2012

Computational Methods for the Differential Profiling of Triacylglycerols Using RP-HPLC/APCI-MS

Margaret Holbrook Broadwater
Medical University of South Carolina

Computational methods for the differential profiling of triacylglycerols using RP-HPLC/APCI-MS

Margaret Holbrook Broadwater

A dissertation submitted to the faculty of the Medical University of South Carolina in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the College of Graduate Studies.

Department of Biochemistry and Molecular Biology

2012

Approved by:

John H. Schwacke, PhD
Chair, Advisory Committee

W. Jim Zheng, PhD

Geoffrey I. Scott, PhD

Elizabeth G. Hill, PhD

L. Ashley Cowart, PhD

Acknowledgements

My advisor, Dr. John Schwacke, guided my way as a student and encouraged every step of this research, pushing me beyond what I thought I could do. My committee members, Dr. Geoff Scott, Dr. Jim Zheng, Dr. Ashley Cowart, and Dr. Elizabeth Hill, each played a role in seeing this project through. Dr. Cowart's research provided new goals and challenges to address through this project. Dodie Weise and June Watson made life much easier by helping with administrative tasks and always having answers related to the graduate school paper trail. Jan Carson always found time for me on Dr. Scott's calendar, and offered words of encouragement along the way.

Gloria Seaborn has been a mentor in lipid chemistry and a true friend, and Joe Wade made days spent in the lab fun and interesting. Kathy Moore, Trey Knott and Ron Lundstrom have been very supportive of the forensics part of this research; Dr. John Ramsdell has provided support and encouragement, and has patiently allowed me to finish this project as I transition into work in a new field at NOAA. Additionally, this work would not have been possible without the contributions of Dr. Craig Byrdwell, who pioneered analysis of TAGs using APCI-MS; the R Development Core, who have made it possible to do absolutely anything related to data analysis at no cost; Dr. Colin Smith and colleagues, who developed the xcms data processing package; and Dr. Josef Cvacka and colleagues, whose work was the basis of the TAG structure identification algorithm that I implemented in R.

I would also like to thank my mother, Dr. Carolyn Jenkins, and my wonderful husband Ryan and son Thomas, and so many other family and friends for their love, understanding, support and encouragement during this time, and for providing many much-needed distractions along the way.

Table of Contents

List of Tables

List of Figures

Key to Symbols and Abbreviations

APCI-MS – Atmospheric pressure chemical ionization mass spectrometry
AT – Adipose tissue
ATGL – Adipose triglyceride lipase, also called desnutrin
BHT – Butylated hydroxytoluene
CoA – Coenzyme A
CD – Low-fat control diet
CVD – Cardiovascular disease
DAG – Diacylglycerol, also called diglyceride
DIO – Diet-induced obesity
ECN – Equivalent carbon number
FA – Fatty acid, fatty acyl
FAME – Fatty acid methyl ester
FID – Flame ionization detection
GC – Gas chromatography
HSL – Hormone sensitive lipase
LCFA – Long chain fatty acid (12 or more acyl carbons)
LC/MS – Liquid chromatography/Mass spectrometry
LCT – Long chain triacylglycerol
LD – Lard-based high-fat diet
LPL – Lipoprotein lipase
MAG – Monoacylglycerol, also called monoglyceride
MCFA – Medium chain fatty acid (fewer than 12 acyl carbons)
MCT – Medium chain triacylglycerol
MD – Milkfat-based high-fat diet
MetS – Metabolic syndrome
MGL – Monoglyceride lipase
MUFA – Monounsaturated fatty acid
NEFA – Non-esterified fatty acid
PL – Pancreatic lipase
PUFA – Polyunsaturated fatty acid
RP-HPLC – Reversed-phase high performance liquid chromatography
SEM – Standard error of the mean
SFA – Saturated fatty acids
TAG – Triacylglycerol, also called triglyceride
TFA – *trans* fatty acid
UFA – Unsaturated fatty acid

MARGARET HOLBROOK BROADWATER. Computational methods for the differential profiling of triacylglycerols using RP-HPLC/APCI-MS. (Under the direction of JOHN H. SCHWACKE).

Reversed phase liquid chromatography with atmospheric pressure chemical ionization mass spectrometry (RP-HPLC/APCI-MS) was employed for the analysis of natural mixtures of triacylglycerols. An integrated framework for data analysis, including preprocessing, statistical analysis and automated structure identification, was implemented in the R statistical program. Raw data stored as mzXML, mzData, or mzML files are preprocessed using a series of steps for peak detection, chromatographic alignment, and normalization. Targeted and non-targeted feature selection steps are employed to filter the data for features that are relevant and informative for a particular biological question. Triacylglycerol structures are identified by evaluating relationships between the diacylglycerol fragment ions and protonated molecules observed in APCI mass spectra, and suggested structures are evaluated using a correlation-based score that reflects whether structure-associated ions are concurrently eluting over the retention-time course of the analysis. The algorithm was tested using five soybean oils and triacylglycerol structure identifications were verified from literature references. We employed the developed methodology for classification of plant oils and marine oils to their biological source, and also to determine structural differences in triacylglycerols in adipose tissue from mice fed different high-fat diets in studies of diet-induced obesity.

INTRODUCTION, BACKGROUND, AND DESIGN

Reversed-phase high performance liquid chromatography coupled to atmospheric-pressure chemical ionization mass spectrometry (RP-HPLC/APCI-MS) has shown promise in the analysis of triacylglycerols (TAGs) in recent years and has been used extensively in the analysis of plant oil TAGs (1-4). Fats and oils from animal sources are more chemically complex than plant oils by nature, as they incorporate dietary fatty acids (FAs) into storage fat TAGs in addition to FAs resulting from biosynthesis via cellular metabolic pathways. This complexity creates a challenge in TAG LC/MS data analysis, as it is not possible to separate individual TAG components in complex mixtures using current chromatographic capabilities (5). The fields of proteomics and metabolomics have addressed similar issues, and a variety of tools useful to process data resulting from the analysis of complex mixtures of molecules have been developed in recent years. Programs such as SpecArray (6), msInspect (7), MZmine (8), xcms (9), and OpenMS (10) are freely available and facilitate the processing of proteomic and metabolomic LC/MS data, but have been not yet been applied specifically to the analysis of complex mixtures of TAGs, such as those found in natural fats and oils. The present study aims to address some of the challenges associated with TAG LC/MS analysis using cross-disciplinary methods, and to develop a high-throughput data processing and analysis pipeline for TAG RP-HPLC/APCI-MS data that can be used to address biological questions related to the TAG composition of storage fats in plants and animals.

This project comprises three independent studies using RP-HPLC/APCI-MS analysis of TAGs. We began with the analysis of plant oils for the purpose of classifying oils to their biological source. We used these data primarily for method validation, as similar data were previously reported to achieve acceptable classification results (4). Two freely available metabolomics tools, MZmine and xcms, were compared with a previously published manual processing methodology (1, 4). We found the xcms results were more highly correlated with the manually processed data, and achieved classification accuracy similar to that of the manually processed data. We then sought to combine xcms data processing with targeted and non-targeted feature selection approaches to address current biological issues. We applied xcms processing of marine oil RP-HPLC/APCI-MS data in combination with two feature selection steps to confirm FA profile identifications of seal oil dietary supplements in forensic analyses. Lastly, we applied xcms processing of RP-HPLC/APCI-MS of mouse adipose tissue samples in combination with two feature selection steps to examine specific differences in the adipose tissue TAG composition in mice fed milkfat- and lard-based high-fat diets to induce obesity. These three studies collectively address the following specific aims.

Specific Aim 1: *To evaluate current methodology and the application of existing computational tools (MZmine and xcms) for processing TAG RP-HPLC/APCI-MS data.*

Computational tools may be used to facilitate high-throughput processing of plant oil TAG RP-HPLC/APCI-MS data and achieve results comparable to conventional data-processing methodologies in terms of identifying and measuring features within an LC/MS image. We examined correlations between plant oil RP-HPLC/APCI-MS data that were manually processed to determine the relative amounts of 12 known analytes (4) and the same analytes measured using MZmine (8) and xcms (9), and compared classification accuracies of the resulting data sets using a random forest model to classify the plant oils to their biological source.

Specific aim 2: *To select relevant and informative features from preprocessed TAG RP-HPLC/APCI-MS data for classification.*

The xcms program and similar tools provide a comprehensive list of all peaks detected in a set of samples, defined by retention time and mass-to-charge ratio (RT, m/z) coordinates; many of these peaks may not be relevant to a specific biological question. We implemented a combined approach of targeted and non-targeted feature selection to reduce the size of the peak list generated using xcms. First, FA compositional data were used to predict TAG species that may be present and to generate a list of ions (m/z values) representative of such species. We then searched the xcms peak list specifically for intensities at these values. This targeted peak selection strategy narrowed our peak list to features that represent relevant analytes. We then aimed to select a list of informative TAG features that may be used to answer specific biological questions, and addressed examples for classification and difference detection in two applications.

Specific aim 3: *To apply high-throughput RP-HPLC/APCI-MS analysis of TAGs using xcms preprocessing combined with feature selection to the forensic identification of seal oils.*

Seal oil dietary supplements, commonly used to increase consumption of omega-3 polyunsaturated fatty acids (PUFAs), are legal in Canada, but prohibited in the U.S. and E.U. FA profiles have been used to distinguish between marine oils harvested from fish and seals for forensic purposes, but FA composition may not be sufficient to determine that such oils are non-synthetic. TAG composition of marine oils is considerably more complex than FA composition, and current technology does not permit separation and identification of TAG molecular species in these samples. A metabolomics approach combining xcms preprocessing of RP-HPLC/APCI-MS data with feature selection was used to obtain a list of features representing TAG molecular species for use as predictor variables to classify samples to fish versus seal sources. A random forest classifier successfully confirmed classification results using FA profile data from the same set of samples. This additional analysis provides a two-tiered approach to identifying seal oils for forensic purposes.

Specific aim 4: *To apply an integrated framework for processing high-throughput RP-HPLC/APCI-MS analysis of TAGs using xcms preprocessing combined with feature selection and TAG structure identification to examine differences in adipose tissue TAG composition in mice fed milkfat- and lard-based high-fat diets.*

Rodent diet-induced obesity models are commonly used to study human disease. Dietary FA composition affects the relative availability of FAs to tissues, and individual FAs have diverse effects on health. A novel milkfat-based high-fat diet (MD) resulted in more obese and more insulin-resistant mice, compared with a traditional lard-based high-fat diet (LD) and isocaloric low-fat control diet (11). TAG RP-HPLC/APCI-MS data were obtained from adipose tissue sampled from MD- and LD-fed mice at 8 and 16 weeks. A metabolomics approach combining xcms preprocessing of RP-HPLC/APCI-MS data with feature selection was used to obtain a list of features representing TAG molecular species that differed in adipose tissue sampled from MD- and LD-fed mice at the two time points. We determined the structures of TAG species that differed between diet groups using an algorithm that exploits the relationships between ions in TAG mass spectra and correlations among groups of ions over retention time windows associated with individual features.

Background

Fats and oils are of great economic importance in agriculture, international commerce, and as ingredients in foods. Fats comprise 40% of dietary energy intake in Western Europe and North America. Composition of dietary fats is vital to good nutrition and contributes to the palatability, taste, and structure of foods (5). Oils are simply fats that are found in the liquid phase at room temperature; fats and oils are similar in chemical structure and together form the biochemicals known as lipids that are found in all living organisms. The term lipid includes FAs and their derivatives, and substances related biosynthetically or functionally to these compounds. FAs are compounds that are naturally synthesized via condensation of malonyl coenzyme A units by a fatty acid synthase enzyme complex (5). FAs found in plant and animal lipids are typically 14 to 22 carbons in length and are saturated or contain one to six *cis* double bonds separated by methylene (-CH_2-) groups. Naturally occurring short-chain FAs, branched-chain FAs, hydroxy FAs, and FAs containing other functional groups including *trans* double bonds are found less frequently and in lesser abundance. FAs are identified by systematic or trivial names, and using the notation A:Bn-C, where A is the number of carbon atoms, B is the number of double bonds, and C is the position of the first double bond from the terminal methyl group. For example, 18:2n-6 is an 18-carbon FA with two methylene-interrupted *cis* double bonds, the first at the 6th carbon from the terminal methyl group; 18:2n-6 has the trivial name linoleic acid. The structures of several common polyunsaturated fatty acids (PUFAs) are shown in Figure 1. Glycerolipids are compounds consisting of one or more FAs esterified to glycerol, and include the glycerophospholipids, glyceroglycolipids, and mono-, di-, and triacylglycerols. Most commercially important fats and oils consist primarily of TAGs, and TAGs in plant and animal storage fats are the main focus of this research.

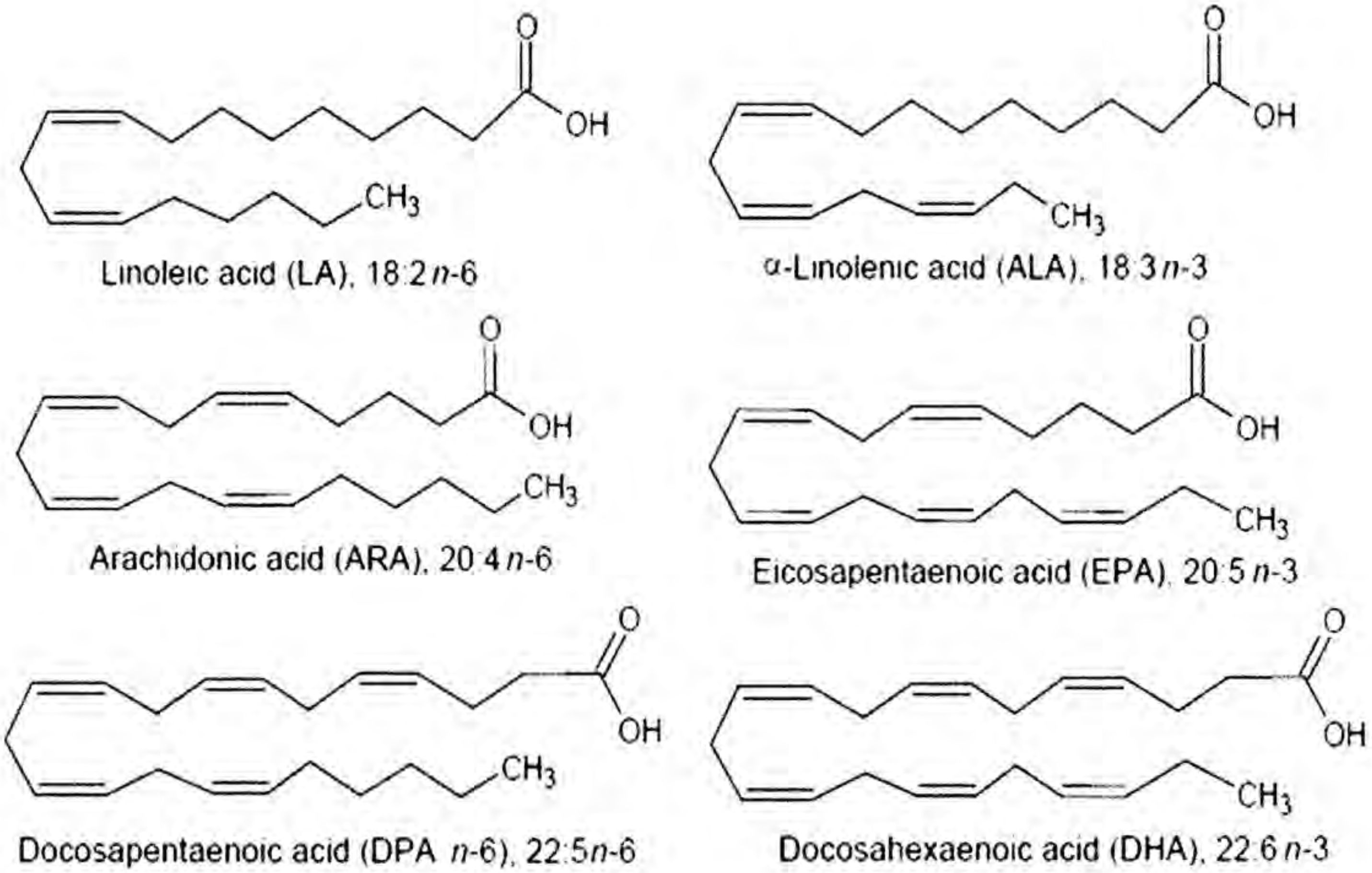

O
OH
CH3
Linoleic acid (LA), 18:2 n-6
O
OH
CH3
α-Linolenic acid (ALA), 18:3 n-3
O
OH
CH3
Arachidonic acid (ARA), 20:4 n-6
O
OH
CH3
Eicosapentaenoic acid (EPA), 20:5 n-3
O
OH
CH3
Docosapentaenoic acid (DPA n-6), 22:5n-6
O
OH
CH3
Docosahexaenoic acid (DHA), 22:6 n-3

Figure 1. Common n-6 and n-3 polyunsaturated fatty acids.

TAGs are the most abundant class of lipid molecules (5). They serve primarily as an energy source, but also are important for insulation and protection (12). At 37.6 kJ (8.98 kcal) per gram, TAGs are the most concentrated form of biological energy. In animals, storage fat TAGs are typically located in well-defined tissues, but also occur as droplets within cells (13). Adipose tissue, in which TAGs are stored within specialized cells called adipocytes, is unique to vertebrates and constitutes the primary energy reserve in mammals, birds, reptiles, and amphibians (14). Skeletal muscle and liver are major lipid storage sites in fish, and are highly variable among species. Lean fish, such as cod, store large amounts of lipid in the liver, while fatty species such as herring and anchovies deposit TAG in the skeletal muscle (15). Oils extracted from select fish and seal species are valued for their high proportion of n-3 long-chain PUFAs, specifically eicosapentaenoic acid (20:5n-3 or EPA) and docosahexaenoic acid (22:6n-3 or DHA), which may provide significant health benefits in humans (16-18). Marine mammals have a specialized tissue called blubber which lies under the skin and consists primarily of fibrous proteins and adipocytes. Blubber is extremely important both physiologically and metabolically, as most marine mammals have very few internal adipose depots (14, 19). In plants, TAGs are stored within lipid bodies in fruits and seeds and used as energy reserves for the next generation. Because plants are not mobile and can use photosynthesis to fix carbon, their energy storage requirements are substantially less than those for animals. The relative proportion of TAGs in seeds from different plant species varies widely, ranging from 1-2% in grasses to 60% dry weight in the castor seed. Greater than 75% of commercial oils are derived from plant oils, with two-thirds of this used for food purposes.

TAGs are synthesized by enzyme systems in living organisms as L-glycerol derivatives, with a center of asymmetry around carbon-2 of glycerol. A stereospecific numbering (*sn*) system is used to describe the stereochemistry of TAGs and other glycerolipids. The structure of 1-palmitoyl-2-docosahexaenoyl-3-oleoyl-*sn*-glycerol is shown in Figure 2. The three FAs in a TAG molecule may vary in carbon chain length; number, position, or configuration of double bonds; and may have branched chains or hydroxyl or other functional groups. Hundreds of different FAs occur naturally in plant and animal TAGs, and FA composition is often distinctive for different species. The FA composition of storage fats also depends on an animal's diet and can thus distinguish among members of the same species who consume different diets (5). FA composition is typically determined by analyzing derivatives of FAs using gas chromatography with flame ionization detection (GC/FID) and/or mass spectrometry (GC/MS).

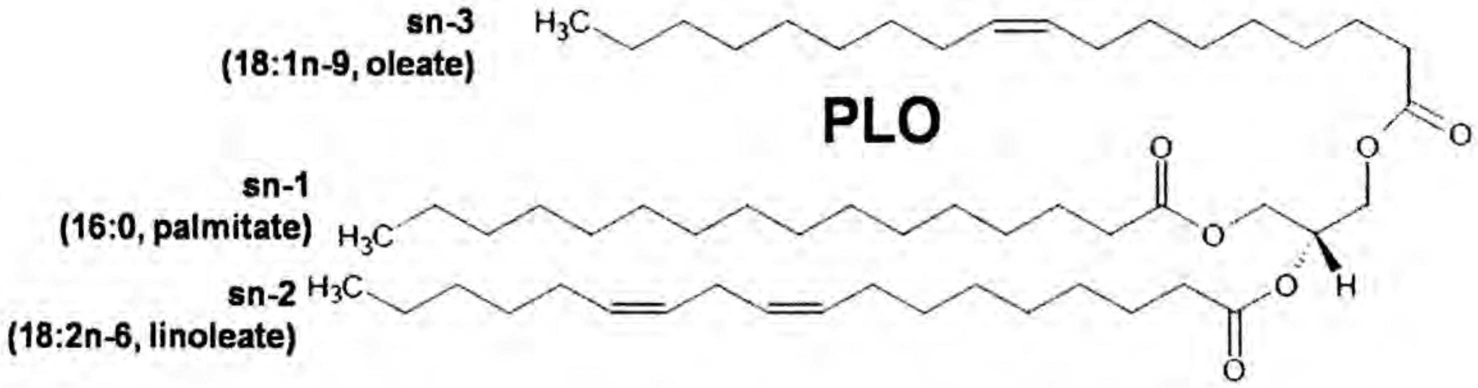


Figure 2. Structure of 1-palmitoyl-2-linoleoyl-3-oleoyl-sn-glycerol (PLO), with acyl chains 16:0 (palmitic acid) at sn-1, 18:2n-6 (DHA) at sn-2, and 18:1n-9 (oleic acid) at sn-3.

Brockerhoff observed that "the positional distribution of fatty acids in triglycerides of animals is nonrandom;" he noted that while the principles regulating this distribution are unknown, general patterns are characteristic of taxonomically related organisms (20). Multiple studies of TAGs from marine animals indicate that DHA is primarily found in the *sn*-2 position in fish and marine invertebrates and in the primary positions (*sn*-1 and *sn*-3) in marine mammals (18, 21). This difference may play a key role in determining the authenticity high *n*-3 PUFA of marine oil supplements, and may provide a means to distinguish between oils from fish and marine mammal sources. TAG composition maybe studied by determining the positional distribution of FAs, i.e. the FA composition at each position on the glycerol backbone, using enzymatic techniques in combination with FA compositional analysis, or by analyzing intact TAGs using LC/MS with the aim of determining molecular species composition, i.e. the TAG profile, or the relative amounts of individual TAG molecules.

Fatty acid profile analysis

Shortly after the advent of gas-liquid chromatography (GLC, 22), the newly developed technique was applied to resolve a mixture of volatile fatty acids ranging in length from 1-12 carbons (23). Lipid chemists have been at the forefront of most of the advances leading to modern GC, which can be used to separate most types of lipid molecules (24). Modern GC instruments use capillary columns to separate individual FAs as their methyl ester derivatives (FAMEs), and components may be identified by retention times in comparison with known standards, and the use of electron-impact MS. Other types of FA derivatives (e.g. butyl or picolinyl esters, 4,4-dimethyloxazoline) may also be used for similar purposes or to determine specific double bond positions in individual FAs. Upon separation, each FAME peak is integrated as a time signal and converted to a weight percent of the total FA composition using a correction factor accounting for differential response; GC/FID is widely believed to provide accurate quantitation when correction factors are properly employed (5, 24). The resulting FA profile is a list of the relative quantities of FAs in a sample, typically expressed as percents. These compositional data must be treated appropriately using statistical methods, as individual FA variables are not independent.

Analysis of triacylglycerols

The general structure of TAG molecules was established during the 19th century, but separation and analysis of TAGs and their component FAs were not possible until chromatographic methods were developed (13). In the first stereospecific analyses of TAGs, enzymatic and chemical digestive techniques were combined with GLC of FA derivatives to quantify the positional distribution of FAs on the glycerol backbone (25, 26). While these methods and their modifications have been widely used, they have been subject to criticism for a variety of reasons. Enzymatic and chemical hydrolysis procedures are time-consuming; and possible complications due to limitations on the selectivity of the lipase, selectivity for chain length and number of double bonds, and acyl migration during analysis have raised doubts about their accuracy. This bottom-up approach to TAG analysis does not provide information about individual TAG molecular species, only the overall distribution of FAs in each position (27).

RP-HPLC of TAGs is a top-down approach, analyzing mixtures of whole molecules. RP-HPLC has proven to be extremely useful in the separation of TAGs, but has not yet succeeded in the resolution of individual TAG molecules. While modern chromatographic and spectrometric instrumentation methods can successfully separate and identify FA derivatives from lipid extracts to determine the FA composition of an oil or fat, characterization of the TAG molecular species is a more difficult process. For n fatty acids, $\frac{n^3+n^2}{2}$ TAG molecular species are possible, not including enantiomers (when enantiomers are considered this number jumps to n^3). The position of fatty acyl moieties on the glycerol backbone of the molecule may yield considerable information about metabolic and nutritional properties of fats and oils, and has been a continuous analytical challenge to lipid chemists (27). TAGs elute on an octadecyl-siloxane (ODS, C18) column approximately in the order of their equivalent carbon number (ECN), defined as the number of fatty acyl carbon atoms minus two times the number of carbon-carbon double bonds. TAGs with the same ECN are termed "critical pairs" and tend to elute close together (5). Recent developments in the field of MS may help to address challenges related to coeluting TAG molecules. APCI-MS used with RP-HPLC allows for partial identification of the FA (number of carbons and double bonds, e.g. 18:3) from individual TAGs in a mixture and provides information on positions of FAs on the TAG molecule (27).

RP-HPLC/APCI-MS was first applied to the analysis of TAGs in 1995 (28). In this study, APCI-MS analysis of a mixture of monoacid TAG standards separated by RP-HPLC revealed minimal fragmentation, resulting in the formation of diacylglycerol ions, $[M-RCO_2]^+$ or $[DAG]^+$, and protonated molecules, $[M+H]^+$, as shown in Figure 3. A later study analyzed ABC-type TAG of known regiospecific compositions and observed that the relative intensities of DAG fragment ions could reveal information on the positions of FA on the TAG molecule (29). Specifically, the least abundant $[DAG]^+$ ion resulted from loss of the FA from the secondary position (*sn*-2). Individual TAG isomers (e.g. LPL and LLP/PLL), which shared the same retention time by RP-HPLC, could be distinguished using the ratio of $[DAG]^+$ ions observed in APCI-MS spectra. This allowed the identification and quantification of mixtures containing only one TAG positional isomer, which is the case in many plant oils. Further research has established that it is possible to construct calibration curves from known TAG standards by which the relative proportion of TAG positional isomers can be calculated. Beef, pork, and chicken fats as well as several plant oils have been characterized in such a manner (30, 31).

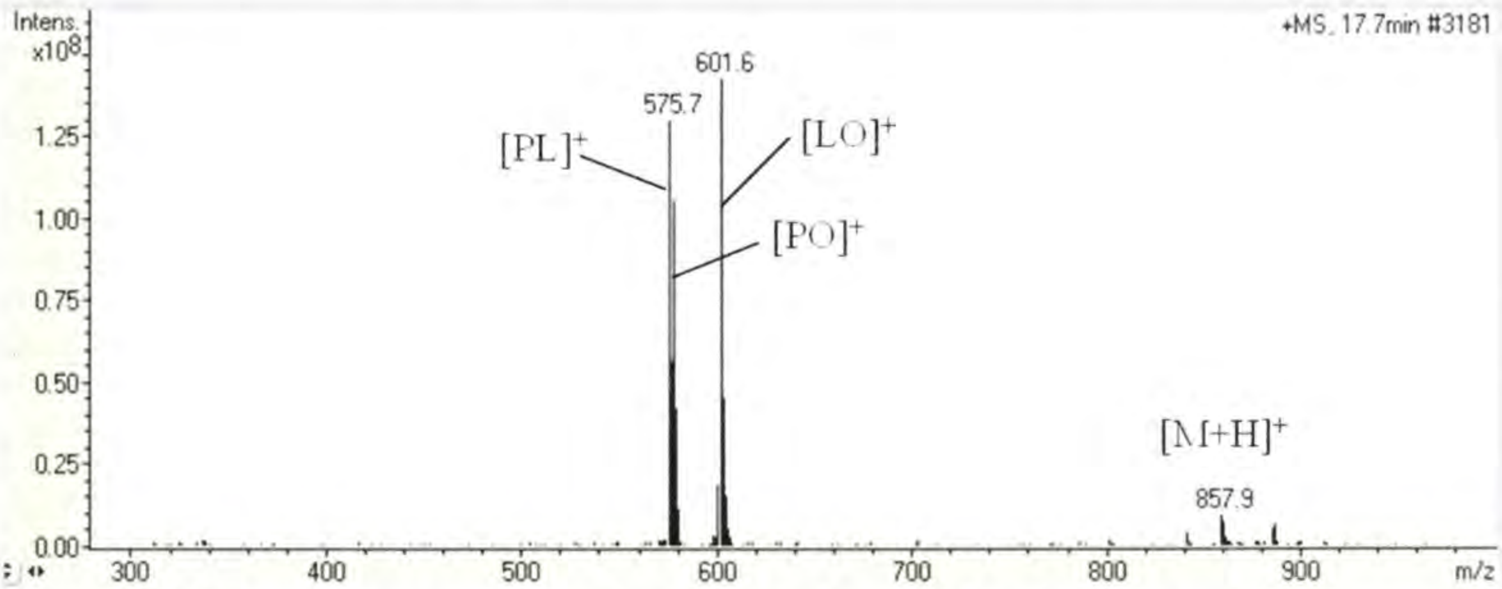


Figure 3. APCI mass spectrum of 1-palmitoyl-2-linoleoyl-3-oleoyl-sn-glycerol (PLO).

The main challenge in analysis of TAGs lies in the lack of ability to resolve individual molecular species using chromatographic separation techniques. Different molecules with the same number of acyl carbons and double bonds may coelute and resolution of TAG molecular species using MS involves calibration procedures requiring standard compounds for accurate quantitation. While it is possible to obtain standards for many of the TAG molecular species observed in plant oils, these standards are expensive. Marine oils are more complex and obtaining standards for each TAG molecule would likely be cost-prohibitive even if such compounds were made available. As chemists, we have been trying to implement TAG profiling experiments in the manner of FA profile analysis, and this may not be possible for complex samples. While our ultimate goal is complete characterization of TAG molecular species, many problems can be addressed without these comprehensive analyses. Jakab and colleagues used relative peak areas of TAG molecular species, calculated from $[DAG]^+$ or $[M+H]^+$ ion intensities from RP-HPLC/APCI-MS analysis, to differentiate plant oils (n = 42) from twelve different biological sources with 97.6% accuracy reported from a linear discriminant model (4). This type of analysis may be very useful for similar classification problems, and could potentially be automated for high-throughput analysis.

A new approach for the RP-HPLC/APCI-MS analysis of TAGs

Bottom-up profiling experiments are common in the fields of proteomics, where complex mixtures of digested peptides isolated from blood or tissue are routinely analyzed using LC/MS. Such mixtures have complexity similar to (or greater than) a mixture of TAG molecules in an extracted oil. A signal-based processing methodology can be used with these data as opposed to the more traditional peak detection, identification, and quantification procedure discussed above. Data are treated as a two-dimensional signal matrix or image, as shown in Figure 4, and established methods in signal processing, statistics, and machine learning are used to find patterns that are characteristic of a particular sample or class of samples (32). Data must be treated in a manner to assure consistency across experiments, and a great deal of effort has been put toward this goal in recent years in proteomics and metabolomics research, and the underlying and supporting discipline of bioinformatics (32-36). Software programs designed for proteomics and/or metabolomics experiments, including the publicly available programs msInspect (7), MZmine (8), SpecArray (6), and xcms (9), have been devised to combine preprocessing and differential analysis of proteomic and metabolomic data in a suite of algorithms. MZmine and xcms, both designed specifically for data resulting from the analysis of metabolites such as lipids using LC/MS or GC/MS, provide a framework from which we can devise a high-throughput methodology for the RP-HPLC/APCI-MS analysis of TAGs, taking the analysis from the raw data stage through classification or difference detection and ultimately to determine the FAs present in individual TAG species, depending on the question at hand. Additionally, such programs may be useful for detecting specific differences in TAG profiles between sample classes that may be informative from a clinical perspective.

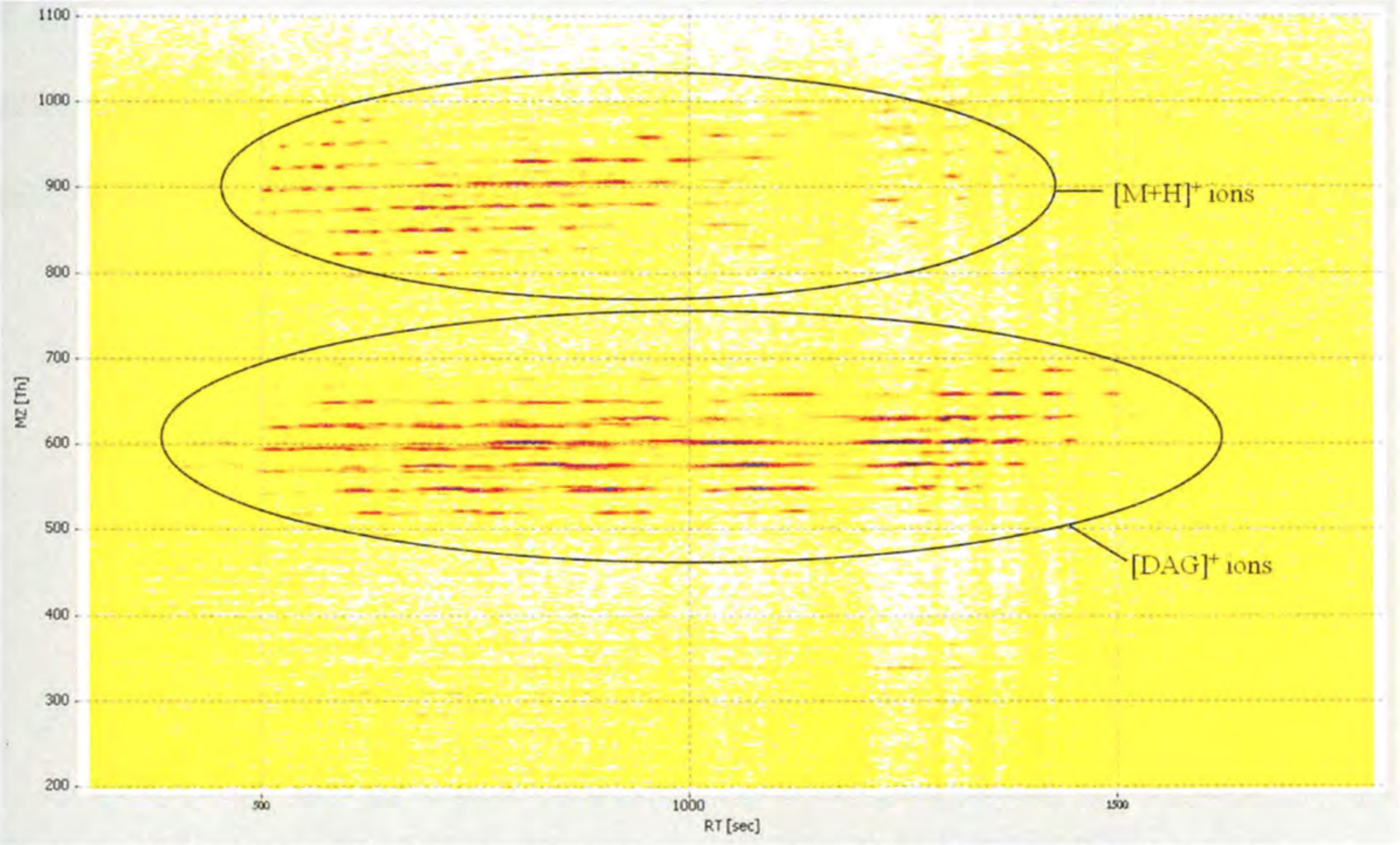


Figure 4. RP-HPLC/APCI-MS of seal oil triacylglycerols. The data are presented as an image of intensity values, indexed by retention time and mass-to-charge ratio.

Data mining is an iterative process, as shown in Figure 5. Data analysis and acquisition, preparation, feature selection, model development, model assessment, and generalization steps all play an important role in the process, and errors can occur at any of these levels (37, 38). Listgarten and Emili divided LC/MS data processing into low-, mid-, and high-level stages (32). Low- and mid-level steps include preprocessing, or assimilating the data into an accessible format, filtering, baseline subtraction, normalization, alignment in time and peak detection and quantification; the goal of these steps is to format the data such that different profiles, or LC/MS experiments, can be compared for classification or difference detection purposes. We want to minimize random and systematic differences between experiments such as RT shifting and changes in intensity measurements. High-level processing encompasses feature selection, difference detection, and classification. Goals of high-level data processing for TAG analysis, in order of complexity, are:

1. Classification of samples, e.g. classifying an oil sample to its biological source.
2. Low-level biomarker discovery, where particular regions of the data matrix are selected for their ability to discriminate among classes.
3. High-level biomarker discovery, where specific TAG peaks corresponding to discriminating regions are identified.
4. Complete identification and quantitative resolution of the full set of TAG molecular species present in a particular oil.

The current research will address the first three goals, with the aim of contributing to the advancement of the fourth. All four may be possible for well-studied groups of samples such as plant oils, while the complexity of animal lipids may be prohibitive to this aim given current technologies. It is important to note that classification is often a simple problem when compared with the challenges of difference detection and biomarker discovery (39). The following processing steps are incorporated to achieve a high-throughput, automated analysis that takes a set of RP-HPLC/APCI-MS experiments from raw data to a final informative product, which may be the output from a classification model or difference detection between sample groups resulting in a list of features, and ideally the related TAG structures, that differ between sample groups.

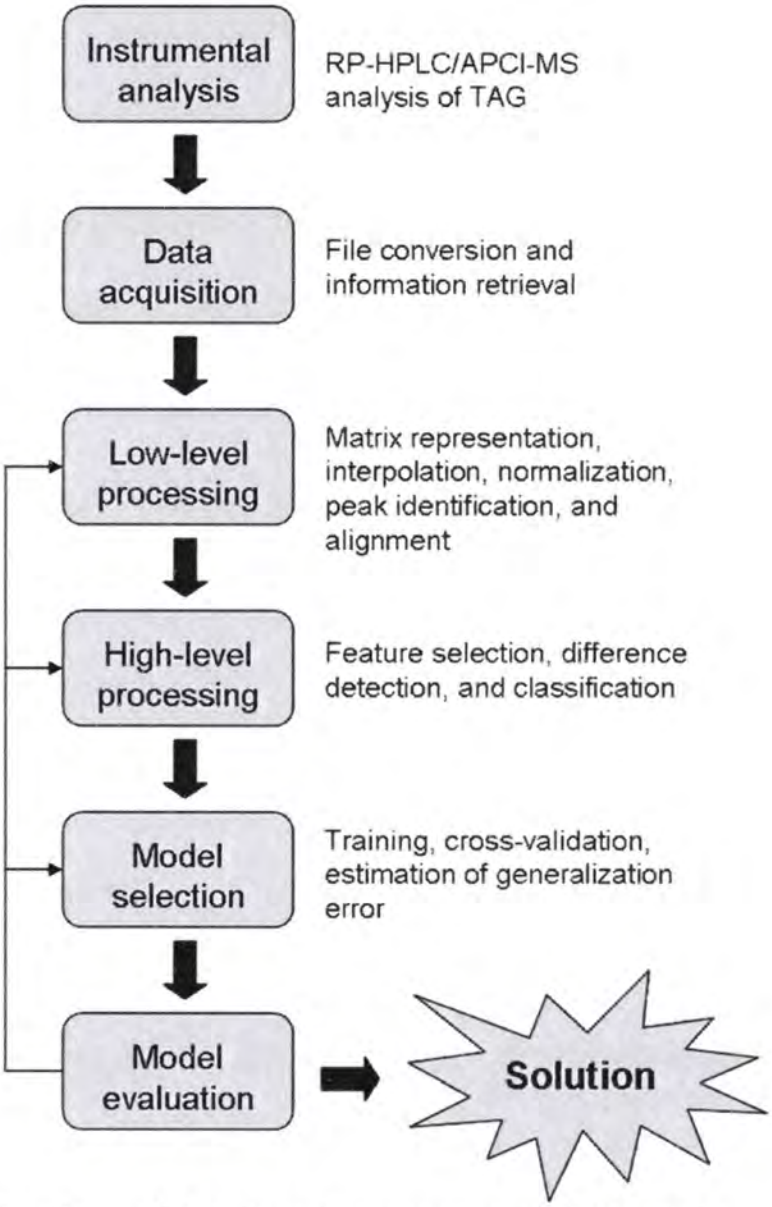

Instrumental analysis
RP-HPLC/APCI-MS analysis of TAG
Data acquisition
File conversion and information retrieval
Low-level processing
Matrix representation, interpolation, normalization, peak identification, and alignment
High-level processing
Feature selection, difference detection, and classification
Model selection
Training, cross-validation, estimation of generalization error
Model evaluation
Solution

Figure 5. A schematic view of the iterative process of data mining, adapted from Polikar (37).

