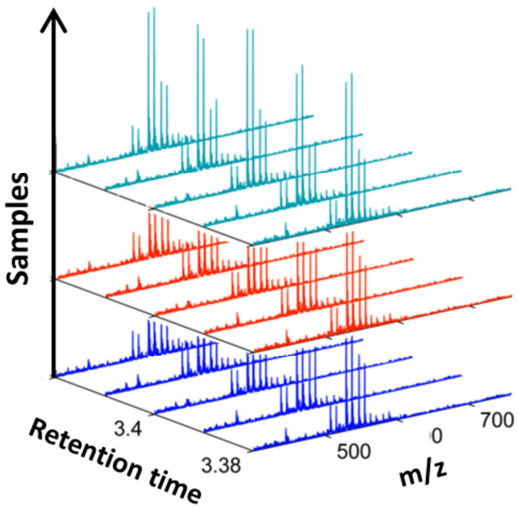
FACULTY OF SCIENCE
UNIVERSITY OF COPENHAGEN

PHD THESIS · GÖZDE GÜRDENIZ

Data Handling Strategies in Nutritional Metabolomics

– illustrated using metabolic states and trans-fat exposures.



FACULTY OF SCIENCE
UNIVERSITY OF COPENHAGEN

PhD thesis

Gözde Gürdeniz

Data Handling Strategies in Nutritional Metabolomics

– illustrated using metabolic states and trans-fat exposures

Academic advisors: Lars Ove Dragsted & Rasmus Bro

Submitted: 07/12/12

Name of department:	Department of Nutrition, Exercise and Sports Department of Food Science
Author:	Gözde Gürdeniz
Title / Subtitle:	Data Handling Strategies in Nutritional Metabolomics – illustrated using metabolic states and trans-fat exposures
Academic advisors:	Professor Lars Ove Dragsted, Department of Nutrition, Exercise and Sports, University of Copenhagen Professor Rasmus Bro, Department of Food Science, University of Copenhagen
Assessment committee:	Professor Søren Balling Engelsen, Department of Food Science, University of Copenhagen Dr Claudine Manach, Human Nutition Unit, University of Auvergne Dr Lars Nørgaard, FOSS, Denmark

ISBN 978-87-7611-561-6
Printed by SL grafik, Frederiksberg C, Denmark

Preface and Acknowledgements

This PhD has been conducted at the Department of Nutrition, Exercise and Sports, Faculty of Science in collaboration with Quality and Technology (Q&T) section, Department of Food Science, Faculty of Science in University of Copenhagen. The grant was provided by the Danish Obesity Research Centre (DanORC, see www.danorc.dk). DanORC is supported by the Danish Council for Strategic Research. The project has been supervised by Professor Lars Ove Dragsted and Professor Rasmus Bro from University of Copenhagen.

I am grateful to my two supervisors for giving me the opportunity to conduct this research. I would like to thank Lars for introducing me to the world of metabolomics. He was always open to scientific discussions with his never ending enthusiasm. Thank you for supporting me when it is most needed. I thank Rasmus for inspiring me with new ideas in multivariate data analysis and challenging me to improve my research skills.

I was really lucky to work with great people in our small metabolomics group; Daniela, Maj-Britt, Jan and Mette. Thank you Jan for all our scientific discussions, Maj-Britt for being supportive both as a colleague and as a friend. Special thanks to Daniela for providing me her unconditional friendship and support.

From Q&T, I am grateful for the support that I had from Thomas Skov and Evrim Acar in improving my multivariate data analysis skills. Also I would like to thank Frans van den Berg for keeping his door open for any kind of scientific questions.

I would like to thank my friends Daniela Rago, Hamid Babamoradi, Bekzod Khakimov, Anita Nielsen, Stinne Larsen, Jose Manuel Amigo, Francesco Savorani and others for reminding me that there are other things than work in Copenhagen.

Finally I'd like to thank my parents, who tolerated my absence for many years. They always put their children before everything else and I would not have made it this far without them.

Gözde Gürdeniz

Copenhagen, December, 2012

Summary

Metabolomics provides a holistic approach to investigate the perturbations in human metabolism with respect to a specific exposure. In nutritional metabolomics, the research question is generally related to the effect of a specific food intake on metabolic profiles commonly of plasma or urine. Application of multiple analytical strategies may provide comprehensive information to reach a valid answer to these research questions. In this thesis, I investigated several analytical technologies and data handling strategies in order to evaluate their effects on the biological answer.

In metabolomics, one of the crucial steps is data preprocessing, which is particularly cumbersome for complex liquid chromatography mass spectrometry (LC-MS) data. Accordingly, in PAPER I, different LC-MS data preprocessing tools, MarkerLynx, MZmine, XCMS and a customised method (spectral binning and chromatographic collapsing) were compared using a simple dataset of plasma samples collected from rats in the fed or fasting state. The methods were compared in terms of the total number and identity of the features discriminating the metabolic states (markers). 32 to 40 % of the markers were selected by all three tools (MarkerLynx, MZmine and XCMS) and 16 to 40 % were specific to each tool. Two reasons for these differences were pointed out: (1) changing the parameter settings of each software tool has a great impact on the number of detected features; (2) each software tool employs different methods in their peak detection and alignment algorithms, such that each has pros and cons. Thus, the use of more than one software tool and/or the use of several parameter settings during data preprocessing are likely to decrease the risk of failing to detect features (potential marker candidates) in untargeted metabolomics. On the other hand, customised methods lead to many false positives and negatives.

Data preprocessing is followed by data analysis. In metabolomics, large amount of complex data characterise few samples, thus data analysis becomes a critical step as well. Principal component analysis (PCA) is useful for exploratory purposes and partial least squares discriminant analysis (PLSDA) for classification and variable selection purposes; both have been used in PAPER I and II. In PAPER III, the application potential of sparse principal component analysis (SPCA) on LC-MS based metabolomics data as a pattern recognition and variable selection tool have been investigated. The results suggested SPCA performs well in terms of extracting time since last meal related patterns, yet it provides more easily interpreted loadings for selection of relevant metabolites.

One of the conclusions of this thesis is that the data handling strategy influences the patterns identified as important for the nutritional question under study. Therefore, in depth understanding of the study design and the specific effects of the analytical technology on the produced data is extremely important to achieve high quality data handling.

Besides data handling, this thesis also deals with biological interpretation of postprandial metabolism and trans fatty acid (TFA) intake. Two nutritional issues were objects of investigation: 1) metabolic states as a function of time since the last meal and 2) markers related to intakes of cis- and trans-fat.

Plasma samples are usually taken in the fasting state, typically following an overnight fast, as it is considered to be more reproducible and it can be defined as a baseline level for metabolic studies. On the other hand, the postprandial response reveals multiple aspects of metabolic health that would not be apparent from studying the fasting state. To investigate this issue two studied are involved, initially LC-MS plasma profiles of rats at fasting and fed states are compared (PAPER I), and later LC-MS plasma profiles of subjects from as observational study has been explored with the aim of identifying the overall response to food intake and its clearance rate in free-living humans (PAPER III).

The adverse health effects of industrial TFA is accepted, still the responsible physiological mechanisms are not fully understood. With the aim of contributing to this issue, the changes in plasma LC-MS profiles due to TFA intake (16 weeks) and its depletion (12 weeks) were examined in order to identify metabolic patterns affected by this potentially toxic fat using a parallel intervention study. In addition, the impact of cis- vs. trans-fat intake on a glucose challenge was investigated (PAPER II).

The postprandial state has been identified with a higher abundance of lyso-lipids and amino acids in plasma LC-MS profiles of rats compared to the fasting state in a controlled study design (PAPER I) and the same metabolites were characterised in humans in an observational study (PAPER III). The higher amino acid concentration after the meal is linked to the protein source present in the last meal and the declining trend thereafter is related to insulin stimulation of amino acids uptake from the plasma to liver and muscles for protein synthesis. The other typical fasting state metabolites such as fatty acids, acyl-carnitines and ketone bodies were only detected in the rat study (PAPER I). The study group in PAPER III were from a largely un-controlled observational setting with varying quality and quantity of food intake as well as varying time from last meal. This

may be the cause why fewer compounds were extracted in this study but differences between rats and humans may also influence the findings.

In PAPER II, Nuclear magnetic resonance spectroscopy (NMR) plasma profiles revealed increased LDL-C (low-density lipoprotein cholesterol) levels and increased unsaturation after TFA intake. LC-MS profiles, on the other hand, demonstrated elevated levels of a few specific polyunsaturated (PUFA) long chain phosphatidylcholines (PCs) and a sphingomyelin (SM). The preferential integration of *trans*18:1 into the sn-1 position of PCs all containing PUFA in the sn-2 position may be explained by a general up-regulation in the formation of long-chain PUFAs after TFA intake and/or by specific mobilisation of these fats into PCs as a result of TFA exposure. These findings provide a unique insight to morphological abnormalities in membrane lipids caused by TFA intake which may lead to a better understanding of its detrimental impact upon health.

List of Publications

PAPER I

Gürdeniz G, Kristensen M, Skov T, Dragsted LO (2012) The Effect of LC-MS data preprocessing methods on the selection of plasma biomarkers in fed vs. fasted rats. *Metabolites*, 2:77-99.

PAPER II

Gürdeniz G, Rago D, Bendsen NT, Savorani F, Astrup A, Dragsted LO, Effect of trans fatty acid intake on LC-MS and NMR plasma profiles. *PLoS One*. Submitted.

PAPER III

Gürdeniz G, Hansen L, Rasmussen MA, Olsen A, Christensen J, Acar E, Barri T, Tjønneland A, Dragsted LO, Patterns of time since last meal revealed by sparse PCA in an observational LC-MS based metabolomics study. *Metabolomics*. Submitted.

Supplemental Material

Acar E, **Gürdeniz G**, Rasmussen AM, Rago D, Dragsted LO, Bro R (2012) Coupled Matrix Factorization with Sparse Factors to Identify Potential Biomarkers in Metabolomics. *ICDM 2012 workshop on Biological Data Mining and its Applications in Healthcare (BioDM)*. 10th December 2012, Brussels, Belgium. Workshop Paper.

Other Publications by the Author

Gurdeniz G, Ozen B, Tokatli F (2010) Comparison of fatty acid profiles and mid-infrared spectral data for classification of olive oils. *European Journal of Lipid Science and Technology*, 112: 218–226.

Gurdeniz G and Ozen B (2009) Detection of adulteration of extra-virgin olive oil by chemometric analysis of mid-infrared spectral data. *Food Chemistry*, 116: 519-525.

Gürdeniz G, Tokatlı F, Özen B (2007) Differentiation of mixtures of monovarietal olive oils with mid-infrared spectroscopy and chemometrics. *European Journal of Lipid Science and Technology*, 109: 1194–1202.

Gurdeniz G, Ozen B, Tokatli F (2008) Classification of Turkish olive oils with respect to cultivar, geographic origin and harvest year using fatty acid profile and mid-IR spectroscopy. *European Food Research and Technology*, 227:1275–1281.

List of Abbreviations

CHD	Coronary heart disease
ESI	Electrospray ionization
GC	Gas chromatography
HDL-C	High-density lipoprotein cholesterol
HMDB	Human metabolome database
HPLC	High performance liquid chromatography
LDL-C	Low-density lipoprotein cholesterol
LPC	Lysophosphatidylcholine
m/z	Mass to charge ratio
MLR	Multiple linear regression
MS	Mass spectrometry
NMC	Number of misclassifications
NMR	Nuclear magnetic resonance spectroscopy
OGTT	Oral glucose tolerance test
OPLSDA	Orthogonal partial least squares discriminant analysis
PCA	Principal component analysis
PLS	Partial least squares analysis
PLSDA	Partial least squares discriminant analysis
QTOF	Quadruple time of flight
SPCA	Sparse principal component analysis
TC	Total cholesterol
TFA	Trans fatty acid
UPLC	Ultra performance liquid chromatography
VIP	Variables importance for the projection

List of Figures

List of Tables

Table of Contents

1 Aim of the Thesis

In this thesis the aim was to establish the metabolomics workflow starting with data handling, through identification of relevant metabolites, to interpretation of results in biological terms for three independent studies, which are presented as three papers. The aim of each paper is given in Table 1.

Table 1. The purpose of each study involved in PAPER I, II & III.

PAPERS	DATA	PURPOSE
PAPER I 	LC-MS plasma profiles of rat collected in the fasted and fed states.	• To investigate the effect of different LC-MS data preprocessing tools on the selection of metabolites representing fasting and fed states. • To identify the relevant metabolites. • To interpret the patterns in relation to fasting and fed state metabolism.
 **PAPER II**	LC-MS and NMR plasma profiles of overweight subject from double-blinded parallel intervention study where subjects received either oil containing TFA or control oil with mainly oleic and palmitic acid for 16 weeks.	• To extract the metabolic patterns associated with TFA intake from plasma LC-MS and NMR profiles. • To identify the relevant metabolites • To interpret the patterns related to TFA in terms of its adverse health effects.
PAPER III 	Plasma LC-MS profiles of subjects from a cross-sectional cohort where each subject's time since last meal is recorded.	• To evaluate the applicability of SPCA as a pattern recognition and metabolite selection tool for LC-MS based metabolomics data. • To interpret the patterns related to time since last meal in biological terms.

2 Metabolomics

Metabolomics is the field concerned with the systemic quantification of small molecule intermediates and products of metabolism present in a given biospecimen (e.g. biofluid, biological tissue etc.). By measuring and evaluating the alterations in the levels of small molecules in biological samples involved in biochemical processes, metabolomics provides a new perspective to the effects of diet, drugs and disease. The idea that biological fluids reflect the health of an individual has existed for a long time. In the Middle Ages, "urine charts" were used to link the colours, tastes and smells of urine to various medical conditions, which are metabolic in origin.

The term ‘metabolomics’ was introduced by Oliver Fiehn, in 2001, as ‘a comprehensive and quantitative analysis of all metabolites’ [1]. Formerly, the term ‘metabonomics’ is defined in 1999 by Jeremy Nicholson and colleagues as ‘the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification’ [2]. Initially, the term ‘metabolomics’ has been applied more to plant science, whereas ‘metabonomics’ has referred to studies involving animal models. Nowadays, the distinction between two terms blurred, with ‘metabolomics’ emerging as the most widely accepted term in the literature.

The modern-day metabolomics came with the advances in analytical platforms. By mid-1980s, NMR was sensitive enough to detect metabolites in biological fluids. Development of mass spectrometry in the early 20th century, as well as of different molecular separation techniques such as gas chromatography (GC) and liquid LC, afforded the detection of small molecules in biological matrices. Initial studies leading to modern metabolomics date to back to the 1970s when Arthur B. Robinson and his colleagues profiled urine vapour by GC. They related the chemical profile differences of the urine to gender and other variables [3]. The idea behind this study coincides with the aims of modern-day metabolomics.

Metabolites are the small molecule intermediates and products of metabolism. Metabolomics allows identification and quantification of low molecular weight (<1500 Da) metabolites in biological samples. The current estimate of identified metabolites in the human metabolome is over 8,500, given by the human metabolome database (HMDB) [4]. This is not a number that will remain unchanged. The improvement of analytical technologies will allow detection of lower abundance metabolites, leading to new entries. The human metabolome can be divided into compartments as described by Manach et al. [5]: 1) the endogenous metabolome covers all metabolites produced by a cell, a tissue or an organism 2) the microbial metabolome produced by

the microbiota and 3) the xenometabolome, which includes all foreign metabolites derived from drugs, pollutants and dietary compounds.

Metabolomics have been applied in different fields such as disease diagnosis, toxicology, plant science and pharmaceutical, environmental and human nutrition research. This thesis covers some metabolomics applications in relation to human nutrition.

3 Metabolomics in Human Nutrition

In human nutrition, the classical approaches are hypothesis-driven. Basically, the human biofluids are routinely analysed for a range of physiological markers, including many macro- and micronutrients as well as a limited number of known metabolites. Later, evaluating these markers, the information on nutritional status and risk factors for health and disease can be assessed. However, nutrients interact with a number of metabolic pathways which may induce alterations in many other metabolites than traditionally targeted ones. Metabolomics offer a more holistic approach with the potential of measuring hundreds of metabolites in a given biological sample with the perspective of metabolic pathway analyses. Its application in nutrition may lead to disclosure of underlying patterns in the interface between the nutrients and biological systems to understand the nutrient influence in assessment of individual's health and disease status. Furthermore, metabolomics is a promising tool for discovery of new biomarkers. In nutrition biomarkers are used as a biochemical indicator of dietary intake/nutritional status (recent or long term), or an index of nutrient metabolism, or a marker of the biological consequences of dietary intake [6]. Metabolomics may provide new biomarkers as specific metabolites or even metabolic profiles which are specific to various dietary/nutrient intake patterns or dietary changes.

Many studies have demonstrated the potential of metabolomics in the nutrition field. It has been employed to characterize the effects of specific foods such as apples [7], both a deficiency of and supplementation with different nutrients [8,9], the influence of specific diets or food components particularly on the gut microbiota [10,11], to compare the metabolic effect of closely related foods such as whole grain and refined flours [12]. Metabolomics also has been suggested as a prospective tool for differentiation of individuals' diet and their effects on specific health outcomes [13].

4 Metabolomics Pipeline

Metabolomics involves multiple steps for investigation of specific research questions. The typical workflow of a nutrition-based metabolomics study is illustrated in Figure 1. Some of these steps are particularly important for this thesis will be further discussed in detail.

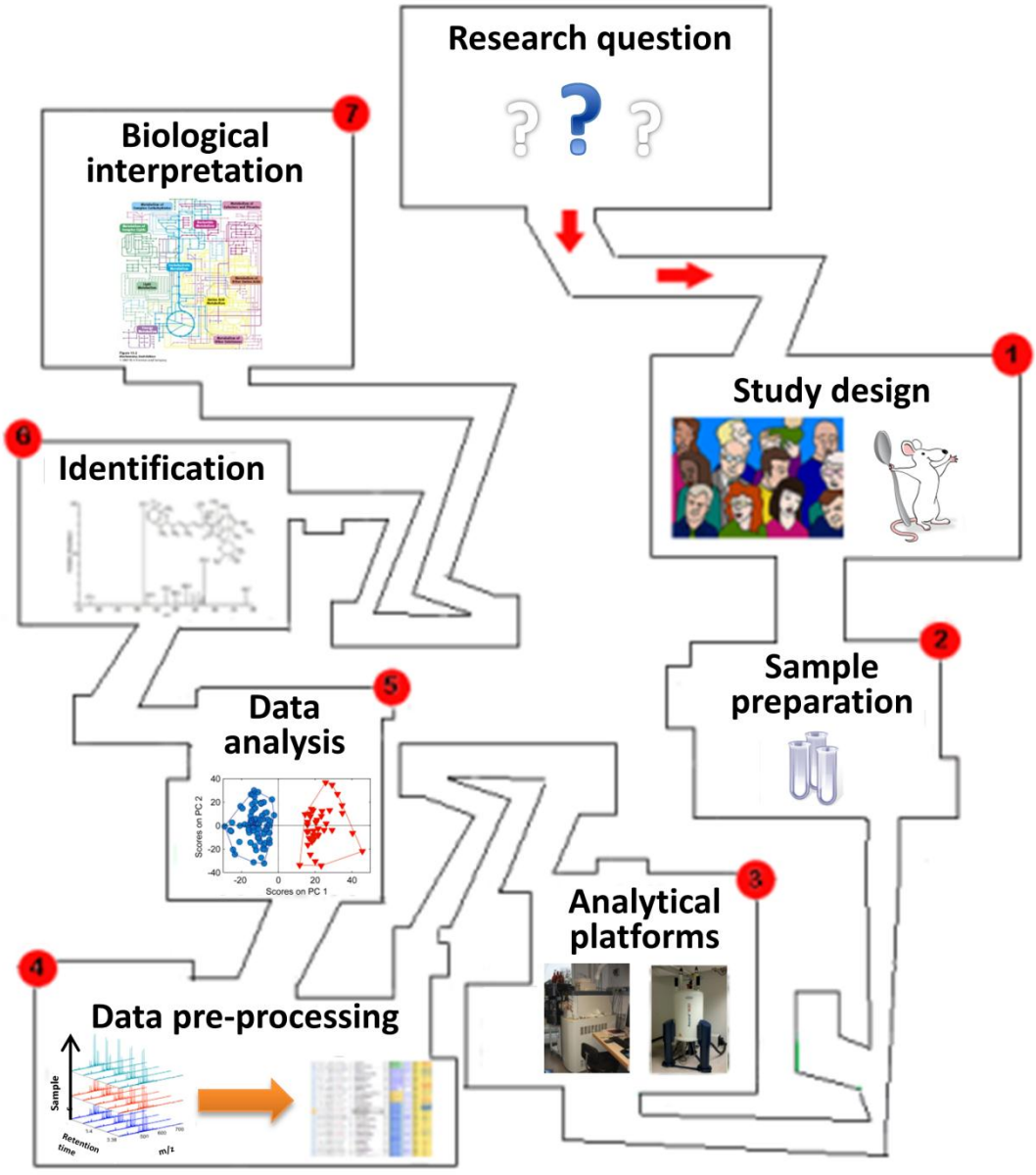


Figure 1. Metabolomics workflow pipeline.

4.1 Study Design

In nutritional studies, the selection of particular study design depends on the nature of the question, time needed and resources available. The same criteria works as well for metabolomics based nutritional studies. Most importantly, nutritional metabolomics studies deal with subtle changes, thus, it is very important to be aware of the factors contributing to the variation in the data while deciding on study design. Furthermore, the number of samples used in metabolomics is usually much smaller than the number of variables. This can be problematic, particularly in data analysis. Thus, the highest possible number of samples should be included in the study design.

The two commonly applied dietary intervention studies, *parallel* and *cross-over* designs, are both suitable in metabolomics. In *parallel* design, subjects are randomly assigned to one of at least two groups, one of these acting as control, and subjects are followed up over a time period [14]. The effect of the intervention in human studies is preferably assessed as the change between selected parameters from start to the end of the intervention period compared with the control group, whereas in animal studies samples are most often only collected at the termination of the study. However, the parallel study design which is essentially the same as the first period of a cross-over design, does not consider the possible variation between subjects response to the treatment. In a cross-over study, each subject receives all treatments, so that inter-individual variation is reduced [15]. On the other hand, cross over studies require longer time and are more sensitive to drop-outs compared to parallel studies. PAPER I in this thesis serves an example of metabolomics applications on an animal study with parallel study design. Compared to humans, animal models offer full control of the food intake. Since inter-individual variation in animal model are less pronounced than human (e.g. standardised phenotype rats in PAPER I), the patterns of interest becomes less cumbersome to extract. On the other hand, the genetic differences between rats and humans may result in some effects and physiological responses of interest to be covered. For instance, basic metabolic rate varies roughly with surface area in mammals and an overnight fasting period in rats having an eight times higher rate of energy metabolism than humans may therefore represent a more extreme condition than overnight fasting in humans (PAPER II). PAPER II represents an example to a parallel design human intervention study.

In cohort studies the dietary intake and other relevant exposures are measured in a population of people identified at baseline and they are followed to determine diet-disease associations. Cohort studies can also be investigated by metabolomics, either cross-sectionally by comparing with other data collected at baseline, or prospectively by comparing with endpoints measured at a later time point, e.g. a disease outcome. Metabolic profiles from cohort studies offer many patterns from

several exposures, yet it is more challenging due to the tremendous individual diversity and lack of dietary control at the time of sampling. In PAPER III, we have performed a metabolomics based profiling strategy to investigate cross-sectionally the time since last patterns in a pilot cohort study. Metabolomics profiles combined with genetic information, dietary and other lifestyle exposure data from large cohorts provide invaluable information. The number of studies investigating this issue is rapidly increasing and the findings are promising [16–19].

Considering the study designs within the publications included in this thesis the inter-individual variations is in increasing order, PAPER I, a well-controlled animal study; PAPER II, double-blind parallel intervention study; PAPER III, a cross-sectional cohort study.

4.2 Biological Samples

In metabolomics, the selection of biological samples depends on a number of factors: (1) accessibility, (2) relevance to biochemical question, (3) the previous knowledge of the biological system, and (4) suitability for the available analytical platform.

Biological fluids, particularly plasma, serum and urine, are relatively easy to obtain and have been used in the majority of nutrition-based metabolomics studies. They are particularly informative as they reflect the global state of an individual. Plasma/serum carries the small molecules informative in relation to the metabolic state at the time of collection, and reflecting catabolic and anabolic processes, whereas urine provides an averaged pattern of easily excreted polar metabolites discarded from the body as a result of catabolic processes [20]. Only plasma samples were used as biological samples for the studies discussed in this thesis.

Many other liquid and non-liquid biological samples such as saliva, breast milk, seminal plasma, bile, digestive fluids, cerebrospinal fluid, or tissue can provide valuable information, in the discovery of patterns for diet-disease associations [21]. Indeed each biological fluid has its own characteristic metabolic profile, recovering many different biological fluids from specific samples offering complementary information and leading to more comprehensive overview of metabolic perturbations.

4.2.1 Sample preparation

Sample preparation for further analysis depends on type of sample, the analytical method and whether a *targeted* or a *non-targeted* approach is of interest. The *targeted* approach is a method involving quantitative measurement of only a specific, pre-defined group of metabolites as the

interest is to examine one or more related metabolic pathways. It is often used to test a specific biological question or hypothesis rather than hypothesis generating. As the focus is to quantify a specific group of metabolites, sample preparation builds on extraction of those. The *non-targeted* approach has a global scope; to analyse as many metabolites as possible in a sample with at least differential quantification. So, the sample preparation should be suitable for global profiling, considering the analytical platform. Under *non-targeted* approach, also a group of metabolites preferred to be profiled in more targeted applications, such as in lipidomics studies, may be selected by special sample preparation. This thesis focuses on only non-targeted applications in metabolomics.

Plasma sample preparation for LC-MS basically involves precipitation of proteins. The large number of proteins in plasma samples interferes with MS, causing metabolite losses. The most commonly used method is protein precipitation with organic solvent. The method developed by our laboratory uses 90% methanol 0.1% formic acid solution, and details are described in PAPER I. An overview of methods for sample preprocessing prior to LC-MS analysis have been described by Vuckovic [22]. Additional internal and external standards may be added and a pooled sample may be included in each batch to assess the data quality.

Regarding samples for NMR, the initial step may be protein precipitation but it is optional. Next step is the optimization of plasma samples for NMR by buffering the sample pH to stabilize chemical shifts. For this, deuterated water is added to provide frequency lock for the spectrometer, followed by addition of a reference compound such as 3-trimethylsilylpropionic acid (TSP) chemical standard [23]. This procedure is also used for plasma sample preparation for NMR analysis in PAPER II [24].

4.3 Analytical Platforms

The rapid development of the metabolomics field is linked to the advances in analytical methodologies, making it possible to qualify/quantify the metabolites in biological samples. A wide range of analytical platforms such as infrared spectroscopy and fluorescence spectroscopy has been employed in metabolomics studies, yet NMR spectroscopy and MS have the leading role today. These two technologies outperform the others because they offer the possibility of measuring the largest number of metabolites.

The basic principle of NMR spectrometry relies on utilization of magnetic spinning properties of certain atomic nuclei to determine the physical and chemical properties of atoms or molecules.

For biological samples the most commonly used nuclei are ^{1}H and 13C. The chemical environment of the nuclei in different chemical environments absorbs energy at slightly different resonance frequencies, and this effect is referred to as the chemical shift [25]. Ultra performance liquid chromatography-quadruple time of flight/mass spectrometry (UPLC-QTOF/MS) utilizing electro-spray ionisation (ESI) and NMR spectroscopy were the platforms utilised in this research project. These two techniques will be briefly discussed.

UPLC-QTOF-MS is a hyphenated technique, utilizing physical separation capabilities of LC and the mass analysis capabilities of mass spectrometry. UPLC provides sharper peaks, high sensitivity and high resolution columns by using columns packed with smaller particles and/or higher flow rates for increased speed compared to high performance liquid chromatography (HPLC) [26]. The analytes eluting from the column are ionised (i.e. ESI). A wide range of biomolecules can be readily ionised by ESI so that LC-MS employing this technology has become commonly used in metabolomics. Some characteristic properties of NMR and UPLC-MS have been given in Table 2.

Table 2. A comparison of NMR and UPLC-MS approaches used in metabolomics [27].

	High Resolution NMR	**LC-MS**
Metabolome coverage	Less sensitive 30–40 in blood plasma 40–100 in urine	Higher sensitivity Hundreds to thousands?
Throughput	~10 min	6-25 min (UPLC vs. HPLC)
Reproducibility	High reproducibility. NMR spectrometer does not get dirty - the sample is physically isolated from the instrument.	Lower reproducibility
Identification	Good libraries of spectra	Metabolite identification is a major challenge.
Destructiveness	Nondestructive. Sample can be reanalysed.	Sample destructive
Disadvantages	Crowded spectra - discrimination of resonances from the various compounds in complex mixtures can be difficult	ion suppression where a high-abundance analyte reduces, or eliminates, the response for a weaker analyte

Nevertheless none of the single technologies can detect the entire range of metabolites within a biological sample. Therefore, integration of MS, NMR, and other modern analytical techniques provides much broader information of analysed samples and thus, leads to a better understanding of biological interest. Utilizing combined targeted and non-targeted NMR, GC-MS and LC-MS methods, Psychogios et al. [4] performed comprehensive identification and quantification of the human serum metabolome. Recently, many more studies have emerged in metabolomics studies employing more than one analytical platform to investigate varying scientific questions of interest [28–30]. In PAPER II, both plasma NMR and LC-MS profiles has been utilised to uncover the effect of TFA intake. The complimentary findings from these two analytical platforms strengthen the identification of trans-fat related patterns.

4.4 Data Preprocessing

4.4.1 LC-MS

LC-MS based metabolomics experiments usually produce large amounts of complex data. Due to its complexity, it is not suitable for application of any data analysis tool in its raw form for extraction of relevant information. This brings out the data preprocessing concept which aims to extract easy-to-access characteristics of each detected ion. These characteristics include m/z, retention time of the compound, and intensity measurement. This representation of an ion is denoted as a 'feature' (Figure 2, preprocessed data).

Data preprocessing is crucial for the quality of the identification and quantification relevant information, and therefore the resulting biological interpretation. Typical data preprocessing includes multiple steps as shown in Figure 2. Many software packages are available for preprocessing LC-MS based metabolomics data both commercial (MarkerLynx) and freely available such as XCMS and MZmine [31–35]. Some performs only specific steps in the preprocessing pipeline, whereas others cover many steps. A nice overview of the algorithms and tools for preprocessing of LC-MS metabolomics data has been given by Castillo et al. [36]. In the next sections each step of the preprocessing pipeline will be discussed briefly. In this thesis, the aim was to have a general understanding of each preprocessing software to achieve correctly preprocessed data. There was no intention to dig into details of the algorithms. As PAPER II focuses on MZmine, XCMS and MarkerLynx software, those will be referred in the discussions.

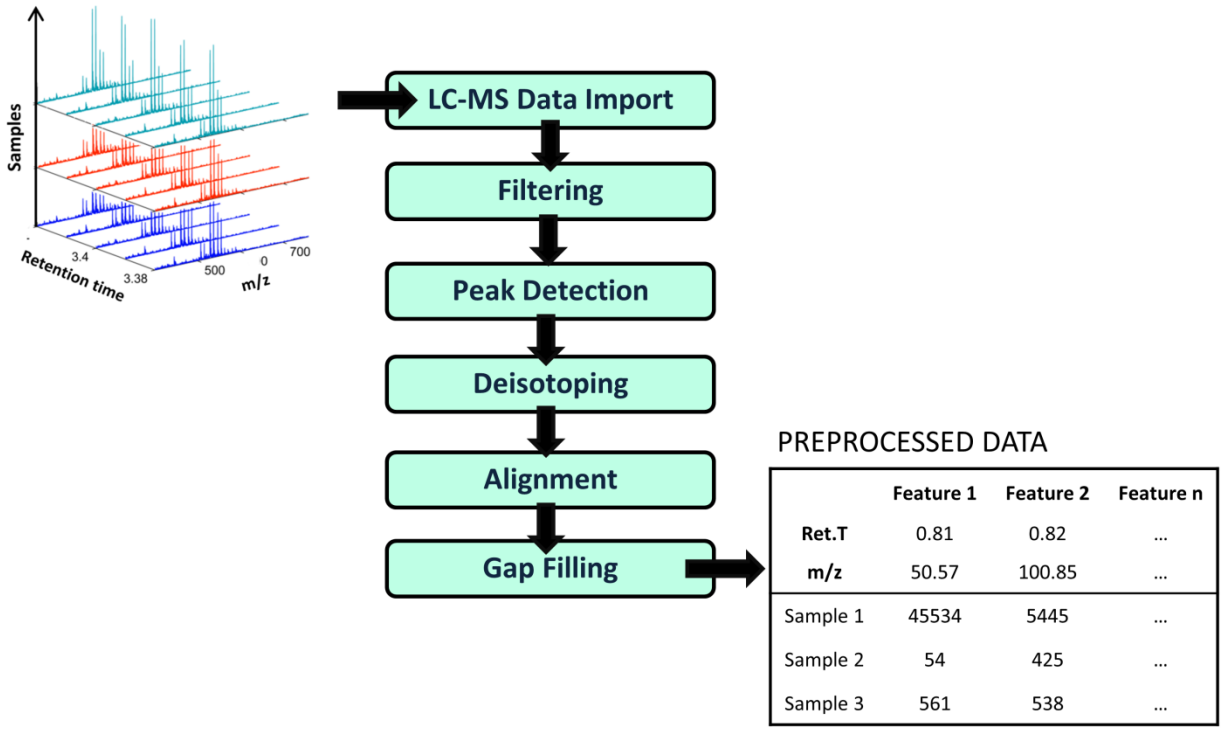


Figure 2. Schematic overview of the metabolomics pipeline in LC–MS data preprocessing from raw data to preprocessed data.

4.4.1.1 Raw data

The raw data can be acquired in continuum mode or centroided mode. In continuum mode each spectrum is represented by a distribution of m/z (mass to charge ratio) values, each representing the hits of ionized molecules on the detector. Due to large size and complexity of continuum data, centroiding is frequently applied either during data acquisition or preprocessing. Centroiding aims to convert multiple data points representing the same peak in the distribution into a single data point with a single m/z and intensity value. After centroiding the data size per sample is reduced approximately 7 fold.

The raw data file of each sample consists of a set of mass spectra, each recorded at a given time point or scan. Each scan point in time is represented by a pair of m/z and intensity vector. Handling LC–MS data in its raw form is difficult because the vector length varies from scan to scan based on the number of detected peaks. Furthermore, the four decimal digit m/z values of an ion deviate in subsequent scans even if the same compound(s) is represented.

The raw data proprietary format differs between instrument vendors but there are different tools available for the conversion of the raw data to an open format such as mzXML, NetCDF or mzML [37,38].

4.4.1.2 Filtering

Filtering aims to remove the noise so that in the subsequent peak detection step less false positives are detected. Most methods combines filtering with peak detection steps on extracted ion chromatograms with signal processing techniques such as Gaussian filtering (XCMS) and others applies a user defined cut off value on each mass spectrum (MZmine, XCMS, MarkerLynx).

4.4.1.3 Peak detection

Peak detection is one of the most crucial steps in LC-MS data preprocessing. It aims to characterise each ion in a sample with one m/z, retention time and intensity value (a feature). A peak detection algorithm should detect true signals while avoiding the noise. It should be flexible to detect peaks in varying shapes.

Generally, the initial step of peak detection is combining the ions representing the same compound in subsequent scans. This can be performed by binning such that the m/z axis is divided into equal size intervals. As a result the data for each sample is transformed into a two-dimensional matrix. However, defining a fixed bin window has some drawbacks. If the bin size is too small, the ion representing the same compound may split into adjacent bins, thus chromatographic peak shape is lost; or if the bin size is too large more than one compound and/or noise can be included in the same bin so that the chromatographic peak shape may be distorted [39]. Some examples of these issues have been shown in PAPER I.

In general, software tools utilise parameters such as minimum peak width, mass accuracy (or m/z window), minimum peak height or signal to noise ratio while combining m/z values in subsequent scans.

MZmine initially creates continuous chromatograms within a user defined minimum time range for each m/z value (within a user defined window), and then deconvolutes peaks based on local minimums (or by another of three deconvolution methods available). The XCMS centWave [40] algorithm does not require a fixed bin size, instead directly provides the potential region of interest (Figure 3) based on mass accuracy and minimum peak width. Later, it constraints the peaks' shapes with a Gaussian filter or continuous wavelet transform. This method has also been

recently implemented in MZmine. As the last step, the peak can be either integrated or peak height can be used as its magnitude.

MarkerLynx initially determines the regions of interest in the m/z domain based on mass accuracy (mass tolerance). The ApexPeakTrack algorithm controls peak detection by peak width (peak width at 5% height) and baseline threshold (peak to peak baseline ratio) parameters which can be either set by the user or calculated automatically. The algorithm finds the inflection points (peak width at 5% height), local minima and peak apex to decide peak area and height. It also calculates the baseline noise level using the slope of inflection points. Compared to peak detection algorithms of other software, the ApexPeakTrack algorithm produces much higher number of peaks, so an additional peak removal step (denoted by user defined peak intensity threshold and noise elimination level parameters) is implemented into the alignment algorithm by its developers.

Basically a few variables should be defined for the peak detection step of any preprocessing tool. Selection of reasonable parameters is vital for the detection of peaks representing the true signals.

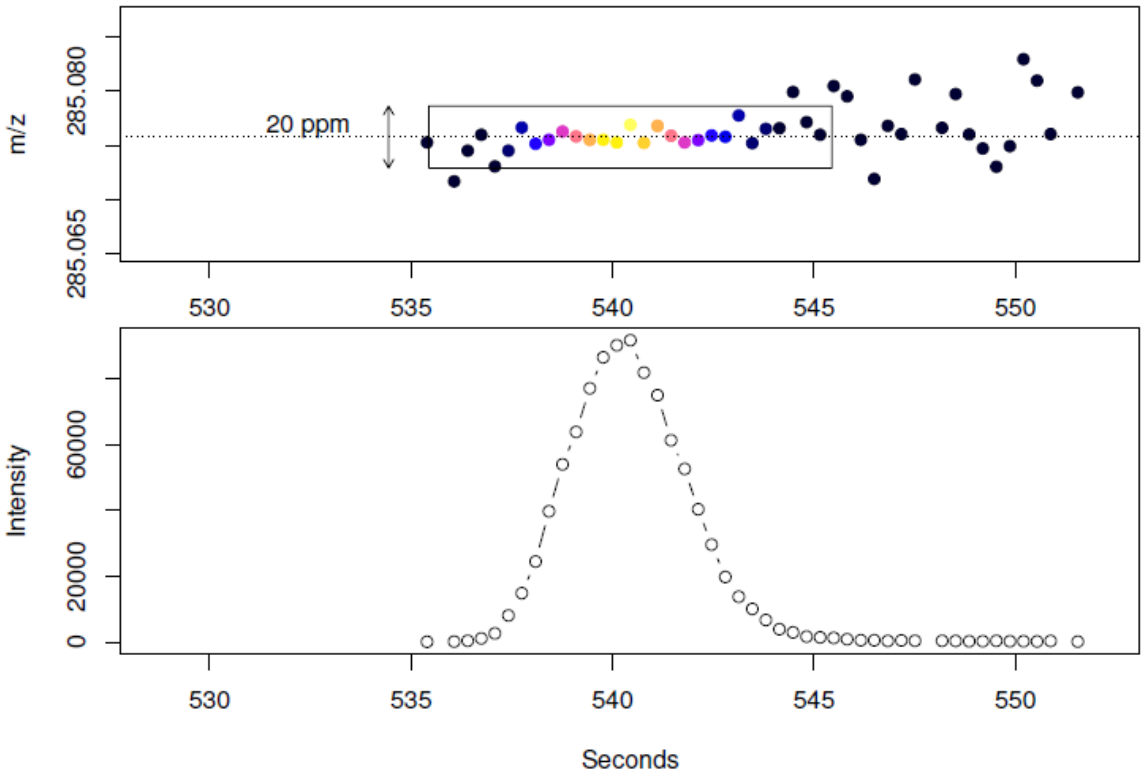


Figure 3. Detection of a peak with a specific mass accuracy (top) and chromatographic width (bottom) [40].

4.4.1.4 Deisotoping

Most elements are naturally present in different isotopic forms. A particular metabolite may produce an isotopic pattern and the relative heights and m/z difference between isotopic peaks may provide valuable information for identification of elemental composition of the metabolite. On the other hand inclusion of all isotopic peaks during data analysis increases the size of the data with redundant information. Therefore, deisotoping can be preferably performed prior to alignment. In general, for each charge state, peaks within the m/z and retention time limits are considered as isotopes, and the most abundant isotope is kept. MZmine and MarkerLynx have optional deisotoping step, yet they store the isotopic patterns for identification purposes, whereas XCMS allows annotation of isotopes but keep them in the preprocessed data.

4.4.1.5 Alignment

Peak detection is performed sample-wise. Thus, a feature representing the same compound may have different m/z and retention time values in different samples due to small random shifts. Alignment aims to match features across the samples so that the whole data can be transferred into a two dimensional matrix for subsequent data analysis. The m/z shifts are easier to deal with as the accuracy of the MS is known. XCMS utilise this fact and initially groups the features only using their m/z values within a fixed bin of 0.25 m/z. If the retention time shifts are linear and relatively small, a peak matching algorithm with defined m/z and retention time window can be sufficient (MZmine). For instance, the retention time of the peaks of the samples analysed with UPLC (PAPER I, II and III) have only slight drifts between samples, even across two years of analysis. However, retention time shifts can be more problematic for HPLC. It is frequently caused by multiple factors such as pressure, temperature and flow rate fluctuations. Thus, an alignment algorithm that can deal with non-linear shifts is usually required to correct retention time differences between samples. Different methodologies have been proposed for alignment of non-linear shifts. XCMS calculates the overall retention time distribution of peaks in each m/z bin in order to estimate the boundaries of regions where many peaks have similar retention times (Figure 4). In addition, XCMS provides an optional algorithm to deal with non-linear shifts, using a group of ‘well-behaved’ peaks as temporary standards to calculate the retention time for each sample and correct it [31]. For non-linear shifts, MZmine creates a model of the retention time shift for each peak list (i.e. sample) with respect to a master peak list. Using that model it estimates the corrected retention time for each sample [41].

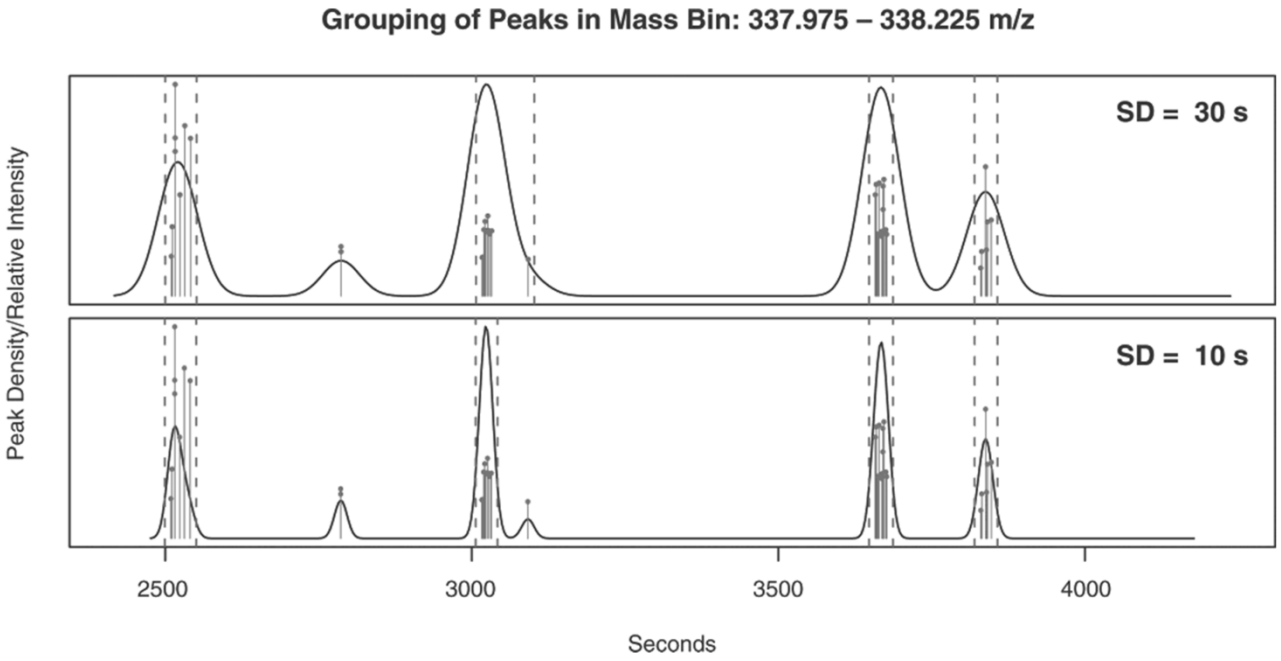


Figure 4. Example of retention time matching across samples (12 samples) within one m/z bin. Individual peaks are shown as sticks indicating relative intensity. The peak density profiles were smoothed with Gaussian functions of SD 30 and 10 s, respectively (solid line). Identified groups are flanked by dashed lines. Note how decreased smoothing eliminates a peak from the second group [31].

4.4.1.6 Gap filling

In the final feature table, some features will be missing in some samples. The missing features occur in peak detection due to low intensity peaks, bad peak shapes, and peak detection mistakes. The missing values are usually zeros where 'true' zeros as well as smaller and larger peaks missed by the algorithm are given the same zero value. This may cause misinterpretations in the data analysis part. Some examples illustrating this situation have been shown in PAPER II. Both XCMS and MZmine fill the gaps from the raw data. MarkerLynx, on the other hand, does not have a gap filling algorithm; approximately 50% of the values in the feature table are filled with zeros.

4.4.1.7 Traditional methods for data preprocessing (customised methods)

Initially, the m/z axis is binned, for instance the bin size used in PAPER I was 0.1 min. As a result data for each sample is transformed into a two-dimensional matrix. Later, a rough filtering is applied to eliminate very low signals. Then, the matrix is summed across all retention times, which eliminates the necessity of retention time alignment. The final data matrix includes samples in its rows and m/z bins in its columns filled with summed intensities [42]. There are some drawbacks of this method:

(1) Selection of the bin size is critical. If the bin size is too small, the ion representing the same compound may split into adjacent bins, or else - if the bin size is too large - more than one compound can be included in the same bin [39]. In both cases the final identification of relevant information can be effected.
(2) After chromatographic collapsing the compounds with the same m/z will be concatenated together (e.g. isomers). Furthermore, if the baseline level within one bin varies sample to sample, it may bury the relevant information.
(3) After extraction of significant bins with data analysis, for identification of interesting features, the accurate m/z and retention time is necessary. Thus, the bins should be resolved in order to determine the peak retention time.

In PAPER I, it is shown that this method leads to identification of many false positives and false negatives. On the positive side it should be mentioned that this approach is all-purpose and allows a fast data preprocessing.

4.4.1.8 Comparison of Software tools

Considering the large number of peaks with varying peak shapes, so far there is no common method to evaluate the preprocessing algorithms from different software. Even with the same feature detection algorithm, using different parameter settings usually lead to different results. In PAPER I, the number of features detected by MZmine, XCMS and MarkerLynx are compared. As shown in Figure 5, 22 to 42% of the features detected by one of the software, also detected by the other two, whereas 14 to 42% of the features were software specific. This was not surprising since not only the peak detection methods of preprocessing tool differ but also their parameter settings. Tautenhahn et al. [40] found a higher number of common features (80 %) from leaf and seed extracts comparing MZmine and XCMS (centWave) peak detection algorithms. The difference can be a result of the more complex nature of plasma samples compared to plant extracts or the chromatographic method employed.

