophospholipids and sphingolipids embedded in the cell membrane are often used in this manner. Some lipids serve as direct messengers. Phosphatidylinositols, which have a head group containing multiple phosphorylation sites and thus multiple configurations (phosphatidylinositides), comprise one such example[3]. Lipids also serve as the primary structural members of cellular membranes[2]. This includes not only the plasma membrane but all the internal membranes surrounding the organelles and their derivatives, such as transport vesicles. Membranes allow for the compartmentalization of both molecules and reactions within a cell and are, in many cases, an active participant in the functioning of the cell. The lipid profile of a given membrane will depend on the role that particular compartment plays but will generally be comprised of glycerophospholipids, sphingolipids, and sterols. Glycerophospholipids and sphingolipids are amphipathic meaning they have both hydrophobic and hydrophilic moieties, an essential property for the formation of a lipid bilayer.

Given the diversity of roles played by lipids it is not surprising that they are also associated with a number of diseases. Patients with type 1 diabetes, for example, have been shown to have altered lipid profiles. Lipids containing a phosphatidylcholine head-group, including glycerophosphocholine, lysoglycerophosphocholine, and sphingomyelin were significantly decreased in this group of patients[4]. Patients with early signs of Alzheimer's disease were shown to have altered levels of lipids from the sphingolipid class[5]. In particular, the levels of sphingomyelin had decreased significantly while levels of ceramide had risen. Sphingomyelin is metabolized into ceramide through cleavage of the phosphocholine head-group, possibly indicating a connection between this mechanism and the disease. Cardiolipins are found in the

inner membranes of mitochondria and play important roles in maintaining the membrane potential and stabilizing the quaternary structure of a number of enzymes and proteins[6]. Alterations in the cardiolipin content of mitochondrial membranes are associated with a number of conditions including heart failure, Barth syndrome, and Parkinson disease[6]. Tumor cells exhibit changes in cell signaling, membrane composition, and metabolic activity. Lipids play a central role in all of these functions making the lipidome an attractive place to look for cancer biomarkers[7-9].

Current Lipidomics Tools

The diversity of cellular functions and diseases in which lipids play a role substantiates their importance and the necessity of their study. There is a need for better bioinformatics software to do these analyses as the currently available tools often cater to specific applications or come with costly licensing fees. There are two basic approaches currently employed to run mass spectrometry based lipidomics experiments: direct infusion or "shotgun" lipidomics and chromatographic separation.

Direct infusion experiments, in which the sample is injected directly into the mass spectrometer, are often used along with precursor ion and neutral loss scanning[10]. Here, the precursor ions are scanned over a mass range at the MS1 stage and then fragmented. A diagnostic ion is then scanned for at the MS2 stage. In the case of a neutral loss scan the m/z value for the diagnostic ion is an offset of the precursor m/z and represents the neutral loss of some component of the lipid. For a precursor ion scan the diagnostic ion is a specific m/z that represents the mass of a known fragment of the lipid, such as a head group or fatty acid component. Software tools have been developed for such an analysis include AMDMS-SL[11] and

LipidView[12]. Direct infusion lipidomics can also be used with data-dependent acquisition scanning. Tools such as LipidQA[13], LipidExplorer[14], and LipidInspector[15] are designed for this.

Liquid chromatography is often used along with a soft-ionization technique such as electrospray ionization (ESI) or matrix-assisted laser desorption (MALDI) which allow for the ionization of intact lipids. Using such a setup in an LC-MS experiment an investigator can use the chromatographic retention time and detected m/z value to produce a list of putative identifications for a mass spectrum. A number of software tools are available for such analyses including lipid specific software such as Lipid Data Analyzer and general metabolomics software such as MZmine2[16, 17].

Unfortunately many lipid classes have a chemical structure that permits a multitude of isobaric species. This is particularly true for lipid categories that contain fatty acid constituents like those targeted by the software outlined in this proposal. Take, for example, the phosphatidylcholines PC(14:0/18:0) and PC(16:0/16:0). Both of these have a total of 32 carbons in their two fatty acid substituents, but these carbons are differentially distributed. The solution to this problem is to use tandem mass spectrometry. In an MS/MS experiment the m/z value for the precursor ion, the intact ionized lipid, is detected at the MS1 stage and then fragmented through collision induced dissociation (CID). The fragment m/z values are then detected at the MS2 stage with a second mass analyzer. There are several types of experiments that can be run with this setup including neutral loss, precursor ion and data-dependent acquisition (DDA) MS/MS experiments.

In a DDA experiment a tandem mass spectrum for every observed precursor is acquired. The detection of diagnostic product or neutral loss ions and the subsequent identification of lipids is left to the user. In recent years several tools have been developed to make this process more

manageable. Some, like LipidView (AB Sciex) and LipidSearch (Thermo Scientific) are powerful but costly commercial products[18, 19]. Other software tools lack the flexibility, depth of fragmentation models, robust scoring algorithms and false discovery detection that our software will provide.

Challenges in Lipidomics

Although lipidomics software has advanced in recent years, several challenges bar the path to routine interpretation of these data sets. In proteomics, peptide sequences are routinely identified from MS/MS spectra through database search algorithms that employ simple rules for predicting peptide fragmentation patterns. Lipids exhibit greater structural complexity and diversity than peptides which makes their computational analysis more difficult. The LIPID MAPS database categorizes more than 37000 lipids into eight lipid categories that are further divided into dozens of subclasses, each of which ionizing and fragmenting in a distinct manner[20]. As a result, predicting the appearance of a MS/MS spectrum for a given lipid requires greater attention than in proteomics, and the number of fragments seen for a given compound is generally smaller. MS/MS spectral libraries for lipids are also far less populated and diverse than those in proteomics, although there has been some recent work in this area with the development of LipidBlast[21], a collection of predicted MS/MS spectra for many lipid classes. The literature detailing the MS/MS mechanisms for lipids is considerably less fully developed than it is for proteomics. In proteomics, extensive spectral libraries and sequence databases have yielded powerful statistical scoring algorithms and methods for estimating false discovery rates. Leveraging methods developed for proteomics to lipidomics may begin to address the shortage of tools in this space.

Research Goals

This research project provides a bioinformatics tool set for the identification of phospholipids from MS/MS experiments. The software allows for the automated identification of lipids from several important classes in the glycerophospholipid and sphingolipid categories by matching the tandem mass spectra to thousands of theoretical spectra from a user defined phospholipid search space. The software incorporates methods currently utilized in the more mature field of proteomics for scoring these matches and estimating the false discovery rate. The aims included:

1. Construction of the lipid search space and associated theoretical spectra
2. Design and implementation of a set of scoring algorithms
3. Design and implementation of a method for the estimation of the false discovery rate
4. Validation of the software against a number of data sets and a commercial lipid identification product

Aim 1 had two parts. First, software was created to build a search space of lipids based on user defined parameters. An extensive search of the literature was then made to find suitable fragmentation models to represent theoretical spectra for the lipids in that search space. For aim 2 we implemented both peak counting and intensity scoring algorithms that are combined to produce a final score. The false discovery rate estimation of aim 3 utilizes mixture modeling and a density estimation algorithm. Finally the software was validated using a number of data sets including the NIST 2014 Small Molecule MS/MS library as well as two experimental data sets. The software is also compared to Thermo Scientifics LipidSearch software.

CHAPTER 2

Greazy: Open-Source Software for Automated Phospholipid MS/MS Identification

Introduction

We believe that the lessons learned and methods developed in proteomics can help to advance the field of lipidomics. To that end we have developed the software "Greazy" for the high-throughput identification of phospholipid lipid tandem mass spectrometry data through several techniques borrowed from proteomics (Figure 1). Greazy incorporates Proteowizard[22], enabling it to accept many vendor-specific and open file formats in use today. It accepts a set of parameters that define a phospholipid search space and then builds a comprehensive literature-derived fragmentation model for each lipid in that space. These predicted spectra are compared to experimental ones through two probability-based scores. Lastly, a mixture modeling approach using density estimation is used to estimate the false discovery rate and to generate a cutoff score. Successful matches are written to the easily reviewed mzTab file format[23].

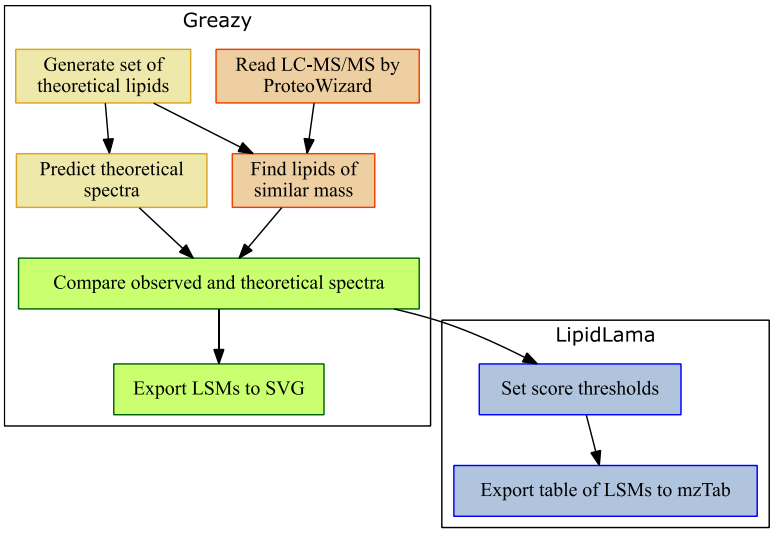


Figure 1. Schematic of the Greazy/LipidLama Workflow

Methods

Software

Greazy is software for the high-throughput identification of phospholipids from MS/MS spectra. The software must be supplied with experimental MS/MS data files and a user defined list of parameters that define a phospholipid search space and experimental variables. Integrated into the software is Proteowizard[22] enabling Greazy to handle a number of open and vendor specific file formats. The search space parameters are supplied through a graphical user interface and are used to build a list of phospholipid precursor ions that are then fragmented in silico for comparison to the experimental spectra.

Search Space and Experimental Parameters. The software generates fragmentation models for phospholipids from the glycerophospholipid and sphingolipid categories. Cardiolipins, more complex variants of the glycerophospholipids, require additional attention and are considered a separate category here. The categories are broken down into classes, each of which has its own structure and requires a distinct model when building the search space. Each lipid is constructed by piecing it together from various backbone, head group, and fatty acid substituents as well as the linkage types for fatty acid attachment that are included in the supplied parameters.

Glycerophospholipids consist of a glycerol backbone, a phosphate-containing head group esterified at the sn-3 position of the backbone, and two fatty acid substituents linked at the sn-1 and sn-2 positions. Several classes of lipid, defined by their head groups, can be specified for inclusion in the search space including glycerophosphocholines (PC), glycerophosphoethanolamines (PE), glycerophosphoinositols (PI), glycerophosphoglycerols (PG), glycerophosphoserines (PS), glycerophosphates (PA) and glycerophosphoinositol mono-, bis-, and trisphosphates (PIP). Parameters required to specify the fatty acid substituents include

ranges for the carbon chain lengths and number of double bonds for each fatty acid. It can also be specified that only fatty acids with an even number of carbon atoms are to be included. Users can specify whether fatty acids may be linked by acyl or ether bonds for each position. The software does not distinguish between ether and vinyl ether linkages, and so the latter are excluded. Lysoglycerophospholipids can also be selected for inclusion in the search space.

Phosphosphingolipids consist of a sphingoid backbone, a phosphate-containing head group, and a single amide linked fatty acid. Four sphingoid backbones, sphingosine, sphinganine, phytosphingosine, and sphingadienine can be selected for inclusion in the search space. The user must provide a range of lengths for the included backbones. Three classes of phosphosphingolipids can be specified for inclusion based on the type of head group. These include ceramide phosphocholines (sphingomyelins), ceramide phosphoethanolamines, and ceramide phosphoinositols. As in the case of glycerophospholipids, the user must specify a range of carbon lengths and saturation levels for the fatty acid.

Cardiolipins consist of a glycerophosphoglycerophosphoglycerol backbone and four acyl-linked fatty acids. The user must provide a range of carbon chain lengths and saturation levels for each of the fatty acids. As in the glycerophospholipid case, the user may specify that only lipids with an even number of carbons in its fatty acids be included in the search. Lysocardiolipins can also be selected for inclusion.

Following the construction of the lipid search space, precursor ions are generated for each included lipid using a number of experimental parameters provided by the user. The software must know the ion mode, positive or negative, employed in the mass spectrometry experiment, and the adducts expected to be present in solution. The adducts available in the program include sodium, potassium, lithium, and ammonium for positive ion mode and chlorine, acetate, and

formate for negative ion mode. Alkali metals are also used to generate negative ion cardiolipin precursors. Protonated and deprotonated precursors are assumed to be present in every positive and negative ion mode search, respectively. Lastly, the mass accuracy tolerances for the precursor and fragment ions must be supplied and can be specified in either ppm or Daltons.