Preprocessing

The first step in data processing is to convert the data from the typical LC/MS vendor-specific proprietary format into a more accessible format such as mzXML (40), mzData (http://www.psidev.info/index.php?q=node/80#mzdata), or the more recently developed mzML (http://www.psidev.info/index.php?q=node/257). These formats use an XML schema to represent raw instrument data and can be read using metabolomic software tools such as MZmine (8), xcms (9), and OpenMS (10), where data can be accessed and manipulated accordingly. Often this involves interpolation along the RT and mass spectral dimensions so that the images are the same dimensional size and range. This step is specific for the instrument used, as conversion programs differ for different proprietary raw data formats. A local smoothing in time and m/z may be used to overcome any issues that arise due to interpolation (39). Additionally, data images may be cropped in RT and/or m/z dimensions to remove regions that do not contain viable information, e.g. the beginning and end of an analysis or areas outside of the range of interest in the m/z dimension.

Low-level processing

Peak detection aims to find informative regions in the LC/MS data image (37). These regions indicate chromatographic elution of one or more compounds of interest. The goal is to distinguish signal from noise. It is important to note that this is different from feature or variable selection, where we select a subset of features from those identified here. Local maximum methods search the data for local intensity maxima, while recursive threshold methods require a width parameter to differentiate actual peaks from noise spikes in the data; both of these are available in MZmine (41). Wavelet transform methods use time-frequency analysis to find changes in signal frequency that are indicative of peaks (42). Methods such as the translation-invariant wavelet transform (TIWT) can be used to extract features from mass spectral data even in situations with experimental variability in background noise and measurement intensities, and before smoothing, estimating signal-to-noise ratio, or modeling a baseline (43). The xcms package uses matched filtering with background suppression (44) to detect peaks in extracted ion base-peak chromatograms using a second-derivative Gaussian function (9). While xcms detects peaks in the RT dimension of the LC/MS image, MZmine finds peaks in individual spectra (the m/z dimension) and connects peaks from successive spectra when they form good continuity (41).

Alignment in chromatographic time has been cited as a major obstacle to reliable detection of differences among sample groups using LC/MS data and is an important step in preprocessing (39). Shifting of retention times can occur for many reasons and must be corrected before comparing data from different LC/MS experiments. The goal of alignment is to match corresponding features from different experiments to minimize RT variation and experimental noise. Time warping methods choose one experiment as a template and warp the time coordinates of each of the other experiments to maximize similarity between the two images. The theory behind dynamic time warping algorithms is similar—each point in RT space can be moved. The alignment problem is more complicated when mass spectral space is considered. TAG with different m/z values may have differential RT shifting in different experiments, e.g. two TAG that coelute in one profile may be separate peaks in another profile. Thus, aligning based only on RT coordinates may not be sufficient to align individual TAG molecular species. Piecewise methods divide the m/z domain into bins and fit piecewise linear time warping functions specific to each bin. These methods rely on the (RT, m/z) intensity values of detected features. Alignment methods that rely on detected peaks as opposed to raw spectral information may be affected by errors in the peak detection step. Another issue to consider is the use of a template profile, as mentioned above, to which all other experiments are aligned. This approach is prone to error when samples from different classes with different features must be aligned. MZmine and xcms use detected peaks to align samples simultaneously. MZmine employs a master list for all peaks detected in all samples, considered one at a time, and matches peaks from each new sample to the master list based on a scoring function that compares isotope patterns. In xcms, peaks are first matched using fixed-interval overlapping bins; the algorithm determines the overall distribution of peaks in chromatographic time and then dynamically identifies boundaries of regions where many peaks have similar retention times. "Well-behaved" peak groups, in which very few samples have no peaks or more than one peak assigned, are used to create a nonlinear (loess) RT deviation contour for each sample. The resulting deviation profiles are used to correct the RTs of the original peak lists, and the corrected lists must be matched into groups again.

High-level processing

The goal of feature selection is to select a subset of relevant peaks that will best discriminate among samples or sample classes. It is important to note that this differs from feature extraction steps such as peak detection, described above. Feature extraction methods search for informative regions within a signal while feature selection methods find variables which discriminate among signals that represent different sample classes. Feature selection can be performed on independent variables using filter methods such as receiver operating curves (ROC), statistical tests, wavelet transforms, or information gain criteria (37). Filters such as wavelet transforms are useful in both feature extraction and selection steps. Grouped feature selection such as stepwise methods, genetic algorithms, correlation-based methods, and principal components analysis (PCA) consider all features simultaneously, though not in the manner of an exhaustive search of all possible subsets of variables. Grouped methods account for correlations among variables, where some variables are better able to discriminate classes when considered together than independently. Wrapper methods perform feature selection in conjunction with classification algorithms such that classifier performance is used as a measure of variable strength for inclusion/exclusion (38). The variable subset is "wrapped around" the classifier, and in this sense feature selection is optimized with regard to a specific classifier. It is important to note that this may result in model overfitting, where generalization error is underestimated and the model does not perform well on new data.

Difference detection procedures are common when data are high-dimensional, as in the case of LC/MS and in microarray analysis. Such procedures aim to identify biomarkers, or individual features that differ most among sample classes. Classical statistics (e.g. t, F, or χ^2) and statistical and permutation test p-values can be used to rate features according to their ability to distinguish among sample groups, though data do not always meet the assumptions associated with these tests (32). Each feature is evaluated independently and thus correlations among features are not considered. To identify as many features as possible while incurring the lowest proportion of false positives, Storey and Tibshirani devised the q-value, which provides a measure of each feature's significance while accounting for the fact that many variables are being tested simultaneously (45, 46). The q-value is based on the false discovery rate (FDR, 47) in contrast to the traditional p-value's false positive rate (FPR). The FDR is the rate that features deemed significant are truly null, while the FPR is the rate that null features are deemed significant. The FDR is a measure of how many selected variables, or "hits" are likely false.

Classification algorithms such as Breiman's random forest (RF) algorithm (48) can be used with independent or grouped variable filters, or can be implemented as wrapper methods. The general idea behind the RF algorithm is to combine random feature selection and bagging (bootstrap aggregation) to improve sample classification. In theory, a RF consists of many decision trees built on independent and identically distributed random samples; the classifier output is the most popular class, or the mode of individual tree outputs. The generalization error converges to a limit as the number of trees becomes large, but convergence depends on the strength of individual trees and the degree of correlation among them. Random feature selection is used to split each node, yielding error rates that compare favorably to boosting and are more robust with respect to noise. Individual decision trees differ from each other due to random selection of features at nodes, reducing correlation to prevent overfitting. So-called "out-of-bag data" is used to estimate the error rate for each tree and for the full forest, providing an unbiased estimate of generalization error and eliminating the need for cross-validation (48-50). The random forest algorithm implements the following steps:

1. Sample with replacement N bootstrap samples, (B_1,...,B_N).
2. For each sample (k = 1:N), construct a decision tree (T_k) without pruning, with the following modification: M randomly selected variables are used to select the best split for the decision at each node. Use T_k to predict the out-of-bag samples (samples not included in B_k).
3. Data are predicted by aggregating the predictions of the N decision trees, i.e. the mode for classification. An estimate of the generalization error (the out-of-bag error rate) is obtained by aggregating the out-of-bag predictions in step 2.

Additionally, the out-of-bag samples are used to assess variable importance: for m = 1:M (randomly selected variables at each node), we randomly permute values of the mth variable and run the out-of-bag data down T_k, saving outputs. Intuitively, the prediction error will increase proportionally to the importance of the variable. Percent increase in misclassification rate with respect to out-of-bag rate (all variables intact) reflects the importance of the variable (48, 51). RF has been shown to be a highly accurate and stable classifier that outperforms other classifiers when used as a wrapper method and performs similarly when used with independent variable selection in classifying samples using MS data (49).

Research design and methods

Three independent studies were designed to achieve the aims of this project. The first of these applied RP-HPLC/APCI-MS analysis of TAGs in plant oils to classify oils to their biological source, with the primary goal of validating the use of a high-throughput processing method for the analysis of targeted TAG analytes (Aim 1). Results from MZmine and xcms programs were compared with results achieved when data were processed manually by a trained analyst using a published method that had successfully classified the plant oils by type based on the relative intensities of twelve TAG features (4). The second study implemented a high-throughput approach for processing marine oil TAG RP-HPLC/APCI-MS data with xcms, combined with two feature selection steps (Aim 2), to classify marine oil dietary supplements to their biological source (fish vs. seal; Aim 3). These data were successfully used to verify classification results from fatty acid profile data for the same samples. The final study applied an integrated framework for the analysis of TAGs using RP-HPLC/APCI-MS that combined xcms preprocessing with statistical analysis and automated structure determination to study TAGs in mouse adipose tissue from mice fed milkfat- and lard-based high-fat diets to induce obesity (Aim 4). The RP-HPLC/APCI-MS data processing workflow for these experiments can be divided into a series of discrete steps, as illustrated in Figure 6. Details for individual experiments are provided in the following chapters.

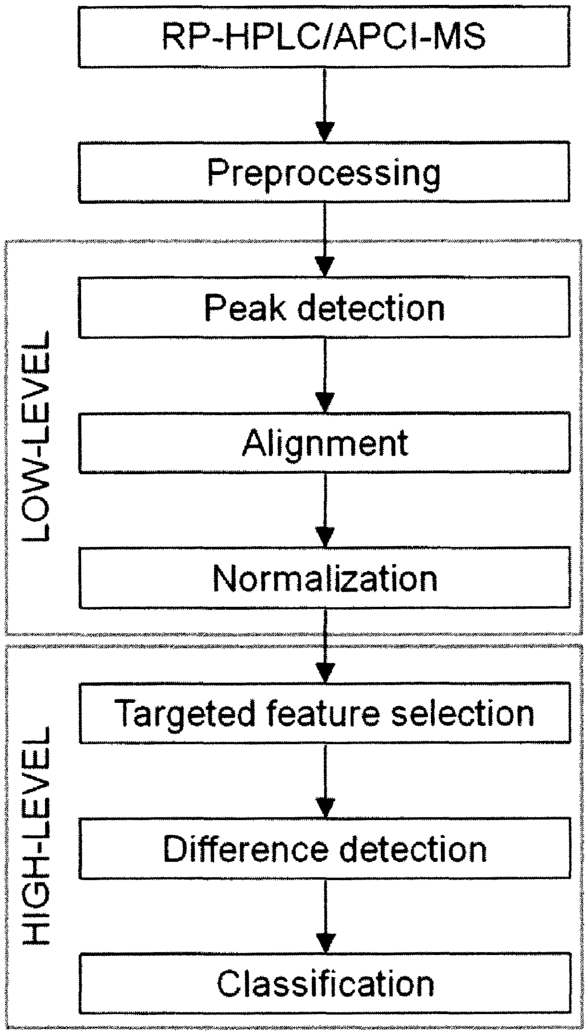


Figure 6. Schematic diagram of data processing workflow, adapted from Katajamaa and Oresic (34).

Rationale and Innovation

The goal of this research is to address the current challenges associated with TAG analysis by applying techniques used in proteomic and metabolomic analyses to solve biological problems. A high-throughput method for processing TAG RP-HPLC/APCI-MS data was applied to three independent studies. Raw data, in the form of preprocessed mzData files, were taken through a series of low- and high-level processing steps in the R computing environment, using xcms combined with normalization, targeted and non-targeted feature selection, difference detection, and (optionally) classification. An integrated framework was developed for the analysis of TAG RP-HPLC/APCI-MS data from complex lipid samples, beginning with raw data stored in mzXML, mzData, or mzML formats and performing a series of steps that ultimately lead to classification and/or difference detection with automated identification of TAG structures for selected features (Appendix). The research described here provides a means for high-throughput analysis of complex TAG samples and will bring lipid chemists closer to achieving the ultimate goal of complete identification and quantitative resolution of the full set of TAG molecular species present in fats and oils.

PAPER 1: COMPUTATIONAL METHODS FOR THE DIFFERENTIAL PROFILING OF TRIACYLGLYCEROLS IN PLANT OILS USING RP-HPLC/APCI-MS

Margaret H. Broadwater[1,2] and John H. Schwacke[2]

[1]Center for Coastal Environmental Health and Biomolecular Research (CCEHBR), National Ocean Service (NOS), National Oceanic and Atmospheric Administration (NOAA), 219 Fort Johnson Road, Charleston, SC 29412.

[2]Department of Biochemistry and Molecular Biology, Medical University of South Carolina, 173 Ashley Avenue, Charleston, SC 29425.

ABSTRACT

A signal-based approach to data analysis, commonly used in proteomics and metabolomics experiments, was applied for high-throughput processing of plant oil TAG profile data obtained using RP-HPLC/APCI-MS. Relative peak areas for twelve targeted TAG features were obtained manually and selected from data processed using the freely available computational tools, MZmine and xcms. Linear discriminant analysis and a random forest classifier were used to classify the plant oils (n = 30) to six different biological sources, and success was measured with classification accuracy. Manual processing, MZmine, and xcms resulted in a random forest model that classified 97%, 87%, and 93% of samples correctly, respectively, based on out-of-bag error rates using the same twelve analytes. TAG structures associated with the targeted features were examined using a tool for automated structure assignment based on mass spectral data obtained via xcms. Most of the TAG structures identified from the targeted features for the different oil types were consistent with previously assigned structures; however, TAG assignments for three targeted features differed between the linseed oils and other groups. Automated data processing with xcms provides a viable high-throughput alternative to traditional time-intensive manual processing of RP-HPLC/APCI-MS triacylglycerol data, and plant oil TAG structures may be determined from xcms output using a new tool for TAG structure assignment.

INTRODUCTION

Plant oils are of great economic importance in agriculture, international commerce, and as cooking materials and ingredients in foods. Most commercially available plant oils are composed of mixtures of triacylglycerols (TAGs), and a great deal of effort has been put forth in the past decade to characterize the TAGs present in these oils using various methods including reversed-phase high performance liquid chromatography with atmospheric pressure chemical ionization mass spectrometry (RP-HPLC/APCI-MS) (1, 3, 4, 52-55). Such characterization is important from a nutritional standpoint, and in establishing the authenticity of commercial oils (17).

TAGs are the most abundant class of lipid molecules and are comprised of L-glycerol esterified to three fatty acyl (FA) groups of varying carbon chain length and degree of unsaturation (5). The three sites of esterification are stereospecifically numbered *sn*-1, -2, and -3. Short-hand notation for TAGs uses the initials or identifiers of the fatty acid trivial names in order of their positions on the glycerol molecule (52).

For example, the notation for 1-palmitoyl-2-linoleoyl-3-oleoyl-*sn*-glycerol is PLO; this molecule is shown in Figure 1 (p. 7) and a list of identifiers for fatty acids commonly identified in plant oils are listed in Table 1.

Table 1. Trivial names and identifiers for fatty acids commonly found in plant oil TAGs. Note: The C:DB notation for each fatty acid is the number of fatty acyl carbon atoms followed by the number of double bonds.

Trivial name	Initial/Identifier	C:DB
Palmitic	P	16:0
Stearic	S	18:0
Oleic	O	18:1
Linoleic	L	18:2
Linolenic	Ln	18:3

TAG molecular species are separated by RP-HPLC and elute in order of their equivalent carbon number (ECN), which is approximately equal to the number of FA carbon atoms minus two times the number of carbon-carbon double bonds (ECN ≈ C – 2·DB). Components with the same ECN are called 'critical pairs' and tend to have similar retention times. For example, the molecules OOO (18:1/18:1/18:1), POO (16:0/18:1/18:1), and POP (16:0/18:1/16:0) all have ECN ≈ 48 and elute in the same region. Modern chromatographic equipment may separate these three components, but regioisomers such as POP and PPO typically coelute, and the complexity of TAGs in many naturally-occurring oils requires the added dimensionality of mass spectrometric detection (5, 56). Several studies have successfully quantified such regioisomers using RP-HPLC/APCI-MS (27, 30, 31, 57, 58), but quantification procedures require standards for all of the individual TAG molecular species present in an oil. These standards are expensive and are not yet available for all known TAG molecular species, and such an analysis would be time consuming on the part of the analyst and may be prohibitively expensive. While the ultimate aim of these analyses is to obtain a qualitative and quantitative profile of all of the TAG molecular species in an oil, this type of comprehensive analysis is not necessary for classification purposes, such as determining the biological source of an oil.

APCI-MS is a "soft" ionization technique that produces relatively simple spectra from TAGs with base peaks consisting of either the protonated molecule, $[M+H]^+$, or diacylglycerol ions, $[DAG]^+$ or $[M-RCO_2]^+$, that result from the loss of a FA moiety (27, 56). The presence of the protonated molecule depends on the degree of unsaturation of the acyl moieties, with the relative abundance of the $[M+H]^+$ ion increasing with the number of double bonds in the molecule; this ion may be absent in saturated TAGs (27). Relative intensities of $[DAG]^+$ ions provide information on the positions at which FAs are attached to the glycerol backbone. $[DAG]^+$ ions resulting from the loss of the FA moiety at position *sn*-1 and -3 are observed in greater abundance than those resulting from a loss at *sn*-2. Lesser-abundant acylium ions, $[RCO]^+$, that correspond to the individual FA moieties may also be present in TAG spectra. These ions, together with HPLC retention time information, allow identification of the FAs attached to glycerol in each TAG species.

Jakab et al. (4) used linear discriminant analysis (LDA) (59) to classify 42 plant oils to twelve different biological sources based on twelve relative peak areas calculated from RP-HPLC/APCI-MS extracted ion chromatograms with a reported classification accuracy of 97.6%. Only one peanut oil (of two) was misclassified and the authors suggested that the oils may have had different geographical origins that resulted in differing TAG profiles. The reported classification accuracy is based on the resubstitution error of the classifier, or the ability of the model to classify the training data, as opposed to an estimate of the generalization error, which would provide more information on the ability of this method to determine the biological source of new samples that were not analyzed in this study. Reanalysis of these data using LDA with leave-one-out cross-validation (CV) resulted in a reduced accuracy of 76.2%; and a random forest (RF) classifier (48) resulted in 85.7% classification accuracy based on the out-of-bag error rate. In this case, the random forest result is likely a better indicator of model performance, because uneven sample class sizes and a large number of classes (12) relative to the number of samples (42) may interfere with the performance of LDA. These results indicate that data from RP-HPLC/APCI-MS analyses of TAGs in plant oils, used as inputs for multivariable classification methods such as LDA and RF, may be used to accurately classify plant oils to their biological source.

Developments in the fields of proteomics and metabolomics have employed a signal-based approach for the automated processing of raw instrument data to circumvent the many challenges associated with non-targeted profiling of peptides and metabolites using comparative LC/MS (34, 39, 60, 61). The individual data files from LC/MS experiments are treated as a two-dimensional signal matrix, or image. Such an image is shown in Figure 7, with columns composed of mass spectral scans over the duration of the chromatographic run and rows of extracted ion chromatograms; each cell in the matrix is a mass-to-charge ratio (m/z) abundance measured within a scan at a particular retention time (RT). Established methods in signal processing, statistics, and machine learning are used to find patterns in the image that are characteristic of samples or sample classes (32).

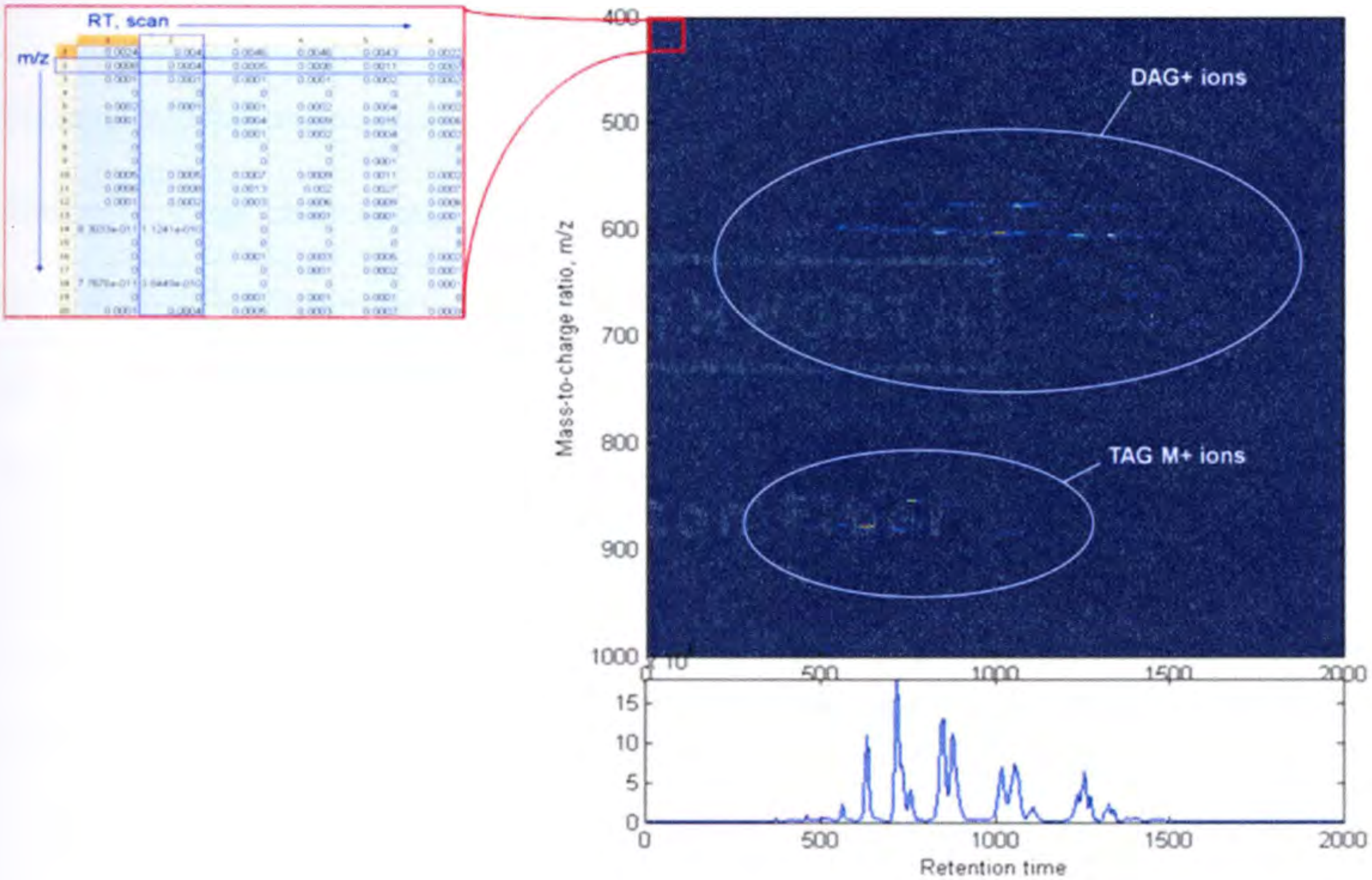


Figure 7. Image representation of an LC/MS analysis. Each cell in the matrix on the left is a mass-to-charge ratio (m/z) abundance measured within a mass spectral scan at a particular retention time; columns contain data from mass spectral scans and rows represent extracted ion chromatograms. The LC/MS image for a soybean oil is shown on the right, with DAG ions and protonated molecules labeled. The total ion chromatogram is shown below the image, and is simply a plot of the sum of column intensity values.

The freely available computational tools MZmine (8, 41) and xcms (9) contain methods for spectral filtering, peak detection and chromatographic alignment that allow the analyst to set data-specific processing parameters. Such tools will ideally provide a list of peak areas corresponding to m/z and RT indices that are quantitatively similar to those obtained by manual data processing, using a method such as the one used by Jakab et al. (4). MZmine and xcms have been reviewed (34, 36) and used successfully in several published metabolomics studies (62-65); the two programs have been found to yield comparable results (65).

The current study set out to evaluate the ability of the computational tools MZmine and xcms to process TAG RP-HPLC/APCI-MS data, compared with manual processing of the same data by a trained analyst. We sought to develop a high-throughput data processing strategy to discriminate among commercially-available plant oils from different biological sources. To this aim, semi-quantitative data for twelve targeted variables were obtained using the two computational tools MZmine and xcms, and compared with results obtained by manually processing the same data using the method of Jakab et al. (4). These twelve variables were used as inputs for LDA and RF classification models. Misclassification rates (%) were used as a metric to evaluate MZmine and xcms data, compared with data that were manually processed. Pearson correlation coefficients between values for twelve specific variables in the manually processed data and MZmine and xcms were examined to determine whether the chemometric data sets reflected values determined from manual processing. Additionally, we evaluated TAG structures associated with the targeted features used for classification using a new tool (described in the Appendix) that determines the FA substitutions for specific features from mass spectral data using xcms. The use of computational tools provides a high-throughput data processing methodology that may be widely employed for authentication of commercial oil samples and for identifying TAG structures associated with specific features useful for classification of oils to their biological source.

METHODS

Sample preparation

Thirty commercial plant oil samples (almond, grapeseed, linseed, olive, peanut, and soybean oils, 5 samples per class) were purchased from local grocery stores and on-line vendors. Samples were diluted in acetone/acetonitrile (2:1, v/v) to a concentration of 1%. Burdick and Jackson solvents were obtained from VWR (West Chester, PA); all solvents were HPLC grade or the highest purity available, and were used without further purification.

RP-HPLC/APCI-MS

RP-HPLC/APCI-MS analyses were performed using an Agilent 1100 quaternary pump HPLC system and Agilent XCT ion trap MS equipped with APCI source (Agilent Technologies, Palo Alto, CA). Separation of TAG was achieved using a Restek Allure C18 column (5 μm, 250 × 2.1 mm, Restek Corporation, Bellefonte, PA) with a two-stepped linear gradient of acetone in acetonitrile at flow rate 0.6 mL/min. Both solvents contained 0.1% acetic acid to facilitate ionization. Acetone concentration was held at 20% for 1 min, stepped to 66% at 4 min and held for 13.5 min, then stepped from 66% to 90% in 1 min and held at 90% until 45 min. Autosampler and column temperatures were 20°C and 35°C, respectively. The injection volume was 3 μL. Direct infusion MS was performed in Ultrascan mode with the following parameters: APCI temperature, 350°C; vaporizer temperature, 500°C; corona current, 5000 nA; nitrogen sheath and auxiliary gas, 60 psi and 5 L/min, respectively. Mass spectra were collected in positive ion mode from m/z 100-1200 with a scan time of 300 ms.

Manual data processing

Samples were processed manually using DataAnalysis software (Bruker Daltonics, Ver. 3.3), as described by Jakab et al. (2002). Briefly, peak areas for twelve TAG analytes were calculated from extracted ion chromatograms of either the protonated molecule $[M+H]^+$ or one of the DAG fragment ions $[M-RCO_2]^+$ and converted to area percent values. Data including the m/z values used for extracted ion chromatograms for each of the twelve TAG are listed in Table 2. The total ion and extracted ion chromatograms for a typical soybean oil are shown in Figure 8. While the area percent value of a particular peak is not a true representation of the concentration, we can use this information to describe the TAG profile of a particular oil sample and such a profile is sufficient for classifying the oil to its biological source (4).

Table 2. Twelve TAG analytes. TAG identifiers refer to the following fatty acids: P=16:0, S=18:0, O=18:1, L=18:2, and Ln=18:3, where C:DB indicates the number of fatty acyl carbon atoms and double bonds. The equivalent carbon number (ECN) for a TAG is approximately equal to the total number of FA carbons minus 2 times the total number of double bonds in the acyl chains.

TAG*	ECN	RT/(m)†	Ion	EIC m/z (± 0.5)	Assigned TAG
LLLn	40	10.6 ± 0.1	$[M+H]^+$	877.7	LLLn / OLnLn (1L)‡
LLL	42	12.1 ± 0.2	$[M+H]^+$	879.7	LLL / OLLn (5L,2O)‡
LnLP	42	12.6 ± 0.1	$[M+H]^+$	853.7	LnLP
LLO	44	14.2 ± 0.2	$[M+H]^+$	881.8	LLO / OLnO (5L)‡
PLL	44	14.8 ± 0.2	$[M+H]^+$	855.7	PLL / OLnP (5L,1O)‡
OOL	46	17.1 ± 0.2	$[OL]^+$	601.5	OOL
PLO	46	17.8 ± 0.2	$[PO]^+$	577.5	PLO
PLP	46	18.6 ± 0.2	$[PP]^+$	551.5	PLP
OOO	48	20.6 ± 0.1	$[OO]^+$	603.5	OOO / SOL (1G)‡
POO	48	21.0 ± 0.1	$[PO]^+$	577.5	POO
POP	48	21.3 ± 0.1	$[PO]^+$	577.5	POP
SOO	50	22.1 ± 0.1	$[SO]^+$	605.5	SOO

* Actual TAG structure, from (4).

† Mean ± SEM.

‡ Numbers and letters in parentheses indicate number and type of oils for which the second TAG structure was assigned. L = linseed, O = olive, G = grapeseed.

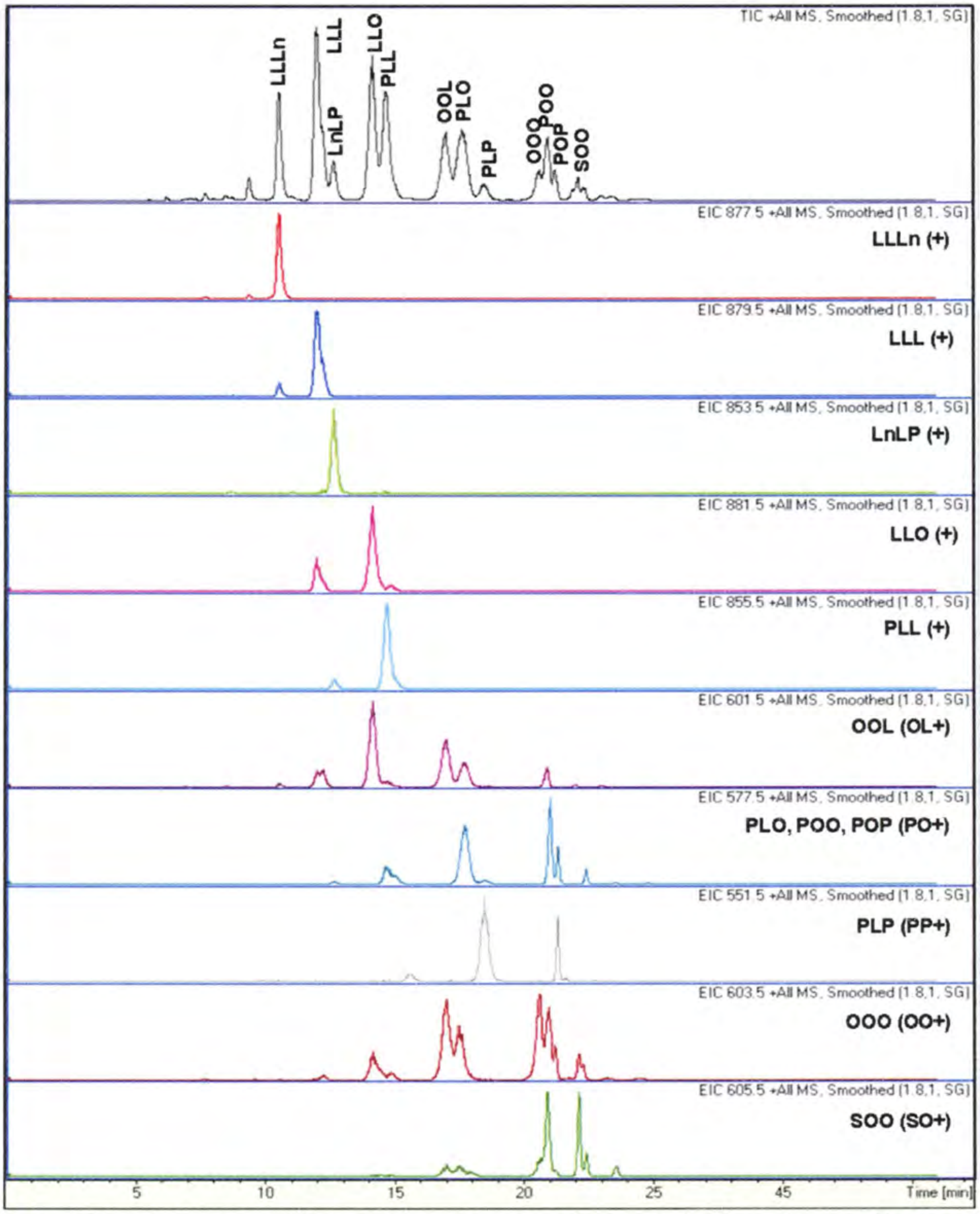


Figure 8. Total and extracted ion chromatograms for a typical soybean oil sample.

MZmine

Data files were converted from the Bruker MS proprietary data file format to mzXML using the CompassXport program (Bruker Daltonics, Ver. 1.3); mzXML files were read directly into MZmine (Ver. 0.60). The data were cropped to m/z 200-1000 and RT 250-

1600 s and a chromatographic median filter (m/z tolerance = 0.2, scan windows = 6) was used to smooth data in the chromatographic direction, and peaks were detected using the recursive threshold method (m/z bin size = 0.25, chromatographic threshold = 5%, noise level = 2,000,000, minimum peak height = 5,000,000, minimum peak duration = 5 s, minimum m/z peak width = 0.5, maximum m/z peak width = 25, m/z tolerance = 1.2, intensity tolerance = 40%). Peak lists from the different files were aligned by matching each peak list to a master list using the Fast aligner algorithm (balance = 10, m/z tolerance = 0.5, RT tolerance = 5%). Peaks that were found in less than three samples were removed, and empty slots in alignment results were filled by searching for a local maximum over the region of raw data where a peak was likely to be located (intensity tolerance = 100%, m/z tolerance = 0.5, RT tolerance = 5%). Peaks were normalized by the total raw signal to remove systematic variation in intensity levels between different data files. The twelve peaks used for manual data processing were selected from the MZmine output and converted to area percent values.

Xcms

The OpenMS TOPPView program (http://open-ms.sourceforge.net, Ver. 1.2) was used to crop the LC/MS image size to m/z 200-1000 and RT 250-1600 s; files were imported into TOPPView in mzXML and exported as mzData files for processing with the xcms package (Ver. 1.14.1) in R (Ver. 2.7.2). Matched filter peak detection was used with the following parameters: sigma = 6.5, max = 25, step = 0.1, steps = 5. Peaks were grouped together across samples using fixed-interval overlapping m/z bins (mzwid = 0.25) and calculation of smoothed peak distributions in chromatographic time using a Gaussian kernel density estimator (bw = 10). RTs were corrected for all samples simultaneously using loess regression to model nonlinear RT deviation contour profiles based on peak group median RTs and deviations from the median. The corrected peak lists were re-grouped using a smaller Gaussian kernel (bw = 5), and samples with missing peak values within a group were filled by integrating raw data in the peak group region using corrected start and ending retention time points defined by the peak group medians. A matrix of peak area values with rows for every group, indexed by m/z and RT, and columns for every sample were generated. Samples with multiple peaks per group were resolved by choosing the peak closest to the median RT. Xcms does not provide a function for normalization, so this was performed manually by dividing each peak by the total area for the sample. The same twelve peaks used for manual and MZmine data processing were selected from the XCMS output and converted to area percent values. We performed a recursive search of the peak list for m/z values we would expect to observe in TAGs resulting from combinations of the five most prominent FAs observed in the FA profile analysis (Table 1).

Statistical analysis and classification

All statistical procedures, classification, and data manipulation were performed using R (Ver. 2.7.2) and Microsoft Office Excel (2007). RF classifiers were generated using the R randomForest package (Ver. 4.5-28) and linear discriminant analyses were performed with the R MASS package (Ver. 7.2-45). RF classifiers were built using 5,000 decision trees with 3 (of 12) variables randomly selected for differentiation at each node. As the RF algorithm uses bootstrap samples to create individual decision trees and implements a stochastic variable selection at the nodes of individual trees, results were confirmed with multiple analyses for all RF models reported. RF output provides the "out-of-bag" error obtained by classifying samples which were not a part of the bootstrap data used to build a specific tree. This provides an unbiased estimate of the generalization error, or the error we can expect when trying to use this model to classify new data, and eliminates the need for cross-validation (48). We also calculated the resubstitution error by using the RF to classify the same data that were used to generate the RF model. As LDA only provides the resubstitution error, we used leave-one-out cross-validation to estimate the generalization error (59).

RESULTS AND DISCUSSION

Area percent values for the twelve TAG analytes listed in Table 2 were obtained by manually processing RP-HPLC/APCI-MS data, and from the peak lists resulting from processing the same data with MZmine and xcms programs. Data (mean ± SEM) for each of the three processing methods are displayed separately and grouped by biological source of the oil samples, in Figure 9. It is apparent in the bar plots that the three processing methods result in similar information for these twelve variables, and that the feature values vary among the different types of oils.

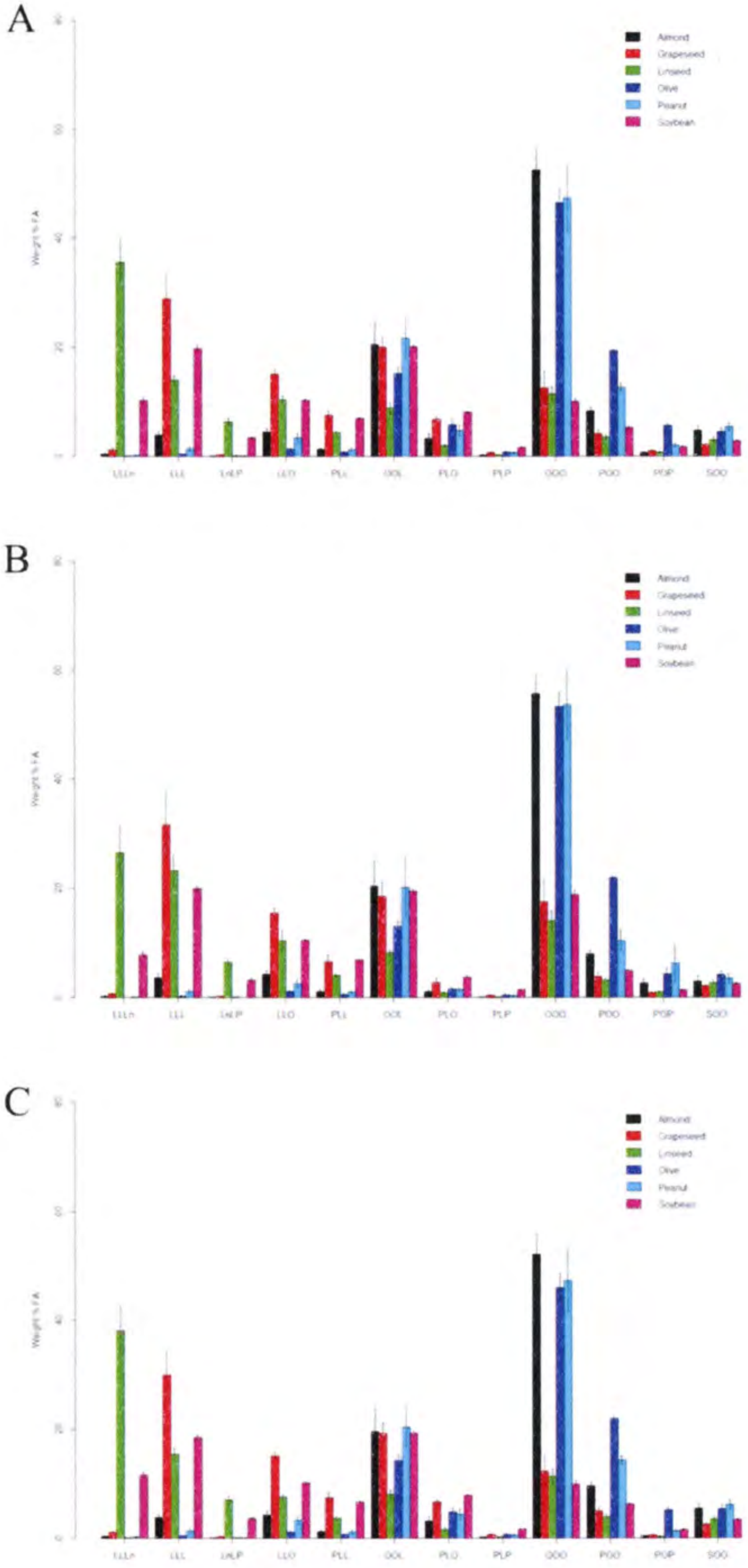


Figure 9. Barplots (Mean ± SEM) of (A) manually processed, (B) MZmine, and (C) xcms data sets, grouped by the biological source of the oil samples.