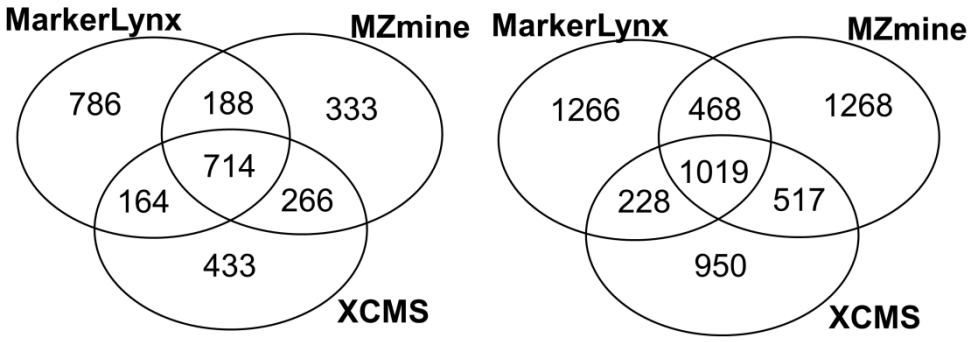


Figure 5. Venn diagrams illustrating the number of common and method specific features extracted from three software tools (right: positive mode; left: negative mode). (Data from PAPER I)

Based on the complex nature of the analyzed samples, the number of features is not known, more features may mean inclusion of false positives (e.g. noise) or true positives. Therefore, the absolute number of detected features is not really suitable to characterize data preprocessing software. Accordingly, in PAPER I from three different software tools the number of features selected as representative of relevant phenomena (markers) have been compared. 32 to 40 % of markers are in common whereas 16 to 40 % are specific to each tool. The potential sources of low number of overlapping markers can be listed as follows:

- Software specific detected features. Major cause is peak detection algorithms and its parameter settings.
- Differences in peak height assignments or errors that may occur during gap filling.

The loss of information and potential introduction of noise during feature selection by a single preprocessing method would therefore seem to be a potential source of error in metabolomics. Thus, the use of more than one software and/or the use of several settings during data preprocessing with any software is likely to improve marker detection in untargeted metabolomics.

The selection of a specific preprocessing software tool depends on programming skills, and easy visualization of the results to allow optimal parameter settings, quality control and coverage of steps in the pipeline. Furthermore, its ability to make use of available memory and CPU of the PC is another important factor which is particularly important when larger number of samples is to be preprocessed. Some of the practical properties of XCMS, MZmine and MarkerLynx have been given in Table 1.

Table 3. Practical properties of MZmine, XCMS and MarkerLynx

	MZmine	XCMS	MarkerLynx
Availability	Free	Free	Commercial
User interface	• GUI* • No requirement of programming skills	• R software command line • Some programming skills is required	• GUI* • No requirement of programming skills
Memory usage	• Adjustable to maximum available memory in the PC. • Less efficient than XCMS e.g. 16 GB RAM = maximum ~2000 samples	• Adjustable to maximum available memory in the PC • e.g. 16 GB RAM = maximum ~5000 samples	• Fixed • e.g. maximum ~1000 samples
CPU usage	• Adjustable to maximum available CPU in the PC	• Adjustable to maximum available CPU in the PC	• Fixed
Identification	• Basic identification tools. • Automated advanced tool CAMERA is incorporated from XCMS	• Automated advanced identification tool CAMERA	• Basic identification tools
Coverage of preprocessing pipeline	• All steps	• Final feature table includes isotopic peaks	• All steps
Visualization of the results	Yes	Yes	No

* Graphical User Interface

Recently, a web based Graphical User Interface version of XCMS has been released [43]. It allows the application of R package based XCMS tools. It requires the data to be uploaded which can be time demanding with large sample sizes.

4.4.2 NMR

NMR signals are collected as a function of time. The chemical shift can be derived from free induction decay, which is the decaying signal that follows a pulse, by utilizing a Fourier transformation [25]. The first step of preprocessing is phase correction and baseline correction which can be performed by the tools provided by machine-vendor software.

Due to pH, overall dilution of samples and relative concentrations of specific metabolites, the chemical shift of the same analyte peak usually varies across the samples. In order to correct for these variations a simple and common approach, spectral binning, has been widely used. Note that the same approach has been used during LC-MS data preprocessing to correct m/z variation as mentioned in the previous section. On the NMR side, the main disadvantage of binning is loss of spectral resolution. To avoid this problem more sophisticated alignment tools have been proposed, utilizing varying procedures such as genetic algorithms [44], partial linear fit [45] and correlations [46,47]. Correlation based alignment methods recursive segment-wise peak alignment [46] and interval correlated shifting (icoshift) [47] use the efficient fast Fourier transform engine to handle the large data sets. Icoshift splits the spectra into intervals and shifts the spectra to get the maximum correlation toward a target (reference or an average spectrum) in that interval. The recursive segment-wise peak alignment also relies on maximizing the correlation between target and spectra on interval-wise basis, but it relies on peak-picking.

Icoshift has been employed for spectral alignment in PAPER II. As an example, NMR spectra before and after alignment with icoshift has been shown in Figure 6.

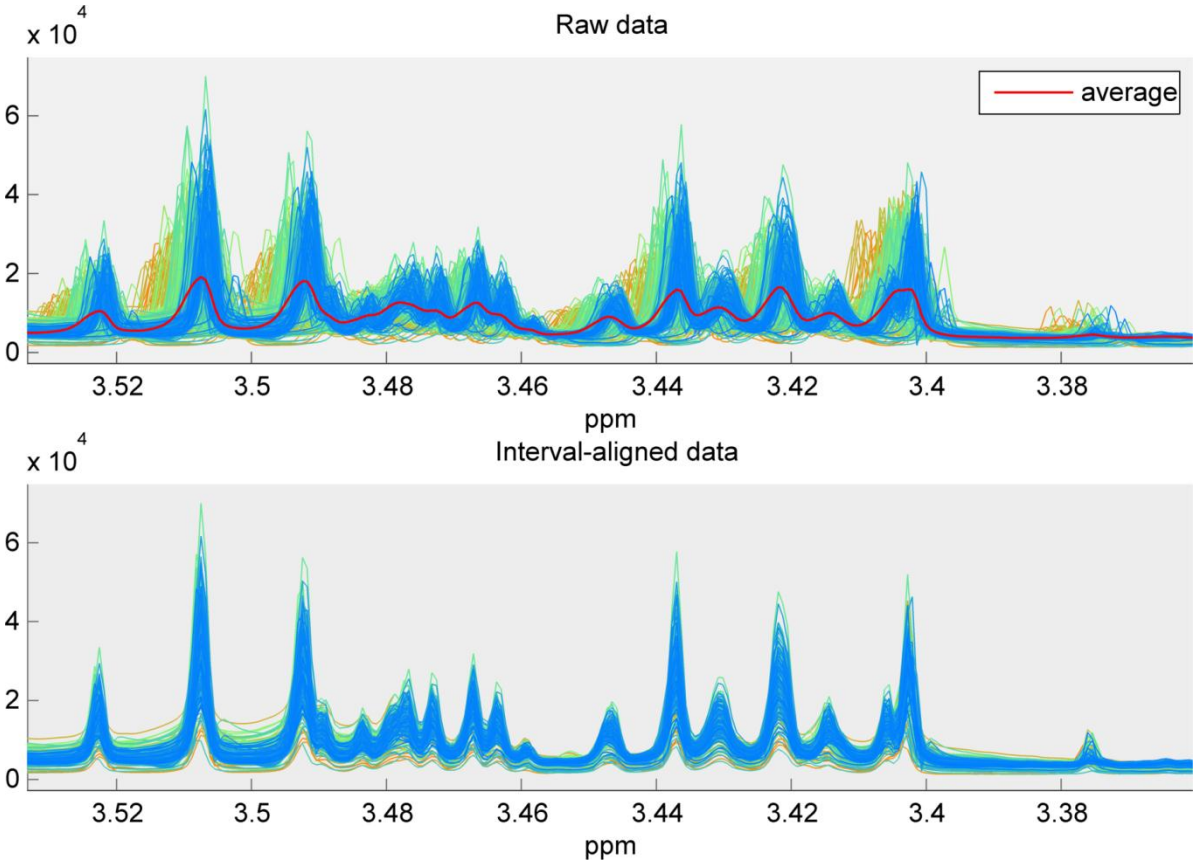


Figure 6. 600 MHz 1H NMR spectra of human plasma pre- and post-alignment using icoshift function. (Data from PAPER II)

Icoshift requires user defined interval boundaries requiring careful examination of the data whereas for recursive segment-wise peak alignment the choice of interval boundaries is often only decided by the total number of segments desired, potentially resulting in a boundary dividing a resonance leading to significant peak distortion. In a recent study, a method for automated determination of intervals has been proposed. This protocol aims to generate spectral intervals sharing a common target spectrum. Its potential application for icoshift and recursive segment-wise peak alignment has been demonstrated [48].

4.5 Data Analysis

In general, the central aim of data analysis in metabolomics is to extract the metabolites that specify the difference between sample groups. Metabolomics data can be analysed with a variety of chemometric and statistical tools [49,50]. Indeed, PCA and PLS are the most widely used ones and those were employed to analyse the data in the PAPERS I, II and III.

4.5.1 Data pre-treatment

Prior to PCA and PLS, normalization can be applied, if necessary, to correct for unwanted variation between samples. Later, the data is commonly transformed to a more suitable format for PCA and PLS based methods by scaling and centering procedures.

4.5.2 LC-MS - normalization

Systematic error arising from sample preparation and/or instrumental issues can bring out unwanted variation between samples that may hinder the extraction of relevant biological variation. Sample preparation related issues can be caused by inhomogeneity of the samples, concentration differences (e.g. common problem, particularly for urine samples), different recoveries during sample extraction and other inevitable minor differences in sample preparation (e.g. pipetting errors). Ion suppression/enhancement or ions source variations constitute the major part of the instrumental issues. This is mainly caused by matrix effect which is defined as alterations of ionization efficiency of analytes by the presence of coeluting substances [51]. Matrix effects vary between samples based on differences in their biological constituents or differences during sample preparation.

The unwanted variation may appear in two forms:

(1) Overall concentration variations between samples, i.e. the signal increase in all analytes of one sample, compared to another sample.
(2) Analyte specific fluctuations between samples, i.e. a signal increase for one analyte while decreasing for another analyte, compared to the same analytes in another sample.

If the first situation is the case, scaling factor based normalization methods can be used for correction of between sample variations. Scaling factor based normalization is performed by dividing each analyte in a sample by a factor such as unit norm, total area, or total sum of intensities calculated for that sample. For instance while acquiring the data from PAPER I, a potential signal suppression effect from a build-up of non-volatile contaminants in the ionization

source throughout the course of the entire analysis caused a steady decrease in the signal. Normalization to unit length lead to partially removal of the pattern related to signal loss, so that the fasting vs. fed pattern became easier to capture by PCA (Figure 7).

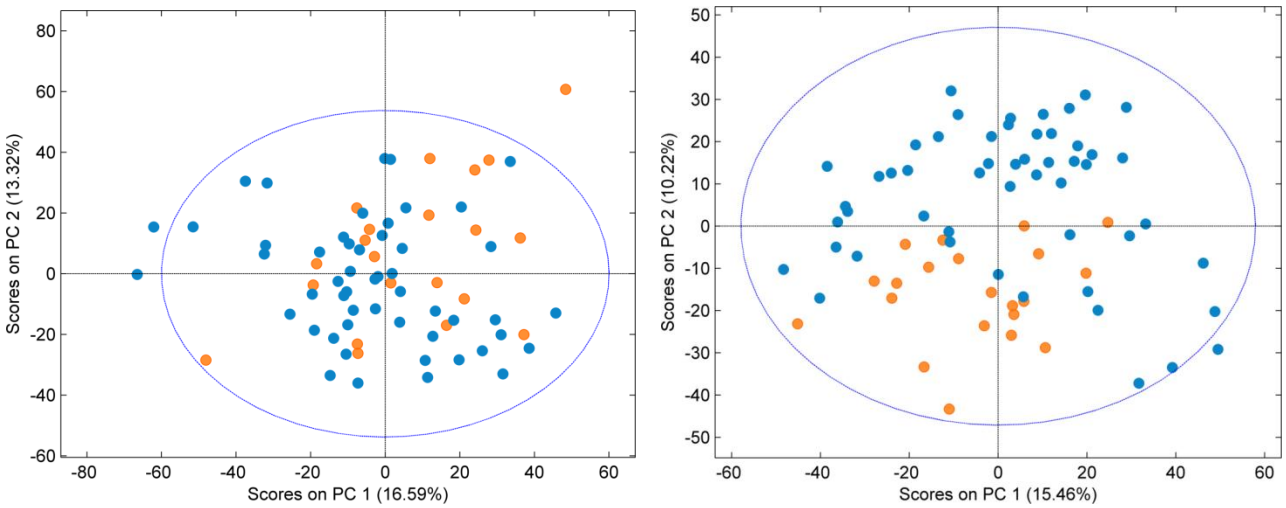


Figure 7. PC1 vs. PC2 scores plots illustrating the effect normalisation on the extraction of fasting vs. fed related patterns. Before normalization (left), after normalization to unit norm (right). Blue: fasting state, orange: fed state. (Data from PAPER I, negative mode data)

However, analyte specific fluctuations are more frequently observed. This is mainly caused by analyte specific ion suppression issues due to the complex nature of the biological samples and cannot be corrected by scaling factor based normalization methods. An example illustrating this issue is shown in Figure 8.

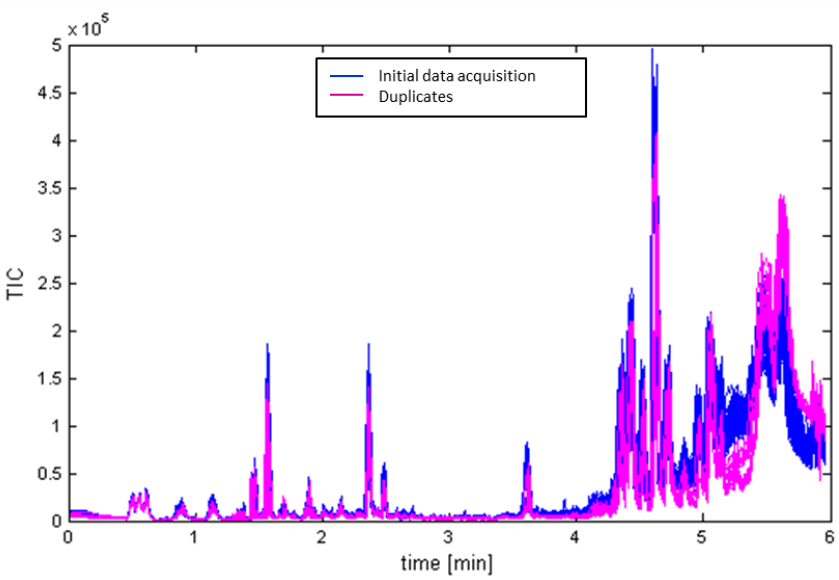


Figure 8. Total ion chromatogram of duplicative measurements shown in different colours. The initially acquired samples had higher signal between 4 to 5.5 min, but lower 5.5 to 6 min compared to their duplicates. Thus, scaling factor (total signal basis) is not representative of the between sample variation and does not correct for it. (Data from PAPER I, positive mode data)

In some situations, applying scaling factor normalization may also introduce further obscuring variation. For instance, for the data from PAPER II, the unit length normalization caused an increase in coefficient of variation (CV) for three out of four internals standards (Figure 9). Thus, careful examination of the dataset is required when using scaling factor based normalization.

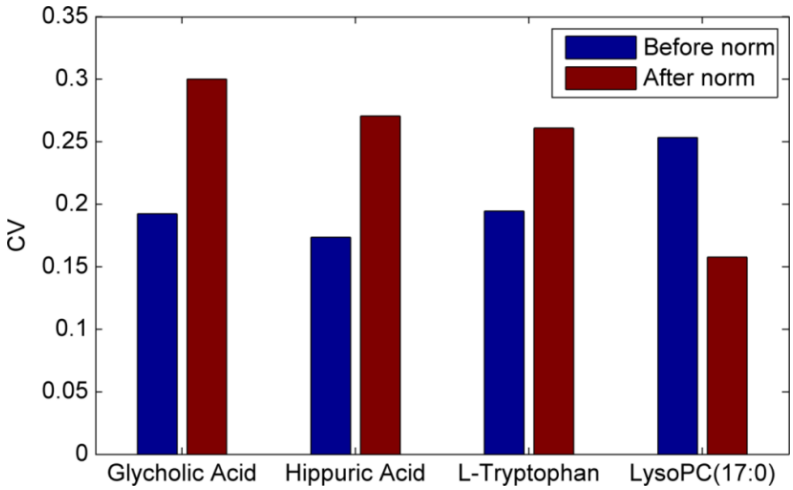


Figure 9. Effect of unit length normalization on CV of internal standards, glycholic acid, hippuric acid, L-tyroptaphan, lysophosphatidylcholine (LPC17:0). (Data from PAPER II).

In order deal with metabolite level fluctuations between samples, utilization of isotopically labelled internal standards has been suggested where each internal standard is added to each sample in identical concentration. As the internal standard represents a known quantity, the estimated analyte signal can be expressed as relative to the internal standard with the aim of

removing the error. However, a fully labelled reference metabolome is not feasible. Instead, multiple internal standards, each representing metabolites from chemically related groups, can be used for correction of systematic errors on metabolite level. Although there is no specific method to employ those to correct for each metabolite, some studies have explored the issue. Bijlsma et al. [52] performed internal standards based normalization by using each standard to correct the metabolites in its corresponding retention time region. Sysi-Aho et al. [53] utilized multiple linear regression (MLR) to remove the correlated variance between the metabolites and the internal standards. A modified version has been suggested by Redestig et al. [54]. In their method it is assumed that the variation between samples on analyte level can be represented by internal standards. Initially, structured variance in internal standards (Z) is estimated by PCA. Later, PC scores (T_Z) were used to remove the Z from the samples (Y_A) by MLR. The final equation evaluated for normalization can be expressed as

$$Y_{A,Norm} = Y_A - T_Z \, (T_Z^T T_Z)^{-1} T_Z^T Y_A$$

where they used PCA and to extract the variation in internal standards and MLR to remove the correlated variance in samples. It has been argued that this method is efficient to remove between batch and within batch differences. The batch-to-batch variation is quite problematic particularly when dataset includes many batches. As batch-to-batch variation is not caused by overall response differences between batches, scaling factor based normalization does not correct for it. The previously mentioned method by Redestig et al. [54] has been applied to the data from PAPER II, where four internal standards were previously added (given internal standards in Figure 9). The data included 12 batches where PCA explained batch differences in the first eight components. The batch-to-batch variation was partially removed, yet the results indicated that proper correction requires a larger number of internal standards representing chemically related metabolite groups. In this case, after normalization, batch-to-batch related variation decreased captured by the first four components.

In summary, two different strategies are suggested for correction of unwanted variation between samples: (1) scaling factor and (2) internal standard based normalizations. Scaling factor based normalization assumes the unwanted variation is caused by overall concentration changes between samples. Thus, it uses the same scaling factor to correct for each analyte in a sample (row-wise correction). On the other hand, internal standard based normalization performs the correction on metabolite level which means metabolites are corrected with their representative internal standards in a sample. As metabolite level fluctuations are very common for LC-MS based

metabolomics data, internal standard based normalization may be a more reasonable choice but a large number of internal standards providing overall coverage of metabolite groups is necessary.

4.5.2.1 LC-MS - Scaling and centering

Particularly in LC-MS based metabolomics, the metabolite levels differ in a wide range, yet this may not correspond to the biological interest. For instance, two metabolites with signals of 5000 and 50 are usually of equal importance. However, PCA tends to gravitate upon the larger variation that is provided by larger peaks. Thus, scaling is necessary prior to PCA or PLS, to put metabolites on similar or equal basis.

Centering adjusts for differences in the offset between high and low abundant metabolites. Mean centering forces the corrected (centered) metabolite concentrations to fluctuate around zero as the mean (Figure 10B). In most cases, centering is applied in combination to scaling.

Autoscaling and pareto scaling are the most commonly employed scaling strategies in metabolomics. Autoscaling, which is a combination of unit scaling and mean centering, uses standard deviation as the scaling factor (Figure 10C). After unit scaling, all metabolites have standard deviation of one so that they have equal chance to influence the model. The main disadvantage of autoscaling is that it also inflates the noise, thus it may complicate the extraction of relevant patterns. Pareto scaling utilises square root of standard deviation as scaling factor. As a result, it reduces large scale differences between metabolites but still they are close to the original measurements (Figure 10D).

Although some studies in LC-MS based metabolomics pareto scales the data, in most cases autoscaling has been shown as a better choice unless there is a specific interest or situation (e.g. very noisy data) [55]. The reason is that the magnitude of metabolite concentration differences are not representative of biological relevance and that can only be provided by autoscaling. Accordingly, the datasets investigated in PAPER I, II and III were autoscaled.

Note that after preprocessing of LC-MS data, elution profile of each analyte is converted to a discrete value such that each chromatographic peak is represented by its height or area. In cases where elution profiles are used (e.g. LC-FID), autoscaling may inflate the baseline and is not recommended.

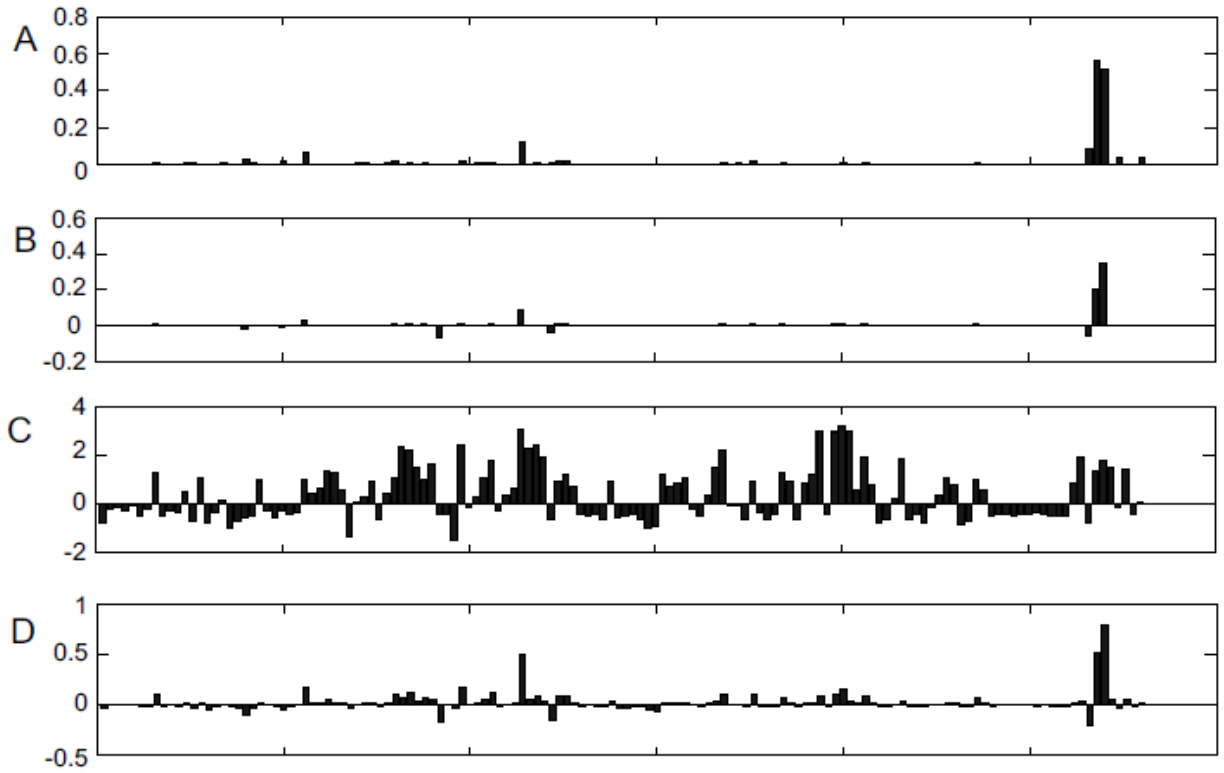


Figure 10. Effect of mean centering (B), autoscaling (C), pareto scaling (D) on deconvoluted GC-MS based metabolomics data (A) [55].

4.5.3 NMR - Normalization

Normalization of NMR spectra is especially important for urine samples to correct for variations of the overall concentrations of samples caused by different dilutions**.** The scaling factors mentioned in the previous section can also be applied to NMR data. Some other more advanced procedures includes Probabilistic Quotient Normalization [56] and Quantile Normalization [57].

Probabilistic Quotient Normalization was utilised for NMR data in PAPER II. Scaling factor based normalizations calculate the scaling factor for each sample based on contributions from all signals in that sample. On the other hand, Probabilistic Quotient Normalization calculates a most probable quotient between the signals of the corresponding spectrum and of a reference spectrum (mean or median of spectrum in the study) and uses that as scaling factor. Probabilistic Quotient Normalization can be applied to raw spectra or binned spectra. Its algorithm has been summarized as:

(1) Perform scaling factor normalization (described in section 2.5.1.1).
(2) Choose/calculate the reference spectrum (median or mean spectrum).
(3) Calculate the quotients of all variables of interest of the test spectrum with those of the reference spectrum.

(4) Calculate the median of these quotients.

(5) Divide all variables of the test spectrum by this median.

It has been shown that compared other scaling factor normalizations, Probabilistic Quotient Normalization is more robust against strong metabolite specific changes as it does not have constraints such as a total integral or a total vector length [56].

4.5.3.1 NMR - Scaling and centering

In case NMR spectral profiles are used, autoscaling of NMR spectra leads to inflated noise as shown in Figure 11, PCA loadings plot. Thus, it may become difficult to extract the relevant biological phenomena. On the other hand, as peak shapes are already distorted and the data is reduced, binned spectral profiles can be autoscaled.

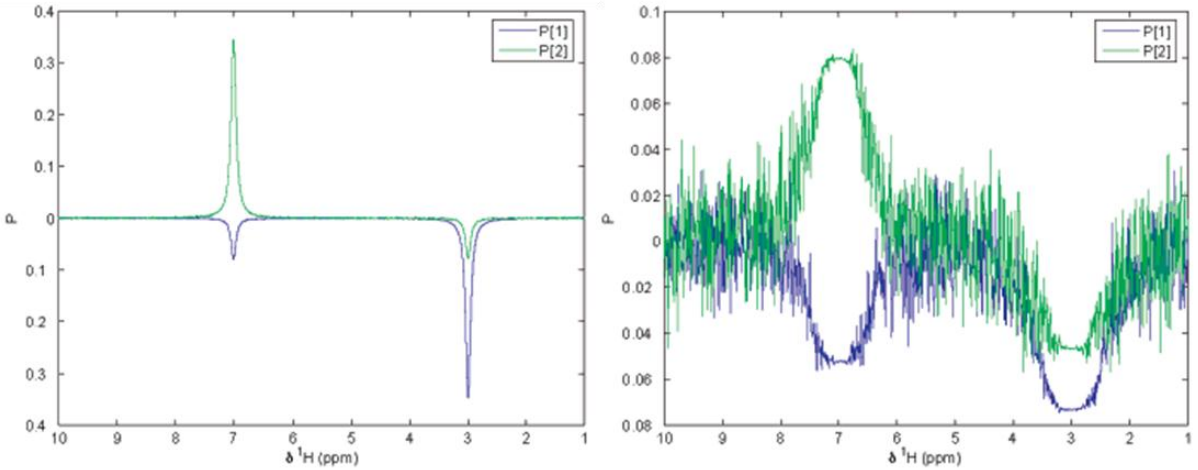


Figure 11. PC1 and PC2 loadings vs. chemical shifts for centered simulated spectral data (left). PC1 and PC2 loadings vs. chemical shifts loadings for the autoscaled true data (right) [58].

4.5.4 Principal Component Analysis

PCA was first defined in statistics as finding 'lines and planes of the closest fit to systems of points in space' by Pearson in 1901 [59] and further developed by Hotelling to its present stage [60]. Since then, PCA has been employed in a wide range of scientific fields as a well-established multivariate data analysis method.

PCA aims to extract the dominant patterns in a data matrix consisting of a large number of interrelated variables in terms of lower dimensional variables called principal components. Principal components represent linear combinations of original variables. The components are approximated as orthogonal directions in original variable space with the aim of capturing

maximum variance. Considering a data matrix X with n rows and k columns, PCA decomposes X into linear sum vector products, $t \cdot p'$. In general we need more than one component to explain the data matrix. For i components, PCA can be formulated as

$$X = t_1 \cdot p_1' + t_2 \cdot p_2' + \ldots + t_i \cdot p_i' + E \qquad i = 1,2, \ldots, i$$

where t_i is the score vector (n X 1), p_i is the loading vector with (k X 1) for each component and E contains the residuals, the part of the data that is not explained by principal components. The loading vector p defines the new directions in original variable space and the projection of samples onto that provides the score vector t. A more compact representation of PCA is given as

$$min(X - T \cdot P')$$

where T is the score matrix (n X i) and P is the loadings matrix (k X i)

The sample patterns are commonly visualized by a scatter plot of scores, for instance t_1 *vs.* t_2 for the first two components. The corresponding variable patterns are represented by a loadings plot p_1 vs. p_2. Principal components are orthogonal to each other which means they are uncorrelated so that we can talk about one component independently from the others.

4.5.4.1 Application of PCA in metabolomics

Wold et al. [61] listed the goals of PCA on a data matrix as simplification, data reduction, modelling, outlier detection, classification, prediction, classification and unmixing. On the basis of metabolomics, PCA has been used for data reduction [62], outlier detection [7], classification [63] and variable selection [64,65].

PCA provides an overview of the data and gives an idea about the dominating patterns. This is usually done by a visual inspection of scores and loading plots. For any kind of metabolomics data, it is a very good idea to start with PCA, since it will help you to get to know your data. In addition, PCA is very useful to identify potential outliers, which you can decide to exclude or not after inspecting those in the raw data. The data from PAPER I, II and III were subjected to initial PCA for outlier detection and explorative purposes.

Variations caused by sample collection/preparation or instrumental issues will also be reflected on PCA. For instance, Rasmussen et al. [66] used PCA on urine samples analysed by NMR to evaluate the effects of sample storage conditions. PCA has been widely utilised to assess the analytical performance in metabolomics studies [67,68]. As an example, Eva et al. [69] evaluated PCA in

terms of repeatability and robustness of quality control samples in order to optimise the UPLC-MS method for metabolomics analysis. Another example for application of PCA for analytical quality check has been shown in PAPER I. As shown in Figure 12, some samples were positioned apart from their instrumental replicates according to PC1, which was further justified to be caused by instrumental signal drift.

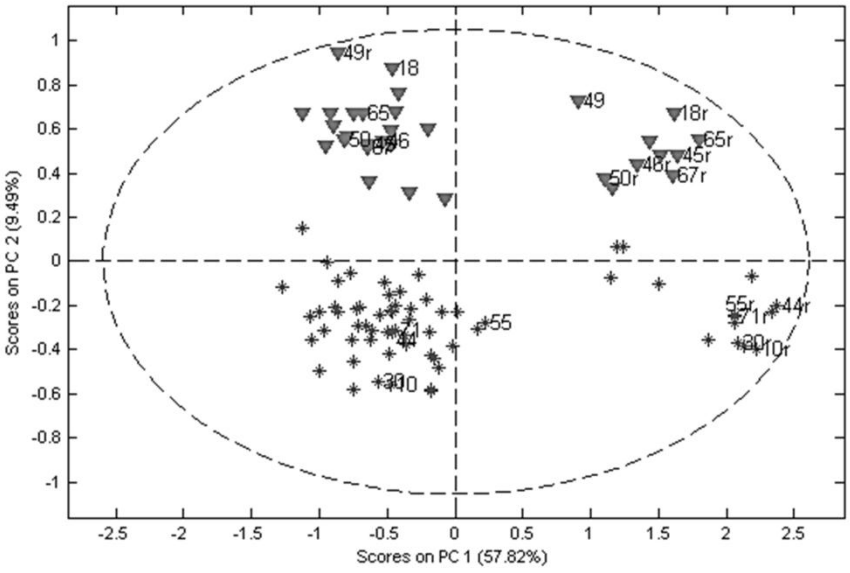


Figure 12. Score plot of PC1 vs. PC2. The instrumental replicates are separated in PC1. The same numbers indicate the replicative measurements. (Data from PAPER I)

The above mentioned applications of PCA are either for optimisation of analysis methods or for evaluating the normalization. Indeed, the core aim of data analysis in metabolomics is to extract metabolites related to specific exposure (e.g. disease *vs.* healthy, case *vs.* control), which boils down to application of PCA for variable selection purposes. For instance, OuYang et al. [70] analysed NMR profiles of serum samples from cancer patients and healthy controls with PCA. PCA score plot revealed a clear distinction between the control and cancer groups so that the representative metabolites were selected from the corresponding loading plot.

However, there are not so many PCA based metabolomics applications for selection of significant metabolites. The reasons for this are based on two drawbacks of PCA:

(1) PCA searches the global patterns and it is not efficient in finding local patterns which is very common in metabolomics data due to its complex nature [71].

(2) Principal components are linear combinations of all variables, thus, considering the large number of variables, it is not easy to point out a group of metabolites among many irrelevant

ones. Sparse PCA has been suggested to overcome this issue by forcing less effective metabolites to have zero loadings [72].This method has been employed in PAPER III and is further discussed in the proceeding section.

4.5.5 Sparse Principal Component Analysis

In 1996, Tibshirani [73] developed a method called Lasso, for estimation of linear models. The lasso is a penalized least squares method, imposing a constraint on the L_1 norm of the regression coefficients. Bounding L_1 norm of PCA model parameters results in a sparse model which makes it favourable for variable selection. Several methods have been proposed for estimating SPCA, based on either the regression error property [74] or the maximum variance property of principal components [75]. In the context of maximizing variance, SPCA can be formulated as a penalized optimization problem with the main objective being a minimization problem similar to PCA but with L_1 norm penalties imposed on the loadings:

$$\text{argmin}(\|X - TP^T\|_F^2)$$

$$\text{subject to } \|p_i\|_1^1 \leq c \text{ and } \|p_i\|_2^2 = 1, \text{ for i = 1,...,k}$$

where X(n x p), is the data matrix, $\|\mathrm{p_i}\|_1^1$ is the sum of absolute values (L_1 norm) of the columns of loading matrix P, and T is the score matrix. The tuning parameter c is a positive penalty parameter bounding the sum of absolute values of the normalized loading vector ($\|p_i\|_1 \leq c$). Thus, it leads to some loadings being exactly zero [76]. If c is chosen large enough, it will lead to unconstrained solution, which will be identical to PCA decomposition. A meaningful sparse solution can be found when c is chosen in between 0 and the sparsity level producing unconstrained solution [76].

Solution of the constrained optimization problem can be solved by deflation where calculation of components is based on the current residual [75]. Alternatively, the calculation of the entire set of components can be done simultaneously by iterating between scores and loadings [76]. For the latter, an alternating least squares-based approach with induced L_1 norm penalty is used for component estimation [76]. Nevertheless, the alternating least square solution may provide local minima. In order to avoid the local minimum, in PAPER III we initialized multiple times with random loadings. It is assumed that a global minimum is achieved if the solution with maximum explained variation (or minimum loss function) is observed multiple times.

Unlike PCA, SPCA does not impose othogonality constraint between components. In SPCA, components are correlated and the loadings are not orthogonal.

Cross-validation has been suggested for the optimization of sparsity penalty selection [76].

4.5.5.1 Application of SPCA in metabolomics

As mentioned previously, PCA with the aim of variable selection can perform poorly for metabolomics data due to large number of irrelevant variables. SPCA allows selection of a limited number of metabolites by penalizing many irrelevant ones to have zero loadings.

Allen et al. [77] developed a modified form of SPCA with non-negativity constraints and showed its potential on NMR based metabolomics data. Furthermore, in Paper III, LC-MS data is subjected to SPCA and PCA. Based on our findings, both SPCA and PCA capture time since last meal patterns from plasma LC-MS profiles (Figure 2-3-4, PAPER III). However, SPCA provided results that were easier to interpret compared to PCA.

Not SPCA, but coupled matrix factorization with imposed sparsity in the variable modes has been developed and the application potential of this tool has been demonstrated on a metabolomics study utilizing two analytical platforms LC-MS and NMR, and a dataset including several clinical end points such as lipoproteins and lipids [78].

4.5.6 Partial Least Square

PLS is a linear regression based method for relating a set of predictor variables, X, with one or more independent variables, Y [79]. The significance of PLS is related to its ability to deal with strongly collinear X variables which makes it suitable for analysis of metabolomics data such as spectral and chromatographic profiles. Like PCA, PLS is a projection based method. PCA aims to find a subspace that explains the maximum amount of variation in X. PLS, on the other hand, tries to find a small dimensional subspace that describes the X well but at the same time the coordinates of this new subspace are good predictors of Y. Similar to PCA, the components are orthogonal. PLS can be formulated as

$$X = T \cdot P' + E$$

$$Y = T \cdot C' + F$$

such that *X* loadings, *P*, are good summaries of *X* and *X* scores, *T*, are good predictors of *Y*.

In metabolomics, PLS has been applied in classification problems where class labels (e.g. case vs. control, exposed vs. unexposed) are used as Y vector. In this case, it is called PLSDA. For the two-

class case, the Y variable is set to have 0 and 1 entries for each class, respectively. PLSDA aims to improve the separation between the two groups by using the class information.

The orthogonal PLSDA (OPLSDA) has been developed as an extension of PLS and it is extensively used in metabolomics [80]. In OPLSDA, the Y unrelated (orthogonal) variation has been removed from X. In this way, OPLSDA attempts to describe classification information in one component. However, the prediction power of PLS and OPLS are usually the same [80,81].

In metabolomics datasets, irrelevant variation is dominant in many cases. As mentioned earlier, PCA tends to gravitate towards to that variation, whereas PLS provides more discriminating latent variables. Thus, PLSDA have been extensively applied in metabolomics both for classification and variable selection.

4.5.6.1 Model validation

Unlike PCA, PLSDA is a supervised data analysis method. Particularly, for metabolomics data, where the number of variables is much larger than the number of samples, there is a potential danger of over-fitting. Thus, careful validation is critical.

In some metabolomics papers, PLSDA scores and loadings plots from models without any indication of validation diagnostic statistics have been presented. However, scores and loadings cannot be trusted if validation shows that the model is not valid. To point out this issue, Westerhuis et al. [82] illustrated that cross-validation of NMR spectra of 23 health volunteers arbitrarily divided into two classes revealed Q^2 values of -0.18, which is considered not to be a good classification. However the PLSDA scores plot showed a clear separation.

The initial step, while building a PLSDA model, is the selection of component number providing the optimal model complexity. Cross-validation has been as a standard tool to determine the number of components. In cross-validation, the samples are divided into training and validation sets. The training set is used to develop models with different number of components (i.e. from 1 to n). These models are evaluated based on their performance for correctly classifying the validation set. Then, the number of components providing the minimum number misclassifications (NMC) is selected. However, assessment of the classification performance of the final model by NMC of the training may lead to over-optimistic validation results. The model is optimised for the samples that are left out, so, those do not assess the validity of the final model [83]. For proper validation, the total data can be divided into training, validation and test sets. The model optimisation is done on training and validation sets and the test set is used to evaluate model performance.

Double cross-validation has been suggested for reducing over-optimism in cross-validation [84,85]. It consists of two nested cross-validation loops as shown in Figure 13. The inner loop cross-validation is used to determine the number of components providing the model with optimum complexity. Then, the final model prediction performance is evaluated by the samples in the test set. The inner and outer loops are repeated N and M times while training, validation and test sets were selected randomly from each class. The number of samples within each set is kept the same for each repetition.

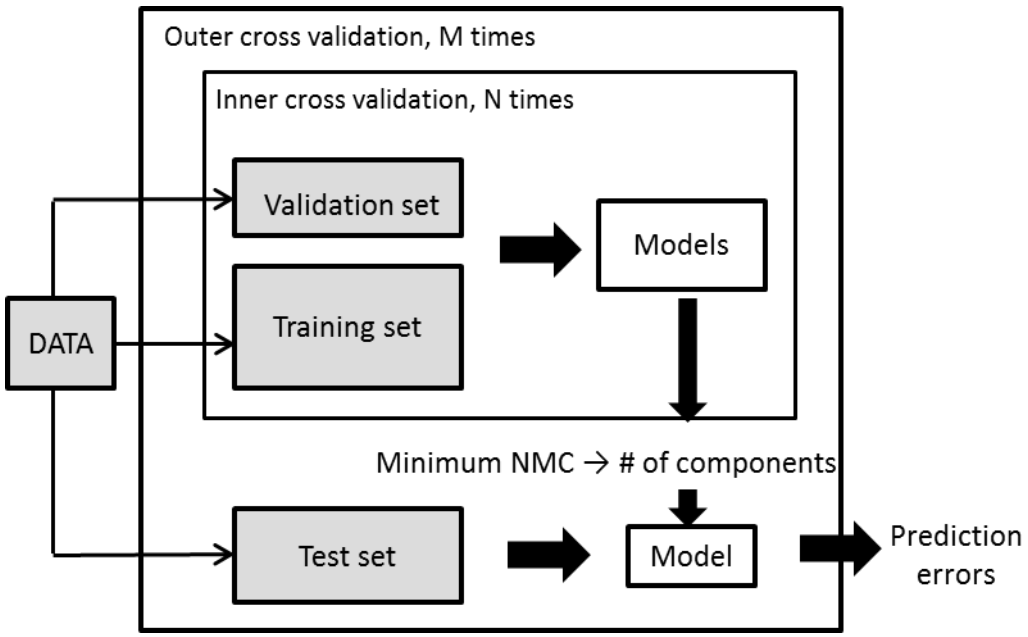


Figure 13. Double cross-validation scheme. In the inner loop, the number of components is determined based on minimum validation set NMC from PLSDA models that are constructed with N different validation and test sets. The number of components that leads to the lowest cross-validation NMC is selected and used to build a model with the corresponding training set. Later, the test set in the outer loop is predicted with this model to give an NMC. The NMC calculated in the M different outer test sets are combined.

In order to decide whether there is a difference between the groups, NMC or other statistic diagnostics (e.g. Q^2 and area under receiver operating characteristic curve) are evaluated. Although it is said that if NMC is lower than 0.5, then there is a difference, it is not known which value of these NMC really corresponds to a good discrimination between groups. Comparison of original classification (two classes) NMC with NMC obtained from the same data but with randomly assigned class labels may provide a better assessment of PLSDA classification performance. This procedure is called permutation test [82]. In general, PLSDA is calculated with random class assignment many times, so that a distribution of NMC can be obtained. The significance level of original classification can be calculated compared to random ones. In case of significance, it can be concluded the original model performs better than random classification. In

permutation test the models with permuted classes is obtained from the same number of samples that also show the same amount of variation, outliers and missing data. Thus, it provides a strong comparison basis. An example of permutation test histogram is given in Figure 14.

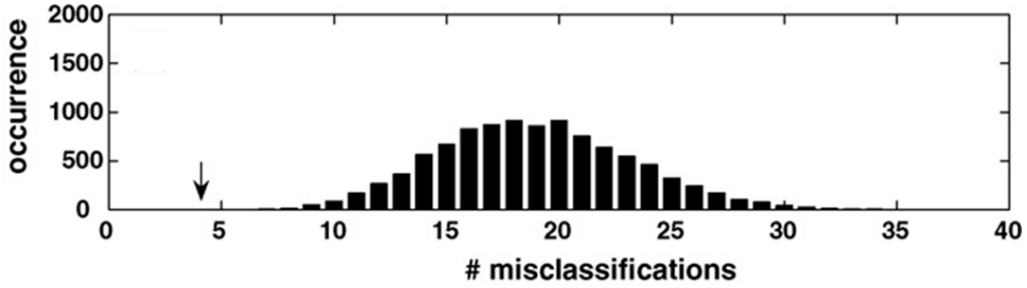


Figure 14. Permutation test. Histogram of the number of misclassifications in 10,000 permutations. Misclassifications are obtained from double cross-validation. The arrows indicate the number of misclassifications in the original problem. From 10,000 permutations none had NMC lower than the original classification [85].

Westerhuis et al. [86] employed permutation test, to illustrate the over-optimistic results when validation set based NMC is used to assess the overall model performance. They used urine NMR profiles of 22 subjects. When there is no difference between the classes, half of the samples are expected to be misclassified. Thus the validation procedure should on average give 11.5 misclassifications for the permuted datasets. As shown in Figure 15, validation set based model evaluation provided over-optimistic results whereas the double cross validation provided the expected NMC.

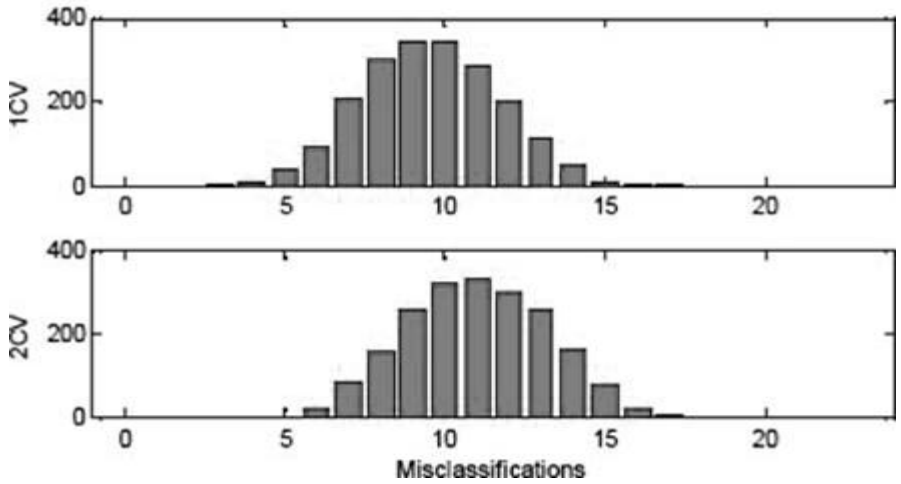


Figure 15. Permutation test applied on a proteomics data. Histogram of the number of misclassifications in 2,000 permutations. NMC are obtained from cross-validation (up) and double cross-validation (down). The expected permuted data NMC is estimated correctly, only by double cross-validation [82].

Both for the datasets PAPER I and II (LC-MS and NMR data) double cross-validation and permutation tests were employed to evaluate the classification performance of the PLSDA models.

4.5.6.2 Variable selection

The metabolite selection in PLSDA is usually based on regression coefficients and variables importance for the projection (VIP) [79]. Regression coefficients represent the importance of a given metabolite for modelling class assignments (Y) whereas VIP summarise its importance for both metabolic profiles (X) and class assignment (Y). Regression coefficients have been employed for selection of discriminating metabolites for PAPER I.

As described in the model validation section, when double cross-validation is employed, the assessment of the model has been performed on multiple subsets of samples, each with its own number of components, variables selected, scaling etc. However, there is no consensus on how to choose the overall model based on sub-model results or which model is to be used for variable selection [83]. In this sense, in PAPER I, firstly, the rank of each feature is recorded based on its absolute regression coefficients from each calculated sub-model (double cross-validation). Then, for each feature, the rank product from all sub-models is calculated which is used to demonstrate the feature's overall importance [85]. This perspective allows each sub-model to contribute in variable selection, yet if one out of many sub-models performs poorly, its effect will be depreciated. Thereby, the features that appear as influential for classification in many models will be selected.