Precursors and Fragmentation patterns. Fragmentation patterns for each of the lipid classes above were gleaned from the literature, or extrapolated based on it[24-35]. Experimental conditions inform the software of the expected precursors for each class.

Table 1: Glycerophospholipids with Metal Adducts in Positive Ion Mode

	PC	PE	PS	PG	PA	PI	PIP
HG and adduct		X	X	X		X	X
loss of HG	X	X	X	X	X	X	X
loss of HG and addition of water		X	X	X	X	X	X
loss of HG and adduct	X	X	X	X	X	X	X
loss of FA1 as ketene	X	X	X	X	X	X	X
loss of FA1 as carboxylic acid	X	X	X	X	X	X	X
loss of HG, adduct and FA1 as ketene		X	X	X	X	X	X
loss of HG, adduct and FA1 as carboxylic acid		X	X	X	X	X	X
R1CO	X	X	X	X	X	X	X
loss of FA2 as ketene	X	X	X	X	X	X	X
loss of FA2 as carboxylic acid	X	X	X	X	X	X	X
loss of HG adduct and FA2 as ketene	X	X	X	X	X	X	X
loss of HG, adduct and FA2 as carboxylic acid		X	X	X	X	X	X
R2CO	X	X	X	X	X	X	X
loss of head		X	X	X		X	X
loss of head and FA1 as carboxylic acid	X	X	X	X		X	X
loss of head and FA1 as ketene		X	X	X		X	X
loss of head and FA2 as carboxylic acid		X	X	X		X	X
loss of head and FA2 as ketene		X	X	X		X	X
loss of phosphoric acid			X				
Choline	X						
loss of NC3H9	X						
PO4C2H5 & adduct	X						
loss of FA1 as carboxylic acid, loss of adduct	X						
loss of FA2 as carboxylic acid, loss of adduct	X						
loss of choline & adduct, loss of FA1 as acid	X						
loss of NC3H9, loss of FA1 as acid	X						
loss of NC3H9, loss of FA2 as acid	X						
loss of HG and FA2 as acid with added D. bond	X						

Fragmentation patterns for the glycerophospholipids in positive ion mode are divided into two groups: those with alkali metal adducts and those with nonmetal adducts. These are then subdivided into classes based on the head group substituent. The alkali metal group includes the precursor ions $[M + Na]^+$, $[M + K]^+$, and $[M + Li]^+$ (Table 1) while the nonmetal group includes $[M + H]^+$ and $[M + NH_4]^+$ ions (Table A1). Fragmentation is expected to be more extensive with the use of alkali metal adducts. The possible precursors in negative ion mode include $[M - H]^-$, $[M + Cl]^-$, $[M + CHOO]^-$, and $[M + CH_3COO]^-$ (Tables A2 and A3). In negative ion mode PCs are only expected to ionize in the presence of an adduct, and when they do, two precursors are generated: $[M + adduct]^-$, and $[M – CH_3]^-$.

Fragmentation patterns for the phosphosphingolipids in positive ion mode are likewise divided into metal (Table A4) and nonmetal (Table A5) precursor adducts and then broken down further based on the head groups. The same precursors as above are found here, and the metal adducts are again expected to fragment more extensively. In negative ion mode the fragmentation models for these lipids are sparse, with fragments pertaining only to the head group (Table A6). Like PCs, sphingomyelins are only expected to ionize in the presence of an adduct, and when they do, two precursors are generated: $[M + adduct]^-$, and $[M - CH3]^-$.

Cardiolipin precursors generated in positive mode include $[M + H]^+$, $[M + Alk]^+$, $[M - H + 2Alk]^+$, and $[M - 2H + 3Alk]^+$ where Alk is Na, K, or Li. The $[M + H]^+$, $[M + Alk]^+$, and $[M - H + 2Alk]^+$ precursors are assumed to have similar fragmentation patterns, with the exception of the loss of alkali metal adducts along with glycerol-1,3-diphosphate (Table A7). The expected number of fragments for the $[M - 2H + 3Alk]^+$ ion is comparatively high (Table A7). Cardiolipin precursors generated in negative mode include $[M - H]^-$, and the doubly charged ions $[M - 2H]^{2-}$, and $[M - 2H + Alk]^{2-}$ (Tables A8-A10).

Preprocessing and Scoring Algorithms. Preprocessing of an MS/MS spectrum begins with removal of peaks within the predefined tolerance of the precursor m/z. The total ion current is then normalized and the peak intensities are accordingly adjusted. In matching the experimental MS/MS spectra to fragmentation models in the search space, the first step is to match the observed and expected precursor m/z values. When the respective m/z values are within the user provided tolerance, the associated MS/MS spectrum is scored against the corresponding theoretical fragmentation pattern for the search space precursor using the scoring algorithm below. The scoring algorithm will consider separately the number of matching peaks and the proportion of total ion intensity found in those peaks. These scores are then combined to determine a final score.

Peak scoring makes use of the hypergeometric distribution (HGD) to calculate the probability of randomly matching a given number of peaks. The HGD is used when one is sampling without replacement from a discrete population for which members are one of two types. Consider for example a jar containing N marbles M of which are green and the remainder are white. We would like to know the probability of finding x green marbles when selecting K marbles from the jar without replacement. This can be calculated using the HGD which takes the form

$$P(x|N,M,K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}.$$

In order to use the HGD for peak scoring, the continuous m/z range is divided into a discrete number of bins based on the scan range of the experiment and the mass tolerance provided by the user. We can then use the HGD to compute the probability of finding x matching peaks given N bins, K peaks in the spectrum, and M peaks in the fragmentation model. To calculate the

probability of at least x+1 matches occurring by chance we sum from x+1 to M and obtain the peaks score

$$s_1 = \sum_{i=x+1}^{M} P\,(i|N, M, K).$$

Now consider a spectrum with k number of matched peaks containing a proportion of the total ion intensity for that spectrum. The intensity score is the probability of finding more ion intensity than the matched peaks when randomly choosing k peaks from the spectrum. Let i be the summed intensity of the matched peaks and i_j be the summed intensity of the jth combination of k peaks. The intensity score is defined as the fraction of combinations that hold more ion intensity than the matched peaks. This given by the equation

$$s_2 = \frac{\sum_{j=1}^{\binom{n}{k}} p_j}{\binom{n}{k}}, \quad p_j = \begin{cases} 1 if\, i_j > i \\ 0 if\, i_j < i \end{cases}.$$

Combined Score: Fishers method is employed to combine the peak and intensity scores into a final score[36]. The result is a χ^2 statistic and is given by

$$\chi^2 = -2\sum_{i=1}^{2} ln(s_i).$$

Error Estimation. Greazy produces a list of putative matches to lipids in the search space by first matching the precursor m/z and then producing a score as detailed above. Many of these matches may have a limited number of matched fragments and are likely to be false discoveries. To eliminate such low quality matches LipidLama was developed. LipidLama is a software component within Greazy that employs a mixture modeling approach to the problem of false discoveries (Figure 2). It is assumed that the distribution of scores produced by Greazy is a combination of a distribution for true matches and one for false matches. LipidLama uses the kernel density estimation method to fit two such distributions to the data. In this method the

distribution of true matches is estimated using the distribution of high scores and a known distribution that is assumed to be representative of the false matches. LipidLama uses the distribution of second highest scores as the model for the distribution of false matches.

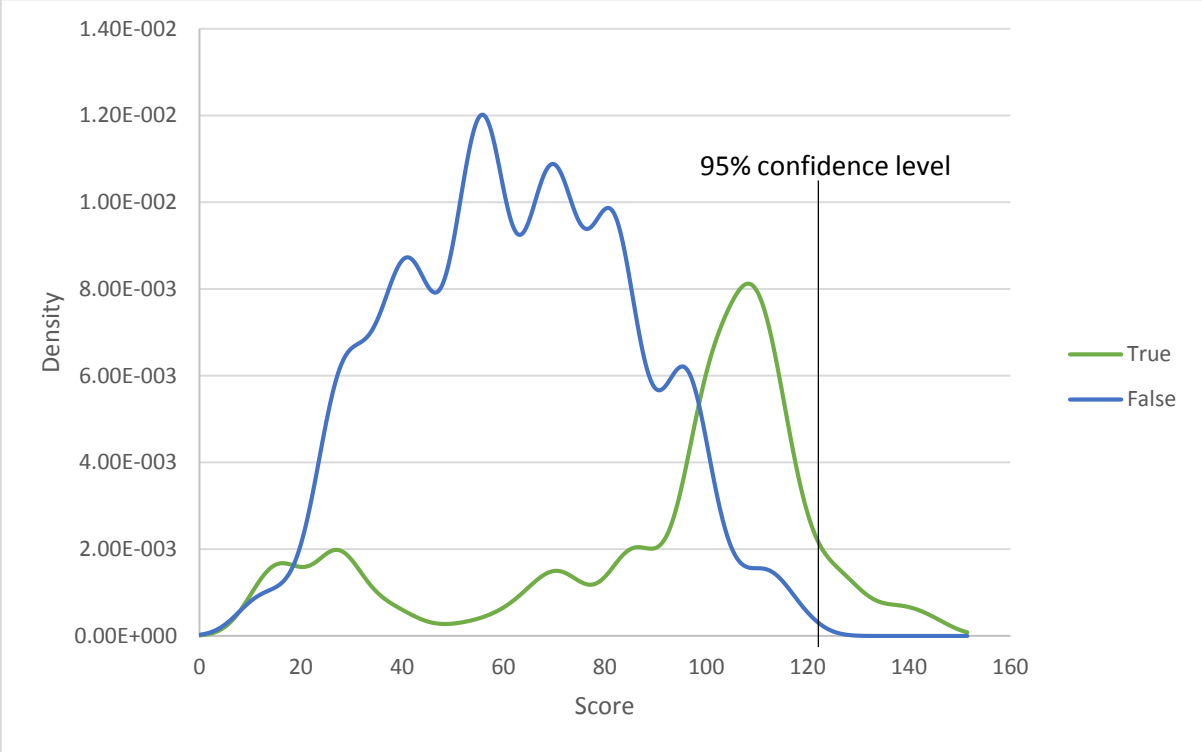


Figure 2. The kernel density estimation algorithm separates the score distribution from Greazy into components for the True (green) and False (blue) positives. A cutoff score for any given confidence level is then calculated.

The algorithm used by LipidLama largely follows Choi et al.,[37] with a few exceptions. Bandwidth selection, for example, is calculated using the method of Sheather and Jones[38]. Also, the proportion of the high score distribution that is accounted for by the false match distribution must be estimated empirically. Let i represent evenly spaced points along the Greazy score axis, f_1 and f_2 represent the estimated distributions for the highest and second highest scores

respectively, and w_i be the difference between these two distributions at the point i. Then the proportion p of the full distribution that is accounted for by the false hits is estimated to be

$$\min_{p} \sum_{\substack{i \\ f_{2,i} > f_{1,i}}} w_i (p f_{2,i} - f_{1,i})^2.$$

Once the distributions are constructed the tail areas of the true and combined distributions are used to determine a cutoff score for any desired confidence level. Figure 2, for example, shows the cutoff at an estimated 5% FDR.

The mixture modeling approach assumes that the score distribution is a composite of two distributions: one true and one false. In practice there could be many distinct true and false distributions that represent different lipid classes, or adduct additions. In LipidLama this is accounted for by applying the algorithm to each of these classes separately and computing a distinct cutoff for each one. If for example PCs are included in a multiple class search the top two PC scores for each spectra will be recorded and used by LipidLama to calculate a cutoff for that class. Lipids with multiple charge states are similarly dealt with. Glycerophosphoinositol mono- and bisphosphates for example can be found as a 1^- or 2^- precursor each of which will have its own fragmentation pattern and set of true and false distributions. LipidLama takes advantage of the fact that a second highest hit for a glycerophospholipid is usually an isomer of the highest ranked hit. Cardiolipins and sphingolipids do not fit this assumption and are not subject to the mixture modeling algorithm. Instead a percentage cut off is specified that is applied to the top ranking spectra for these two classes.

Output. Greazy produces an intermediary text format that is read by LipidLama. LipidLama presents the final identification results in the form of a summary mzTab file[23]. The mzTab file consists of a list of identified precursor ions (lipids and their modifications). For each precursor on this list the scan number, retention time, and score for each spectrum whose best match was to

that precursor are presented. The list is in descending order based on the score of the best matching spectrum for each lipid. The final set of identified precursors are those that were a best match for at least one spectrum in the data set. Additional information for each identified precursor includes the chemical formula, experimental and expected precursor m/z values, and the type of precursor that was analyzed.