While manual processing targeted specific peaks representing the twelve TAG analytes, MZmine and xcms are non-targeted processing methods. Both of these programs first identify all peaks in each RP-HPLC/APCI-MS image, and then align the peaks from different samples that represent the same analytes in the retention time dimension. MZmine allows the user to filter out rare peaks, estimate areas for peaks that were missed on the first round based on user-specified parameters, and normalize each peak value to the total signal to remove systematic variation in intensity levels. Xcms also estimates peak areas for features missed during peak detection, and the group function may be used to filter out rare peaks after alignment; xcms does not include a function for normalization. MZmine processing resulted in a list of 129 features, including the twelve targeted variables of interest. Xcms returned a longer peak list, consisting of 2408 features (including the twelve targeted features), and peaks were normalized manually by dividing by the total area. The different number of features obtained using MZmine vs. xcms is likely a result of peak detection parameters used for the different methods. The xcms peak list was filtered to obtain a peak list of 161 relevant ions representing [DAG]+ and [M+H]+ ions from possible TAG analytes resulting from all combinations of the FAs listed in Table 1.

LDA resubstitution and leave-one-out CV error rates and RF resubstitution and out-of-bag error rates are listed in Table 3. The twelve targeted features (Table 2) produced classification results similar to those reported by Jakab et al. (1) for all three processing methods. As noted previously, reanalysis of those data using LDA with leave-one-out CV and the RF algorithm yielded much higher estimates of generalization error (23.8% for LDA with CV and 14.3% for RF out-of-bag error) than the reported resubstitution error (2.4% for LDA). We analyzed 30 oils from six different biological sources (n=5 per class), while Jakab et al. analyzed 42 oils from twelve biological sources, with only two samples in the smallest class. The lower generalization error estimates that we report may be a result of balancing the class sizes for our data and looking at fewer classes overall. For our oil samples, manual processing of the RP-HPLC/APCI-MS data yielded the lowest classification error rates, followed by using relative values for the twelve targeted features from automated processing with xcms and then MZmine. XCMS values for the twelve targeted features were more highly correlated with manually-processed data than MZmine values (Table 4). XCMS peak integrations for the twelve targeted features are shown in Figure 10. Dot plots of the mean decrease in accuracy in the RF model for the xcms values of the 12 targeted features are shown in Figure 11. POO (m/z 577.5, RT 21.0 ± 0.1) was the "most important" variable in this model.

Table 3. Misclassification rates for LDA (resubstitution), LDA with leave-one-out cross-validation (CV), and random forest classifiers (resubstitution and out-of-bag error rates).

Data	Features	LDA	LDA (CV)	RF	RF (OOB)
Manual	12*	0% (0/30)	6.67% (2/30)	0% (0/30)	3.33% (1/30)
MZmine	12*	3.33% (1/30)	13.3% (4/30)	0% (0/30)	13.3% (4/30)
MZmine	129	--	--	0% (0/30)	3.33 % (1/30)
XCMS	12*	0% (0/30)	10.0% (3/30)	0% (0/30)	6.67% (2/30)
XCMS	2408	--	--	0% (0/30)	3.33% (1/30)
XCMS	161†	--	--	0% (0/30)	3.33% (1/30)

* These peaks represent the same twelve analytes: LLLn, LLL, LnLP, LLO, PLL, OOL, PLO, PLP, OOO, POO, POP, and SOO, with m/z values and RTs corresponding those listed in Table 2. LDA was not applied to data sets where the number of features exceeded the number of observations (n = 30).

†Xcms data screened specifically for [DAG]$^+$ and [M+H]$^+$ ions in TAGs containing the FAs 16:0, 18:0, 18:1, 18:2, and 18:3.

Table 4. Pearson correlation coefficients between feature values from manually processed data and the two chemometric processing methods.

Feature	MZmine	xcms
LLLn	0.989	0.994
LLL	0.954	外.995
LnLP	0.993	0.998
LLO	0.961	0.976
PLL	0.932	0.995
OOL	0.930	0.996
PLO	0.813	0.982
PLP	0.966	0.989
OOO	0.982	0.999
POO	0.939	0.999
POP	0.362	0.986
SOO	0.515	0.997

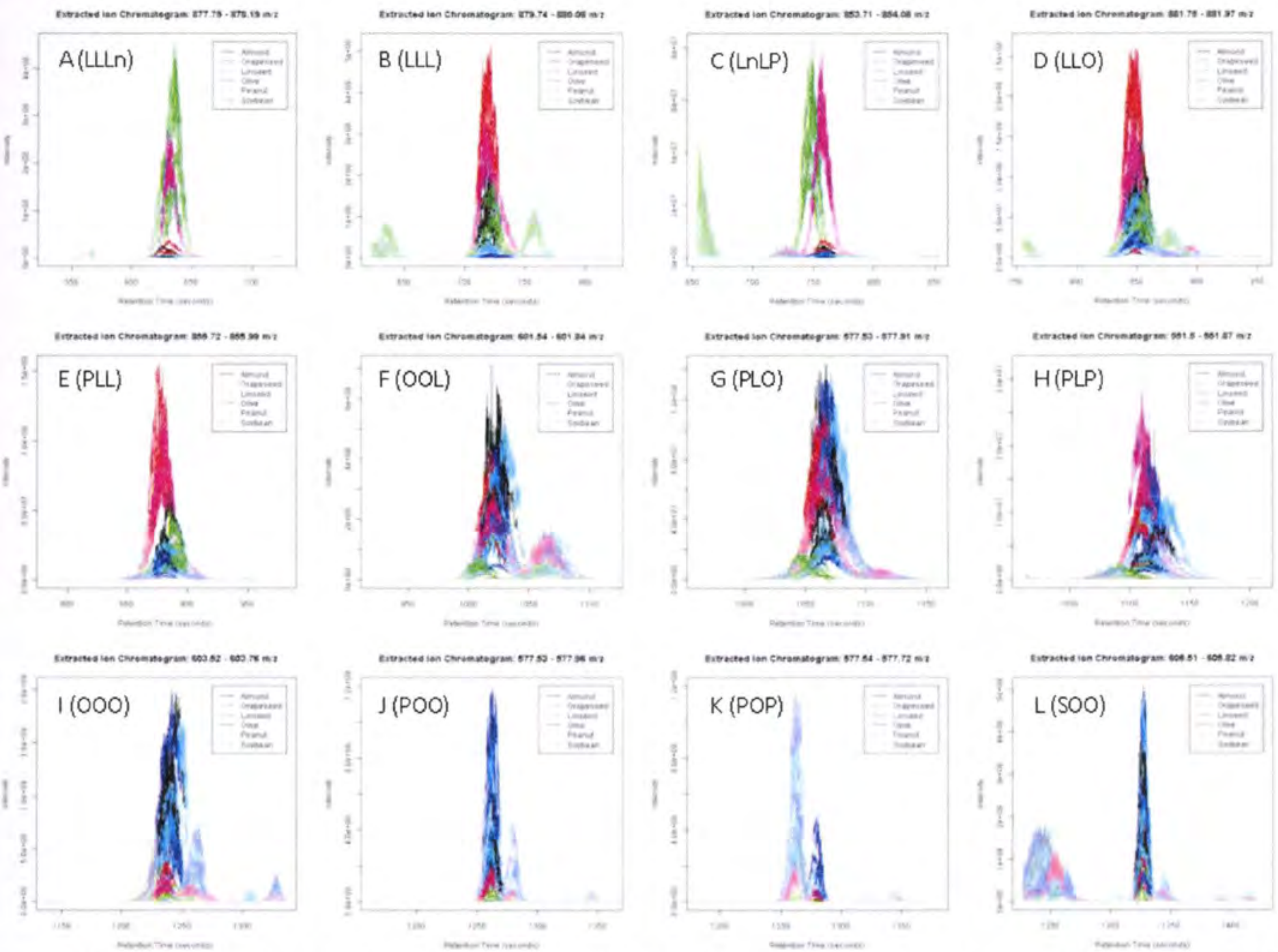


Figure 10. Extracted ion chromatograms for twelve targeted features; data processed with xcms. Integrated areas for each peak are indicated by darker colors, with lighter areas outside the area of integration.

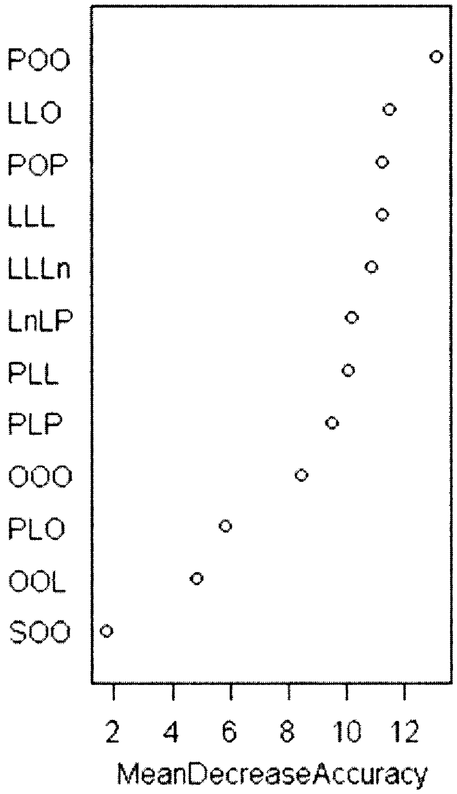


Figure 11. Variable importance of the xcms values for the twelve targeted features, estimated by the mean decrease in accuracy of the RF classifier.

Non-targeted data from MZmine (129 features) and xcms (2408 features) and xcms output filtered for peaks representing $[M+H]^+$ and $[DAG]^+$ ions from possible TAG analytes (161 features) performed similarly to each other, with all three models having 97% classification accuracy based on the RF out-of-bag error rates. As LDA must have fewer variables than observations (n=30), we could not use this model for these data sets. Multidimensional scaling (MDS) plots based on the RF proximity matrix were used to visualize groupings among the different classes, or biological sources of the oils, in the different feature sets (Figure 12). All of the RF models using twelve features produced similar MDS plots, though Coordinate 1 and 2 axes were reversed in the xcms plots [Figure 12 (C) and (D)].

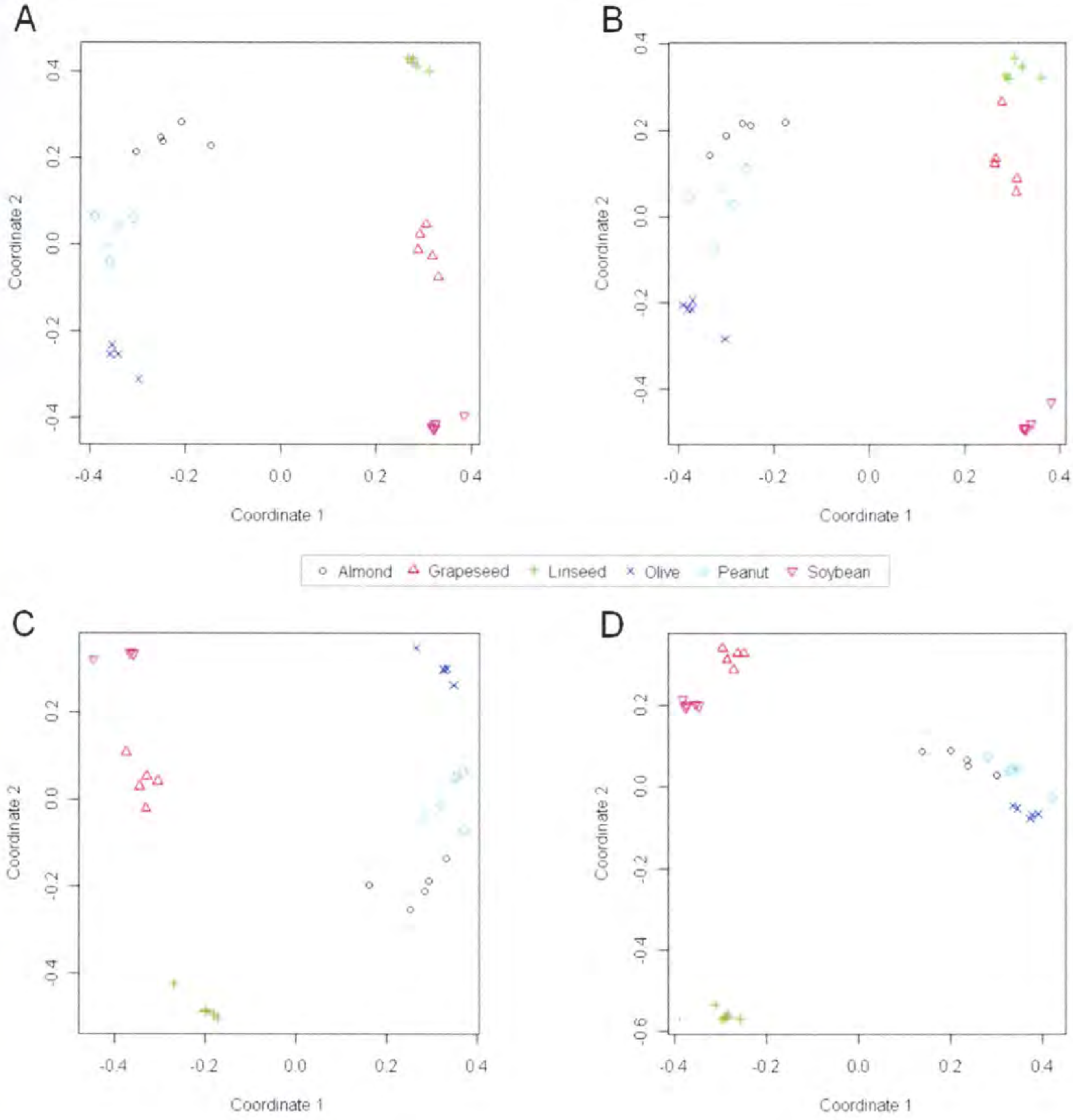


Figure 12. Multidimensional scaling plots of the RF proximity matrix from different processing methods for the twelve targeted features processed (A) manually, using (B) MZmine and (C) xcms, and (D) 161 TAG-filtered xcms features.

Automated TAG structure assignments are listed in Table 2. All twelve targeted features were assigned the same TAG structures that were documented by Jakab and colleagues for most of the plant oil samples (4). Five features were assigned an alternate structure in one or more samples; these are noted in Table 2. Notably, all five linseed oils were assigned alternate structures for LLL (m/z 879.7, RT 12.1 m), LLO (m/z 881.8, RT 14.2 m), and PLL (m/z 855.7, RT 14.8 m); the structures assigned for these peaks in the linseed oils were OLLn, OLnO, and OLnP, respectively. Examination of the target feature mass spectra for LLL, LLO, and PLL in soybean and linseed oil samples shown in Figure 13 supports the alternative TAG structure assignments in the linseed oils. Ions observed in the soybean oil LLL spectrum are 599.5 and 879.7; additional peaks at m/z 597.4 and 601.5 observed in linseed oil indicate the presence of OLLn. Similar patterns are observed in LLO (vs. OLnO) and PLL (vs. OLnP) spectra: LLO m/z 599.5, 601.5, and 881.8, OLnO m/z 599.5, 603.5, and 881.8; PLL m/z 575.5, 599.5, and 855.7, OLnP m/z 573.5, 577.5, 599.5, and 855.7. So, for linseed oil, the amounts of target features m/z 879.7, 881.8, and 855.7 actually represent different TAG species from the other plant oils. It is unclear what effect this may have on classification results. TAG structure assignments may be used to select features that represent the same structure in different samples, assuring that we compare the same elements in different types of oils. The presence of these alternative TAG species in appreciable quantities in linseed oil is confirmed by the observations of Lisa, et al. (66).

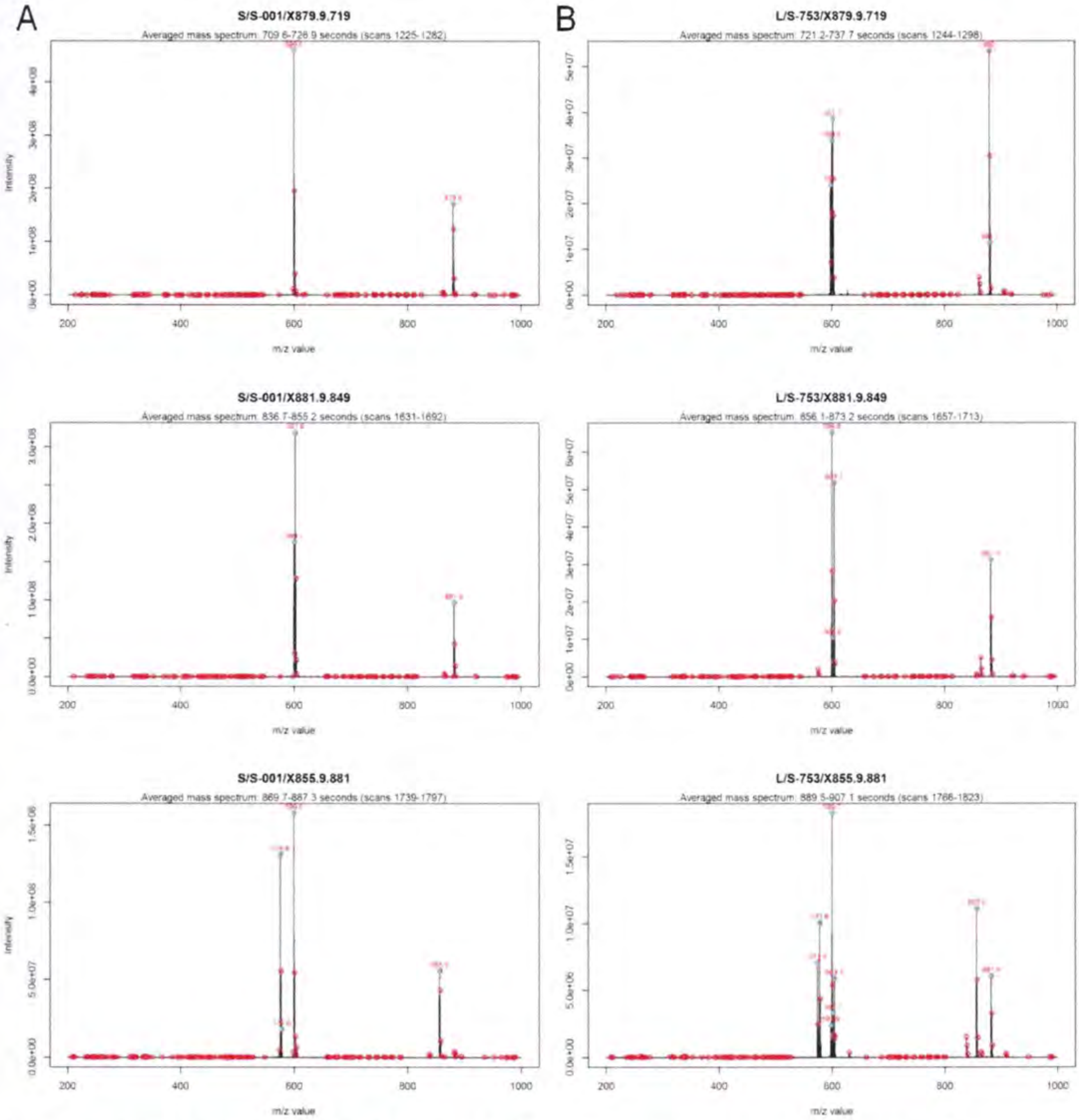


Figure 13. (A) Soybean and (B) linseed oil mass spectra for targeted peaks LLL (X879.9.719), LLO (X881.9.849), and PLL (X855.9.881).

Overfitting becomes a problem when model complexity increases with the number of features (59, p. 194) such that the model does not accurately classify new data. Non-targeted peak detection programs, such as MZmine and xcms, generate large lists of peaks with many redundancies in the data. As each analyte may produce several peaks, correlations may exist among variables. Searching for useful features among these data is the focus of many recent studies. Resubstitution error rates are poor estimates of a model's ability to generalize, or to classify samples that were not used in generating the model. For all feature sets that we studied, the generalization error estimates were higher than (or equal to) resubstitution errors, as would be expected. This illustrates the need to perform some sort of estimation of the generalization error for all classification models. The RF out-of-bag error rate was lower than the cross-validated LDA error rate for both manually processed data and XCMS, implying that the RF algorithm may be superior to LDA for these data. RF has been demonstrated to outperform other classifiers for MS data, and is not subject to the stringent assumptions of the LDA model (49).

One objective of this research was to evaluate the data produced by MZmine and xcms, quantitatively and in terms of classification of oils to their biological source, compared with manual processing of the same data by a trained analyst. Of the three processing methods, the manually processed data were best able to classify the plant oils to their biological source based on the twelve targeted features, followed closely by xcms and then MZmine (Table 3). Examining scatter plots and correlations among the manual, MZmine, and xcms processing methods for each of the twelve variables revealed that xcms values were more closely related to the manually processed data than MZmine values (Table 4), and therefore we determined that xcms outperforms MZmine for processing these data with the parameters employed. It is possible that MZmine performance could be improved with a different parameter set. A great deal of time can be spent optimizing parameters and we did not perform any true optimization procedures. We settled on the parameters that gave peak lists containing all twelve of the targeted features and produced peak integrations that were acceptable based on a visual inspection. Additionally, when using xcms, once the original files are imported to generate an xcmsSet variable, all further statistical analyses and TAG structure assignment can be performed in R. We found this to be a great advantage, and used the xcms data to develop a TAG structure assignment tool for this reason (see Appendix). The results of this study indicate that data from RP-HPLC/APCI-MS analyses of TAGs in plant oils, used as inputs for multivariable classification methods such as LDA and RF, may be used to accurately classify plant oils to their biological source with high accuracy and that programs such as xcms and MZmine allow analysts to devise a high-throughput methodology to achieve this end.

PAPER 2: FORENSIC IDENTIFICATION OF SEAL OILS USING LIPID PROFILES AND STATISTICAL MODELS

Margaret H. Broadwater[1,2], Gloria T. Seaborn[1] and John H. Schwacke[2]

[1]Center for Coastal Environmental Health and Biomolecular Research (CCEHBR), National Ocean Service (NOS), National Oceanic and Atmospheric Administration (NOAA), 219 Fort Johnson Road, Charleston, SC 29412.

[2]Department of Biochemistry and Molecular Biology, Medical University of South Carolina, 173 Ashley Avenue, Charleston, SC 29425.

ABSTRACT

Seal blubber oils are used as a source of omega-3 polyunsaturated fatty acids in Canada but prohibited in the United States and European Union. Thus, a reliable method is needed to identify oils originating from seals vs. fish. Two lipid profiling methods, fatty acid analysis using gas chromatography and triacylglycerol analysis using liquid chromatography and mass spectrometry, were applied with statistical models to discriminate commercial oils and blubber samples harvested from marine fish and seals. Significant differences were observed among fatty acid profiles, and seal samples differed from each of the fish oils ($p \leq 0.001$). Fatty acid and triacylglycerol profiles were used to discriminate sample groups using a random forest classifier; all samples were classified correctly as seals vs. fish using both methods. We propose a two-step

method for the accurate identification of seal oils, with preliminary identification based on fatty acid profile analysis and confirmation with triacylglycerol profiles.

INTRODUCTION

The recent focus of the biomedical community on dietary fats and their relation to health and disease states has brought to light the necessity of including polyunsaturated fatty acids (PUFA) in a healthy diet, and in particular the long-chain omega-3 (n-3) PUFA eicosapentaenoic (EPA, 20:5n-3) and docosahexaenoic (DHA, 22:6n-3) acids found in fatty fish. For those unable or unwilling to increase dietary fatty fish intake, fish oil omega-3 PUFA dietary supplements are available. A national survey conducted in 2007 by the Centers for Disease Control and Prevention (CDC) revealed that fish oil/omega-3/DHA supplements were the most popular products used by consumers for health reasons (67). In Canada, seal oil supplements manufactured primarily from harp seal blubber are available to consumers (68-71), but the Marine Mammal Protection Act of 1972 (MMPA; 72) prohibits buying and selling of seal products in the United States (US). The European Union (EU) also banned commercial seal products in May 2009 (73). A reliable method to distinguish between marine oils harvested from fish and seals is necessary for law-enforcement purposes.

While many fish products may be identified to species using DNA analysis, verifying the biological source of a marine oil is more difficult because oils are composed of lipids and typically do not contain amplifiable DNA. Authentication and determination of the biological source of marine oil dietary supplements may be achieved via compositional analyses of the lipids present in such oils (17). Natural marine oils harvested from fish and seals are typically composed of triacylglycerols (TAGs), in which three fatty acids (FAs) are esterified to glycerol. The FAs vary in carbon chain length and the number of double bonds, with most naturally occurring FAs having 14-24 carbons and 0-6 methylene-interrupted double bonds in the *cis* configuration (5, 74). FAs are identified using the notation A:Bn-C, where A is the number of carbon atoms, B is the number of double bonds, and C is the position of the first double bond from the terminal methyl group. The FA profile of an oil is a quantitative list of the FAs present, measured as fatty acid methyl esters (FAMEs) using gas chromatography with flame ionization detection (GC/FID); amounts are relative and sum to 100%.

FA profiles have been used to reliably distinguish among different fish species (75-77) and marine mammal populations and subspecies (78, 79), and to differentiate between wild and cultured fish for forensic purposes (80, 81). Marine oil FA profiles are species-specific with some overlap and variation due to age and differences in diet (17, 82). Characteristics of seal oil FA profiles noted in the literature include a 16:1n-7/16:0 ratio greater than unity (83, 84) and high levels of 18:1n-11 (68) and 22:5n-3 (18, 71, 82) when compared with other marine oils. Additionally, FA distribution on the glycerol backbone in TAG molecules differs between seals and fish. Studies of positional distribution of FAs have consistently shown that long-chain PUFA including 20:5n-3, 22:5n-3, and 22:6n-3 are located in the TAG *sn*-2 position in fish oils and the *sn*-1/3 positions in seals and other marine mammals (68, 71).

Reversed-phase high performance liquid chromatography with atmospheric-pressure chemical ionization mass spectrometry (RP-HPLC/APCI-MS) can be used to obtain semi-quantitative TAG profile data for the purpose of classifying oils to their biological source. RP-HPLC has proven useful in the separation of TAG molecular species, but cannot resolve individual TAGs in complex mixtures. TAGs elute on an octadecylsiloxane (ODS) column in order of their equivalent carbon number (ECN), approximated by the number of fatty acyl carbon atoms minus two times the number of carbon-carbon double bonds. TAGs with the same ECN tend to elute close together (5). APCI-MS of TAGs produces protonated molecules, $[M+H]^+$, and one to three diacylglycerol (DAG) fragment ions, $[M-RCO_2]^+$, depending on the number of different FAs attached to glycerol (28). The relative intensities of DAG fragment ions reveal information on the positions of FAs on the TAG molecule (29). Specifically, the least abundant DAG ion resulted from loss of the FA from the secondary position (*sn*-2). RP-HPLC/APCI-MS allows for partial identification of the FAs (number of carbons and double bonds, e.g. 18:3) from individual TAGs in a mixture and also provides pertinent information on positions of FAs on the glycerol backbone in the TAG molecule (27). Because differences exist in TAG positional distribution of long-chain PUFA between seal and fish oils, RP-HPLC/APCI-MS should be useful in discriminating these oils. While the complexity of marine oil TAGs does not allow quantification of individual TAG molecular species, TAG RP-HPLC/APCI-MS data may be used with classification models to discriminate between marine oils from fish and seals. Such data must be subjected to preprocessing steps including peak detection, alignment, and normalization prior to classification (39). This study determines FA composition for commercial marine oils and marine oil dietary supplements from five different biological origins (Cod liver, herring, salmon, generically labeled "fish," and seals; to evaluate differences among oils from different marine sources and address the feasibility of using such data to identify seal oil using classification models for forensic purposes. We report the FA composition for 45 commercial marine oils and 3 extracted seal blubbers and use these data to classify marine oils to five different biological sources. RP-HPLC/APCI-MS of TAG in marine oils is implemented as a confirmatory analysis procedure to verify classification and to ascertain that the oils are not synthetic in origin. We propose that FA and TAG profiling methodologies should be used together for the forensic identification of seal oils, and assert that this two-tiered analysis approach, when used with the appropriate statistical models, can identify seal oil definitively and without bias on the part of the analyst.

METHODS

Harp seal (*Phoca groenlandica*) blubbers and commercially marketed fish and seal oils were obtained from government agencies, academic institutions, and commercial vendors as described in Table 5. Possession of marine mammal specimens was authorized under National Marine Fisheries Service (NMFS) Permit No. 13599. Burdick and Jackson solvents were obtained from VWR (West Chester, PA); all solvents were HPLC grade or the highest purity available, and were used without further purification.

Table 5. List of samples.

Sample	Source
Cod-1	Norwegian Cod liver oil, 0.5 g capsules, Dale Alexander, Norway
Cod-2	Natural Cod liver oil, 0.65 g capsules, Sundown, USA
Cod-3	Cod liver oil (with vitamins A & D), 0.5 g capsules, Premier Value, USA
Cod-4	Cod liver oil (with vitamins A & D), Squibb, USA
Cod-5	Natural Cod liver oil, 0.6 g capsules, Sundown, USA
Cod-6	Arctic Cod liver oil, 1 g capsules, Nordic Naturals, Norway
Cod-7	Norwegian Cod liver oil, 0.39 g capsules, Carlson, Norway
Cod-8	Norwegian Cod liver oil, 0.52 g capsules, Spectrum Essentials, Norway
Fish-1[*]	18/12 Fish oil, Biooriginal Food & Science Corp., Norway
Fish-2†	Res-Q Omega 3 supplement (pelagic fish), 1.25 g capsules, N3 Oceanic, Norway
Fish-3[*]	18/12 Fish oil, ICE, Department of Homeland Security, USA
Fish-4[*]	18/12 Fish oil, ICE, Department of Homeland Security, USA
Fish-5	Nutra Sea EPA/DHA rich omega-3 supplement (sardine, anchovy), Ascenta, Canada
Fish-6	Omega-3 Fish oil (sardine, anchovy), 1.2 g capsules, Nature Made, USA
Fish-7	Wild Fish oil (sardine, anchovy, mackerel, herring), 1 g capsules, Physician Formulas, Inc., USA
Fish-8†	Fisol Enteric-coated fish oil (sardine, anchovy, mackerel), 0.5 g capsules, Nature's Way, USA
Fish-9	Omega-3 Fish oil (sardine, anchovy), 1.2 g capsules, Sundown, USA
Fish-10[*]	Marine body oil (menhaden), Omega Protein, USA
Fish-11[*]	Marine body oil (menhaden), Omega Protein, USA
Fish-12[*]	Fish oil (menhaden), Biomedical Test Materials Program, NOAA, USA
Herring-1[*]	Atlantic Herring oil, Omega Protein, USA
Herring-2[*]	Atlantic Herring oil, Omega Protein, USA
Herring-3[*]	Atlantic Herring oil, Dalhousie University, Canada
Herring-4[*]	North Atlantic Herring oil, Noble, Canada
Herring-5[*]	Atlantic Herring oil, Dalhousie University, Canada
Herring-6[*]	Atlantic Herring oil, Zapata, USA
Salmon-1[*]	Sockeye Salmon oil, NOAA-NMFS, USA
Salmon-2[*]	Salmon oil, NOAA-NMFS, USA
Salmon-3	Salmon oil + Vitamin E, 1 g capsules, Solaray, USA
Salmon-4	Pure concentrated Salmon oil, 1 g capsules, Neolife, USA

Table 5. – continued.

Sample*	Source
Salmon-5	Pure Norwegian Salmon oil, 1 g capsules, Health Laboratories of America, Norway
Salmon-6	Salmon oil, 1 g capsules, Sundown, USA
Salmon-7	Wild Alaskan Sockeye Salmon oil, 1 g capsules, Natural Factors, Canada
Salmon-8	Norwegian Salmon oil, 1 g capsules, Carlson, Norway
Seal-1*	Marine oil (seal), 0.5 g capsules, Healthy Life Co., Canada (NOAA-NMFS)
Seal-2*	Omega-3 DPA EPA DHA, 0.5 g capsules, Super Natural, USA (NOAA-NMFS)
Seal-3*	Seal oil, 0.5 g capsules (NOAA-OLE)
Seal-4*	Commercial Harp seal oil (Dalhousie University, Canada)
Seal-5*	Omega-3 Seal oil, 0.5 g capsules, Terra Nova, Canada (NOAA-OLE)
Seal-6*	Seal oil, 0.5 g capsules, Creators Own, Canada (NOAA-NOS)
Seal-7*‡	Harp Seal oil, Spirulina Bio-Lab Co., Osaka, Japan (FWS-NWPR)
Seal-8*‡	Omega Plus Harp seal oil, 0.5 g capsules, Terra Nova, Canada (FWS-NWPR)
Seal-9*‡	Harp seal oil, 0.5 g capsules, BEC, Canada (FWS-NWPR)
Seal-10*‡	Harp seal oil omega-3 plus, 0.5 g capsules (FWS-NWPR)
Seal-11*	Commercial Harp seal oil (Memorial University of Newfoundland, Canada)
Pgro-1*	*Phoca groenlandica*, Harp seal blubber, NMMTB, NIST, USA
Pgro-2*	*Phoca groenlandica*, Harp seal blubber, NMMTB, NIST, USA
Pgro-3*	*Phoca groenlandica*, Harp seal blubber, NMMTB, NIST, USA

Note: Abbreviations for indicated U.S. government sources: FWS = Fish and Wildlife Service; ICE = Immigrations and Customs Enforcement; NIST = National Institute of Standards and Technology; NMFS = National Marine Fisheries Service; NMMTB = National Marine Mammal Tissue Bank; NOAA = National Oceanic and Atmospheric Administration; NOS = National Ocean Service; OLE = Office of Law Enforcement; NWPR = National Wildlife Property Repository.

*Samples were purchased at local (Charleston, SC) grocery or health food stores unless so indicated.

†Fish-2 and -8 were composed of fatty acid ethyl esters (as opposed to TAGs), so RP-HPLC/APCI-MS analysis was not performed on these samples and they were not used for classification.

‡Seal-7, -8, -9, and -10 were not available for RP-HPLC/APCI-MS analysis and data from these samples are only reported in summary tables and plots for FA profile data; these samples were not used for classification.

Homogenized blubber samples were extracted in hexane (1:30 sample-to-hexane volume). Lipid class composition and sample quality were assessed using thin-layer chromatography (TLC; 85). FAME derivatization according to Metcalfe et al. was followed with FA profile analysis using GC with mass spectrometry (MS) and FID (85, 86). Mass spectra were used to identify individual FAME peaks, in conjunction with comparison of retention times (RTs) with those of known standards. Empirical correction factors determined from quantitative standards (GLC-85, 411, 566, and 617, NuChek Prep, Elysian, MN) were applied to integrated peak areas from the FID chromatogram and compositions reported as weight percent FA. Sample order was randomized prior to analysis to account for systematic changes in instrumental conditions.

Analysis of TAGs using an Agilent 1100 quaternary pump HPLC system and Agilent XCT ion trap MS equipped with APCI source (Agilent Technologies, Palo Alto, CA) was implemented as a secondary step to verify results of the FA profile analysis. Detector optimization was performed using trilinolein (NuChek Prep, Elysian, MN). Lipid extracts were dissolved in acetone-acetonitrile (2:1; 1 mg/mL). TAGs were separated on a Restek Allure C18 column (5 μm, 250 × 2.1 mm; Restek Corporation, Bellefonte, PA) with a two-stepped linear gradient of acetone in acetonitrile (with 0.1% acetic acid) at 0.6 mL/min. Acetone concentration was held at 20% for 1 min, stepped to 66% at 4 min and held for 13.5 min, then stepped from 66% to 90% in 1 min and held at 90% until 45 min, adapted from Jakab et al. (4). Autosampler and column temperatures were 20°C and 35°C, respectively. The injection volume was 3 μL. Direct infusion MS was performed in Ultrascan mode with the following parameters: APCI temperature, 350°C; vaporizer temperature, 500°C; corona current, 5000 nA; nitrogen sheath and auxiliary gas, 60 psi and 7 L/min, respectively. Mass spectra were collected in positive ion mode from mass-to-charge ratio (m/z) 100-1200 with a scan time of 300 ms. Samples were analyzed in random order to account for systematic differences in LC/MS profiles over the time of the analysis, and gradient blanks were analyzed between samples to prevent column contamination.

LC/MS data files were converted to mzXML using CompassXport (Bruker Daltonics, Ver. 1.3); OpenMS TOPPView (http://open-ms.sourceforge.net, Ver. 1.2) was used to crop the LC/MS image to m/z 200-1100 and RT 300-1800 s; files were imported into TOPPView in mzXML and exported as mzData files for processing with the XCMS package, ver. 1.14.1, in R, ver. 2.9.1 (9, 10, 87). Matched filter peak detection was used with the following parameters: sigma = 6, step = 0.25, mzdiff = 0.6. Peaks were grouped together across samples using fixed-interval overlapping m/z bins (mzwid = 0.25) and calculation of smoothed peak distributions in chromatographic time using a Gaussian kernel density estimator (bw = 10). RTs were corrected for all samples simultaneously using loess regression to model nonlinear RT deviation contour profiles based on peak group median RTs and deviations from the median. The corrected peak lists were re-grouped using a smaller Gaussian kernel (bw = 5), and samples with missing peak values within a group were filled in by integrating raw data in the peak group region using corrected RT start and end points defined by the peak group medians. A matrix of peak area values with rows for every group, indexed by m/z and RT, and columns for every sample was generated. Samples with multiple peaks per group were resolved by choosing the peak closest to the median RT. XCMS does not provide a function for normalization, so this was performed manually by dividing each peak by the total area for each sample.

To reduce the number of LC/MS peaks identified using the preprocessing steps above, we implemented two additional steps. First, we performed a recursive search of the peak list for m/z values we would expect to observe in TAGs resulting from combinations of the most prominent FAs observed in the FA profile analysis. The resulting peak list was reduced to 100 variables using minimum redundancy, maximum relevance (mRMR) feature selection algorithm with the mutual information difference scheme and a threshold of 1 to discretize data. The mRMR program selects the features that best discriminate among classes (maximum relevance) while reducing correlations among these features (minimum redundancy) (88, 89).