Rajalahti et al. [87] developed a new tool, selectivity ratio, for variable selection in spectral data. Selectivity ratio of a metabolite is calculated as the ratio between explained and residual variance on the target projected component which is a single latent variable explaining the covariance of the X variables with the Y. They have shown on spectral data that variables selected by regression coefficients compared to selectivity ratio may include a larger number of false discoveries. Thus, we have used selectivity ratio as a variable selection tool in PAPER II.

4.6 Identification

In metabolomics studies, data analysis provides a number of metabolites with known m/z and retention time for LC-MS or with chemical shifts for NMR, related to specific exposures. In order to interpret and understand the associated metabolic perturbations in biological systems, the chemical identity of these metabolites should be determined. Four levels of chemical compound identification have been defined by the Metabolomics Standards Initiative as shown in Table 4 [88].

Table 4. Levels for validation of non-novel compounds defined by Metabolomics Standards Initiative [88].

Level	Name	Minimum requirements
1	Identified compounds	At least two independent and orthogonal data relative to an authentic compound analysed under identical experimental conditions. (e.g. retention time/index and mass spectrum, retention time and NMR spectrum, accurate mass and tandem MS, accurate mass and isotope pattern, full ^{1}H and/or ^{13}C NMR, 2-D NMR spectra)
2	Putatively annotated compounds	Without chemical reference standards. Based upon physicochemical properties and/or spectral similarity with public/commercial spectral libraries.
3	Putatively characterized compound classes	Based upon characteristic physicochemical properties of a chemical class of compounds, or by spectral similarity to known compounds of a chemical class
4	Unknown compounds	These metabolites can still be differentiated and quantified based upon spectral data

4.6.1 LC-MS

The identification of the compounds from LC-MS based methods is both analytically and computationally challenging. It is often time-consuming, laborious and considered as a bottleneck in interpretation of metabolomics data. The three main strategies for identification of LC-MS based metabolites include:

- Accurate mass based identification using high-resolution instruments. In combination with accurate mass isotopic distribution may reveal elemental composition of the compound.
- Application of tandem MS where the instrument performs an MS^1 survey scan, and selects one or more ions for subsequent MS^2 or even MS^n scans. This provides the structural information of a compound by exploiting the fragmentation patterns.
- Comparison of retention time and spectra with authentic standards.

Each specific chemical compound gives rise to one or more ion species during ESI, which are included in the same mass spectrum. Those ion species include isotope, fragment, adduct and cluster ions. Inclusion of all ions representing one compound brings out redundancy issues in the data analysis part. Recently, a new R based package called CAMERA [89] has been released which aims to automatically group the features derived from the same analyte and annotates isotope and adducts peaks by utilizing correlations across the samples and similarity of the peak shapes. The assignments may also ease the identification step in the sense that the ion species to search for the accurate mass in spectral databases will be known.

In order to cope with the challenges in metabolite identification, many compound databases have been developed including chemical and physical properties of the compounds. For accurate mass and spectral search the databases, HMDB [4], METLIN [90] and Lipid Maps [91] have been systematically searched in this thesis. The other databases such as Manchester Metabolomics Database [92] contains 42,687 endogenous and exogenous metabolites retrieved from primary sources such as HMDB, Lipid Maps, BioCyc [93] and DrugBank [94]. The MassBank [95] database maintains the spectral information from a wide variety of commonly used mass spectrometry platforms. The spectral database and visualization tools are publicly available and web-accessible which was regularly used for the present work.

For identification of the LC-MS based features from metabolomics studies, additional experiments have been performed. For instance in PAPER I, the authentic standard of sn-2 LPC(18:1) was produced by phospholipase A1 based hydrolysis. Furthermore, post-column lithium infusion was

performed to increase the abundance of PC fragments with lithium adduct formation and thereby improve the structural characterization of PC species in PAPER II.

4.6.2 NMR

NMR identification in PAPER II has been limited to peak assignment from comparison of chemical shift with previously published blood plasma metabolites [96]. However, database libraries used for MS searches such as HMDB [4], KEGG [97] and MetaCyc [98] also contain ^{1}H and ^{13}C NMR assignment of the metabolites. These sources provide reliable assignments of NMR spectra for identification of metabolites.

Nevertheless, 1D ^{1}H NMR spectra suffers from peak overlap thereby complicating the identification and quantification of metabolites. 2-D NMR methods offer the benefits of 1-D NMR but additionally resolving the overlapping resonances into a second dimension, and increasing metabolite specificity. Thus 2-D NMR methods have the potential for application in metabolomics with the advantage of improved identification. Recently, Birmingham Metabolite Library have been established with the database of 1-D and 2-D *J-resolved* NMR spectra [99].

4.7 Biological Interpretation

The final step in the metabolomics pipeline is interpretation of the identified compounds reflecting a specific exposure in biological terms. In the next sections, the biology behind the identified compounds from studies involved in PAPER I, II and III will be discussed.

4.7.1 Meal responses

Human metabolism shifts constantly between anabolic (fed) conditions after food intake and catabolic states between meals or during extended starvation periods. Insulin is the main coordinator of this shift where high levels of insulin modulate energy storage in the anabolic state and low levels of insulin and high levels of glucagon control energy expenditure in the catabolic state [100].

In the anabolic state, after food intake, insulin enhances utilization of glucose as a prime energy substrate by muscle, adipose tissue and liver and promotes hepatic synthesis of glycogen while inhibiting gluconeogenesis and glycogenolysis. Furthermore, triacylglyceride formation is favoured with uptake of fatty acids from plasma for energy storage. Also, protein synthesis increases by amino acid uptake from plasma into muscle and liver. The catabolic state involves a series of

adaptations to ensure adequate fuels for body tissues in the absence of exogenous substrate. When the insulin level drops, the liver becomes an organ of glucose production to provide energy. In addition, lipolysis and proteolysis increases for energy production. Plasma fatty acid levels rise to maintain energy levels via ß-oxidation and acetyl-CoA production. Branched chain amino acids are fuels for energy production and plasma levels decrease initially during the first hours after a meal whereas increased levels after prolonged fasting (>8-16hrs) are indicators of a high rate of protein breakdown [100,101]. The schematic representation of the dynamics of plasma metabolite changes is given in Figure 16.

Carnitine is required to assist the transport and metabolism of long-chain fatty acids in mitochondria, where they are oxidized as a major source of energy. Thus, during fasting, long-chain and short chain acylcarnitines increase with a decrease in free carnitine [102]. After oxidation of fatty acids, acetyl-CoA is produced and gives rise to formation of the so-called ketone bodies, acetone, acetoacetate and β-hydroxybutyric acid. Ketone bodies provide an alternative fuel to body tissues, especially to the brain during fasting. The brain can utilise only ketone bodies as an energy source when glucose levels are not sufficient [103].

The above mentioned metabolic patterns reflecting the body's shift from anabolism to catabolism has been confirmed in PAPER I. During fasting state compounds characteristic to lipolysis such as fatty acids, β-hydroxybutyric acid (ketone body), acetyl-carnitine and acyl-carnitines are increased, whereas L-carnitine is decreased. The amino acids and lyso-lipids were higher in plasma at fed state.

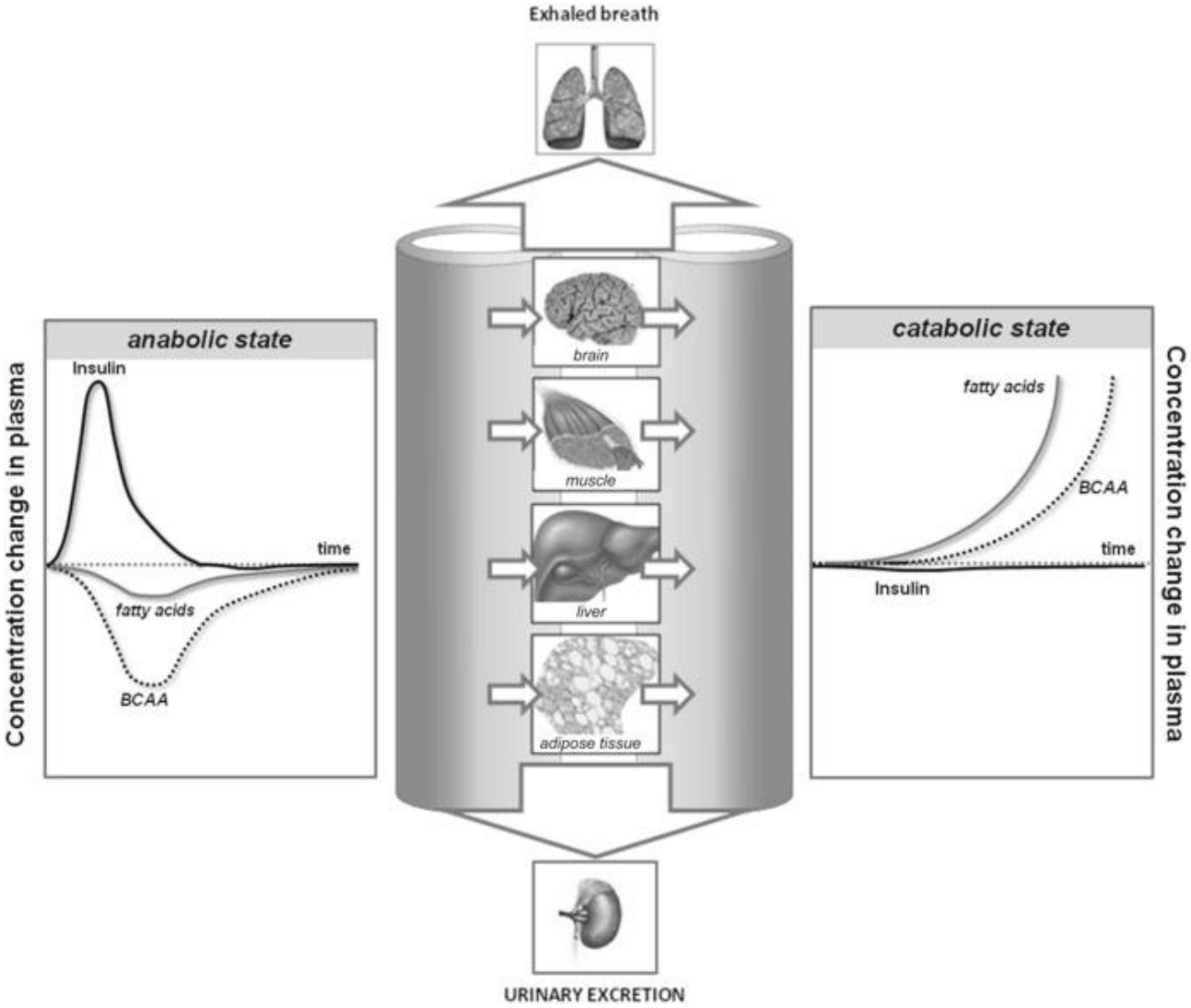


Figure 16. Dynamics of plasma metabolite changes between anabolic (fed) and catabolic (fasted) states. Detailed explanation is in the text [100].

The response to food intake and metabolite clearance rates vary depending on the quality and quantity of the food source [104] and the physiological differences between subjects such as gender, age and weight [105]. The metabolic responses to food intake and metabolite clearance rates are usually measured by postprandial challenge tests. These may be performed by glucose tolerance tests (OGTT or clamps), lipid challenges, or by specific foods or whole meal challenges, depending on the specific metabolite group of interest. However, time-resolved changes of the human metabolome in response to a challenge have been very rarely investigated in metabolomics studies. Instead, fasting state plasma samples have been used, typically following an overnight fast, as it is considered to be more reproducible. Nevertheless, recent metabolomics studies have demonstrated that challenge tests increase metabolite variability between volunteers, allowing discrete metabotypes to be identified that would not be seen in normal

fasting conditions. Zhao et al. [106] and Shaham et al. [107] were the first to utilize metabolomics to investigate the physiological changes during an OGTT. They identified major concentration changes in compounds, such as bile acids, that have not been reported previously. In addition, Shaham et al. [107] demonstrated that time-resolved metabolic profiling has the potential to define an individual's 'insulin response profile', which could have value in predicting diabetes. This has been shown by pre-diabetic individuals' selective resistance to suppression of either proteolysis or lipolysis. Wopereis et al. [108] have shown the effect of the diclofenac treatment can only be revealed by investigating metabolic patterns with OGTT and time course. Krug et al. [109] submitted 15 young healthy male volunteers to a highly controlled 4 d challenge protocol, including 36 h fasting, OGTT and lipid test, liquid test meals, physical exercise, and cold stress. They have shown that physiological challenges increased inter-individual variation even in phenotypically similar volunteers, revealing metabotypes not observable in baseline metabolite profiles. Another study investigated the metabolic perturbation in response to a postprandial challenge in a controlled intervention study [29]. All these studies provided unique findings illustrating that the profiles obtained from metabolic challenge tests are more informative than using fasting state profiles.

In PAPER III, based on a cross-sectional study group, time since last meal related pattern, revealed higher levels of amino acids and LPCs in volunteers who were considered to be in postprandial state, so even under free-living conditions it is possible to reproduce part of the patterns observed in controlled settings (PAPER I).

4.7.2 Trans fatty acids

Industrially produced TFAs are formed during the partial hydrogenation of vegetable oil that changes *cis* configuration of double bond(s) to *trans*, resulting in semi-solid fats for use in margarines, commercial cooking, and manufacturing processes. Partially hydrogenated vegetable oils are appealing because of their long shelf life, their stability during deep-frying, and their semi-solidity, which is utilised to enhance the palatability of baked goods and sweets.

Partially hardened vegetable oils mainly contain *trans* isomers of oleic acid (Figure 17, left), the major one being C18:1 *trans*-9 or elaidic acid (Figure 17, right) and C18:1 *trans*-10. In addition, smaller amounts of C18: 1 *trans*-8, and C18:1 *trans*-11, and *trans* isomers of alpha-linolenic acid may arise during deep-fat frying [110].

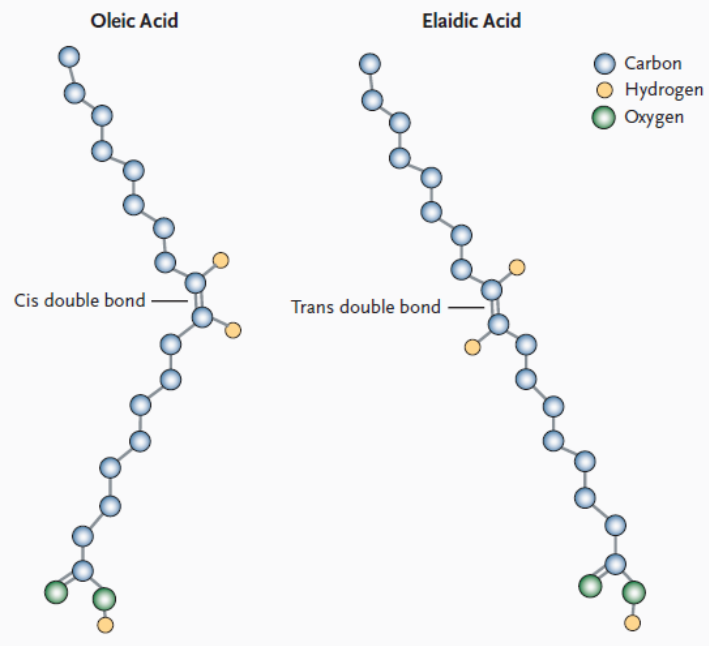


Figure 17. Chemical structure of *cis* and TFA. Both oleic acid and elaidic acid has 18 carbons. In *cis* configuration (e.g. oleic acid) the carbon chain extends from the same side of the double bond, causing a bent molecule, whereas in trans configuration (e.g. elaidic acid) the carbon chain extends from opposite sides of the double bond, providing a straight molecule [111].

4.7.2.1 Health effects

TFA intake has been identified as a modifiable dietary risk factor of coronary heart disease (CHD). Consumption of TFA, on a per-calorie basis, potentially increases the risk of CHD more than any other micronutrient. In a meta-analysis of four prospective cohort studies involving nearly 140,000 subjects Mozaffarian et al. [111] have demonstrated that a 2 % increase in energy intake from TFAs raised the incidence of CHD with 23 %.

The adverse effects of TFA consumption on serum lipids in humans has been demonstrated by randomised, controlled trials. In a meta-study of eight selected trials [112], isoenergetic replacement of saturated or *cis* unsaturated fats with TFAs raised the level of total cholesterol (TC) to high-density lipoprotein cholesterol (HDL-C) in the blood. In relation to that, in PAPER II, LDL-C is increased with TFA intake based on plasma NMR profiles. Also, TFA intake has been shown to have unfavourable effects on triglycerides, apolipoprotein (Apo) B/ApoAI ratio and C-reactive protein [113]. Although alteration of blood lipids, particularly an increase in TC/HDL-C ratio, is associated with CHD, the relation of TFA intake with the incidence of CHD has been greater than that predicted by changes in blood lipid levels alone [113,114]. This implies that the mechanisms behind the adverse effects of TFAs are not fully understood. Bendsen et al. [115] have shown that TFA consumption may involve in activation of TNF-α as a possible mechanism leading to

development of CHD (study group and design from PAPER II). In some studies TFA intake in the human diet has also been associated with type 2 diabetes. Two prospective studies found positive associations of TFA intake and type 2 diabetes [116,117] whereas no association has been observed in two other prospective cohort studies [118,119]. Moreover, in a 16-week randomised controlled trial, there was no relation between TFA intake and glucose metabolism (study group and design from PAPER II) [120]. Furthermore, higher plasma phospholipid and erythrocyte membrane particularly including 18:2 TFA (*trans*-18:2) are associated with higher risks of fatal ischemic heart disease [121] and sudden cardiac death [122].

4.7.2.2 TFAs effects membrane properties

Fatty acids are incorporated into phospholipids in all cell membranes of the body so dietary TFA level was reported to directly reflect the TFA uptake to the membrane [123]. The fatty acid composition of the membrane can strongly influence its physical characteristics. It has been shown that TFAs convey membrane properties such as lateral lipid packing, fluidity and permeability more similar to saturated fatty acids than their *cis* forms [124]. The *trans* double bond (Figure 17) produces a linear conformation resembling more a saturated chain, which provides better chain packing than a *cis* double bond [125]. It may be assumed that more tightly packed (*trans* isomer) membranes should be less permeable than membranes whose lipids are loosely packed (*cis* isomer). Depending on the similar basis, *cis*-PC membranes are more 'fluid' than *trans* containing membranes. The efficiency of molecular signal transduction is highly dependent on the orientation and positioning of various proteins within the cell membrane, which can be related to adverse health effects of TFAs.

In PAPER II, indications of preferential incorporation of TFAs to PCs with longer chain and higher saturation have been observed which could potentially cause membranes dysfunctioning of the cell.

5 Conclusions

This thesis aimed to examine the whole process involved in LC-MS and NMR based nutritional metabolomics studies, from data preprocessing, through data analysis and compound identification to interpretation of results in biological terms. For this purpose three independent datasets involving different study designs - an animal study, a human intervention with parallel design and a prospective cohort - were analysed. In the first study (PAPER I), the influence of LC-MS data preprocessing on the marker selection has been investigated. The data was preprocessed with three different tools MZmine, XCMS, MarkerLynx and a customised method (binning and summation through retention time index). The main conclusions from the study are as follows:

- A customised method which is considered as a more primitive data preprocessing approach leads to identification of a few false positives and false negatives but at the same time allows fast preprocessing.
- Each software tool employs different methods in their peak detection and alignment algorithms such that each has pros and cons to detect specific features.
- The selection of proper parameters for each tool based on the characteristics of the dataset is the key for obtaining high quality preprocessed data. Furthermore, the use of more than one software and/or the use of several settings during data preprocessing with any software is likely to improve marker detection in untargeted metabolomics.

Analysis of metabolomics data is challenging because large amounts of complex data is generated from relatively few samples. In order to analyse the complex datasets in this thesis, PCA has been an extremely useful tool not only for providing an overview of the dominant patterns but also for detection of outliers, and evaluation of preprocessing and pre-treatment methods. However, PCA is not very efficient for extraction of relevant metabolites from the vast number of irrelevant ones which is the core aim of data analysis in metabolomics studies. PAPER III explores this issue and aims to compare PCA with its modified version SPCA. The results suggest that SPCA and PCA are equally good to capture relevant patterns, yet the selection of representative metabolites is much easier with SPCA. Therefore, SPCA can potentially be applied for variable selection purposes in LC-MS based metabolomics. In metabolomics studies, the relevant patterns are rarely the dominant ones, thus unsupervised methods such as PCA (or SPCA) do not always work. Therefore, a supervised approach, PLSDA, has been employed for PAPER I and II. PLSDA is prone to overfitting particularly for datasets where the number of variables is much larger than the number of

samples. To overcome this problem, double cross-validation routine is applied. The selection of relevant patterns is based on regression coefficients and selectivity ratio.

In order to identify the selected compounds from LC-MS data besides the routines, enzymatic reactions has been performed when the compounds were not commercially available (PAPER I). Furthermore, it has been demonstrated that lithiated adducts of phospholipids have enhanced ionization and class specific fragmentation in MS/MS scan modes (PAPER II).

The identified compounds have been interpreted in terms of biological interest. In the first study (PAPER I) the aim was to extract the metabolic patterns related to fasting (12h) and fed states from rat LC-MS plasma profiles. This was further extended to a prospective cohort with the slightly different focus to identify time since last meal related patterns (PAPER III). Although similar purposes were involved, a rat model offers full control of the food intake whereas a cohort study provides un-controlled observational settings. The major conclusions from these two studies can be summarized as follows:

- Only for fasted rats, compounds such as fatty acids, β-hydroxybutyric acid (ketone body), acetyl-carnitine and acyl-carnitines in plasma increased, which suggests an upregulated energy production via lipolysis. The promoted lipolysis indicates body's shift to catabolism. However in the cohort study, the few subjects had the last meal more than 12 h, yet most of them had a drink independently of their recorded TSLM. Thus, most were probably not in the fasting state. On the other hand, the rats have higher rate of energy metabolism than humans and for that reason, overnight fasting represents a more extreme condition than in humans.
- In both studies, high levels of amino acids with recent food intake (fed state) were found, which is linked to protein sources introduced from the last meal.
- In both studies, lyso-lipids were higher after food intake and decreased with time.

In PAPER II, NMR and LC-MS untargeted metabolomics has been used as an approach to explore the effect of industrially produced TFA intake on plasma metabolites. The well-known adverse effects of TFA on serum lipids were confirmed by NMR in terms of increased LDL cholesterol levels. On the other hand, LC-MS findings have demonstrated that in overweight healthy women, intake of industrially produced TFA affects lipid metabolism by increasing the concentration of specific PCs and an SM. The indications for preferential integration of trans18:1 into the sn-1 position of phosphatidylcholines, all containing PUFA in the sn-2 position, could be explained by a general up-regulation in the formation of long-chain PUFAs after TFA intake and/or by specific mobilisation of

these fats into phosphatidylcholines as a result of TFA exposure. NMR supported these findings by revealing increased unsaturation of plasma lipids in the TFA group.

In conclusion, the utilization of metabolomics to disentangle the metabolic perturbations requires detailed understanding of the system under study, of the analytical technologies and their specific effects on the data produced, so that suitable data preprocessing and analysis strategies can be applied.

In terms of biology, lyso-lipids and amino acids emerged as the most dominating patterns for identification of recent food intake. On the other hand, TFA intake caused specific changes in membrane lipid species which may be related to the mechanisms of trans fat-induced diseases.

6 Future Perspectives

In this thesis LC-MS and NMR based metabolomics has been demonstrated as a powerful tool to disclose the underlying metabolic patterns reflecting the (1) postprandial response to food intake and its clearance rate and (2) TFA intake.

The postprandial response reveals multiple aspects of metabolic health that would not be apparent from studying the fasting parameters. Thus, investigating the effect of specific food intake or disease by utilising postprandial response of individuals has the potential to identify the discrete metabolic profiles that would not be seen in normal fasting conditions. In fact, this issue was explored with data from pilot DCH (Diet, Cancer and Health) cohort with the aim of resolving cancer and postprandial response interactions (PAPER III). However, the number of subjects was not sufficient to describe the group specific trends. Potentially, the same principle can be applied to the larger cohort of DCH, where LC-MS plasma profiles of ~3000 individuals have been recorded. Indeed using this set, metabolic profiles in terms of postprandial response can be examined to assess the incidence of diseases such as overweight, diabetes and CHD. In order to analyse this data advanced methods (e.g. multi-way data analysis tools) is required, as the purpose is the identification of time series metabolite evaluations in discrete classes (e.g. healthy/disease).

The samples of DCH cohort were collected in the '90s before TFA was banned in Denmark, making this data set suitable to investigate the impacts of long term exposure to TFAs. TFA intake of each individual can be assigned from plasma LC-MS profiles using previously identified TFA exposure markers (PAPER II). Then, the associations between long term exposure to TFA and the incidence of diseases such as CVD and diabetes can be explored.

7 Reference List

1. Fiehn O (2001) Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. Comp Funct Genom 2: 155-168.

2. Nicholson JK, Lindon JC, Holmes E (1999) 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica 29: 1181-1189.

3. Pauling L, Robinson AB, TERANISH.R, Cary P (1971) Quantitative Analysis of Urine Vapor and Breath by Gas-Liquid Partition Chromatography. P Natl Acad Sci USA 68: 2374-2376.

4. Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, Hau DD, Psychogios N, Dong E, Bouatra S, Mandal R, Sinelnikov I, Xia JG, Jia L, Cruz JA, Lim E, Sobsey CA, Shrivastava S, Huang P, Liu P, Fang L, Peng J, Fradette R, Cheng D, Tzur D, Clements M, Lewis A, De Souza A, Zuniga A, Dawe M, Xiong YP, Clive D, Greiner R, Nazyrova A, Shaykhutdinov R, Li L, Vogel HJ, Forsythe I (2009) HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res 37: D603-D610.

5. Manach C, Hubert J, Llorach R, Scalbert A (2009) The complex links between dietary phytochemicals and human health deciphered by metabolomics. Mol Nutr Food Res 53: 1303-1315.

6. Primrose S, Draper J, Elsom R, Kirkpatrick V, Mathers JC, Seal C, Beckmann M, Haldar S, Beattie JH, Lodge JK, Jenab M, Keun H, Scalbert A (2011) Metabolomics and human nutrition. Br J Nutr 105: 1277-1283.

7. Kristensen M, Engelsen SB, Dragsted LO (2012) LC-MS metabolomics top-down approach reveals new exposure and effect biomarkers of apple and apple-pectin intake. Metabolomics 8: 64-73.

8. Duggan GE, Miller BJ, Jirik FR, Vogel HJ (2011) Metabolic profiling of vitamin C deficiency in Gulo-/- mice using proton NMR spectroscopy. J of Biomol NMR 49: 165-173.

9. Nieman DC, Gillitt N, Jin F, Henson DA, Kennerly K, Shanely RA, Ore B, Su M, Schwartz S (2012) Chia Seed Supplementation and Disease Risk Factors in Overweight Women: A Metabolomics Investigation. J Altern Complem Med 18: 700-708.

10. Kau AL, Ahern PP, Griffin NW, Goodman AL, Gordon JI (2011) Human nutrition, the gut microbiome and the immune system. Nature 474: 327-336.

11. Moco S, Martin FP, Rezzi S (2012) Metabolomics View on Gut Microbiome Modulation by Polyphenol-rich Foods. J Proteome Res 11: 4781-4790.

12. Fardet A, Canlet C, Gottardi G, Lyan B, Llorach R, Remesy C, Mazur A, Paris A, Scalbert A (2007) Whole-grain and refined wheat flours show distinct metabolic profiles in rats as assessed by a H-1 NMR-based metabonomic approach. J Nutr 137: 923-929.

13. Mcniven EMS, German JB, Slupsky CM (2011) Analytical metabolomics: nutritional opportunities for personalized health. J Nutr Biochem 22: 995-1002.

14. Astley, S. and Penn, L. (2009) Design of human nutrigenomics studies. Wageningen: Wageningen Academic Publishers.

15. Westerhuis JA, van Velzen EJJ, Hoefsloot HCJ, Smilde AK (2010) Multivariate paired data analysis: multilevel PLSDA versus OPLSDA. Metabolomics 6: 119-128.

16. Holmes E, Loo RL, Stamler J, Bictash M, Yap IKS, Chan Q, Ebbels T, De Iorio M, Brown IJ, Veselkov KA, Daviglus ML, Kesteloot H, Ueshima H, Zhao LC, Nicholson JK, Elliott P (2008) Human metabolic phenotype diversity and its association with diet and blood pressure. Nature 453: 396-U50.

17. Wang-Sattler R, Yu Z, Herder C, Messias AC, Floegel A, He Y, Heim K, Campillos M, Holzapfel C, Thorand B, Grallert H, Xu T, Bader E, Huth C, Mittelstrass K, Doring A, Meisinger C, Gieger C, Prehn C, Roemisch-Margl W, Carstensen M, Xie L, Yamanaka-Okumura H, Xing G, Ceglarek U, Thiery J, Giani G, Lickert H, Lin X, Li Y, Boeing H, Joost HG, de Angelis MH, Rathmann W, Suhre K, Prokisch H, Peters A, Meitinger T, Roden M, Wichmann HE, Pischon T, Adamski J, Illig T (2012) Novel biomarkers for pre-diabetes identified by metabolomics. Mol Syst Biol 8: 615.

18. Floegel A, Stefan N, Yu Z, Muhlenbruch K, Drogan D, Joost HG, Fritsche A, Haring HU, Hrabe de AM, Peters A, Roden M, Prehn C, Wang-Sattler R, Illig T, Schulze MB, Adamski J, Boeing H, Pischon T (2012) Identification of Serum Metabolites Associated With Risk of Type 2 Diabetes Using a Targeted Metabolomic Approach. Diabetes 62(2):639-48.

19. Menni C, Zhai G, MacGregor A, Prehn C, R+Âmisch-Margl W, Suhre K, Adamski J, Cassidy A, Illig T, Spector T, Valdes A (2012) Targeted metabolomics profiles are strongly correlated with nutritional patterns in women. Metabolomics 1-9.

20. Alvarez-Sanchez B, Priego-Capote F, de Castro MDL (2010) Metabolomics analysis I. Selection of biological samples and practical aspects preceding sample preparation. Trac-Trend Anal Chem 29: 111-119.

21. Zhang AH, Sun H, Wang P, Han Y, Wang XJ (2012) Recent and potential developments of biofluid analyses in metabolomics. J Proteomics 75: 1079-1088.

22. Vuckovic D (2012) Current trends and challenges in sample preparation for global metabolomics using liquid chromatography-mass spectrometry. Anal Bioanal Chem 403: 1523-1548.

23. Griffiths WJ (2008) Metabolomics, metabonomics and metabolite profiling. Cambridge: RSCPublishing.

24. Barri T, Holmer-Jensen J, Hermansen K, Dragsted LO (2012) Metabolic fingerprinting of high-fat plasma samples processed by centrifugation- and filtration-based protein precipitation delineates significant differences in metabolite information coverage. Anal Chim Acta 718: 47-57.

25. Lambert, Joseph B. and Mazzola, Eugene P. (2004) Nuclear magnetic resonance spectroscopy an introduction to principles, applications, and experimental methods. Upper Saddle River, N.J: Pearson Education.

26. MacNair JE, Lewis KC, Jorgenson JW (1997) Ultrahigh-pressure reversed-phase liquid chromatography in packed capillary columns. Anal Chem 69: 983-989.

27. Griffin JL, Atherton H, Shockcor J, Atzori L (2011) Metabolomics as a tool for cardiac research. Nat Rev Cardiol 8: 630-643.

28. Rubio-Aliaga I, de Roos B, Duthie SJ, Crosley LK, Mayer C, Horgan G, Colquhoun IJ, Le Gall G, Huber F, Kremer W, Rychlik M, Wopereis S, van Ommen B, Schmidt G, Heim C, Bouwman FG, Mariman EC, Mulholland F, Johnson IT, Polley AC, Elliott RM, Daniel H (2011) Metabolomics of prolonged fasting in humans reveals new catabolic markers. Metabolomics 7: 375-387.

29. Pellis L, van Erk MJ, van Ommen B, Bakker GCM, Hendriks HFJ, Cnubben NHP, Kleemann R, van Someren EP, Bobeldijk I, Rubingh CM, Wopereis S (2012) Plasma metabolomics and proteomics profiling after a postprandial challenge reveal subtle diet effects on human metabolic status. Metabolomics 8: 347-359.

30. Wu Z, Li M, Zhao C, Zhou J, Chang Y, Li X, Gao P, Lu X, Li Y, Xu G (2010) Urinary metabonomics study in a rat model in response to protein-energy malnutrition by using gas chromatography-mass spectrometry and liquid chromatography-mass spectrometry. Mol Biosyst 6: 2157-2163.

31. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem 78: 779-787.

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

33. Lommen A (2009) MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. Anal Chem 81: 3079-3086.

34. Yu TW, Park Y, Johnson JM, Jones DP (2009) apLCMS-adaptive processing of high-resolution LC/MS data. Bioinformatics 25: 1930-1936.

35. Jaitly N, Mayampurath A, Littlefield K, Adkins JN, Anderson GA, Smith RD (2009) Decon2LS: An open-source software package for automated processing and visualization of high resolution mass spectrometry data. BMC Bioinformatics 10: 87.

36. Castillo S, Gopalacharyulu P, Yetukuri L, Oresic M (2011) Algorithms and tools for the preprocessing of LC-MS metabolomics data. Chemom Intell Lab Syst 108: 23-32.

37. Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, Tang WH, Ropp A, Neumann S, Pizarro AD, Montecchi-Palazzi L, Tasman N, Coleman M, Reisinger F, Souda P, Hermjakob H, Binz PA, Deutsch EW (2011) mzML-a Community Standard for Mass Spectrometry Data. Mol Cell Proteomics 10.

38. Deutsch EW (2010) Mass spectrometer output file format mzML. Methods Mol Biol 604: 319-331.

39. Smedsgaard J, Nielsen J (2005) Metabolite profiling of fungi and yeast: from phenotype to metabolome by MS and informatics. J Exp Bot 56: 273-286.

40. Tautenhahn R, Bottcher C, Neumann S (2008) Highly sensitive feature detection for high resolution LC/MS. BMC Bioinformatics 9: 504.

41. Sandra K, Pereira AD, Vanhoenacker G, David F, Sandra P (2010) Comprehensive blood plasma lipidomics by liquid chromatography/quadrupole time-of-flight mass spectrometry. J Chromatogr A 1217: 4087-4099.

42. Nielsen NJ, Tomasi G, Frandsen RJN, Kristensen MB, Nielsen J, Giese H, Christensen JH (2010) A pre-processing strategy for liquid chromatography time-of-flight mass spectrometry metabolic fingerprinting data. Metabolomics 6: 341-352.

43. Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G (2012) XCMS Online: a web-based platform to process untargeted metabolomic data. Anal Chem 84: 5035-5039.

44. Forshed J, Schuppe-Koistinen I, Jacobsson SP (2003) Peak alignment of NMR signals by means of a genetic algorithm. Anal Chim Acta 487: 189-199.

45. Vogels JTWE, Tas AC, Venekamp J, VanderGreef J (1996) Partial linear fit: A new NMR spectroscopy preprocessing tool for pattern recognition applications. J Chemometr 10: 425-438.

46. Veselkov KA, Lindon JC, Ebbels TMD, Crockford D, Volynkin VV, Holmes E, Davies DB, Nicholson JK (2009) Recursive Segment-Wise Peak Alignment of Biological H-1 NMR Spectra for Improved Metabolic Biomarker Recovery. Anal Chem 81: 56-66.

47. Savorani F, Tomasi G, Engelsen SB (2010) icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. J Magn Reson 202: 190-202.

48. MacKinnon N, Ge W, Khan AP, Somashekar BS, Tripathi P, Siddiqui J, Wei JT, Chinnaiyan AM, Rajendiran TM, Ramamoorthy A (2012) Variable Reference Alignment: An Improved Peak Alignment Protocol for NMR Spectral Data with Large Intersample Variation. Anal Chem 84: 5372-5379.

49. Blekherman G, Laubenbacher R, Cortes DF, Mendes P, Torti FM, Akman S, Torti SV, Shulaev V (2011) Bioinformatics tools for cancer metabolomics. Metabolomics 7: 329-343.

50. Sugimoto M, Kawakami M, Robert M, Soga T, Tomita M (2012) Bioinformatics Tools for Mass Spectroscopy-Based Metabolomic Data Processing and Analysis. Current Bioinformatics 7: 96-108.

51. Taylor PJ (2005) Matrix effects: the Achilles heel of quantitative high-performance liquid chromatography-electrospray-tandem mass spectrometry. Clin Biochem 38: 328-334.

52. Rubingh CM, Bijlsma S, Derks EPPA, Bobeldijk I, Verheij ER, Kochhar S, Smilde AK (2006) Assessing the performance of statistical validation tools for megavariate metabolomics data. Metabolomics 2: 53-61.

53. Sysi-Aho M, Katajamaa M, Yetukuri L, Oresic M (2007) Normalization method for metabolomics data using optimal selection of multiple internal standards. BMC Bioinformatics 8: 93.

54. Redestig H, Fukushima A, Stenlund H, Moritz T, Arita M, Saito K, Kusano M (2009) Compensation for systematic cross-contribution improves normalization of mass spectrometry based metabolomics data. Anal Chem 81: 7974-7980.

55. van den Berg RA, Hoefsloot HCJ, Westerhuis JA, Smilde AK, van der Werf MJ (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics 7: 142.

56. Dieterle F, Ross A, Schlotterbeck G, Senn H (2006) Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in H-1 NMR metabonomics. Anal Chem 78: 4281-4290.

57. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185-193.

58. Craig A, Cloareo O, Holmes E, Nicholson JK, Lindon JC (2006) Scaling and normalization effects in NMR spectroscopic metabonomic data sets. Anal Chem 78: 2262-2267.

59. Pearson K (1901) On lines and planes of closest fit to systems of points in space. Philos Mag 2: 559-572.

60. Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24: 498-520.

61. Wold S, Esbensen K, Geladi P (1987) Principal Component Analysis. Chemom Intell Lab Syst 2: 37-52.

62. Wang ZG, Chen Z, Yang SS, Wang Y, Yu LF, Zhang BC, Rao ZG, Gao JF, Tu SH (2012) H-1 NMR-based metabolomic analysis for identifying serum biomarkers to evaluate methotrexate treatment in patients with early rheumatoid arthritis. Experimental and Therapeutic Medicine 4: 165-171.

63. Wang X, Yang B, Sun H, Zhang A (2012) Pattern recognition approaches and computational systems tools for ultra performance liquid chromatography-mass spectrometry-based comprehensive metabolomic profiling and pathways analysis of biological data sets. Anal Chem 84: 428-439.

64. Turner E, Brewster JA, Simpson NAB, Walker JJ, Fisher J (2007) Plasma from women with Preeclampsia has a low lipid and ketone body content - A nuclear magnetic resonance study. Hypertens Pregnancy 26: 329-342.

65. Park Y, Jones DP, Ziegler TR, Lee K, Kotha K, Yu TW, Martin GS (2011) Metabolic effects of albumin therapy in acute lung injury measured by proton nuclear magnetic resonance spectroscopy of plasma: A pilot study. Crit Care Med 39: 2308-2313.

66. Rasmussen LG, Savorani F, Larsen TM, Dragsted LO, Astrup A, Engelsen SB (2011) Standardization of factors that influence human urine metabolomics. Metabolomics 7: 71-83.

67. Theodoridis G, Gika HG, Wilson ID (2008) LC-MS-based methodology for global metabolite profiling in metabonomics/metabolomics. Trac-Trend Anal Chem 27: 251-260.

68. Bruce SJ, Tavazzi I, Parisod V, Rezzi S, Kochhar S, Guy PA (2009) Investigation of Human Blood Plasma Sample Preparation for Performing Metabolomics Using Ultrahigh Performance Liquid Chromatography/Mass Spectrometry. Anal Chem 81: 3285-3296.

69. Zelena E, Dunn WB, Broadhurst D, Francis-McIntyre S, Carroll KM, Begley P, O'Hagan S, Knowles JD, Halsall A, Wilson ID, Kell DB (2009) Development of a Robust and Repeatable UPLC-MS Method for the Long-Term Metabolomic Study of Human Serum. Anal Chem 81: 1357-1364.

70. OuYang D, Xu JJ, Huang HG, Chen Z (2011) Metabolomic Profiling of Serum from Human Pancreatic Cancer Patients Using H-1 NMR Spectroscopy and Principal Component Analysis. Appl Biochem Biotech 165: 148-154.

71. van der Greef J, Smilde AK (2005) Symbiosis of chemometrics and metabolomics: past, present, and future. J Chemometr 19: 376-386.

72. Zou H, Hastie T, Tibshirani R (2006) J Comput Graph Stat 15: 265-286.

73. Tibshirani R (1996) Regression shrinkage and selection via the Lasso. J Roy Stat Soc B Met 58: 267-288.

74. Zou H, Hastie T, Tibshirani R (2006) Sparse principal component analysis. J Comput Graph Stat 15: 265-286.

75. Witten DM, Tibshirani R, Hastie T (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10: 515-534.

76. Rasmussen MA, Bro R (2012) A tutorial on the Lasso approach to sparse modeling. Chemom Intell Lab Syst 119: 21-31.

77. Allen GI, Maletic-Savatic M (2011) Sparse non-negative generalized PCA with applications to metabolomics. Bioinformatics 27: 3029-3035.

78. Acar E, Gurdeniz G, Rasmussen MA, Rago D, Dragsted LO, Bro R (2012) Coupled Matrix Factorization with Sparse Factors to identify Potential Biomarkers in Metabolomics. Proceedings of the 2012 IEEE International Conference on Data Mining Workshops .

79. Wold S, Sjostrom M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst 58: 109-130.

80. Trygg J, Wold S (2002) Orthogonal projections to latent structures (O-PLS). J Chemometr 16: 119-128.

81. Kemsley EK, Tapp HS (2009) OPLS filtered data can be obtained directly from non-orthogonalized PLS1. J Chemometr 23: 263-264.

82. Westerhuis JA, Hoefsloot HCJ, Smit S, Vis DJ, Smilde AK, van Velzen EJJ, van Duijnhoven JPM, van Dorsten FA (2008) Assessment of PLSDA cross validation. Metabolomics 4: 81-89.

83. Brereton RG (2006) Consequences of sample size, variable selection, and model validation and optimisation, for predicting classification ability from analytical data. Trac-Trend Anal Chem 25: 1103-1111.

84. Anderssen E, Dyrstad K, Westad F, Martens H (2006) Reducing over-optimism in variable selection by cross-model validation. Chemom Intell Lab Syst 84: 69-74.

85. Smit S, van Breemen MJ, Hoefsloot HCJ, Smilde AK, Aerts JMFG, de Koster CG (2007) Assessing the statistical validity of proteomics based biomarkers. Anal Chim Acta 592: 210-217.

86. Szymanska E, Saccenti E, Smilde AK, Westerhuis JA (2012) Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. Metabolomics 8: 3-16.

87. Rajalahti T, Arneberg R, Berven FS, Myhr KM, Ulvik RJ, Kvalheim OM (2009) Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. Chemom Intell Lab Syst 95: 35-48.

88. Sumner L, Amberg A, Barrett D, Beale M, Beger R, Daykin C, Fan T, Fiehn O, Goodacre R, Griffin J, Hankemeier T, Hardy N, Harnly J, Higashi R, Kopka J, Lane A, Lindon J, Marriott P, Nicholls A, Reily M, Thaden J, Viant M (2007) Proposed minimum reporting standards for chemical analysis. Metabolomics 3: 211-221.

89. Kuhl C, Tautenhahn R, Bottcher C, Larson TR, Neumann S (2012) CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. Anal Chem 84: 283-289.

90. Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, Custodio DE, Abagyan R, Siuzdak G (2005) METLIN: a metabolite mass spectral database. Ther Drug Monit 27: 747-751.

91. Sud M, Fahy E, Cotter D, Brown A, Dennis EA, Glass CK, Merrill AH, Jr., Murphy RC, Raetz CR, Russell DW, Subramaniam S (2007) LMSD: LIPID MAPS structure database. Nucleic Acids Res 35: D527-D532.

92. Brown M, Dunn WB, Dobson P, Patel Y, Winder CL, Francis-McIntyre S, Begley P, Carroll K, Broadhurst D, Tseng A, Swainston N, Spasic I, Goodacre R, Kell DB (2009) Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. Analyst 134: 1322-1332.

93. Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahren D, Tsoka S, Darzentas N, Kunin V, Lopez-Bigas N (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic Acids Res 33: 6083-6089.

94. Wishart DS (2008) DrugBank and its relevance to pharmacogenomics. Pharmacogenomics 9: 1155-1162.

95. Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, Ojima Y, Tanaka K, Tanaka S, Aoshima K, Oda Y, Kakazu Y, Kusano M, Tohge T, Matsuda F, Sawada Y, Hirai MY, Nakanishi H, Ikeda K, Akimoto N, Maoka T, Takahashi H, Ara T, Sakurai N, Suzuki H, Shibata D, Neumann S, Iida T, Tanaka K, Funatsu K, Matsuura F, Soga T, Taguchi R, Saito K, Nishioka T (2010) MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom 45: 703-714.

96. Nicholson JK, Foxall PJ, Spraul M, Farrant RD, Lindon JC (1995) 750 MHz 1H and 1H-13C NMR spectroscopy of human blood plasma. Anal Chem 67: 793-811.

97. Wixon J, Kell D (2000) The Kyoto encyclopedia of genes and genomes--KEGG. Yeast 17: 48-55.

98. Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C, Walk TC, Zhang P, Karp PD (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res 36: D623-D631.

99. Ludwig C, Easton J, Lodi A, Tiziani S, Manzoor S, Southam A, Byrne J, Bishop L, He S, Arvanitis T, G++nther U, Viant M (2012) Birmingham Metabolite Library: a publicly accessible database of 1-D 1H and 2-D 1H J-resolved NMR spectra of authentic metabolite standards (BML-NMR). Metabolomics 8: 8-18.

100. Suhre, Karsten and SpringerLink (Online service) (2012) Genetics Meets Metabolomics from Experiment to Systems Biology. New York, NY: Springer New York.

101. Griffin, James E. and Ojeda, Sergio R. (2000) Textbook of endocrine physiology. Oxford England: Oxford University Press.

102. Hoppel CL, Genuth SM (1980) Carnitine Metabolism in Normal-Weight and Obese Human-Subjects During Fasting. Am J Physiol 238: E409-E415.

103. Mcgarry JD, Foster DW (1980) Regulation of Hepatic Fatty-Acid Oxidation and Ketone-Body Production. Annual Rev Biochem 49: 395-420.

104. Bondia-Pons I, Nordlund E, Mattila I, Katina K, Aura AM, Kolehmainen M, Oresic M, Mykkanen H, Poutanen K (2011) Postprandial differences in the plasma metabolome of healthy Finnish subjects after intake of a sourdough fermented endosperm rye bread versus white wheat bread. Nutr J 10: 116.

105. Basu R, Dalla MC, Campioni M, Basu A, Klee G, Toffolo G, Cobelli C, Rizza RA (2006) Effects of age and sex on postprandial glucose metabolism: differences in glucose turnover, insulin secretion, insulin action, and hepatic insulin extraction. Diabetes 55: 2001-2014.

106. Zhao X, Peter A, Fritsche J, Elcnerova M, Fritsche A, Haring HU, Schleicher ED, Xu G, Lehmann R (2009) Changes of the plasma metabolome during an oral glucose tolerance test: is there more than glucose to look at? Am J Physiol Endocrinol Metab 296: E384-E393.

107. Shaham O, Wei R, Wang TJ, Ricciardi C, Lewis GD, Vasan RS, Carr SA, Thadhani R, Gerszten RE, Mootha VK (2008) Metabolic profiling of the human response to a glucose challenge reveals distinct axes of insulin sensitivity. Mol Syst Biol 4: 214.