Greazy also produces a series of spectral visualization files for further analysis (Figure 3). For each spectrum with at least one lipid match, the spectrum is reproduced along with the theoretical spectrum of the top scoring lipid. Peaks that match between the two spectra are colored green. Also found here are the fragment descriptions for the theoretical spectrum, along with their m/z values and the intensity of any fragments matched to the experimental spectrum. The precursor description and mass are also given here, as is the final score. These files are labeled by their spectrum number and are written as SVG embedded html files.

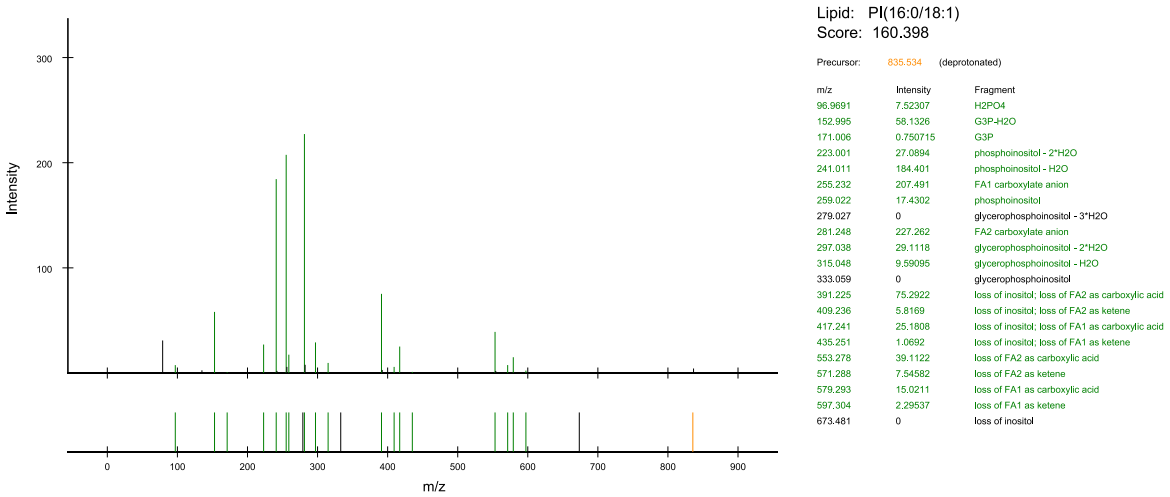


Figure 3. For each spectra with a match the spectrum is reproduced (top left) along with the theoretical spectrum of the lipid (bottom left). For each of the peaks in the theoretical lipid spectrum a description of associated fragment is provided as is the intensity of any of those peaks that match (green). The name of the lipid, the score, and the precursor m/z and description are also given.

Data Sets

NIST Tandem MS Library. The 2014 release of the NIST small molecule MS/MS library consists of 193,119 spectra for a number of different compound types including metabolites, drugs, phospholipids, and more. The spectra for these compounds were acquired under a broad range of experimental conditions. Ion trap and collision cell instrument types were used; data were acquired in both positive and negative ion mode. A number of precursor adducts were used, along with a wide range of collision energies, resulting in differing degrees of fragmentation. This data set was split into two files based on the ion mode of each entry. This resulted in a positive ion file containing 161,355 entries and a negative ion file containing 31,764. Next, glycerophospholipids, cardiolipins, and sphingolipids were extracted from these files using a string search. Those entries containing the strings "glycero-3-phos", "sphingosyl" or "sphingenyl" were kept and the rest discarded. This reduced the positive ion file to 3,107 entries and the negative ion file to 1,232. Represented in these entries were 141 distinct compounds in the positive ion file and 175 in the negative. A number of compounds were present in both files. The total number of distinct compounds was 214. The resulting files were reformatted as MGF files for use with Greazy.

Bone Marrow Stem Cells. In the Department of Biochemistry at the University of Texas Health Science Center at San Antonio lipids were extracted from cultured bone marrow stem cells collected from wild-type mice. The lipid extracts were subjected to reverse phase liquid chromatography with a Waters Atlantis dC18 column and a 55-min gradient. A PicoChip II nanospray ion source then led to a Thermo Fisher Q exactive tandem mass spectrometer. Analysis was done in both positive and negative ion modes resulting in 13362 and 12438 MS/MS spectra respectively. Lipid identifications were made with Thermo Scientific's

LipidSearch software version 4.0.14 and provided to us for comparison purposes. The raw files were converted to mzML files using Proteowizard's MSconvertGUI tool for analysis with Greazy and LipidLama.

Calu-3 2B4 Cell Line. At the Pacific Northwest National Laboratory six replicates each of icSARS, icSARS-DORF6, and mock infected Calu-3 2B4 cells were plated and samples were taken at 0, 12, 24, 36, 48, 60, and 72 hours. The series for four of the six replicates were subjected to capillary ultra-performance reverse phase liquid chromatography and introduced via a nano-electrospray ion source to a Thermo Scientific LTQ Orbitrap Velos. They were analyzed in both positive and negative ion mode for a total of 168 data sets. The raw files were converted to mzXML files via Proteowizard and uploaded to the MassIVE repository. We chose to work with the four icSARS 0 hour data sets in negative ion mode (MassIVE ID MSV000078781). These files contained 8748, 8826, 8831, and 8536 MS/MS spectra.

Results and Discussion

Greazy is designed to make high quality identifications of phospholipids from high throughput tandem mass spectrometry data sets. We used the NIST 2014 small molecule tandem MS library as a set of reference spectra with which to validate the fragmentation models and scoring algorithms within Greazy. We then compared the identifications made by Greazy to those of LipidSearch, a leading commercial lipid identification software platform, in an experimental data set derived from a bone marrow cell line. Lastly, we gauged the software's performance in the analysis of data from several biological replicates from a human bronchial epithelium cell line.

Characterization of the Search Space

The definition of the search space is the greatest difference between proteomic database search and a lipid search with Greazy. In Greazy the search space is defined by the parameters set by the user (Figure 4) and reflects the lipid composition expected for a given sample as well as the experiment that measured them. The size of the search space should also be considered, as it can grow exponentially if every possibility is allowed.

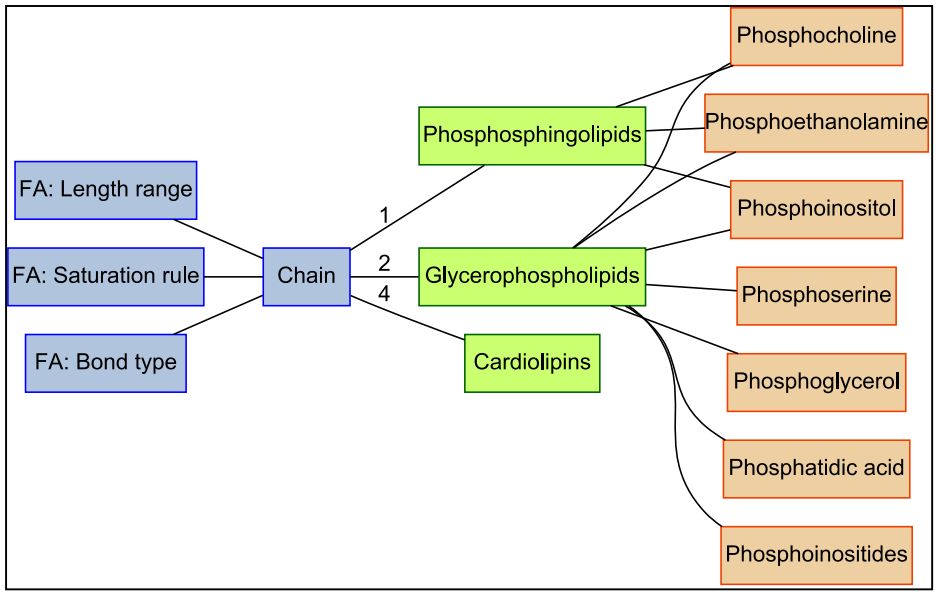


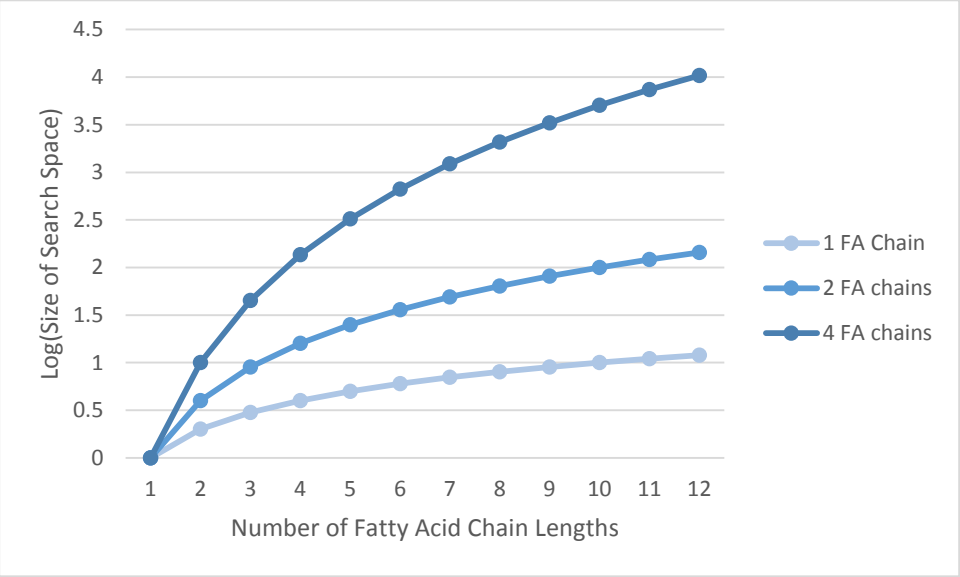
Figure 4. Schematic by which Greazy constructs potential lipids from components

Fatty acid configurations impact the search space size most significantly. Supporting Figure 5a illustrates the effect of increasing the number of possible chain lengths for one, two,

and four fatty acid chains while holding all other parameters constant. The search space for lipids with one fatty acid chain grows linearly, but the search space for lipids with two fatty acids grows with the square of possible chain lengths. The search space for cardiolipins with their four

fatty acid chains grows at the rate of ½(N^4 + N^2) where N is the number of possible fatty acid chain lengths. This reflects the elimination of identical species that arise due to the symmetry of this class of lipid. Allowing for unsaturations in the fatty acids has approximately half the impact of fatty acid chain length in search space scaling (Figure 5b). Typical configurations without CL generate search spaces of between 10,000 and 100,000 distinct lipids.

(a) Effect of Fatty Acid Chain Length on Search Space Size



(b) Effect of Chain Length and Double Bond Content on Search Space Size

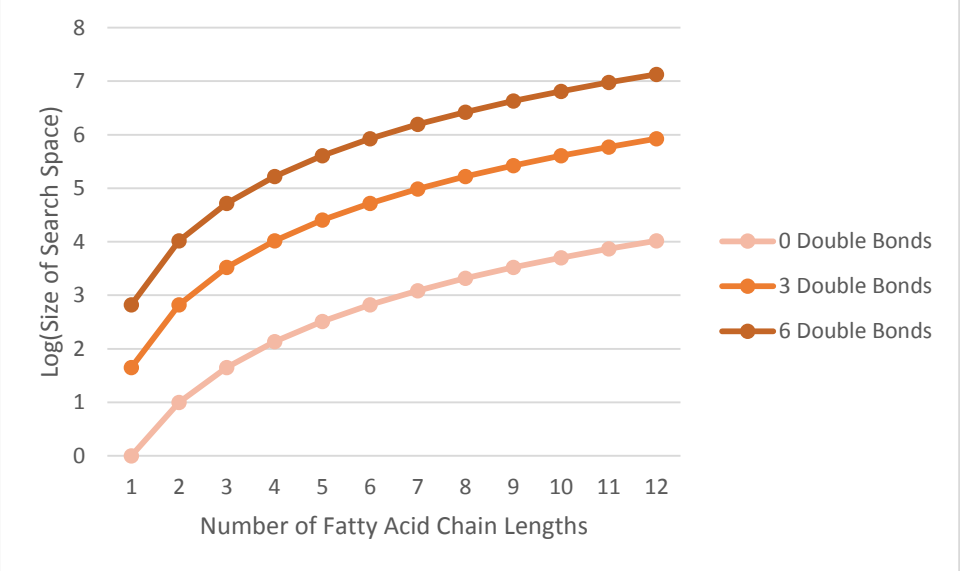


Figure 5. The search space grows rapidly as the number of possible fatty acid chain lengths increases, particularly for lipids with a greater number of fatty acid substituents (a). The number of possible saturation levels similarly affects the size of the search space and when combined can make the search space infeasibly large.