All statistical procedures, classification, and data manipulation were performed using R, ver. 2.9.1 (87), and Microsoft Excel (2007). Multivariate analysis of variance using distance matrices was performed using the *adonis* function in the R vegan package, ver. 1.17-0 (90), to test for differences between seal oils and blubbers, and for differences among oils from five different biological sources (Cod, "Fish", Herring, Salmon, Seal). This procedure is analogous to a nonparametric MANOVA performed on a distance matrix calculated from the raw FA profile data, where the ratio of the F-statistic is used to compare variability observed among samples from different biological sources versus within-source variability. This test is ideal for the statistical analysis of FA profile data, as it may be applied to non-independent multivariable data and can handle situations where the number of variables is large relative to the number of observations (91, 92). Variables with mean and median weight percent values below 0.5% were eliminated from the FA profile data prior to performing this analysis. The semi-metric Bray-Curtis distance measure was used to calculate a matrix of pairwise distances between samples (93). As it is possible to determine significant differences among groups that are due to differences in dispersion as opposed to differences in location, non-metric multidimensional scaling (nMDS) of Bray-Curtis distances was used to visualize multivariate patterns among observations. nMDS was performed using the R MASS package, ver. 7.2-49 (94). Hierarchical cluster analysis was performed on the distance matrix obtained from unsupervised random forest proximities ($distance = 1 - \sqrt{proximity}$) as an exploratory technique to examine unsupervised groupings among the samples.

The random forest (RF) algorithm (48) was used to classify oil samples to their biological source using FA and TAG profile data, separately. RF has been shown to be a highly accurate and stable classifier that performs well when variables are not independent (95), as is often the case with FA profile data, and specifically for MS data (49). RF classifiers were generated using the R randomForest package, ver. 4.5-34 (51), using 5,000 decision trees. RF output provides the "out-of-bag" (OOB) error obtained by classifying samples which were not a part of the bootstrap data used to build a specific classification tree. This provides an unbiased estimate of the generalization error and eliminates the need for cross-validation (48). All RF classification results were checked against classification using the same sample data set with permuted group labels to assure that classification error was similar to what would be expected due to chance in the permuted model. Code for all procedures performed in R is available from the authors upon request.

RESULTS AND DISCUSSION

TLC analysis revealed that sample quality was intact; all samples were composed primarily of TAGs with the exception of two fish oils (Fish-2 and Fish-8), which were ethyl esters. FA profile analysis allowed identification of 98 individual FAME peaks; with 24 FAMEs having mean and median values greater than or equal to 0.5% by weight for all samples (n=48). These 24 FAMEs accounted for 93.2 ± 0.2% (mean ± SE) of total FA weight and were retained for statistical analyses. Data are listed by group in Table 6; medians and ranges were reported due to the presence of outliers for many variables. Individual FAs varied within and among the different groups (Figure 14); such variability is not surprising in commercial marine oils, as fish are harvested from different locations and different methods may be used to render and process oils (96). Many of the oils in the generic "fish" class are mixtures of different species, e.g. menhaden, sardines and anchovies (Table 5).

Table 6. Group median (range) FA weight percent data for 24 FAMEs with overall mean or median ≥ 0.5%.

	Cod liver oils (n=8)	**Fish oils (n=12)**	**Herring oils (n=6)**	**Salmon oils (n=8)**	**Seal oils (n=11)**	**Harp seal blubbers (n=3)**
14:0	4.1 (3.1-5.6)	7.3 (0.1-9.7)	6.4 (4.3-6.8)	5.5 (3.8-7.1)	4.6 (4.3-4.7)	4.0 (3.3-4.2)
16:0	11.7 (9.7-13.9)	16.1 (1.2-20.8)	10.0 (7.4-14.0)	13.7 (11.5-16.3)	8.0 (6.7-10.2)	5.3 (2.7-8.0)
16:1n-9*	0.6 (0.2-0.7)	0.6 (0.1-0.8)	0.5 (0.3-0.6)	0.6 (0.5-0.7)	0.5 (0.5-0.6)	0.6 (0.5-0.7)
16:1n-7	6.8 (3.9-9.8)	8.0 (0.5-11.4)	7.2 (6.8-10.7)	7.3 (4.6-8.5)	15.9 (12.6-16.9)	16.5 (10.6-19.3)
16:2n-4	0.5 (0.3-0.8)	1.3 (0.1-1.5)	0.5 (0.4-0.6)	0.7 (0.3-1.2)	0.6 (0.4-0.6)	0.7 (0.4-0.8)
16:3n-4*	0.6 (0.4-1.1)	1.7 (0.1-2.5)	0.5 (0.5-0.6)	0.8 (0.3-1.8)	0.4 (0.3-0.5)	0.6 (0.5-0.6)
16:4n-1	0.4 (0.3-1.1)	1.7 (0.1-2.7)	0.7 (0.5-0.9)	0.9 (0.3-2.5)	0.4 (0.3-0.5)	0.5 (0.2-0.6)
18:0	2.3 (2.1-3.2)	3.1 (2.2-3.8)	0.9 (0.7-1.5)	2.7 (2.1-3.3)	1.1 (0.9-1.4)	0.7 (0.6-1.4)
18:1n-11	1.4 (0.5-2)	0.1 (0.0-0.2)	0.6 (0.4-0.8)	0.8 (0.1-1.6)	4.4 (3.2-4.6)	4.4 (2.8-5.5)
18:1n-9	15.0 (10.8-18.5)	8.2 (3.8-9.3)	7.6 (4.4-9.1)	12.9 (8.2-19.1)	16.0 (13.7-21.5)	15.5 (15.4-20.7)
18:1n-7	3.2 (2.4-4.9)	2.8 (1.2-3.3)	2.0 (1.6-2.7)	3.1 (2.5-5)	4.3 (4.0-4.4)	4.7 (3.4-4.9)
18:2n-6	2.7 (1.5-21.5)	1.4 (0.7-1.9)	1.0 (0.6-2.2)	1.7 (1.3-3.2)	1.9 (1.5-2.1)	1.4 (1.3-1.5)
18:3n-3	1.1 (0.6-3.2)	0.7 (0.3-1.7)	0.4 (0.2-1.4)	0.7 (0.6-1.1)	0.6 (0.5-0.8)	0.4 (0.4-0.5)
18:4n-3	2.2 (1.3-2.6)	2.6 (1.0-3.2)	1.6 (0.8-4.3)	2.1 (1.1-2.7)	1.5 (1.2-2.2)	0.9 (0.7-1.1)
20:1n-11*	1.2 (0.7-1.5)	0.2 (0.1-0.6)	1.3 (0.8-1.9)	0.9 (0.1-9.1)	2.3 (1.9-2.6)	2.3 (1.6-3.9)
20:1n-9	7.4 (4.2-10.6)	1 (0.9-3.4)	16.3 (9.1-19.1)	3.7 (0.9-8.4)	9.0 (7.3-11.3)	8.7 (7.1-15)
20:4n-6	0.5 (0.3-0.7)	1.1 (0.8-1.9)	0.2 (0.1-0.3)	0.6 (0.3-1.1)	0.4 (0.4-0.5)	0.4 (0.3-0.4)
20:4n-3	0.9 (0.6-1.2)	0.9 (0.7-1.7)	0.4 (0.2-0.8)	0.9 (0.6-1.6)	0.5 (0.4-0.7)	0.3 (0.3-0.4)
20:5n-3 (EPA)	9.0 (4.8-11.4)	17.3 (8.5-34.5)	4.8 (3.6-7.8)	9.9 (8.5-17.9)	6.7 (6.0-8.2)	4.2 (3.5-8.5)
22:1n-11*	5.4 (3.8-9.3)	0.9 (0-3.4)	24.3 (12.3-33)	5 .0 (0.3-9.4)	2.1 (1.3-4.6)	3 .0 (1.9-3.9)
22:1n-9	0.7 (0.5-0.9)	0.2 (0.1-0.8)	2.2 (0.9-3.5)	0.6 (0.0-1.1)	0.5 (0.3-0.9)	0.6 (0.5-1.1)
21:5n-3	0.4 (0.3-0.6)	0.8 (0.5-2)	0.2 (0.0-0.4)	0.5 (0.3-0.8)	0.4 (0.4-0.5)	0.5 (0.3-0.6)
22:5n-3 (DPA)	1.5 (1.2-3.0)	2 .0 (1.7-5.0)	0.6 (0.4-0.7)	2.4 (1.5-3.8)	4.0 (2.4-4.3)	5.3 (3.2-8.1)
22:6n-3 (DHA)	10.1 (6.3-12.7)	11 (8.2-29.9)	2.7 (1.9-9.4)	9.9 (5.6-14.9)	8.4 (7.5-9.7)	8.6 (7.1-9.8)
Total	93.7 (92.6-95.3)	92.1 (87.5-93.9)	93.7 (93-94.5)	92.9 (91.4-93.7)	94.4 (94.0-94.9)	93.1 (93.0-93.8)

* Indicates possible coelution with another FA.

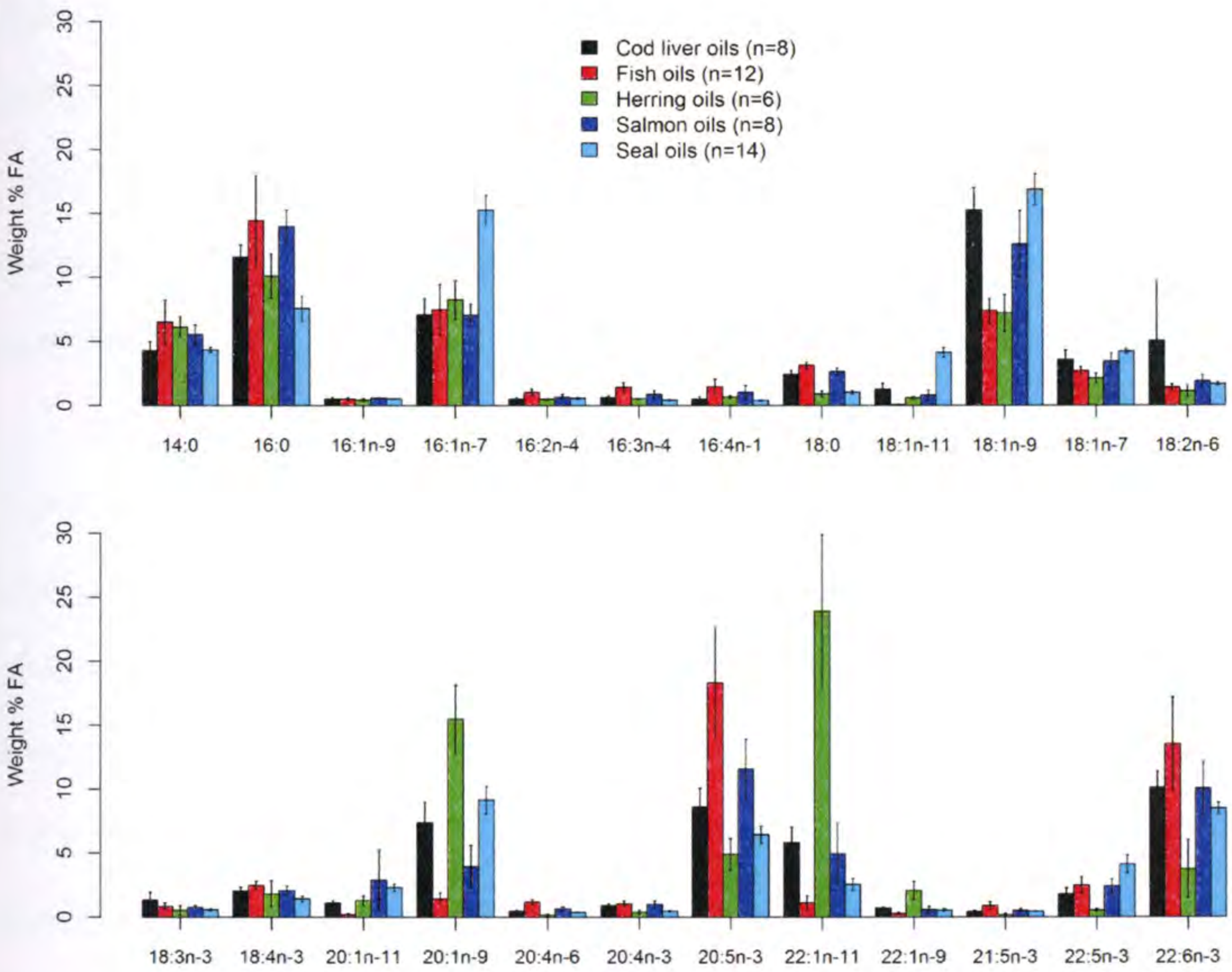


Figure 14. FA profile data (Mean ± 95% CI) for 24 FA with mean or median ≥ 0.5% (n=48). Note that seal oils represent combined data from extracted blubbers (n=3) and commercial oils (n=11).

Palmitic acid (16:0), oleic acid (18:1n-9), EPA (20:5n-3), and DHA (22:6n-3) were the predominant FA among the cod liver, "fish" and salmon oils. Herring oils were also high in palmitic acid, but herring oil FA profiles had lower levels of EPA and DHA and very high levels of 20- and 22-carbon monoenoic FAs (20:1n-9 and 22:1n-11). These long-chain monoenes were also high in some, but not all, cod liver and salmon oils. One cod liver oil (Cod-1) was notably different from the others within its group, with very high values for linoleic (18:2n-6, 21.5%) and linolenic (18:3n-3, 3.2%) acids.

The commercial seal oils and Harp seal blubbers were very similar to each other, with profiles dominated by high values for 16:1n-7 and 18:1n-9, and to a lesser extent 20:1n-9 and 22:6n-3. Multivariate analysis of variance using distance matrices verified that these two groups were not significantly different (F=2.34, R^2=0.1630, p=0.104), and could be combined into one group of seal oils (n=14).

Multivariate analysis of variance using distance matrices revealed an overall significant difference among the five groups, and between the seal oils and each of the fish oil groups (Table 7). Non-metric multidimensional scaling using Bray-Curtis distances from FA profiles verified that differences were due to location and not dispersion (Figure 15). While a great deal of variability and overlap was observed among and within the different fish classes, the seal oils grouped together and clearly separated from the fish classes in the two-dimensional nMDS plot. One cod liver oil (Cod-1) separated from the others in its class, likely due to the higher levels of 18:2n-6 and 18:3n-3 noted previously. Two "fish" oil samples (Fish-2 and Fish-8) separated from the other "fish" oils; these two samples were composed of ethyl esters rather than TAGs and had extremely high levels of EPA (>30%) and DHA (>20%). One herring oil (Herring-6) grouped with most of the cod liver and salmon oils; the other five herring oils grouped together and were separated from all other samples. The characteristics attributed to commercial seal oils, including high 16:1n-7/16:0, high 18:1n-11, and high 22:5n-3, were present in all of the seal oil and blubber samples and differed significantly from each of the fish oil groups in pairwise one-sided Wilcoxon tests (Table 8).

Table 7. Multivariate analysis of variance using distance matrices for FA profile data.

	F	R^2	p
Overall (5 classes)	21.17	0.6633	0.001
Seal vs. Cod	20.18	0.5022	0.001
Seal vs. "Fish"	39.49	0.6220	0.001
Seal vs. Herring	58.85	0.7658	0.001
Seal vs. Salmon	28.72	0.5895	0.001

Note: 24 FAs with mean or median weight % values > 0.5%.

Table 8. P values from pairwise one-sided Wilcoxon tests.

H_A	16:1n-7/16:0	18:1n-11	22:5n-3 (DPA)
Seal > Cod	3.13×10^{-6}	7.56×10^{-5}	2.19×10^{-5}
Seal > "Fish"	1.04×10^{-7}	8.54×10^{-6}	8.27×10^{-4}
Seal > Herring	2.58×10^{-5}	3.07×10^{-4}	2.58×10^{-5}
Seal > Salmon	3.13×10^{-6}	7.56×10^{-5}	4.25×10^{-4}

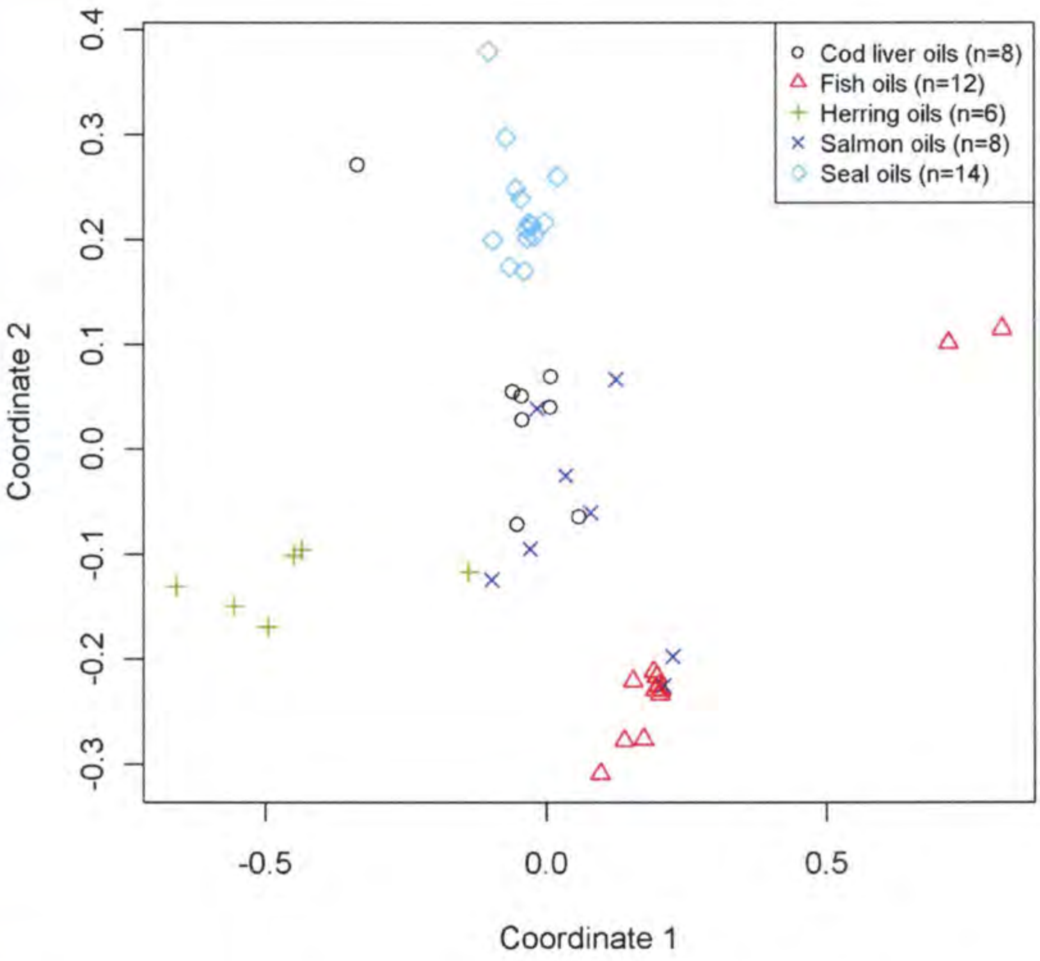


Figure 15. Non-metric multidimensional scaling plot of Bray-Curtis distances from FA profile data (n=48).

Hierarchical cluster analysis of the FA profile data using unsupervised random forest proximities resulted in a primary split between seal oils and fish oils, and then split the fish oils into four overlapping groups containing "fish"/salmon, salmon/cod liver, salmon/cod liver/herring, and herring oils (Figure 16). Sample groupings were similar to those observed in the nMDS plot (Figure 15). A random forest classifier was used to classify the oil samples to their biological source using 24 FA with an out-of-bag error estimate of 26.2% (Table 9). The algorithm was unable to classify the fish samples to their respective classes, likely due to the high variability observed within groups and similarity among the fish groups. The seal samples were classified with 0% class error and none of the fish samples were classified as seals, indicating that FA profile analysis may be used to identify seal oils for forensic purposes.

Analysis of TAGs using RP-HPLC/APCI-MS was implemented as a secondary step to determine whether the oils were natural (non-synthetic) marine oils and to verify the results of the FA profile analysis. Because of the complex nature of marine oils, the actual TAG molecular species composition cannot be determined using available technologies (10); it is therefore extremely unlikely that a synthetic oil would meet the specifications of this analysis. Manual inspection of the LC/MS chromatograms and spectral data indicated that the oils were complex mixtures of TAGs. Preliminary processing of all samples simultaneously with XCMS resulted in 1886 peaks. It is important to note that many of these peaks are redundant, as each "peak" is defined by its (RT, m/z) coordinates in an image, as opposed to the two-dimensional peaks we use for

FA profile analysis that represent individual chemical components. One TAG molecular species may yield several highly correlated ions in this analysis (e.g. the protonated molecule and up to three DAG fragment ions and corresponding isotope peaks), and each is processed as a separate peak using XCMS. The peak list was first filtered by screening for only DAG ions and protonated molecules expected from TAGs resulting from combinations of the 24 FA selected in the FA profile analysis, and then using mRMR feature selection. The TAG screening step resulted in 669 peaks, and we used the first 24 variables selected by the mRMR algorithm as inputs to the random forest, such that results could be compared with those from the classification using FA profile data.

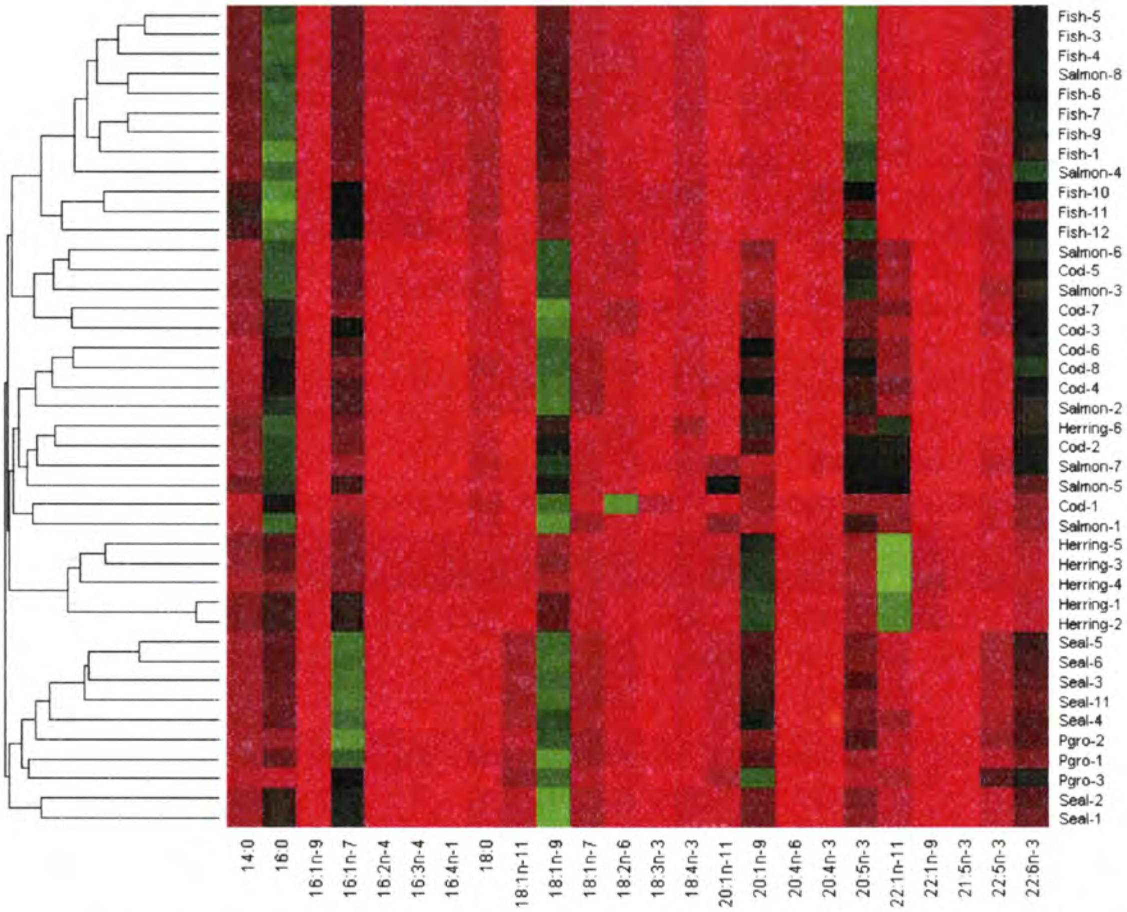


Figure 16. Dendrogram and heatmap of cluster analysis from unsupervised random forest, FA profile data (n=42).

Table 9. Random forest classification results for FA profile data.

Actual	Predicted					
	Cod	"Fish"	Herring	Salmon	Seal	Class error
Cod	6	0	0	2	0	0.2500
"Fish"	0	9	0	1	0	0.1000
Herring	1	0	5	0	0	0.1667
Salmon	5	2	0	1	0	0.8750
Seal	0	0	0	0	10	0.0000

Notes: OOB estimate of error rate: 26.19%. Weight % values for 24 FAs with mean or median > 0.5%; n=42 (Fish-2 and -8, and Seal-7, -8, -9, and -10 were not included, see Table 5 for explanation).

Table 10. Random forest classification results for TAG data.

Actual	Predicted					
	Cod	"Fish"	Herring	Salmon	Seal	Class error
Cod	7	0	0	1	0	0.1250
"Fish"	0	10	0	0	0	0.0000
Herring	0	0	5	1	0	0.1667
Salmon	3	2	0	3	0	0.6250
Seal	0	0	0	0	10	0.0000

Notes: OOB estimate of error rate: 16.67%. Normalized peak areas for 24 TAG peaks; n=42 (Fish-2 and -8, and Seal-7, -8, -9, and -10 were not included, see Table 5 for explanation).

The random forest classified the oil samples to their biological source with an out-of-bag error estimate of 16.7% (Table 10). As with the prior FA profile data, the random forest could not discriminate well among the fish samples, but seal samples were classified with 0% class error and there were no false positives, or samples classified incorrectly as seal oils. Multidimensional scaling plots of the random forest proximities for FA profile and TAG RP-HPLC/APCI-MS data are shown in Figure 17. The seal samples separate clearly from the fish groups, but as in the previous analyses, some overlap can be observed among the fish groups. It is noteworthy that the FA and TAG plots are nearly identical when Coordinate 2 is inverted on either plot; this indicates that the underlying structure of the classifiers, as indicated by the sample proximities, is similar even though the data are based on different measurements. Both of the lipid profiling methods, FA by GC/FID and TAG using LC/MS may be used to successfully identify commercial seal oils.

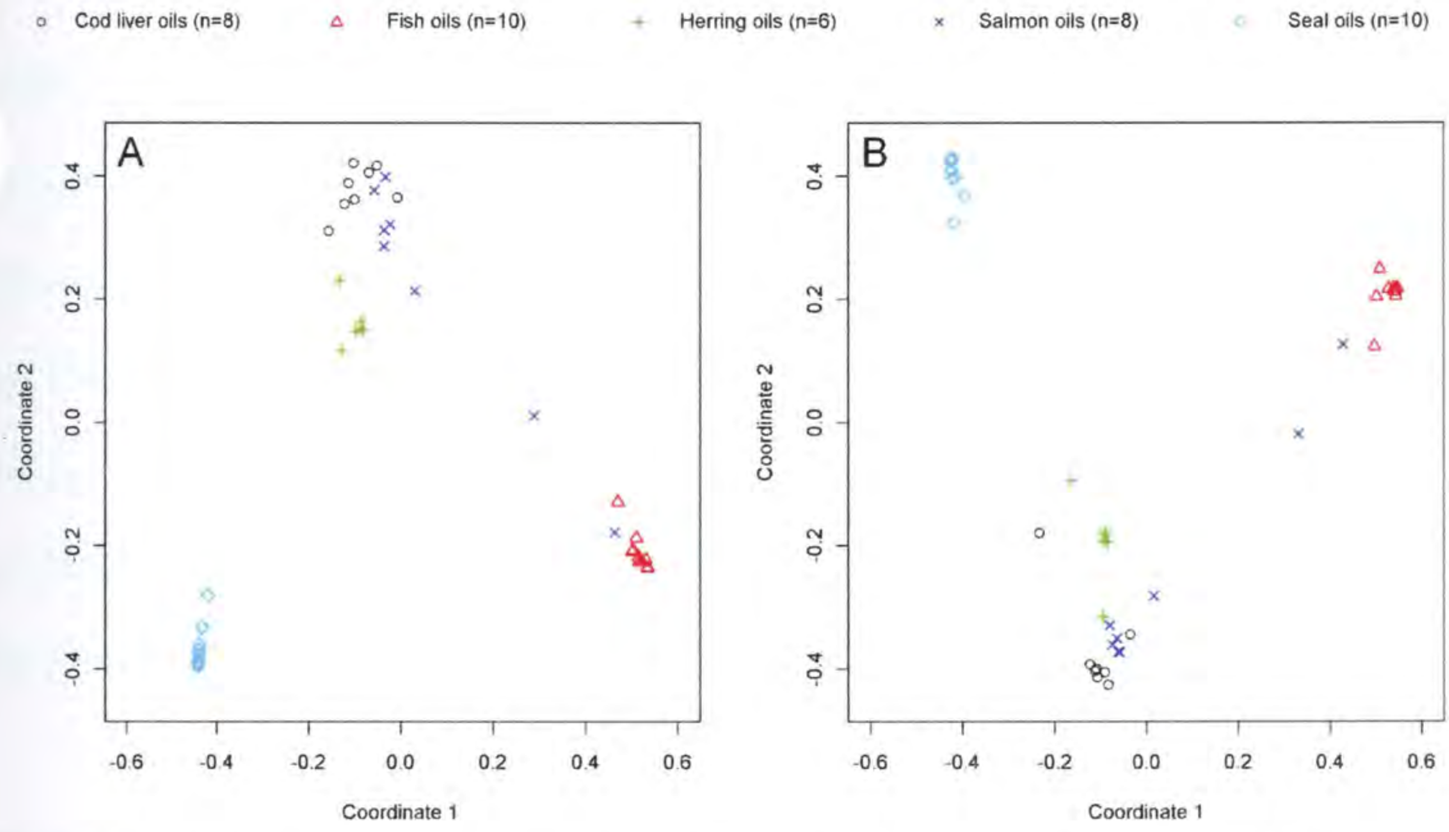


Figure 17. Multidimensional scaling plots of random forest proximities for (A) FA profile data and (B) TAG RPHPLC/APCI-MS data.

FA profiling methodologies are well-established and provide consistent data for relative quantitation of individual FA that may be identified using retention time and mass spectral data along with known standards for verification. Thus, a database of FA profile information may be stored and used for classification of unknown samples as needed. Due to the semi-quantitative nature of the TAG RP-HPLC/APCI-MS analysis, a set of standard oils must be analyzed with any unknown samples and a classifier must be built from data obtained in real-time. As true amounts of individual TAG molecular species are unknown, the peak values cannot be verified and standards and samples should be analyzed together and in random order with gradient blanks run in between each data collection.

It is important to note that this study is based on commercially available oil supplements that were purchased or obtained from academic and government institutions and private companies. The biological source for many samples was determined from information provided on the sample label. As many of the samples are not authentic forensic standards, it will be important to continue this research with appropriate fish and seal oil standards. The difficulty of obtaining authentic standards is a recurrent problem in wildlife forensics, and particularly when working with species like marine mammals that cannot be legally taken in the US.

We propose a two-tiered analysis to accurately identify seal oil dietary supplements for forensic purposes. While FA profile analysis is sufficient to conclude that a suspect sample is consistent with seal oil, it does not eliminate the remote possibility of a synthetic oil mixture that was made to mimic the FA composition of seal oil. Natural marine oils comprise a mixture of TAGs that is considerably more complex than the FAMEs observed when the acyl chains are transesterified in FA profiling experiments. Moreover, the underlying structures and quantities of individual TAG molecular species in marine oils are not known. Thus, LC/MS of TAGs provides an additional analysis step that will reliably identify omega-3 dietary supplements that were originally harvested from seals, and should be used to confirm any unknown samples that are identified as seal oils using FA profiling. Because seal products are illegal in the US and EU, this methodology may be useful for international law enforcement purposes. This research provides a solid foundation for the development and validation of a method to enforce existing laws prohibiting the possession of seal oil dietary supplements in the US and EU.

PAPER 3: DIFFERENTIAL PROFILING OF ADIPOSE TISSUE TRIACYLGLYCEROLS IN MICE FED MILKFAT- AND LARD-BASED HIGH FAT DIETS

Margaret H. Broadwater[1,2], Tuoyu Geng[2], L. Ashley Cowart[2,3], and John H. Schwacke[2]

[1] National Oceanic and Atmospheric Administration, Center for Coastal Environmental Health and Biomolecular Research, 219 Fort Johnson Road, Charleston, SC 29412.

[2] Medical University of South Carolina, Department of Biochemistry and Molecular Biology, 173 Ashley Avenue, Charleston, SC 29425.

[3] Ralph H. Johnson Veteran's Affairs Medical Center, 109 Bee Street, Charleston, SC 29403.

ABSTRACT

Lipid profiles of diets and adipose tissue from mice fed milkfat- and lard-based high-fat diets were studied with the objective to determine whether differences exist between adipose tissue lipid profiles from mice on the different diets, and whether the profiles resemble those of the diets. Palmitate (16:0) and oleate (18:1n-9) were the most prominent fatty acids in feed and adipose tissue samples. Multivariable analysis based on eight fatty acids indicated significant differences between diets ($p < 0.05$) and adipose tissue sampled at eight and sixteen weeks from mice fed different diets ($p < 0.005$). The milkfat-based diet was higher in medium chain (< 12 carbons) and saturated fatty acids and lower in unsaturated fatty acids than the lard-based diet. Overall trends observed in diet FA profiles were reflected in adipose tissue composition. Multivariable analysis

based on twenty triacylglycerol species in adipose tissue sampled from mice fed different diets revealed differences between diet groups at both time points ($p < 0.005$), and we determined the structures of triacylglycerol species that differed between groups. As differences in adipose tissue lipid composition may result in differential plasma fatty acid composition and ultimately affect the availability of specific fatty acids to peripheral tissues, high levels of saturated fatty acids in adipose tissue may result in a chronic disease phenotype.

INTRODUCTION

The increasing prevalence of obesity and obesity-related disease in humans worldwide has fueled much new research into the mechanisms by which obesity abets pathological outcomes. Rodent models of diet-induced obesity (DIO) have been demonstrated to be useful in the study of human disease, and particularly for the combination of insulin resistance, abnormal blood lipid levels, hypertension, and central obesity known collectively as the metabolic syndrome (MetS) (97-99). MetS is associated with an increased risk of type 2 diabetes and non-alcoholic fatty liver disease; the primary clinical outcome associated with MetS is morbidity and mortality due to atherosclerotic cardiovascular disease (CVD). As MetS risk factors are modifiable, dietary interventions are recommended for both prevention and treatment (100).

The relationship between dietary fat, obesity, and human disease is well-established, but recently the focus in human nutrition has shifted from the quantity of dietary fat to its quality (101). There is a general consensus that saturated and *trans* fatty acids (SFAs and TFAs) increase disease risk, while monounsaturated and polyunsaturated fatty acids (MUFAs and PUFAs) have a protective effect on both CVD and type 2 diabetes (102, 103). Epidemiological studies, as well as recent laboratory studies using rodent models and cell cultures, have shed light on the diverse effects individual FAs have on MetS risk factors, and specifically on the counteracting roles of SFAs and MUFAs in the development of insulin resistance (11, 102, 104-106).

Elevated plasma lipids resulting from obesity lead to increased levels of ceramide and diacylglycerol (DAG) in liver, heart, pancreas, skeletal muscle, and adipose tissues. These bioactive lipid mediators mechanistically link dietary SFAs to insulin resistance and other pathologies associated with the obese state (107). Hu et al. (104) observed that the MUFA oleate (18:1n-9) attenuated ceramide production induced by the SFA palmitate (16:0). These two FAs differentially regulated dihydroceramide desaturase 1 (DES1) at the mRNA level; up-regulation of DES1 by palmitate may drive sphingolipid-modulated insulin resistance. Additionally, stearate (18:0), which differs in chain length from palmitate by only two carbons, had no effect on DES1, leading the researchers to conclude that individual FAs function as discrete chemical species and can mediate distinct actions (104). As the fats ingested by humans and the rodents used to model human disease are primarily mixtures of triacylglycerols (TAGs) that comprise a variety of FAs differing in chain length and degree of unsaturation, it is noteworthy that the composition of dietary fat may profoundly influence the results observed in scientific studies (103, 107).

Rodent DIO models have employed a variety of high-fat diets from animal and plant sources (e.g. lard, butter, milkfat, coconut fat, corn oil, safflower oil) with relative fat ranging 20-60% of total energy; these diets have very different FA compositions and have led to considerable variability in reported results (99). Geng et al. (11) compared mice fed a novel milkfat-based diet (MD) with a more traditional lard-based diet (LD) and a low-fat, isocaloric control diet (CD). Mice fed the MD developed an obese, severely insulin-resistant phenotype compared to both LD and CD, but did not exhibit increased levels of DAG or ceramide in muscle or liver tissues. Instead, the MD (and LD, to a lesser extent) promoted expression of the mammalian homolog of drosophila tribbles 3 (TRIB3), which binds to and prevents phosphorylation of PKB/AKT in response to insulin (108). The authors demonstrated dose- and time-dependent differential expression of TRIB3 in response to different FAs, reiterating the distinct actions of individual dietary FAs. SFAs (14:0 and 16:0), which were higher in the MD than LD, promoted expression of TRIB3, and UFAs (18:1n-9 and 18:2n-6), which were lower in the MD than LD, attenuated SFA-induced TRIB3 expression. Postprandial plasma non-esterified FA (NEFA) composition in MD- and LD-fed mice was affected by diet. Thus, the dietary FA composition in rodent DIO studies may affect the phenotype observed and this may occur via different mechanisms. Over time, a high-SFA diet may affect the composition of fat storage depots and contribute to a chronic disease phenotype.

It is generally accepted that the relative availability and storage of FAs in tissues depends on dietary FA composition (103, 109). A simplified model of the fate of dietary FAs is depicted in Figure 18. Dietary TAGs are digested by lipases, releasing FAs and monoacylglycerol (MAG). The digestion process differs for medium-chain vs. long-chain TAGs, as shown in Figure 19. FAs from medium-chain TAGs are shuttled to the liver via the portal vein, while those from long-chain TAGs (containing FAs with ≥ 12 carbons) are reconstituted as TAGs and packaged in chylomicrons, entering the circulation via lymph after a meal (110, 111). TAGs in postprandial plasma are delivered to adipose tissue, liver, and peripheral tissues, where lipoprotein lipase (LPL) facilitates transport of FAs across the cell membrane. Within the cell, FAs are used for energy or reconstituted as TAGs for storage. FA and TAG deposition in adipose tissue is selective and depends on the diet (112). When energy levels are low, e.g. fasting, TAGs are mobilized from adipose tissue stores; FAs are released in three steps by desnutrin/adipose triglyceride lipase (ATGL), hormone-sensitive lipase (HSL) and monoglyceride lipase (MGL) (113). This process, termed lipolysis, is also selective and is based on substrate availability to lipolytic enzymes, which depends on the molecular polarity of TAG molecules within the lipid droplet (114). NEFAs cross the cell membrane and are bound to serum albumin for delivery to the liver and other tissues via the circulatory system (Figure 18). Thus, postprandial plasma FAs reflect a recent meal, while FAs in adipose tissue and fasting plasma are representative of FAs in the diet over a longer period of time. The relationship between dietary, adipose tissue, and plasma FAs is further complicated by the convergence of metabolic pathways; *de novo* FA biosynthesis of palmitate by the fatty acid synthase complex occurs when an excess of acetyl-CoA is present from the glycolytic pathway, e.g. dietary carbohydrates (115). Palmitate and dietary fatty acids may also undergo elongation and desaturation, as shown in Figure 20.