108. Wopereis S, Rubingh CM, van Erk MJ, Verheij ER, van VT, Cnubben NH, Smilde AK, van der Greef J, van OB, Hendriks HF (2009) Metabolic profiling of the response to an oral glucose tolerance test detects subtle metabolic changes. PLoS One 4(2): e4525.

109. Krug S, Kastenmuller G, Stuckler F, Rist MJ, Skurk T, Sailer M, Raffler J, Romisch-Margl W, Adamski J, Prehn C, Frank T, Engel KH, Hofmann T, Luy B, Zimmermann R, Moritz F, Schmitt-Kopplin P, Krumsiek J, Kremer W, Huber F, Oeh U, Theis FJ, Szymczak W, Hauner H, Suhre K, Daniel H (2012) The dynamic range of the human metabolome revealed by challenges. Faseb Journal 26: 2607-2619.

110. Brouwer IA, Wanders AJ, Katan MB (2010) Effect of Animal and Industrial Trans Fatty Acids on HDL and LDL Cholesterol Levels in Humans - A Quantitative Review. PLoS One 5.

111. Mozaffarian D, Katan MB, Ascherio A, Stampfer MJ, Willett WC (2006) Trans fatty acids and cardiovascular disease. N Engl J Med 354: 1601-1613.

112. Mensink RP, Zock PL, Kester AD, Katan MB (2003) Effects of dietary fatty acids and carbohydrates on the ratio of serum total to HDL cholesterol and on serum lipids and apolipoproteins: a meta-analysis of 60 controlled trials. Am J Clin Nutr 77: 1146-1155.

113. Mozaffarian D, Clarke R (2009) Quantitative effects on cardiovascular risk factors and coronary heart disease risk of replacing partially hydrogenated vegetable oils with other fats and oils. Eur J Clin Nutr Suppl 2: S22-S33.

114. Micha R, Mozaffarian D (2008) Trans fatty acids: Effects on cardiometabolic health and implications for policy. Prostaglandins Leukotrienes and Essential Fatty Acids 79: 147-152.

115. Bendsen NT, Stender S, Szecsi PB, Pedersen SB, Basu S, Hellgren LI, Newman JW, Larsen TM, Haugaard SB, Astrup A (2011) Effect of industrially produced trans fat on markers of systemic inflammation: evidence from a randomized trial in women. J Lipid Res 52: 1821-1828.

116. Carey VJ, Walters EE, Colditz GA, Solomon CG, Willett WC, Rosner BA, Speizer FE, Manson JE (1997) Body fat distribution and risk of non-insulin-dependent diabetes mellitus in women - The Nurses' Health Study. Am J Epidemiol 145: 614-619.

117. Salmeron J, Hu FB, Manson JE, Stampfer MJ, Colditz GA, Rimm EB, Willett WC (2001) Dietary fat intake and risk of type 2 diabetes in women. Am J Clin Nutr 73: 1019-1026.

118. Meyer KA, Kushi LH, Jacobs DR, Folsom AR (2001) Dietary fat and incidence of type 2 diabetes in older Iowa women. Diabetes Care 24: 1528-1535.

119. van Dam RM, Rimm EB, Willett WC, Stampfer MJ, Hu FB (2002) Dietary patterns and risk for type 2 diabetes mellitus in US men. Ann Intern Med 136: 201-209.

120. Bendsen NT, Haugaard SB, Larsen TM, Chabanova E, Stender S, Astrup A (2011) Effect of trans-fatty acid intake on insulin sensitivity and intramuscular lipids-a randomized trial in overweight postmenopausal women. Metabolism 60: 906-913.

121. Lemaitre RN, King IB, Raghunathan TE, Pearce RM, Weinmann S, Knopp RH, Copass MK, Cobb LA, Siscovick DS (2002) Cell membrane trans-fatty acids and the risk of primary cardiac arrest. Circulation 105: 697-701.

122. Lemaitre RN, King IB, Mozaffarian D, Sotoodehnia N, Siscovick DS (2006) Trans-fatty acids and sudden cardiac death. Atheroscler Suppl 7: 13-15.

123. Morgado N, Galleguillos A, Sanhueza J, Garrido A, Nieto S, Valenzuela A (1998) Effect of the degree of hydrogenation of dietary fish oil on the trans fatty acid content and enzymatic activity of rat hepatic microsomes. Lipids 33: 669-673.

124. Roach C, Feller SE, Ward JA, Shaikh SR, Zerouga M, Stillwell W (2004) Comparison of cis and trans fatty acid containing phosphatidylcholines on membrane properties. Biochemistry 43: 6344-6351.

125. Soni SP, Ward JA, Sen SE, Feller SE, Wassall SR (2009) Effect of Trans Unsaturation on Molecular Organization in a Phospholipid Membrane. Biochemistry 48: 11097-11107.

PAPER I

The Effect of LC-MS data preprocessing methods on the selection of plasma biomarkers in fed vs. fasted rats.

Gürdeniz G, Kristensen M, Skov T, Dragsted LO

Metabolites, (2012), 2:77-99

Metabolites **2012**, *2*, 77-99; doi:10.3390/metabo2010077

OPEN ACCESS

metabolites

ISSN 2218-1989

www.mdpi.com/journal/metabolites/

Article

The Effect of LC-MS Data Preprocessing Methods on the Selection of Plasma Biomarkers in Fed *vs.* Fasted Rats

Gözde Gürdeniz [1,*], Mette Kristensen [1], Thomas Skov [2] and Lars O. Dragsted [1]

[1] Department of Human Nutrition, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, 1958, Frederiksberg C, Denmark; E-Mails: mkri@life.ku.dk (M.K.); ldra@life.ku.dk (L.O.D.)

[2] Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, 1958, Frederiksberg C, Denmark; E-Mail: thsk@life.ku.dk

* Author to whom correspondence should be addressed; E-Mail: gozg@life.ku.dk; Tel.: +45-29-176-389; Fax: +45-35-332-483.

Received: 30 November 2011; in revised form: 6 January 2012 / Accepted: 6 January 2012 / Published: 18 January 2012

Abstract: The metabolic composition of plasma is affected by time passed since the last meal and by individual variation in metabolite clearance rates. Rat plasma in fed and fasted states was analyzed with liquid chromatography quadrupole-time-of-flight mass spectrometry (LC-QTOF) for an untargeted investigation of these metabolite patterns. The dataset was used to investigate the effect of data preprocessing on biomarker selection using three different softwares, MarkerLynx™, MZmine, XCMS along with a customized preprocessing method that performs binning of *m/z* channels followed by summation through retention time. Direct comparison of selected features representing the fed or fasted state showed large differences between the softwares. Many false positive markers were obtained from custom data preprocessing compared with dedicated softwares while MarkerLynx™ provided better coverage of markers. However, marker selection was more reliable with the gap filling (or peak finding) algorithms present in MZmine and XCMS. Further identification of the putative markers revealed that many of the differences between the markers selected were due to variations in features representing adducts or daughter ions of the same metabolites or of compounds from the same chemical subclasses, e.g., lyso-phosphatidylcholines (LPCs) and lyso-phosphatidylethanolamines (LPEs). We conclude that despite considerable differences in the performance of the preprocessing tools we could extract the same biological information by any of them. Carnitine, branched-chain amino acids, LPCs and LPEs were identified by all methods as markers of the fed state whereas acetylcarnitine was abundant during fasting in rats.

Keywords: rat plasma; biomarkers; LC-QTOF; data pre-processing; MarkerLynx; MZmine; XCMS

1. Introduction

In nutritional studies, blood samples are frequently collected in order to relate dietary conditions with metabolic markers. Blood may be obtained either in the fasted or postprandial state, depending on the hypothesis being tested. The fasting state, typically following an overnight fast, is considered to be more reproducible and can be defined as a baseline level for metabolic studies. However, imbalances in diet-dependent metabolism may not be detectable in the fasted state [1]. On the other hand, determination of the metabolic response in the extended postprandial state, which is the normal metabolic situation of human beings throughout the day, is more challenging as individual variability is high [2]. The basic metabolic rate varies roughly with surface area in mammals and an overnight fasting period in rats having an eight times higher rate of energy metabolism than humans may therefore represent a more extreme condition than overnight fasting in humans. A rat model may therefore be convenient to study the major differences between fasting and fed states, the latter defined as the state of rats following a normal *ad libitum* meal pattern. A rat model also offers full control of the food intake in the study subjects.

In this study, an untargeted metabolomics based approach to study the metabolic differences between rat plasma at fasted and fed states was performed. Metabolomics is defined as the process of monitoring and evaluating changes in metabolites during biochemical processes and has become an emerging tool to understand responses of cells and living organisms with respect to their gene expression or alterations in their lifestyles and diets of biochemical variation, during or after food intake [3].

A wide range of metabolites and other compounds can be detected in various biofluids by nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS). These approaches can be either untargeted through total data capture or highly targeted, such as measuring a large number of defined lipids. MS based instruments, with higher sensitivity compared to NMR [4,5], have become a widely used technique in metabolomics studies. Liquid chromatography (LC) coupled with time-of-flight (TOF) MS offers high resolution, reasonable sensitivity and improved data acquisition for complex sample mixture analyses. The system has served as a powerful tool in many other studies focusing on untargeted metabolic profiling of biofluids [6–8].

LC-MS analysis produces large amounts of data with complex chemical information. An important task is to arrange data in a way so that relevant information can be extracted. The complexity of LC-MS data brings out the concept of data handling which can be roughly summarized in two basic steps: data preprocessing and data analysis. *Data preprocessing* covers the methods to go from complex raw data to clean data. Raw data are comprised of retention times and mass to charge ratios of thousands of chemical compounds. Several software tools (commercial or freely available) have emerged for LC-MS data preprocessing. These tools typically include specific algorithms for the two key steps in data preprocessing, (1) peak detection and (2) alignment. Each software tool creates a list of peaks denoted by a specific mass and retention time. Each entry has a signal intensity denoting peak height or area.

Alignment corrects retention time and mass differences across samples so that a peak, considered as one chemical compound, is represented by the same mass and retention time across all samples. The peak detection and alignment result in a data table providing the detected peaks across samples which can be denoted as clean data. All of these tools aim to provide high speed, automated data preprocessing. The basic principles of the many LC-MS data preprocessing software tools have recently been summarized [9,10]. To be able to obtain high efficiency in data preprocessing, the software tool employed should have the parameter settings required to match the structure of the specific dataset.

Existence of various data preprocessing tools brings out concerns about what are the characteristics of the software tools and what are the pros and cons of their algorithms. There are some studies attempting to define quality parameters for comparison of peak detection [11,12] or alignment [13] algorithms of different data preprocessing tools, but a direct comparison of the overall performance of the most commonly used data preprocessing tools has not so far been attempted. The question to be addressed in this study is whether there is agreement between the biological information as represented by the biomarkers extracted by preprocessing the same dataset with different data preprocessing methods. Therefore we compare here the potential biomarkers extracted from the current small dataset using four different softwares for preprocessing; (1) MarkerLynx^{TM} (MassLynx (Waters, Milfold, MA, USA)); (2) MZmine [14]; (3) XCMS [15,16] and (4) a customized method that is implemented in MATLAB (The Mathworks, Inc., MA, USA). MarkerLynx^{TM} is a commercial software whereas XCMS and MZmine are freely available software tools. The customized method included *m/z* binning and retention time collapsing which can be considered as a more old-fashioned method for LC-MS data preprocessing. The applicability of this method for LC-MS data has been evaluated in other studies [17] but an extensive comparison with other approaches has not been published previously.

Thus, in this study the UPLC-QTOF profiles of rat plasma collected in the fasted and fed states were analyzed for two different purposes: (1) to investigate the effect of different data preprocessing tools on biomarker selection; and (2) to interpret the biology behind the biomarkers identified for the two states.

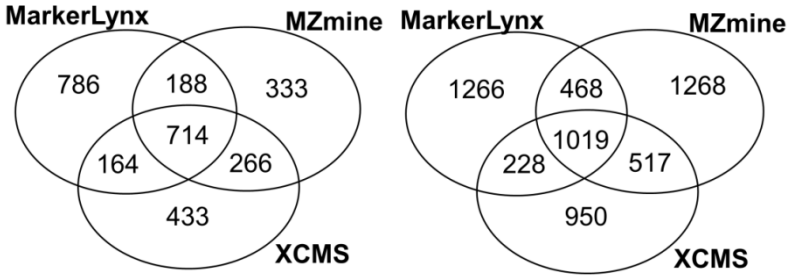
2. Results and Discussion

2.1. Comparison of Data Preprocessing Methods

The number of features obtained from each preprocessing method is given in the Supplementary information 3. We succeeded in extracting a similar number of features with optimized parameter settings (positive or negative), except for the custom method in negative mode where we have an approximate doubling compared with the other software tools.

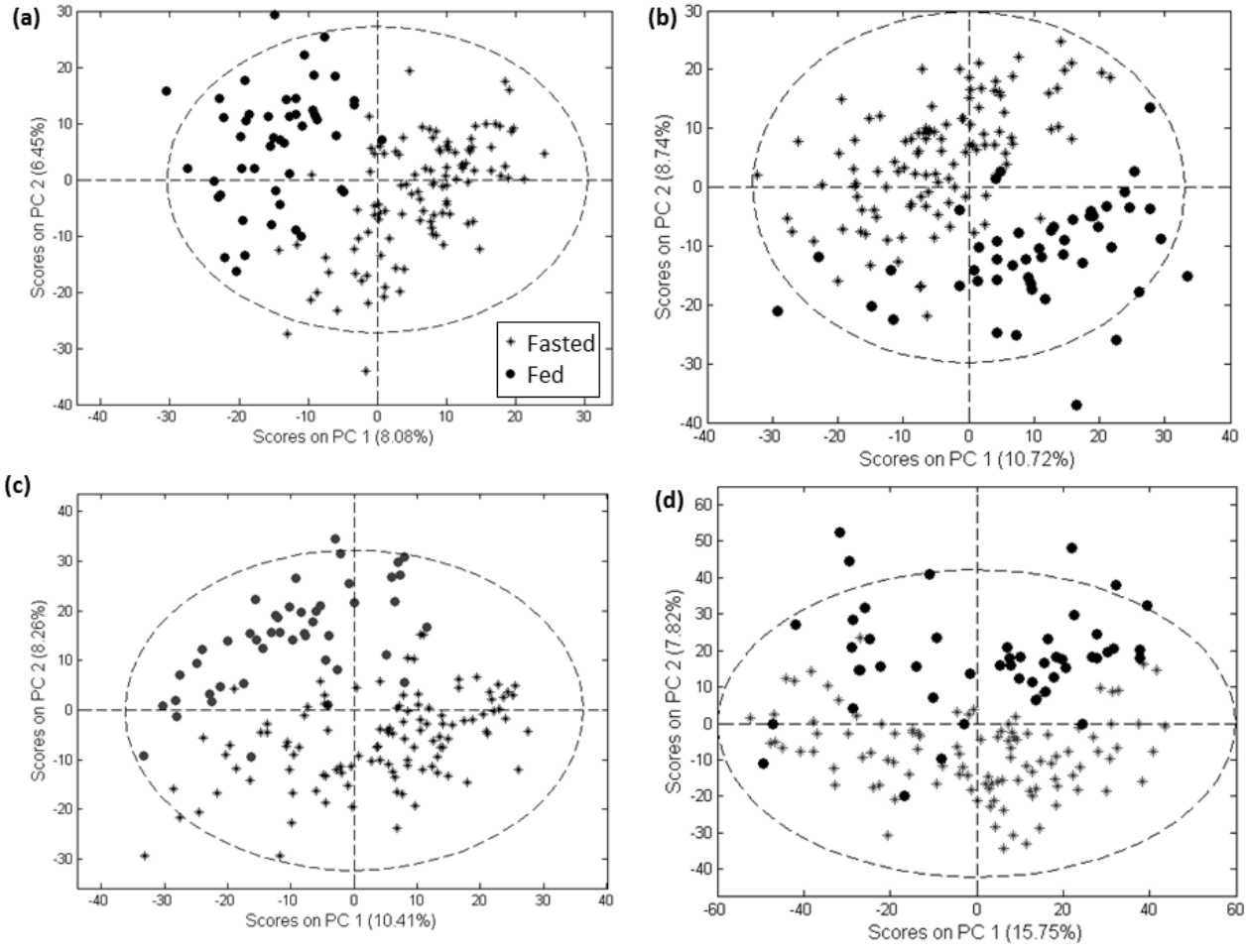
Primarily, common and unique extracted features from three different softwares were illustrated in Figure 1. We found 37%–46% of the features extracted by each software to be in common. Rauf *et al.* [16] found higher number for common features (46%–52%) from leaf and seed extracts comparing MZmine and XCMS (centWave) peak detection algorithms. The difference can be the result of more complex nature of plasma samples compared to plant extracts.

Figure 1. Venn diagrams illustrating the number of common and method specific features extracted from three software tools (right: positive mode; left: negative mode).



All three software tools and the customized method employed here were able to produce a feature set showing substantial separation of samples from the fasted and fed states in a PCA scores plot (Figure 2 for negative mode data and Supplementary information 5 for positive mode data).

Figure 2. PCA scores plots of negative mode data processed with MarkerLynx (**a**), MZmine (**b**), XCMS (**c**) and customized methods (**d**).



PLSDA model of each preprocessed data on independent test sets provided an average classification error rate of 0–0.02 (Supplementary information 6) indicating that all models resulted in good classification performance. The classification error rates were very similar for datasets obtained from

different data preprocessing methods. On the other hand, the average classification error rates of datasets where classes were permuted were calculated as 0.49–0.51, corresponding to misclassification of half of the samples, which is an expected value for permuted data [18]. None of the 2,000 permutations had classification error lower than 0.00–0.02, indicating original fasted *vs.* fed discrimination was significant. Histograms of permutation test are given in Supplementary Information 7.

As previously mentioned, autoscaling is applied in this study to detect possible variation between two states for any feature, regardless of its concentration. Nevertheless, autoscaling complicates variable selection as it gives the same chance to all peaks to influence the PLSDA model, and the decision of a regression coefficient cut-off value for selection of important features becomes difficult. Hereby, we decided to select of 25 features only but there is no proof to say the feature with 26th highest regression coefficient was not a potential biomarker. Thus, the 25 markers from each method and their various ranks from other softwares were included in Tables 1 and 2 for the negative and positive modes, respectively. While there is no way to say which software is the more correct, the consequence of the differences observed here is that there is no basis for putting too much emphasis on the rank in PLS-DA methods. Howeverin many metabolomics studies, PLS-DA regression coefficients or VIP cut-offs have commonly been employed for marker selection, even without the rigorous iteration used here.

Table 1. Retention times and measured masses of the markers obtained from MarkerLynx, MZmine, XCMS and custom data processing of negative mode data that contributed most to the separation of samples in fasted and fed states.

NO	RT (min)	Measured *m/z*	MX Rank	MZ Rank	XCMS Rank	Custom rank	Group	Suggested Compound	Adduct	Monoisotopic mass
1	0.64	105.02	57	17	14	194	fed	U1		
2	0.82	116.07	91	26	17	507	fed	U2		
3	1.15	180.06	67	28	21	27	fed	U3		
4	1.15	383.12	40	80	25	624	fed	U3		
5	1.36	59.01	21	34	9	7	fasted	3-hydroxybutanoic acid F		104.0473
6	1.36	260.00	49	68	nd	22	fasted	3-hydroxybutanoic acid F		104.0473
7	1.37	229.07	20	35	nd	72	fasted	3-hydroxybutanoic acid A	[2M+Na-H]	104.0473
8	1.37	103.04	39	15	nd	20	fasted	3-hydroxybutanoic acid	[M-H]	104.0473
9	1.37	261.18	1424	nd	18	14	fed	Isoleucine	[2M-H]	131.0946
10	1.37	130.09	25	nd	24	65	fed	Isoleucine	[M-H]-	131.0946
11	1.80	178.05	nd	22	nd	166	fed	U4		

Table 1. *Cont.*

NO	RT (min)	Measured *m/z*	MX Rank	MZ Rank	XCMS Rank	Custom rank	Group	Suggested Compound	Suggested Adduct	Monoisotopic mass
12	1.88	134.06	14	9	6	40	fasted	Hippuric acid * F		179.0582
13	1.88	178.05	15	7	4	116	fasted	Hippuric acid *	[M-H]	179.0582
14	2.02	344.10	383	nd	222	12	none	U5		
15	2.46	365.07	3	6	nd	43	fed	U6		
16	2.46	623.36	8	nd	3	94	fed	U6		
17	2.46	343.08	2	2	1	6	fed	U6		
18	2.47	623.87	4	nd	nd	16	fed	U7		
19	3.00	185.12	793	23	77	284	fed	U8		
20	3.50	505.30	1833	nd	nd	10	none	U9		
21	4.11	586.31	nd	13	nd	13	fed	LPC(20:5)	[M+FA-H]	541.3168
22	4.12	309.20	1	10	7	1802	fed	LPC(20:5) F		541.3168
23	4.15	452.28	22	30	22	1006	fed	LPC(14:0) F		467.3012
24	4.16	512.30	17	21	19	45	fed	LPC(14:0) A	[M+FA-H]	467.3012
25	4.16	979.60	19	nd	nd	33	fed	LPC(14:0) A	[2M+FA-H]	467.3012
26	4.17	502.29	13	11	nd	25	fed	LPC(18:3) F		517.3168
27	4.18	562.31	5	8	51	17	fed	LPC(18:3)	[M+FA-H]	517.3168
28	4.18	818.50	16	nd	nd	1672	fed	U10		
29	4.18	526.30	11	19	11	912	fed	LPC(20:5) F		541.3168
30	4.19	586.31	7	18	8	13	fed	LPC(20:5)	[M+FA-H]	541.3168
31	4.23	563.32	nd	nd	13	15	fed	U11		
32	4.34	476.28	23	1	nd	1	fed	2-acyl LPC(18:2) F		519.3325
33	4.35	564.33	10	12	nd	3	fed	2-acyl LPC(18:2)	[M+FA-H]	519.3325
34	4.35	504.31	147	3	nd	2	fed	2-acyl LPC(18:2) F		519.3325
35	4.35	578.30	nd	5	nd	35	fasted	U12		
36	4.36	632.33	120	25	nd	113	fed	U13		
37	4.38	281.25	33	nd	15	nd	fasted	U14		
38	4.43	476.28	105	4	2	1	fed	1-acyl LPC(18:2) F		519.3325
39	4.44	168.35	6	nd	nd	1512	fed	1-acyl LPC(18:2) F		519.3325
40	4.44	995.59	60	nd	nd	4	fed	1-acyl LPC(18:2) F		519.3325
41	4.44	168.63	18	nd	nd	170	fed	1-acyl LPC(18:2) F		519.3325
42	4.44	504.31	65	14	32	2	fed	1-acyl LPC(18:2) F		519.3325
43	4.45	457.10	12	nd	561	2332	fasted	U15		
44	4.45	564.33	32	31	20	3	fed	1-acyl LPC(18:2)	[M+FA-H]	519.3325
45	4.45	335.40	nd	nd	nd	8	none	none		
46	4.45	335.70	nd	nd	nd	9	none	none		
47	4.45	477.28	nd	nd	nd	21	fed	1-acyl LPC(18:2) **iso1**		
48	4.45	564.10	nd	nd	nd	23	none	none		
49	4.45	565.34	nd	nd	nd	5	fed	1-acyl LPC(18:2) iso2		
50	4.45	587.30	nd	nd	nd	11	none	none		

Table 1. *Cont.*

NO	RT (min)	Measured *m/z*	MX Rank	MZ Rank	XCMS Rank	Custom rank	Group	Suggested Compound	Suggested Adduct	Monoisotopic mass
51	4.45	996.59	nd	nd	nd	19	fed	1-acyl LPC(18:2) iso3		
52	4.50	552.33	24	46	63	320	fed	U16		
53	4.62	452.28	48	55	23	1006	fasted	U17		
54	4.65	566.35	374	24	nd	138	fed	1-acyl LPC(18:1)	[M+FA-H]	521.3481
55	4.73	478.29	9	16	12	18	fed	LPE(18:1) *	[M-H]	479.3012
56	4.88	445.33	76	20	10	1206	fasted	U19		
57	5.14	277.22	85	106	5	98	fasted	Gamma-Linolenic acid *	[M-H]	278.2246
58	5.22	338.30	100	nd	nd	24	none	U20		
59	5.38	279.23	145	nd	16	177	fasted	Linoleic acid *	[M-H]	280.2402

MX: MarkerLynx; MZ: MZmine; 'U', Unidentified compound; A: Adduct; F: Fragment *, identity confirmed with authentic standards; 'nd', not detected by the software peak-finding algorithm.

Table 2. Retention times and measured masses of the markers obtained from MarkerLynx, MZmine, XCMS and custom data processing of positive mode data that contributed most to the separation of samples in fasted and fed states.

NO	RT (min)	Measured *m/z*	MX Rank	MZ Rank	XCMS Rank	Custom rank	Group	Suggested Compound	Suggested Adduct	Monoisotopic mass
1	0.53	112.11	nd	12	13	301	fasted	U1		
2	0.57	730.70	276	nd	nd	25	fasted	U2		
3	0.61	103.04	46	nd	19	2901	fed	L-Carnitine *F		161.1052
4	0.61	102.09	1368	nd	21	481	fed	L-Carnitine *F		161.1052
5	0.61	162.11	31	41	10	10	fed	L-Carnitine *	[M+H]	161.1052
6	0.66	70.07	12	11	25	22	fed	D-proline *F		115.0633
7	0.66	116.07	13	14	12	11	fed	D-proline *	[M+H]	115.0633
8	0.86	130.09	24	521	44	838	fasted	U3		
9	0.90	144.10	23	nd	16	455	fasted	L-Acetylcarnitine*F		203.1158
10	0.90	204.12	28	18	6	8	fasted	L-Acetylcarnitine*	[M+H]	203.1158
11	0.90	145.05	21	13	11	41	fasted	L-Acetylcarnitine*F		203.1158
12	1.17	248.15	49	23	7	38	fasted	U4		
13	1.64	231.12	nd	100	1	649	fasted	U5		
14	1.90	105.03	1	17	2	78	fasted	Hippuric Acid*F		179.0582
15	1.90	77.04	3	19	3	578	fasted	Hippuric Acid*F		179.0582
16	2.23	316.21	19	46	nd	179	fasted	U6		
17	2.42	899.43	nd	nd	nd	17	fed	U7		
18	2.42	287.20	nd	nd	nd	1	fed	U7		
19	2.42	286.20	7	3	50	4	fed	U7		

Table 2. *Cont.*

NO	RT (min)	Measured *m/z*	MX Rank	MZ Rank	XCMS Rank	Custom rank	Group	Suggested Compound	Suggested Adduct	Monoisotopic mass
20	3.42	536.34	35	nd	nd	24	fed	U8		
21	3.49	158.16	338	222	63	19	fasted	U9		
22	4.11	542.33	16	16	nd	21	fed	LPC(20:5)	[M+H]	541.3168
23	4.12	564.31	nd	15	nd	43	fed	LPC(20:5) A	[M+Na]	541.3168
24	4.16	312.03	151	nd	17	2659	fed	U10		
25	4.16	468.31	20	24	23	15	fed	LPC(14:0)	[M+H]	467.3012
26	4.19	540.31	25	64	nd	47	fed	LPC(18:3) A	[M+Na]	517.3168
27	4.19	518.33	15	6	81	62	fed	LPC(18:3)	[M+H]	517.3168
28	4.23	445.40	nd	nd	nd	12	fasted	octadecanoylcarnitine^{Iso}		
29	4.23	444.37	18	33	47	33	fasted	octadecanoylcarnitine		
30	4.35	337.28	9	9	5	57	fed	2-acyl LPC(18:2) F		519.3325
31	4.35	520.34	6	1	nd	2	fed	2-acyl LPC(18:2)	[M+H]	519.3325
32	4.36	542.33	4	2	nd	21	fed	2-acyl LPC (18:2) A	[M+Na]	519.3325
33	4.36	819.96	22	nd	nd	950	fed	U11		
34	4.36	502.33	nd	10	nd	28	fed	2-acyl LPC(18:2) F	[M+Na]	479.3376
35	4.42	566.32	1024	2058	15	50	fasted	U12		
36	4.42	844.47	219	233	20	1312	fasted	U13		
37	4.44	519.90	nd	nd	nd	18	fed	U14		
38	4.44	521.35	nd	nd	nd	5	fed	1-acyl LPC(18:2) ^{Iso1}	[M+H]	519.3325
39	4.45	523.35	nd	7	nd	89	fed	1-acyl LPC(18:2)^{Iso2}	[M+H]	519.3325
40	4.45	519.70	316	nd	nd	7	fed	U15		
41	4.45	997.64	14	20	9	3	fed	1-acyl LPC(18:2) A		519.3325
42	4.45	819.97	2	21	835	950	fasted	U16		
43	4.45	520.34	8	4	18	2	fed	1-acyl LPC(18:2)	[M+H]	519.3325
44	4.45	998.64	30	nd	nd	6	fed	U17		
45	4.45	460.29	59	54	14	612	fed	1-acyl LPC(18:2) F		519.3325
46	4.45	520.10	nd	nd	nd	13	none	U18		
47	4.45	520.90	nd	nd	nd	23	none	U18		
48	4.45	521.55	nd	nd	nd	20	none	U18		
49	4.45	521.80	nd	nd	nd	16	none	U18		
50	4.45	807.97	5	8	4	2664	fed	U19		
51	4.63	949.64	34	25	48	85	fasted	U20		
52	4.64	454.30	32	22	22	1425	fasted	U20		
53	4.65	975.70	76	nd	nd	14	fed	U21		
54	4.65	522.36	10	nd	nd	70	fed	2-acyl LPC(18:1) *	[M+H]	521.3481
55	4.65	339.29	17	5	8	573	fed	2-acyl LPC(18:1) *F		
56	4.68	520.34	11	nd	24	2	fed	U22	[M+H]	519.3325

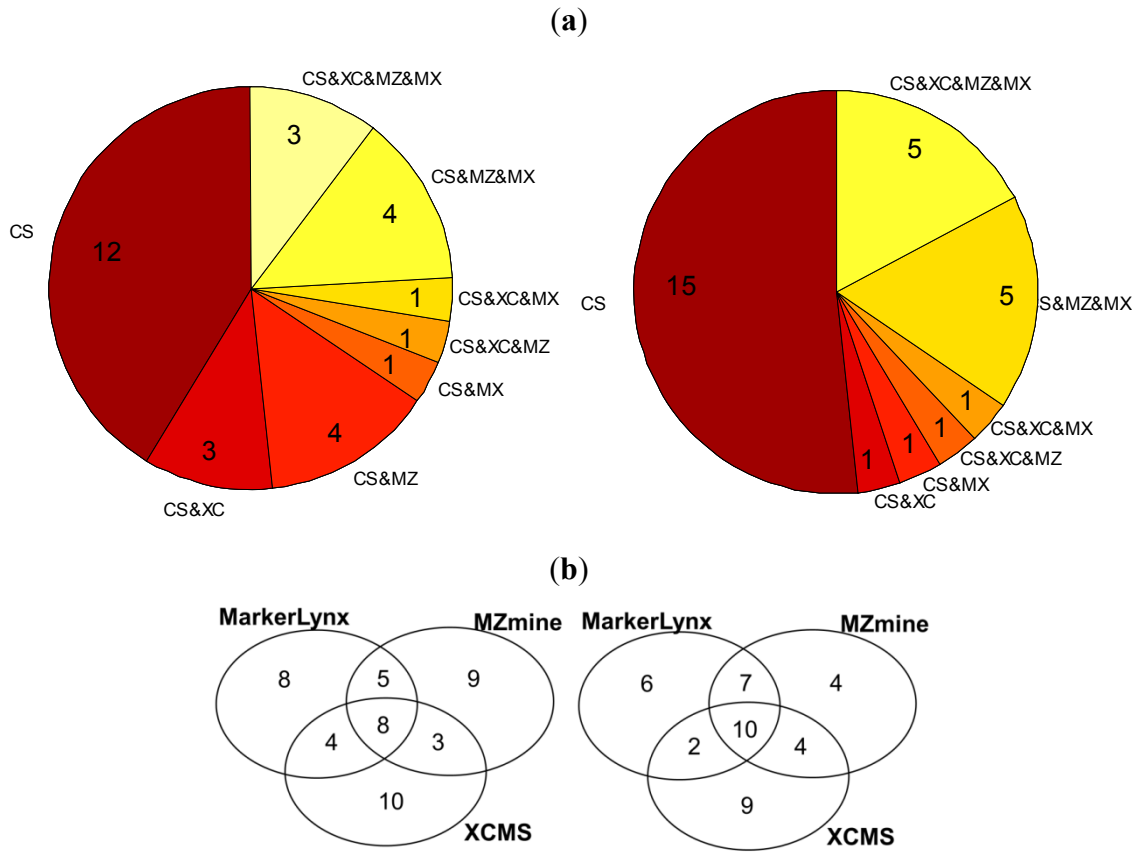
MX: MarkerLynx; MZ: MZmine; 'U', Unidentified compound; A: Adduct; F: Fragment *, identity confirmed with authentic standards; 'nd', not detected by the software peak-finding algorithm.

2.2. Custom Method vs. *Software Tools*

The algorithm of the custom preprocessing method differs from the others by not having any peak detection and alignment steps. It can therefore be considered as more independent, albeit more primitive and simple.

We compared first the *m/z* bins selected by the custom method with the markers from the three dedicated softwares (Figure 3a). Out of 25, only five of them were common for all data preprocessing tools in the positive mode and three in the negative. On the other hand, 48% (positive mode) and 58% (negative mode) of the *m/z* bins were identified also as markers by at least one of the software tools.

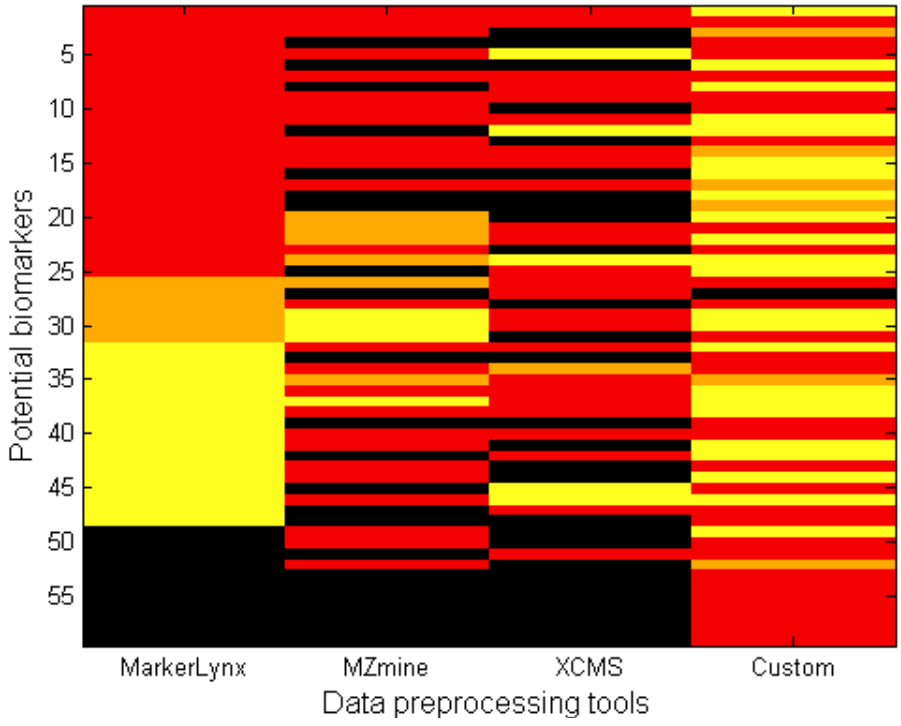
Figure 3. (**a**) Pie chart illustrating the number of custom data preprocessing markers that are unique and that are detected as markers by the other software tools (CS:Custom, MZ:MZmine, XC:XCMS, MX:Markerlynx); (**b**) Venn diagrams illustrating the number of common and method specific markers extracted from three software tools (right: positive mode; left: negative mode).



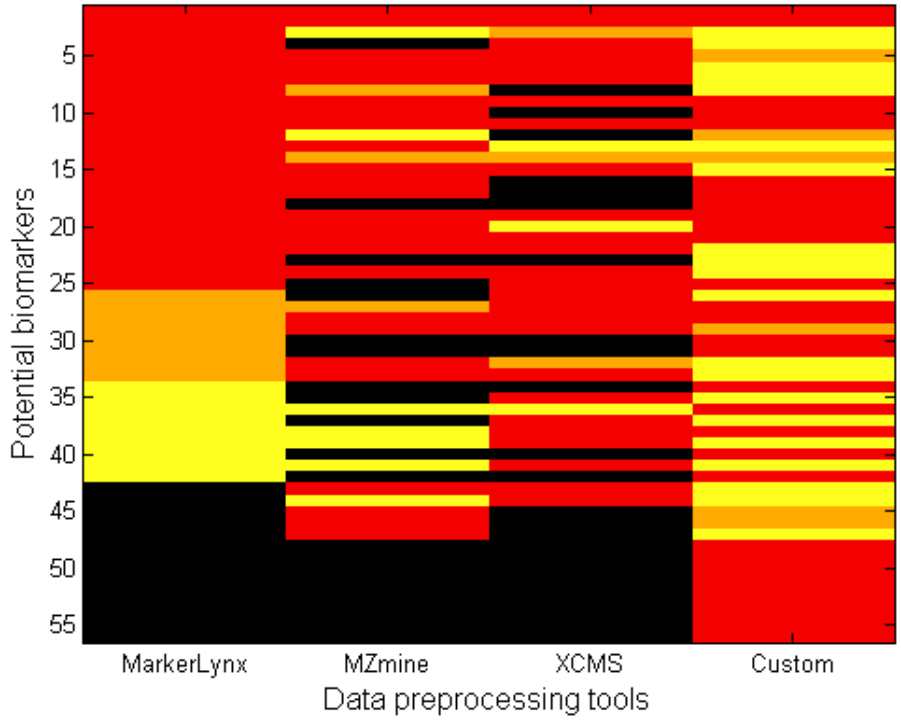
Another perspective in the comparison of different data preprocessing methods is illustrated in Figure 4 where, each row represents the rank of one marker from Table 1 (columns 4–7) for all four different data preprocessing methods. The first impression from this figure may be that the number of black regions (undetected peaks) might seem alarmingly high for some of the methods. It is important here to state that the custom data preprocessing leads to a number of false positives. The major causes of false positives are splitting of analytes into two adjacent bins or chromatographic collapsing.

Figure 4. Heatmap comparing the importance of each marker based on four different data preprocessing tools for (**a**) negative and (**b**) positive mode data. Each row represents the lowest value rank of a metabolite for four different methods (Table 1, 3rd column). The markers were sorted in ascending order based on the rank obtained with MarkerLynx (red: rank 1–25; orange: rank 26–50; yellow: rank > 50; black: not detected).

(**a**)



(**b**)



An additional point from Figure 4 is the large area of yellow regions for the custom method, which presents markers detected with higher than 50 as rank in PLS-DA. This is explained mainly by retention time collapse causing peaks to be added with other peaks having the same mass but different retention time. For instance, the chromatogram of *m/z* bin = 820 as illustrated in Supplementary information 8 includes two peaks. The sample track signals of the peak at retention time = 4.32–4.38 (No. = 33, Table 2) is higher in the fed state while the peak with retention time = 4.4–4.48 38 (No. = 42, Table 2) is higher in the fasted state, indicating that they are actually markers. As these two different peaks are in the same *m/z* bin, the retention times collapsing leads to the loss of these markers. In other cases a small peak representing a marker is added with a larger one without marker characteristics thereby diluting the effect so that the bin escapes selection.

2.3. Comparison of the Dedicated Software Tools

Further comparison of the 25 markers for positive and negative mode data from each of the three dedicated software tools is illustrated in two Venn diagrams (Figure 3b). In general these three tools seem to have 8–10 markers in common among the selected 25 markers detected in the negative and positive mode (Figure 3b). There is a trend towards a larger difference between XCMS and each of the other methods in the pairwise comparisons. So all of the data preprocessing methods seem to miss out potentially important markers observed to be ranked among the top-25 markers by the other methods. In fact, only 8–10 markers would be observed to be in common if three different research groups were to investigate the same biological phenomenon using different softwares for data preprocessing, provided they had recorded similar LC-MS data. There are three possible explanations of the differences between detected markers:

(1) The marker is not included in the feature list of the other softwares. The potential cause is differences between peak detection algorithms. The number of detected features is different as shown in Figure 1. This condition is illustrated by Figure 4 as black regions.

(2) The marker is detected but the peak height assignment was not the same among software tools, which did not result in significant difference between fasted and fed states. One reason of this is shown in the next section as influence of gap filling. This condition is illustrated as yellow in Figure 4.

(3) The data analysis method affected the marker selection. This was discussed as an effect of autoscaling previously. This condition is illustrated by orange in Figure 4.

Additional differences might be caused by optimization of parameter settings and other factors from the metabolomics experiment. The loss of information and potential introduction of noise from feature selection by a single preprocessing method would therefore seem to be a potential source of error in metabolomics.

2.4. The Influence of Gap Filling

An important drawback for MarkerLynx^{TM} is that it does not contain any gap filling algorithm resulting in many zero values in the final extracted feature set. Zeros may obscure the later data analysis step and may result in incorrect grouping of ‘effect markers’ and ‘exposure markers’, because ‘true’ zeros as well as smaller and larger peaks missed by the algorithm are given the same zero value [19]. Consequences of this lacking gap filling algorithm is illustrated with two real cases. In the first case,

MarkerLynx™ algorithm records the signal of some samples from the group with a lower signal as zero, thereby increasing the differences between groups and the chance that the feature is selected as a marker. For instance, marker number 42 (Table 2) has rank 2 for MarkerLynx™ whereas it came out with higher ranks by the others (Supporting Information 6) due to this phenomenon. In the second case, the signal of some samples recorded as zero while those samples belong to the group with higher signal. In this way, the true difference between the two groups was deflated and those markers had higher rank number (lower importance) with MarkerLynx™. Many examples (Supplementary information 9) of this situation is observed, particularly in the negative mode data where the signal intensity is generally lower, thereby explaining the large yellow region for MarkerLynx™ in Figure 4a.

Another observation particularly in Figure 4a is that MarkerLynx™ has fewer black regions, meaning very few undetected peaks and several markers that are detected by MarkerLynx™ but not by the other two softwares. Since the total number of features obtained from preprocessing the data was similar for all three softwares, one possible explanation could be the differences in the filtering step. The 80% rule applied to the MarkerLynx™ dataset differs from that of the others by retaining features with many non-zero observations in at least one sample group. The filtering algorithm of MZmine does not allow the user to define the filter for each sample group. By filtering away features with many zeros, there is a risk of removing perfect markers that appear only in one of the sample groups. Therefore the filter has to be set to no more than 80% of the number of observations in the smallest sample group in order to be equivalent to the 80% rule. Another possible reason could be the differences between the peak detection algorithms. MarkerLynx™ provides an automated peak detection algorithm whereas many parameters are user-defined for the others. Although we optimized the selection of parameters carefully by testing several settings, we cannot rule out that better overlap could have been obtained with a different parameter set.

2.5. Software Preprocessing Settings

The number of detected peaks depends very much on the data preprocessing settings of each software algorithm. Although we attempted to attain the largest possible similarity in the preprocessing parameters of MarkerLynx™, MZmine and XCMS, we were aware that it is not possible to obtain exactly the same results, since each method is based on different algorithms. To illustrate this point, we preprocessed the data with MZmine using less conservative settings for many peak detection parameters and constructed the heat map again, leading to a new pattern much more similar to XCMS (figure not shown). So, in reality, it may be possible to obtain similar patterns, at least with MZmine and XCMS where gap filling is available, depending on their individual parameter settings.

In this study the contrasts between the fasted and fed states were very clear, whereas such strong contrasts may not be seen in many other metabolomics studies. Improper settings of data preprocessing parameters may therefore obscure the extraction of relevant information, and several settings and/or softwares should be applied. Proper settings are based on careful inspection of raw data as well as insight into the functionalities of software parameters. It could seem like an appealing option to allow a much larger number of peaks by being less conservative with many peak detection parameters. However, the consequence of detecting many peaks will be the inclusion of more noise and will complicate not only the alignment but also the data analysis step for the detection of biomarkers.

MarkerLynx™ and MZmine are both user friendly tools for users who do not want to go into R, MATLAB, or similar programming tools. Preprocessing data with MarkerLynx™ requires just a few user-defined settings. However the software does not provide any possibility for checking the success of any data preprocessing step. In comparison, MZmine provides a powerful visualization side that can be considered as quite useful for tuning the settings. Algorithms for visualization of peak detection results are also included in the XCMS package in R.

2.6. Biomarker Patterns

Three patterns are immediately visible for markers of the fed state in Tables 1 and 2. The first of these is the presence of sets of isomers having very similar masses but slightly different retention times, indicating that some specific groups of isomers are typical markers. The slight mass difference may be attributed to the mass accuracy of the instrument. Some examples are clusters at 512.29, 478.29 and 590.35 in the negative mode, and at 468.32, 520.34, and 522.36 in the positive mode. In many cases the earlier eluting isomeric form was not detected in the XCMS preprocessed dataset, possibly because they are much smaller peaks. Considering the parameters set while preprocessing the data with XCMS (Supplementary Information 10), additional filtering or a too high *bw* parameter (for setting the RT shift) might be the cause of not detecting those peaks. Furthermore, these patterns are always spotted with the custom data preprocessing as they were included into the same *m/z* bin, thereby intensifying their relative importance. As can be seen from Tables 1 and 2, the possible isomers were therefore given the same rank for the custom data preprocessing.

Another pattern in the marker sets is the presence of peaks with mass differences corresponding to 2 or 4 hydrogen atoms but with different retention times. These pairs are observed in both modes (e.g., 476/478, 562/564/566 in the negative mode, and 506/508 or 520/522 in the positive, Tables 1 and 2). These clusters and patterns are all observed for compounds with retention times in the same (unpolar) range pointing towards a series of lipids with varying levels of saturation (2 for each double bond).Similar patterns can also be observed for changes in chain lengths (+26 for adding –CH=CH–) as the underlying biomarkers.

Pattern recognition therefore identified lipids as potential discriminative markers between plasma samples collected at fasted and fed states. This confirms an expected finding and further identification of some of the lipids as well as some of the more polar peaks was therefore perused.

2.7. Biomarkers of Fasted and Fed State

Most of the masses belonging to the lipid-related patterns and clusters in the positive mode fit with the masses expected for positively charged lysophosphatidylcholines (LPCs) of varying chain lengths and degrees of saturation. LPC is a plasma lipid that has been recognized as an important cell signaling molecule and it is produced by the action of phospholipases A1 and A2, by endothelial lipase or by lecithin-cholesterol acyltransferase (LCA).LCA has a well-known function in catalyzing the transfer of fatty acids from phosphatidylcholine to free cholesterol in plasma for the formation of cholesteryl esters [20]. In the rat, the LPCs with more saturated acids are formed mainly in the plasma whereas unsaturated LPC is formed from PCs in the liver. We observe here a mixture of both saturated and unsaturated LPCs, indicating that the source may be dual. The cytolytic and pro-inflammatory effects

of LPCs are well-known so their level is closely regulated. However, in blood plasma the LPCs form complexes with albumin and lipoproteins, especially LDL, and are therefore not as likely to cause direct cell injury [21]. Another action of LPCs seems to be related to increased insulin resistance [22]. A slow clearance of postprandial lipids is known to be a risk factor for diabetes but the LPCs might be a lipid fraction contributing more strongly to this action. It is interesting in this context to note that Kim *et al.* identified LPCs as the major discriminative compounds of plasma species separating fasting plasma from obese/overweight and lean men [7]. They reported lower levels of saturated LPCs and higher level of unsaturated LPCs in the plasma of lean as compared to obese or overweight men. We found a similar profile here in lean rats. The unsaturated LPCs have also been found to pass the blood-brain barrier and to be important vehicles for delivering unsaturated lipids to the brain [23]. We speculate that the high level of unsaturated LPCs in the postprandial state of healthy individuals might be part of the satiety signaling system which is malfunctioning in obesity.