Validation against the NIST 2014 small molecule MS/MS Library

We sought to validate the fragmentation models and scoring algorithms of Greazy by testing the software against the NIST small molecule MS/MS library. We filtered the library for spectra that correspond to phospholipids and split the data into two sets based on the polarity of the ions (see Methods). The chosen search space covered the majority of phospholipids known to be in the NIST data set. Searches were performed in both positive and negative ion mode. Glycerophospholipids in the negative mode search included glycerophosphoethanolamines (PE), glycerophosphoglycerols (PG), glycerophosphoserines (PS), glycerophosphates (PA), glycerophosphoinositols (PI), and Phosphoinositides (PIP) with fatty acid chains ranging from 4 to 24 carbons and 0 to 6 double bonds, allowing for the possibility of lysoglycerophospholipids. The negative mode search also included CLs with fatty acid chains of 16-20 carbons and 0-4 double bonds. In total, the negative mode search included 315,673 distinct lipids. The positive mode search space included the glycerophosphocholines (PC) with the same fatty acid configurations as the glycerophospholipids in negative mode as well as sphingomyelins (SM) with an 18 carbon sphingosine base, 12-24 fatty acid carbons and 0-1 double bonds, for a total of 15,032 distinct lipids. The mass accuracy tolerances for these searches were 0.5 Daltons for both precursor and fragment ions because the data included spectra from ion trap, QTOF, and FTMS instruments. Only protonated and deprotonated precursors were included in the search.

Overall Identification Results. Of the 214 distinct lipids with MS/MS spectra in at least one of the two data sets, 170 were included within the search space for Greazy. While considering only the best scoring spectrum for each lipid on the list, Greazy correctly identified 157 of these lipids (92%). These results were consistent across lipid classes as Greazy correctly identified at least 87.5% of the detectable precursors for each class in the search space (Figure 6).

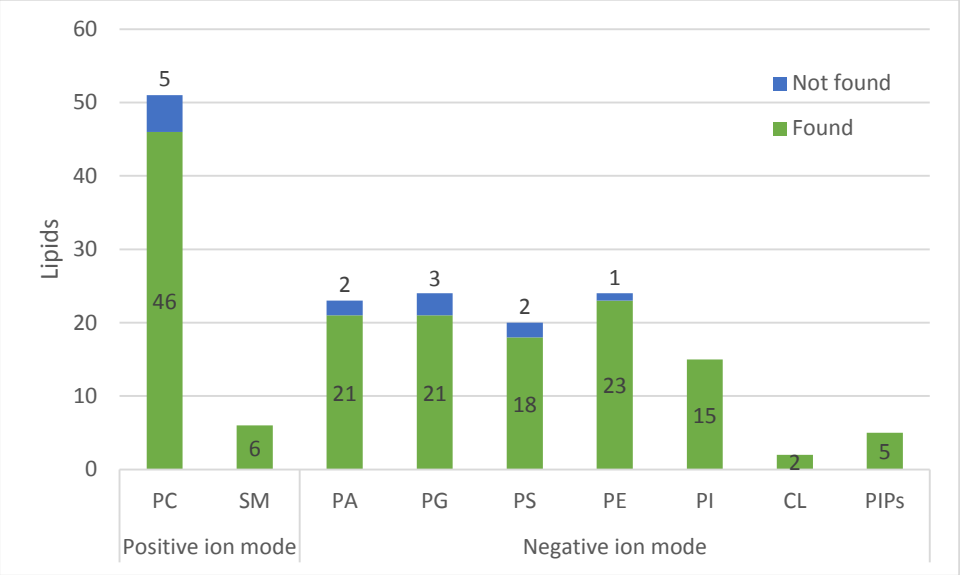


Figure 6. The filtered NIST small molecule MS/MS library contained 170 lipids that Greazy was capable of identifying. Greazy identified 157 of them.

Combined and Peak Only Scoring Accuracy. The match score employed by Greazy to match predicted spectra to experimental ones combines a peak matching score based on the hypergeometric distribution (HGD) with an intensity score that takes into account the percentage of the total ion current (TIC) accounted for in those peaks that match. In practice, the HGD produces lower p-values, dominating the TIC in combination. Users may specify that only the HGD score should be used. The NIST database searches were repeated using only HGD. Figures 7a and 7b display the numbers of correct and incorrect lipid identifications at different score thresholds. Using only the HGD score increased the number of false compounds identified when no threshold was applied (by 21.5% and 31.9% in positive and negative ion mode, respectively) while reducing the number of true compounds identified in negative ion mode (by 4.8%). The information of the TIC score is most useful in negative ion mode searches.

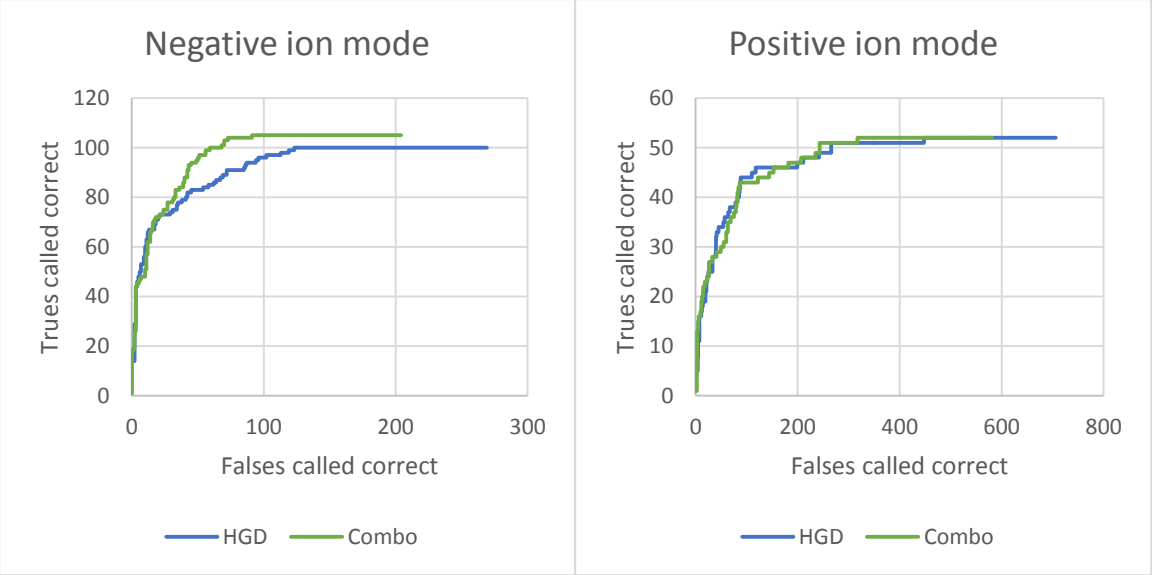


Figure 7. Correct vs incorrect matches for Greazy search and scoring of the NIST library using both the combined and HGD-only scoring algorithms in both positive (a) and negative (b) ion mode.

Spectrum Identification Accuracy. Greazy's final results emphasize the best evidence, or highest scoring spectrum, for a given lipid. The analysis shown in Figure 8, for example, ranks the lipids by the best spectrum match score for each. In this test, however, we evaluate the ability of the Greazy combined score to find the correct match for each spectrum. Figure 4 shows the rank of the correct matches for identifiable spectra, where rank 1 hits reflect the spectra that produce the highest score when matched to the correct lipids. The correct matches for 86% of spectra in negative mode and 59% in positive mode were ranked as the top match. In positive mode, a substantial number of spectra produced tie scores for multiple lipids; if the correct identity was tied with other lipids for the top score, it was treated as first rank. Spectra that contain a limited number of fragment ions allow many isomeric species to match equally well.

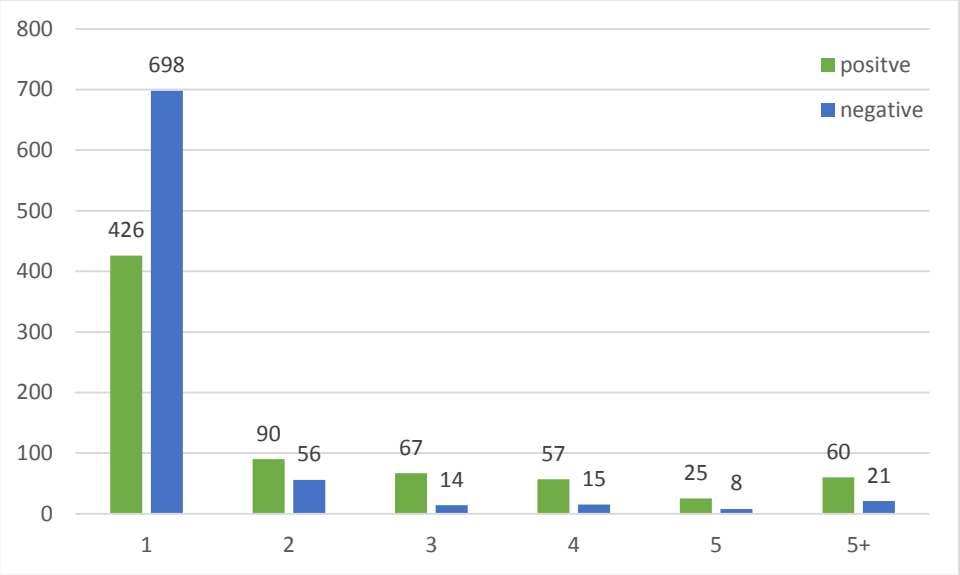


Figure 8. Correct match ranks for spectra in the NIST library whose identity falls within the search space.

Greazy at Various Energy Levels. Many of the phospholipids within the NIST metabolomics library were measured at various collision energies in a Thermo Finnigan LTQ Orbitrap Elite by ion trap collision-induced dissociation (CID) and higher-energy collisional dissociation (HCD), allowing us to test Greazy with different degrees of fragmentation. Under HCD conditions we looked at several classes of glycerophospholipids and a cardiolipin species in negative ion mode (Figure 9) and considered a PC and an SM in positive ion mode (Figure 10). In negative ion mode the collision energy ranged from 11 to 127 (normalized), and every lipid surveyed in this mode had its best score in the first half of this range. In positive mode both PCs and SMs scored poorly at all energy levels. By contrast, spectra for these species within the NIST library that were generated by ion trap CID produced much higher scores, recommending a different experimental setup for identification of PCs and SMs in positive mode. Higher scores in Greazy result from matching larger numbers of fragments; this relationship is illustrated in Figure 11.

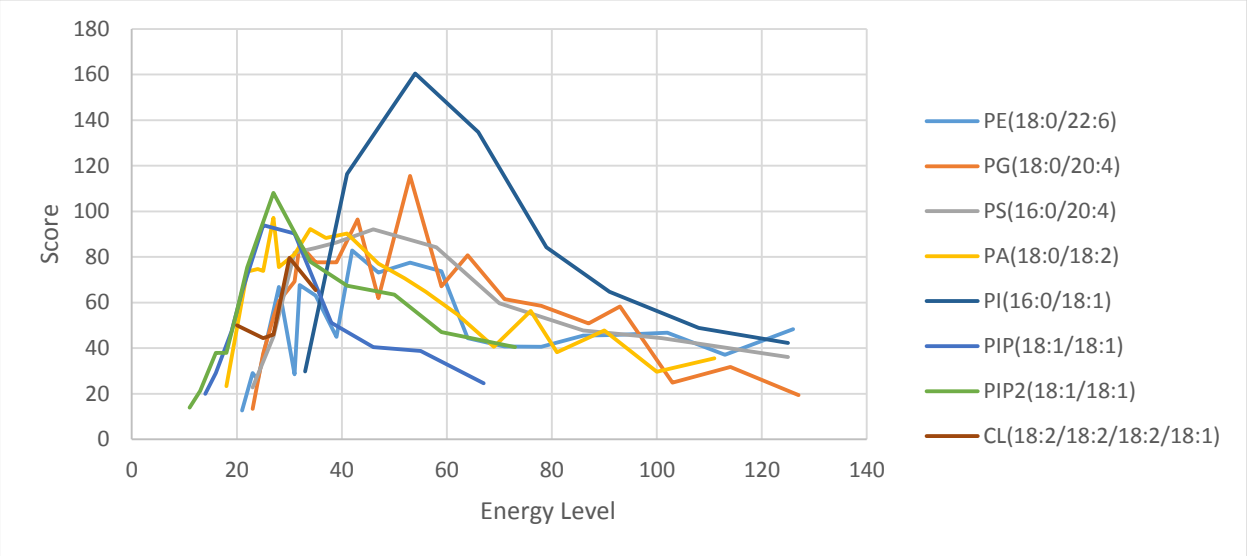


Figure 9. Score vs collision energy for examples of several lipid classes in negative ion mode. In each case the highest score was attained within the first half of the given energy range.