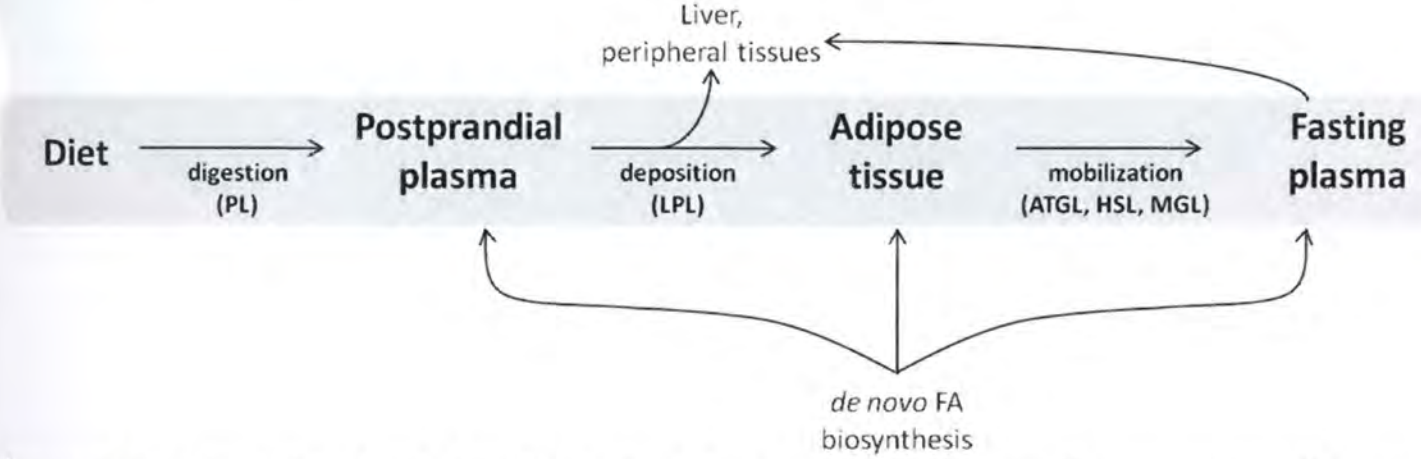


Figure 18. Simplified model of the fate of dietary fatty acids. FAs are released from dietary TAGs by lipases including pancreatic lipase (PL) during digestion, and reconstituted as TAGs and packaged in chylomicrons for circulation via lymph and plasma during the postprandial state. These TAGs are delivered to the adipose tissue, liver, and peripheral tissues, where lipoprotein lipase (LPL) facilitates breakdown for transport across the cell membrane. The FAs are stored as TAGs in the form of a lipid droplet inside the adipocyte. During the fasting state, TAGs are mobilized from the adipose tissue and broken down into FAs and glycerol by desnutrin/adipose triglyceride lipase (ATGL), hormone-sensitive lipase (HSL) and monoglyceride lipase (MGL); the non-esterified FAs are bound to serum albumin for delivery to the liver and other tissues via the circulatory system.

Adipose tissue, once thought to be an inert storage depot for lipids, is now known to be a complex organ with endocrine and paracrine functions. Adipose tissue comprises adipocytes and other cell types that serve to maintain homeostasis. Adipocytes have a unique organelle, the lipid droplet, that stores FAs in the form of TAGs and can account

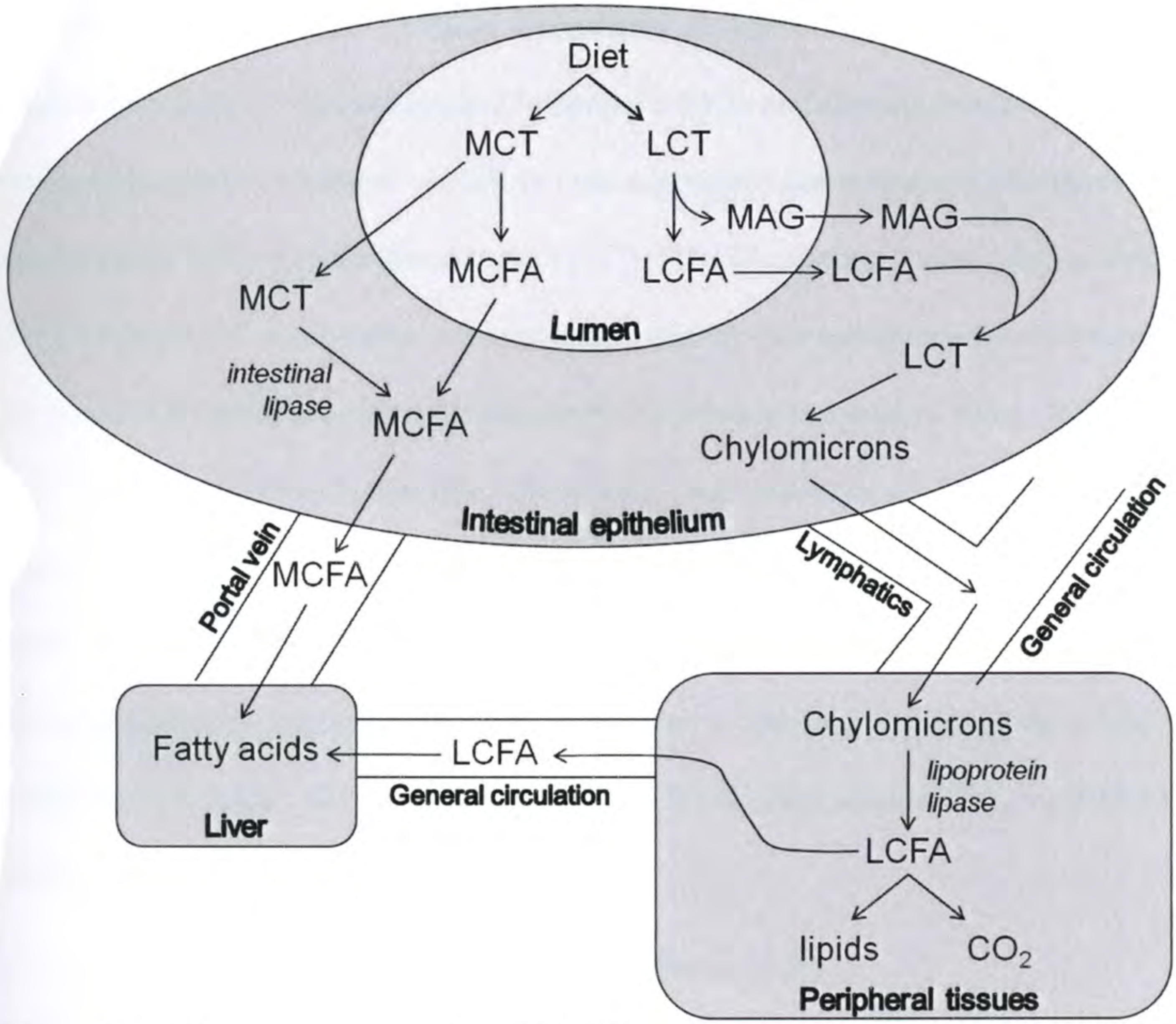


Figure 19. Schematic representation of the different paths of dietary medium-chain (MCT) and long-chain triacylglycerols (LCTs), adapted from Bach and Babayan (111). MCTs and LCTs are both hydrolyzed to release fatty acids (MCFAs and LCFAs) by pancreatic lipase in the lumen of the intestine. MCT hydrolysis is faster and more complete, and MCFAs diffuse rapidly across the intestinal epithelium. MCTs are also absorbed as TAGs, and can undergo hydrolysis by an intestinal lipase within enterocytes. MCFAs leave the intestine via the portal vein and are transported to the liver as non-esterified FAs bound to serum albumin. LCTs must be hydrolyzed to LCFAs and monoacylglycerols (MAG) and organized as micelles to cross the epithelium. A fatty acyl-CoA synthetase specific for FAs with 12 or more carbons converts LCFAs to fatty acyl-CoAs, and TAGs are resynthesized from fatty acyl-CoAs and digested MAGs. The TAGs are packaged in chylomicrons with dietary cholesterol and fat-soluble vitamins, which are then released into the lymph and then the blood for distribution to peripheral tissues for storage and/or energy.

for 95% of cell mass (116). Adipose tissue can be thought of as a buffer for the flux of plasma FAs, both as chylomicron and lipoprotein TAGs and albumin-bound NEFAs. As such, it suppresses the release of NEFAs from adipocytes and increases TAG clearance from plasma during the postprandial state (117). The FA composition of adipose tissue reflects dietary FA composition over the long term, and thus influences the composition of depot FAs available to tissues as plasma NEFAs during low energy states. Release of FAs from adipose tissue is selective. More polar TAG molecular species comprise more polar FAs, i.e. FAs with fewer carbons and more double bonds; these molecules have better access to lipolytic enzymes and are thus more easily hydrolyzed (114). Because the composition of adipose tissue affects the release of FAs into the bloodstream, it is important to consider TAG molecular species composition of adipose tissue in addition to the FA profile (112).

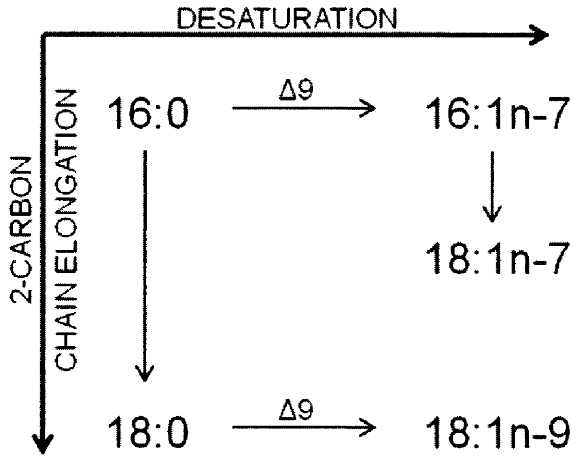


Figure 20. Mammalian fatty acid pathway, adapted from Cook (118). Palmitate (16:0) is the product of the *de novo* biosynthesis by the fatty acid synthase complex. Elongation and desaturation of dietary and biosynthetic palmitate are carried out by other enzyme systems. Elongase enzymes add 2 carbons to the acyl chain at the carbonyl end of the molecule; a specific Δ9-desaturase enzyme can add a double bond to form palmitoleate (16:1n-7) and oleate (18:1n-9) from palmitate and stearate (18:0), respectively.

Analysis of fatty acid methyl ester (FAME) derivatives is commonly performed using gas chromatography with flame ionization detection (GC/FID) to determine the FA profile, or the relative amounts of fatty acids, in a lipid extract. FAME peaks are identified using retention times (RTs) of known standards and mass spectral data (i.e. concurrent analysis by GC/MS). This well-established procedure is quantitative, and can be validated using standard mixtures or samples with known composition to assure data quality (5). TAG molecular species analysis using reversed-phase high performance liquid chromatography with atmospheric pressure chemical ionization mass spectrometry (RP-HPLC/APCI-MS) is a more recent analytical development, and can be used to obtain semi-quantitative TAG profile data (119). Using RP-HPLC, TAG molecular species elute depending on polarity, ordered by equivalent carbon number (ECN), which is approximately equal to the number of acyl carbons minus twice the number of carbon-carbon double bonds ($C - 2 \cdot DB$). TAGs with the same ECN are termed "critical pairs" and elute together (5). APCI-MS is a 'soft' ionization technique that produces relatively simple spectra from TAGs with base peaks consisting of either the protonated molecule, $[M+H]^+$, or diacylglycerol ions, $[M-RCO_2]^+$ or $[DAG]^+$, that result from the loss of a FA moiety (120). The presence and intensity of the protonated molecule is related to the saturation of FA moieties attached to the glycerol backbone of the TAG molecule, with the $[M+H]^+$ ion intensity increasing with the number of double bonds in the molecule. The protonated molecule may not be observed in TAGs with three saturated acyl chains (27).

RP-HPLC/APCI-MS allows for partial identification of the FAs (number of carbons and double bonds, e.g. 18:2) from individual TAGs in a mixture and also can provide information on the positions of FAs on the glycerol backbone in the TAG molecule (27). RP-HPLC/APCI-MS data may be treated as a three-dimensional image (RT × m/z × intensity); raw instrument data must be subjected to preprocessing steps including peak detection, alignment, and normalization prior to statistical analysis (39). The metabolomics software program xcms (9) is a freely available tool that runs in the R statistical program (87) and combines matched filtration peak detection, nonlinear chromatographic alignment, and peak matching in one package. As xcms identifies all peaks present in the samples, it is necessary to filter xcms output data specifically for TAG peaks. TAGs have an advantage over many other molecules in that their structures vary only by the FA moieties attached to the glycerol molecule. Knowledge of FA profile data for a sample or group of samples enables the prediction of all possible TAG molecules based on FAs present. Knowing TAG molecular structure, we can make a list of $[M+H]^+$ and $[DAG]^+$ peaks that may be observed and recursively search the xcms output data for these specific peaks. We can use feature selection methods to search specifically for peaks that differ between sample groups, and return to the original RP-HPLC/APCI-MS data to identify the TAG molecules that these peaks represent. These processing steps have been combined as an integrated framework for the analysis of natural TAG mixtures using RP-HPLC/APCI-MS (Appendix), where raw data in the form of mzXML, mzData, or mzML files undergo preprocessing with xcms, statistical analysis, and structure determination within the platform of the R statistical program (87).

We examined lipid profiles of diets and adipose tissue from mice fed milkfat- and lard-based high-fat diets (MD and LD) and an isocaloric low-fat control diet (CD) at eight and sixteen weeks. The objectives of this study were to determine whether differences existed between adipose tissue lipid profiles from MD- and LD-fed mice, and whether the profiles resembled those of the diets. Theoretically, differences in adipose tissue lipid composition result in differential plasma FA composition and ultimately affect the availability of specific FAs to peripheral tissues. High SFAs in adipose tissue may result in a chronic disease phenotype. We know that the MD-fed mice are more obese and more insulin-resistant than LD-fed mice (11); and hypothesize that FA profiles of adipose tissue from these mice will reveal that MD-fed mice have higher palmitate (16:0) and lower oleate (18:1n-9), reflecting dietary FA composition. We used a targeted metabolomics approach, described in detail in the Appendix, to determine differences in adipose tissue TAGs from MD- and LD- fed mice by searching the RP-HPLC/APCI-MS data image for masses of interest and identifying specific features that differed between the diet groups.

METHODS

C57bl/6J male mice maintained on high-fat lard (LD), milkfat (MD), or isocaloric low-fat control diet (CD) for eight or sixteen weeks (11). Adipose tissue was harvested at eight- and sixteen-week time points. To harvest adipose tissue, mice were euthanized with isoflurane (Hospira, Inc., Lake Forest, IL) and cervical dislocation. Abdominal cavity was subsequently cut open, and epididymal white adipose tissue was then carefully separated by scissors from each mouse and immediately frozen in liquid nitrogen. All tissues were transferred to -80°C freezer for long-term storage. Animal protocols were reviewed and approved by the Institutional Animal Care and Use Committee of the Medical University of South Carolina and the VA Medical center in accordance with the *Guide for the care and use of laboratory animals* (NIH Publication No. 86-23, revised 1996). Burdick and Jackson solvents (VWR, West Chester, PA) were HPLC-grade or the highest purity available, and were used without further purification.

Lipid extraction

Lipids were extracted from diets and adipose tissue samples in chloroform-methanol (2:1) using the method of Folch et al. (121). Diet pellets (2.4-3.9 g) were softened in 2 mL water, then extracted in a volume of chloroform-methanol (2:1, 25 mg/L BHT) 20 times the sample volume. Diet extracts were filtered and a volume of aqueous sodium chloride (0.73%) was added to achieve a ratio of chloroform-methanol-water of 8:4:3. After thorough mixing, diet extracts were stored at 4°C overnight to facilitate phase separation. The aqueous layer was removed and discarded. Frozen adipose tissue samples (~ 40 mg) were transferred to glass conical vials and 0.4 mL methanol and 0.8 mL chloroform (containing 25 mg/L BHT) were added in sequence, topped with nitrogen, vortexed 10 s and sonicated for 15 min, then left at room temperature for 24 hours. Adipose extracts were again sonicated for 15 min and vortexed 10 s, then 0.29 mL aqueous sodium chloride (0.73%) was added, then vortexed again and centrifuged 3 min at 1300 rpm to induce phase separation. The aqueous layer was removed and discarded. The adipose lipid in chloroform was dried with anhydrous sodium sulfate and allowed to settle for 15 min, then filtered.

FA profile analysis

Fatty acid methyl esters (FAMEs) were prepared according to Metcalfe et al. (122). Briefly, 25 mg of extracted lipid was dissolved in 1 mL hexane, to which 1.5 mL sodium hydroxide in methanol (0.5 N) were added. The samples were vortexed for 10 s and heated to 100°C for 5 min, then cooled in a water bath. Two mL boron trifluoride in methanol (10%) were added, and samples were vortexed for 10 s, heated to 100°C for 20 min, and cooled in a water bath. Addition of one mL hexane was followed by vortexing for 10 s, and washing with basic saturated sodium chloride. Samples were centrifuged 3 min at 1300 rpm and the hexane layer removed and diluted to approximately 0.7 mg/mL FAMEs in hexane for fatty acid profile analysis using gas chromatography with mass spectrometry (MS) and FID. FAMEs were analyzed on an Agilent 6890 gas chromatograph with splitless injection equipped with FID and a 5973 MS (Agilent Technologies, Inc., Palo Alto, CA). With the use of dual injection, each sample was simultaneously analyzed on two DB225-MS columns (50%-cyanopropylphenyl-50%-methylpolysiloxane, 30 m × 0.25 mm, J&W Scientific Inc., Folsom, CA, USA). Separation was achieved with oven temperature programming as follows: 50°C held for 2 min, ramped at 20°C/min to 150°C followed by a 1°C/min ramp to 220°C. Mass spectra were used to identify individual FAME peaks, in conjunction with comparison of retention times with those of known standards. Empirical correction factors determined from quantitative standards (GLC-85, 411, 566, and 617; NuChek Prep, Elysian, MN) were applied to integrated peak areas from the FID chromatogram and compositions reported as weight percent of fatty acids (85). Sample order was randomized prior to analysis.

A combination of univariate and multivariable statistical methods were used to evaluate differences in FA profiles between the milkfat and lard diets and adipose tissue samples harvested from the MD- and LD-fed mice at eight and sixteen weeks. The eight most prominent FAs in adipose tissue samples harvested from MD- and LD-fed mice were used for all statistical analyses. Univariate t-tests were used to evaluate the null hypothesis of no difference between MD and LD using a Bonferroni-adjusted Type I error rate to account for multiple tests ($\alpha = 0.05/8 = 0.00625$). A nonparametric permutational MANOVA was performed on the Bray-Curtis distance matrix using the *adonis* function in the R vegan package, ver. 1.17-0 (90). As it is possible to determine significant differences among groups that are due to differences in dispersion as opposed to differences in location, non-metric multidimensional scaling (NMMDS) of Bray-Curtis distances was used to visualize multivariate patterns among observations. All significant results from nonparametric permutational MANOVA tests were confirmed using NMMDS plots. NMMDS was performed using the R MASS package, ver. 7.2-49 (94).

Pattern similarities in FA composition between diet and adipose tissue from MD-, LD-, and CD-fed mice were calculated using the cosine similarity measure (123).

TAG profile analysis

Lipid extracts, composed primarily of TAGs, were analyzed on an Agilent 1100 quaternary pump HPLC system and Agilent XCT ion trap MS equipped with APCI source (Agilent Technologies, Palo Alto, CA). Detector optimization was performed using trilinolein (NuChek Prep, Elysian, MN; 5 µg/mL in 2:1 acetone-acetonitrile). Lipid extracts were diluted to approximately 1 mg/mL in acetone-acetonitrile (2:1). TAGs were separated on a Restek Allure C18 column (5 µm, 250 × 2.1 mm, Restek Corporation, Bellefonte, PA) with a two-stepped linear gradient of acetone in acetonitrile at flow rate 0.6 mL/min. Solvents contained 0.1% acetic acid to facilitate ionization. Acetone concentration was held at 20% for 1 min, stepped to 66% at 4 min and held for 13.5 min, then stepped from 66% to 90% in 1 min and held at 90% until 45 min, adapted from Jakab et al. (4). Autosampler and column temperatures were 20°C and 35°C, respectively. The injection volume was 3 µL. Direct infusion MS was performed in Ultrascan mode with the following parameters: APCI temperature, 350°C; vaporizer temperature, 500°C; corona current, 5000 nA; nitrogen sheath and auxiliary gas, 60 psi and 7 L/min, respectively. Mass spectra were collected in positive ion mode from mass-to-charge ratio (m/z) 100-1200 with a scan time of 300 ms. Samples were analyzed in random order with a gradient blank run between each analysis to prevent systematic differences in LC/MS profiles over the time of the analysis and column contamination.

LC/MS data files were converted from the Bruker MS proprietary format to mzXML using CompassXport (Bruker Daltonics, Ver. 1.3); OpenMS TOPPView (10; http://open-ms.sourceforge.net, Ver. 1.2) was used to crop the LC/MS image to m/z 250-900 and RT 300-1800 s; files were imported into TOPPView in mzXML (40) and exported as mzData files for processing with the xcms package (9; Ver. 1.14.1) in R (87; Ver. 2.9.1). Matched filter peak detection was used with the following parameters: sigma = 8, max = 25, step = 0.1, steps = 3, mzdiff = 0.7. Peaks were grouped together across samples using fixed-interval overlapping m/z bins (mzwid = 0.25) and calculation of smoothed peak distributions in chromatographic time using a Gaussian kernel density estimator (bw = 10). RTs were corrected for all samples simultaneously using loess regression to model nonlinear RT deviation contour profiles based on peak group median RTs and deviations from the median. The corrected peak lists were re-grouped using a smaller Gaussian kernel (bw = 5), and samples with missing peak values within a group were filled in by integrating raw data in the peak group region using corrected RT start and end points defined by the peak group medians. A matrix of peak area values with rows for every group, indexed by m/z and RT, and columns for every sample was generated. Samples with multiple peaks per group were resolved by choosing the peak closest to the median RT. Xcms does not provide a function for normalization, so this was performed manually by dividing each peak by the total area for each sample.

We implemented a targeted variable selection approach to identify relevant features in the xcms output, and then searched for peaks that differed between MD- and

LD-fed mice. Using our knowledge of the FA profiles of the adipose tissue samples, we used the most prominent FAs to generate a list of TAG molecular species structures that were likely to be observed; FAs are listed in Table 11. The expected ions from these structures were predicted and used to recursively search the xcms feature list for relevant peaks, or those likely to represent TAG analytes. Because the presence and intensity of the $[M+H]^+$ ion depends on the degree of unsaturation in the acyl chains of individual TAG molecules (27), we chose to limit our search to $[DAG]^+$ ions. The resulting peak list was reduced to twenty variables by using q-values to rank features according to differences in integrated peak values between MD and LD groups. The q-value is based on the false discovery rate, and provides a measure of each variable's significance, accounting for the fact that many variables are being tested simultaneously. A low q-value signifies a low probability that a feature was falsely deemed significant (45). The qvalue package in R was used to obtain q-values from p-values generated by individual t-

Table 11. FAs used for TAG screening. Note that we cannot differentiate among isomers because masses were used as search criteria.

Common name	Abbreviation	C:DB	ECN*
Laurate	La	12:0	12
Myristate	M	14:0	14
Palmitate	P	16:0	16
Stearate	S	18:0	18
Palmitoleate	Po	16:1	14
Oleate	O	18:1	16
Linoleate	L	18:2	14
Linolenate	Ln	18:3	12

*ECN ≈ C – 2·DB; TAG molecules elute by RP-HPLC in order of their molecular ECN values (the sum of three FA ECNs).

tests of the identified $[DAG]^+$ ions. The twenty features with the lowest q-values were identified using RTs and mass spectral data. All statistical procedures, data manipulation, and graphics were performed using R (87; Ver. 2.9.1) and Microsoft Excel (2007).

RESULTS

The FA composition of milkfat, lard, and control diets is listed in Table 12; our analysis revealed that the diet compositions were consistent with expected values (Harlan Laboratories, Madison, WI). Total SFAs were higher in MD than LD samples; total MUFAs were higher in LD than MD samples (Table 12). The FA composition of adipose tissue harvested from MD-, LD-, and CD-fed mice at eight and sixteen weeks is listed in Table 13. Palmitate (16:0) and oleate (18:1n-9) were the most prominent FAs in all feed and adipose samples; linoleate was also very high in CD and LD samples. The following eight FAs were present at > 1.0% by weight on average in adipose samples from MD- or LD-fed mice: 12:0, 14:0, 16:0, 18:0, 16:1n-7, 18:1n-9, 18:1n-7, and 18:2n-6. These FAs were used in all statistical analyses of the FA profile data.

All eight FAs differed significantly between milkfat and lard diets (Figure 21 A). Multivariable analysis using a nonparametric permutational MANOVA results indicated that the FA profiles differed significantly between the two diets ($F = 29{,}324$, $R^2 = 0.99986$, $p = 0.035$). Four FAs differed significantly between adipose tissue samples from MD- and LD-fed mice at eight weeks (Figure 21 B), and seven FAs were

Table 12. Feed fatty acid profile data (mean ± SEM).

FAME	Control (n=3)	Lard (n=3)	Milkfat (n=3)
Saturated and branched-chain saturated fatty acids (SFA)			
4:0	0.9 ± 0.0	0.0 ± 0.0	2.4 ± 0.0
6:0	0.8 ± 0.0	0.0 ± 0.0	2.0 ± 0.0
8:0	0.4 ± 0.0	0.0 ± 0.0	1.2 ± 0.0
10:0	1.0 ± 0.0	0.1 ± 0.0	2.7 ± 0.0
11:0	0.0 ± 0.0	0.0 ± 0.0	0.1 ± 0.0
12:0	1.2 ± 0.0	0.1 ± 0.0	3.1 ± 0.0
13:0	0.0 ± 0.0	0.0 ± 0.0	0.1 ± 0.0
iso 14:0	0.0 ± 0.0	0.0 ± 0.0	0.1 ± 0.0
14:0	4.2 ± 0.0	1.3 ± 0.0	9.9 ± 0.0
iso 15:0	0.1 ± 0.0	0.0 ± 0.0	0.2 ± 0.0
anteiso 15:0	0.2 ± 0.0	0.0 ± 0.0	0.4 ± 0.0
15:0	0.5 ± 0.0	0.1 ± 0.0	1.1 ± 0.0
iso 16:0	0.1 ± 0.0	0.0 ± 0.0	0.2 ± 0.0
16:0	21.3 ± 0.0	22.3 ± 0.1	27.7 ± 0.1
iso 17:0	0.1 ± 0.0	0.0 ± 0.0	0.3 ± 0.0
anteiso 17:0	0.2 ± 0.0	0.0 ± 0.0	0.4 ± 0.0
17:0	0.4 ± 0.0	0.4 ± 0.0	0.5 ± 0.0
18:0	9.8 ± 0.0	12.7 ± 0.1	10.5 ± 0.0
20:0	0.2 ± 0.0	0.2 ± 0.0	0.2 ± 0.0
22:0	0.1 ± 0.0	0.0 ± 0.0	0.1 ± 0.0
Monounsaturated fatty acids (MUFA)			
10:1	0.1 ± 0.0	0.0 ± 0.0	0.2 ± 0.0
12:1(1)	0.0 ± 0.0	0.0 ± 0.0	0.1 ± 0.0
12:1(2)	0.0 ± 0.0	0.0 ± 0.0	0.1 ± 0.0
14:1*n*-5	0.3 ± 0.0	0.0 ± 0.0	0.8 ± 0.0
16:1*n*-10(9)	0.2 ± 0.0	0.3 ± 0.0	0.2 ± 0.0
16:1*n*-7	1.2 ± 0.0	1.8 ± 0.0	1.2 ± 0.0
17:1*n*-8	0.2 ± 0.0	0.3 ± 0.0	0.2 ± 0.0
18:1*n*-9	27.0 ± 0.1	35.7 ± 0.1	22.0 ± 0.1
18:1*n*-7	1.7 ± 0.0	2.3 ± 0.0	1.1 ± 0.0
18:1*n*-6	0.2 ± 0.1	0.1 ± 0.0	0.9 ± 0.4
18:1*n*-5	0.1 ± 0.0	0.0 ± 0.0	0.3 ± 0.0
20:1*n*-13(11)	0.0 ± 0.0	0.0 ± 0.0	0.1 ± 0.0
20:1*n*-9	0.3 ± 0.0	0.7 ± 0.0	0.1 ± 0.0
Polyunsaturated fatty acids (PUFA)			
18:2*n*-6	23.3 ± 0.1	19.3 ± 0.1	7.6 ± 0.0
18:2 (conj)	0.2 ± 0.0	0.1 ± 0.0	0.5 ± 0.0
18:3*n*-3	2.9 ± 0.0	1.2 ± 0.0	1.1 ± 0.0
20:2*n*-6	0.2 ± 0.0	0.6 ± 0.0	0.0 ± 0.0
20:3*n*-6	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0
20:4*n*-6	0.1 ± 0.0	0.2 ± 0.0	0.1 ± 0.0
Sums and ratios			
Σ SFA	41.7 ± 0.1	37.3 ± 0.2	63.2 ± 0.2
Σ MUFA	31.5 ± 0.2	41.1 ± 0.1	27.3 ± 0.3
Σ *n*-3 PUFA	2.9 ± 0.0	1.2 ± 0.0	1.1 ± 0.0
Σ *n*-6 PUFA	23.9 ± 0.1	20.3 ± 0.1	8.8 ± 0.3
16:0/18:1*n*-9	0.8 ± 0.0	0.6 ± 0.0	1.3 ± 0.0
n-3/*n*-6	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0

Table 13. Adipose tissue fatty acid profile data (mean ± SEM).

	8 weeks			16 weeks		
	Control	Lard	Milkfat	Control	Lard	Milkfat
FAME	n=6	n=6	n=6	n=3	n=6	n=6
Saturated and branched-chain saturated fatty acids (SFA)						
10:0	0.1 ± 0.0	0.0 ± 0.0	0.3 ± 0.0	0.1 ± 0.0	0.0 ± 0.0	0.2 ± 0.0
12:0	0.4 ± 0.0	0.1 ± 0.0	1.0 ± 0.0	0.3 ± 0.0	0.1 ± 0.0	0.9 ± 0.1
14:0	2.0 ± 0.1	1.0 ± 0.1	4.7 ± 0.1	2.2 ± 0.0	0.9 ± 0.0	4.7 ± 0.3
anteiso 15:0	0.1 ± 0.0	0.0 ± 0.0	0.1 ± 0.0	0.1 ± 0.0	0.0 ± 0.0	0.1 ± 0.0
15:0	0.3 ± 0.0	0.1 ± 0.0	0.7 ± 0.0	0.3 ± 0.0	0.1 ± 0.0	0.8 ± 0.0
iso 16:0	0.2 ± 0.0	0.1 ± 0.0	0.2 ± 0.0	0.2 ± 0.0	0.1 ± 0.0	0.2 ± 0.0
16:0	18.6 ± 0.5	19.3 ± 0.7	21.1 ± 0.8	19.4 ± 0.5	18.9 ± 0.4	22.8 ± 0.5
iso 17:0	0.1 ± 0.0	0.0 ± 0.0	0.2 ± 0.0	0.1 ± 0.0	0.0 ± 0.0	0.2 ± 0.0
anteiso 17:0	0.1 ± 0.0	0.1 ± 0.0	0.2 ± 0.0	0.1 ± 0.0	0.0 ± 0.0	0.3 ± 0.0
17:0	0.2 ± 0.0	0.2 ± 0.0	0.3 ± 0.0	0.3 ± 0.0	0.3 ± 0.0	0.3 ± 0.0
iso 18:0	0.1 ± 0.0	0.0 ± 0.0	0.1 ± 0.0	0.1 ± 0.0	0.0 ± 0.0	0.1 ± 0.0
18:0	2.8 ± 0.2	3.2 ± 0.2	3.3 ± 0.2	2.6 ± 0.4	3.5 ± 0.3	2.9 ± 0.1
20:0	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0
Monounsaturated fatty acids (MUFA)						
14:1*n*-5	0.2 ± 0.0	0.1 ± 0.0	0.4 ± 0.0	0.2 ± 0.0	0.0 ± 0.0	0.5 ± 0.0
16:1*n*-10(9)	0.8 ± 0.1	0.7 ± 0.0	0.7 ± 0.0	0.9 ± 0.0	0.7 ± 0.0	0.8 ± 0.1
16:1*n*-7	6.1 ± 0.5	4.7 ± 0.2	5.7 ± 0.5	6.8 ± 1.6	3.8 ± 0.4	6.6 ± 0.5
17:1*n*-8	0.4 ± 0.0	0.4 ± 0.0	0.5 ± 0.0	0.4 ± 0.0	0.4 ± 0.0	0.6 ± 0.0
18:1*n*-9	38.4 ± 0.6	45.2 ± 0.8	42.4 ± 1.3	40.9 ± 1.6	47.0 ± 0.6	42.6 ± 1.0
18:1*n*-7	2.6 ± 0.1	2.4 ± 0.0	1.9 ± 0.0	3.0 ± 0.2	2.5 ± 0.0	1.8 ± 0.1
18:1*n*-5	0.2 ± 0.1	0.0 ± 0.0	0.6 ± 0.0	0.1 ± 0.0	0.0 ± 0.0	0.6 ± 0.0
19:1	0.1 ± 0.0	0.1 ± 0.0	0.2 ± 0.0	0.2 ± 0.0	0.1 ± 0.0	0.2 ± 0.0
20:1*n*-9	0.8 ± 0.1	0.7 ± 0.0	0.5 ± 0.1	0.9 ± 0.2	0.8 ± 0.1	0.4 ± 0.0
Polyunsaturated fatty acids (PUFA)						
16:2	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0
18:2*n*-7	0.2 ± 0.1	0.1 ± 0.0	0.5 ± 0.0	0.2 ± 0.1	0.1 ± 0.0	0.5 ± 0.0
18:2*n*-6	22.7 ± 0.7	19.2 ± 0.5	11.6 ± 0.3	17.4 ± 0.1	18.5 ± 0.3	9.7 ± 0.1
18:2 (conj)	0.3 ± 0.1	0.2 ± 0.0	1.0 ± 0.0	0.4 ± 0.0	0.1 ± 0.0	1.0 ± 0.0
18:3*n*-3	1.1 ± 0.1	0.8 ± 0.0	0.6 ± 0.1	0.9 ± 0.1	0.6 ± 0.0	0.6 ± 0.0
20:2*n*-6	0.2 ± 0.0	0.3 ± 0.0	0.1 ± 0.0	0.2 ± 0.0	0.4 ± 0.0	0.0 ± 0.0
20:3*n*-6	0.2 ± 0.0	0.2 ± 0.0	0.2 ± 0.0	0.2 ± 0.0	0.2 ± 0.0	0.1 ± 0.0
20:4*n*-6	0.4 ± 0.1	0.3 ± 0.0	0.2 ± 0.0	0.3 ± 0.0	0.3 ± 0.0	0.2 ± 0.0
22:5*n*-3	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0
22:6*n*-3	0.3 ± 0.0	0.2 ± 0.0	0.1 ± 0.0	0.2 ± 0.0	0.2 ± 0.0	0.1 ± 0.0
Sums and ratios						
Σ SFA	25.0 ± 0.7	24.3 ± 0.9	32.5 ± 0.7	26.1 ± 0.3	24.0 ± 0.5	33.5 ± 0.7
Σ MUFA	49.5 ± 0.2	54.2 ± 0.7	53.0 ± 1.0	53.7 ± 0.2	55.3 ± 0.5	54.1 ± 0.7
Σ *n*-3 PUFA	1.4 ± 0.1	1.1 ± 0.1	0.9 ± 0.1	1.2 ± 0.1	0.8 ± 0.1	0.8 ± 0.1
Σ *n*-6 PUFA	23.5 ± 0.8	20.1 ± 0.5	12.1 ± 0.3	18.3 ± 0.1	19.5 ± 0.3	10.0 ± 0.1
16:0/18:1*n*-9	0.5 ± 0.0	0.4 ± 0.0	0.5 ± 0.0	0.5 ± 0.0	0.4 ± 0.0	0.6 ± 0.0
n-3/*n*-6	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0	0.0 ± 0.0	0.1 ± 0.0

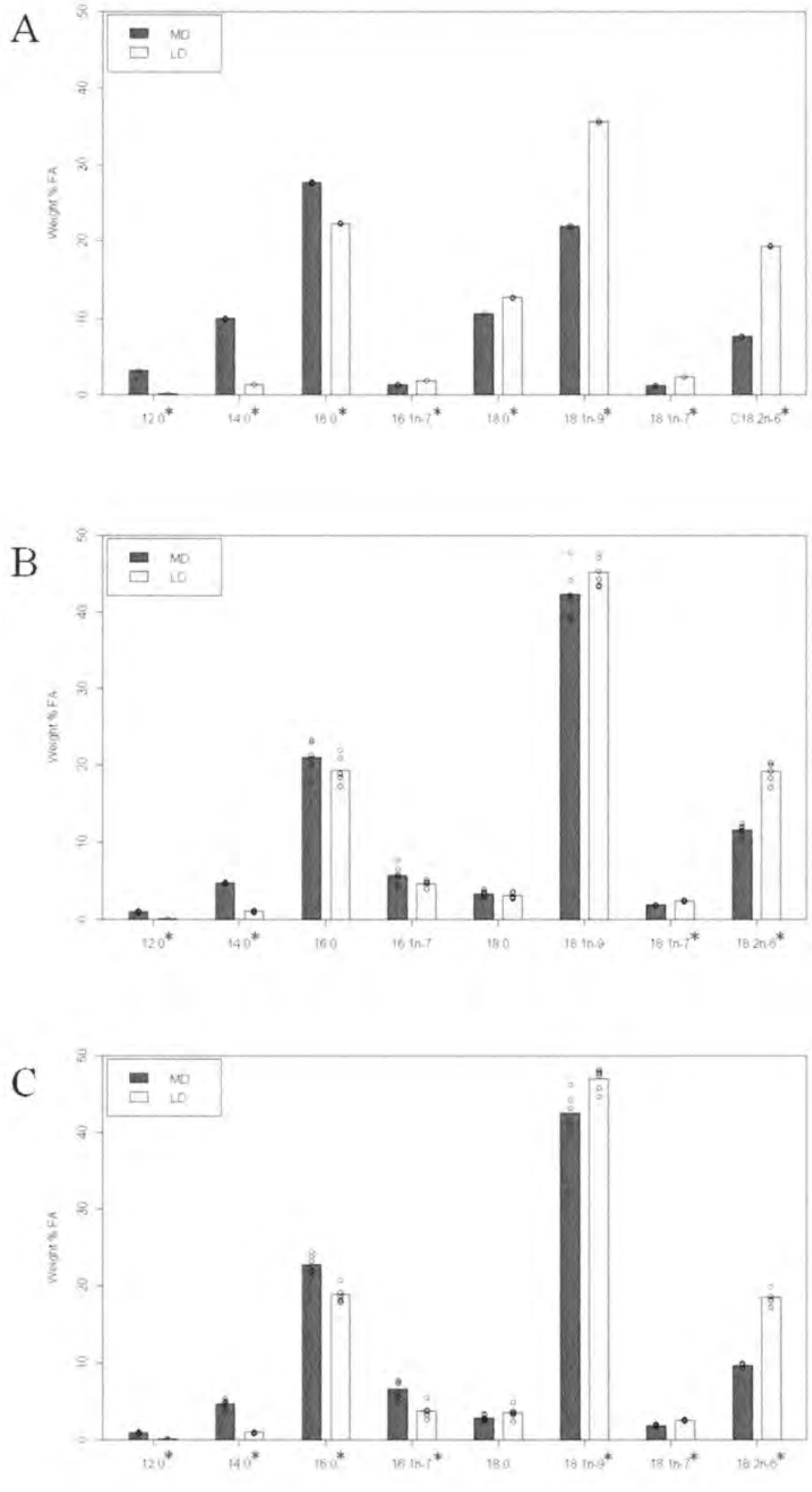


Figure 21. Profiles of major FAs in (A) milkfat (MD) and lard (LD) diets (n=3 per diet) and adipose tissue from mice fed milkfat (MD) and lard (LD) diets sampled at (B) 8 weeks (n=6 per diet) and (C) 16 weeks (n=6 per diet). Means and actual data are shown; * indicates $p < 0.00625$.

significantly different at sixteen weeks (Figure 21 C). Nonparametric permutational MANOVA revealed that FA profiles of adipose samples differed between diet groups at both time points (8 weeks, F = 26.1, $R^2 = 0.72$, p = 0.004; 16 weeks, F = 64.4, $R^2 = 0.85$, p = 0.003). Differences in adipose tissue palmitate (16:0) and oleate (18:1n-9) between diet groups were observed at sixteen weeks, but not at eight weeks; this suggests a delay in the onset of effects. FA profiles did not differ between the eight- and sixteen-week time points, but linoleic acid (18:2n-6) decreased in MD-fed mice (p = 0.0003) and palmitoleic acid (16:1n-7) decreased in LD-fed mice (p = 0.0035) between eight and sixteen weeks. Pattern similarities in FA profiles between diet and adipose tissue samples were highest in LD samples, followed by CD and MD samples at both time points, as shown in Figure 22. Because mean values were used to calculate pattern similarities, statistical significance could not be assessed between the time points.