The LPCs appear usually in two isomeric forms, as 1-acyl or 2-acyl LPCs. The true separation of isomeric groups of LPC(18:1) in a fed state plasma sample is illustrated in Supplementary information 11. These isomers were unstable and spontaneously isomerized positionally, as also recognized in 1-acyl authentic standards of LPC and LPE(18:1), where 9% of the authentic standard was detected as the peak belonging to the 2-acyl form. For the confirmation of the 2-acyl LPC form, standards of PC and PE(16:0/18:1) were hydrolyzed by phospholipase A1. In addition to the 2-acyl LPC and LPE(18:1) we observed that 7% of the acyl group had spontaneously migrated to the 1-acyl position (Supplementary information 11). Croset *et al.* studied the significance of positional acyl isomers of unsaturated LPCs in blood [24]. They concluded that 50% of PUFA was located at the 2-acyl position where they are available for tissue uptake, and that they can be re-acylated at the 1-acyl position to form membrane phospholipids.

With the applied methodology we would only be able to extract the more polar lipids and detect lipids with *m/z* below 1,000 daltons. Therefore, we cannot conclude here that the LPCs, LPEs and free fatty acids are the major discriminative lipid species. Lipidomics studies have previously reported less polar lipid classes which may have *m/z* above 1,000 daltons, such as PCs, sphingomyelins and triacylglycerols as potentially reflecting the time since last meal [25,26]. With our current method, we were able to identify PCs but they were not discriminative in this study, possibly due to incomplete extraction.

A group of carnitine based compounds was also detected as markers in the positive mode data. The main function of carnitine is to assist the transport and metabolism of fatty acids in mitochondria, where they are oxidized as a major source of energy [27]. In the plasma samples from the fasting state, the level of L-carnitine was found to be lower whereas acetyl-L-carnitine was higher. During fasting an elevated concentration of acetyl coenzyme A favors the production of acetyl-L-carnitine and the ketone body, 3-hydroxybutanic acid [28], and these were identified as characteristic markers for the fasting state.

Two of the amino acids, isoleucine and proline, were found to be strongly discriminating between the fed and fasted states. Isoleucine belongs to the group of branched-chain amino acids which have been implicated in altered protein catabolism, insulin resistance and obesity [29,30]. However, leucine may have contributed to the signal since separation by our current UPLC-method was not efficient. It

seems therefore that isoleucine, and possibly other specific amino acids, may be markers of recent food ingestion and decrease with fasting.

Many adduct or daughter ions were also observed among our markers as shown in Tables 1 and 2. In many cases, different adducts or fragment ions of the same metabolite may emerge with a higher or lower rank than the parent ion, and this is an important cause of differences in the ranking orders between the preprocessing softwares. So at the metabolite level, the differences between the preprocessing methods are actually much smaller. To illustrate the higher concordance at the metabolite level, we established a new rank for each metabolite (giving each metabolite the lowest rank value from among its representative adducts, fragments or isomers). The unidentified features were considered as representing the same metabolite as long as they are within the range of 0.02 min retention time window. The metabolite ranks of different methods are represented in Supplementary information 12, which illustrates that the rank patterns were much more similar between different methods at the metabolite level than at the feature level (Figure 4). Thus, it seems reasonable to conclude that different data preprocessing methods employed in this study provide around 50% common markers, but the agreement is actually much higher at the metabolite level since different markers (adducts or fragment ions) selected from the different preprocessing softwares represent the same metabolites.

The observation that all these related ions come up with low rank numbers, *i.e.*, high importance, and that their low ranks are shared between positive and negative modes as in this study strengthens not only the confidence in the identification step but also in our variable selection method.

3. Experimental Section

3.1. Animal Study and Sample Collection

Eighty male Fisher 344 rats (4 weeks old) were obtained from Charles River (Sulzfeld, Germany). The animals had a one week run-in period to adapt to the standardized diet. The rats were subsequently randomized into five groups of 16 rats, each with equal total body weights and then fed five different diets which were all nutritionally balanced to give exactly the same amounts of all important macro- and micronutrients [31]. After 16 weeks, all rats were sacrificed by decapitation after CO_2/O_2 anesthesia. Before sacrifice, 56 of the animals had fasted for 12 h and 24 of the animals were given access to food up until termination. Blood samples were collected immediately after sacrifice directly from the *vena jugularis* into a heparin coated funnel drained into 4 mL vials containing heparin as an anticoagulant. The blood was centrifuged at 3,000 g, 4°C for 10 min. The plasma fraction was aliquoted into 2 mL cryotubes and stored at −80°C until further processing. The animal experiment was carried out under the supervision of the Danish National Agency for Protection of Experimental Animals.

3.2. Plasma Preprocessing and LC-QTOF Analysis

Removal of plasma proteins was performed before LC-MS analysis of the plasma metabolites. The plasma samples were thawed on ice and 40 µL of each sample was added into a 96-well Sirocco™ plasma protein filtering plate (#186002448, Waters) containing 180 µL of 90% methanol 0.1% formic acid solution, and the plates were vortexed for 5 min to extract metabolites from the plasma protein

precipitate. A 96-well plate for the ultra-performance liquid chromatograms UPLC autosampler (Waters, cat # 186002481) was placed underneath the protein filtering plate and vacuum was applied to the plates (using a manifold) whereby the rubber wells in the Sirocco™ plates opened and the crash solvent including metabolites dripped into the 96-well UPLC plate. When the filtering plates were dry, 180 µL of a 20:80 acetone/acetonitrile solution containing 0.1% formic acid was added to each well to further extract metabolites from the precipitated protein and vacuum was connected until dryness. The solvent was evaporated from the UPLC plates by using a cooled vacuum centrifuge and the dry samples were redissolved in 200 µL milliQ acidic water before analysis. A blank sample (0.1% formic acid) and a standard sample containing 40 different physiological compounds (metabolomics standard) was also added to spare wells to evaluate possible contamination and/or loss of metabolites in the filtering procedure.

Each sample (10 µL) was injected into the UPLC equipped with a 1.7 µm C18 BEH column (Waters) operated with a 6.0 min gradient from 0.1% formic acid to 0.1% formic acid in 20:80 acetone/acetonitrile. The eluate was analyzed in duplicates by TOF-MS (QTOF Premium, Waters). The instrument voltage was 2.8 or 3.2 kV to the tip of the capillary and analysis was performed in negative or positive mode, respectively. In the negative mode desolvation gas temperature was 400 °C, cone voltage 40 V, and Ar collision gas energy 6.1 V; in the positive mode we used the same settings except for collision energy of 10 V. A blank (0.1% formic acid) and the metabolomics standard were analyzed after every 50 samples during the run.

3.3. Authentic Standards

L-carnitine, linoleic acid and gamma-linolenic acid were purchased from Sigma Aldrich (Copenhagen, Denmark). 1-acyl LPC(18:1), 1-acyl LPE(18:1), PC(16:0/18:1) and PE(16:0/18:1) were obtained from Avanti Lipids (Alabaster, AL, USA). For the synthesis of acetyl L-carnitine, carnitine acetyltransferase from pigeon and acetyl coenzyme A were purchased from Sigma Aldrich. Acetylation of L-carnitine was performed as described by Bergmeyer *et al.* [32]. The 2-acyl lyso-forms were synthesized with phospholipase A1 from Thermomyces lanuginosus (Sigma Aldrich). Phospholipase A1 hydrolyzes the acyl group attached to the 1-position of PC(16:0/18:1) and PE(16:0/18:1) so that acyl-2 LPC(18:1) and LPE(18:1) were produced. The description of the method has been given by Pete *et al.* [33]. For the chemical verification of identified metabolites, one plasma sample from a rat in the fasted and another from the fed state were spiked with LPC(18:1) and LPE(18:1) individually, before analysis by the procedure outlined above.

3.4. Raw Data

The MassLynx™ (Version 4.1, Waters, Milford, MA, USA) software collected centroided mass spectra in real time using leucine-enkephalin as a lock-spray standard injected every 10 s to calibrate mass accuracy. Each of the 80 samples was analyzed in duplicates. For negative mode both measurements were included in the data analysis. However, for positive mode 64 sample measurements were excluded, which leaves 65 and 31 sample measurements for fasting and fed states, respectively. The exclusion criterion was based on an instrumental error occurred during analysis. In this case, the outliers had very low intensity due to injection errors.

The software stores data as non-uniform sample data files, each comprised of three vectors; retention time (0–6 min), *m/z* and intensity. The raw data was converted to an intermediate netCDF format with the DataBridge™ utility provided with the MassLynx software.

3.5. Software Tools for Data Preprocessing

Raw data was transferred to MarkerLynx™ (Version 4.1, Waters, Milford, MA, USA) directly from MassLynx whereas netCDF files were imported to MZmine [14] and XCMS [15].

The available information regarding the principle of algorithms used in MarkerLynx™, MZmine and XCMS and the selected data preprocessing parameters are shown in electronic Supplementary information 1. The raw data was inspected while selecting the parameters for each software tool. For the peak detection step parameters such as minimum peak width included in MZmine (minimum and maximum peak width included in XCMS) and *m/z* tolerance included in MZmine (ppm in XCMS) were chosen by inspecting the raw data in a 2D sample plot (retention *vs. m/z*). For the alignment step (or peak grouping) TIC of at least 10 samples were overlapped to decide maximum retention time shift between samples. On the other hand, some parameters such as noise level or required peak shape were not straightforward to decide. Thus, at least 10 different parameter settings slightly varying were evaluated for each software tool. The optimum parameters were selected based on the best separation in a PCA scores plot. Deisotoping is performed in MATLAB for XCMS preprocessed data. The final outcome from each software tool is a feature set where each feature is denoted by the mass over charge (*m/z*) ratio and a retention time. The feature sets from the three software tools were transferred to MATLAB for further data analysis.

3.6. Custom Methods for Data Preprocessing

An alternative data preprocessing was performed directly on the raw data using MATLAB (Version 7, The Mathworks, Inc., MA, USA). To import netCDF files to MATLAB, the iCDF function [17,34] was employed. The steps of the custom data preprocessing are shown in Figure 5. As the first step, binning was performed on the *m/z* dimension as described by Nielsen *et al.* [17].

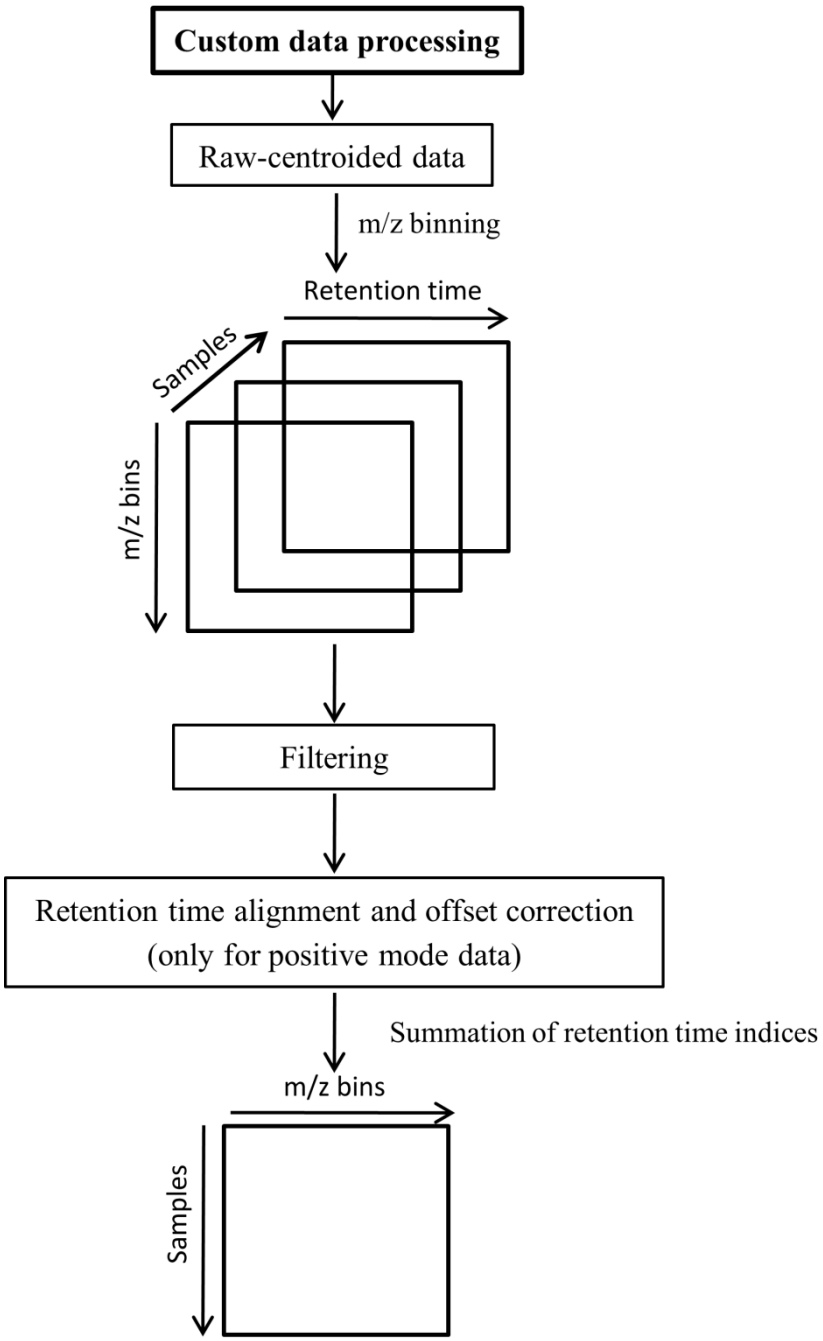
Alignment and offset correction were applied only to positive mode data as the instrumental response was observed to be significantly lower during the duplicate runs in the positive mode. To correct for instrumental response differences, prior alignment was performed using ICOshift [35]. The lower response of duplicates was corrected by calculating the difference matrices between each duplicate set, averaging and adding the average difference to the matrix with the lower response. Here it is assumed that the first injection of a sample holds the correct instrumental response whereas its duplicate with lower response is the one being corrected. The effect of this procedure is shown in Supplementary Information 2.

A threshold level was applied for the elimination of small peaks/intensities lower than the analytical detection level. Values lower than a certain threshold level were considered as zero. The strategy to define the threshold was as follows: (1) The first median value of the whole dataset (excluding zeros) was calculated; (2) That median was evaluated as a threshold (by the ability of principal component analysis (PCA) score plots to fully separate the fasted *vs.* the fed state (data not shown); (3) The next median was calculated by using only those data from the whole dataset that were higher than the

previous median, and again the corresponding PCA scores plot (not shown) was evaluated; (4) This procedure was iterated until an improved separation was achieved by PCA. The threshold levels of the fourth median with the value of 16.17 cps (count per second) in the negative mode and 24.85 cps in the positive mode were selected as adequate.

To enable the application of subsequent two-dimensional data analysis methods, the intensity values of each sample matrix were summed (or collapsed) throughout the retention time index. The resulting data matrix (two-dimensional) is described by samples *vs. m/z* bins (Figure 5) and is also referred to as feature sets throughout this paper.

Figure 5. Custom data preprocessing scheme.



3.7. Data Analysis

The feature sets preprocessed by the three different softwares and the customized method were normalized to unit length and autoscaled. Autoscaling refers to combination of mean centering and unit scaling.

The PLS_Toolbox (Version 5.3, Eigenvector Research, Inc., MA, USA) was used to implement the data analysis. PCA [36] was applied individually on feature sets obtained from each data preprocessing method for general visualization of discrimination of samples from rats in fasted *vs.* fed state.

PLS-DA is based on the development of a PLS model [37] to predict class membership of a dataset X with a y vector including only 0 and 1 (1 indicates that one sample belongs to a given class). Validation of PLSDA classification models was performed by cross model validation as recommended by Westerhuis *et al.* [18]. 25% of the samples were divided as an independent test set. The remaining samples were cross validated (4-fold) to determine optimal number of latent variables that offers minimum cross validation classification errors. In addition, permutation test is applied with 2,000 random assignments of classes. The test set sample classification errors were evaluated to qualify the classification results.

3.7.1. Variable Reduction

A rough and effective variable reduction procedure was performed specifically during MarkerLynx^{TM} and custom data preprocessing by only keeping a feature if it had a nonzero measurement in at least 80% of the intensity values recorded within one of the sample groups (fasting *vs.* fed in this case); otherwise the feature was removed (80% rule) [38]. Gap filling (or peak finding) algorithms implemented in MZmine and XCMS softwares resulted in few zero entries. However, additional filtering algorithm was enabled in MZmine and XCMS prior to gap filling, which removes any feature if it appears in less than 10 samples (settings are defined in Supplementary information 3).

3.7.2. Variable (Feature) Selection

Further variable selection was performed with PLS-DA. The features or *m/z* bins with larger regression coefficients were considered as more discriminative between fasted and fed states and were regarded as potential biomarkers. Due to the fact that PLS-DA is very prone to overfitting, instead of applying only a single cross-validated PLS-DA model for variable selection on all samples, we performed repeated submodel testing. This implies removing samples randomly (here 10% were taken out at a time), constructing a PLS-DA model on the remaining 90% samples and repeating this 1,000 times. By performing many models the importance of each feature for class separation is tested. The number of latent variables (LV) for each model was determined to minimize the classification errors using cross validation (CV). For each model the features are given a ‘rank’ in the order of their regression coefficients and the final rank of each feature for all the 1,000 submodels were summarized with one number using the median of the 1,000 ranks per feature. This method has the potential of reducing false positives so that the features appearing with higher rank in only a few of the submodels were not considered as markers. We arbitrarily selected the 25 top rank features from each feature set, *i.e.*, those with highest absolute regression coefficient products as potentially representing biomarkers.

However, since these features might be daughter ions, adducts, summed ions, *etc.*, we chose here to simply call them 'markers' whereas after identification the compounds represented by these markers in the top rank feature sets will be termed 'biomarkers'.

3.8. Marker Identification

The initial identification of markers was performed according to their exact mass compared with those that were registered in the Human Metabolome Database [39]. Possible fragment ions were investigated by an automated tool using a mol-file format of a candidate compound (MassFragment™, Waters). Further confirmation of candidate biomarkers was obtained by verification of the retention time and fragmentation pattern of an authentic standard (see authentic standards section above). The authentic standards were in some cases selected as one representative of biomarkers belonging to the same chemical compound class, *i.e.*, only one LPC out of a series was confirmed by a standard. Additionally, acyl-1 and acyl-2 LPC(18:1) and LPE(18:1) were spiked into two plasma samples collected in the fed and fasted states, respectively, at a concentration of 0.5 mg/L for a more reliable confirmation.

4. Conclusions

We aimed here to explore the effect of four data preprocessing methods on the pattern of final biomarkers for the fasting and fed states in a small rat study. In our custom method, the binning followed by collapsing across retention time gives rise to false positives and negatives. Even so, half of the marker bins selected contained markers detected by at least one of the other softwares.

The less selective peak picking algorithm for Markerlynx™ and the avoidance of peak picking algorithms for the custom method gave rise to detection of some markers that could not be detected by MZmine or XCMS. On the other hand, the gap filling algorithms in MZmine and XCMS improves marker selection because the true signal differences between groups becomes more correct, *i.e.*, in accordance with the raw data.

The selection of proper software parameters based on the specifics of the dataset is the key for obtaining a high quality data analysis, regardless of the applied software. The better parameter setting is a matter of experience and wrong settings may obscure the extraction of relevant information. The use of more than one software and/or the use of several settings during data preprocessing with any softwareare likely to improve marker detection in untargeted metabolomics.

Although the comparison of the selected marker ions from different data preprocessing methods revealed some differences, further chemical identification revealed that they were often just adducts or daughter ions representing the same biomarker compound. Many of the biomarkers identified were chemically closely related so that any of the softwares and procedures applied here could identify biomarkers explaining a major part of the biological processes differing between the fasting and the fed states in our dataset. Thus, all data preprocessing methods agree that specific lipids, carnitines and amino acids are of importance for discriminating plasma samples from the fed and fasting states. Three major lipid classes, LPCs, LPEs and free fatty acids, emerged as discriminative markers in the rats. The high level in the postprandial state of LPCs, generally known to be pro-inflammatory, is interesting and their possible importance for low-grade inflammation in humans should be further

explored. L-carnitine and acyl carnitines were also found as important markers and the shift from free to acylated carnitine during fasting might be useful as a marker to follow the switch from postprandial lipid storage to the lipid degradation during fasting. Finally, proline and possibly branched chain amino acids seem to be important amino acid markers that decrease in the fasting state when protein catabolism is necessary for their availability.

Supplementary Materials

Supplementary materials can be accessed at: http://www.mdpi.com/2218-1989/2/1/77/s1.

Acknowledgements

This study has been funded in part by the DanORC project that is funded by Danish Research Council and the ISAFRUIT project that was funded by the European Commission (Contract No. FP6-FOOD 016279).

References

1. Zivkovic, A.M.; Wiest, M.M.; Nguyen, U.; Nording, M.L.; Watkins, S.M.; German, J.B. Assessing individual metabolic responsiveness to a lipid challenge using a targeted metabolomic approach. *Metabolomics* **2009**, *5*, 209–218.
2. Sharman, M.J.; Gomez, A.L.; Kraemer, W.J.; Volek, J.S. Very low-carbohydrate and low-fat diets affect fasting lipids and postprandial lipernia differently in overweight men. *J. Nutr*. **2004**, *134*, 880–885.
3. Lindon, J.C.; Nicholson, J.K.; Holmes, E. *The Handbook of Metabonomics and Metabolomics*; Elsevier: Amsterdam, The Netherlands, 2007.
4. Brindle, J.T.; Nicholson, J.K.; Schofield, P.M.; Grainger, D.J.; Holmes, E. Application of chemometrics to H-1 NMR spectroscopic data to investigate a relationship between human serum metabolic profiles and hypertension. *Analyst* **2003**, *128*, 32–36.
5. Constantinou, M.A.; Tsantili-Kakoulidou, A.; Andreadou, I.; Iliodromitis, E.K.; Kremastinos, D.T.; Mikros, E. Application of NMR-based metabonomics in the investigation of myocardial ischemia-reperfusion, ischemic preconditioning and antioxidant intervention in rabbits. *Eur. J. Pharm. Sci*. **2007**, *30*, 303–314.
6. Fardet, A.; Llorach, R.; Martin, J. F.; Besson, C.; Lyan, B.; Pujos-Guillot, E.; Scalbert, A. A liquid chromatography-quadrupole time-of-flight (LC-QTOF)-based metabolomic approach reveals new metabolic effects of catechin in rats fed high-fat diets. *J. Proteome Res.* **2008**, *7*, 2388–2398.
7. Kim, J.Y.; Park, J.Y.; Kim, O.Y.; Ham, B.M.; Kim, H.J.; Kwon, D.Y.; Jang, Y.; Lee, J.H. Metabolic profiling of plasma in overweight/obese and lean men using ultra performance liquid chromatography and Q-TOF mass spectrometry (UPLC-Q-TOF MS). *J. Proteome Res*. **2010**, *9*, 4368–4375.

8. Wilson, I.D.; Nicholson, J.K.; Castro-Perez, J.; Granger, J.H.; Johnson, K.A.; Smith, B.W.; Plumb, R.S. High resolution "Ultra performance" liquid chromatography coupled to oa-TOF mass spectrometry as a tool for differential metabolic pathway profiling in functional genomic studies. *J. Proteome Res*. **2005**, *4*, 591–598.
9. Katajamaa, M.; Oresic, M. Data processing for mass spectrometry-based metabolomics. *J. Chromatogr. A* **2007**, *1158*, 318–328.
10. Castillo, S.; Gopalacharyulu, P.; Yetukuri, L.; Oresic, M. Algorithms and tools for the preprocessing of LC-MS metabolomics data. *Chemom. Intell. Lab. Syst*. **2011**, *108*, 23–32.
11. Yu, T.W.; Park, Y.; Johnson, J.M.; Jones, D.P. apLCMS-adaptive processing of high-resolution LC/MS data. *Bioinformatics* **2009**, *25*, 1930–1936.
12. Schulz-Trieglaff, O.; Hussong, R.; Gropl, C.; Leinenbach, A.; Hildebrandt, A.; Huber, C.; Reinert, K. Computational quantification of peptides from LC-MS data. *J. Comput. Biol.* **2008**, *15*, 685–704.
13. Lange, E.; Tautenhahn, R.; Neumann, S.; Gropl, C. Critical assessment of alignment procedures for LC- MS proteomics and metabolomics measurements. *BMC Bioinformatics* **2008**, *9*, 375.
14. Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **2010**, *11*, 395.
15. Smith, C.A.; Want, E.J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem*. **2006**, *78*, 779–787.
16. Tautenhahn, R.; Bottcher, C.; Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* **2008**, *9*, 504.
17. Nielsen, N.J.; Tomasi, G.; Frandsen, R.J.N.; Kristensen, M.B.; Nielsen, J.; Giese, H.; Christensen, J.H. A pre-processing strategy for liquid chromatography time-of-flight mass spectrometry metabolic fingerprinting data. *Metabolomics* **2010**, *6*, 341–352.
18. Westerhuis, J.A.; Hoefsloot, H.C.J.; Smit, S.; Vis, D.J.; Smilde, A.K.; van Velzen, E.J.J.; van Duijnhoven, J.P.M.; van Dorsten, F.A. Assessment of PLSDA cross validation. *Metabolomics* **2008**, *4*, 81–89.
19. Kristensen, M.; Engelsen, S.B.; Dragsted, L.O. LC-MS metabolomics top-down approach reveals new exposure and effect biomarkers of apple and apple-pectin intake. *Metabolomics* **2011**, in press.
20. Subbaiah, P.V.; Liu, M. Comparative studies on the substrate specificity of lecithin:cholesterol acyltransferase towards the molecular species of phosphatidylcholine in the plasma of 14 vertebrates. *J. Lipid Res.* **1996**, *37*, 113–122.
21. Weltzien, H.U. Cytolytic and Membrane-Perturbing Properties of Lysophosphatidylcholine. *Biochim. Biophys. Acta* **1979**, *559*, 259–287.
22. Han, M.S.; Lim, Y.M.; Quan, W.; Kim, J.R.; Chung, K.W.; Kang, M.; Kim, S.; Park, S.Y.; Han, J.S.; Park, S.Y.; *et al.* Lysophosphatidylcholine as an effector of fatty acid-induced insulin resistance. *J. Lipid Res.* **2011**, *52*, 1234–1246.
23. Sekas, G.; Patton, G.M.; Lincoln, E.C.; Robins, S.J. Origin of plasma lysophosphatidylcholine: Evidence for direct hepatic secretion in the rat. *J. Lab. Clin. Med*. **1985**, *105*, 190–194.

24. Croset, M.; Brossard, N.; Polette, A.; Lagarde, M. Characterization of plasma unsaturated lysophosphatidylcholines in human and rat. *Biochem. J.* **2000**, *345*, 61–67.
25. Seppanen-Laakso, T.; Oresic, M. How to study lipidomes. *J. Mol. Endocrinol*. **2009**, *42*, 185–190.
26. Sandra, K.; Pereira, A.D.; Vanhoenacker, G.; David, F.; Sandra, P. Comprehensive blood plasma lipidomics by liquid chromatography/quadrupole time-of-flight mass spectrometry. *J. Chromatogr. A* **2010**, *1217*, 4087–4099.
27. Kerner, J.; Hoppel, C. Fatty acid import into mitochondria. *Biochimica et Biophysica Acta-Mol. Cell Biol. Lipids* **2000**, *1486*, 1–17.
28. Pearson, D.J.; Tubbs, P.K. Carnitine and Derivatives in Rat Tissues. *Biochem. J*. **1967**, *105*, 953–963.
29. Shaham, O.; Wei, R.; Wang, T.J.; Ricciardi, C.; Lewis, G.D.; Vasan, R.S.; Carr, S.A.; Thadhani, R.; Gerszten, R.E.; Mootha, V.K. Metabolic profiling of the human response to a glucose challenge reveals distinct axes of insulin sensitivity. *Mol. Syst. Biol.* **2008**, *4*, 214.
30. Pietilainen, K.H.; Naukkarinen, J.; Rissanen, A.; Saharinen, J.; Ellonen, P.; Keranen, H.; Suomalainen, A.; Gotz, A.; Suortti, T.; Yki-Jarvinen, H.; *et al.* Global transcript profiles of fat in monozygotic twins discordant for BMI: Pathways behind acquired obesity. *PLoS Med.* **2008**, *5*, 472–483.
31. Poulsen, M.; Mortensen, A.; Binderup, M.L.; Langkilde, S.; Markowski, J.; Dragsted, L.O. The effect of apple feeding on markers of colon carcinogenesis. *Nutr. Cancer* **2011**, *63*, 402–409.
32. Bergmeyer, H.U.; Gawahn, G.; Grassl, M. *Methods of Enzymatic Analysis*, 2nd ed.; Academic Press Inc.: New York, NY, USA, 1974.
33. Pete, M.J.; Ross, A.H.; Exton, J.H. Purification and Properties of Phospholipase-A(1) from Bovine Brain. *J. Biol. Chem.* **1994**, *269*, 19494–19500.
34. Skov, T.; Bro, R. Solving fundamental problems in chromatographic analysis. *Anal. Bioanal. Chem.* **2008**, *390*, 281–285.
35. Savorani, F.; Tomasi, G.; Engelsen, S.B. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *J. Magn. Reson*. **2010**, *202*, 190–202.
36. Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemom. Intell. Lab. Syst*. **1987**, *2*, 37–52.
37. Wold, S.; Sjostrom, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst*. **2001**, 58, 109–130.
38. Bijlsma, S.; Bobeldijk, I.; Verheij, E. R.; Ramaker, R.; Kochhar, S.; Macdonald, I. A.; van, O. B.; Smilde, A. K. Large-scale human metabolomics studies: A strategy for data (pre-) processing and validation. *Anal. Chem.* **2006**, *78*, 567–574.
39. Wishart, D.S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A.C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; *et al.* HMDB: The human metabolome database. *Nucleic Acids Res.* **2007**, *35*, D521–D526.

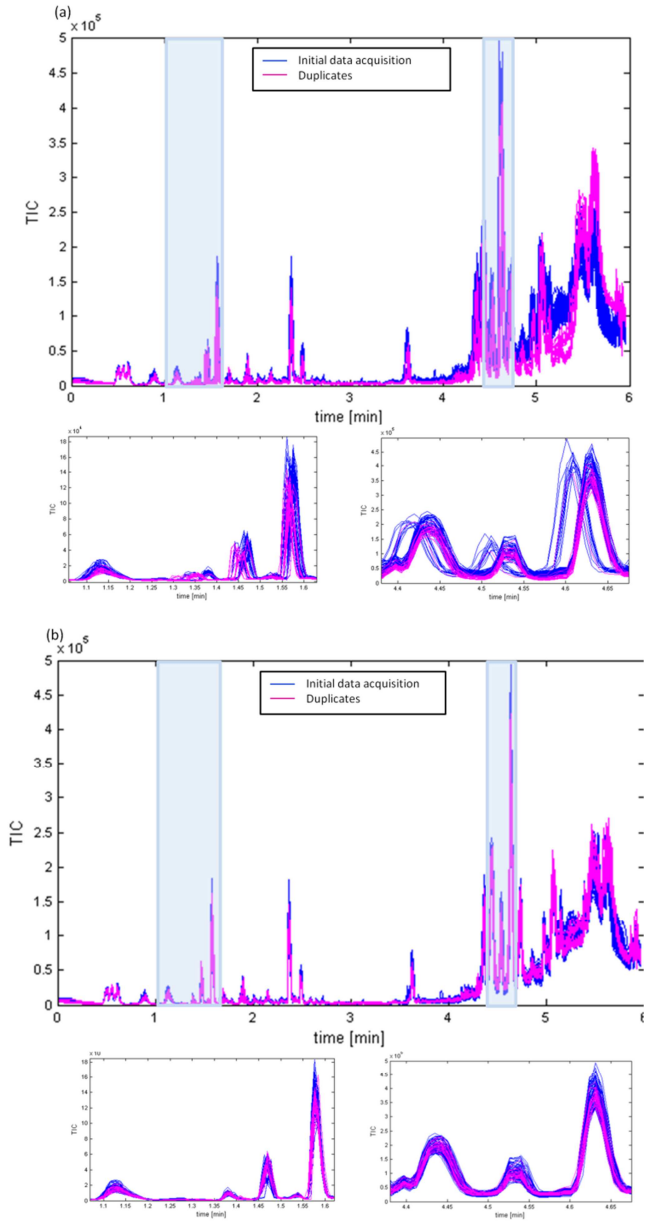
Supplementary information1: Principles of algorithms of data preprocessing softwares

XCMS offers two different peak detection algorithms, *matchedFilter* and *centWave*. The latest developed, *centWave,* was recommended for very complex mixtures which can be represented as plasma in our case. The algorithm first detects the regions of interest in m/z domain based on user defined parameters for mass accuracy (ppm) and maximum and minimum expected chromatographic peak width (peakwidth). Next, chromatographic peaks with different widths were detected using continuous wavelet transform. Finally, features are excluded based on user defined signal to noise ratio (snth). The XCMS alignment algorithm groups peaks together across samples using overlapping m/z bins and calculation of smoothed peak distributions in chromatographic time [9,10] .

MZmine performs peak detection in two steps. The first step is chromatogram builder, which creates continuous chromatograms for each mass within the user-defined mass range (m/z tolerance) based on mass accuracy of the employed instrument. The width of each peak is determined within the range of the chromatogram limited by the user-defined minimum peak width (min time span) and its absolute height is determined with a restriction on height (min absolute height. Each chromatogram is then deconvoluted using one of the four available algorithms. In this study we applied *local minimum search* for deconvolution of the chromatograms. This algorithm is based on separation of peaks based on their local minima. For alignment MZmine offers linear (*join aligner*) and nonlinear (*ransac peak list aligner*) methods. In this study, limited shifts in retention time favored the use of *join aligner* where its algorithm requires user-defined mass and retention time windows (m/z and retention time tolerance). The algorithm tries to match each peak in a master peak list with the peaks in the sample lists and finds the best match based on the retention time and mass windows [14].

MarkerLynx as a commercial software is using algorithms which are not publicly revealed and is thus a kind of black box. In the manual it is stated that the software is applying peak detection by the *ApexPeakTrack* peak detection algorithm. MarkerLynx initially determines the regions of interest in the m/z domain based on mass accuracy (mass tolerance). The ApexPeakTrack algorithm controls peak detection by peak width (peak width at 5% height) and baseline threshold (peak to peak baseline ratio) parameters which can be either set by user or calculated automatically. The algorithm finds the inflection points (peak width at 5% height), local minima

and peak apex to decide peak area and height. It also calculates the baseline noise level using the slope of inflection points. Compared to peak detection algorithms of other softwares, the ApexPeakTrack algorithm produces a much higher number of peaks, so an additional peak removal step (denoted by user defined peak intensity threshold and noise elimination level parameters) is conjugated to the alignment algorithm by its developers. The basic principle of peak removal is described in the accompanying materials: “If a peak is above threshold in one sample and if it is lower than threshold in another sample it lowers the threshold for that sample until it reaches the noise elimination level”. The MarkerLynx alignment algorithm performs alignment of peaks across samples within the range of user-defined mass and retention time windows. (MassLynx (Waters, Milfold, MA))



Supplementary information 2: TIC of the samples obtained from positive mode (a) before and (b) after alignment and normalization using the average difference between replicates that are shifted. The two small inserts below each graph show zoomed parts at two retention time intervals.

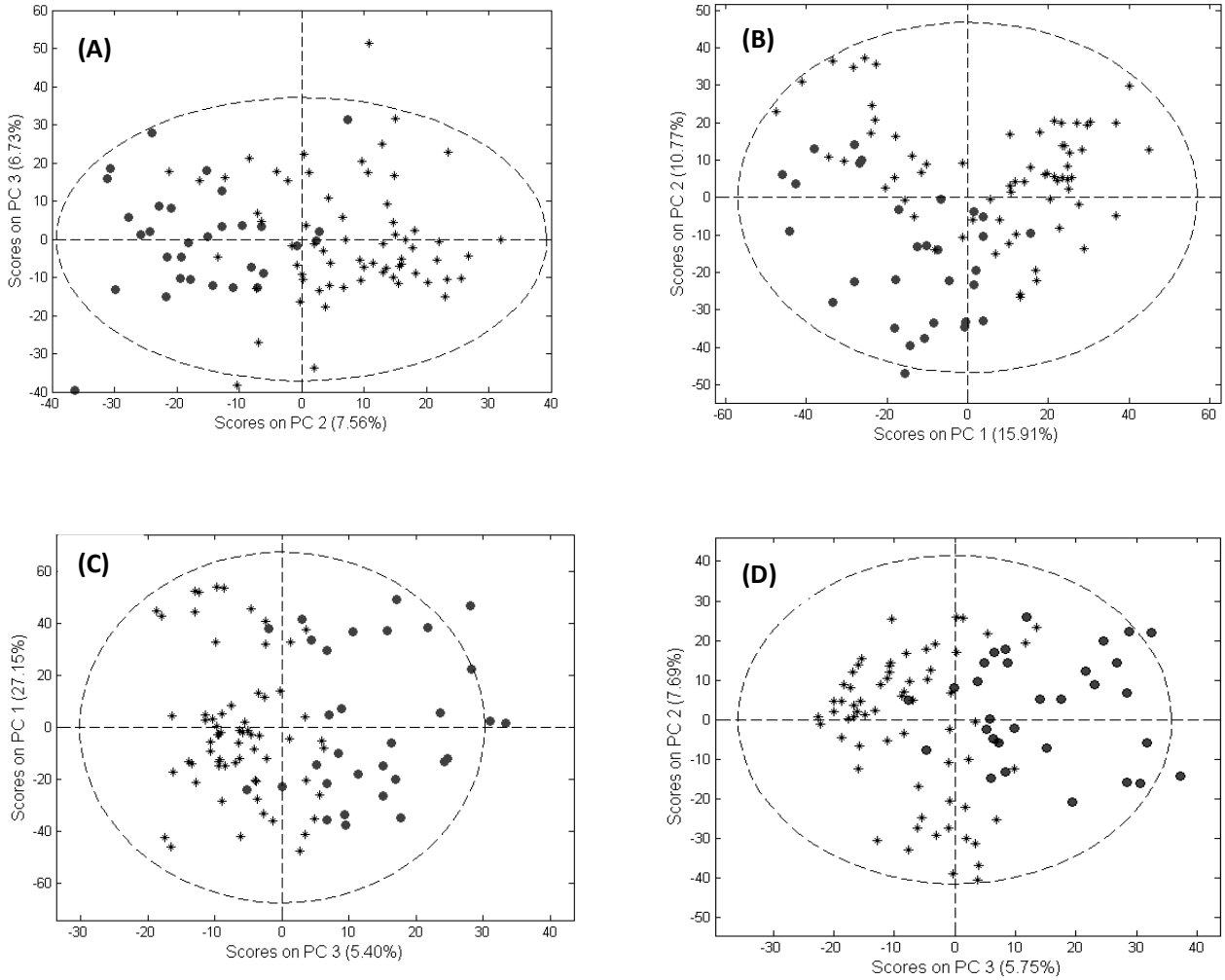
Supplementary information 3. Data preprocessing steps and its parameters settings for MarkerLynx,MZmine and Custom preprocessing (pos: positive mode data; neg: negative mode data).

	MarkerLynx	**MZmine**	**XCMS**	**Custom**
Peak Detection	ApexPeakTrack Peak width at 5% height =	Highest data Point Min time span = 0:01	centWave ppm = 30 (neg);	No peak detection

	default Peak-to-peak baseline ratio = default Noise elimination = 4 Intesity threshold = 30 (neg); 60 (pos)	Min absolute height = 20 (neg); 60 (pos) m/z tolerance = 0.04 (neg); 0.03 (pos) <u>Local minimum search</u> Min RT range = 0:01 Min absolute height = 30 (neg) - 60 (pos) Min peak top/edge = 1.5	40 (pos) peakwidth = (2,10) snth = 4 (neg); 5 (pos) prefilter = c(1,40) (neg); c(1,80) (pos)	
Normalization	✓	-	-	-
Deisotoping	✓	✓	✓	✓
Alignment	m/z window = 0.05 r/t window = 0.05	<u>Join aligner</u> M/Z tolerance = 0.05 RT tolerance = 0:03	<u>Group</u> bw = 4 mxwid = 0.05 <u>Retcor</u> obiwarp profStep = 0.1	ICOshift (pos)
Filtering	80 % rule	<u>Peak list row filter</u> Min peaks in a row = 10 <u>Duplicate peak filter</u> M/Z tolerance = 0.01 RT tolerance = 0.01	Implemented in previous group function minfrac = 0.1	80 % rule
Gap filling	-	<u>Peak finder</u> M/Z tolerance = 0.02 RT tolerance = 0:02	<u>Fillpeaks</u>	-

Supplemantary infomation4: Number of features extracted from each data processing method.

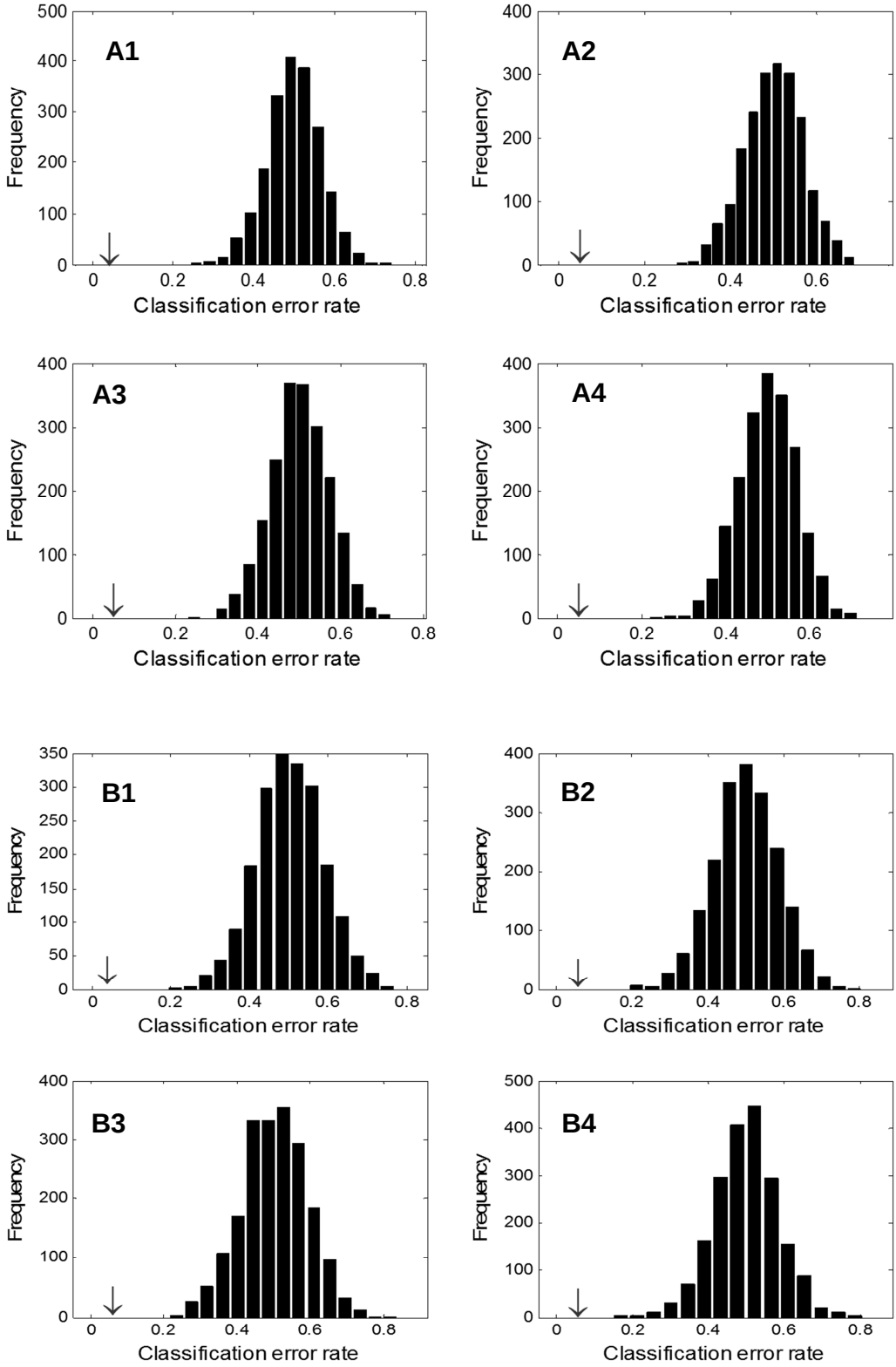
	MarkerLynx		MZmine	XCMS	Custom	
NEG	Before 80% rule:	3780	1501	1562	Before 80% rule:	9500
	After 80% rule:	1852			After 80% rule:	3700
POS	Before 80% rule:	6065	3272	2714	Before 80% rule:	9500
	After 80% rule:	2981			After 80% rule:	3894



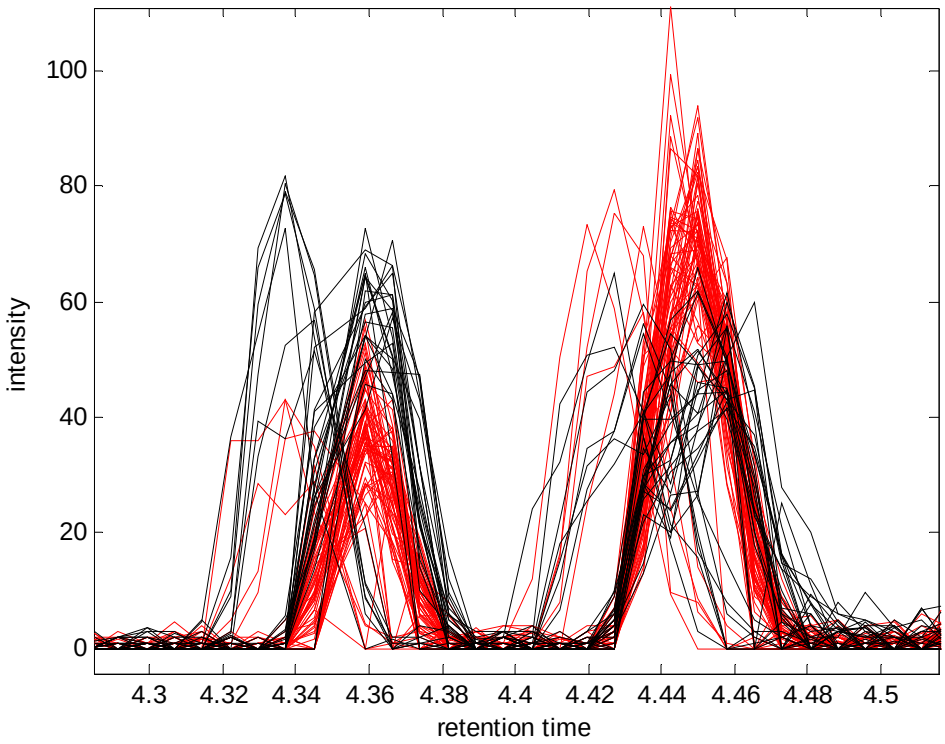
Supplemantary infomation5: PCA scores plot of MarkerLynx (A), MZmine (B), XCMS (C) custom preprocessed (D) positive mode data.