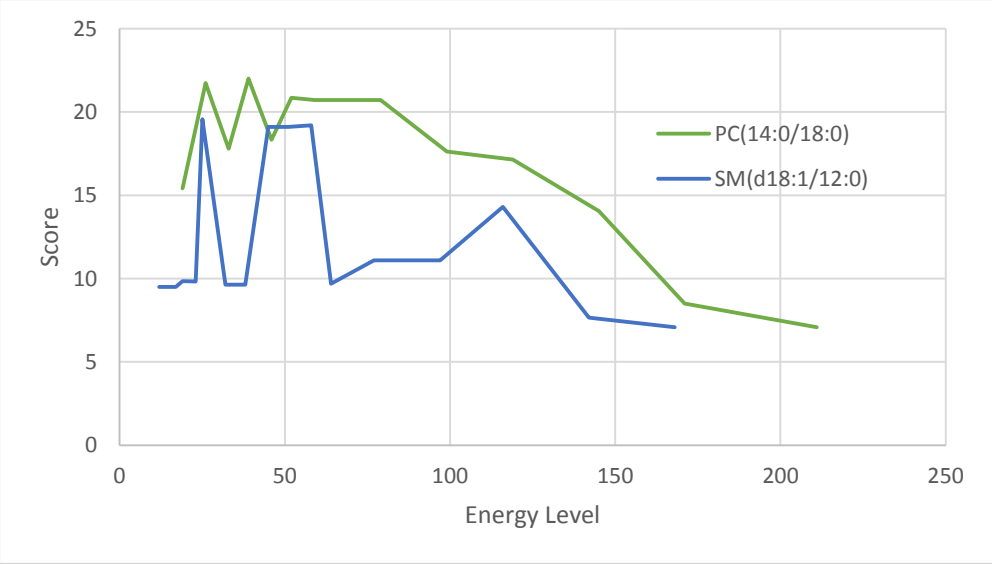


Figure 10. Score vs collision energy for a PC and an SM in positive ion mode. In both cases the high score was attained at the low end of the energy range.

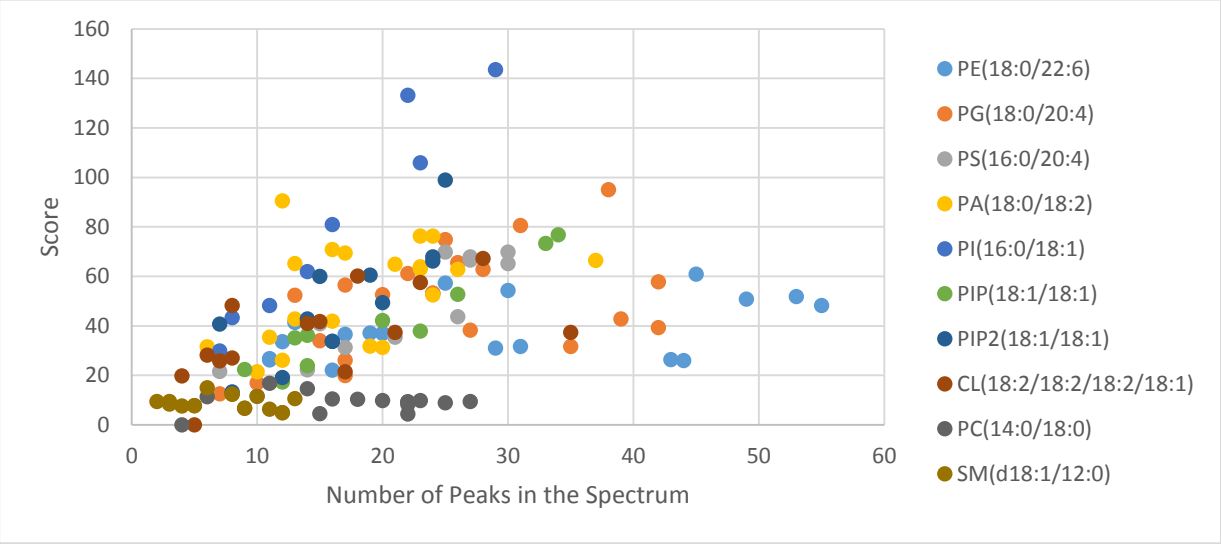


Figure 11. Score vs spectrum peak count. For those lipids run in negative ion mode the score obtained with Greazy generally rises with the number of peaks in the spectrum. For those run in positive mode the score obtained is flat regardless of the number of peaks in the spectrum.

Comparison of Greazy/LipidLama to LipidSearch

While the above tests were intended to evaluate the ability of Greazy to match the correct lipid composition to spectra, tools are also needed to determine which spectra have been successfully identified. The LipidLama tool has the same relationship to Greazy that PeptideProphet[39] does to SEQUEST[40] and other database search tools developed for peptide MS/MS data. Because decoy lipids are less easily generated than decoy peptide sequences, however, we opted to compare the distribution of best scores for spectra to the distribution of second-best scores. As such, LipidLama can be quite conservative in assessing lipid-spectrum matches (LSMs). LSMs are evaluated separately by lipid class since scores for each may have different distributions.

Greazy/LipidLama and LipidSearch (Thermo Scientific)[19] are both intended to discover the best lipid to explain each MS/MS spectrum. We compared the algorithms using LC-MS/MS data from samples of mouse bone marrow stem cells acquired on a Q Exactive mass spectrometer (see methods). Two raw files, one for positive mode and one for negative, were

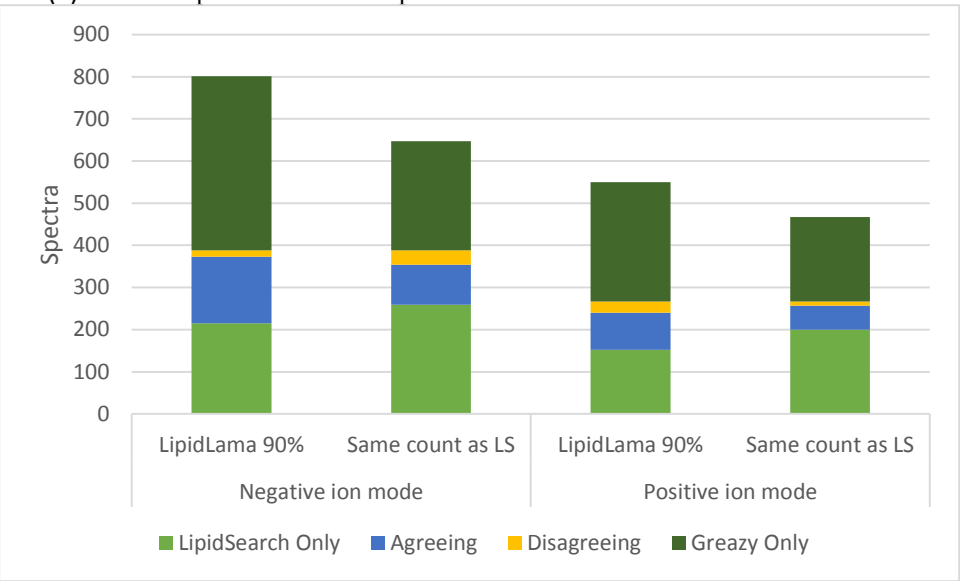
analyzed by both tools. The Greazy search space was chosen to capture the majority of the lipids identified by LipidSearch. The search space for the positive and negative runs were identical. Glycerophospholipids in the search space included the PCs, PEs, PIs, PGs, PSs, and PAs with fatty acid chains of 12-24 carbons and 0-6 double bonds. Lysoglycerophospholipids and lipids with an ether-linked fatty acid at one position were also included. SMs in the search space could incorporate a sphingosine or sphinganine base containing 16-24 carbons and a fatty acid chain of length 14-28 with 0-4 double bonds. CLs in the search space had fatty acid chains with 16-20 carbons and 0-4 double bonds. NH_4 and CH_3COO adducts were allowed in positive and negative mode, respectively. The mass accuracy tolerances for these searches were set to 20 ppm for both precursor and fragment ions. LipidLama was configured for an estimated 10% FDR, while the LipidSearch software was set to an m-score threshold of 5.0 (the m-score measures the fit between experimental and expected product ion spectra and is based on the number of matching fragments).

Comparison at the spectrum level. LipidSearch reported LSMs that were within the Greazy search space for 388 and 267 spectra in negative and positive ion mode, respectively. In negative ion mode an unfiltered Greazy search returned 2352 identified spectra, agreeing in 297 of the 388 LSMs (76.5%) identified by LipidSearch. In positive ion mode Greazy returned 1569 spectral identifications, agreeing in 132 of the 267 LipidSearch identifications (49.4%). Filtering Greazy's search results with LipidLama returned 586 and 398 LSMs in negative and positive ion mode, respectively. Figure 12a investigates the overlap between identifications in Greazy/LipidLama and LipidSearch. For those spectra with an identification reported by both platforms, 91.3% of LSMs agreed in negative ion mode and 76.5% in positive. When Greazy results were filtered to produce the same overall number of identifications as LipidSearch, a

similarly small overlap in identified spectra results. FigureS 13a and 13b separate the spectra by lipid class. Algorithmic differences may explain the divergence in these two sets of results. LipidSearch appears to be reporting only a subset of spectra for which the best match is to a particular lipid. Of the 413 negative ion and 283 positive ion LSMs reported only by Greazy/LipidLama 278 (67.3%) and 183 (64.7%) matched to lipids that both systems had identified in other data. LipidSearch can report identifications based only on an accurate mass and retention time. Greazy, however, requires the presence of MS/MS spectral data for identification; this difference would result in many LipidSearch-only identifications.

Comparison of lipid identifications. At the lipid level, LipidSearch identified 386 distinct lipids while Greazy/LipidLama yielded 286. They agreed on 187 of these lipids. Figure 12b gives the class breakdown of these results. Agreement between the two platforms is much better for the glycerophospholipids than for the CLs or SMs, whether one is comparing sets of identified spectra or sets of distinct lipids. There were several CLs identified by Greazy with the same complement of fatty acid substituents but in different positions than those identified by LipidSearch. Greazy's fragmentation model for CL is position-specific with regard to fatty acids while those for glycerophospholipids are not, except in the case of ether lipids. The added complexity and more rigid matching criteria hinders agreement between the two platforms. The SM model predicts far fewer fragment ions in these MS/MS spectra, decreasing the discrimination possible for this class. It should also be noted that LipidLama employs an alternative approach for CL and SM; the comparison of the distribution of top match scores and second-best match scores is not viable. Instead, LipidLama was configured to accept the top 10 percent of LSMs by score for these two classes.

(a) Overall Spectra Search Comparison



(b) Class by Class Lipid Search Comparison

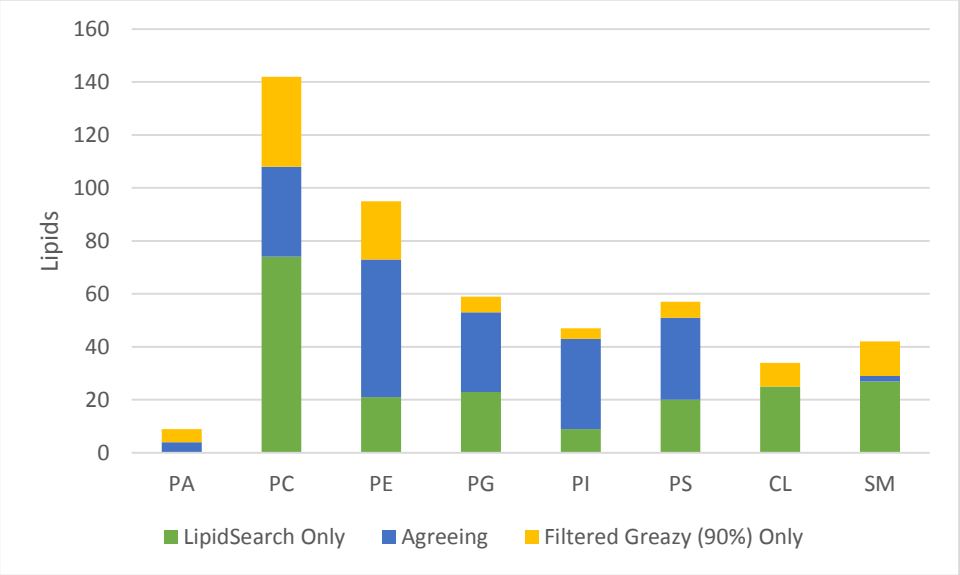
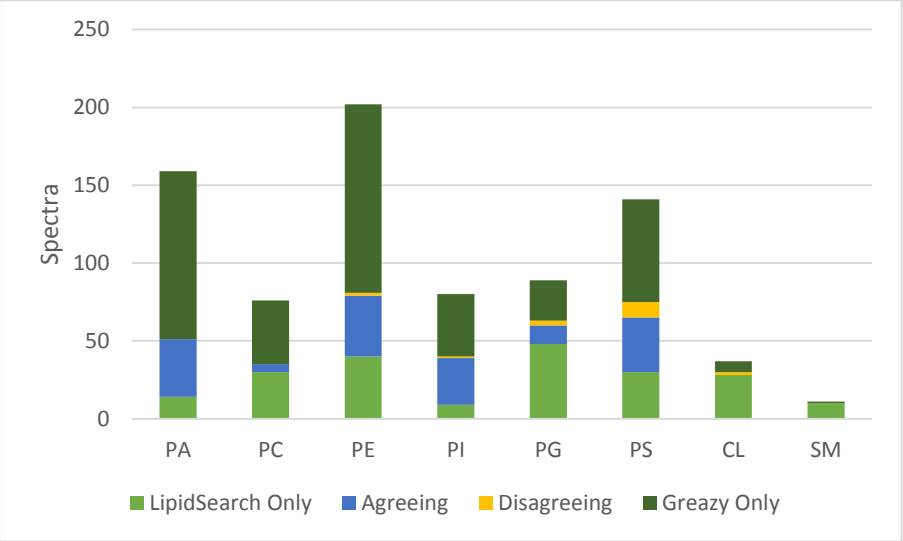


Figure 12. (a) Overlap of spectral identifications made by LipidSearch and Greazy in negative and positive ion mode. Greazy's results were filtered by LipidLama at a 10% FDR or by matching the number of hits reported by LipidSearch. (b) Overlap of class specific lipid identifications made by LipidSearch and Greazy.