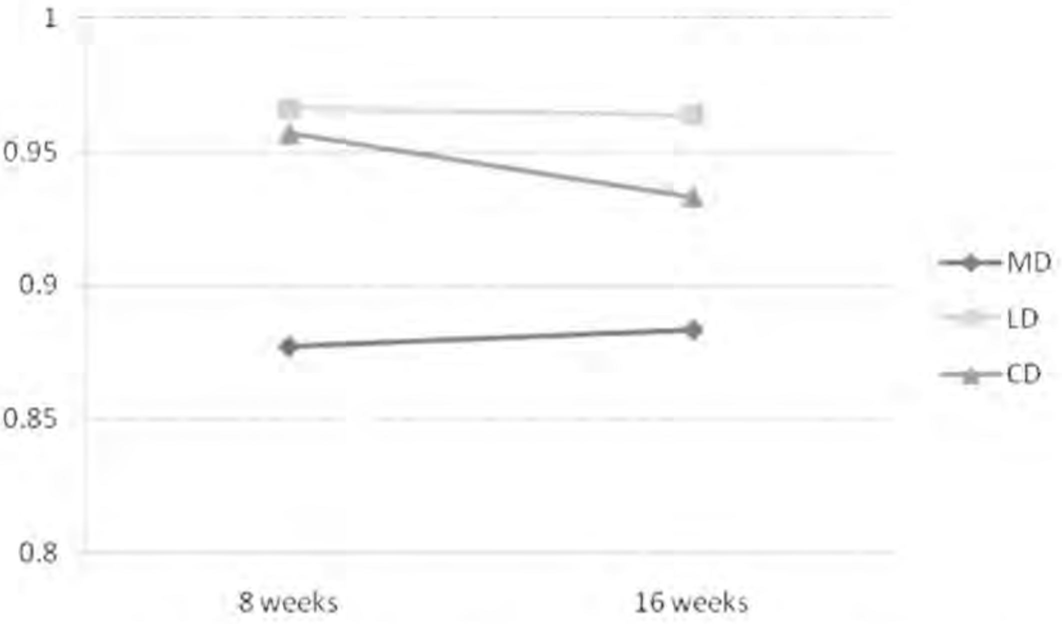


Figure 22. Pattern similarities (123) in FA composition between diet and adipose tissue from mice fed milkfat (MD), lard (LD), and control (CD) diets at 8 and 16 weeks. Note: pattern similarities calculated using diet and AT means of 8 FA, cosine similarity measure.

RP-HPLC/APCI-MS data from the analysis of TAGs in adipose tissue sampled from MD- and LD-fed mice were processed as described in Methods; eight- and sixteen-week samples were processed separately. We detected 2285 and 2020 peaks, respectively, in mouse adipose tissue sampled at eight and sixteen weeks (Table 14). As xcms detects all peaks present in the samples using the specified parameters (signal-to-noise ratio, etc.), non-TAG peaks were included in this list. Also, several peaks may represent the same analyte because TAG molecules may be represented by $[M+H]^+$ and as many as three $[DAG]^+$, $[MAG]^+$ and $[RCO]^+$ fragment ions, depending on the different FAs attached to glycerol. We filtered this peak list for $[DAG]^+$ ions representing the TAGs likely to be observed based on the feed and adipose tissue FA profile data. The eight FAs listed in Table 11 were used to generate a list of 120 possible TAG molecular species by considering all possible combinations by choosing three FAs with replacement from the list of 8 FAs. Because we use masses to screen for specific ions, we cannot differentiate between FAs with the same number of carbons and double bonds; i.e. 18:1n-9 and 18:1n-7 are both 18:1 or O. We do not account for positional differences in TAG molecules, though it is possible that positional isomers are present. Filtering the xcms peak list for the 24 unique $[DAG]^+$ m/z values that we would observe for the possible TAG species resulted in 85 and 80 $[DAG]^+$ peaks for eight- and sixteen-week samples, respectively (Table 14). We ranked these features in order of ascending q-values identified using p-values from individual t-tests to evaluate the null hypothesis of no difference between MD- and LD- fed mouse adipose tissue samples and observed that 56

(8-week) and 58 (16-week) features had q-values less than 0.01 (Table 14), indicating a probability less than 0.01 that these features were falsely deemed significant.

Table 14. Peak counts and description for results of TAG RP-HPLC/APCI-MS analysis of MD- and LD-fed mouse AT sampled at 8 and 16 weeks.

	8-week AT samples	16-week AT samples
xcms peaks	2285	2020
$[DAG]^+$ ion peaks	85	80
q-value < 0.01	56	58
q-value < 0.001	44	47
q-value < 0.0001	28	30

Note: Data were processed using xcms to identify all peaks present, peaks were screened for specific $[DAG]^+$ peaks based on all possible combinations of eight FAs, and q-value ranks were used to select 20 peaks that differed between feed groups.

We selected twenty features with the lowest q-values, separately for eight- and sixteen-week samples, for further analysis. Heatmaps and dendrograms illustrating hierarchical cluster analysis using all detected peaks, $[DAG]^+$ peaks, and the twenty features selected using q-values are shown in Figure 23. The three feature sets clearly separate the samples based on feed groups for adipose tissue TAG data sampled at both time points; distances between samples from the different feed groups increase as we decrease the number of features used in the analysis. Nonparametric permutational MANOVA based on the twenty features selected at each time point revealed that TAG profiles of adipose samples differed between diet groups (8 weeks, $F = 271.1$, $p = 0.002$, $R^2 = 0.96$; 16 weeks, $F = 177.9$, $p = 0.003$, $R^2 = 0.95$). Multidimensional scaling of Bray-Curtis distances of the twenty peaks selected at each time point clearly separated samples based on diet group in two dimensions for adipose tissue samples at both time points (Figure 24).

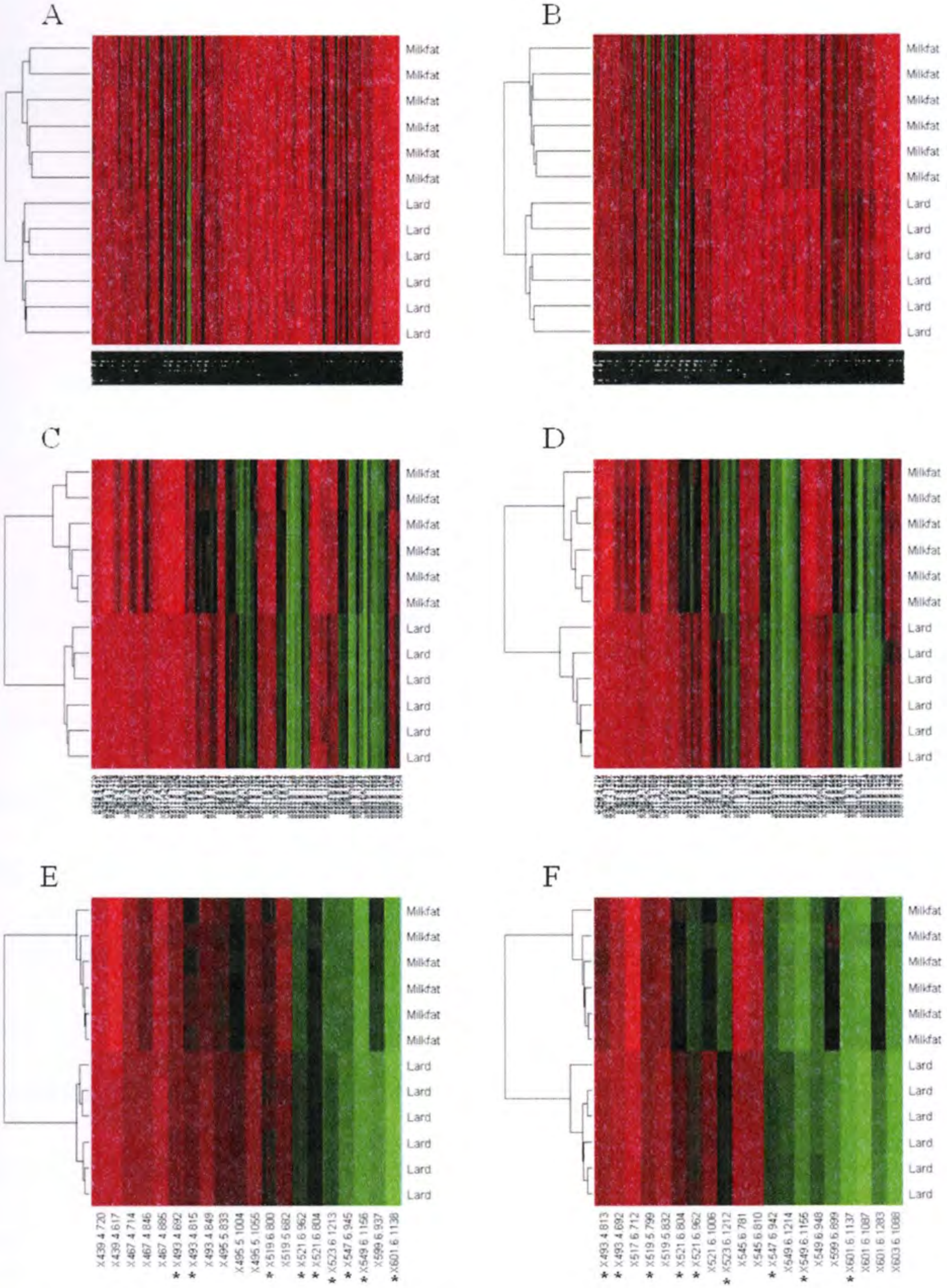


Figure 23. Heatmaps and dendrograms showing results of hierarchical cluster analysis (Euclidean distance of log values, complete linkage) of (A) 2285 TAG RP-HPLC/APCI-MS features resulting from preprocessing 8-week AT samples with xcms, (B) 2020 features identified similarly in 16-week samples, (C) 85 features resulting from screening xcms output for $[DAG]^+$ ions in 8-week AT samples, (D) 80 features identified similarly in 16-week samples, (E) twenty features selected using q-value ranks in 8-week and (F) 16-week AT samples. *Nine features were identified in samples at both time points.

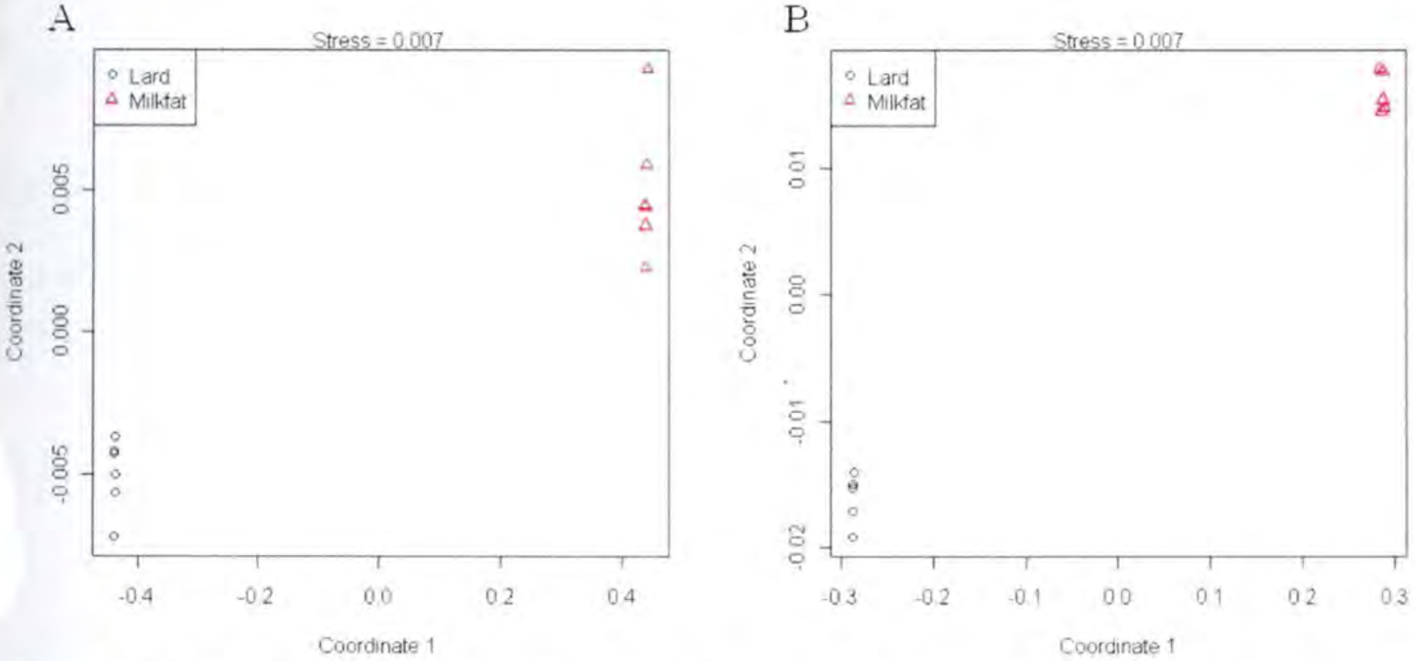


Figure 24. Multidimensional scaling of TAG data using Bray-Curtis distances for twenty TAG features selected using q-value ranks, adipose tissue sampled at (A) 8 weeks and (B) 16 weeks.

Automated TAG structure analysis was performed on the 20 selected features in eight- and sixteen-week samples using the algorithm described in the Appendix. TAG structures assigned to the features are listed in Table 15 (8-week) and Table 16 (16 week). Nine features were common to data from both time points and TAG structure assignments were consistent for these peaks. Structures were not assigned for two features identified in the eight-week samples, and one in the sixteen-week samples. In the eight-week data, X467.4.885 eluted with OOL, and X495.5.1055 eluted with OLL; these TAG are present in large quantities and the ions observed for these components were much larger than any others in the spectra for the two features. The feature X517.6.712, identified in the sixteen-week samples was a low-intensity feature with coelution from multiple other TAG species leading to poor spectral interpretation for structure determination for this feature. Extracted ions for 20 features are presented in Figure 25 (8 weeks) and Figure 26 (16 weeks).

Table 15. Twenty TAG features identified by q-value ranks in adipose tissue sampled from MD- and LD-fed mice at 8 weeks. Note: Diet indicates which diet group had higher relative amounts of each feature.

Peak	Feature	RT/(s)	m/z	Ion	ECN	TAG ID	Diet
1	X439.4.617	617	439.4	$[\mathrm{LaLa}]^+$	38	LLaLa	MD
2	X519.5.682	682	519.5	$[\mathrm{LaL}]^+$	40	LLaL + LPoLa	MD
3*	X493.4.692	692	493.4	$[\mathrm{LaPo}]^+$	40	LPoLa	MD
4	X467.4.714	714	467.4	$[\mathrm{LaM}]^+$	40	PoMLa	MD
5	X439.4.720	720	439.4	$[\mathrm{LaLa}]^+$	40	OLaLa	MD
6*	X519.6.800	800	519.6	$[\mathrm{LaL}]^+$	42	OLLa	MD
7*	X521.6.804	804	521.6	$[\mathrm{MPo}]^+$	42	LPoM	MD
8*	X493.4.815	815	493.4	$[\mathrm{LaPo}]^+$	42	OPoLa	MD
9	X495.5.833	833	495.5	$[\mathrm{LaP}]^+$	42	LPLa + PPoLa	MD
10	X467.4.846	846	467.4	$[\mathrm{LaM}]^+$	42	OMLa	MD
11	X493.4.849	849	493.4	$[\mathrm{LaPo}]^+$	42	PPoLa	MD
12	X467.4.885	885	467.4	$[\mathrm{LaM}]^+$	--	"NA"	MD
13	X599.6.937	937	599.6	$[\mathrm{LL}]^+$	44	LPL	LD
14*	X547.6.945	945	547.6	$[\mathrm{PoPo}]^+$	44	OPoPo	MD
15*	X521.6.962	962	521.6	$[\mathrm{MPo}]^+$	44	OPoM	MD
16	X495.5.1004	1004	495.5	$[\mathrm{LaP}]^+$	44	OPLa	MD
17	X495.5.1055	1055	495.5	$[\mathrm{LaP}]^+$, $[\mathrm{MM}]^+$	--	"NA"	MD
18*	X601.6.1138	1138	601.6	$[\mathrm{OL}]^+$	46	OLP	LD
19*	X549.6.1156	1156	549.6	$[\mathrm{PPo}]^+$	46	OPPo	MD
20*	X523.6.1213	1213	523.6	$[\mathrm{MP}]^+$	46	OPM	MD

[*] Peaks included in top twenty features identified by q-value ranks in adipose tissue sampled at both 8 and 16 weeks.

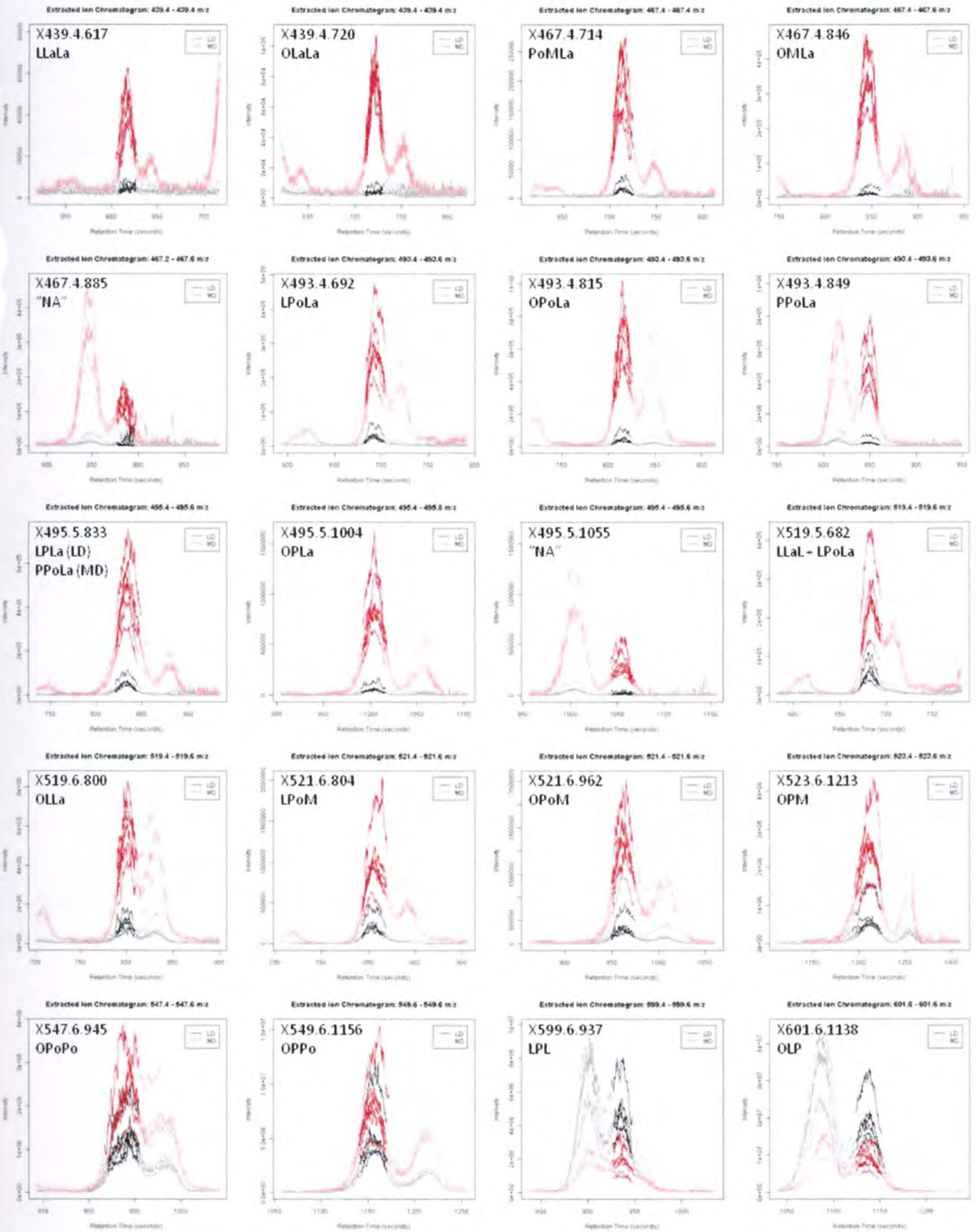


Figure 25. Extracted ion chromatograms for twenty TAG features selected using q-value ranks, adipose tissue sampled at 8 weeks from mice fed lard- and milkfat-based high fat diets. Integrated areas, used for quantitation, are shown with a darkened line.

Table 16. Twenty TAG features identified by q-value ranks in adipose tissue sampled from MD- and LD-fed mice at 16 weeks. Note: Diet indicates which diet group had higher relative amounts of each feature.

Peak	Feature	RT/(s)	m/z	Ion	ECN	TAG ID	Diet
1*	X493.4.692	692	493.4	$[LaPo]^+$	40	LPoLa	MD
2	X517.6.712	712	517.6	$[LaLn]^+$	--	"NA"	MD
3	X545.6.781	781	545.6	$[MLn]^+$	40	LLnM	MD
4*	X519.5.799	799	519.5	$[LaL]^+$	42	OLLa	MD
5*	X521.6.804	804	521.6	$[MPo]^+$	42	LPoM	MD
6	X545.6.810	810	545.6	$[MLn]^+$	42	OLnM	MD
7*	X493.4.813	813	493.4	$[LaPo]^+$	42	OPoLa	MD
8	X519.5.832	832	519.5	$[LaL]^+$	42	LPLa	MD
9	X599.6.899	899	599.6	$[LL]^+$	44	OLL	LD
10*	X547.6.942	942	547.6	$[PoPo]^+$	44	OPoPo	MD
11	X549.6.948	948	549.6	$[PPo]^+$	44	LPPo	MD
12*	X521.6.962	962	521.6	$[MPo]^+$	44	OPoM	MD
13	X521.6.1006	1006	521.6	$[LaO]^+$	44	OPLa	MD
14	X601.6.1087	1087	601.6	$[OL]^+$	46	OLO	LD
15	X603.6.1088	1088	603.6	$[OO]^+$	46	OLO, OPoO	LD
16*	X601.6.1137	1137	601.6	$[OL]^+$	46	OLP	LD
17*	X549.6.1155	1155	549.6	$[PPo]^+$	46	OPPo	MD
18*	X523.6.1212	1212	523.6	$[MP]^+$	46	OPM	MD
19	X549.6.1214	1214	549.6	$[MO]^+$	46	OPM	MD
20	X601.6.1283	1283	601.6	$[OL]^+$	48	SOL	LD

* Peaks included in top twenty features identified by q-value ranks in adipose tissue sampled at both 8 and 16 weeks.

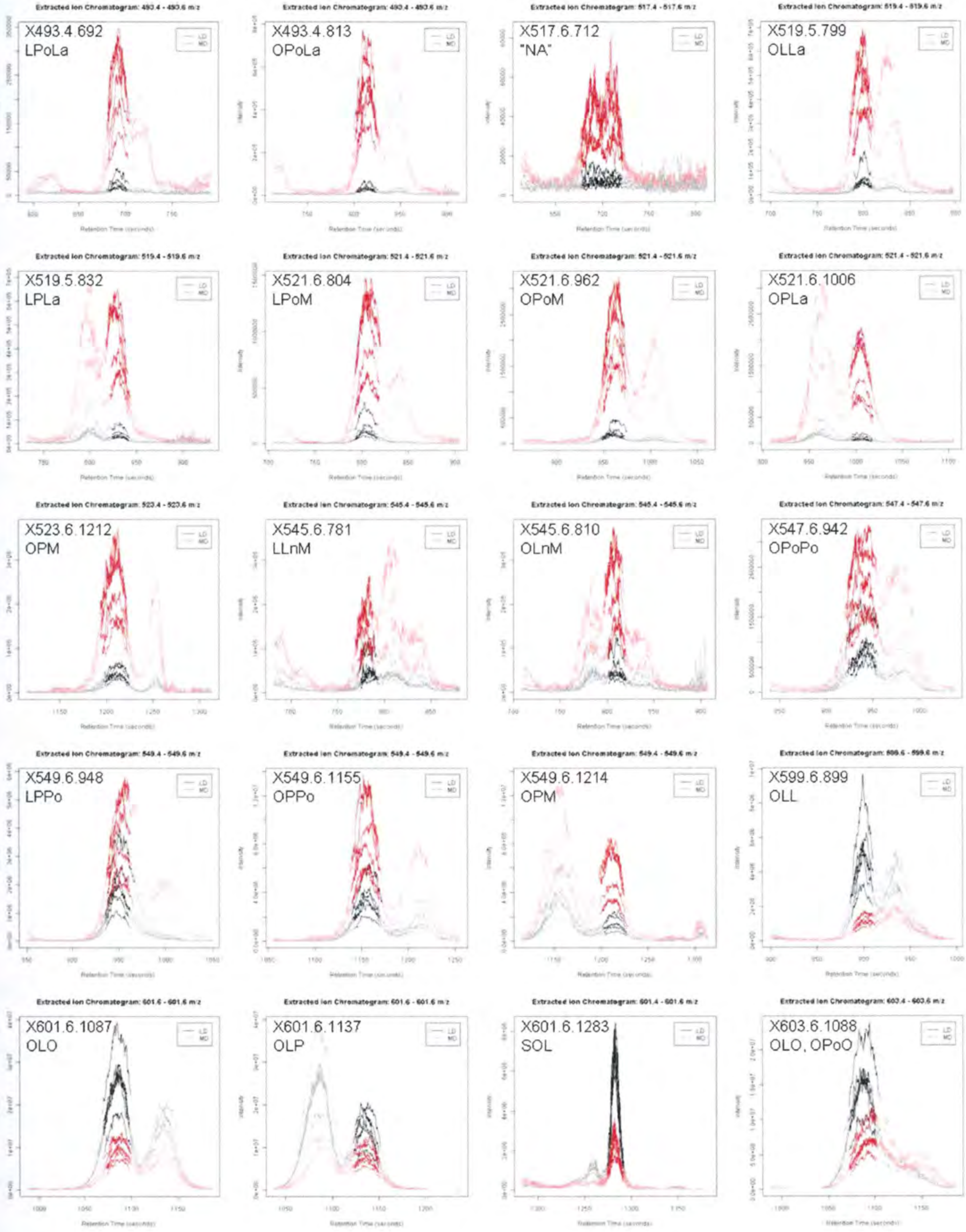


Figure 26. Extracted ion chromatograms for twenty TAG features selected using q-value ranks, adipose tissue sampled at 16 weeks from mice fed lard- and milkfat-based high fat diets. Integrated areas, used for quantitation, are shown with a darkened line.

DISCUSSION

Rodent DIO models allow researchers to study MetS and related diseases, including type 2 diabetes and CVD, in a controlled laboratory setting using animals with relatively few genetic variations. The amount and type of dietary fat used in such studies often has an effect on the results observed (99). The MD used in our study was higher in SFAs and lower in UFAs than the LD (Table 12, Figure 21 A). MCFAs were also found at higher levels in the MD than LD. Trends observed in FA profiles of the diets were similarly observed in post-prandial plasma (11) and adipose tissue sampled at eight and sixteen weeks from MD- and LD-fed mice (Table 13, Figure 21 B-C). These results indicate that the different diets determine plasma and adipose tissue FA profiles, to some extent, as expected. A simplified model for the fate of dietary FAs (Figure 18) illustrates that dietary FAs enter the circulatory system in the postprandial state, then are either distributed to the liver and/or peripheral tissues or deposited as TAGs in adipose tissue; our data support this model. While substantial amounts of MCFAs (4:0, 6:0, 8:0 and 10:0 > 1.0%, Table 12) were measured in the MD, these FAs were not present in measurable amounts in adipose tissue of MD-fed mice (except 10:0, 0.2%, Table 13). This supports the proposed different pathways taken by dietary MCTs and LCTs during digestion (Figure 19); it is likely that these MCFAs are sent directly to the liver. MCFAs are rapidly oxidized by the liver and have a very low tendency to deposit in AT (111). Because adipose tissue TAGs are mobilized during low energy states, releasing FAs that are circulated to the liver and peripheral tissues via plasma (Figure 18), changes in adipose tissue FA profiles may have long-term implications on an organism's overall health.

Palmitate (16:0) is the major end-product of *de novo* FA biosynthesis, but various elongation and desaturation steps convert 16:0 to 18:0, 16:1n-7, 18:1n-9, and 18:1n-7 (Figure 20). These processes may affect FA composition in plasma and adipose tissue (Figure 18). Regulation of FA biosynthesis and related steps may control the amounts of these FAs under specific conditions, and perhaps in response to dietary FA composition. While all eight FAs differed significantly between MD and LD (Figure 21 A), only four (12:0, 14:0, 18:1n-7, and 18:2n-6) were different in eight-week adipose tissue samples (Figure 21 B), and seven (all except 18:0) in sixteen-week adipose tissue samples (Figure 21 C). The four FAs that did not differ between MD- and LD-fed mouse adipose tissue sampled at eight weeks were all FAs that could be synthesized via the *de novo* pathway (Figure 20). It is likely that the relative amounts of these FAs in plasma and adipose tissue are only partially dependent on diet, and are controlled metabolically. Any difference in adipose tissue levels of these FAs due to diet may have been offset by *de novo* FA biosynthesis and related pathways to maintain homeostatic control over TAG and FA composition. Further research will be necessary to determine whether, and how much, relative amounts of these biosynthetic FAs in adipose tissue can be influenced by diet.

Lipolysis, when FAs are released from adipose tissue TAGs into the bloodstream, is a selective process and depends on the polarity of individual TAG molecules (114). As the more polar TAGs (i.e. having lower ECN values, or eluting more quickly from the non-polar column during RP-HPLC) have better access to lipolytic enzymes, FAs from these TAG molecules are preferentially released during a state of low energy and may be more available to tissues. Perona et al. suggested that TAG molecular species composition should be considered in addition to FA composition. We used a metabolomics approach to identify twenty TAG molecules that differed between adipose tissue sampled from MD- and LD-fed mice at eight and sixteen weeks. We used eight FAs, with ECN values from 12 to 18, to generate a list of possible TAG molecules used to search RP-HPLC/APCI-MS data for $[DAG]^+$ peaks. ECN values of these TAGs ranged from 36 to 54, but actual peaks that differed in adipose tissue from animals in different feed groups had ECN values between 38 and 48. Based on the relative amounts of twenty TAGs, we found significant differences between TAG profiles from the different diet groups at both time points. Of the twenty TAG features selected from data at each time point, nine were identified in both data sets. While these data show clear differences between relative amounts of TAG molecules in MD- and LD-fed mice (Figure 25, Figure 26), how these differences translate to an effect on overall health is not clear.

Hu and colleagues recently revealed a mechanism whereby palmitate specifically drives sphingolipid-mediated insulin resistance via up-regulation of DES1, and demonstrated that oleate attenuated palmitate-induced overexpression of DES1, preventing an increase in cellular ceramide levels (104). Inconsistencies in results between cell culture studies, using the direct addition of palmitate, and rodent DIO models seeking to demonstrate a role for ceramide in insulin resistance had been noted previously with the suggestion that mechanisms for insulin resistance may depend on the relative amounts of FAs available to cells (107). Geng et al. found that MD-fed mice were more obese and more insulin resistant compared to LD-fed mice, though ceramide levels were similar, and identified a separate mechanism for insulin resistance (expression of TRIB3) that was promoted by SFAs and attenuated by UFAs (11). The higher SFAs and lower UFAs observed in adipose tissue from MD-fed mice may promote a chronic disease state resulting from constitutively high levels of enzymes such as DES1 and TRIB3. As dietary fats are mixtures of TAGs comprised of a variety of FAs, we can make changes in our diets that change the FAs available to cells and tissues and ultimately regulate metabolic processes that determine health and disease. This new mechanistic understanding of the effects of dietary fats on our health may provide insight for the prevention, control and treatment of MetS and obesity-related disease.

In conclusion, we determined that differences observed in FA profiles between milkfat- and lard-based high fat diets do result in differences in adipose tissue FA and TAG profiles in C57bl/6J adult male mice fed these diets. Greater obesity and more severe insulin resistance observed in MD-fed mice may result from higher relative amounts of SFAs in the diet. Future research should address other possible mechanisms for SFA-induced insulin resistance, and how such mechanisms are modulated by the mixtures of TAGs that make up dietary fat as opposed to individual FAs. Also, can subtle changes in adipose tissue composition influence an individual's propensity for disease? Specifically, why does the MD incite severe insulin resistance compared with the LD? This diet differs from the LD both in relative amounts of SFA/MUFA and also MCFAs. High-energy MCFAs are not deposited in AT, but are sent directly to the liver for immediate metabolism and serve as a quick energy source that is likely necessary for the young mammals for which it is intended. Milk is the only known food that is created in nature for no other reason except to be used for fuel—it is typically used to sustain young mammals and is necessary for their growth and viability. Further research should address the mechanism by which the MD induces MetS symptoms and work to determine whether an MD-induced disease state results from changes in adipose tissue, and therefore the relative quantities of specific FAs available to tissues, or a possible burden placed on the liver in metabolizing MCFAs.

SUMMARY AND CONCLUSIONS

This project began, as most do, with a simple scientific question. Having worked with FA profiles for some time, we wanted to explore whether TAG analyses by RP-HPLC/APCI-MS could yield similar, or even better, information for classification of marine samples for forensic purposes. RP-HPLC/APCI-MS has shown great promise in the analysis of plant oil TAGs (1, 2, 4), and we wanted to apply similar data processing methods to the analysis of marine oils, which are a great deal more complex in terms of FA and TAG composition. As it was not possible to separate individual TAG components in complex mixtures such as marine oils using current chromatographic capabilities (5), we had to explore other options for processing data from these samples.

We found that bottom-up profiling experiments common to proteomics, where complex mixtures of digested peptides isolated from blood or tissue are routinely analyzed using LC/MS, had similar complexity to TAG mixtures in extracted marine oils. A signal-based processing methodology was typically employed with these data as opposed to the more traditional peak detection, identification, and quantification steps that we were accustomed to in processing FA profile data. Data were treated as a two-dimensional signal matrix or image, as shown in Figure 4 (p. 16), and established methods in signal processing, statistics, and machine learning were used to find patterns characteristic of a particular sample or class of samples (32). Data undergo a series of preprocessing steps to assure consistency across experiments, and much effort has been focused on this goal in proteomics and metabolomics research, and the underlying and supporting discipline of bioinformatics (32-36).

We set out first to verify that metabolomics data processing tools such as xcms and MZmine would work similarly to the manual data processing procedures that we were familiar with, and that we could achieve similar classification performance with data produced by these tools. We successfully implemented data processing using the xcms package in the R statistical environment, and results were comparable with manual processing. This process is described in Paper 1: Computational methods for the differential profiling of plant oils. We then focused on our original aim to classify marine oils to their biological source, specifically to determine whether marine oil dietary supplements contained oils harvested from seals versus fish. We found that TAG RP-HPLC/APCI-MS data classified marine oils to their biological source with accuracies similar to those observed using FA profile data. TAG data were processed with xcms and we employed both targeted and non-targeted feature selection, based on TAG species that may be present for a known set of FAs in the samples and differences between groups. This research is presented in Paper 2: Forensic identification of seal oils using lipid profiles and statistical models.

Classification is a relatively easy task compared with difference detection and high-level biomarker discovery, so we moved our focus to these tasks to address questions related to observed differences in the health of mice fed different high-fat diets in a diet-induced obesity study. We obtained adipose tissue from mice fed lard- and milkfat-based high-fat diets for eight and sixteen weeks. As with the marine oils, we analyzed both FA and TAG profiles and used FA composition to determine a list of targeted features to filter from the xcms output. Difference detection (q-values) based on the null hypothesis of no difference between diet groups was used to identify a list of features for which to determine TAG structures. TAG structures were identified for nine features that were common to data from both time points. Paper 3: Differential profiling of adipose tissue triacylglycerols in mice fed milkfat- and lard-based high-fat diets describes this research.

During analysis of these data, we developed an automated tool in the R computing environment that combined xcms preprocessing, statistical analysis and difference detection techniques, and TAG structure identification (i.e. the numbers of carbons and double bonds in FAs attached to glycerol). As many of the steps involved in processing TAG data were common to the three experiments described above, a set of R functions was written to facilitate high-throughput processing of TAG RP-HPLC/APCI-MS data. Raw data stored in mzXML, mzData, or mzML formats are processed using a series of steps that ultimately lead to classification and/or difference detection with automated identification of TAG structures for selected features. TAG structures were identified by evaluating the relationships between $[DAG]^+$ and $[M+H]^+$ ions observed in APCI mass spectra according to Cvacka et al. (125), and a score was devised for individual TAG structures that evaluated multiple correlations among ions related to a particular TAG structure over the retention time window of the peak. A full description of this process is provided in the Appendix.

The processing of raw instrument data as opposed to determining quantitative values of molecular species is necessary in order to differentiate among complex mixtures of compounds, as observed for peptide digests and complex mixtures of TAGs. A limitation of this approach is the ability to determine true compositional data. We can only address classification and difference detection, when the ultimate goal is often to characterize the chemical composition of a sample. One limitation of this approach is that the use of raw data limits the analyst to comparison of samples analyzed on one instrument. Quantitative values for TAG composition would be necessary to compare sets of data among different studies.

Throughout these studies, we sought to address the challenges associated with TAG RP-HPLC/APCI-MS analysis using cross-disciplinary methods, and to develop a high-throughput data processing and analysis pipeline for these data that can be used to address biological questions related to the TAG composition of storage fats in plants and animals. The automated framework developed here integrates data processing, statistical analysis, and structure determination of the FA chain structures of individual TAG species. While manual processing may be feasible for the analysis of TAGs in plant oils, a high-throughput methodology is essential for more complicated projects such as differential profiling of TAGs in animal tissues. LC/MS data processing has begun a migration from proprietary instrument vendor software to open-source packages such as xcms that work on data stored in universal formats (e.g. mzXML, mzML, mzData), and tools for high-level data analysis must be developed that work with the data produced from these packages. We have developed one such tool that is specific to the analysis of TAGs using RP-HPLC/APCI-MS, and we envision expanding this tool to a general lipids analysis pipeline universal to all lipid classes composed of FAs analyzed using both GC/MS and LC/MS.

This study has opened the door to greater possibilities in the analysis of lipids using LC/MS and, as is often the case in scientific research, more questions related to what we can learn from these analyses. The first extensive characterization of menhaden oil TAGs was published in March 2012 (124). In this study, researchers used reversed phase ultra-high pressure liquid chromatography with atmospheric pressure chemical ionization-ion trap-time of flight mass spectrometry (RP-UHPLC/APCI-IT-TOF) with four serially coupled shell-packed octadecylsilyl columns (60cm total length) to separate 137 TAGs containing nineteen different FAs in 224 min. The data were processed manually, which is a very time-consuming and analyst-intensive process. We plan to apply the data processing methodology described here (see Appendix) to check TAG structure identifications in these data. This will be the first automated characterization of TAGs in a marine oil.