Supplementary information 6: PLSDA model classification error rates of test sets.

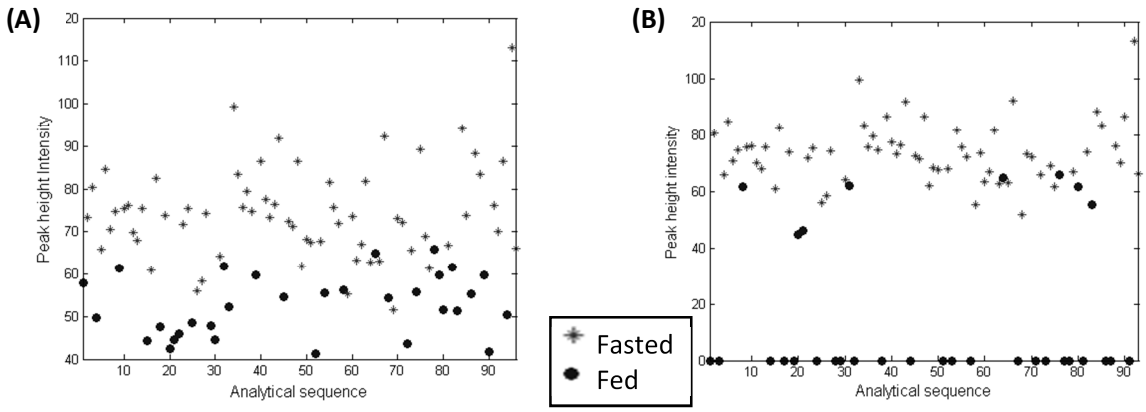
	Classification Error Rates			
Negative mode data	0.00	0.01	0.01	0.01
Positive mode data	0.00	0.02	0.02	0.02


A1
A2
A3
A4
B1
B2
B3
B4
Frequency
Classification error rate

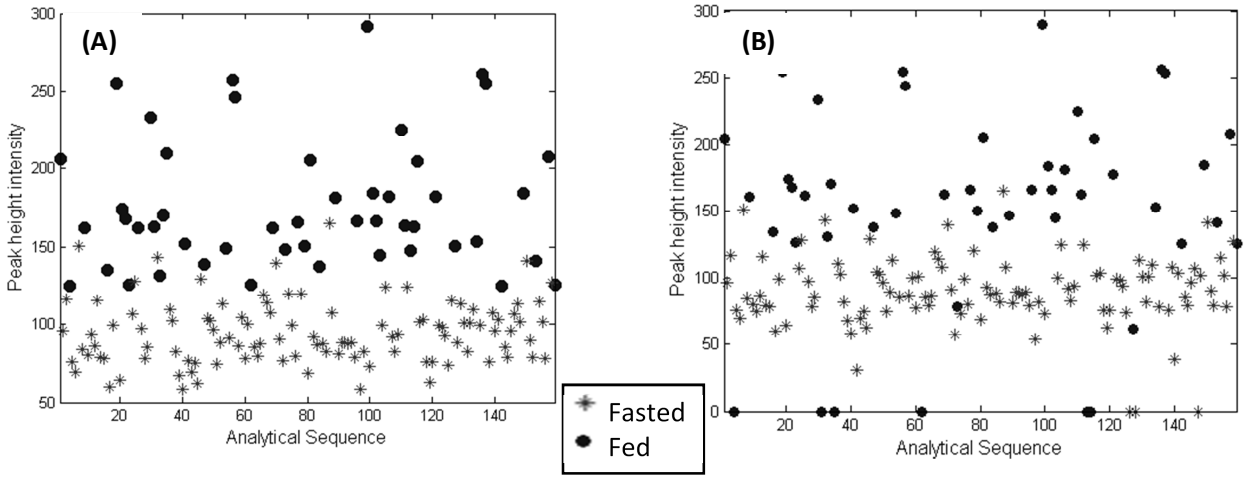
Supplemantary information 7: Classification error rates based on cross model validation predictions of the correct classes (gray arrow) and permuted class labels (black bars) for MarkerLynx (1), MZmine (2), XCMS (3) custom preprocessed negative (A) and positive mode (B).



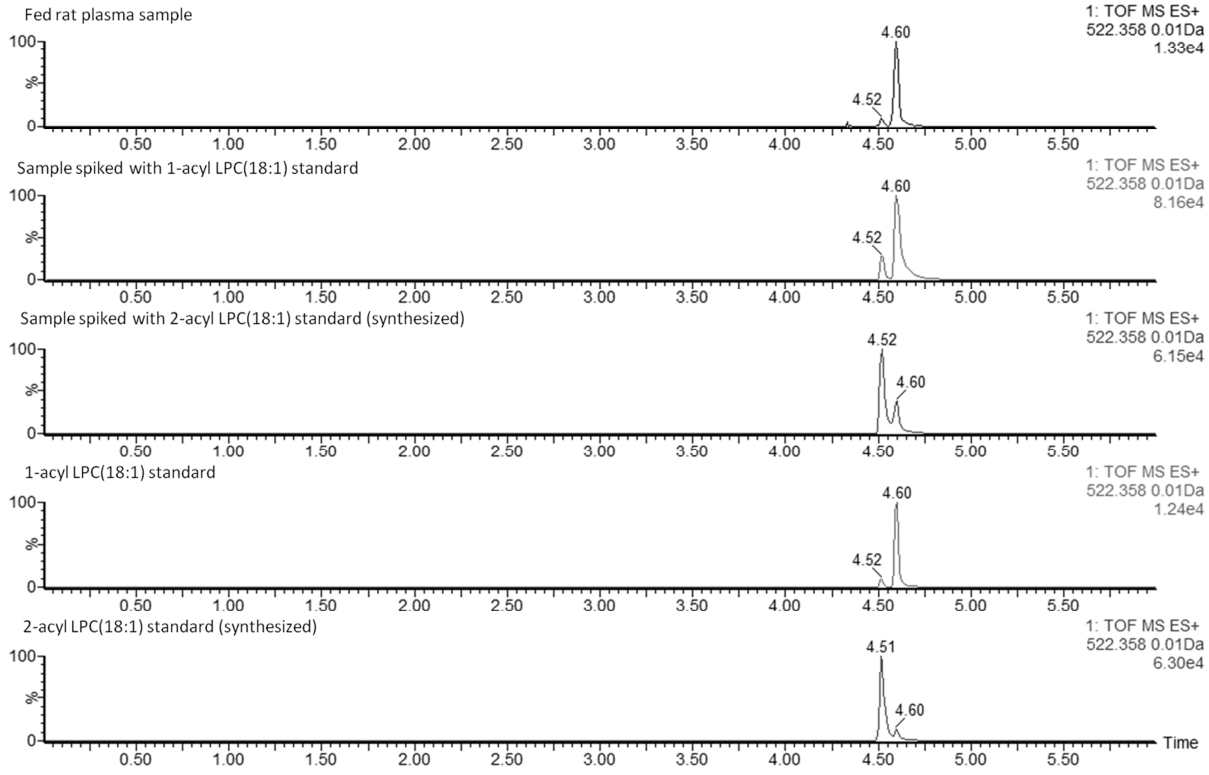
Supplemantary infomation8: The chromatogram of m/z bin = 819.6 from 4.3 to 4.5 min. The peaks are detected as two separate features by the other softwares (Peak no: 42 and 33, positive mode). Red tracks, fasting state; black tracks fed state.



Supplemantary infomation9: (A) MZmine and (B) MarkerLynx recorded peak heights of samples in fasted and fed groups for marker number 42. The difference between the two groups in (B) is inflated as MarkerLynx recorded the signal as zero for many of the samples in fed group. Thus this marker has lower rank in MarkerLynx.

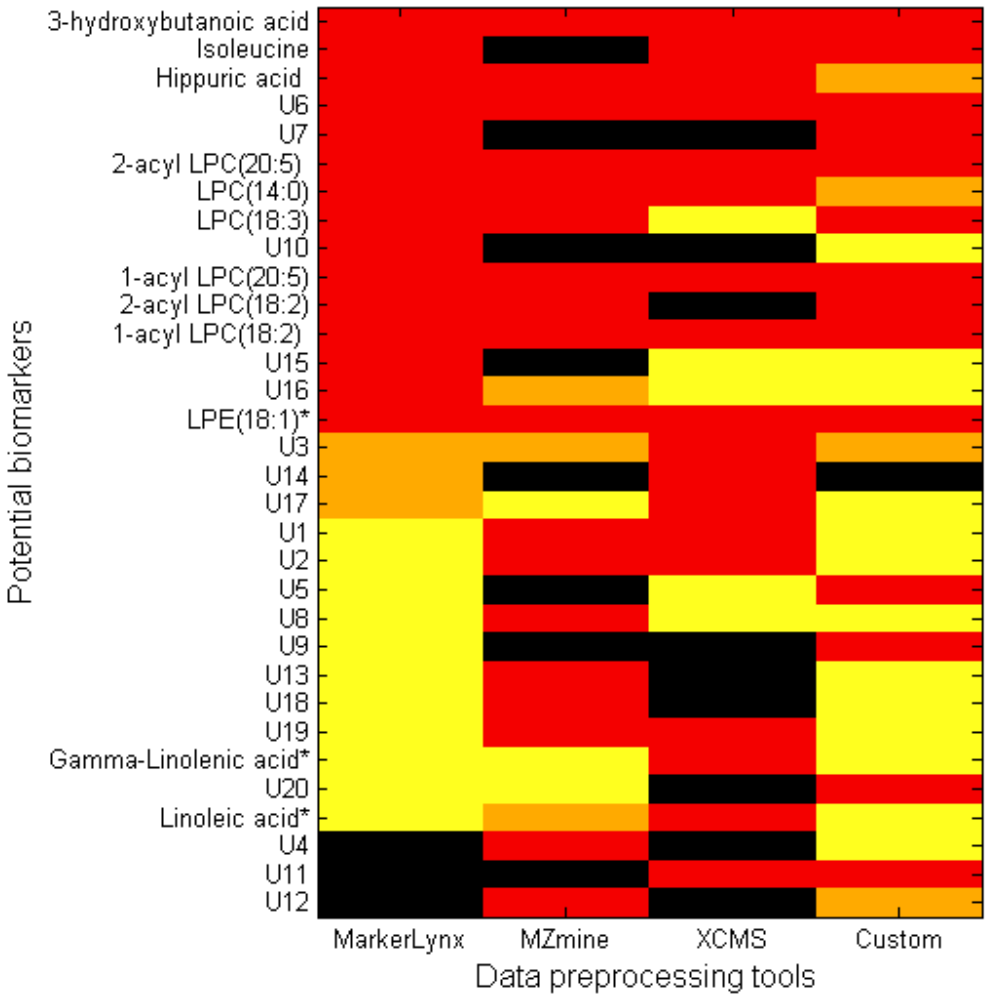


Supplemantary infomation10: (A) MZmine and (B) MarkerLynx recorded peak heights of samples in fasted and fed groups for marker number 38. The difference between the two groups in (B) is deflated as MarkerLynx recorded the signal as zero for some of the samples in the fed group. Thus this marker has higher rank in MarkerLynx.

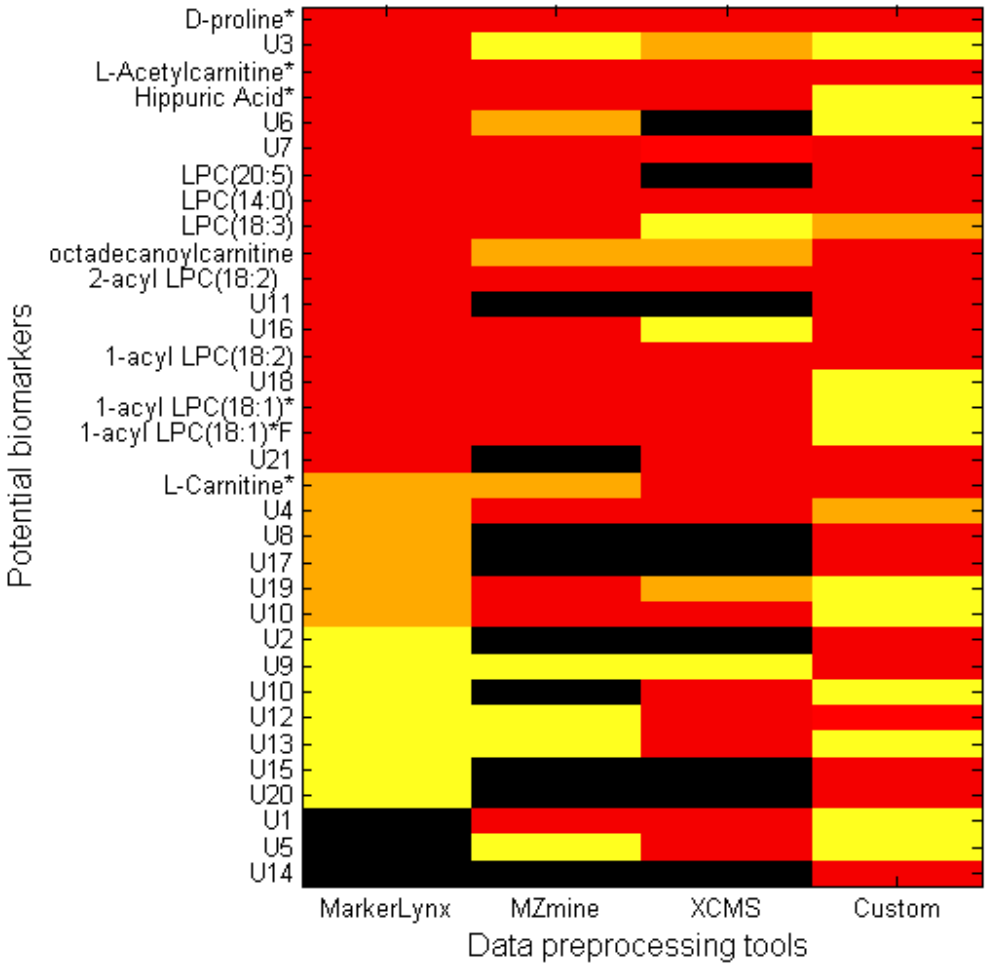


Supplemantary infomation11：XIC of 1-acyl and 2-acyl LPC(18:1) detected in positive mode ionization. The panels in sequence from top to buttom show the extracted ion chromatogram for m/z 522.358 of 1. an authentic rat plasma sample; 2. The same sample spiked with 1-acyl LPC(18:1); 3. The sample spiked with 2-acyl LPC(18:1); 4. A 1-acyl LPC(18:1) standard; 5. A 2-acyl LPC(18:1) standard.

(A)


3-hydroxybutanoic acid
Isoleucine
Hippuric acid
U6
U7
2-acyl LPC(20:5)
LPC(14:0)
LPC(18:3)
U10
1-acyl LPC(20:5)
2-acyl LPC(18:2)
1-acyl LPC(18:2)
U15
U16
LPE(18:1)*
U3
U14
U17
U1
U2
U5
U8
U9
U13
U18
U19
Gamma-Linolenic acid*
U20
Linoleic acid*
U4
U11
U12
Potential biomarkers
MarkerLynx
MZmine
XCMS
Custom
Data preprocessing tools

(B)



Supplemantary infomation12: Heatmap comparing the importance of metabolite based on four different data preprocessing tools. (MarkerLynx, MZmine, XCMS and Custom data processing) for (a) negative and (b) positive mode data. Each row represents the rank (importance) of a marker for four different methods (from Table 1 or 2, 3^{rrd} column). The markers selected had a rank below 25 with at least one of the four methods. The markers were sorted in ascending rank order of MarkerLynx. (red: rank 1-25; orange: rank 26-50; yellow: rank>50; black: not detected).

PAPER II

Effect of trans fatty acid intake on LC-MS and NMR plasma profiles

Gürdeniz G, Rago D, Bendsen NT, Savorani F, Astrup A, Dragsted LO

PLoS One. Submitted.

Effect of *trans* fatty acid intake on LC-MS and NMR plasma profiles

Short title: Trans fat intake: Metabolomics based investigation

Gözde Gürdeniz[1*], Daniela Rago[1], Nathalie Tommerup Bendsen[1], Francesco Savorani[2], Arne Astrup[1] and Lars O. Dragsted[1]

[1]Department of Nutrition, Exercise and Sports, Faculty of Science, Faculty of Science, University of Copenhagen/Rolighedsvej 30, 1958, Frederiksberg C, Denmark

[2]Department of Food Science, Faculty of Science, University of Copenhagen/Rolighedsvej 30, 1958, Frederiksberg C, Denmark

Trial registration: Registered at clinicaltrials.gov as NCT00655902.

E-mail: gozg@life.ku.dk

* Author to whom correspondence should be addressed; Tel.: +45 29176389; Fax: +45 35332483 email : gozg@life.ku.dk

Abstract

Background: The consumption of high levels of industrial *trans* fatty acids (TFA) has been related to cardiovascular disease, diabetes and sudden cardiac death but the causal mechanisms are not well known. In this study, NMR and LC-MS untargeted metabolomics has been used as an approach to explore the impact of TFA intake on plasma metabolites.

Methodology/Principle Findings: In a double-blinded randomized controlled parallel-group study, 52 overweight postmenopausal women received either partially hydrogenated soybean oil, providing 15.7 g/day of TFA (*trans*18:1) or control oil with mainly oleic acid for 16 weeks. Subsequent to the intervention period, the subjects participated in a 12-week dietary weight loss program. Before and after the TFA intervention and after the weight loss programme, volunteers participated in an oral glucose tolerance test. PLS-DA revealed elevated lipid profiles with TFA intake. NMR pointed out an up-regulated LDL cholesterol levels and unsaturation. LC-MS profiles demonstrated elevated levels of specific polyunsaturated (PUFA) long-chain phosphatidylcholines (PCs) and a sphingomyelin (SM) which were confirmed with a lipidomics based method. Plasma levels of these markers of TFA intake declined to their baseline levels, after the weight loss program for the TFA group and did not fluctuate for the control group. The marker levels were unaffected by OGTT.

Conclusions/Significance: This study demonstrated that intake of TFA affects lipid metabolism. The preferential integration of *trans*18:1 into the sn-1 position of PCs, all containing PUFA in the sn-2 position, could be explained by a general up-regulation in the formation of long-chain PUFAs after TFA intake and/or by specific mobilisation of these fats into PCs as a result of TFA exposure. NMR supported these findings by revealing increased unsaturation of plasma lipids in the TFA

group. These specific changes in membrane lipid species may be related to the mechanisms of *trans* fat-induced disease.

Introduction

Industrially produced *trans* fatty acids (TFA) are formed by partial hydrogenation of vegetable oil that changes *cis* configuration of double bond(s) to *trans*, resulting in solid fat for use in margarines, commercial cooking, and manufacturing processes. Partially hardened oils are appealing for food industry owing to their properties such as long shelf life, their stability during deep-frying and their semisolidity. However, consumption of TFA in the human diet have been shown to increase the individual's risk for developing cardiovascular disease [1,2], diabetes [3], and sudden death from cardiac causes [4]. This increased risk has been linked to the impact of TFA on lipoprotein metabolism, inflammation, and endothelial function [5]. It has been well documented that TFA intake increases low-density lipoprotein (LDL) cholesterol, reduces high-density lipoprotein (HDL) cholesterol, and increases the risk of cardiovascular disease [6,7]. Nevertheless, the incidence of CHD reported in prospective studies has been greater than that predicted by serum lipids alone. Thus, the observed associations between TFA consumption and cardiovascular disease events cannot be explained only by changes in lipoprotein levels, triglycerides, apolipoprotein (Apo) B/ApoAI ratio and C-reactive protein [8], implying that the mechanisms behind the adverse effects of TFAs are not fully understood. TFA exposure has also been associated with a higher risk of fatal ischemic heart disease [9] and sudden cardiac death [10]. Although the potential mechanism between TFA and sudden cardiac death is unclear, some have suggested that TFA may modulate cardiac membrane ion channel function [11] or have proarrhythmic properties, affecting cardiovascular electrophysiology [2].

In order to fill the gap between TFA intake and its detrimental health impacts, an untargeted metabolomics approach by allowing quanlification/quatification of hundreds of metabolites can provide a unique insight to potential underlying mechanisms. Many studies have demonstrated metabolomics as a powerful tool to understand responses of individuals with respect to their gene expression or alterations in their lifestyles and diets [12]. The application of liquid chromatography mass spectrometry (LC-MS) and nuclear magnetic resonance (NMR) in metabolomics for measurement of a wide range of metabolites in various biofluids has been well established. NMR provides high reproducibility and is a powerful tool in terms of quantification, whereas LC-MS is more sensitive, allowing detection of a larger number of chemical compounds.

Here, we aimed to contribute to the on-going research interest for identifying the adverse effects of TFA intake by introducing an LC-MS and NMR based metabolomics investigation of a specific TFA intake through 16 weeks. The dietary intervention study was conducted by Bendsen et al. [13] for examining the effect of a high intake of industrially produced TFAs (*trans*18:1) compared to their *cis* analogs (*cis*18:1). Our results revealed an increased presence of membrane-derived, specific long chain polyunsaturated fatty acid (PUFA)-containing PCs and SM with TFA intake, suggesting the possibility of using those compounds as individual markers of TFA integration into plasma membranes.

Materials and Methods

Subjects

52 healthy, moderately overweight (body mass index between 25 and 32 kg m^{-2}), postmenopausal women, between 45 to 70 years of age, were recruited in this study. Detailed description of participant recruitment and enrolment, inclusion and exclusion criteria, and compliance were published previously [13].

The subjects were given both verbal and written information, whereupon all gave written consent. The study was carried out at the Department of Human Nutrition, Faculty of Life Sciences, University of Copenhagen, Frederiksberg, Denmark, between April 2008 and March 2009 and was approved by the Municipal Ethical Committee of The Capital Region of Denmark in accordance with the Helsinki-II declaration (H-B_2007-089) [13].

Study Design

The dietary intervention study had a randomized, double-blind, parallel design. Subjects were given 26 g/d of partially hydrogenated soybean oil with approximately 60% *trans* fats (TFA group; n = 25) or 50/50% mix of palm oil and high oleic sunflower oil as the control oil (CTR group; n = 27) for 16 weeks. Both test oils were supplied by Aarhus Karlshamn, Aarhus C, Denmark. The fatty acid composition in the oils has been described elsewhere [13]. Briefly, the two fats differed in the content of TFA (18:1 *trans*-9, 18:1 *trans*-8, 18:1 *trans*-7), palmitic (16:0), oleic (18:1 *cis*-9) and linoleic acid (18:2 *cis*-6). The fats were incorporated into bread rolls providing a total of 600 kcal/d (41 E% from fat), equivalent to 28% of the subjects' energy requirements on average. Frozen rolls were handed out to the subjects every 1–4 weeks from the department for consumption at home.

The women visited the department for four examinations during the study: at screening (1–8 weeks prior to baseline), baseline (w0), mid-intervention (week 8) and at the end of treatment (w16). In addition, the subjects attended the department for control weighing at weeks 4 and 12. Subjects were instructed to maintain their habitual activity level throughout the dietary intervention period. Subsequent to the dietary intervention period, the subjects participated in a 12-week (w28) dietary weight loss program. The blood samples for metabolomics analysis were collected only at w0, w16 and w28.

Dietary intake was measured using 3-day weighed food records at baseline and in the last week of the intervention. The only significant dietary differences between diet groups during the intervention were the contributions of energy from monounsaturated fatty acids (MUFA) and TFA, indicating that the diets were overall comparable apart from the fatty acid composition. The intake of TFA was higher (7.0 ± 0.2 E% [mean ± SEM] vs. 0.3 ± 0.0 E%) and the intake of MUFA was lower (10.3 ± 0.4 E% vs. 13.4 ± 0.8 E%) in the TFA group compared with the CTR group [13]. The trial was registered at clinicaltrials.gov as NCT00655902.

Ethics statement

The subjects were given both verbal and written information, whereupon all gave written consent. The study was carried out at the Department of Human Nutrition, Faculty of Life Sciences, University of Copenhagen, Frederiksberg, Denmark, between April 2008 and March 2009 and was approved by the Municipal Ethical Committee of The Capital Region of Denmark in accordance with the Helsinki-II declaration (H-B_2007-089). Subjects received B900 US$ as compensation on completion of all the tests. Lean reference subjects received B500 US$.

Blood sampling

Prior to each visit, the subjects were told to fast for at least 10 hours (except for 0.5 L water). They were instructed to avoid alcohol consumption and vigorous exercise on the day before and to consume similar carbohydrate-rich evening meals on the evening before each visit. Body weight and height were measured by standard procedures.

Insulin sensitivity was assessed by use of frequent sampling 3-hour oral glucose tolerance tests (OGTTs) where subjects ingested a solution of 75 g glucose dissolved in 300 mL water. Venous blood samples were collected before and during the OGTT at −10, 30 and 120 minutes into 4 mL

coated tubes. The blood was centrifuged at 3000 g at 4^{o}C for 10 min. The plasma fraction was portioned into 2 mL cryotubes and stored at -80^{o}C until further processing.

Chemicals

Authentic standards PC(18:0/18:2), PC(*cis*18:1/*cis*18:1), *trans* PC(*trans*18:1/*trans*18:1), PC(18:0/18:2) and PC(18:0/20:4), PC(18:0/22:6) were purchased from Avanti Polar Lipids Inc. (Alabaster, AL, USA).

LC-QTOF-MS analysis

Plasma protein precipitation was performed, as described earlier [14]. An ultra-performance liquid chromatography (UPLC) system coupled to quadruple time-of-flight (Premier QTOF) mass spectrometer (Waters Corporation, Manchester, UK) was used for sample analysis. The mobile phase was 0.1% formic acid in water (A) and 0.1% formic acid in 70% acetonitrile and 30% methanol (B). Five µL of each sample were injected into a HSS T3 C_{18} column (2.1 x 100 mm, 1.8µm) coupled with a VanGuard HSS T3 C_{18} column (2.1 x 5mm, 1.8µm) operated for 7.0 min. The eluate was analyzed by electrospray ionization (ESI)-QTOF-MS (Premium QTOF, Waters) in positive and negative mode, applying a capillary voltage of 3.2 kV and 2.8 kV, respectively and cone voltage of 20 kV. Ion source and desolvation gas (nitrogen) temperatures were set at 120 and 400°C, respectively. More detailed UPLC-QTOF analysis conditions were explained previously [15]. Blanks (5% of acetonitrile:methanol 70:30 v/v in water) and external metabolomics standard mixtures were injected every 30 plasma samples throughout each analytical batch.

In order to identify relevant metabolites, MS/MS fragmentation analyses were performed by post-column infusion experiments conducted as follows: 1.6 mM solution lithium formate dissolved in 1:1 mixture water-propanol was infused at 4uL/min using a Waters built-in syringe pump. Both

flows, from the UPLC column and the infusion pump, were combined using a zero-dead-volume 'T' union and introduced into the mass spectrometer. The MS/MS experiment was conducted in positive ion mode operating in product ion scan. The collision-induced dissociation (CID) energy was set at 25 eV and the MS/MS scan range at m/z 100-850. All other parameters were set to the same values with the MS experiment.

In order to verify the findings of lipophilic markers, we performed a lipidomics analysis of 12 samples from each treatment group at baseline and at the end of the intervention. Each sample was added with the internal standards, PC(17:0/0:0), PC(17:0/17:0), PE (17:0/17:0), PG(17:0/17:0), Cer(d18:1/17:0), PS(17:0/17:0), PA(17:0/17:0) (Avanti Polar Lipids, Inc., Alabaster, AL, USA), recemic MG(17:0/0:0/0:0), racemic DG(17:0/17:0/0:0) and TG(17:0/17:0/17:0) (Larodan Fine Chemicals, AB, Malmö, Sweden). The concentration of each standard was approximately 0.1 µg/sample. The samples were extracted as described previously [14], but an additional extraction with 200uL chloroform:methanol (2:1 v/v) was performed on the Sirocco^{R} filter support by gentle shaking with the precipitated protein for 5 min followed by opening of the valves to collect the additional extract. The combined extract was evaporated to dryness and redissolved in 190uL water-saturated chloroform-methanol (2:1). Before injection 0.1 µg of the following additional standards were added in 10µL of the same solvent: PC(16:1/0:0-D3), PC(16:1/16:1-D6), and TG(16:0/16:0/16:0-13C3) (Larodan Fine Chemicals), as described by Nygren et al (2011). The samples were injected on the UPLC-QTOF system using a HSS T3 C_{18} column (2.1 x 100 mm, 1.8µm) coupled with a VanGuard HSS T3 C_{18} column (2.1 x 5mm, 1.8µm). Solvent A was 1% 1 M NH_4Ac and 0.1% HCOOH in water and solvent B was acetonitrile:2-propanol (1:1, v/v), 1% 1 M NH4Ac and 0.1% HCOOH. A 6 min gradient from 100% A to 100% B was used. A gradient in flow was also applied starting from 0.2mL/min, increasing to 0.5mL/min over 3min and going back to starting conditions at 10min with 2min re-equilibration time before next injection.

Identification of lipids. Authentic standards PC(18:0/18:2), PC(18:1/18:1), PC(*trans*18:1/*trans*18:1), PC(18:0/18:2), PC(18:0/20:4), and PC(18:0/22:6) purchased from were analysed by LC-MS with sample analysis instrumental conditions. As it was not possible to purchase the standard compound for each PC and SM, we developed a simple algorithm to extract the PCs and SMs species utilizing their retention time and m/z. The algorithm was based on well-known principle of reversed phase chromatography and predicted based on the m/z values observed for plasma samples. An increased number of carbon atoms results in decreased polarity and increased retention time. In addition, for a PC, SM or lysophosphatidylcholine (LPC) with a specific carbon number, an increasing number of double bonds in the fatty acyl chain reduce the retention time. Since each of the lipid species appear with its Na^+ adduct, this information is utilized to remove irrelevant matches for positive mode data. The samples were analysed two years prior to the authentic standards which resulted in +0.1min linear shift in retention time. Thus, 0.1 min was added to the retention time of each compound in the data set. As shown in Figure 1 for PCs, the retention times of authentic standards were matching almost precisely with the predicted ones (+0.1 min), validating the model. Equally good matching was observed for retention times of authentic standard of SM (36:2) and the observed SM (36:2) (not shown). A few PCs appeared as two isomers, illustrated in Figure 1, corresponding to structural differences.

The structural characterization of compounds reflecting TFA intake was performed by their parent mass information and characteristic fragments in the CID spectrum of their lithated ions. PC(18:1/20:3) and PC(18:1/22:5) are identified with two orthogonal data; retention time and spectral information. PC(18:1/22:6) is putatively annotated whereas SM(18:1/18:2) is putatively characterised. Further information about spectral fragmentation patterns (MS/MS) of the PC(18:1/20:3), PC(18:1/22:5), PC(18:1/22:6) and SM(18:1/18:2) were explained in detail in as follows.

The product ion spectrum of lithated [M+Li]+ ions for PC(18:0/22:6) and PC(18:0/20:4) standards were comparable with the product ion spectrum of lithated [M+Li]+ ions of those, in the samples, confirming their structural identity. A few PCs appeared as two isomers, illustrated in Figure 1, corresponding to structural differences. The ions arising from loss of trimethylamine [M-59], ethylene phosphate [M-183] and lithium ethylene phosphate [M-189] were common fragments for CID spectra of PCs and SMs. The earlier eluting isomer of PC(38:4) gave rise to the fragment ions 504.3, 528.3 and 534.3, matching with the potential fragmentation pattern of PC(18:1/20:3). The ions 504.3 and 528.3 represents the neutral loss of sn-2 fatty acyl substituent as a lithium salt [M+Li-R2CO2Li]+ and as a ketene [M+Li-R2'CHCO]+, respectively [25,26]. Moreover, the loss of sn-1 fatty acyl as a free fatty acid yielded the ion 534.3 corresponding to ([LPC(20:3)-H2O+Li]+). The later eluting isomer of PC(38:4) coeluted with our standard, PC(18:0/20:4). MS/MS spectra of the earlier eluting isomer of PC(40:6) implied contribution of multiple species (PC(18:1/20:5) and PC(20:2/20:4)) to a single chromatographic peak. The ions 552.3 and 526.3 resulted from loss of the sn-1 acyl group as a lithium salt from PC(18:1/22:5) and PC(20:2/20:4), respectively. The most abundant fragment was arising from the removal of the sn-1 substituent as a ketene. The later eluting isomer of PC(40:6) coeluted with our standard, PC(18:0/22:6). MS/MS fragmentation of PC(40:7) lead to its identification as PC(18:1/22:6) based on the fragment ions 550.3 ([M+Li-R1'CHCO]+) and 556.3 ([M+Li-R1CO2Li]+). MS/MS fragmentation of pseudomolcular ion of PC(40:7) [M+H] on Waters Synapt of supported its identity with fragments 445.3 ([LPC(18:1)-OH]+), 504.3 ([M+H-R1CO2H] +), 568.3([M+H-R1'CHCO]+) and 522.3 ([M+H-R2'CHCO]+). However, the CID spectrum of SM(36:3) did not reveal any abundant ions that identify the fatty acyl substituents.

We putatively characterized PC(44:9), which is observed only with potassium adduct, and we could therefore not include it into our prediction model in Figure 1. However, extrapolation of the model agrees with the observed retention time, 5.34 min.

^{1}H NMR Analysis

Plasma samples were slowly thawed overnight at 4 °C. Samples where then centrifuged 20 min at 12k RPM and 300 µl plasma were transferred into a 5 mm NMR tube together with 300 µl of phosphate buffer at pH 7.4 containing at least 10% w/w D_2O and gently mixed in order to avoid formation of bubbles/foam. 1D NOESY ^{1}H NMR spectra were acquired on a Bruker DRX spectrometer (Bruker Biospin Gmbh, Rheinstetten, Germany) operating at 600,00 MHz for protons (14.09 Tesla) using a TCI cryo-probe head and equipped with a SampleJet autosampler. All samples were individually and automatically tuned, matched and shimmed. FIDs were Fourier transformed using a 0.3 Hz line broadening. The resulting spectra were automatically phased and baseline corrected using Topspin™ (Bruker Biospin), and the ppm scale was referenced towards the TSP peak at 0.00 ppm [16]. Assignment of resonances was done by comparison to literature values [17].

Data Preprocessing

LC-MS. The raw data was converted to an intermediate netCDF format with the DataBridge^{TM} utility provided with the MassLynx software. MZmine 2.7 [18] was employed for data preprocessing including following steps: mass detection, chromatogram builder, chromatogram deconvolution (local minimum search), isotopic peaks grouper, peak alignment (join aligner) and gap filling. The final outcome from MZmine is a feature set where each feature is denoted by the mass over charge (m/z) ratio and a retention time.

MZmine preprocessed data was imported to MATLAB (Version 7.2, The Mathworks, Inc., MA, US). Peak filtering was applied based on two criteria. First, if a feature has a reasonable peak area (>60) in the first run blank sample, it is removed. Second, if a feature has a peak area lower than 5 (considered as noise level or gap filling errors), in more than 60% of the samples within both sample groups (TFA vs. CTR, in this case), it is excluded (percent rule, [19]).

To remove intra-individual variation, each feature is normalized with the mean of the two recordings (before and after intervention) for each subject at each OGTT time point (-10, 30 or 120 min) [19].

^{1}H NMR. The spectral alignment was performed by icoshift algorithm [20]. Only the spectral region between 8.5 and 0.2 ppm was considered, and the spectral region containing the residual resonance from water (4.7-5.1 ppm) was removed. Spectral data set was normalized by using probabilistic quotient normalization [21] and reduced by an in-house implementation of the adaptative intelligent binning algorithm [22]. Varying bin size, within the boundaries of minimum 0.002 to a maximum 0.02 ppm, was used, depending on the peak width.

Data Analysis. The PLS_Toolbox (version 6.5, Eigenvector Research, Inc., MA, US) was used to implement the data analysis. Initially, principal component analysis (PCA) was applied to visualize grouping patterns and detection of outliers as an unsupervised multivariate data analysis method. Then, data was subjected to partial least squares-discriminant analysis (PLS-DA) for classification purposes. PLS-DA attempts to separate two groups of samples by regressing on a so-called dummy y-vector consisting of zeros and ones in the PLS decomposition. Permutation test [23] was applied with 1000 random assignment of classes. The test set sample classification errors were evaluated to qualify the classification results. Selectivity ratio [24], which provides a simple numerical assessment of the usefulness of each variable in a regression model, was chosen as the criteria for

variable selection. Briefly, using the y-vector as a target, PLS components (in many cases more than one) are transformed into a single target-projected component. The variance explained by the target component is calculated for each variable and compared with the residual variance for the same variable. The ratio between explained and residual variance, called the selectivity ratio, represents a measure of the ability of a variable to discriminate different groups [24].

Data analysis was performed on baseline adjusted metabolite levels after intervention (w16-w0). Figure 2A illustrates data structure and baseline adjustment schema.

Results

Three subjects did not complete the intervention, eventually resulting in 24 for the TFA group and 25 for the CTR group.

Plasma ^{1}H NMR profiles – extraction of TFA related patterns

Based on sample preparation issues, 42 NMR spectra were excluded, leaving 105 spectra (51 for TFA group, 54 for CTR group) for further analysis. Subsequent to binning, spectral data set was condensed into 1493 binned ppm regions. Primarily, PLS-DA was applied individually for the data including only one OGTT time point with the aim of discriminating of CTR and TFA groups. The original classifications errors were barely significantly lower than the permuted ones (not shown). The classification performance was improved when we concatenated OGTT time points in the sample direction. The original and permuted data classification errors are given in Figure 3, none of the permutations had lower classification errors than the original ones.

The resonances reflecting TFA intake was selected based on evaluation of selectivity ratios from PLS-DA model (i.e. the resonances that have high selectivity ratio are more influential in discriminating between TFA and CTR groups). Annotation of discriminative resonances revealed

elevated unsaturated lipids (δ 5.3) and LDL & VLDL (δ 1.28) methylenic protons for TFA group and an unassigned quartet (δ 3.23) for CTR group.

Plasma LC-MS profiles – extraction of TFA related patterns

Based on instrumental issues, 29 sample measurements were excluded. In all, 59 and 60 samples measurements were remained for TFA group and CTR groups, respectively. A total of 2260 features in ESI positive mode and 1689 in ESI negative mode were detected by MZmine. After exclusion of noise and irrelevant features, by using blank and the percent rule, 767 and 710 features for positive and negative modes, respectively, remained for data analysis.

Initially, samples from each OGTT time point was analysed individually by PLSDA with the aim of discriminating CTR and TFA groups. Permutation test was applied to investigate the potential PLSDA over-fitting issues. Classification error distributions from models with 1000 times permuted class identifiers together with the original classification error are presented in Figure 4. In case there were no differences between the groups, the expected classification error would be 0.5. Figure 4 perfectly matches this requirement. The comparison of classification error of the original model against the permutations was evaluated on the basis of p-values. Original classification errors were significantly lower than the permutations with p-values of 0.01 for T_{OGTT} = -10, 0.04 for T_{OGTT} = 30, and 0.03 for T_{OGTT} = 120 ($\alpha = 0.05$).

Variable selection was performed based on the selectivity ratio from the PLSDA model using datasets from each OGTT time point. Features with the highest selectivity ratio were extracted (Table 1). Many of the discriminating features were common for the three OGTT time points indicating that TFA related patterns were not affected by OGTT. Identification of these features (as described in Materials and Methods section) pointed out that those were compounds from the lipid classes; PCs and SMs.

A similar variable selection procedure was applied for negative mode, though PLSDA classification performance was lower compared to positive mode. Still, identical PC species (Table 1) were associated with TFA intake (not shown). However, SM(36:3) was not detected in the negative mode which could be a potential reason for the lower classification performance.

Since metabolites responding to the TFA exposure did not seem to be affected by OGTT, we concatenated the time points into a new data set, to increase the power of classification model with larger number of samples. In this case each subject was represented by three time points from OGTT measurements as illustrated in Figure 2B. Furthermore, as we have already demonstrated that only lipids were associated with TFA intake, features from the lipid classes (PC, SM and LPC) were included as variables (Figure 2B). The idea behind targeting the lipids was to explore whether only the specific PCs and the SM mentioned in Table 1 respond to TFA intake or if there are other lipids that could be blurred due to the large number of variables. The PCA scores plot is shown in Figure 5. The control group clearly separated from the TFA group in the second principal component. Samples from different time points were quite spread in both CTR and TFA clusters and none of the principal components explained OGTT (not shown). Later, PLS-DA was applied to select the main contributing lipids. The classification errors, sensitivity and specificity of cross validated samples were 0.04, 0.85 and 0.88, respectively. The calculated selectivity ratios were the largest for the lipid compounds given in Table 1 (Figure 6) which were all increased with TFA intake.

In order to investigate whether the increase in specific lipids is temporary or those are remained for longer period, the measurements at w28 (i.e.12 weeks after the end of the intervention) were included. As mentioned earlier in this period all subjects were under a weight loss program. The levels of SM(36:3) and PC(40:7) were increased at w16 and declined to levels before intervention (w0) at w28 for the TFA group, whereas there was no clear fluctuation for the CTR group (Figure

7). The other markers in Table 1 exhibited similar trends (not shown). The standard deviation for the TFA group was higher at w16, which is related to varying individual responses to TFA intake.

Finally, to ascertain that the two major markers identified, SM(36:3) and PC(40:7), were genuine TFA markers in plasma we quantified them by a targeted lipidomics analysis of a subset of 12 samples from each group using appropriate internal standards. Under the lipidomics conditions used here the two markers emerged as significantly higher by factors of 16-40 in the period with trans-fat exposure and were very low before intervention or during control conditions. No other features emerged with similar strong contrasts and other PCs such as two PC(36:2) isomers did not differ between the two treatments. Some weaker markers of TFA exposures may possibly exist but that would need more extensive analysis of the full set to ascertain.

Discussions

TFA has been banned in Denmark since 2003 and background levels of TFA in Danish citizens are therefore low, resulting only from residual exposures from ruminant fats [27]. This has made Denmark an ideal place for interventions to investigate the short-term effects of TFA with a low background exposure. Several TFAs exist and in the current study, *trans*18:1 was almost exclusively given as the intervention [13]. From this well-controlled study of *trans vs. cis* C18:1 fat in overweight women we report that both ^{1}H NMR and LC-MS plasma metabolic profiles were altered with TFA intake. In this as in many other studies consumption of TFA is related with an increased LDL to HDL ratio and it is considered as a powerful predictor of cardiovascular disease [28]. Another outcome from NMR was elevated unsaturated lipid signals for the TFA group, which can be attributed to an increased level of unsaturated fatty acyl side chains in lipid species. The fatty acid composition of phospholipids in red blood cell membranes was reported by Bendsen et al. [13] Their results did not reveal any significant alteration between the CTR and TFA groups with respect

to the PUFA (or monounsaturated) levels, except for a different content of TFA. Thus, this difference may be arising from unsaturation of other lipid groups such as triglycerides in the lipoproteins. Similarly, an elevated unsaturation in the NMR spectrum (δ 5.3-5.4 and δ 1.9-2.5) of HepG2 cell extracts exposed to TFA was mentioned by Najbjerg et al. [29] in which they concluded disturbed lipid storage efficiency with TFA intake.

The LC-MS profiles demonstrated elevated levels of a limited number of polyunsaturated long chain PCs (PC(40:7), PC(40:6), PC(38:4)) and of SM(36:3), which has the longest chain and the highest unsaturation among all detected SMs. Increased double bond formation has also been supported by NMR results. None of these markers were affected by the OGTT test, revealing that they are not necessarily fasting state markers. TFA intake did not seem to have long term effects on the composition of plasma lipids, as their levels at w28 after intervention (weight loss period) were comparable to baseline (w0) levels as shown in Figure 7. We observed here a SM as a marker of TFA intake. This SM was almost not present in the non-TFA group indicating a special structure. An increased level of total plasma SMs has been associated with increased risk of atherosclerosis [30,31] although the consequence in terms of cardiovascular risk has been debated [32]. In this study we observed an increase in only a single, minor SM having two C18 chains with one and two double bonds, respectively, either SM(d18:2/18:1) or SM(d18:1/18:2). The configuration (*cis* or *trans*) around the double bonds in these markers is unresolved and so is the atherogenic potential of this specific SM. There was no correlation between the concentration of this SM and any other compound from this class. We speculate that the marker observed here has a Δ9 or Δ 11 *trans*-fatty sphingosine chain containing a *cis*-double bond in the 3-position. This would result from Δ7 or Δ9- and other trans-hexadecanoic acids being a substrate for the slightly promiscuous serine palmitoyltransferase (EC 2.3.1.50) [33] to form a 3-ketodehydrosphingosine which would then be reduced and acylated by oleyl-CoA followed by desaturation to form Cer(d18:2/18:1). This

ceramide would act as a precursor to the SM(d18:2/18:1) formed by SM synthase (EC 2.7.8.27). We are not able to see the less polar products postulated here not even by lipidomics, but the consequence of this hypothesis would be that after intakes of *trans*16:1 fatty acids it would be possible to observe the formation of a whole series of sphingolipids containing the unusual Δ 9 or Δ 11- *trans*- Δ3-*cis* C18:2 and other similar sphingosines with the *trans* double bond in other positions. In the current study *trans*16:1 was below detection limit in the diet but it is likely that it is formed by β-oxidation of Δ9 or Δ 11-*trans*-18:1. In a study of 16:1 ruminant TFAs, the Δ9 was the dominating isomer but *trans* double bond isomers with the double bond at any carbon from position 3 up to14 also existed [34]. The identity of our SM(d18:2/18:1) marker needs to be finally proven in separate studies, and if the assignment is correct the biological and especially neurological consequence of changing the usual *cis*-Δ3-sphingosines by an aberrant backbone must be elucidated.

We succeeded in identifying several PCs based on authentic standards and by a systematic pattern of RTs depending on chain length and saturation. Based on this pattern we could identify two PC's, PC(40:6) and PC(40:7), which were specifically increased in plasma following dietary TFAs, and PC(38:4), which tended to be increased as well. These PCs carry a C18:1 acyl side chain in one position and a long-chain PUFA chain in the other based on their CID fragmentation patterns. Since C20 and C22 acyl side chains in PCs are almost exclusively found in the sn-2 position in humans [35], it is most likely that the 18:1 is found in sn-1. TFAs, including *trans*-vaccenic acid (Δ11-*trans*-18:1), sterically resemble saturated fatty acids and might therefore substitute for these in the sn-1 position. In agreement, the preferential incorporation of elaidic acid to sn-1 chain of phospholipids have been reported in hepatocytes by Woldseth et al. [36]. In accordance, Wolf and Entressangles et al. [37] showed that phospholipids from rat liver mitochondria modified *in vivo* had large quantities of elaidic acid esterified at the sn-1 position. We therefore propose that the

species observed here are PC(*trans*18:1/22:5), PC(*trans*18:1/22:6) and PC(*trans*18:1/20:3). This hypothesis is supported by the previously reported elevated trans-18:1 residue levels in red blood cell phospholipids in the TFA group [13].

It is well known that TFA incorporate membrane phospholipids into plasma altering the packing of phospholipid and influencing the physical properties and responses of membrane receptors [38,39]. TFA produce membrane properties more similar to those of saturated chains than those of acyl chains containing *cis* double bonds [39]. When incorporated into membrane phospholipids, TFA either replace existing saturated or *cis* unsaturated acyl chains. Harvey et al. [40] showed that both elaidic and linoelaidic acid integrated into phospholipids, mainly in the expense of myristic, palmitic, and stearic acids, without causing any net gain in total fatty acid levels. In our study, the published membrane phospholipids levels [13] revealed significantly decreased stearic acid (P=0.04) and oleic acid (P=0.02) in the TFA group compared to CTR, suggesting replacement of those with elaidic acid. Although LC-MS based metabolomics did not show any decrease in PCs having one saturated fatty acyl chain, elaidic acid-containing specific PCs potentially increased in the TFA group. Many other researchers have investigated the variation of fatty acid composition in red blood cells PCs with TFA intake; however none of them reported the effect of TFA intake on specific PCs. Here, LC-MS based metabolomics demonstrated up-regulation of specific PCs with TFA.