(a) Breakdown of negative ion mode spectral identifications



(b) Breakdown of positive ion mode spectral identifications

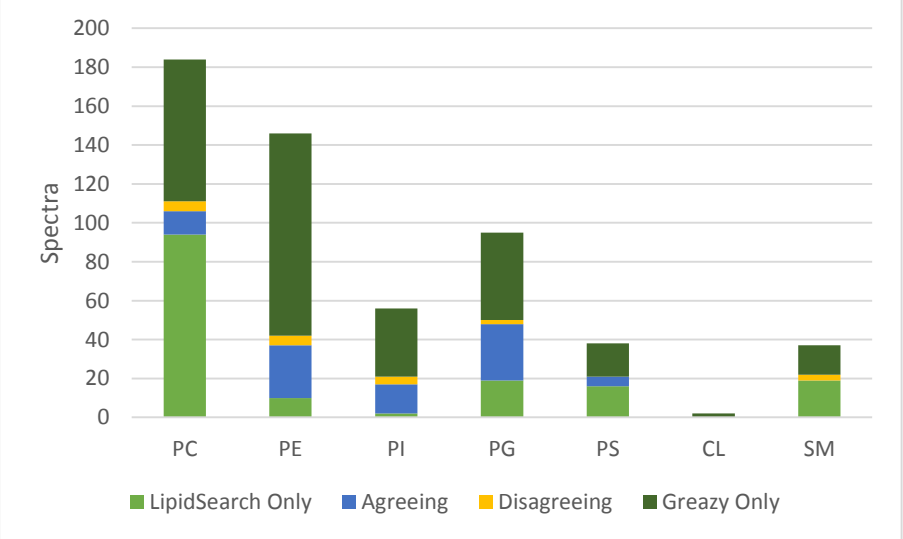


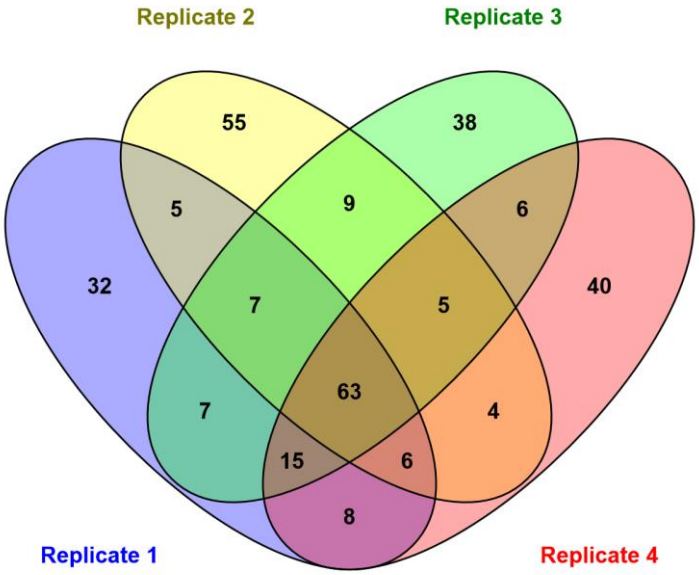
Figure 13. Overlap of class specific spectral identifications made by LipidSearch and Greazy in negative (a) and positive (b) ion mode. Greazy's results were filtered by LipidLama at a 90% confidence level.

Analysis of Biological Replicates with Greazy and LipidLama

We next used Greazy/LipidLama to analyze four biological replicates of SARS infected Calu-3 2B4 cells. These data sets, one for each replicate, were produced at PNNL (see methods). The search space included the glycerophospholipids PA, PE, PI, PG, PS and their lyso- counterparts with fatty acid chains of 12-24 carbons and 0-6 double bonds. Cardiolipins were included with fatty acid chains of 16-20 carbons and 0-4 double bonds. The search was made in negative mode and included only singly deprotonated precursors. The 20 ppm mass accuracy tolerance for this search applied to both precursor and fragment ions. LipidLama was applied at 10% FDR for the glycerophospholipids and retained the top 10 percent of cardiolipins.

There were 143, 154, 150, and 147 distinct lipids reported in replicates 1-4 respectively. 300 distinct lipids were found among all replicates, with 63 (21%) of them present in all four (Figure 14a). Of these 300 lipids, 277 came from the classes PE, PI, PG, and PS, with only 19 PAs and 4 CLs. (Figure 14b) Those lipids common to all replicates are displayed in Supporting Table S1 and include species from the PE, PI, PG, and PS classes. Most of these have fatty acid configurations that one would expect with chain lengths of 16, 18, 20, or 22 and an appropriate number of double bonds (16:0, 20:4, 22:6, etc.). Another 165 of the 300 lipids fell into only one of the four replicates, suggesting a degree of variability among biological replicates that is reminiscent of that seen in proteomics[41].

(a) Four replicates of SARS infected Calu-3 2B4 cells



(b) Lipid class replicate count

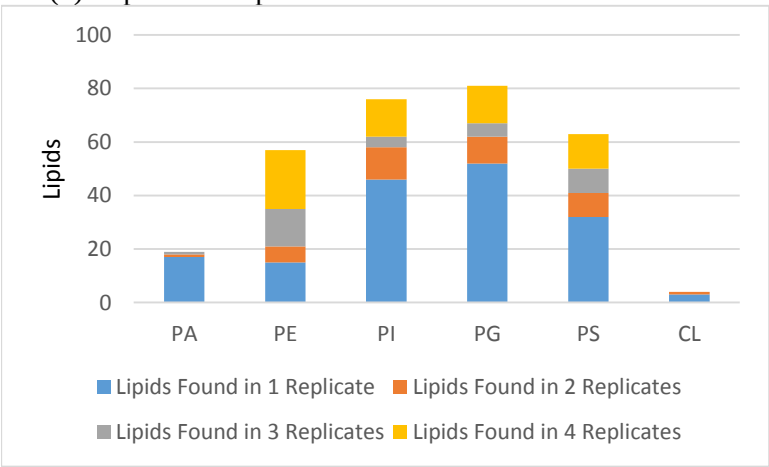


Figure 14. (a) A Greazy search of four biological replicates of SARS infected Calu-3 2B4 cells. 300 distinct lipids were identified among all four data sets with 63, or 21% of them, present in each. (b) A lipid class breakdown of the number of lipids found in 1, 2, 3, or 4 of the biological replicates.

Table 2. Lipids present in all four biological replicates of SARS infected Calu-3 2B4 cells.

PE	PG	PI	PS
PE(16:0/18:1)	PG(16:0/18:0)	PI(16:0/18:1)	PS(16:0/18:1)
PE(16:0/20:4)	PG(16:0/18:1)	PI(16:0/20:3)	PS(16:0/22:6)
PE(16:0/20:5)	PG(16:0/22:6)	PI(16:0/20:4)	PS(16:1/18:0)
PE(16:1/18:0)	PG(16:1/18:0)	PI(16:1/18:0)	PS(18:0/18:0)
PE(16:1/18:1)	PG(16:1/18:1)	PI(17:2/21:2)	PS(18:0/18:1)
PE(16:1/20:4)	PG(18:0/18:1)	PI(18:0/18:1)	PS(18:0/19:1)
PE(17:1/18:1)	PG(18:0/18:2)	PI(18:0/18:2)	PS(18:0/20:3)
PE(18:0/18:1)	PG(18:1/18:1)	PI(18:0/20:2)	PS(18:0/20:4)
PE(18:0/18:2)	PG(18:1/18:2)	PI(18:0/20:3)	PS(18:0/22:4)
PE(18:0/20:3)	PG(18:1/20:4)	PI(18:0/20:4)	PS(18:0/22:5)
PE(18:0/20:4)	PG(18:1/22:5)	PI(18:0/22:4)	PS(18:0/22:6)
PE(18:0/20:5)	PG(18:1/22:6)	PI(18:0/22:5)	PS(18:1/0:0)
PE(18:0/22:4)	PG(18:2/22:6)	PI(18:0/22:6)	PS(18:1/20:1)
PE(18:0/22:6)	PG(20:4/23:5)	PI(18:1/18:1)	
PE(18:1/18:2)			
PE(18:1/20:1)			
PE(18:1/20:4)			
PE(20:0/20:4)			
PE(20:4/0:0)			
PE(20:4/22:0)			
PE(20:4/24:0)			
PE(20:5/0:0)			

Conclusions and Future Work

Greazy and LipidLama are intended to remove a key bottleneck to lipidomic experiments by automating the process of identification from tandem mass spectrometry data. Their approach borrows heavily from comparable algorithms in proteomics. Even when a hundred thousand distinct lipids are included in the search space, the algorithm can complete its analysis of an LC-MS/MS experiment in much less time than was required to collect the data at the mass spectrometer. Tests in the gold-standard NIST MS/MS library establish that the software is highly effective in recognizing lipids. A relatively low intersection of identifications with the commercial LipidSearch software likely reflects differences in data processing. Nevertheless, agreement was high within this intersection. The release of a quality open-source codebase for lipid identification may be a powerful enabler for laboratories seeking to apply lipidomics to biological problems.

Although Greazy and LipidLama successfully bring a number of methods from proteomics to the analysis of phospholipids there are a number of ways in which the software could be improved or extended. The fragmentation models, for example, should be continually updated as new information appears in the literature. It may also be advantageous to allow for their customization in order to account for experimental conditions. The scope of the software could also be expanded. Currently, Greazy focuses only on a limited number of phospholipid classes. A number of additional glycerophospholipid and sphingolipid classes could easily be added as could classes from the glycerolipid category. It may also be prudent to separate the software into two components: generation of theoretical spectra and scoring/FDR estimation. This would allow for easier integration with other toolsets.

The scoring and FDR estimation algorithms used in Greazy and LipidLama are new to the field of lipidomics and are good candidates for improvement. Greazy, for example, does not predict the

expected intensities of peaks in theoretical spectra. Including such information and developing a scoring algorithm to utilize it could be extremely useful in cases where particular fragments dominate a lipids spectrum. The inclusion of a method to estimate the false discover rate is, to our knowledge, a first in lipidomics. Although the mixture modeling approach, as implemented in LipidLama, works well for the glycerophospholipids it is not suitable for use with the cardiolipins or sphingolipids. A general method for FDR estimation that is applicable to all lipid classes would clearly be advantageous. Much more work in this area is warranted.

Finally, identification and quantification are equally important in the field of lipidomics. Incorporation of a software component to accurately quantify the identified lipids will be necessary if Greazy/LipidLama is going to be a true alternative to available commercial software.