The mouse adipose tissue study revealed differences in FAs and TAGs in storage fats of mice fed lard- and milkfat-based high fat diets, which may be associated with higher insulin resistance and greater obesity observed in the milkfat-fed mice. An alternative hypothesis is that higher levels of medium chain TAGs (containing FAs with fewer than twelve carbons) present in the milkfat-based diet are placing a burden on the liver that could also contribute to the milkfat-associated disease state. We will analyze FA and TAG profiles of liver samples from these mice to provide insight into the mechanism of obesity-related disease.

Additional future research will include the comparison of the TAG structure score discussed here with scoring algorithms for similar programs and adaptation of the algorithm for the analysis of other lipid classes containing FAs (e.g. phosphoglycerolipids, wax and sterol esters, sphingolipids). Another goal will be to interface with tandem MS (MS^n) methods that are possible using ion trap MS. The xcms program has already established the capability for processing tandem MS data to provide structural information from unknown metabolites (125). We may be able to use RP-HPLC/APCI-MS^n analyses to confirm structure assignments and to determine TAG structures of multiple coeluting TAG species. Lastly, obtaining qualitative and quantitative data for the complete characterization of TAGs present in a fat or oil sample is the ultimate goal in these types of studies, and we will continue to work toward resolving the challenges associated with this process.

APPENDIX: AN INTEGRATED FRAMEWORK FOR THE ANALYSIS OF NATURAL MIXTURES OF TRIACYLGLYCEROLS USING RP-HPLC/APCI-MS—PREPROCESSING, STATISTICAL ANALYSIS AND AUTOMATED STRUCTURE DETERMINATION

Margaret H. Broadwater[1,2] and John H. Schwacke[2]

[1]Center for Coastal Environmental Health and Biomolecular Research (CCEHBR), National Ocean Service (NOS), National Oceanic and Atmospheric Administration (NOAA), 219 Fort Johnson Road, Charleston, SC 29412.

[2]Department of Biochemistry and Molecular Biology, Medical University of South Carolina, 173 Ashley Avenue, Charleston, SC 29425.

ABSTRACT

We present an integrated framework for the processing, statistical analysis, and structure determination of triacylglycerols (TAGs) occurring in natural oils and fats analyzed using reversed-phase high performance liquid chromatography with atmospheric pressure chemical ionization mass spectrometry (RP-HPLC/APCI-MS). This methodology combines the existing metabolomic preprocessing platform xcms with higher-level processing steps that include screening specifically for peaks resulting from TAGs, classification and/or difference detection, and determination of the fatty acyl chain structures of individual TAG molecules based on the observed masses of diacylglycerol fragment ions and molecular adducts. Suggested TAG structures are evaluated with a

correlation-based score that reflects whether structure-associated peaks are concurrently eluting over the retention-time course.

INTRODUCTION

Many useful tools have been developed in recent years for global metabolomic applications using gas and liquid chromatography with mass spectrometric detection (GC/MS and LC/MS). Several of these are freely available under a GNU General Public License [e.g. MZmine, MZmine 2, xcms, OpenMS (8-10, 126)], and allow users to process data from any LC/MS system via the conversion of proprietary instrument output files to open-source mzXML, mzData, or mzML formats (40, 127). All of these tools aim to treat these data as an image, as shown in Figure 4 (p. 16), and use signal processing, statistics, and machine learning methods to find patterns that are characteristic of a particular sample or class of samples (32). LC/MS data must undergo peak detection, chromatographic alignment, and normalization steps to assure consistency across experiments.

While automated data processing tools have been demonstrated to be useful in the analysis of lipids (64, 128, 129), in many cases lipids have specific structural advantages that global metabolomic tools do not exploit. For example, all triacylglycerol (TAG) molecules comprise glycerol bound to three fatty acyl (FA) groups via ester linkages. The FAs vary in chain length and number of double bonds, and we may determine the FA composition of a mixture of triacylglycerols using well-established fatty acid profiling methodologies such as gas chromatography of fatty acid methyl ester derivatives (5). We can use FA compositional information to predict the possible TAG structures present in a sample, i.e. a list of known TAG molecules defined by the number of carbons and double bonds in the three FA chains, and search LC/MS data from extracted lipid samples (e.g. plasma, tissues, and natural oils and fats) for peaks (ions) specific to these molecules. This targeted approach yields a list of the peak intensities representing selected TAG molecules, and differs from the non-targeted approach common in global metabolomics experiments (130, 131). Yetukuri and colleagues employed such a structure-based informatics strategy to facilitate the global lipidomic analysis of ob/ob and wild type mouse livers by using combinations of structural variants, i.e. FA chains and polar head groups, to determine theoretically possible lipid structures (129). However, in this study and others (64, 128), actual identification of lipid FA substitutions is achieved only by performing additional experiments using tandem MS.

RP-HPLC/APCI-MS is a popular technique for the analysis of mixtures of triacylglycerols found in natural oils and fats (2, 27, 30, 52, 56, 132, 133). TAG molecular species are separated by RP-HPLC and elute in order of their equivalent carbon number (ECN), which is approximately equal to the number of FA carbon atoms minus two times the number of carbon-carbon double bonds (ECN $\approx$ C – 2·DB). As TAG species elute from the HPLC, relatively simple spectra are produced via APCI-MS with base peaks consisting of either the protonated molecule, $[M+H]^+$, or diacylglycerol ions, $[DAG]^+$ or $[M-RCO_2]^+$, that result from the loss of a FA moiety (27, 56). These ions, together with HPLC retention time (RT) information, allow identification of the number of carbons and double bonds in the FAs attached to glycerol in each TAG species. The intensity of the protonated molecule depends on the degree of unsaturation of the FA moieties, with the relative abundance of the $[M+H]^+$ ion increasing with the number of double bonds in the molecule, and may be absent in some fully saturated TAGs (27).

RP-HPLC/APCI-MS experiments generate an abundance of data that can be overwhelming to an analyst looking at multiple samples or aiming to compare differences between groups or classes of samples. These data have traditionally been processed manually (1, 2, 4, 133), and reported results may be oversimplified due to the selective nature of the analysis and the limited capacity of the chromatographic system to separate individual TAG molecular species (5). We present here an integrated framework for the processing and statistical analysis of TAG RP-HPLC/APCI-MS data that combines the existing metabolomic preprocessing platform xcms with higher-level processing steps that include screening for specific $[DAG]^+$ and $[M+H]^+$ ions, classification and/or difference detection, and determination of the FA chain structures (number of carbons and double bonds) of individual TAG molecules based on observed ion masses. A score is calculated for each proposed structure based on the multiple correlation of the primary peak (identified by xcms) with other ions we would expect to observe for a particular TAG structure across the RT window of peak integration.

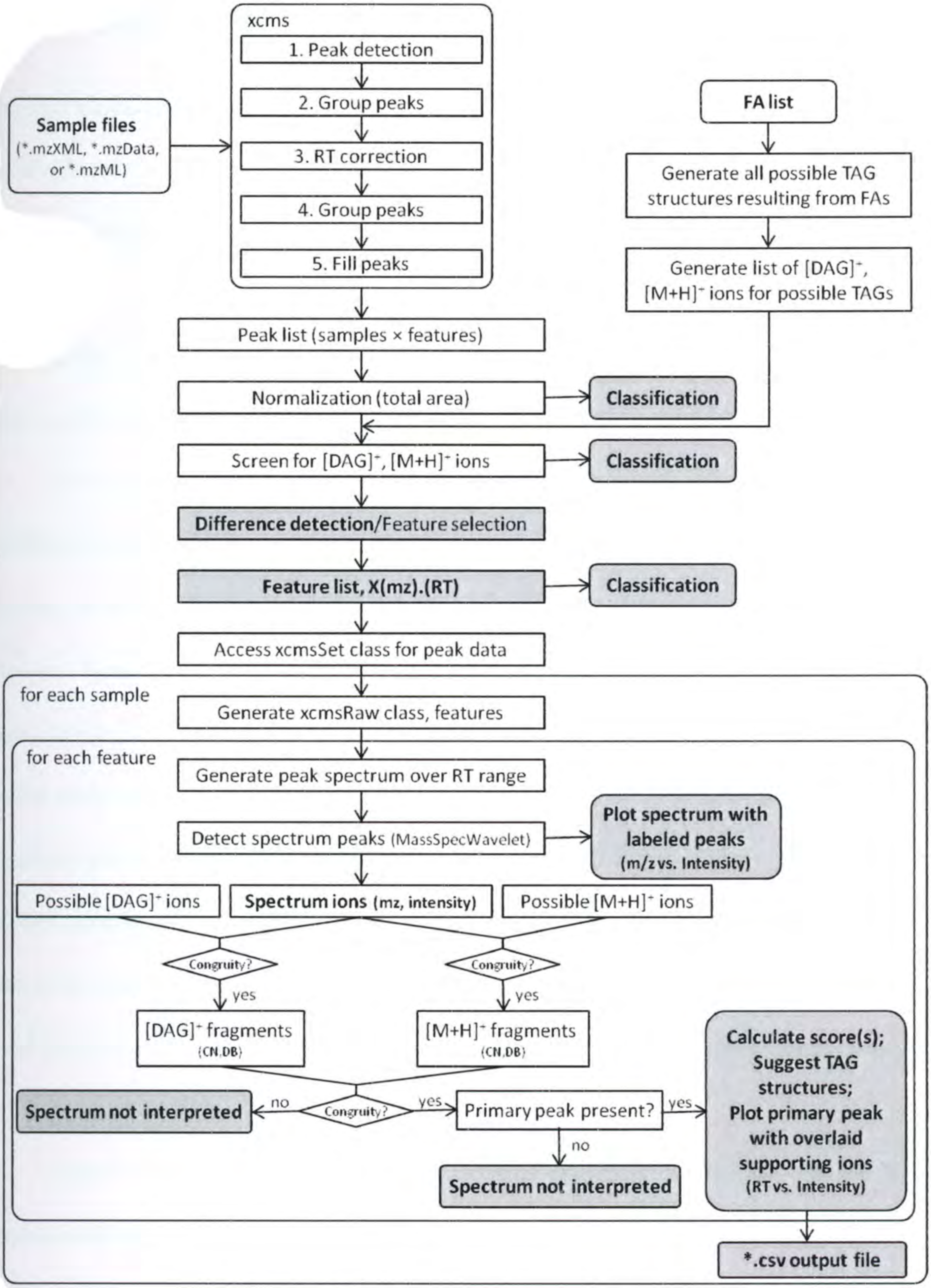

xcms
1. Peak detection
2. Group peaks
3. RT correction
4. Group peaks
5. Fill peaks
Sample files
(*.mzXML, *.mzData, or *.mzML)
FA list
Generate all possible TAG structures resulting from FAs
Generate list of [DAG]⁺, [M+H]⁺ ions for possible TAGs
Peak list (samples × features)
Normalization (total area)
Classification
Screen for [DAG]⁺, [M+H]⁺ ions
Classification
Difference detection/Feature selection
Feature list, X(mz).(RT)
Classification
Access xcmsSet class for peak data
for each sample
Generate xcmsRaw class, features
for each feature
Generate peak spectrum over RT range
Detect spectrum peaks (MassSpecWavelet)
Plot spectrum with labeled peaks (m/z vs. Intensity)
Possible [DAG]⁺ ions
Spectrum ions (mz, intensity)
Possible [M+H]⁺ ions
Congruity?
yes
Congruity?
yes
[DAG]⁺ fragments (CN,DB)
[M+H]⁺ fragments (CN,DB)
Spectrum not interpreted
no
Congruity?
yes
Primary peak present?
yes
no
Spectrum not interpreted
Calculate score(s); Suggest TAG structures; Plot primary peak with overlaid supporting ions (RT vs. Intensity)
*.csv output file

Figure 27. Data processing workflow.

PROGRAM DESCRIPTION

The data processing workflow described here is shown in Figure 27; this process takes a set of raw data files through a series of processing steps to identify relevant TAG features that differ among sample classes and to determine possible TAG structures that these features represent. The process may be used in its entirety as a data processing pipeline for TAG RP-HPLC/APCI-MS data, or steps may be selected and applied individually to solve specific problems related to these data.

We rely on the well-established methodology of xcms (9) for preprocessing raw RP-HPLC/APCI-MS data converted from proprietary instrument output files to mzXML, mzData, or mzML formats; tools for file conversion are provided by most instrument vendors. Xcms has been reviewed (34, 36) and used to process data in several published metabolomics studies (62, 63, 128); results from xcms processing are comparable with similar tools including MZmine (65). Preprocessing steps in xcms include peak detection, grouping, chromatographic alignment, and peak filling that allow the analyst to set data-specific processing parameters. The product of xcms preprocessing is a list of peak areas corresponding to specific m/z and RT indices for each sample; xcms output is in the form of a matrix of samples (rows) × features (columns) with each feature labeled using the notation X[m/z].[RT].

One TAG molecule may yield several highly correlated ions in this analysis (e.g. a protonated molecule and up to three DAG fragment ions and corresponding isotope peaks), and each is a separate peak in xcms. Thus, many of the identified features are redundant, i.e. several features represent one TAG molecule. When we plot a spectrum over a specific RT window for peak X[m/z].[RT], as in Figure 28, the spectrum will contain $[DAG]^+$ and $[M+H]^+$ ion peaks for the for each of the TAGs that elute during the RT window; we refer to the specific m/z value from X[m/z].[RT] as the primary peak. We will exploit redundancies in the data to determine the TAG structure that likely represents the primary peak and assign a score based on the correlations among peaks representing a specific TAG structure. Xcms does not provide a function for normalization, so we divide each peak by the total area for each sample to eliminate differences in peak intensities due to the amount of sample injected on the HPLC.

Most metabolomics tools, including xcms, employ a non-targeted strategy (130) to detect all peaks in a RP-HPLC/APCI-MS image. For a targeted analysis of TAGs, we want to select peaks from this list that represent TAG molecules, i.e. $[DAG]^+$ and $[M+H]^+$ ions. If we know which FAs are present in a specific sample or group of samples, we can predict the TAG structures that may be present and also the $[DAG]^+$ and $[M+H]^+$ ions that may be observed. We can implement a recursive search of the detected features to screen for m/z values we would expect to observe in TAGs resulting from combinations of specific FAs. This will eliminate most peaks that do not represent TAG molecules, i.e. noise and artefacts or contaminants. This list contains peak areas for relevant features, or peaks that represent TAG molecules, indexed by m/z and RT indices for each sample.

We use difference detection techniques to identify peaks that differ between classes of samples, e.g. plant oils from different biological sources or lipids extracted from plasma or tissues in diseased vs. healthy subjects (39). Difference detection can be an endpoint to our statistical data analysis, e.g. biomarker detection, or can be used to select a subset of features that best discriminate among sample classes. We use q-values calculated from p-values resulting from t-tests or ANOVAs (H_0: no difference between/among sample classes) to select a list of features of using a criterion based either on the desired number of features or a threshold q-value. The purpose of using q-values is to identify as many features that differ among sample classes as possible while incurring the lowest proportion of false positives. The q-value provides a measure of each feature's significance while accounting for the fact that many variables are being tested simultaneously (45, 46). Each feature is evaluated independently, thus redundant features may be included in this list of relevant features that differ among sample classes.

At this point, we have narrowed the original xcms output to relevant (TAG) features that differ among sample classes, and we can further examine these features by looking at each selected feature independently for each sample analyzed. Original peak data for each feature identified in each sample may be obtained by accessing the original xcmsSet() class variable created using xcms during preprocessing. We can generate an xcmsRaw() class for each individual sample to access raw data observed during the RT window of each identified feature. The MassSpecWavelet package may be used to detect spectrum peaks and plot spectra across the specified RT window (134). These spectrum ions can then be used to determine which possible TAG structures are eluting during the RT window.

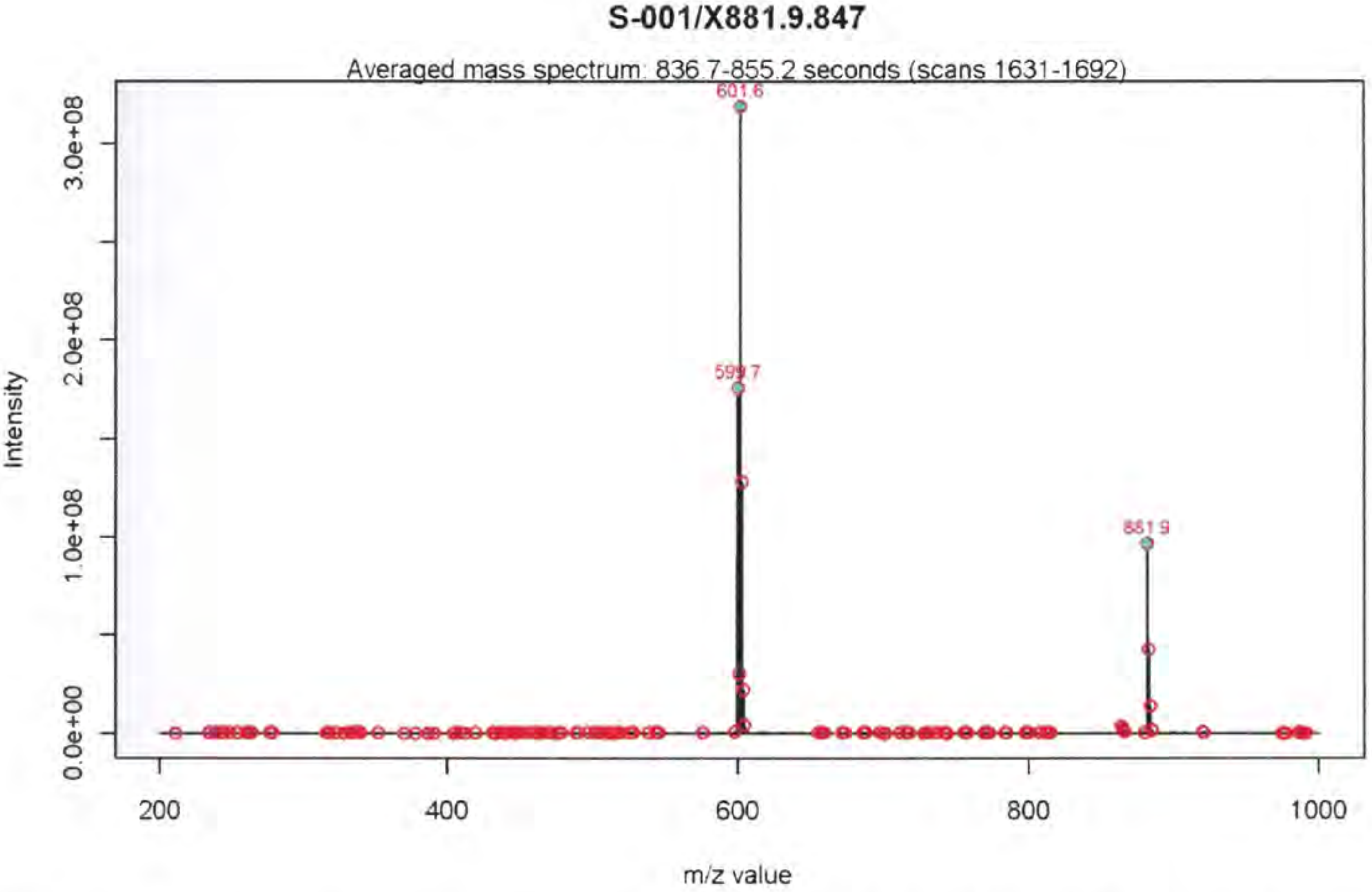


Figure 28. Spectrum for feature X881.9.847 identified in soybean oil sample S-001. Spectral peaks detected using MassSpecWavelet are marked with red circles; labeled peaks with cyan-filled circles are detected peaks that represent $[DAG]^+$ or $[M+H]^+$ ions associated with possible TAG structures.

We use an algorithm based on the *TriglyAPCI* program developed by Cvačka and colleagues (135) for the automated interpretation of TAG APCI mass spectra. The algorithm characterizes FAs, and also TAGs, by the number of FA carbons and double bonds and determines relations among ions in the spectra using these two parameters. The following equations define the relationship between the masses of $[DAG]^+$ fragment ions and $[M+H]^+$ molecular adducts in terms of the number of FA carbons (CN) and double bonds (DB).

$$CN_{(M+H)^+} = \frac{CN_{(M-R_1COO)^+} + CN_{(M-R_2COO)^+} + CN_{(M-R_3COO)^+}}{2}$$

$$DB_{(M+H)^+} = \frac{DB_{(M-R_1COO)^+} + DB_{(M-R_2COO)^+} + DB_{(M-R_3COO)^+}}{2}$$

When these equations are satisfied simultaneously by the ions in a spectrum, we can calculate the number of carbon atoms and double bonds in the individual FA groups to obtain the associated TAG structure using the following equations:

$$CN_{R_1COOH} = CN_{(M+H)^+} - CN_{(M-R_1COO)^+}$$

$$CN_{R_2COOH} = CN_{(M+H)^+} - CN_{(M-R_2COO)^+}$$

$$CN_{R_3COOH} = CN_{(M+H)^+} - CN_{(M-R_3COO)^+}$$

$$DB_{R_1COOH} = DB_{(M+H)^+} - DB_{(M-R_1COO)^+}$$

$$DB_{R_2COOH} = DB_{(M+H)^+} - DB_{(M-R_2COO)^+}$$

$$DB_{R_3COOH} = DB_{(M+H)^+} - DB_{(M-R_3COO)^+}$$

As we are only interested in TAG structures that generate the primary peak, we consider only structures that produce this peak in their spectra. We then calculate a correlation-based score using the multiple regression model,

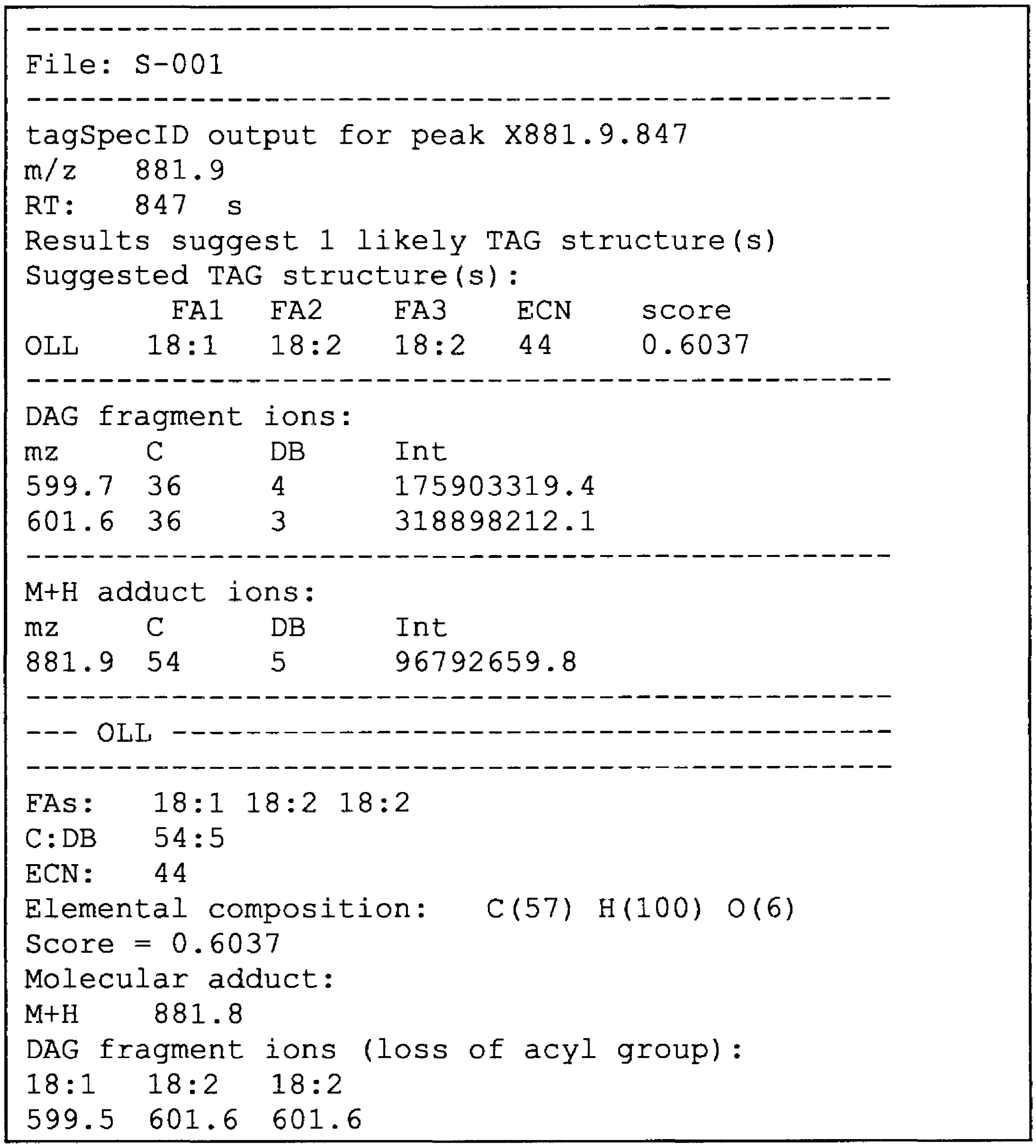
$$Y = \beta_0 + \beta_1 X_1 [+ \beta_2 X_2 + \beta_3 X_3] + \epsilon ,$$

where Y is the primary peak (m/z value) and $X_{1[,2,3]}$ are the other ions present for a particular TAG structure. We use the adjusted R^2 value calculated from the multiple regression model as the TAG structure score:

$$R_a^2 = 1 - \frac{n-1}{n-m-1}(1 - R^2),$$

where n is the number of observations and m is the number of independent variables in the multiple regression model. While R^2 increases when independent variables are added to the model, R_a^2 increases only when an added variable results in improved model fit (136). Thus, the adjusted R^2 value is a measure of the model fit that may be used to compare models with different numbers of parameters. An output file containing suggested TAG structure(s), with score value(s) and supporting $[M+H]^+$ and $[DAG]^+$ ions specific to each structure, is generated for each feature for each sample (Figure 29), along with diagnostic plots of the mass spectrum observed during the RT window of peak integration (Figure 28) and RT vs. intensity for the primary feature (m/z) with overlaid plots each of the ions that support suggested TAG structures for that ion (Figure 30). For the feature X881.9.847, identified in soybean oil sample S-001, the TAG structure score is the adjusted R^2 value for the regression model

$$mz881.9 = \beta_0 + \beta_1(mz599.7) + \beta_2(mz601.6).$$



```
------------------------------------------------
File: S-001
------------------------------------------------
tagSpecID output for peak X881.9.847
m/z   881.9
RT:   847  s
Results suggest 1 likely TAG structure(s)
Suggested TAG structure(s):
        FA1   FA2    FA3    ECN    score
OLL    18:1   18:2   18:2   44     0.6037
------------------------------------------------
DAG fragment ions:
mz      C      DB     Int
599.7  36      4      175903319.4
601.6  36      3      318898212.1
------------------------------------------------
M+H adduct ions:
mz      C      DB     Int
881.9  54      5      96792659.8
------------------------------------------------
--- OLL ----------------------------------------
------------------------------------------------
FAs:    18:1 18:2 18:2
C:DB    54:5
ECN:    44
Elemental composition:   C(57) H(100) O(6)
Score = 0.6037
Molecular adduct:
M+H     881.8
DAG fragment ions (loss of acyl group):
18:1   18:2   18:2
599.5  601.6  601.6
```

Figure 29. Text file output for TAG structure identification for peak X881.9.847 detected in soybean oil sample S-001.

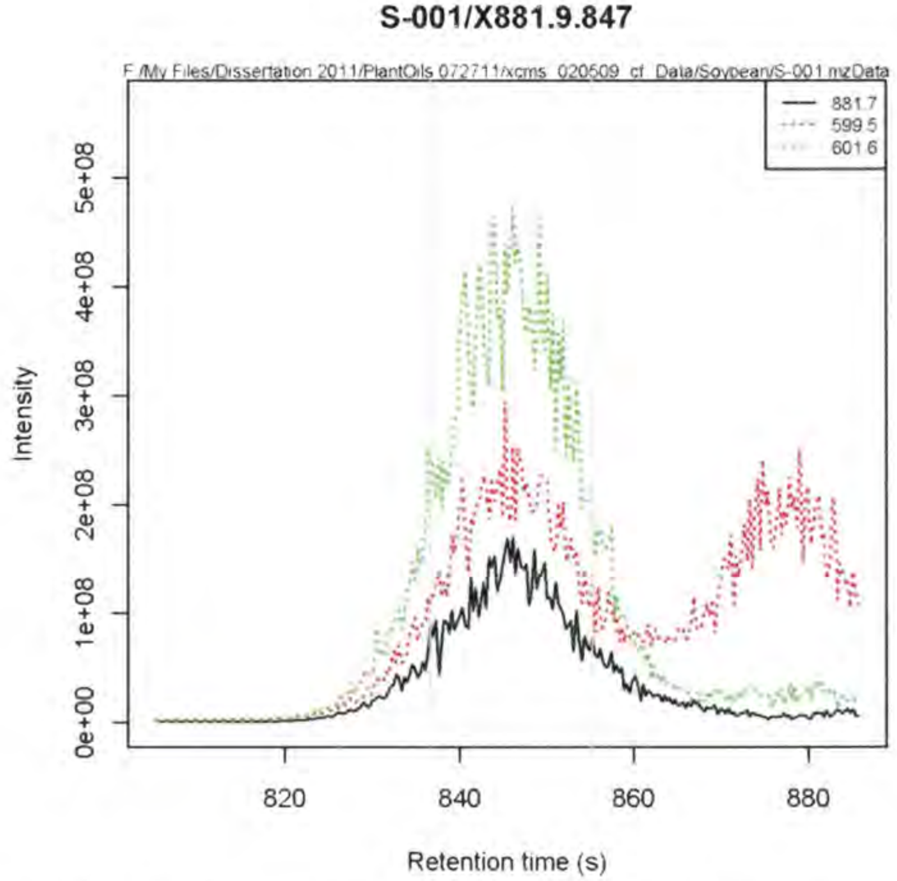


Figure 30. Extracted ion chromatograms for peaks supporting structure identification for feature X881.9.847 detected in sample S-001. The primary peak (m/z 881.9) is shown in solid black, and structure-associated ions are overlaid with dotted lines as indicated in the legend. The grey vertical lines indicate the RT region over which feature X881.9.847 was detected using matched filter peak detection in xcms.

RESULTS AND DISCUSSION

We tested this methodology on two sets of data. In the first example, we analyzed five soybean oils with the aim of identifying major TAG components. As soybean oil TAGs are relatively well-characterized, we can check our TAG structure prediction results against structure assignments from the literature. The second example addresses a differential profiling problem, where we aim to determine the TAGs that differ in adipose tissue samples from mice fed two different high-fat diets in a diet-induced obesity study. We used xcms, ver. 1.14.1, in R, ver. 2.9.1 (9, 10, 87) for the preprocessing of these data.

Xcms functions used for preprocessing, and parameters that differed from the default settings, are listed in Table 17 and Table 18.

Table 17. Xcms functions and associated parameters for preprocessing soybean oils.

Function	Parameters
xcmsSet	sigma=6.5, max=25, steps=5
group	bw=10
retcor	missing=0, extra=0
group	bw=5
fillPeaks	
groupval	method='medret', value='into'

Table 18. Xcms functions and associated parameters for preprocessing mouse AT lipids.

Function	Parameters
xcmsSet	sigma=8, max=25, steps=3, mzdiff=0.7
group	bw=10
retcor	
group	bw=5
fillPeaks	
groupval	method='medret', value='into'

We detected 899 features in the soybean oils (n=5) using xcms, and screened these data for $[DAG]^+$ and $[M+H]^+$ peaks that would be observed for all possible TAG structures containing the most prominent FAs in soybean oil, 16:0 (P), 18:0 (S), 18:1 (O), 18:2 (L), and 18:3 (Ln), to get a list of 77 relevant features. We then applied the structure identification algorithm to these 77 features and compared results among the five samples and with TAGs identified in soybean oil from the literature (4, 66). TAG structures for 41 features were consistently assigned in all five soybean oils. Due to data redundancies, i.e. multiple ions related to one TAG species, these features accounted for 23 TAG species. The total ion chromatogram for soybean oil S-001 with identified TAG species labeled is shown in Figure 31; TAG structures and scores are listed in Table 19.

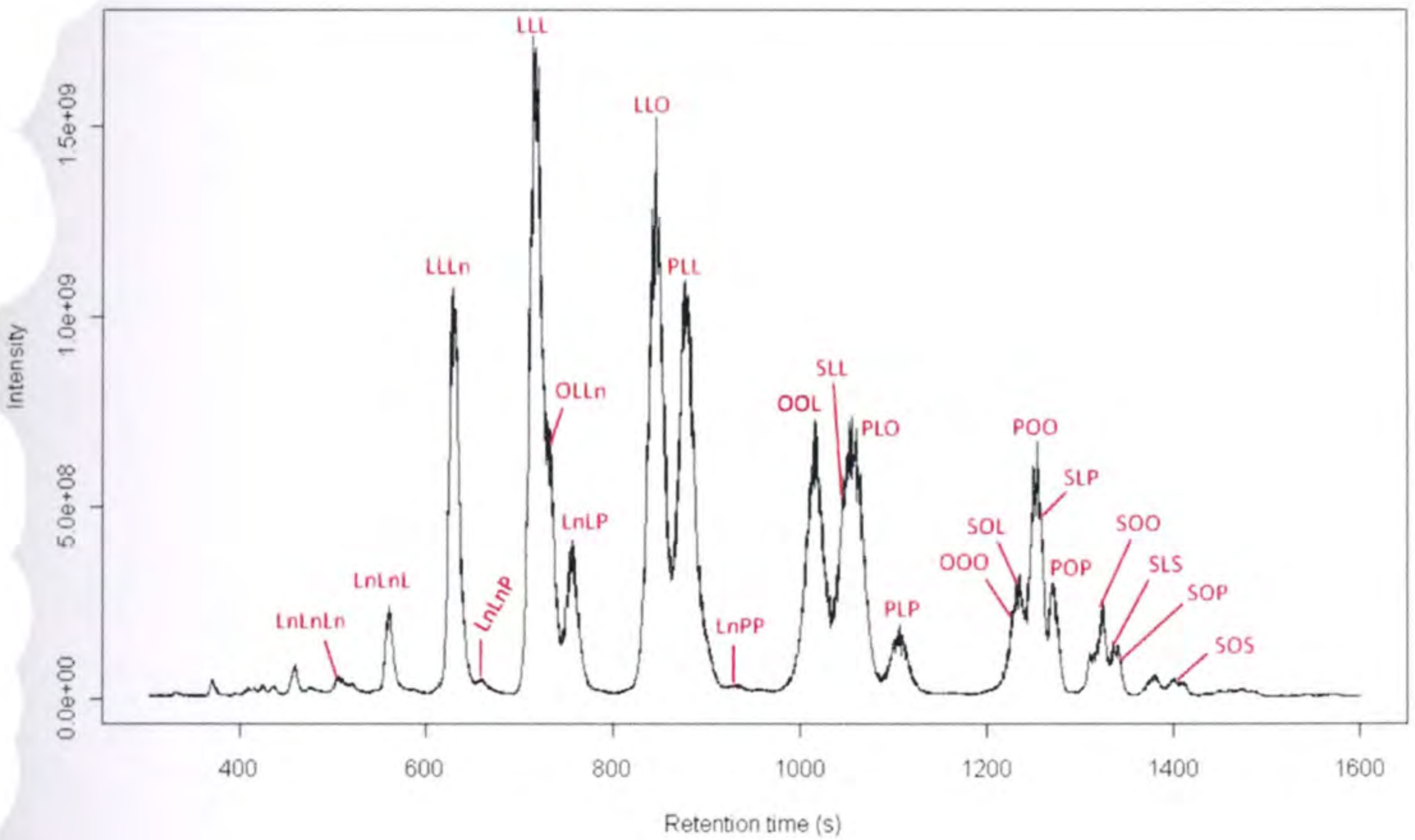


Figure 31. Total ion chromatogram for soybean oil S-001 with peak labels corresponding to the 23 TAG species identified in all five soybean oils. Note: P = palmitate (16:0), S = stearate (18:0), O = oleate (18:1), L = linoleate (18:2), and Ln = linolenate (18:3).

Table 19. TAG structure assignments for features identified in soybean oils.

Feature	RT	m/z	Structure	ECN	Score*	Reference(s)
X595.7.505	505	595.7	LnLnLn	36	0.93 ± 0.03	
X597.6.561	561	597.6	LLnLn	38	0.92 ± 0.02	(66)
X599.6.630	630	599.6	LLnL	40	0.85 ± 0.04	(66)
X877.8.658	658	877.8	LLnL	40	0.92 ± 0.01	(4, 66)
X851.9.659	659	851.9	LnPLn†	40	0.82 ± 0.04	(66)
X573.7.660	660	573.7	LnPLn†	40	0.86 ± 0.02	(66)
X595.6.660	660	595.6	LnPLn†	40	0.87 ± 0.03	(66)
X879.9.717	717	879.9	LLL	42	0.70 ± 0.09	(4, 66)
X599.6.718	718	599.6	LLL	42	0.69 ± 0.08	(66)
X601.6.732	732	601.6	OLLn	42	0.75 ± 0.06	(66)
X575.7.756	756	575.7	LLnP	42	0.90 ± 0.02	(4, 66)
X597.7.757	757	597.7	LLnP	42	0.85 ± 0.05	(66)
X601.6.847	847	601.6	OLL	44	0.73 ± 0.07	(66)
X881.9.847	847	881.9	OLL	44	0.62 ± 0.05	(4, 66)
X575.7.880	880	575.7	LPL	44	0.79 ± 0.02	(66)
X599.6.880	880	599.6	LPL	44	0.71 ± 0.04	(66)
X855.9.881	881	855.9	LPL	44	0.59 ± 0.05	(4, 66)
X573.7.936	936	573.7	LnPP†	44	0.73 ± 0.04	(66)
X829.8.936	936	829.8	LnPP†	44	0.46 ± 0.08	(66)
X603.5.1021	1021	603.5	OLO	46	0.73 ± 0.09	(4, 66)
X883.9.1021	1021	883.9	OLO	46	0.28 ± 0.10	(66)
X603.6.1054	1054	603.6	SLL	46	0.61 ± 0.04	(66)
X857.9.1064	1064	857.9	OLP	46	0.38 ± 0.07	(66)
X601.7.1065	1065	601.7	OLP	46	0.79 ± 0.05	(66)
X577.7.1066	1066	577.7	OLP	46	0.85 ± 0.04	(4, 66)
X831.8.1112	1112	831.8	LPP	46	0.06 ± 0.06	(66)
X551.6.1113	1113	551.6	LPP	46	0.65 ± 0.12	(4, 66)
X575.6.1113	1113	575.6	LPP	46	0.66 ± 0.12	(66)
X603.6.1246	1246	603.6	SOL	48	0.31 ± 0.11	(4, 66)
X885.9.1254	1254	885.9	SOL	48	0.61 ± 0.13	(66)
X605.6.1255	1255	605.6	SOL	48	0.89 ± 0.06	(66)
X577.7.1262	1262	577.7	OPO	48	0.34 ± 0.11	(4, 66)
X859.9.1262	1262	859.9	OPO	48	0.38 ± 0.08	(66)
X579.7.1274	1274	579.7	SLP	48	0.93 ± 0.01	(66)
X833.8.1280	1280	833.8	OPP†	48	0.18 ± 0.10	(66)
X577.6.1281	1281	577.6	OPP†	48	0.88 ± 0.06	(4, 66)
X887.9.1327	1327	887.9	SOO†	50	0.36 ± 0.15	(66)
X605.6.1329	1329	605.6	SOO†	50	0.80 ± 0.15	(4, 66)
X607.6.1337	1337	607.6	SLS	50	0.21 ± 0.05	(66)
X579.7.1345	1345	579.7	SOP	50	0.95 ± 0.01	(66)
X607.6.1414	1414	607.6	SOS	52	0.90 ± 0.01	(66)

* Mean ± SD.