The TFA markers PC(*trans*18:1/20:3), PC(*trans*18:1/22:4) and PC(*trans*18:1/22:5) preferentially integrated into PCs all contain PUFA in the sn-2 position. There was no difference in the dietary intake of PUFAs in the two diet groups [13], so the preferred presence of these specific acyl chains together with *trans*18:1 would need an explanation. The two minor markers have peaks with a RT slightly different from the main, 18:0 containing PC(40:6) and PC(38:4) peaks (Figure 1), indicating that they may be detectable due to better signal-to-noise ratio for these specific

compounds, but the more prominent PC(40:7) marker is actually dominating the only PC(40:7) peak observed and the level in the non-TFA group is quite low. This is not surprising since this compound in general would be a minor PC because it violates the general rule of saturated sn-1 and unsaturated sn-2 acyl chains and because no C22 fatty acid with seven double bonds exists in human lipids. Other minor TFA-containing PCs may therefore exist but with RTs that fall on top of major PCs so that they are not detected as markers. However, it is still noticeable that PC(40:7) is so abundant. It forms a large peak comparable to other major PCs, indicating a facilitated formation. We also found putatively the even longer, PC(44:9). These observations could either indicate that there is a general up-regulation in the formation of long-chain PUFAs after TFA intakes and/or that these fats are specifically mobilised into PC as a result of TFA exposures. It has been shown that the acyl chain distribution is almost completely similar in plasma and erythrocyte membranes, indicating that plasma PCs may be a surrogate marker for membrane composition. Indeed, most plasma PCs may be abstracted from the membranes in contact with blood. Increased formation of long-chain PUFAs has been observed in adipose tissue membranes in overweight individuals [41], resulting from increased elongase and desaturase activities. This phenomenon is likely due to compensation for the increased fat load in the adipocytes in order for them to remain functional, despite their enlargement during weight gain [41]. TFA resembling saturated fatty acids may therefore negatively affect adipose tissue function leading to a response similar to that seen during weight gain with increased formation of long-chain PUFA's. This is supported also by an increased unsaturation in the NMR spectra for the TFA group, yet the FA composition of red blood cell phospholipids did not show any overall significant increase in PUFA [13]. Further investigation of the PUFA distribution among specific membrane PCs is therefore needed in order to confirm this hypothesis.

As previously mentioned, phospholipids with TFA behave similar to saturated fatty acids rather than their *cis* monounsaturated isomers. It has been reported that *trans*-acyl chains adopt extended configurations similar to saturated acyl chains, allowing better interaction with the cholesterol molecule compared with their *cis* analogs [42]. These effects could be contributing factors in modulating cholesterol homeostasis, and as such, may be part of the explanation of the elevation of LDL cholesterol by a TFA-rich diet [42] which was demonstrated by NMR. Although TFA has properties similar to those of saturated fatty acids and also substitute for saturated fatty acids in membrane lipids, it has been confirmed in a meta-analysis that TFA raises levels of LDL more than an equal amount of saturated fatty acids. The effect on LDL levels is much larger when TFAs are compared with their *cis* analogs [6].

Conclusions

We conclude that several specific markers of TFA intake have been observed in this study and propose that SM(d18:2/18:1) may be a general plasma marker of exposure to TFAs as well as that the presence of PC(*trans*18:1/22:6) may be a specific marker of C18:1 TFA exposure. This study was established to investigate the effect of 18:1 TFA intake on plasma metabolites using an untargeted approach. As the results demonstrate, specific lipid molecular species in plasma were formed as a result of TFA exposure and all belong to the SM and PC polar lipids that exist in plasma in equilibrium with the plasma membranes. We could also confirm that TFA exposure leads to increased plasma LDL. Further studies with other specific exposures to 16:1 and 18:2 TFAs would give further insight into the general and specific lipid markers of TFA exposure.

Acknowledgements

We would like to thank to Abdelrhani Mourhrib, for preparing the samples for NMR analysis.

References

1. Oh K, Hu FB, Manson JE, Stampfer MJ, Willett WC (2005) Dietary fat intake and risk of coronary heart disease in women: 20 years of follow-up of the nurses' health study. Am J Epidemiol 161: 672-679.
2. Soares-Miranda L, Stein PK, Imamura F, Sattelmair J, Lemaitre RN, Siscovick DS, Mota J, Mozaffarian D (2012) Trans-Fatty Acid Consumption and Heart Rate Variability in Two Separate Cohorts of Older and Younger Adults. Circ Arrhythm Electrophysiol 5: 728-738.
3. Hu FB, Manson JE, Stampfer MJ, Colditz G, Liu S, Solomon CG, Willett WC (2001) Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. N Engl J Med 345: 790-797.
4. Lemaitre RN, King IB, Raghunathan TE, Pearce RM, Weinmann S, Knopp RH, Copass MK, Cobb LA, Siscovick DS (2002) Cell membrane trans-fatty acids and the risk of primary cardiac arrest. Circulation 105: 697-701.
5. Mozaffarian D, Katan MB, Ascherio A, Stampfer MJ, Willett WC (2006) Trans fatty acids and cardiovascular disease. N Engl J Med 354: 1601-1613.
6. Mensink RP, Zock PL, Kester AD, Katan MB (2003) Effects of dietary fatty acids and carbohydrates on the ratio of serum total to HDL cholesterol and on serum lipids and apolipoproteins: a meta-analysis of 60 controlled trials. Am J Clin Nutr 77: 1146-1155.
7. Flock MR, Kris-Etherton PM (2011) Dietary Guidelines for Americans 2010: implications for cardiovascular disease. Curr Atheroscler Rep 13: 499-507.
8. Mozaffarian D, Clarke R (2009) Quantitative effects on cardiovascular risk factors and coronary heart disease risk of replacing partially hydrogenated vegetable oils with other fats and oils. Eur J Clin Nutr 63 Suppl 2: S22-S33.
9. Yaemsiri S, Sen S, Tinker L, Rosamond W, Wassertheil-Smoller S, He K (2012) Trans fat, aspirin, and ischemic stroke in postmenopausal women. Ann Neurol DOI: 10.1002/ana.23555.

10. Lemaitre RN, King IB, Mozaffarian D, Sotoodehnia N, Siscovick DS (2006) Trans-fatty acids and sudden cardiac death. Atheroscler Suppl 7: 13-15.
11. Katz AM (2002) Trans-fatty acids and sudden cardiac death. Circulation 105: 669-671.
12. Oresic M (2009) Metabolomics, a novel tool for studies of nutrition, metabolism and lipid dysfunction. Nutr Metab Cardiovasc Dis 19: 816-824.
13. Bendsen NT, Chabanova E, Thomsen SB, Larsen TM, Newman JW, Stender S, Dyerberg J, Haugaard SB, Astrup A (2011) Effect of trans fatty acid intake on abdominal and liver fat deposition and blood lipids: a randomized trial in overweight postmenopausal women. Nutrition and Diabetes 1: 1-11.
14. Gürdeniz G, Kristensen M, Skov T, Dragsted LO (2012) The Effect of LC-MS Data Preprocessing Methods on the Selection of Plasma Biomarkers in Fed vs. Fasted Rats. Metabolites 2: 77-99.
15. Barri T, Holmer-Jensen J, Hermansen K, Dragsted LO (2012) Metabolic fingerprinting of high-fat plasma samples processed by centrifugation- and filtration-based protein precipitation delineates significant differences in metabolite information coverage. Analytica Chimica Acta 718: 47-57.
16. Rasmussen LG, Winning H, Savorani F, Ritz C, Engelsen SB, Astrup A, Larsen TM, Dragsted LO (2012) Assessment of dietary exposure related to dietary GI and fibre intake in a nutritional metabolomic study of human urine. Genes Nutr 7: 281-293.
17. Nicholson JK, Foxall PJ, Spraul M, Farrant RD, Lindon JC (1995) 750 MHz ^{1}H and ^{1}H-^{13}C NMR spectroscopy of human blood plasma. Anal Chem 67: 793-811.
18. Pluskal T, Castillo S, Villar-Briones A, Oresic M (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics 11: 395.
19. Bijlsma S, Bobeldijk I, Verheij ER, Ramaker R, Kochhar S, Macdonald IA, van OB, Smilde AK (2006) Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. Anal Chem 78: 567-574.
20. Savorani F, Tomasi G, Engelsen SB (2010) icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. J Magn Reson 202: 190-202.

21. Dieterle F, Ross A, Schlotterbeck G, Senn H (2006) Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in H-1 NMR metabonomics. Anal Chem 78: 4281-4290.
22. De Meyer T, Sinnaeve D, Van Gasse B, Tsiporkova E, Rietzschel ER, De Buyzere ML, Gillebert TC, Bekaert S, Martins JC, Van Criekinge W (2008) NMR-Based Characterization of Metabolic Alterations in Hypertension Using an Adaptive, Intelligent Binning Algorithm. Analytical Chemistry 80: 3783-3790.
23. Westerhuis JA, Hoefsloot HCJ, Smit S, Vis DJ, Smilde AK, van Velzen EJJ, van Duijnhoven JPM, van Dorsten FA (2008) Assessment of PLSDA cross validation. Metabolomics 4: 81-89.
24. Rajalahti T, Arneberg R, Berven FS, Myhr KM, Ulvik RJ, Kvalheim OM (2009) Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. Chemometrics and Intelligent Laboratory Systems 95: 35-48.
25. Ivanova PT, Milne SB, Byrne MO, Xiang Y, Brown HA (2007) Glycerophospholipid identification and quantitation by electrospray ionization mass spectrometry. Methods Enzymol 432: 21-57.
26. Hsu FF, Turk J (2009) Electrospray ionization with low-energy collisionally activated dissociation tandem mass spectrometry of glycerophospholipids: mechanisms of fragmentation and structural characterization. J Chromatogr B Analyt Technol Biomed Life Sci 877: 2673-2695.
27. Stender S, Astrup A, Dyerberg J (2008) Ruminant and industrially produced trans fatty acids: health aspects. Food Nutr Res 52. DOI: 10.3402/fnr.v52i0.1651.
28. Brouwer IA, Wanders AJ, Katan MB (2010) Effect of animal and industrial trans fatty acids on HDL and LDL cholesterol levels in humans--a quantitative review. PLoS One 5: e9434.
29. Najbjerg H, Young JF, Bertram HC (2011) NMR-based metabolomics reveals that conjugated double bond content and lipid storage efficiency in HepG2 cells are affected by fatty acid cis/trans configuration and chain length. J Agric Food Chem 59: 8994-9000.
30. Jiang XC, Paultre F, Pearson TA, Reed RG, Francis CK, Lin M, Berglund L, Tall AR (2000) Plasma sphingomyelin level as a risk factor for coronary artery disease. Arterioscler Thromb Vasc Biol 20: 2614-2618.

31. Schlitt A, Blankenberg S, Yan D, von GH, Buerke M, Werdan K, Bickel C, Lackner KJ, Meyer J, Rupprecht HJ, Jiang XC (2006) Further evaluation of plasma sphingomyelin levels as a risk factor for coronary artery disease. Nutr Metab (Lond) 3: 5
32. Yeboah J, McNamara C, Jiang XC, Tabas I, Herrington DM, Burke GL, Shea S (2010) Association of plasma sphingomyelin levels and incident coronary heart disease events in an adult population: Multi-Ethnic Study of Atherosclerosis. Arterioscler Thromb Vasc Biol 30: 628-633.
33. Merrill AH, Jr., Williams RD (1984) Utilization of different fatty acyl-CoA thioesters by serine palmitoyltransferase from rat brain. J Lipid Res 25: 185-188.
34. Precht D, Molkentin J (2000) Identification and quantitation of cis/trans C16:1 and C17:1 fatty acid positional isomers in German human milk lipids by thin-layer chromatography and gas chromatography/mass spectrometry. Eur J Lipid Sci Technol 102: 102-113.
35. Marai L, Kuksis A (1969) Molecular species of lecithins from erythrocytes and plasma of man. J Lipid Res 10: 141-152.
36. Woldseth B, Retterstol K, Christophersen BO (1998) Monounsaturated trans fatty acids, elaidic acid and trans-vaccenic acid, metabolism and incorporation in phospholipid molecular species in hepatocytes. Scand J Clin Lab Invest 58: 635-645.
37. Wolff RL, Entressangles B (1994) Steady-state fluorescence polarization study of structurally defined phospholipids from liver mitochondria of rats fed elaidic acid. Biochim Biophys Acta 1211: 198-206.
38. Clandinin MT, Cheema S, Field CJ, Garg ML, Venkatraman J, Clandinin TR (1991) Dietary fat: exogenous determination of membrane structure and cell function. FASEB J 5: 2761-2769.
39. Roach C, Feller SE, Ward JA, Shaikh SR, Zerouga M, Stillwell W (2004) Comparison of cis and trans fatty acid containing phosphatidylcholines on membrane properties. Biochemistry 43: 6344-6351.
40. Harvey KA, Walker CL, Xu Z, Whitley P, Siddiqui RA (2012) Trans Fatty Acids: Induction of a Pro-inflammatory Phenotype in Endothelial Cells. Lipids 47: 647-657.
41. Pietilainen KH, Rog T, Seppanen-Laakso T, Virtue S, Gopalacharyulu P, Tang J, Rodriguez-Cuenca S, Maciejewski A, Naukkarinen J, Ruskeepaa AL, Niemela PS, Yetukuri L, Tan CY, Velagapudi V, Castillo S, Nygren H, Hyotylainen T, Rissanen A, Kaprio J, Yki-Jarvinen H, Vattulainen I, Vidal-Puig

A, Oresic M (2011) Association of Lipidome Remodeling in the Adipocyte Membrane with Acquired Obesity in Humans. Plos Biology 9(6):e1000623.

42. Niu SL, Mitchell DC, Litman BJ (2005) Trans fatty acid derived phospholipids show increased membrane cholesterol and reduced receptor activation as compared to their cis analogs. Biochemistry 44: 4458-4465.

Figure Legends

Figure 1. The observed retention time values of identified PCs (empty circles). Filled circles illustrate retention time of the authentic standards, PC(18:1/18:1), PC(18:0/20:4) and PC(18:0/22:6) confirming the predicted pattern

Figure 2. Data structure and arrangement schema. Baseline subtraction (A) concatenation of time points (applied on LC-MS and NMR profiles), and selection of lipid classes (B) (applied on LC/MS data)

Figure 3. Permutation test results for NMR profiles. Class prediction results for NMR profiles based on test set predictions of the original labelling compared to the permuted data assessed using the classification error. P-values were calculated based on the comparison of classification error of the original model against the permutations

Figure 4. Permutation test results for LC-MS profiles at each OGTT time point. Class prediction results for LC-MS profiles based on test set predictions of the original labelling compared to the permuted data assessed using the classification errors. T_{OGTT} = -10 (A) T_{OGTT} = 30 (B) T_{OGTT} = 120 (C)

Figure 5. PC1 vs. PC2 scores plot of LC-MS based lipid profiles. The LC-MS profiles with concatenated time points including only LPCs, PCs and SMs as variables. Filled circles: TFA, empty circles: CTR

Figure 6. Selectivity ratio of each lipid species from PLS-DA model.

Figure 7. Normalized intensity for metabolites reflected by TFA intake. PC(40:7) (A) and SM(36:3) (B) at w0, w16 and w28. The values are the mean of samples in CTR and TFA groups. Each variable is normalized with the mean of the 9 recordings (at week 0, 16 and 28 with three OGTT time point recordings) for each subject

Table 1. Features with the highest selectivity ratio based on PLSDA models. The importance of each feature was represented by its rank. The rank is based on each features sorted selectivity ratio in descending order.

Measured m/z	Retention time* (min)	Suggested Compound	Suggested Adduct	Monoisotopic mass	Rank T_{OGTT}=-10	Rank T_{OGTT}=30	Rank T_{OGTT}=120
749.5614	5.31	SM(36:3)	$[M+Na]^+$	726.5676	1	1	4
727.5781	5.31	SM(36:3)	$[M+H]^+$	726.5676	2	4	7
832.5887	5.33	PC(18:1/22:6)	$[M+H]^+$	831.5778	3	18	6
810.6072	5.52	PC(18:1/20:3)	$[M+H]^+$	809.5934	4	9	25
854.5728	5.33	PC(18:1/22:5)	$[M+Na]^+$	831.5778	5	2	1
922.5617	5.34	PC(44:9)	$[M+K]^+$	883.6091	12	3	2
856.5728	5.41	PC(18:1/22:5)	$[M+Na]^+$	833.5934	6	5	5

*0.1 min was added to the retention time of each compound.

Figures

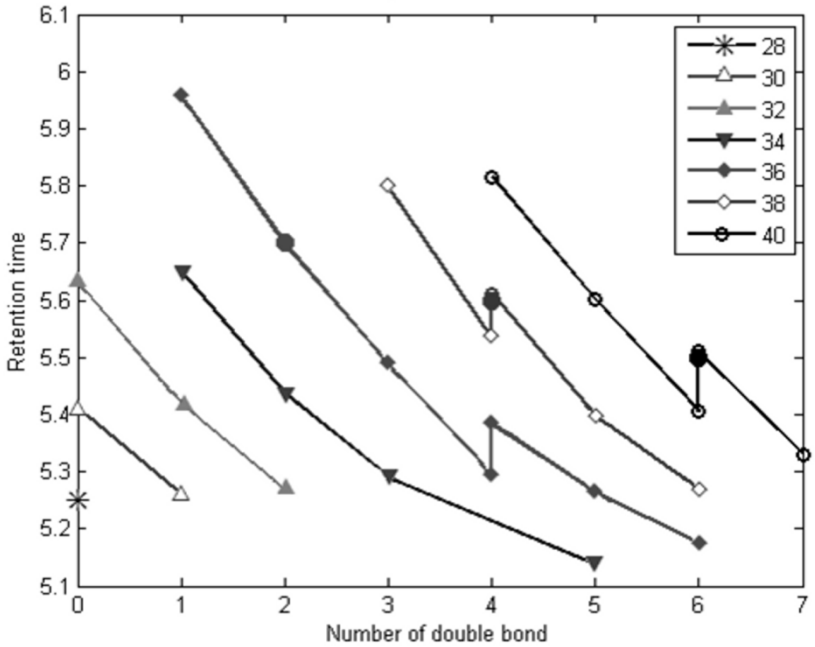


Figure 1

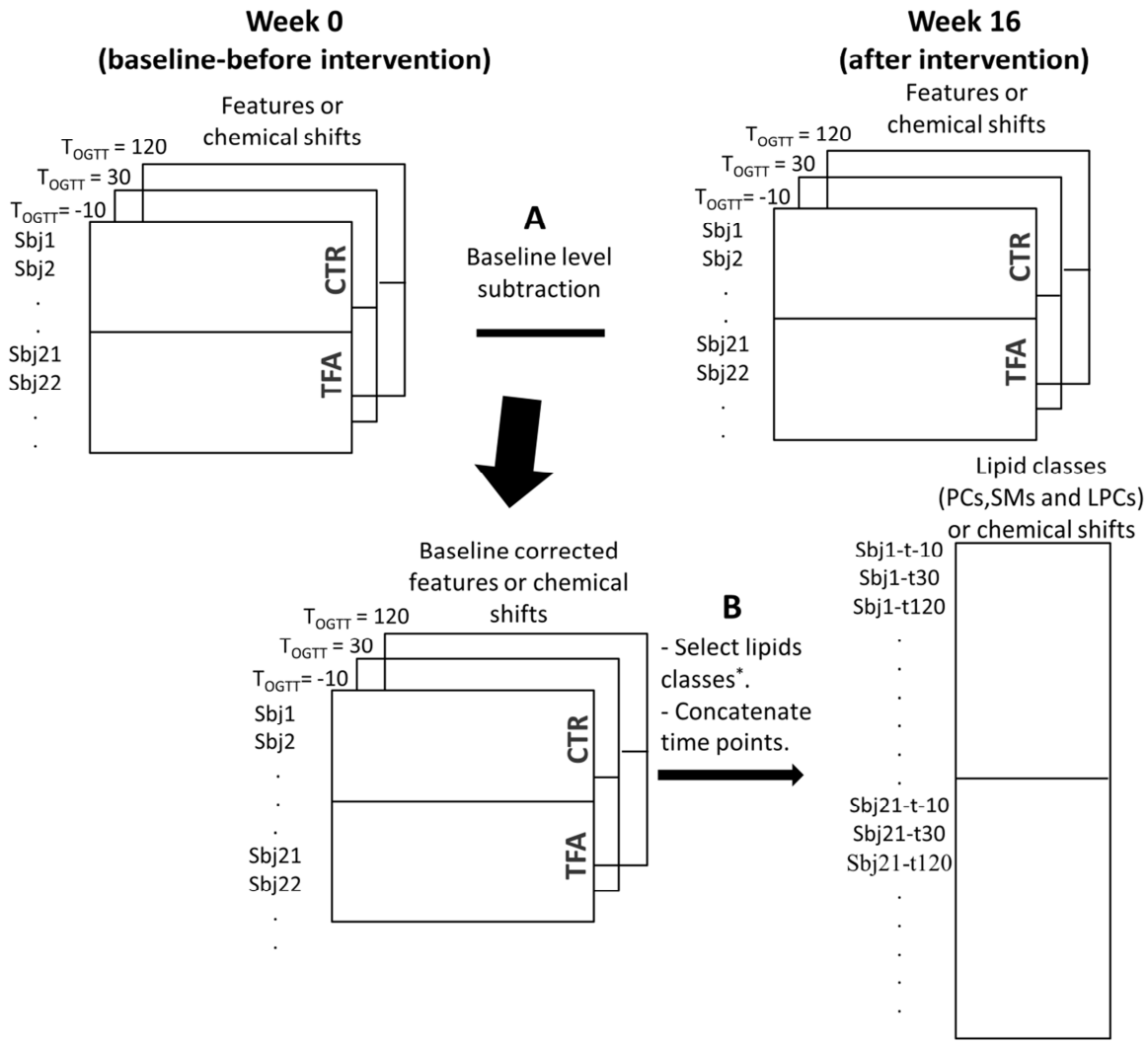

Week 0
(baseline-before intervention)
Features or
chemical shifts
T_OGTT = 120
T_OGTT = 30
T_OGTT= -10
Sbj1
Sbj2
Sbj21
Sbj22
CTR
TFA
A
Baseline level
subtraction
Week 16
(after intervention)
Features or
chemical shifts
T_OGTT = 120
T_OGTT = 30
T_OGTT= -10
Sbj1
Sbj2
Sbj21
Sbj22
CTR
TFA
Baseline corrected
features or chemical
shifts
T_OGTT = 120
T_OGTT = 30
T_OGTT= -10
Sbj1
Sbj2
Sbj21
Sbj22
CTR
TFA
B
- Select lipids
classes*.
- Concatenate
time points.
Lipid classes
(PCs,SMs and LPCs)
or chemical shifts
Sbj1-t-10
Sbj1-t30
Sbj1-t120
Sbj21-t-10
Sbj21-t30
Sbj21-t120

Figure 2

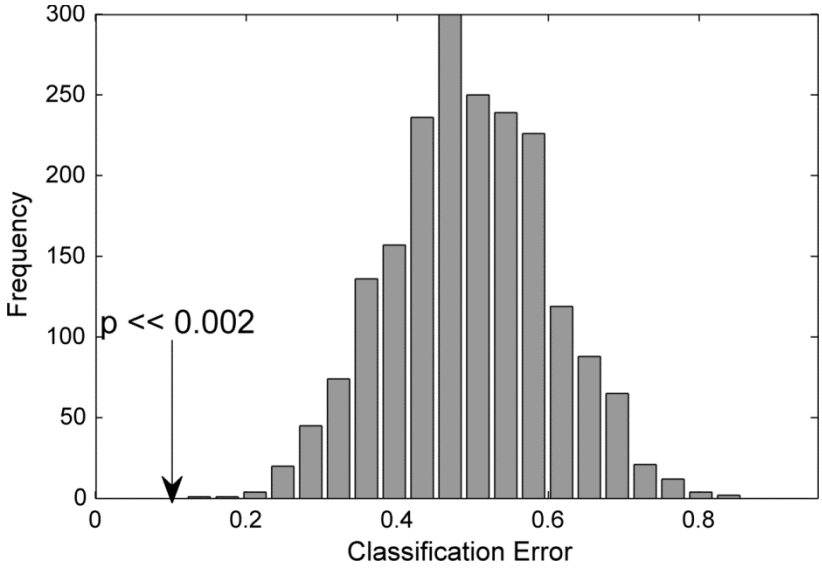

300
250
200
150
100
50
0
Frequency
p << 0.002
0
0.2
0.4
0.6
0.8
Classification Error

Figure 3

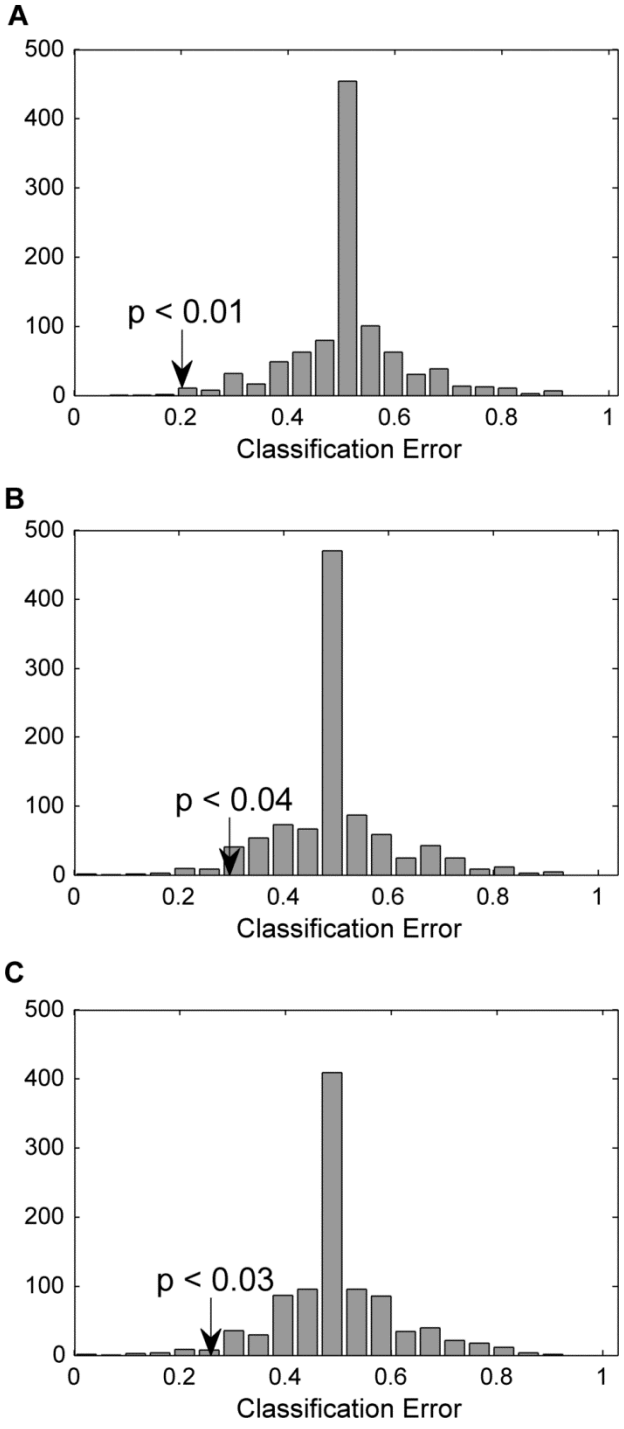

A
500
400
300
200
100
0
p < 0.01
0
0.2
0.4
0.6
0.8
1
Classification Error
B
500
400
300
200
100
0
p < 0.04
0
0.2
0.4
0.6
0.8
1
Classification Error
C
500
400
300
200
100
0
p < 0.03
0
0.2
0.4
0.6
0.8
1
Classification Error

Figure 4

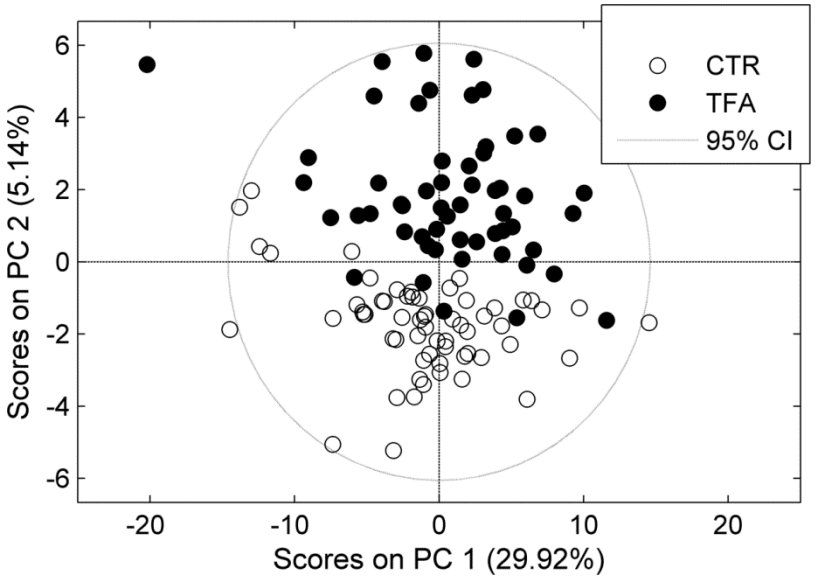


Figure 5

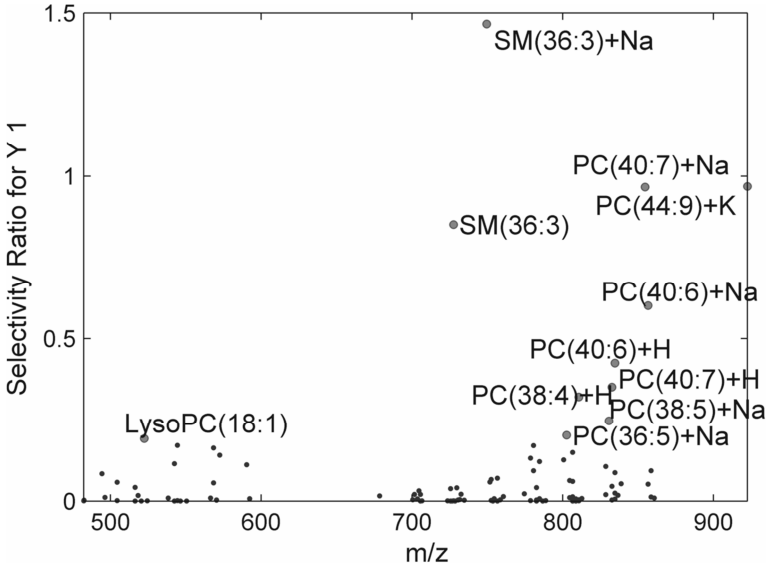


Figure 6

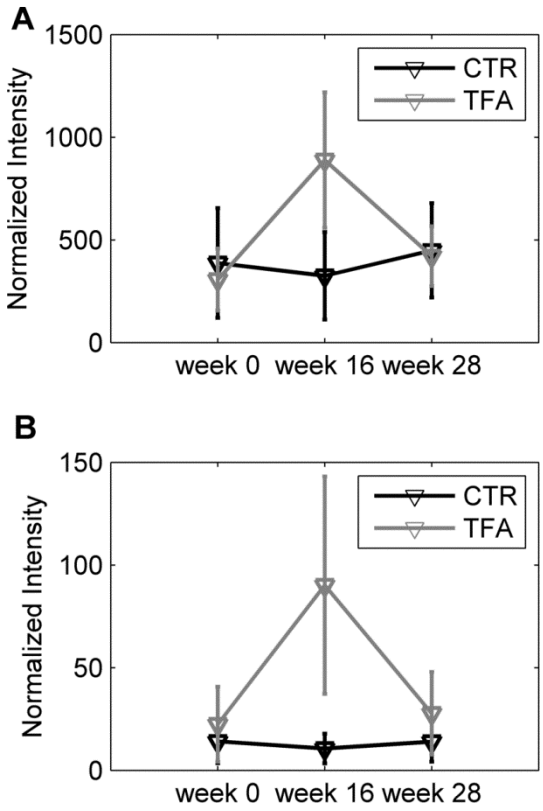

A
1500
1000
500
0
Normalized Intensity
CTR
TFA
week 0
week 16
week 28
B
150
100
50
0
Normalized Intensity
CTR
TFA
week 0
week 16
week 28

Figure 7

PAPER III

Patterns of time since last meal revealed by sparse PCA in an observational LC-MS based metabolomics study.

Gürdeniz G, Hansen L, Rasmussen MA, Olsen A, Christensen J, Acar E, Barri T, Tjønneland A, Dragsted LO

Metabolomics. Submitted.

Patterns of time since last meal revealed by sparse PCA in an observational LC-MS based metabolomics study

Gözde Gürdeniz[1*], Louise Hansen[2], Morten Arendt Rasmussen[3], Evrim Acar[3], Anja Olsen[2], Jane Christensen[2], Thaer Barri[1], Anne Tjønneland[2] and Lars Ove Dragsted[1]

[1]Department of Nutrition, Exercise and Sports, Faculty of Science, University of Copenhagen, DK-1958 Frederiksberg C, Denmark

[2]Danish Cancer Society Research Center, Strandboulevarden 49, DK-2100 Copenhagen, Denmark

[3]Department of Food Science, Faculty of Life Sciences, University of Copenhagen, DK-1958 Frederiksberg C, Denmark

*__Corresponding author:__ Gözde Gürdeniz, Tel.: +45 29176389; Fax: +45 35332483

e.mail: gozg@life.ku.dk

Abstract

In metabolomics studies, liquid chromatography mass spectrometry (LC-MS) provides comprehensive information on biological samples. However, extraction of few relevant metabolites from this large and complex data is cumbersome. To resolve this issue, we have employed sparse principal component analysis (SPCA) to capture the underlying patterns and select relevant metabolites from LC-MS plasma profiles. The study involves a small pilot cohort with 270 subjects where each subject's time since last meal (TSLM) has been recorded prior to plasma sampling. Our results have demonstrated that both PCA and SPCA can capture the TSLM patterns. Nevertheless, SPCA provides more easily interpretable loadings in terms of selection of relevant metabolites, which are identified as amino acids and lyso-lipids.

This study demonstrates the utility of SPCA as a pattern recognition and variable selection tool in metabolomics. Furthermore, amino acids and lyso-lipids are determined as dominating compounds in response to TSLM.

Keywords Metabolomics · SPCA · LC-MS · Plasma · time since last meal · Observational Study

1 Introduction

Based on the recent advances in analytical technologies, metabolomics evolved as a powerful tool, allowing qualification/quantification of hundreds of metabolites in biological samples. Particularly, mass spectrometry (MS)-based methods have been widely employed with the advantage of broader metabolome coverage. However, in MS-based metabolomics, large amount of complex data characterize few samples. Thus, extracting a small set of relevant features from this complex data is challenging.

Principal component analysis (PCA) has been widely used for both dimension reduction and for exploratory analysis of complex datasets (Wold *et al.* 1987). PCA provides features representing the dominating characteristics of the data. Nevertheless, only few LC-MS based metabolomics studies employ PCA for feature selection. One of the reasons for this is that PCA represents each principal component (PC) as the linear combination of all original variables.

Particularly in cases where the number of variables exceeds the number of samples as in metabolomics, this complicates interpretation of PCs such that it becomes unclear to extract only a few relevant features from the many irrelevant ones. To overcome this issue, sparse principal component analysis (SPCA) was introduced by Zou *et al.* (2006a) using the lasso (elastic net) to produce modified PCs with sparse loadings. SPCA allowed the less influential variables to have zero influence on the model by imposing lasso (elastic net) constraint on the regression coefficients. Allen and Maletic-Savatic (2011) have demonstrated applicability of SPCA with non-negativity constraints on an NMR based metabolomics dataset. Sparsity penalty has also been applied for some studies of metabolomics-based multi-block data analysis (Acar *et al.* 2012;Van *et al.* 2011). However, a direct application of SPCA for metabolite selection from LC-MS based metabolomics data has not been shown.

The metabolic responses to food intake and metabolite clearance rates are usually measured by postprandial challenge tests. These may be performed by glucose tolerance tests (OGTT or clamps), lipid challenges, or by specific foods or whole meal challenges, depending on the specific metabolite group of interest. In traditional clinical nutrition, the postprandial tests have been evaluated based on a response pattern of established biomarkers, such as plasma glucose in OGTT and plasma triacylglycerides in lipid challenges. On the other hand, metabolomics offers a more holistic view by comprehensive coverage of metabolites in biological samples. Zhao *et al.* (2009) and Shaham *et al.* (2008) were the first to utilize metabolomics to investigate the physiological changes during an OGTT. They identified major concentration changes in compounds, such as bile acids, that have not been reported previously. A more recent metabolomics study demonstrated the dynamics of the human metabolome in response to diverse challenges, a prolonged fasting period of 36 h, a standard liquid diet, an OGTT, an oral lipid tolerance test, a physical activity test, and a cold pressure stress test (Krug *et al.* 2012). Another study investigated the metabolic perturbation in response to postprandial challenge in a controlled intervention study (Pellis et al. 2012). Our perspective, on the other hand, is to explore whether the plasma LC-MS metabolic profiles reflecting the TSLM can be extracted from a small cohort where subjects had various quality and quantity of food during their last meal. In observational studies, there are many life-style factors which influence each other and that may lead to confounding. Metabolic profiles may be useful to study relationships between diets and metabolism and may eventually reveal disease patterns, yet they may be severely biased by patterns related to sampling such as the TSLM. In the current study, to uncover TSLM-related patterns in a first attempt to disentangle some of the factors affecting the metabolic profiles in the un-controlled observational setting, we employed SPCA as a pattern recognition and feature selection tool and compared its performance with PCA.

2 Materials and Methods

2.1 Subjects

Data from the Danish prospective cohort study, Diet, Cancer and Health, was used for this study. Briefly, a total of 57,053 men and women were enrolled into the cohort between December 1993 and May 1997. Participants were eligible for inclusion if they fulfilled the following criteria: age between 50-64 years, born in Denmark (living in the Copenhagen or Aarhus areas) and no previous cancer diagnosis in the Danish Cancer Registry. A detailed food frequency questionnaire (FFQ) and a lifestyle questionnaire were completed by each participant. Biological and anthropometric measurements were taken, including a non-fasting 30-ml blood sample. The blood samples were centrifuged and divided into fractions of plasma, serum, red blood cells, and buffy coat and stored in 1-ml tubes. All samples were processed and frozen within 2 hours at -20°C and were ultimately transferred to liquid nitrogen vapor (max. -150°C), where they were stored until needed. Citrate was used as the anticoagulant. A thorough description of the data collection procedure has been published elsewhere (Tjonneland et al. 2007).

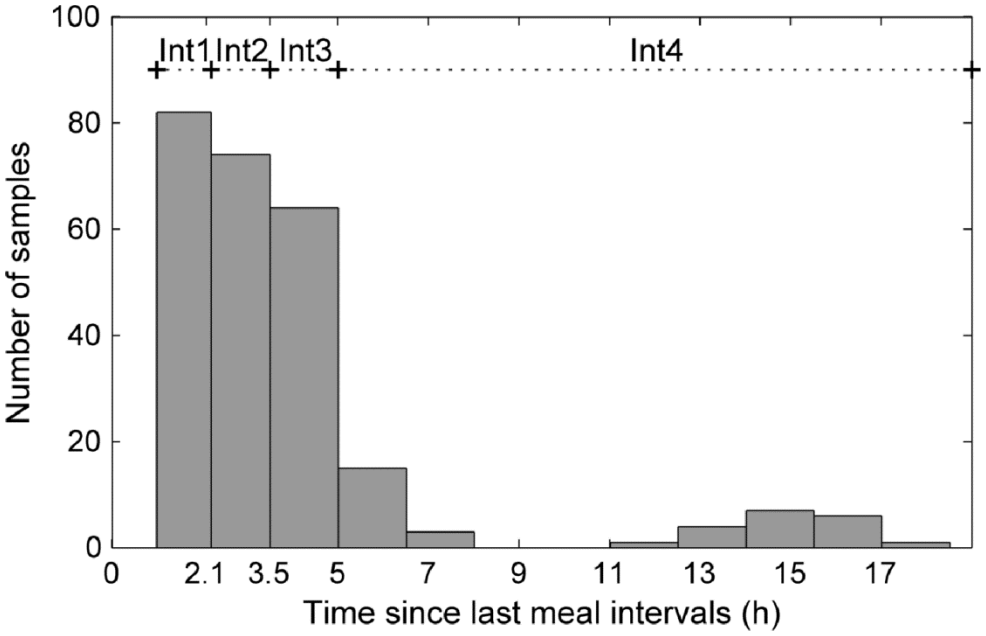


Fig. 1 TSLM distribution of the subjects. Subjects are grouped into four intervals based on the time passed since their last meal has been taken, denoted by int1, int2, int3 and int4.

For the present study, we considered a sub-cohort of female colorectal cancer cases and matched controls. A total of 175 colon and rectal cancer cases among females were identified during a median follow-up of 5.9 years. An identical number of controls was selected randomly.

2.2 Study design

Prior to blood sampling, the number of hours passed since subjects had taken their last meal was recorded. Besides breakfast, lunch and dinner, solid snacks were considered as the last meal but not liquid intake. The subject's whose TSLM has not been reported or with a BMI lower than 20 were excluded. The remaining dataset contained 270 subjects. Fig. 1 represents the TSLM distribution of the subjects. Later, subjects were grouped into four time intervals; 1 to 2.1 h, 2.1 to 3.5 h, 3.5 to 5 h and 5 to 18.6 h (Fig. 1).

2.3 LC-MS analysis

Plasma protein precipitation was performed as described earlier (Gürdeniz et al. 2012). Samples were randomized and placed in 96-well plates. An ultra-performance liquid chromatography (UPLC) system coupled to quadruple time-of-flight (Premier QTOF) mass spectrometer (Waters Corporation, Manchester, UK) was used for sample analyses. Each sample (10 µL) was injected into the UPLC equipped with a 1.7µm C18 BEH column (Waters) operated with a 6-min linear gradient from 0.1% formic acid in water to 0.1% formic acid in 20% acetone: 80% acetonitrile. The capillary probe voltage was set at 2.8 and 3.2 kV for negative (ESI-) and positive (ESI+) electrospray ionization (ESI) modes, respectively. In the ESI- mode, desolvation gas temperature 400°C, cone voltage 40 V, and Ar collision gas energy 6.1 V were used. In the ESI+ mode, we used the same settings except for collision energy of 10 V). Samples of blank (0.1% formic acid) and metabolomics standard mixture of 44 metabolites were analyzed after every 50 samples during the sample sequence.

Amino acids (i.e.tyrosine, leucine/isoleucine, tryptophan, phenylalanine), lysophosphatidylcholines (LPC18:1, LPC18:0, LPC17:0, LPC16:0, and C18:2) and lysophosphatidylethanolamine (LPE18:1) were identified by using an in-house metabolite database containing retention time information and MS spectra of reference substances (Gürdeniz et al. 2012). LPC15:0, LPC18:3, LPC20:2 and LPC20:3 were putatively identified based on spectra and retention time relative to the identified LPCs and LPEs.

2.4 Data pre-processing and pre-treatment

The centroided raw data was converted to an intermediate netCDF format with the DataBridgeTM utility provided with the MassLynx software. MZmine 2.7 (Pluskal et al. 2010) was employed for data preprocessing including the following steps: mass detection, chromatogram builder, chromatogram deconvolution (local minimum search), isotopic peaks grouper, peak alignment (using join aligner) and gap filling. The term 'feature' is used to refer to a chemical compound with a specific retention time and mass over charge ratio (m/z) throughout this paper.

MZmine-preprocessed data was imported into MATLAB R2012a (ver. 7.17.0.739). Peak filtering was applied based on two criteria. First, if a feature has a reasonable peak area (>60) in the first blank sample in at least one of the four analytical batches, the feature is removed from the entire set. Second, if a feature has a peak area lower than 5 (considered as noise level or gap filling errors), in more than 60% of the samples within every sample group (TSLM intervals), the feature is excluded (percent rule, (Bijlsma et al. 2006)). Afterwards, the few remaining missing entries were filled with a number within a random range of 0-70 % of the smallest value for each feature.

Systematic error caused by experimental conditions was corrected based on two normalization approaches. Initially, samples were normalized to unit length to correct for decreasing instrumental response during sample acquisition batches. Second, to remove inter-batch variation, each feature was normalized within each batch with the overall mean of its recordings throughout the entire set. This approach is justified by the randomization of the samples between the plates prior to analysis.

Data preprocessing was automated by a MATLAB function which can be provided upon request.

2.5 Data analysis

Autoscaled data was subjected to PCA (Wold et al. 1987) and SPCA. SPCA can be formulated as a penalized optimization problem with the main objective being a minimization problem similar to PCA with L1 norm penalties imposed on the parameters, in this case the loading vectors, to achieve sparsity. The formulation of SPCA can be shown as:

$$\text{argmin}\ (\|X - TP^T\|_F^2)$$

$$\text{subject to } \|p_i\|_1 \leq c \text{ and } \|p_i\|_2^2 = 1, \text{ for } i=1,\ldots,k$$

where X (n x p), is the data matrix, $\|p_i\|_1$ is the sum of absolute values (L1 norm) of the i'th column of the loading matrix P, and T is the scores matrix. The tuning parameter c is a positive penalty parameter bounding the sum of absolute values of the normalized loading vector ($\|p_i\|_1 \leq c$). It controls the degree of sparsity in the loading vector,

i.e., the number of nonzero loadings. A meaningful sparse solution can be found when the parameter is chosen between 1 (univariate decomposition, one variable pr. component) and the square root of the number of variables (unconstrained PCA decomposition) (Rasmussen and Bro 2012).

The calculation of the entire set of components was done simultaneously by iterating between scores and loadings. An alternating least squares-based approach with induced L1 norm penalty was used for component estimation. Relative change in function values is used as a stopping condition and it is set to 10-10. Each step of the alternating procedure is a convex optimization problem, and hence provides the global minima. However, the entire problem is not convex and in order to avoid local minimum issues, we initialized multiple times with random loadings. Unlike PCA, SPCA does not impose othogonality constraint between components.

SPCA requires the selection of the number of components and the degree of sparsity. The number of components was varied from 4 to 20 incrementing by 3. The tuning parameter c was varied from 1.5 to 6 with 0.5 intervals. In this study, we evaluated SPCA scores and chose the sparsity level and the number of components that sufficiently explained the TSLM patterns.

PCA is implemented in PLS_Toolbox (ver. 6.5.1, Eigenvector Research, Inc., MA, US) for MATLAB® R2012a (ver. 7.17.0.739). SPCA was conducted using a freely available SPCA algorithm from (http://models.life.ku.dk/sparsity) for MATLAB together with a web tutorial, describing details of the algorithm(Rasmussen and Bro 2012).