Appendix

A. Fragmentation Models

Table A1: Glycerophospholipid Fragments with Nonmetal Adducts in Positive Ion Mode

	PC	**PE**	**PS**	**PG**	**PA**	**PI**	**PIP**
HG (head group)	X	X	X				
loss of HG	X	X	X	X	X	X	X
loss of FA1 (fatty acid 1) as ketene	X	X	X	X	X	X	X
loss of FA1 as carboxylic acid	X	X	X	X	X	X	X
loss of HG and FA1 as ketene	X	X	X	X	X	X	X
R1CO	X	X	X	X	X	X	X
loss of FA2 as ketene	X	X	X	X	X	X	X
loss of FA2 as carboxylic acid	X	X	X	X	X	X	X
loss of HG and FA2 as ketene	X	X	X	X	X	X	X
R2CO	X	X	X	X	X	X	X
choline	X						
loss of NC_3H_9	X						

Table A2: Glycerophospholipids in Negative Ion Mode

	PC	PE	PS	PG	PA	PI
G3P-H2O	X	X	X	X	X	X
G3P	X	X	X	X	X	X
H2PO4	X	X	X	X	X	X
PO3	X	X	X	X	X	X
FA1 carboxylate anion	X	X	X	X	X	X
loss of FA1 as ketene	X	X	X	X	X	X
loss of FA1 as carboxylic acid	X	X	X	X	X	X
FA2 carboxylate anion	X	X	X	X	X	X
loss of FA2 as ketene	X	X	X	X	X	X
loss of FA2 as carboxylic acid	X	X	X	X	X	X
loss of head		X	X	X		X
loss of head and FA1 as carboxylic acid		X	X	X		X
loss of head and FA1 as ketene		X	X	X		X
loss of head and FA2 as carboxylic acid		X	X	X		X
loss of head and FA2 as ketene		X	X	X		X
head group		X	X			X
head group - H2O		X				X
head group - 2*H2O						X
demethylation[a]	X					
loss of C3H10N	X					
loss of C5H12N	X					
GPC-CH3-H2O	X					
PC-CH3	X					
demethylation and loss of FA1 as ketene[a]	X					
loss of C3H10N and loss of FA1 as ketene	X					
loss of C5H12N and loss of FA1 as ketene	X					
demethylation and loss of FA1 as carboxylic acid[a]	X					
loss of C3H10N and loss of FA1 as carboxylic acid	X					
loss of C5H12N and loss of FA1 as carboxylic acid	X					
demethylation and loss of FA2 as ketene[a]	X					
loss of C3H10N and loss of FA2 as ketene	X					
loss of C5H12N and loss of FA2 as ketene	X					
demethylation and loss of FA2 as carboxylic acid[a]	X					
loss of C3H10N and loss of FA2 as carboxylic acid	X					
loss of C5H12N and loss of FA2 as carboxylic acid	X					
glycerophosphoinositol						X
glycerophosphoinositol - H2O						X
glycerophosphoinositol - 2*H2O						X
glycerophosphoinositol - 3*H2O						X
glycerophosphoglycerol				X		
glycerophosphoglycerol - H2O				X		
glycerophosphoglycerol - 2*H2O				X		

a. Fragment not present in demethylated precursor

Table A3: Phosphatidylinositides in Negative Ion Mode

	PIP	PIP (-2)	PIP2	PIP2 (-2)	PIP3
G3P-H2O	X	X	X	X	X
G3P					
H2PO4	X	X	X	X	X
PO3	X	X	X	X	X
FA1 carboxylate anion	X	X	X	X	X
loss of FA1 as ketene	X				
loss of FA1 as carboxylic acid	X				
FA2 carboxylate anion	X	X	X	X	X
loss of FA2 as ketene	X	X			
loss of FA2 as carboxylic acid	X				
loss of FA1 as carboxylate anion		X		X	
loss of FA2 as carboxylate anion		X		X	
loss of FA2 as ketene and PO3		X			
loss of FA2 as ketene and FA1 carboxylate anion		X			
loss of inositol phosphate	X				
loss of inositol	X				
loss of H2O	X		X		X
HO6P2	X		X	X	X
H3O7P2	X		X		X
loss of HPO3	X		X		X
loss of HPO3 & inositol			X		
loss of 2*HPO3 & inositol			X		
loss of HPO3 & inositol; addition of H2O			X		
loss of H3PO4	X		X		X
loss of 2*H3PO4					X
loss of H3PO4 & H2O			X		X
loss of inositol & H2O	X				
loss of FA1 as carboxylic acid; loss of H2O	X		X		
loss of FA2 as carboxylic acid; loss of H2O	X		X		
loss of H3PO4 & FA1 as carboxylic acid			X		X
loss of H3PO4 & FA2 as carboxylic acid			X		X
inositol phosphate		X			
inositol phosphate - H2O	X	X	X	X	X
inositol phosphate - 2*H2O	X	X	X	X	X
inositol phosphate - 3*H2O	X	X	X		X
inositol bisphosphate			X		X
inositol bisphosphate - H2O	X		X	X	X
inositol bisphosphate - 2*H2O	X		X	X	X
inositol trisphosphate + H2O			X		
inositol trisphosphate			X		X
inositol trisphosphate - H2O			X		X
inositol trisphosphate - 2*H2O			X		X
inositol tetraphosphate					X
inositol tetraphosphate - H2O					X
inositol tetraphosphate - 2*H2O					X
loss of PO3		X		X	
loss of PO3; loss of H2O				X	
bisphosphate (2-)		X			

bisphosphate - H2O (2-)		X			
loss of HO6P2				X	
loss of HO6P2; loss of H2O				X	
loss of PO3 & inositol; addition of H2O				X	

Table A4: Phosphosphingolipid Fragments with Nonmetal Adducts in Positive Ion Mode

	SM	Cer-PE	Cer-PI
loss of HG	X	X	X
loss of HG and H2O	X	X	X
loss of HG; addition of H2O		X	X
loss of HG and fatty acid	X	X	X
loss of HG and fatty acid; addiction of H2O	X	X	
fatty acid + NC2H3	X	X	X
fatty acid + NH3	X	X	X
HG	X	X	
loss of NC3H9	X		
head	X		
loss of head		X	
loss of H2O			X
loss of 2*H2O			X

Table A5: Phosphosphingolipid Fragments with Metal Adducts in Positive Ion Mode

	SM	Cer-PE	Cer-PI
loss of HG	X	X	X
loss of HG and adduct	X	X	X
loss of HG , adduct, and H2O	X	X	X
loss of HG and H2O	X	X	
loss of HG and CH2O	X	X	X
loss of HG, adduct, and FA	X	X	X
loss of HG and FA; addition of H2O	X	X	
FA + NC2H3	X	X	X
FA + NH3	X	X	X
FA + NH3 and adduct	X	X	
HG	X	X	
loss of NC3H9	X		
head	X		
PO4C2H5 and adduct	X		
loss of HG; addition of water		X	X
loss of head		X	X
loss of head and H2O			X
HG and adduct			X
HG and adduct minus H2O			X
loss of H2O			X

Table A6: Phosphosphingolipid Fragments in Negative Ion Mode

SL negative mode	SM	Cer-PE	Cer-PI
loss of methyl group[a]	X		
loss of C3H10N	X		
loss of head and adduct	X		
loss of head		X	X
HG		X	X
HG – H2O		X	X
HG – 2H2O			X

Table A7: Cardiolipin Fragments in Positive Ion Mode

protonated, + Alkali Metal, - H + 2Alkali Metal	CL - 2H + 3Alkali Metal
loss of FA1, FA2, & glycerol	loss of FA1 as carboxylic acid
loss of FA1, FA2, glycerol, & H2O	loss of FA2 as carboxylic acid
loss of FA3, FA4, & glycerol	loss of FA3 as carboxylic acid
loss of FA3, FA4, glycerol, & H2O	loss of FA4 as carboxylic acid
loss of FA1, FA2, glycerol & glycerol-1,3-diphosphate[b]	loss of FA1 & FA3 as carboxylic acids
loss of FA1, FA2, FA3, glycerol & glycerol-1,3-	loss of FA1 & FA2 as carboxylic acids
FA3 cation	loss of FA3 & FA4 as carboxylic acids
FA4 cation	loss of FA1, FA2, and glycerol
loss of FA3, FA4, glycerol & glycerol-1,3-diphosphate[b]	loss of FA1, FA2, glycerol, and FA3 as a carboxylic
loss of FA3, FA4, FA1, glycerol & glycerol-1,3-	loss of FA1, FA2, glycerol, and FA3 as an alkali salt
FA1 cation	loss of FA1, FA2, glycerol, and FA4 as a carboxylic
FA2 cation	loss of FA1, FA2, glycerol, and FA4 as an alkali salt
	loss of FA3, FA4, and glycerol
	loss of FA3, FA4, glycerol, and FA1 as a carboxylic
	loss of FA3, FA4, glycerol, and FA1 as an alkali salt
	loss of FA3, FA4, glycerol, and FA2 as a carboxylic
	loss of FA3, FA4, glycerol, and FA2 as an alkali salt
	loss of FA1, FA2, and sodiated glycerophosphatidic
	loss of FA1, FA2, sodiated glycerophosphatidic acid, and FA3 as a carboxylic acid
	loss of FA1, FA2, and phosphoglycerophosphate

a. Loss of metal adduct if present (sodium, di-sodium, etc.)

Table A8: Cardiolipin Fragments in Negative Ion Mode – Deprotonated

G3P-H2O
G3P
H2PO4
PO3
FA1 carboxylate anion
FA2 carboxylate anion
FA3 carboxylate anion
FA4 carboxylate anion
loss of FA1 as carboxylic acid
loss of FA2 as carboxylic acid
loss of FA3 as carboxylic acid
loss of FA4 as carboxylic acid
loss of FA1, FA2 & glycerol
loss of FA1, FA2, FA3 as carboxylic acid & glycerol
loss of FA1, FA2, FA4 as carboxylic acid & glycerol
loss of FA1, FA2, FA3 as ketene & glycerol
loss of FA1, FA2, FA4 as ketene & glycerol
loss of FA1, FA2 & phosphoglycerol
loss of FA1, FA2, FA3 as carboxylic acid & glycerophosphoglycerol
loss of FA1, FA2, FA4 as carboxylic acid & glycerophosphoglycerol
loss of FA1, FA2, FA3 as ketene & glycerophosphoglycerol
loss of FA1, FA2, FA4 as ketene & glycerophosphoglycerol
loss of FA1, FA2, phosphoglycerol & FA3 as a carboxylic acid
loss of FA1, FA2, phosphoglycerol & FA4 as a carboxylic acid
loss of FA1, FA2, phosphoglycerol & FA3 as a ketene
loss of FA1, FA2, phosphoglycerol & FA4 as a ketene
loss of FA1, FA2 & glycerophosphoglycerol
loss of FA3, FA4 & glycerol
loss of FA3, FA4, FA1 as carboxylic acid & glycerol
loss of FA3, FA4, FA2 as carboxylic acid & glycerol
loss of FA3, FA4, FA1 as ketene & glycerol
loss of FA3, FA4, FA2 as ketene & glycerol
loss of FA3, FA4 & phosphoglycerol
loss of FA3, FA4, FA1 as carboxylic acid & glycerophosphoglycerol
loss of FA3, FA4, FA2 as carboxylic acid & glycerophosphoglycerol
loss of FA3, FA4, FA1 as ketene & glycerophosphoglycerol
loss of FA3, FA4, FA2 as ketene & glycerophosphoglycerol
loss of FA3, FA4, phosphoglycerol & FA1 as a carboxylic acid
loss of FA3, FA4, phosphoglycerol & FA2 as a carboxylic acid
loss of FA3, FA4, phosphoglycerol & FA1 as a ketene
loss of FA3, FA4, phosphoglycerol & FA2 as a ketene
loss of FA3, FA4 & glycerophosphoglycerol

Table A9: Cardiolipin Fragments in Negative Ion Mode – Doubly Deprotonated

G3P-H2O
G3P
H2PO4
PO3
loss of FA1 carboxylic anion
loss of FA2 carboxylic anion
loss of FA3 carboxylic anion
loss of FA4 carboxylic anion
FA1 carboxylate anion
FA2 carboxylate anion
FA3 carboxylate anion
FA4 carboxylate anion
loss of FA2 as ketene
loss of FA4 as ketene
loss of FA1, FA2 & phosphoglycerol
loss of FA1, FA2, phosphoglycerol & FA3 as a carboxylic acid
loss of FA1, FA2, phosphoglycerol & FA4 as a carboxylic acid
loss of FA1, FA2, phosphoglycerol & FA3 as a ketene
loss of FA1, FA2, phosphoglycerol & FA4 as a ketene
loss of FA1, FA2 & glycerophosphoglycerol
loss of FA1, FA2, FA3 as carboxylic acid & glycerophosphoglycerol
loss of FA1, FA2, FA4 as carboxylic acid & glycerophosphoglycerol
loss of FA1, FA2, FA3 as ketene & glycerophosphoglycerol
loss of FA1, FA2, FA4 as ketene & glycerophosphoglycerol
loss of FA3, FA4 & phosphoglycerol
loss of FA3, FA4, phosphoglycerol & FA1 as a carboxylic acid
loss of FA3, FA4, phosphoglycerol & FA2 as a carboxylic acid
loss of FA3, FA4, phosphoglycerol & FA1 as a ketene
loss of FA3, FA4, phosphoglycerol & FA2 as a ketene
loss of FA3, FA4 & glycerophosphoglycerol
loss of FA3, FA4, FA1 as carboxylic acid & glycerophosphoglycerol
loss of FA3, FA4, FA2 as carboxylic acid & glycerophosphoglycerol
loss of FA3, FA4, FA1 as ketene & glycerophosphoglycerol
loss of FA3, FA4, FA2 as ketene & glycerophosphoglycerol