† Observed at relative amount < 1% (66).

The determined TAG structure assignments were compared with two literature references (4, 66) to evaluate the success of the algorithm. All twelve peaks (indexed by RT, m/z) used for classification of plant oils by Jakab et al. (4) were detected using xcms preprocessing, and structure assignments were consistent with the manual assignments in the reference. Of the 24 TAG species observed at relative amounts ≥ 1.0% by Lisa, et al. (66), we identified 20 and our structure assignments were consistent with those in the reference. We identified one peak (X595.7.505, identified as LnLnLn) that was not identified in either of the references, and we were able to identify several TAG species that were present at relative amounts < 1% (LnPLn, LnPP, OPP, and SOO; 66). We suggest that automated data processing and TAG structure identification may be used to achieve results similar to those in the plant oils literature, and that such an automated method may have particular utility for classification problems, e.g. identification and/or authentication of plant oils.

In our second example, we used a differential profiling methodology to determine LC/MS peaks (indexed by RT, m/z) that differ in adipose tissue samples from mice fed two different high-fat diets, containing lard (LD, n=6) and milkfat (MD, n=6), for eight weeks (11). The aim of this example was to determine which TAG species differ in adipose tissue from mice fed the different diets. We detected 2285 features in these samples using xcms, and screened these data for $[DAG]^+$ peaks that would be observed for all possible TAG structures containing the most prominent FAs observed in the adipose tissue and feed samples, 12:0 (La), 14:0 (M), 16:0 (P), 18:0 (S), 16:1 (Po), 18:1

(O), 18:2 (L), and 18:3 (Ln), to get a list of 85 relevant features. We used q-values (45, 46) to rank features according to differences between groups and selected the 20 "most different" features; q-values were obtained from the p-values associated with t-tests for differences in individual variables in animals fed the two different diets (i.e. H_0: LD = MD) using the qvalue package in R (ver. 1.26.0). We then applied the structure identification algorithm to these 20 features to determine the TAG structures associated with peaks that differed between adipose tissue sampled from LD- and MD-fed mice.

We determined the TAG structures for 18 of the 20 features; extracted ion chromatograms for all 20 peaks are shown in Figure 32. All structure assignments were verified manually by examining the files produced by the TAG structure identification algorithm. We determined that the two peaks that could not be identified, X467.4.885 and X495.5.1055, both labeled with "NA," were minor components coeluting with OLL and OOL, respectively. Spectra for these features are shown in Figure 33, and it is obvious that ions resulting from elution of OLL (m/z 599.6, 601.6, and 881.8) and OOL (m/z 601.6, 603.6, and 883.8) dominate the respective spectra. Also, the highest-scoring TAG structures for peaks X493.4.815 and X493.4.849 did not appear to be correct assignments, based on the associated ECN values. As the ECN determines the order of elution for TAGs using RP-HPLC, ECN values should increase with retention times. We noted that the highest scoring structure for peak X493.4.815 was LPoLa, with ECN = 40, and that this structure was also assigned to a peak with the same m/z value that eluted earlier, X493.4.692. Examination of the output file showed OPoLa as the TAG structure

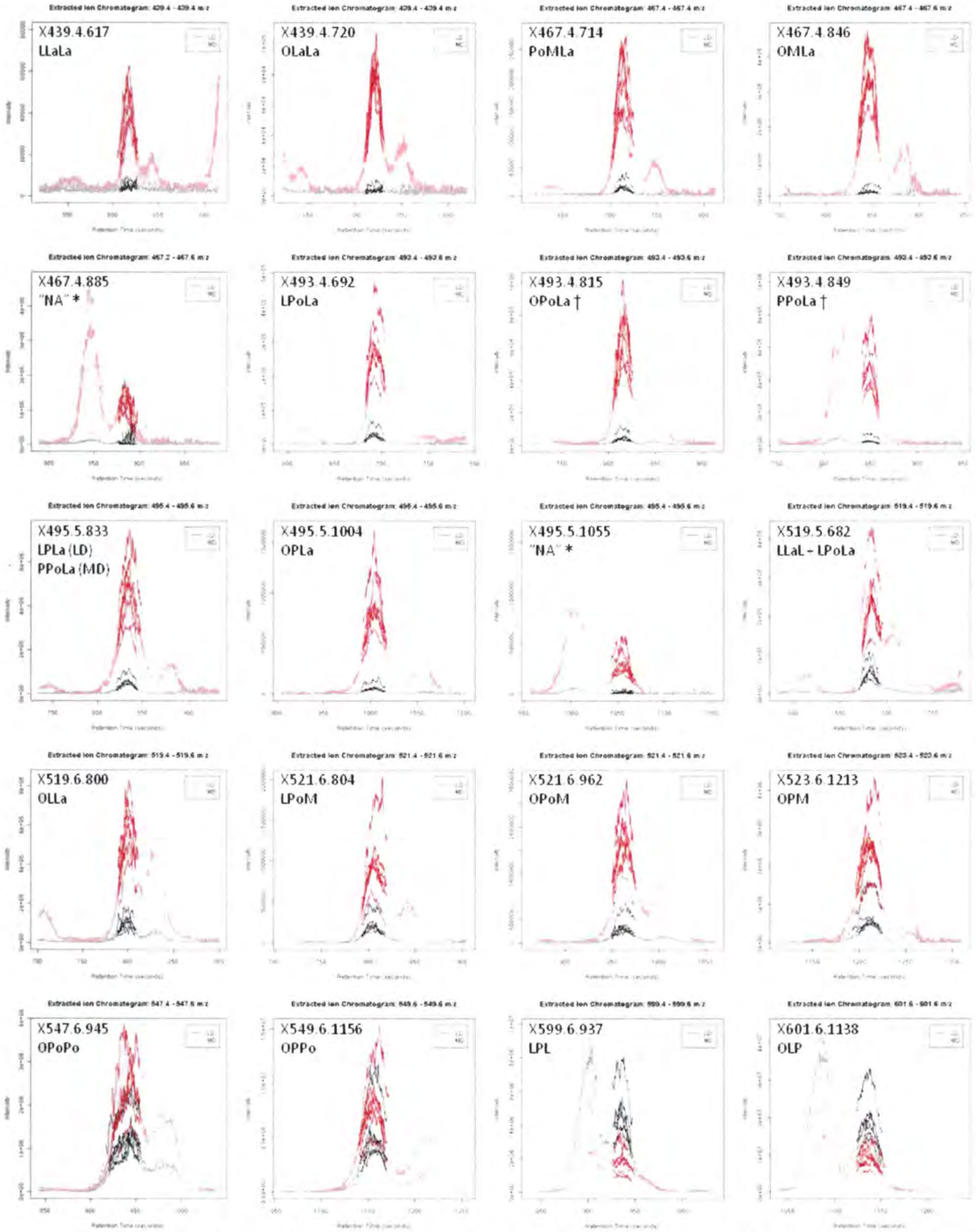


Figure 32. Extracted ion chromatograms for 20 "most different" TAG features identified in adipose tissue samples from MD- and LD-fed mice. * "NA" peak structures could not be determined due to coelution with very prominent TAGs, OLL and OOL. † TAG structure assignments were not structures with highest score.

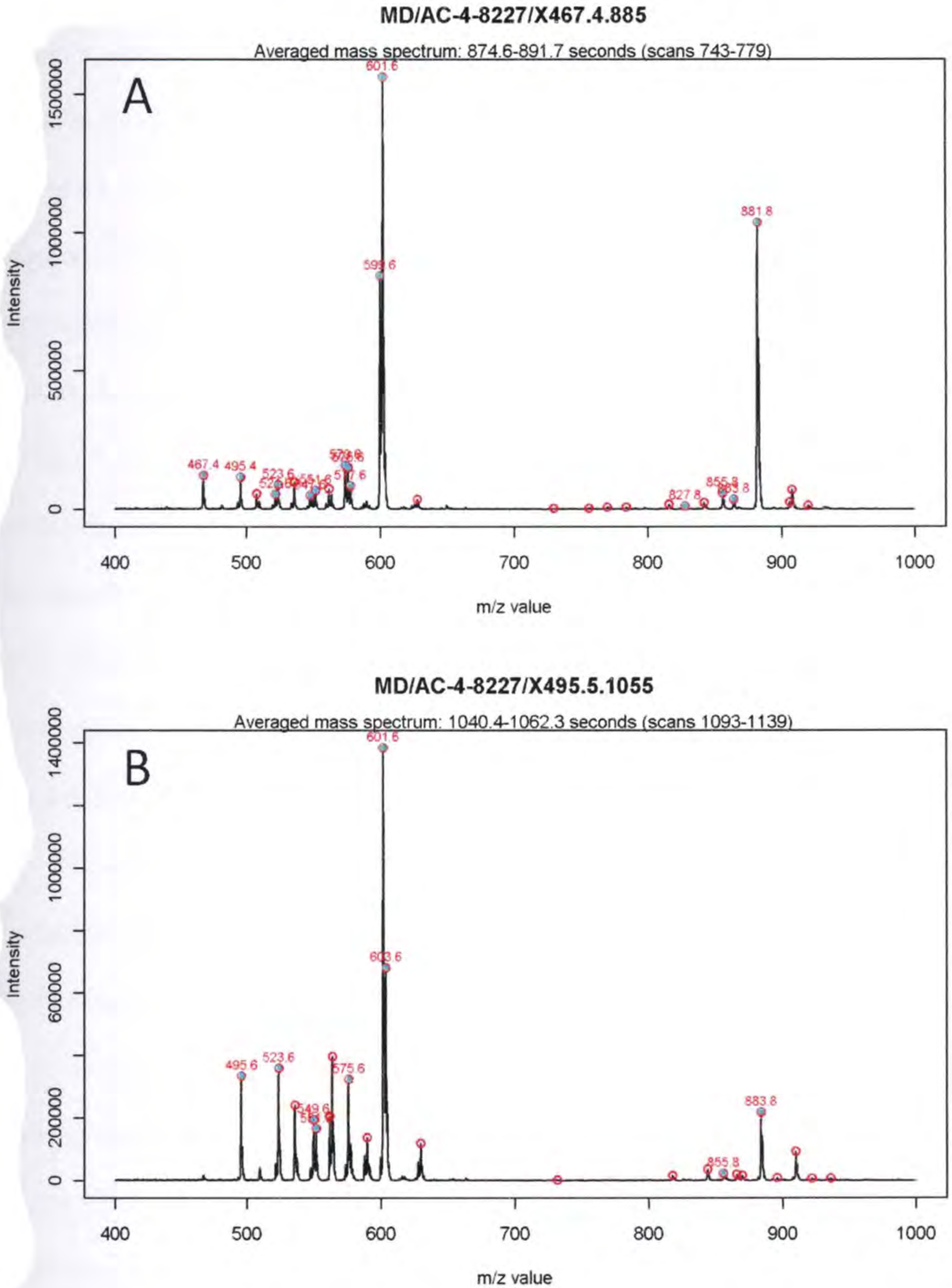


Figure 33. Mass spectra for MD sample AC-4-8227, peaks X467.4.885 and X495.5.1055. Spectra show coelution of multiple TAG species, with large amounts of (A) OLL (m/z 599.6, 601.6, 881.8) and (B) OOL (m/z 601.6, 603.6, 883.8).

with the next highest score and correct ECN (42) for this component. The same process was used to identify PPoLa as the structure for peak X493.4.849. TAG structures for peak X495.5.833 differed by diet group; this peak was identified as LPLa in LD-fed mice and PPoLa in MD-fed mice, though both TAG species may be coeluting in samples from both groups. Peak X519.5.682 appeared to be a mixture of LLaL and LPoLa, with the latter dominating in the MD-fed mice. As we can see from this interpretation, these data are extremely complex. In animal samples, such as the adipose tissue examined here, it is likely that many TAG species are coeluting, and we observe this directly as multiple $[M+H]^+$ ions in the spectra (Figure 33). In this case, the TAG structure assigned to most of the identified peaks is simply the most prevalent of the several TAG species that are present. Plant oils typically have cleaner spectra that are easier to interpret (Figure 28); this is likely due to the fact that plants biosynthesize all of their TAGs, while animals also obtain TAGs (as FAs) from the diet.

The data generated by RP-HPLC/APCI-MS experiments are often overwhelming to researchers analyzing multiple samples or examining differences between groups or classes of samples. While manual data processing may be feasible for studies of plant oil TAG composition or for relatively simple classification tasks (1, 2, 4, 133), an automated framework that integrates data processing, statistical analysis, and structure determination of the FA chain structures of individual TAG species is necessary for more complicated projects such as differential profiling experiments. As LC/MS data processing migrates from proprietary software to open-source packages such as xcms that work on data stored in universal formats (e.g. mzXML, mzML, mzData), tools for high-level data analysis must be developed that work with the data produced from these packages. The integrated framework for the analysis of natural mixtures of TAGs presented here is one such tool. We have combined preprocessing in xcms with subsequent steps for normalization to total area, optional targeted screening for $[DAG]^+$ and $[M+H]^+$ ions observed in TAG structures that may be determined from FA compositional data, difference detection and/or feature selection using q-values (R qvalue package), and structural determination of FA using an automated structure determination algorithm (135) combined with the MassSpecWavelet package in R. We use multiple correlations of the primary peak (m/z) intensity values with structure-supporting ions to assign a score to the identified TAG structure(s). Data are formatted for classification tasks at several points during the analysis workflow (Figure 27). This package was written in R and is available from the corresponding author upon request.

LIST OF REFERENCES

1. Jakab, A., K. Nagy, K. Heberger, et al. "Differentiation of vegetable oils by mass spectrometry combined with statistical analysis." Rapid Communications in Mass Spectrometry. 16:2291-2297, 2002.

2. Lisa, M., M. Holcapek, and M. Bohac. "Statistical evaluation of triacylglycerol composition in plant oils based on high-performance liquid chromatography--atmospheric pressure chemical ionization mass spectrometry data." Journal of Agricultural and Food Chemistry. 57:6888-6898, 2009.

3. Holcapek, M., P. Jandera, P. Zderadicka, et al. "Characterization of triacylglycerol and diacylglycerol composition of plant oils using high-performance liquid chromatography-atmospheric pressure chemical ionization mass spectrometry." Journal of Chromatography A. 1010:195-215, 2003.

4. Jakab, A., K. Heberger, and E. Forgacs. "Comparative analysis of different plant oils by high-performance liquid chromatography-atmospheric pressure chemical ionization mass spectrometry." Journal of Chromatography A. 976:255-263, 2002.

5. Christie, W.W. Lipid Analysis: Isolation, separation, identification and structural analysis of lipids. 3rd ed. Bridgewater, England: PJ Barnes & Associates, 2003, pp. 3, 5, 12, 301.

6. Li, X.-j., E.C. Yi, C.J. Kemp, et al. "A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry." Molecular & Cellular Proteomics. 4:1328-1340, 2005.

7. Bellew, M., M. Coram, M. Fitzgibbon, et al. "A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS." Bioinformatics. 22:1902-1909, 2006.

8. Katajamaa, M., J. Meiettinen, and M. Oresic. "MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data." Bioinformatics. 22:634-636, 2006.

9. Smith, C.A., E.J. Want, G. O'Maille, et al. "XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification." Analytical Chemistry. 78:779-787, 2006.

10. Sturm, M., A. Bertsch, C. Gropl, et al. "OpenMS - an open-source software framework for mass spectrometry." BMC Bioinformatics. 9:163, 2008.

11. Geng, T., W. Hu, J. Snider, et al. "A new diet-induced obesity model reveals dietary fat profile regulates hepatic ER stress and insulin resistance." Diabetes, in revision, 2012.

12. Gunstone, F.D., and B.G. Herslof. A Lipid Glossary. Ayr, Scotland: The Oily Press Ltc., 1992, p. 95.

13. Holmer, G. "Triglycerides." In: Marine Biogenic Lipids, Fats, and Oils, edited by R.G. Ackman. Boca Raton: CRC Press, Inc., 1989, pp. 139-174.

14. Pond, C.M. The Fats of Life. Cambridge: Cambridge University Press, 1998, p. 28.

15. Jobling, M. "Feeding and nutrition in intensive fish farming." In: Biology of Farmed Fish, edited by K.D. Black and A.D. Pickering. Boca Raton: CRC Press LLC, 1998, pp. 67-113.

16. Ackman, R.G. "Marine lipids and omega-3 fatty acids." In: Handbook of Functional Lipids, edited by C.C. Akoh. Boca Raton: Taylor & Francis, 2005, p. 311.

17. Eskin, N.A.M. "Authentication of evening primrose, borage and fish oils." In: Oils and Fats Authentication, edited by M. Jee. Boca Raton: CRC Press LLC, 2002, pp. 95-114.

18. Shahidi, F., and H. Miraliakbari. "Marine Oils: Compositional characteristics and health effects." In: Nutraceutical and Specialty Lipids and their Co-Products, edited by F. Shahidi. Boca Raton: CRC Press Taylor & Francis Group, 2006, pp. 227-250.

19. Koopman, H.N. The structure and function of the blubber of odontocetes: Duke University (Ph.D. Dissertation), 2001.

20. Brockerhoff, H., R.J. Hoyle, P.C. Hwang, et al. "Positional distribution of fatty acids in depot triglycerides of aquatic aminals." Lipids. 3:24-29, 1968.

21. Litchfield, C. Analysis of Triglycerides. New York: Academic Press, Inc., 1972.

22. Martin, A.J., and R.L. Synge. "A new form of chromatogram employing two liquid phases: A theory of chromatography. 2. Application to the micro-determination of the higher monoamino-acids in proteins." Biochemical Journal. 35:1358-1368, 1941.

23. James, A.T., and A.J.P. Martin. "Gas-liquid partition chromatography: the separation and micro-estimation of volatile fatty acids from formic acid to dodecanoic acid." Biochemical Journal. 50:679-690, 1952.

24. Christie, W.W. Gas Chromatography and Lipids. Ayr, Scotland: The Oily Press Ltd., 1989, pp. 1-8.

25. Brockerhoff, H. "A stereo specific analysis of triglycerides." Journal of Lipid Research. 79:10-15, 1965.

26. Brockerhoff, H. "Stereospecific analysis of triglycerides: an alternative method." Journal of Lipid Research. 8:167-169, 1967.

27. Mottram, H.R. "Regiospecific analysis of triacylglycerols using High performance liquid chromatography/Atmospheric pressure chemical ionization mass spectrometry." In: Modern Methods for Lipid Analysis by Liquid Chromatography/Mass Spectrometry and Related Techniques, edited by W.C. Byrdwell. Champaign, IL: AOCS Press, 2005, pp. 276-297.

28. Byrdwell, W.C., and E.A. Emken. "Analysis of triglycerides using atmospheric-pressure chemical-ionization mass spectrometry." Lipids. 30:173-175, 1995.

29. Mottram, H.R., and R.P. Evershed. "Structure analysis of triacylglycerol positional isomers using atmospheric pressure chemical ionisation mass spectrometry." Tetrahedron Letters. 37:8593-8596, 1996.

30. Fauconnot, L., J. Hau, J.-M. Aeschlimann, et al. "Quantitative analysis of triacylglycerol regioisomers in fats and oils using reversed-phase high-performance liquid chromatography and atmospheric pressure chemical ionization mass spectrometry." Rapid Communications in Mass Spectrometry. 18:218-224, 2004.

31. Jakab, A., I. Jablonkai, and E. Forgacs. "Quantification of the ratio of positional isomer dilinoleoyl-oleoyl glycerols in vegetable oils." Rapid Communications in Mass Spectrometry. 17:2295-2302, 2003.

32. Listgarten, J., and A. Emili. "Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry." Molecular & Cellular Proteomics. 4:419-434, 2005.

33. Jonsson, P., S.J. Bruce, T. Moritz, et al. "Extraction, interpretation and validation of information for comparing samples in metabolic LC/MS data sets." The Analyst. 130:701-707, 2005.

34. Katajamaa, M., and M. Oresic. "Data processing for mass spectrometry-based metabolomics." Journal of Chromatography A. 1158:318-328, 2007.

35. Enot, D.P., B. Haas, and K.M. Weinberger. "Bioinformatics for mass spectrometry-based metabolomics." Methods in Molecular Biology. 719:351-375, 2011.

36. Want, E. "Processing and analysis of GC/LC-MS-based metabolomics data." In: Metabolic Profiling: Methods and Protocols, edited by T.O. Metz. New York: Humana Press, 2011, pp. 277-298.

37. Polikar, R. "Pattern recognition." In: Wiley Encyclopedia of Biomedical Engineering, edited by M. Akay. Hoboken: Wiley-Interscience, 2006, pp. 1-22.

38. Thomas, A., G.D. Tourassi, A.S. Elmaghraby, et al. "Data mining in proteomic mass spectrometry." Clinical proteomics. 2:13-32, 2006.

39. Listgarten, J., R.M. Neal, S.T. Roweis, et al. "Difference detection in LC-MS data for protein biomarker discovery." Bioinformatics. 23:e198-e204, 2006.

40. Pedrioli, P.G.A., J.K. Eng, R. Hubley, et al. "A common open representation of mass spectrometry data and its application to proteomics research." Nature Biotechnology. 22:1459-1466, 2004.

41. Katajamaa, M., and M. Oresic. "Processing methods for differential analysis of LC/MS profile data." BMC Bioinformatics. 6, 2005.

42. Morris, J.S., K.R. Coombes, J. Koomen, et al. "Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum." Bioinformatics. 21:1764-1775, 2005.

43. Randolph, T.W., and Y. Yasui. "Multiscale processing of mass spectrometry data." Biometrics. 62:589-597, 2006.

44. Danielsson, R., D. Bylund, and K.E. Markides. "Matched filtering with background suppression for improved quality of base peak chromatograms and mass spectra in liquid chromatography-- mass spectrometry." Analytica Chimica Acta. 454:167-184, 2002.

45. Storey, J.D., and R. Tibshirani. "Statistical significance for genomewide studies." Proceedings from the National Academy of Sciences. 100:9440-9445, 2003.

46. Storey, J.D. "A direct approach to false discovery rates." Journal of the Royal Statistical Society B. 64:479-498, 2002.

47. Benjamini, Y., and Y. Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." Journal of the Royal Statistical Society B. 57:289-300, 1995.

48. Breiman, L. "Random forests." Machine Learning. 45:5-32, 2001.

49. Wu, B., T. Abbott, D. Fishman, et al. "Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data." Bioinformatics. 19:1636-1643, 2003.

50. Yu, W., B. Wu, T. Huang, et al. "Statistical methods in proteomics." In: Springer Handbook of Engineering Statistics, edited by H. Pham. London: Springer, 2006, pp. 623-638.

51. Liaw, A., and M. Wiener. "Classification and regression by randomForest." R News. 2:18-22, 2002.

52. Holcapek, M., M. Lisa, P. Jandera, et al. "Quantitation of triacylglycerols in plant oils using HPLC with APCI-MS, evaporative light-scattering, and UV detection." Journal of Separation Science. 28:315-333, 2005.

53. Cunha, S.C., and M.B.P.P. Oliveira. "Discrimination of vegetable oils by triacylglycerols evaluation of profile using HPLC/ELSD." Food Chemistry. 95:518-524, 2006.

54. Andrikopoulos, N.K. "Triglyceride species compositions of common edible vegetable oils and methods used for their identification and quantification." Food Reviews International. 18:71-102, 2002.

55. Byrdwell, W.C. "Atmospheric pressure chemical ionization mass spectrometry for analysis of lipids." Lipids. 36:327-346, 2001.

56. Byrdwell, W.C. "Qualitative and quantitative analysis of triacylglycerols by Atmospheric pressure ionization (APCI and ESI) mass spectrometry techniques." In: Modern Methods for Lipid Analysis by Liquid Chromatography/Mass Spectrometry and Related Techniques, edited by W.C. Byrdwell. Champaign, IL: AOCS Press, 2005, pp. 298-412.

57. Leskinen, H., J.P. Suomela, and H. Kallio. "Quantification of triacylglycerol regioisomers in oils and fat using different mass spectrometric and liquid

chromatographic methods." Rapid Communications in Mass Spectrometry. 21:2361-2373, 2007.

58. Byrdwell, W.C. "The bottom-up solution to the triacylglycerol lipidome using atmospheric pressure chemical ionization mass spectrometry." Lipids. 40:383-417, 2005.

59. Hastie, T., R. Tibshirani, and J. Friedman. The Elements of Statistical Learning. New York: Springer-Verlag, 2001, pp. 79-94.

60. America, A.H., and J.H. Cordewener. "Comparative LC-MS: a landscape of peaks and valleys." Proteomics. 8:731-749, 2008.

61. Codrea, M.C., C.R. Jimenez, J. Heringa, et al. "Tools for computational processing of LC-MS datasets: a user's perspective." Computational Methods and Programs in Biomedicine. 86:281-290, 2007.

62. Dunn, W.B., D. Broadhurst, M. Brown, et al. "Metabolic profiling of serum using Ultra Performance Liquid Chromatography and the LTQ-Orbitrap mass spectrometry system." Journal of Chromatography B. 871:288-298, 2008.

63. Kind, T., V. Tolstikov, O. Fiehn, et al. "A comprehensive urinary metabolomic approach for identifying kidney cancer." Analytical Biochemistry. 363:185-195, 2007.

64. Schwab, U., T. Seppanen-Laakso, L. Yetukuri, et al. "Triacylglycerol fatty acid composition in diet-induced weight loss in subjects with abnormal glucose metabolism--the GENOBIN study." PLoS ONE. 3:e2630, 2008.

65. Jankevics, A., E. Liepinsh, E. Liepinsh, et al. "Metabolomic studies of experimental diabetic urine samples by 1H NMR spectroscopy and LC/MS method." Chemometrics and Intelligent Laboratory Systems. 97:11-17, 2009.

66. Lisa, M., and M. Holcapek. "Triacylglycerols profiling in plant oils important in food industry, dietetics and cosmetics using high-performance liquid chromatography–atmospheric pressure chemical ionization mass spectrometry." Journal of Chromatography A. 1198-1199:115-130, 2008.

67. Barnes, P.M., B. Bloom, and R. Nahin. Complementary and alternative medicine use among adults and children: United States, 2007. Atlanta: Centers for Disease Control and Prevention, CDC National Health Statistics #12, 2008.

68. Shahidi, F., P.K.J.P.D. Wanasundara, and U.N. Wanasundara. "Seal blubber oil: a novel source of w3 fatty acids." Journal of Food Lipids. 3:293-306, 1996.

69. Ackman, R.G., and W.M.N. Ratnayake. "Fish oils, seal oils, esters and acids - are all forms of omega-3 intake equal?" In: Health effect of fish and fish oils, edited by R.K. Chandra. St. John's, Newfoundland: Arts Biomedical Publishers & Distributors, 1989, pp. 373-393.

70. Fisheries and Oceans Canada. Overview of the Atlantic Seal Hunt 2006-2010. 2010.

71. Ackman, R.G. "Safety of seal oil as a nutritional supplement." Proceedings of the Nova Scotian Institute of Science. 41:103-114, 1997.

72. Marine Mammal Protection Act of 1972 (as amended 2007), 16 U.S.C. § 1361 et seq, 1401-1407, 1538, 4107.

73. Lewanowicz, C., and R. Freedman. European Parliament Press Release: "MEPs adopt strict conditions for the placing on the market of seal products in the European Union." Press Service, Directorate for the Media; 2009.

74. Laakso, P., and P. Manninen. "Mass spectrometric techniques in the analysis of triacylglycerols." In: Spectral Properties of Lipids, edited by R.J. Hamilton and J. Cast. Boca Raton: CRC Press LLC, 1999, pp. 141-190.

75. Recks, M.A., and G.T. Seaborn. "Variation in fatty acid composition among nine forage species from a southeastern US estuarine and nearshore coastal ecosystem." Fish Physiology and Biochemistry. 34:275-287, 2008.

76. Iverson, S.J., K.J. Frost, and S.L.C. Lang. "Fat content and fatty acid composition of forage fish and invertebrates in Prince William Sound, Alaska: factors contributing to among and within species variability." Marine Ecology Progress Series. 241:161-181, 2002.

77. Budge, S.M., S.J. Iverson, W.D. Bowen, et al. "Among- and within-species variability in fatty acid signatures of marine fish and invertebrates on the Scotian Shelf, Georges Bank, and southern Gulf of St. Lawrence." Canadian Journal of Fisheries and Aquatic Science. 59:886-898, 2002.

78. Walton, M.J., R.J. Henderson, and P.P. Pomeroy. "Use of blubber fatty acid profiles to distinguish dietary differences between grey seals *Halichoerus grypus* from two UK breeding colonies " Marine Ecology Progress Series. 193:201-208, 2000.

79. Smith, R.J., K.A. Hobson, H.N. Koopman, et al. "Distinguishing between populations of fresh- and salt-water harbour seals (*Phoca vitulina*) using stable-isotope ratios and fatty acid profiles." Canadian Journal of Fisheries and Aquatic Science. 53:272-279, 1996.

80. Seaborn, G.T., M.L. Jahncke, and T.I.J. Smith. "Differentiation between cultured hybrid striped bass and wild striped bass and hybrid bass using fatty acid profiles." North American Journal of Fisheries Management. 20:618-626, 2000.

81. Jahncke, M.L., T.I.J. Smith, and G.T. Seaborn. "Use of fatty acid profiles to distinguish cultured from wild fish: A possible law enforcement tool." Proceedings of the Annual Conference of Southeast Association of Fish and Wildlife Agencies. 42:546-553, 1988.

82. Grahl-Nielsen, O., T. Haug, U. Lindstrom, et al. "Fatty acids in harp seal blubber do not necessarily reflect their diet." Marine Ecology Progress Series. 426:263-276, 2011.

83. Ackman, R.G., S. Epstein, and C.A. Eaton. "Differences in the fatty acid compositions of blubber fats from northwestern Atlantic finwhales (*Baleanoptera physalus*) and harp seals (*Pagophilus groenlandica*)." Comparative Biochemistry and Physiology B. 40:683-697, 1971.

84. Engelhardt, F.R., and B.L. Walker. "Fatty acid composition of the harp seal, *Pagophilus groenlandicus (Phoca groenlandica)*." Comparative Biochemistry and Physiology. 47B:169-179, 1974.

85. Seaborn, G., T.I.J. Smith, M.R. Denson, et al. "Comparative fatty acid composition of eggs from wild and captive black sea bass, Centropristis striata L." Aquaculture Research. 40:656-668, 2009.

86. Metcalfe, L.D., A.A. Schmitz, and J.R. Pelka. "Rapid preparation of fatty acid esters from lipids for gas chromatographic analysis." Analytical Chemistry. 38:514-515, 1966.

87. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2009.

88. Ding, C., and H. Peng. "Minimum redundancy feature selection from microarray gene expression data." Journal of Bioinformatics and Computational Biology. 3:185-205, 2005.

89. Giskeodegard, G.F., M.T. Grinde, B. Sitter, et al. "Multivariate modeling and prediction of breast cancer prognostic factors using MR metabolomics." Journal of Proteome Research. 9:972-979, 2010.

90. Oksanen, J.F., G. Blanchet, R. Kindt, et al. vegan: Community Ecology Package, R package version 1.17-0. 2010.

91. McArdle, B.H., and M.J. Anderson. "Fitting multivariate models to community data: a comment on distance-based redundancy analysis." Ecology. 82:290-297, 2001.

92. Anderson, M.J. "A new method for non-parametric multivariate analysis of variance." Austral Ecology. 26:32-46, 2001.

93. Faith, D.P., P.R. Minchin, and L. Belbin. "Computational dissimilarity as a robust measure of ecological distance." Vegetatio. 69:57-68, 1987.

94. Venables, W.N., and B.D. Ripley. Modern Applied Statistics with S. 4th ed. New York: Springer, 2002.

95. Cutler, D.R., T.C. Edwards, Jr., K.H. Beard, et al. "Random forests for classification in ecology." Ecology. 88:2783-2792, 2007.

96. Moffat, C.F., and A.S. McGill. "Variability of the composition of fish oils: significance for the diet." Proceedings of the Nutrition Society. 52:441-456, 1993.

97. Gajda, A.M., M.A. Pellizzon, M.R. Ricci, et al. "Diet-induced metabolic syndrome in rodent models." Animal Lab News, March, 2007, p. 5.

98. Panchal, S.K., and L. Brown. "Rodent models for metabolic syndrome research." Journal of Biomedicine and Biotechnology. 2011:1-14, 2011.

99. Buettner, R., J. Scholmerich, and L.C. Bollheimer. "High-fat diets: modeling the metabolic disorders of human obesity in rodents." Obesity. 15:798-808, 2007.

100. Gillingham, L.G., S. Harris-Janz, and P.J. Jones. "Dietary monounsaturated fatty acids are protective against metabolic syndrome and cardiovascular disease risk factors." Lipids. 46:209-228, 2011.

101. Moussavi, N., V. Gavino, and O. Receveur. "Could the quality of dietary fat, and not just its quantity, be related to risk of obesity?" Obesity. 16:7-15, 2008.

102. Kris-Etherton, P., S.R. Daniels, R.H. Eckel, et al. "AHA scientific statement: summary of the Scientific Conference on Dietary Fatty Acids and Cardiovascular Health. Conference summary from the Nutrition Committee of the American Heart Association." Journal of Nutrition. 131:1322-1326, 2001.

103. Riserus, U., W.C. Willett, and F.B. Hu. "Dietary fats and prevention of type 2 diabetes." Progress in Lipid Research. 48:44-51, 2009.

104. Hu, W., J.S. Ross, T. Geng, et al. "Differential regulation of dihydroceramide desaturase by palmitate vs. monounsaturated fatty acids:Implications to insulin resistance." Journal of Biological Chemistry. 286:16596-16605, 2011.

105. Peng, G., L. Li, Y. Liu, et al. "Oleate Blocks Palmitate-Induced Abnormal Lipid Distribution, Endoplasmic Reticulum Expansion and Stress, and Insulin Resistance in Skeletal Muscle." Endocrinology. 152:2206-2218, 2011.

106. Wen, H., D. Gris, Y. Lei, et al. "Fatty acid-induced NLRP3-ASC inflammasome activation interferes with insulin signaling." Nature Immunology. 12:408-415, 2011.

107. Cowart, L.A. "Sphingolipids: players in the pathology of metabolic disease." Trends in Endocrinology and Metabolism. 20:34-42, 2009.

108. Du, K., S. Herzig, R.N. Kulkarni, et al. "TRB3: a tribbles homolog that inhibits Akt/PKB activation by insulin in liver." Science. 300:1574-1577, 2003.

109. Hodson, L., C.M. Skeaff, and B.A. Fielding. "Fatty acid composition of adipose tissue and blood in humans and its use as a biomarker of dietary intake." Progress in Lipid Research. 47:348-380, 2008.

110. Papamandjaris, A.A., D.E. Macdougall, and P.J.H. Jones. "Medium chain fatty acid metabolism and energy expenditure: Obesity treatment implications." Life Sciences. 62:1203-1215, 1998.

111. Bach, A.C., and V.K. Babayan. "Medium-chain triglycerides: an update." American Journal of Clinical Nutrition. 36:950-962, 1982.

112. Perona, J.S., M.P. Portillo, M.T. Macarulla, et al. "Influence of different dietary fats on triacylglycerol deposition in rat adipose tissue." British Journal of Nutrition. 84:765-774, 2000.

113. Ahmadian, M., Y. Wang, and H.S. Sul. "Lipolysis in adipocytes." International Journal of Biochemistry and Cell Biology. 42:555-559, 2010.

114. Raclot, T. "Selective mobilization of fatty acids from white fat cells: evidence for a relationship to the polarity of triacylglycerols." Biochemical Journal. 322:483-489, 1997.

115. Berg, J.M., J.L. Tymoczko, and L. Stryer. Biochemistry. 6th ed. New York: W. H. Freeman and Company, 2007.

116. Trujillo, M.E., and P. Scherer. "Adipose tissue-derived factors: Impact on health and disease." Endocrine Reviews. 27:762-778, 2006.

117. Frayn, K.N. "Adipose tissue as a buffer for daily lipid flux." Diabetologia. 45:1201-1210, 2002.

118. Cook, H. "Faty acid desaturation and chain elongation in eukaryotes." In: Biochemistry of Lipids, Lipoproteins and Membranes, edited by D.E. Vance and J.E. Vance. Amsterdam: Elsevier Science, 1996, pp. 129-152.

119. Byrdwell, W.C., and W.E. Neff. "Qualitative and quantitative analysis of triacylglycerols using Atmospheric-pressure chemical ionization mass spectrometry." In: New Techniques and Applications in Lipid Analysis, edited by R.E. McDonald and M.M. Mossoba. Champaign, IL: AOCS Press, 1997, pp. 45-80.

120. Byrdwell, W.C. "Atmospheric pressure ionization techniques in modern lipid analysis." In: Modern Methods for Lipid Analysis by Liquid Chromatography/Mass Spectrometry and Related Techniques, edited by W.C. Byrdwell. Champaign, IL: AOCS Press, 2005, pp.1-18.

121. Folch, J., M. Lees, and G.H.S. Stanley. "A simple method for the isolation and purification of total lipides from animal tissues." Journal of Biological Chemistry. 226:497-509, 1957.

122. Metcalfe, L.D., and A.A. Schmitz. "The rapid preparation of fatty acid esters for gas chromatographic analysis." Analytical Chemistry. 33:363-364, 1961.

123. Suzuki, S., S. Ishikawa, K. Arihara, et al. "Molecular species-specific differences in composition of triacylglycerols of mouse adipose tissue and diet." Nutrition Research. 28:258-262, 2008.

124. Dugo, P., M. Beccaria, N. Fawzy, et al. "Mass spectrometric elucidation of triacylglycerol content of Brevoortia tyrannus (menhaden) oil using non-aqueous reversed-phase liquid chromatography under ultra high pressure conditions." Journal of Chromatography A, March 27 (epub ahead of print), 2012.

125. Benton, H.P., D.M. Wong, S.A. Trauger, et al. "XCMS2: processing tandem mass spectrometry data for metabolite identification and structural characterization." Analytical Chemistry. 80:6382-6389, 2008.

126. Pluskal, T., S. Castillo, A. Villar-Briones, et al. "MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry based molecular profile data." BMC Bioinformatics. 11:395, 2010.

127. Martens, L., M. Chambers, M. Sturm, et al. "mzML--a community standard for mass spectrometry data." Molecular and Cellular Proteomics. 10:110-133, 2011.

128. Mutch, D.M., G. O'Maille, W.R. Wikoff, et al. "Mobilization of pro-inflammatory lipids in obese Plscr3-deficient mice." Genome Biology. 8:R38, 2007.

129. Yetukuri, L., M. Katajamaa, G. Medina-Gomez, et al. "Bioinformatics strategies for lipidomics analysis: characterization of obesity related hepatic steatosis." BMC Systems Biology. 1:12, 2007.

130. Oresic, M. "Metabolomics, a novel tool for studies of nutrition, metabolism, and lipid dysfunction." Nutrition, Metabolism, and Cardiovascular Disease. 19:816-824, 2009.

131. Oresic, M., V.A. Hanninen, and A. Vidal-Puig. "Lipidomics: a new window to biomedical frontiers." Trends in Biotechnology. 26:647-652, 2008.

132. Byrdwell, W.C. "Atmospheric pressure chemical ionization mass spectrometry (APCI-MS) in lipid analysis." In: Advances in Lipid Methodology-- Five, edited by R.O. Adlof. Bridgewater, England: The Oily Press, 2003, pp. 171-254.

133. Holcapek, M., and M. Lisa. "Statistical evaluation of triacylglycerol composition by HPLC/APCI-MS." Lipid Technology. 21:261-265, 2009.

134. Du, P., Kibbe, Warren A. and Lin, Simon M. "Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching." Bioinformatics:2059-2065, 2006.

135. Cvacka, J., E. Krafkova, P. Jiros, et al. "Computer-assisted interpretation of atmospheric pressure chemical ionization mass spectra of triacylglycerols." Rapid Communications in Mass Spectrometry. 20:3586-3594, 2006.

136. Zar, J.H. Biostatistical Analysis. Fourth ed. Upper Saddle River, NJ: Simon & Schuster, 1999, p. 423.