Table A10: Cardiolipin Fragments in Negative Ion Mode – Doubly Deprotonated plus Alkali Metal

CL - 2H + Alkali Metal
loss of FA1 as a carboxylic acid
loss of FA2 as a carboxylic acid
loss of FA3 as a carboxylic acid
loss of FA4 as a carboxylic acid
loss of FA1 as a ketene
loss of FA2 as a ketene
loss of FA3 as a ketene
loss of FA4 as a ketene
loss of FA1 as an alkali salt
loss of FA2 as an alkali salt
loss of FA3 as an alkali salt
loss of FA4 as an alkali salt
loss of FA1, FA2 & glycerol
loss of FA1, FA2 & glycerol and FA3 as a carboxylic acid
loss of FA1, FA2 & glycerol and FA3 as a ketene
loss of FA1, FA2 & glycerol and FA3 as an alkali salt
loss of FA1, FA2 & glycerol and FA4 as a carboxylic acid
loss of FA1, FA2 & glycerol and FA4 as a ketene
loss of FA1, FA2 & glycerol and FA4 as an alkali salt
loss of FA3, FA4 & glycerol
loss of FA3, FA4 & glycerol and FA1 as a carboxylic acid
loss of FA3, FA4 & glycerol and FA1 as a ketene
loss of FA3, FA4 & glycerol and FA1 as an alkali salt
loss of FA3, FA4 & glycerol and FA2 as a carboxylic acid
loss of FA3, FA4 & glycerol and FA2 as a ketene
loss of FA3, FA4 & glycerol and FA2 as an alkali salt
loss of FA1, FA2 & glycerophosphoglycerol
loss of FA1, FA2 & glycerophosphoglycerol and FA3 as an alkali salt
loss of FA1, FA2 & glycerophosphoglycerol and FA4 as an alkali salt
loss of FA3, FA4 & glycerophosphoglycerol
loss of FA3, FA4 & glycerophosphoglycerol and FA1 as an alkali salt
loss of FA3, FA4 & glycerophosphoglycerol and FA2 as an alkali salt
loss of FA1, FA2, and alkalated glycerophsphatidic acid
loss of FA3, FA4, and alkalated glycerophsphatidic acid

REFERENCES

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

2. Van Meer, G., Voelker, D. R., and Feigenson, G. W. (2008) Membrane lipids: where they are and how they behave. *Nat. Rev. Mol. Cell Biol. 9*, 112–124.

3. Di Paolo, G., and De Camilli, P. (2006) Phosphoinositides in cell regulation and membrane dynamics. *Nature 443*, 651–657.

4. Sorensen, C. M., Ding, J., Zhang, Q., Alquier, T., Zhao, R., Mueller, P. W., Smith, R. D., and Metz, T. O. (2010) Perturbations in the lipid profile of individuals with newly diagnosed type 1 diabetes mellitus: lipidomics analysis of a Diabetes Antibody Standardization Program sample subset. *Clin. Biochem. 43*, 948–956.

5. Han, X., Rozen, S., Boyle, S. H., Hellegers, C., Cheng, H., Burke, J. R., Welsh-Bohmer, K. A., Doraiswamy, P. M., and Kaddurah-Daouk, R. (2011) Metabolomics in early Alzheimer's disease: identification of altered plasma sphingolipidome using shotgun lipidomics. *PLoS ONE 6*, e21643.

6. Yan, Y., and Kang, B. (2012) The Role of Cardiolipin Remodeling in Mitochondrial Function and Human Diseases. *Journal of Molecular Biology Research 2*, p1.

7. Fernandis, A. Z., and Wenk, M. R. (2009) Lipid-based biomarkers for cancer. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. 877*, 2830–2835.

8. Dória, M. L., Cotrim, Z., Macedo, B., Simões, C., Domingues, P., Helguero, L., and Domingues, M. R. (2012) Lipidomic approach to identify patterns in phospholipid profiles and define class differences in mammary epithelial and breast cancer cells. *Breast Cancer Res. Treat. 133*, 635–648.

9. Min, H. K., Lim, S., Chung, B. C., and Moon, M. H. (2011) Shotgun lipidomics for candidate biomarkers of urinary phospholipids in prostate cancer. *Anal Bioanal Chem 399*, 823–830.

10. 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.

11. Yang, K., Cheng, H., Gross, R. W., and Han, X. (2009) Automated Lipid Identification and Quantification by Multi-dimensional Mass Spectrometry-Based Shotgun Lipidomics. Anal Chem 81, 4356–4368.

12. Ejsing, C. S., Duchoslav, E., Sampaio, J., Simons, K., Bonner, R., Thiele, C., Ekroos, K., and Shevchenko, A. (2006) Automated identification and quantification of glycerophospholipid molecular species by multiple precursor ion scanning. Anal. Chem. 78, 6202–6214.

13. Song, H., Hsu, F.-F., Ladenson, J., and Turk, J. (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.

14. Herzog, R., Schwudke, D., Schuhmann, K., Sampaio, J. L., Bornstein, S. R., Schroeder, M., and Shevchenko, A. (2011) A novel informatics concept for high-throughput shotgun lipidomics based on the molecular fragmentation query language. Genome Biology 12, R8.

15. Schwudke, D., Oegema, J., Burton, L., Entchev, E., Hannich, J. T., Ejsing, C. S., Kurzchalia, T., and Shevchenko, A. (2006) Lipid profiling by multiple precursor and neutral loss scanning driven by the data-dependent acquisition. Anal. Chem. 78, 585–595.

16. Lipid Data Analyzer: unattended identification and quantitation of lipids in LC-MS data. Bioinformatics 27, 572–577.

17. Pluskal, T., Castillo, S., Villar-Briones, A., and Orešič, M. (2010) MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics 11, 395.

18. Ejsing, C. S., Duchoslav, E., Sampaio, J., Simons, K., Bonner, R., Thiele, C., Ekroos, K., and Shevchenko, A. (2006) Automated Identification and Quantification of Glycerophospholipid Molecular Species by Multiple Precursor Ion Scanning. *Anal. Chem. 78*, 6202–6214.

19. Taguchi, R., Nishijima, M., and Shimizu, T. (2007) Basic Analytical Systems for Lipidomics by Mass Spectrometry in Japan, in Methods in Enzymology (Brown, H. A., Ed.), pp 185–211. Academic Press.

20. Schmelzer, K., Fahy, E., Subramaniam, S., and Dennis, E. A. (2007) The lipid maps initiative in lipidomics. Meth. Enzymol. 432, 171–183.

21. Kind, T., Liu, K.-H., Lee, D. Y., DeFelice, B., Meissen, J. K., and Fiehn, O. (2013) LipidBlast in silico tandem mass spectrometry database for lipid identification. Nat. Methods 10, 755–758.

22. Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T. A., Brusniak, M.-Y., Paulse, C., Creasy, D., Flashner, L., Kani, K., Moulding, C., Seymour, S. L., Nuwaysir, L. M., Lefebvre, B., Kuhlmann, F., Roark, J., Rainer, P., Detlev, S., Hemenway, T., Huhmer, A., Langridge, J., Connolly, B., Chadick, T., Holly, K., Eckels, J., Deutsch, E. W., Moritz, R. L., Katz, J. E., Agus, D. B., MacCoss, M., Tabb, D. L., and Mallick, P. (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotech 30*, 918–920.

23. Griss, J., Jones, A. R., Sachsenberg, T., Walzer, M., Gatto, L., Hartler, J., Thallinger, G. G., Salek, R. M., Steinbeck, C., Neuhauser, N., Cox, J., Neumann, S., Fan, J., Reisinger, F., Xu, Q.-W., Del Toro, N., Pérez-Riverol, Y., Ghali, F., Bandeira, N., Xenarios, I., Kohlbacher, O., Vizcaíno, J. A., and Hermjakob, H. (2014) The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. Mol. Cell Proteomics 13, 2765–2775.

24. 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. *J Chromatogr B Analyt Technol Biomed Life Sci 877*, 2673–2695.

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

26. Murphy, R. C., Fiedler, J., and Hevko, J. (2001) Analysis of Nonvolatile Lipids by Mass Spectrometry. *Chem. Rev. 101*, 479–526.

27. Ho, Y.-P., Huang, P.-C., and Deng, K.-H. (2003) Metal ion complexes in the structural analysis of phospholipids by electrospray ionization tandem mass spectrometry. *Rapid Commun. Mass Spectrom. 17*, 114–121.

28. Han, X., and Gross, R. W. (1995) Structural determination of picomole amounts of phospholipids via electrospray ionization tandem mass spectrometry. *J. Am. Soc. Mass Spectrom. 6*, 1202–1210.

29. Hsu, F. F., and Turk, J. (2000) Characterization of phosphatidylinositol, phosphatidylinositol-4-phosphate, and phosphatidylinositol-4,5-bisphosphate by electrospray ionization tandem mass spectrometry: a mechanistic study. *J. Am. Soc. Mass Spectrom. 11*, 986–999.

30. Milne, S. B., Ivanova, P. T., DeCamp, D., Hsueh, R. C., and Brown, H. A. (2005) A targeted mass spectrometric analysis of phosphatidylinositol phosphate species. *J. Lipid Res. 46*, 1796–1802.

31. Byrdwell, W. C., and Perry, R. H. (2006) Liquid chromatography with dual parallel mass spectrometry and (31)P nuclear magnetic resonance spectroscopy for analysis of sphingomyelin and dihydrosphingomyelin. I. Bovine brain and chicken egg yolk. *J Chromatogr A 1133*, 149–171.

32. Hsu, F.-F., and Turk, J. (2000) Structural determination of sphingomyelin by tandem mass spectrometry with electrospray ionization. *Journal of the American Society for Mass Spectrometry 11*, 437–449.

33. Hsu, F.-F., Turk, J., Zhang, K., and Beverley, S. M. (2007) Characterization of Inositol Phosphorylceramides from Leishmania major by Tandem Mass Spectrometry with Electrospray Ionization. *J Am Soc Mass Spectrom 18*, 1591–1604.

34. Hsu, F.-F., and Turk, J. (2006) Characterization of Cardiolipin as the Sodiated Ions by Positive-Ion Electrospray Ionization with Multiple Stage Quadrupole Ion-Trap Mass Spectrometry. *Journal of the American Society for Mass Spectrometry 17*, 1146–1157.

35. Hsu, F.-F., Turk, J., Rhoades, E. R., Russell, D. G., Shi, Y., and Groisman, E. A. (2005) Structural characterization of cardiolipin by tandem quadrupole and multiple-stage quadrupole ion-trap mass spectrometry with electrospray ionization. *Journal of the American Society for Mass Spectrometry 16*, 491–504.

36. Mosteller, F., and Fisher, R. A. (1948) Questions and Answers. The American Statistician 2, 30–31.

37. Choi, H., Ghosh, D., and Nesvizhskii, A. I. (2008) Statistical Validation of Peptide Identifications in Large-Scale Proteomics Using the Target-Decoy Database Search Strategy and Flexible Mixture Modeling. J. Proteome Res. 7, 286–292.

38. S J Sheather, M. C. J. (1991) A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society. Series B. Methodological 53, 683–690.

39. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Anal. Chem. 74*, 5383–5392.

40. Eng, J. K., McCormack, A. L., and Yates III, J. R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry 5*, 976–989.

41. Tabb, D. L., Vega-Montoto, L., Rudnick, P. A., Variyath, A. M., Ham, A.-J. L., Bunk, D. M., Kilpatrick, L. E., Billheimer, D. D., Blackman, R. K., Cardasis, H. L., Carr, S. A., Clauser, K. R., Jaffe, J. D., Kowalski, K. A., Neubert, T. A., Regnier, F. E., Schilling, B., Tegeler, T. J., Wang, M., Wang, P., Whiteaker, J. R., Zimmerman, L. J., Fisher, S. J., Gibson, B. W., Kinsinger, C. R., Mesri, M., Rodriguez, H., Stein, S. E., Tempst, P., Paulovich, A. G., Liebler, D. C., and Spiegelman, C. (2010) Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J. Proteome Res. 9*, 761–776.