3 Results and Discussions

Initially we would like to mention the reasons why an unsupervised method has been selected to investigate this data. For instance, a regression method such as PLS (Partial Least Squares) could have been a natural choice to predict TSLM patterns or a classification approach such as PLS-DA (Partial Least Squares Discriminant Analysis) could have been applied using the determined TSLM intervals. However, the subjects had their habitual diets varying in quantity as well as quality of foods and drinks during each meal. Furthermore, many of the subjects additionally had a drink which was not further specified (e.g., water, juice, coffee, etc.) independently of their recorded TSLM. As a matter of fact, for many subjects TSLM was approximate rather than a very certain value. Thus, using PLS with approximate labels corresponding to TSLM was not very accurate. Furthermore, we have attempted to group subjects into intervals based on their TSLM (Fig. 1) in order to ease the visualization. However, the interval boundaries were not clear, as the metabolic response to TSLM is an ongoing process. Therefore, classification-based methods like PLS-DA are not very appropriate. On the other hand, an explorative data analysis method aims to capture underlying dominating patterns

without any predictor variables. Thus, PCA-based methods provide the necessary basis to explore all systematic variation and especially TSLM-related patterns.

3.1 Interpretation of models from SPCA vs. PCA

A total of 1199 features in ESI+ mode and 1324 in ESI- mode were detected by MZmine. After exclusion of noise and irrelevant features as described in the section on data pre-processing, 547 and 681 features for ESI+ and ESI- modes, respectively, remained for data analysis. Furthermore, 10 samples in ESI+ and 13 samples in ESI- were excluded as outliers in PCA. These samples also had either instrumental (i.e. very low response) or sample preparation issues (i.e. too little sample left for analysis).

PCA captured the slight TSLM trend by PC1 (11.3%) and PC4 (3.2%) for ESI- mode, as shown in Fig. 2a and 2c. The TSLM trends are not very obvious, which was expected as each subject's last meal differed in quality and quantity. Based on Fig. 2b and 2d the loading plots are difficult to interpret. LPCs and LPEs for PC1 (Fig. 2b) and a group of amino acids for PC4 (Fig. 2d) tend to have relatively higher loadings (i.e., coefficients with high magnitudes), yet those are not clearly distinguishable from many others.

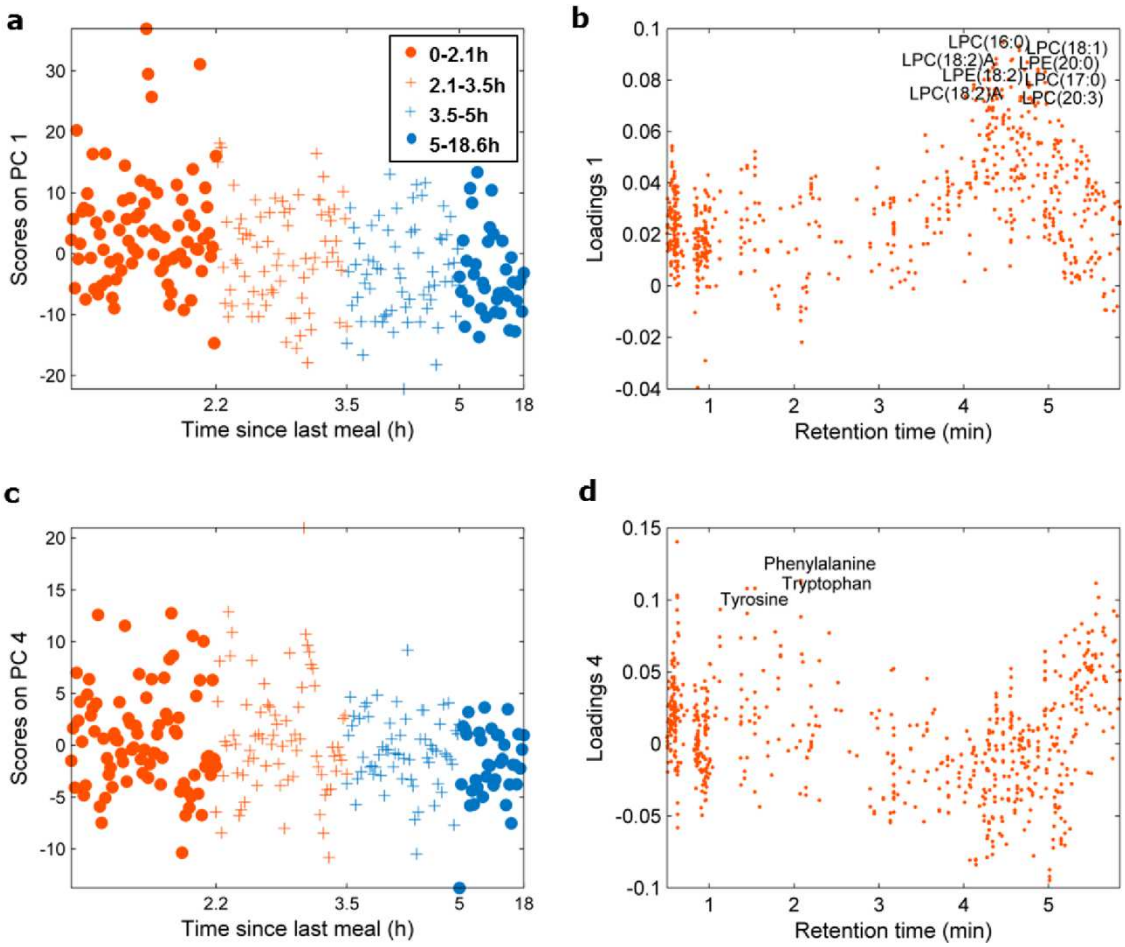


Fig. 2 TSLM vs. scores on PC 1 (a) and PC 4 (c) scores. Retention time vs. PC 1 (c) and PC 4 (d) loadings. (Data acquired in ESI- mode)

Among the different number of components and sparsity levels considered, the SPCA model with 14 components and sparsity level of 2.5 i. e., ($\|p_i\|_1 \leq 2.5$) was determined to be the appropriate model capturing the TSLM trend in ESI- mode. SPCA score plots for SPC2 (0.82%), SPC6 (0.78%) and SPC12 (0.7%) (Fig. 3a, 3c, 3e) illustrate similar trends for TSLM patterns compared to PC1 (Fig. 2a). However, unlike PCA loadings (Fig. 2b), we can clearly see the compounds reflecting the patterns from SPCA loadings (Fig. 3b, 3d, 3f). Furthermore, the TSLM trend is also described by scores on SPC14 (0.62%) (Fig. 4a), even slightly better than the corresponding PC4 (Fig. 2c). It seems PC4 is reflecting also other irrelevant patterns (Fig. 2d). SPC14, on the other hand, is able to extract the TSLM related part.

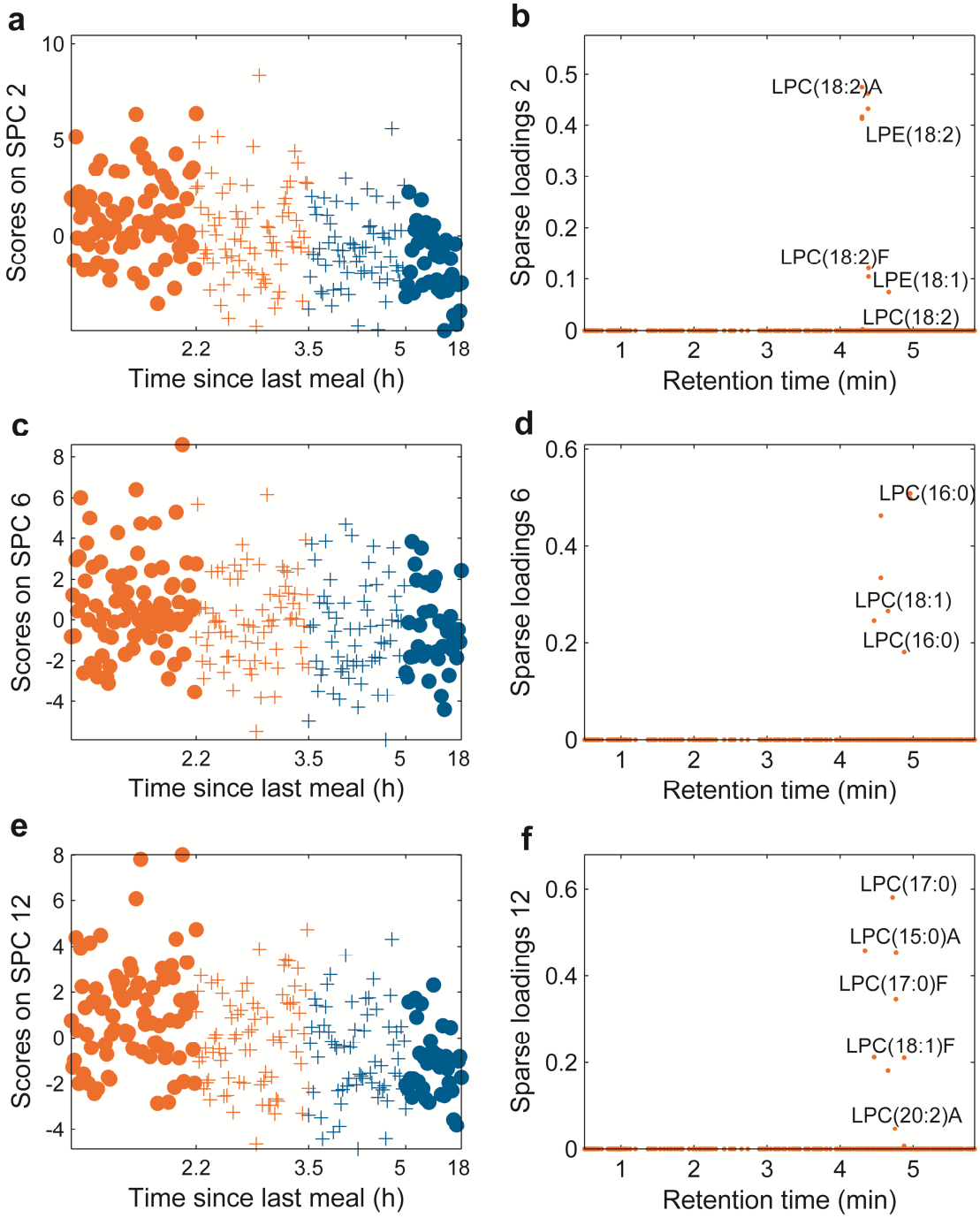


Fig. 3 TSLM vs. scores on SPC 1 (a). Retention time vs. SPC 1 (b) loadings. (Data acquired in ESI- mode)

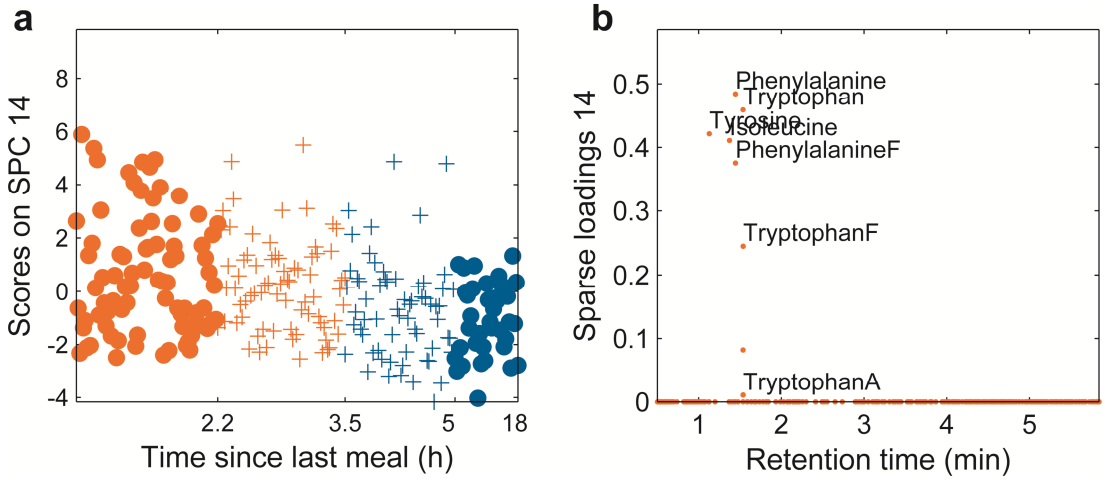


Fig. 4 TSLM vs. scores on SPC 3 (a), SPC 9 (c) and SPC 13 (e). Retention time vs. SPC 3 (b), SPC 9 (d) and SPC 13 (f) loadings. Spearman correlations between components are $r_{2\text{-}6}$= 0.69, $r_{2\text{-}12}$ = 0.72 and $r_{2\text{-}12}$ = 0.73 [CI = 95%]) (Data acquired in ESI- mode)

The explained variation for SPCA is much lower than PCA, yet this was not surprising especially when the level of sparsity is low and large number of component is included.

Scores on SPC2, SPC6 and SPC12 are highly correlated with each other (Fig. 3). When SPCA is performed with decreased degree of sparsity ($\|p_i\|_1 \leq 4$ for i=1,..,14) the same pattern is explained with only one component. Nevertheless, the pattern explained by SPC1 disappeared. This problem can be solved by further improvement of the SPCA algorithm using component-wise sparsity penalties.

For the data acquired in ESI+ mode, we used SPCA with eight components and a sparsity degree of four ($\|p_i\|_1 \leq 4$, for i=1,..,8). The TSLM is captured by SPC8 (1.7%) as shown in Fig. 5. Using lower than 8 components and decreasing the degree of the sparsity penalty, the model did not reveal TSLM related patterns.

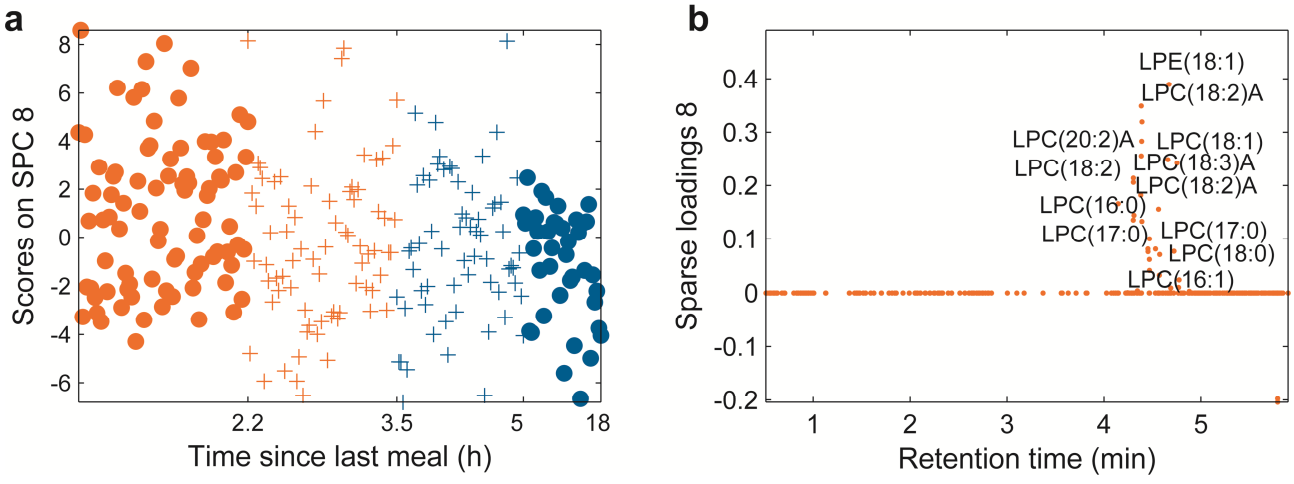


Fig. 5 TSLM vs. scores on SPC1 (a). Retention time vs. SPC 1 (b) loadings. (Data acquired in ESI+ mode)

Our results clearly show that SPCA outperforms PCA by providing more easily interpretable results, yet with the explained variance trade off. SPCA encouraged features with negligible contributions in standard PCA to have zero loadings. Thus, the significant metabolites could be identified easily by the loadings of the SPCA model.

It was interesting to see the groups of metabolites reflecting TSLM patterns, i.e., amino acids and LPCs/LPEs, appearing in different components. The correlation coefficients between the SPC scores corresponding to amino acids and LPCs/LPEs is 0.3 (Spearman's correlations [CI = 95%]), meaning that they are not correlated. This can provide another perspective for interpretation such that the compounds from different chemical groups behaved differently in relation to TSLM. Once one variable is selected, SPCA tends to select a group of variables correlated with that one. In this case, the metabolites from the same chemical class, LPCs/LPEs and amino acids, were correlated within their group and had group specific influence on TSLM patterns (Fig. 3). Rasmussen & Bro (2012) had similar findings where selection of inter-correlated variables was favored by SPCA from proteomics based MS data.

In this study, we have tested the performance of SPCA as a feature selection tool for LC-MS based metabolomics data. The main obstacle is selection of the optimum number of components and the sparsity tuning parameter. In some other studies, cross validation (Rasmussen and Bro 2012) and Bayesian information criteria (Allen and Maletic-Savatic 2011) have been suggested for selection of the sparsity penalty. However, for our problem class boundaries were not very clear, which makes these solutions unsuitable. In this case, we selected the minimum number of components that we can observe TSLM related patterns, yet as a future perspective in-depth sensitivity analysis can be performed to select number of components and sparsity level.

3.2 Metabolic reflections of TSLM

The overview of identified compounds reflecting TSLM in ESI- and ESI+ mode is given in Table 1. As shown in Table 1, SPCA revealed compounds from two different chemical classes, amino acids and lyso-lipids (LPCs and LPEs) as reflecting TSLM patterns.

Table 1 The identified plasma metabolites reflecting the TSLM. The ESI mode in which the metabolites have been found as significant by SPCA is indicated.

Coumpound Name	ESI Mode
Tyrosine	ESI-,ESI+
Leucine/Isoleucine	ESI-,ESI+
Phenylalanine	ESI-,ESI+

Tryptophan	ESI-,ESI+
LPC (18:3)A	ESI+
LPC (16:1)	ESI+
sn2-LPC (18:2)	ESI-,ESI+
sn2-LPE (18:2)	ESI-
LPC (15:0)	ESI-,ESI+
sn1-LPC (18:2)	ESI+
sn1-LPE (18:2)	ESI-,ESI+
sn2-LPC (16:0)	ESI-,ESI+
LPC (20:3)	ESI+
sn1-LPC (16:0)	ESI-
LPC (18:1)	ESI-,ESI+
LPC (17:0)	ESI+
LPE (18:1)	ESI-,ESI+
LPC (20:2)A	ESI-,ESI+
LPC (17:0)	ESI-,ESI+
sn2-LPC (18:0)	ESI-,ESI+
sn1-LPC (18:0)	ESI-

The scores of the components describing each group are observed in different components meaning that the TSLM responses of lyso-lipids and amino acids diverge. The reflected amino acids are four essential amino acids, phenylalanine, leucine/isoleucine and tryptophan and one nonessential amino acid, tyrosine (Fig. 3B). Amino acids increase with recent food intake and decrease until 18h after the last meal has been taken (Fig. 3A) yet, there is a large variation particularly within the 1 to 2.1 h cluster. In fact, it has been shown that plasma amino acid concentration can fluctuate widely in response to many factors such as type of food consumed (Boirie et al. 1997;Wurtman et al. 1968), obesity (Shaham et al. 2008) and diabetes (Wang et al. 2011). Thus, considering the varying characteristics of the subjects as well as the qualitative and quantitative differences of the last meals taken by the subjects in this study, the large variation is not surprising. Nevertheless, in a more controlled intervention study, Pellis et al. (2012) observed approximately the same TSLM responses for a wide range of amino acids (0-6h). The higher amino acid concentration during the first 1-2h is linked to the compositions of the protein source present in the last meal. The declining trend after 2h is related to insulin stimulation of amino acid uptake from the plasma to liver and muscle for protein synthesis (Fukagawa et al. 1985). The decrease in plasma branched-chain amino acids has been shown to start earlier after a glucose challenge without a concomitant protein load, starting at 30 min, which is most likely related to a strong, immediate impact of glucose on insulin secretion (Shaham et al. 2008). Prolonged fasting causes a later increase in branched chain amino acids levels starting at 10-20 hours due to increased proteolysis (Rubio-Aliaga et al. 2011). We

did not observe this phenomenon in the current study although some subjects had their last meal even as much as 12-18 hours earlier; those were very few and they were not subjected to prolonged-fasting as TSLM is determined without considering the last drink. All of these participants reported having consumed a drink but we do not have records of what the participants were drinking in this study and some may have taken nutritious drinks such as milk.

Both sn-1 and sn-2 isomers of a wide range of LPCs and LPEs exhibit a steady decrease with increasing TSLM (Fig. 3a, 3c, 3e). LPC is a plasma lipid that has been recognized as an important cell signaling molecule and it is produced by the action of phospholipases A1 and A2, by endothelial lipase or by lecithin-cholesterol acyltransferase (LCAT) which transfers one of the fatty acids from phosphatidylcholine to cholesterol. LCAT has a well-known function in catalyzing the transfer of fatty acids to free cholesterol in plasma for the formation of cholesteryl esters (Schmitz and Ruebsaamen 2010). In our previous rat study, we have seen a wide range of LPCs and a few LPEs decrease in the fasted state compared to the fed state, in support of our findings in the current study (Gürdeniz et al. 2012). In a study of prolonged fasting (12h to 36h) a reduction in plasma LPCs (C18:0, C18:1 and C18:2) was also observed as shown by (Rubio-Aliaga et al. 2011). Another study investigating the effects of an oral glucose tolerance test, plasma LPCs (C16:0, C18:0, C16:1, C18:1, and C18:2) increased from fasting levels up to 1h with a slight further increase until 2h. (Pellis et al. 2012) in a postprandial challenge test observed an increase in one specific LPC, (C18:2), with a somewhat longer time course of 1 to 6h. In their study fasting for 1-2h response for LPC (18:2), was not clear. This discrepancy might be related to the specific challenge meal that the subjects were given in the intervention study.

LPCs have been related to increased insulin resistance (Han et al. 2011); however, their effect compared to other related lipids such as PCs and SMs has not been reported. A recent lipidomics study demonstrated a reduction of fasting plasma LPC levels in obese and type 2 diabetic obese subjects stronger than for other PCs and SMs (Barber et al. 2012). These findings suggest that LPCs have an important role in insulin regulation. A further investigation of plasma LPC responses to a postprandial challenge test on diabetic obese subjects can reveal if the reduction is specific to the fasting state or if the time course response is affected. The unsaturated LPCs have been found also to pass the blood-brain barrier and to be important vehicles for delivering unsaturated lipids to the brain (Sekas et al. 1985). We speculate that the high level of unsaturated LPCs in the postprandial state of healthy individuals might be a part of the satiety signaling system which is malfunctioning in obesity.

Although LPEs indicate similar trends to LPCs, the previous discussions were attributed to LPCs. The reason is that not so much is known regarding to physiological functions of the plasma LPEs. LPEs, in analogy to LPCs, can be generated from phosphatidylethanolamine (PE), a component of the cell membrane via a phospholipase A-type reaction

(Makide et al. 2009). In this study, LPCs and LPEs seem to have similar functions based on their parallel response to TSLM and LPEs have the same fatty acyl groups as LPCs. The specific chain length and saturation level of LPCs and LPEs in plasma may primarily be related to the distribution of fatty acyl chains in the food consumed most recently.

Conclusions

The results here suggest that SPCA is able to capture TSLM patterns with loadings which are much easier to interpret compared to PCA. Also, it is able to extract inter-correlated variables from the same biochemical classes. Based on these results we believe SPCA can be potentially applied for variable selection in LC-MS based metabolomics studies.

In spite of the variability and the uncontrolled nature of an observational setting, amino acids and LPCs/LPEs emerged as TSLM reflecting patterns in this relatively small pilot study. In larger studies within observational settings it should be possible also to disentangle the influence of factors such as diabetes, waist circumference, or BMI and possibly to find cancer-related patterns. We have recently analyzed more than 3000 samples from the DCH cohort and will proceed to analyze at the metabolome level the confounding effect of recent food intake, food intake patterns and current health status.

Acknowledgements

This work is carried out as a part of the research program of the Danish Obesity Research Centre (DanORC, see www.danorc.dk), funded by the Danish Strategic Research Council. This work is also supported by Nordic Centre of Excellence (NCoE) programme (Systems biology in controlled dietary interventions and cohort studies—SYSDIET, P no. 070014).

References

Acar, E., Gurdeniz, G., Rasmussen, M. A., Rago, D., Dragsted, L. O., and Bro, R. (2012) Coupled Matrix Factorization with Sparse Factors to identify Potential Biomarkers in Metabolomics. Proceedings of the 2012 IEEE International Conference on Data Mining Workshops

Allen, G.I. & Maletic-Savatic, M. (2011) Sparse non-negative generalized PCA with applications to metabolomics. Bioinformatics, 27, 3029-3035

Barber, M.N., Risis, S., Yang, C., Meikle, P.J., Staples, M., Febbraio, M.A., & Bruce, C.R. (2012) Plasma lysophosphatidylcholine levels are reduced in obesity and type 2 diabetes. PLoS One, 7, e41456

Bijlsma, S., Bobeldijk, I., Verheij, E.R., Ramaker, R., Kochhar, S., Macdonald, I.A., van, O.B., & Smilde, A.K. (2006) Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. Anal. Chem., 78, 567-574

Boirie, Y., Dangin, M., Gachon, P., Vasson, M.P., Maubois, J.L., & Beaufrere, B. (1997) Slow and fast dietary proteins differently modulate postprandial protein accretion. Proceedings of the National Academy of Sciences of the United States of America, 94, 14930-14935

Fukagawa, N.K., Minaker, K.L., Rowe, J.W., Goodman, M.N., Matthews, D.E., Bier, D.M., and Young, V.R. (1985) Insulin-Mediated Reduction of Whole-Body Protein Breakdown - Dose-Response Effects on Leucine Metabolism in Post-Absorptive Men. Journal of Clinical Investigation, 76, 2306-2311

Gürdeniz, G., Kristensen, M., Skov, T., and Dragsted, L.O. (2012) The Effect of LC-MS Data Preprocessing Methods on the Selection of Plasma Biomarkers in Fed vs. Fasted Rats. Metabolites, 2, 77-99

Han, M.S., Lim, Y.M., Quan, W., Kim, J.R., Chung, K.W., Kang, M., Kim, S., Park, S.Y., Han, J.S., Park, S.Y., Cheon, H.G., Rhee, S.D., Park, T.S., and Lee, M.S. (2011) Lysophosphatidylcholine as an effector of fatty acid-induced insulin resistance. Journal of Lipid Research, 52, 1234-1246

Krug, S., Kastenmuller, G., Stuckler, F., Rist, M.J., Skurk, T., Sailer, M., Raffler, J., Romisch-Margl, W., Adamski, J., Prehn, C., Frank, T., Engel, K.H., Hofmann, T., Luy, B., Zimmermann, R., Moritz, F., Schmitt-Kopplin, P., Krumsiek, J., Kremer, W., Huber, F., Oeh, U., Theis, F.J., Szymczak, W., Hauner, H., Suhre, K., and Daniel, H. (2012) The dynamic range of the human metabolome revealed by challenges. Faseb Journal, 26, 2607-2619

Makide, K., Kitamura, H., Sato, Y., Okutani, M., and Aoki, J. (2009) Emerging lysophospholipid mediators, lysophosphatidylserine, lysophosphatidylthreonine, lysophosphatidylethanolamine and lysophosphatidylglycerol. Prostaglandins Other Lipid Mediat., 89, 135-139

Pellis, L., van Erk, M.J., van Ommen, B., Bakker, G.C.M., Hendriks, H.F.J., Cnubben, N.H.P., Kleemann, R., van Someren, E.P., Bobeldijk, I., Rubingh, C.M., and Wopereis, S. (2012) Plasma metabolomics and proteomics profiling after a postprandial challenge reveal subtle diet effects on human metabolic status. Metabolomics, 8, 347-359

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

Rasmussen, M.A. and Bro, R. (2012) A tutorial on the Lasso approach to sparse modeling. Chemometrics and Intelligent Laboratory Systems, 119, 21-31

Rubio-Aliaga, I., de Roos, B., Duthie, S.J,, Crosley, L.K., Mayer, C., Horgan, G., Colquhoun, I.J., Le Gall. G., Huber, F., Kremer, W., Rychlik, M., Wopereis, S., van Ommen, B., Schmidt, G., Heim, C., Bouwman, F.G., Mariman, E.C., Mulholland, F., Johnson, I.T., Polley, A.C., Elliott, R.M., and Daniel, H. (2011) Metabolomics of prolonged fasting in humans reveals new catabolic markers. Metabolomics, 7, 375-387

Schmitz, G. and Ruebsaamen, K. (2010) Metabolism and atherogenic disease association of lysophosphatidylcholine. Atherosclerosis, 208, 10-18

Sekas, G., Patton, G.M., Lincoln, E.C., and Robins, S.J. (1985) Origin of plasma lysophosphatidylcholine: evidence for direct hepatic secretion in the rat. J. Lab Clin. Med., 105, 190-194

Shaham, O., Wei, R., Wang, T.J., Ricciardi, C., Lewis, G.D., Vasan, R.S., Carr, S.A., Thadhani, R., Gerszten, R.E., and Mootham, V.K. (2008) Metabolic profiling of the human response to a glucose challenge reveals distinct axes of insulin sensitivity. Molecular Systems Biology, 4, 214

Tjonneland, A., Olsen, A., Boll, K., Stripp, C., Christensen, J., Engholm, G., and Overvad, K. (2007) Study design, exposure variables, and socioeconomic determinants of participation in Diet, Cancer and Health: a population-based prospective cohort study of 57,053 men and women in Denmark. Scand. J. Public Health, 35, 432-441

Van, D.K., Wilderjans, T.F., van den Berg, R.A., Antoniadis, A., and Van, M.I. (2011) A flexible framework for sparse simultaneous component based data integration. BMC Bioinformatics, 12, 448

Wang, T.J., Larson, M.G., Vasan, R.S., Cheng, S., Rhee, E.P., McCabe, E., Lewis, G.D., Fox, C.S., Jacques, P.F., Fernandez, C., O'Donnell, C.J., Carr, S.A., Mootha, V.K., Florez, J.C., Souza, A., Melander, O., Clish, C.B.,

and Gerszten, R.E. (2011) Metabolite profiles and the risk of developing diabetes. Nature Medicine, 17, 448-U83

Wold, S., Esbensen, K., and Geladi, P. (1987) Principal Component Analysis. Chemom. Intell. Lab. Syst., 2, 37-52

Wurtman, R.J., Rose, C.M., Chou, C., and Larin, F.F. (1968) Daily rhythms in the concentrations of various amino acids in human plasma. N. Engl. J Med., 279, 171-175

Zhao, X., Peter, A., Fritsche, J., Elcnerova, M., Fritsche, A., Haring, H.U., Schleicher, E.D., Xu, G., and Lehmann, R. (2009) Changes of the plasma metabolome during an oral glucose tolerance test: is there more than glucose to look at? Am. J Physiol Endocrinol. Metab, 296, E384-E393

Zou, H., Hastie, T., and Tibshirani, R. (2006a) Sparse principal component analysis. Journal of Computational and Graphical Statistics, 15, 265-286

Zou, H., Hastie, T., and Tibshirani, R. (2006b) Sparse principal component analysis. Journal of Computational and Graphical Statistics, 15, 265-286

Supplemental Material

2012 IEEE 12th International Conference on Data Mining Workshops

Coupled Matrix Factorization with Sparse Factors to Identify Potential Biomarkers in Metabolomics

Evrim Acar*, Gözde Gürdeniz†, Morten A. Rasmussen*, Daniela Rago†, Lars O. Dragsted† and Rasmus Bro*
*Department of Food Science, Faculty of Science, University of Copenhagen
Email: {evrim, mortenr, rb}@life.ku.dk
†Department of Human Nutrition, Faculty of Science, University of Copenhagen
Email: {gozg, dara, ldra}@life.ku.dk

***Abstract*—Metabolomics focuses on the detection of chemical substances in biological fluids such as urine and blood using a number of analytical techniques including Nuclear Magnetic Resonance (NMR) spectroscopy and Liquid Chromatography-Mass Spectroscopy (LC-MS). Among the major challenges in analysis of metabolomics data are (i) joint analysis of data from multiple platforms and (ii) capturing easily interpretable underlying patterns, which could be further utilized for biomarker discovery. In order to address these challenges, we formulate joint analysis of data from multiple platforms as a coupled matrix factorization problem with sparsity constraints on the factor matrices. We develop an all-at-once optimization algorithm, called CMF-SPOPT (Coupled Matrix Factorization with SParse OPTimization), which is a gradient-based optimization approach solving for all factor matrices simultaneously. Using numerical experiments on simulated data, we demonstrate that CMF-SPOPT can capture the underlying sparse patterns in data. Furthermore, on a real data set of blood samples collected from a group of rats, we use the proposed approach to jointly analyze metabolomic data sets and identify potential biomarkers for apple intake.**

***Keywords*-Coupled matrix factorization; sparsity; gradient-based optimization; missing data; metabolomics**

I. Introduction

With the ability to collect massive amounts of data as a result of technological advances, we are commonly faced with data sets from multiple sources. For instance, metabolomics studies focus on detection of a wide range of chemical substances in biological fluids such as urine and plasma using a number of analytical techniques including Liquid Chromatography-Mass Spectroscopy (LC-MS) and Nuclear Magnetic Resonance (NMR) Spectroscopy. NMR, for example, is a highly reproducible technique and powerful in terms of quantification. LC-MS, on the other hand, allows the detection of many more chemical substances in biological fluids but only with lower reproducibility. These techniques often generate data sets that are complementary to each other [1]. Data from these complementary methods, when analyzed together, may enable us to capture a larger proportion of the complete metabolome belonging to a specific biological system. However, currently, there is a significant gap between data collection and knowledge extraction: being able to collect a vast amount of relational data from multiple sources, we cannot still analyze these data sets in a way that shows the overall picture of a specific problem of interest, e.g., exposure to a specific diet.

To address this challenge, data fusion methods have been developed in various fields focusing on specific problems of interest, e.g., missing link prediction in recommender systems [2], and clustering/community detection in social network analysis [3], [4]. Data fusion has also been studied in metabolomics mostly with a goal of capturing the underlying patterns in data [5] and using the extracted patterns for prediction of a specific condition [6] (see [1] for a comprehensive review on data fusion in omics).

Matrix factorizations are the common tools in data fusion studies in different fields. An effective way of jointly analyzing data from multiple sources is to represent data from different sources as a collection of matrices. Subsequently, this collection of matrices can be jointly analyzed using collective matrix factorization methods [7], [8].

Nevertheless, applicability of available data fusion techniques is limited when the goal is to identify a limited number of variables, e.g., a few metabolites as potential biomarkers. Matrix factorization methods, without specific constraints on the factors, would reveal dense patterns, which are difficult to interpret. Therefore, motivated by the applications in metabolomics, in this paper, we formulate data fusion as a coupled matrix factorization model with penalties to enforce sparsity on the factors in order to capture sparse patterns. Our contributions in this paper can be summarized as follows:

- Formulating a coupled matrix factorization model with penalties to impose sparsity on factor matrices,
- Developing a gradient-based optimization algorithm for solving the smooth approximation of the coupled matrix factorization problem with sparsity penalties,
- Demonstrating the effectiveness of the proposed model/algorithm in terms of capturing the underlying sparse patterns in data using simulations,
- Identifying potential apple biomarkers based on joint analysis of metabolomics data sets collected on blood samples of a group of rats.

The rest of the paper is organized as follows. In Section II,

978-0-7695-4925-5/12 $26.00 © 2012 IEEE
DOI 10.1109/ICDMW.2012.17

we introduce our coupled matrix factorization model with penalties to impose sparsity and a gradient-based optimization algorithm for fitting the model. Section III demonstrates the performance of the proposed approach on both simulated and real data. In Section IV, we survey the related work, and, finally, conclude in Section V.

II. CMF-SPOPT

In this section, we first introduce our model for coupled matrix factorization (CMF) with penalty terms to enforce sparsity on the factor matrices and discuss the extension of the model to coupled analysis of incomplete data. We then present our algorithmic framework called CMF-SPOPT (Coupled Matrix Factorization with SParse OPTimization), which fits the proposed model using a gradient-based optimization method.

A. Model

We consider joint analysis of multiple matrices with one mode in common using coupled matrix factorization to capture the underlying sparse factors. We first discuss the formulation of coupled matrix factorization, which has previously been studied in various data fusion studies [2], [8], [9]. Without loss of generality, suppose matrices $\mathbf{X} \in \mathbb{R}^{I \times J}$ and $\mathbf{Y} \in \mathbb{R}^{I \times K}$ have the first mode in common. The objective function for their joint factorization can be formulated as:

$$f(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \left\| \mathbf{X} - \mathbf{A}\mathbf{B}^{\mathsf{T}} \right\|^2 + \left\| \mathbf{Y} - \mathbf{A}\mathbf{C}^{\mathsf{T}} \right\|^2 \quad (1)$$

where $||.||$ denotes the Frobenius norm for matrices and the 2-norm for vectors. The goal is to find the matrices $\mathbf{A} \in \mathbb{R}^{I \times R}$, $\mathbf{B} \in \mathbb{R}^{J \times R}$ and $\mathbf{C} \in \mathbb{R}^{K \times R}$ that minimize (1). Note that $\mathbf{A}$, i.e., the factor matrix extracted from the shared mode, is common in factorization of both $\mathbf{X}$ and $\mathbf{Y}$.

In this paper, we extend the formulation in (1) by adding penalty terms in order to impose sparsity on factor matrices $\mathbf{B}$ and $\mathbf{C}$, and reformulate the objective function as:

$$\begin{aligned} f(\mathbf{A}, \mathbf{B}, \mathbf{C}) &= \left\| \mathbf{X} - \mathbf{A}\mathbf{B}^{\mathsf{T}} \right\|^2 + \left\| \mathbf{Y} - \mathbf{A}\mathbf{C}^{\mathsf{T}} \right\|^2 \\ &+ \lambda \sum_{r=1}^{R} \| \mathbf{b}_r \|_1 + \lambda \sum_{r=1}^{R} \| \mathbf{c}_r \|_1 + \alpha \sum_{r=1}^{R} \| \mathbf{a}_r \|^2 \end{aligned} \quad (2)$$

where $\mathbf{b}_r$ and $\mathbf{c}_r$ correspond to the rth column of $\mathbf{B}$ and $\mathbf{C}$, respectively. $\| \mathbf{x} \|_1$ denotes 1-norm of a vector and is defined as $\sum |x_i|$. λ and α are penalty parameters with $\lambda, \alpha \geq 0$.

This formulation is motivated by metabolomics applications, where we often have different types of measurements on the same samples. For instance, $\mathbf{X}$ may correspond to a *samples* by *features* matrix constructed using LC-MS measurements while $\mathbf{Y}$ may be a matrix in the form of *samples* by *chemical shifts* constructed using NMR measurements. In most metabolomics applications, we need the underlying sparse patterns in variables dimensions, e.g., metabolites, in order to relate diseases or dietary interventions with a small set of variables. Therefore, we impose sparsity only in the variables modes by adding the 1-norm penalty, which has shown to be an effective way of enforcing sparsity [10]. The 2-norm penalty on the factors in the samples mode, i.e., the last term in (2), is added to handle the scaling ambiguity. Since there is a scaling ambiguity in the matrix factorization given above, i.e., $\hat{\mathbf{X}} = (\eta\mathbf{A})(\frac{1}{\eta}\mathbf{B}) = \mathbf{A}\mathbf{B}$, without penalizing the norm of the factors in the samples mode, the sparsity penalty would not have the desired effect.

1) Smooth Approximation: In order to minimize the objective function (2), we need to deal with a non-differentiable optimization problem due to the 1-norm terms. However, by replacing the 1-norm terms with differentiable approximations, it can be converted into a differentiable problem. Here, we approximate the terms with 1-norm using the "epsL1" function [11] and rewrite (2) as:

$$\begin{aligned} f(\mathbf{A}, \mathbf{B}, \mathbf{C}) &= \left\| \mathbf{X} - \mathbf{A}\mathbf{B}^{\mathsf{T}} \right\|^2 + \left\| \mathbf{Y} - \mathbf{A}\mathbf{C}^{\mathsf{T}} \right\|^2 \\ &+ \lambda \sum_{r=1}^{R} \sum_{j=1}^{J} \sqrt{b_{jr}^2 + \epsilon} + \lambda \sum_{r=1}^{R} \sum_{k=1}^{K} \sqrt{c_{kr}^2 + \epsilon} \\ &+ \alpha \sum_{r=1}^{R} \| \mathbf{a}_r \|^2 \end{aligned} \quad (3)$$

where b_{jr} denotes the entry in the jth row, rth column of $\mathbf{B}$. Note that, for sufficiently small $\epsilon > 0$, $\sqrt{x_i^2 + \epsilon} = |x_i|$.

2) Missing Data: In the presence of missing data, we can still jointly factorize matrices and extract sparse patterns by fitting the coupled model only to the known data entries. Suppose $\mathbf{X}$ has missing entries and let $\mathbf{W} \in \mathbb{R}^{I \times J}$ indicate the missing entries of $\mathbf{X}$ such that

$$w_{ij} = \begin{cases} 1 & \text{if } x_{ij} \text{ is known,} \\ 0 & \text{if } x_{ij} \text{ is missing,} \end{cases}$$

for all $i \in \{1, \ldots, I\}$ and $j \in \{1, \ldots, J\}$. To jointly analyze matrix $\mathbf{Y}$ and the incomplete matrix $\mathbf{X}$, we can then modify the objective function (3) as

$$\begin{aligned} f_{\mathbf{W}}(\mathbf{A}, \mathbf{B}, \mathbf{C}) &= \left\| \mathbf{W} * (\mathbf{X} - \mathbf{A}\mathbf{B}^{\mathsf{T}}) \right\|^2 + \left\| \mathbf{Y} - \mathbf{A}\mathbf{C}^{\mathsf{T}} \right\|^2 \\ &+ \lambda \sum_{r=1}^{R} \sum_{j=1}^{J} \sqrt{b_{jr}^2 + \epsilon} + \lambda \sum_{r=1}^{R} \sum_{k=1}^{K} \sqrt{c_{kr}^2 + \epsilon} \\ &+ \alpha \sum_{r=1}^{R} \| \mathbf{a}_r \|^2 \end{aligned} \quad (4)$$

where $*$ denotes the Hadamard (element-wise) product.

The formulations in (3) and (4) easily generalize to joint factorization of more than two matrices, each with

underlying sparse factors in the variables mode. In our objectives, we give equal weights to the factorization of each data matrix, and in the experiments, we divide each data set by its Frobenius norm so that the model does not favor one part of the objective. However, determining the right weighting scheme remains to be an open research question.

B. Algorithm

With the smooth approximation, we have obtained differentiable objective functions in (3) and (4), which can be solved using any first-order optimization algorithm [12]. In order to use a first-order optimization method, we only need to derive the gradient. The gradient of $f_{\mathbf{W}}$ in (4), which is a vector of size $P = R(I + J + K)$, can be formed by vectorizing the partial derivatives with respect to each factor matrix and concatenating them all, i.e.,

$$\nabla f_{\mathbf{W}} = \begin{bmatrix} \mathsf{vec}\left(\frac{\partial f_{\mathbf{W}}}{\partial \mathbf{A}}\right) \\ \mathsf{vec}\left(\frac{\partial f_{\mathbf{W}}}{\partial \mathbf{B}}\right) \\ \mathsf{vec}\left(\frac{\partial f_{\mathbf{W}}}{\partial \mathbf{C}}\right) \end{bmatrix}$$

Let $\mathbf{Z} = \mathbf{A}\mathbf{B}^{\mathsf{T}}$. Assuming each term of $f_{\mathbf{W}}$ in (4) is multiplied by $\frac{1}{2}$ for the ease of computation, the partial derivatives of $f_{\mathbf{W}}$ with respect to factor matrices, $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, can be computed as:

$$\frac{\partial f_{\mathbf{W}}}{\partial \mathbf{A}} = (\mathbf{W} * \mathbf{Z} - \mathbf{W} * \mathbf{X})\mathbf{B} - \mathbf{Y}\mathbf{C} + \mathbf{A}\mathbf{C}^{\mathsf{T}}\mathbf{C} + \alpha\mathbf{A}$$

$$\frac{\partial f_{\mathbf{W}}}{\partial \mathbf{B}} = (\mathbf{W} * \mathbf{Z} - \mathbf{W} * \mathbf{X})^{\mathsf{T}}\mathbf{A} + \frac{\lambda}{2}\mathbf{B}/(\mathbf{B} * \mathbf{B} + \epsilon)^{\frac{1}{2}}$$

$$\frac{\partial f_{\mathbf{W}}}{\partial \mathbf{C}} = -\mathbf{Y}^{\mathsf{T}}\mathbf{A} + \mathbf{C}\mathbf{A}^{\mathsf{T}}\mathbf{A} + \frac{\lambda}{2}\mathbf{C}/(\mathbf{C} * \mathbf{C} + \epsilon)^{\frac{1}{2}}$$

where the operator / denotes element-wise division.

Traditional approaches for coupled matrix factorizations are based on alternating algorithms [8], [9], where the optimization problem is solved for one factor matrix at a time by fixing the other factor matrices. While alternating algorithms are widely-used, direct nonlinear optimization methods solving for all factor matrices simultaneously have better convergence properties within the context of matrix factorizations with missing entries [13] and shown to be more accurate in the case of tensor factorizations [14]. Therefore, we use a gradient-based optimization algorithm to solve the non-convex optimization problem in (4). Neither alternating nor all-at-once approaches can guarantee to reach the global optimum. The computational cost *per iteration* is the same for both alternating and gradient-based approaches (See [13], [14] for in-depth comparison of alternating and all-at-once approaches).

Once the gradient, $\nabla f_{\mathbf{W}}$, is computed, we then use the Nonlinear Conjugate Gradient (NCG) method with Hestenes-Steifel updates [12] and the Moré-Thuente line search as implemented in the Poblano Toolbox [15].

III. Experiments and Results

In this section, performance of the proposed approach in terms of capturing the underlying sparse patterns in coupled data sets, is demonstrated using both simulated and real data.

A. Simulated Data

The goal of simulations is two-fold: (i) to demonstrate that underlying sparse factors used to generate coupled data sets can be accurately captured using the proposed model/algorithm (ii) to study the sensitivity of the proposed approach to different parameter values.

1) Experimental Set-up: We generate coupled matrices, $\mathbf{X} \in \mathbb{R}^{I \times J}$ and $\mathbf{Y} \in \mathbb{R}^{I \times K}$ computed as $\mathbf{X} = \mathbf{A}\mathbf{B}^{\mathsf{T}}$ and $\mathbf{Y} = \mathbf{A}\mathbf{C}^{\mathsf{T}}$, where $\mathbf{A} \in \mathbb{R}^{I \times R}$ has entries randomly drawn from the standard normal distribution; matrices $\mathbf{B} \in \mathbb{R}^{J \times R}$ and $\mathbf{C} \in \mathbb{R}^{K \times R}$, similarly, have entries randomly drawn from the standard normal but $S\%$ of the entries in each column of $\mathbf{B}$ and $\mathbf{C}$ is set to zero to have sparse factors. Columns of $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ are normalized to unit norm.

We then add noise to $\mathbf{X}$ and $\mathbf{Y}$ to form coupled noisy matrices, i.e., $\mathbf{X}_{noisy} = \mathbf{X} + \eta \frac{\mathbf{N}_1}{\| \mathbf{N}_1 \|} \| \mathbf{X} \|$ and $\mathbf{Y}_{noisy} = \mathbf{Y} + \eta \frac{\mathbf{N}_2}{\| \mathbf{N}_2 \|} \| \mathbf{Y} \|$, where entries of $\mathbf{N}_1 \in \mathbb{R}^{I \times J}$ and $\mathbf{N}_2 \in \mathbb{R}^{I \times K}$ are randomly drawn from the standard normal.

In order to assess the performance of CMF-SPOPT in terms of capturing the underlying sparse patterns, we generate data sets with (i) sparsity levels: $S = 30, 50, 70$, (ii) noise levels: $\eta = 0.1, 0.5$, and (iii) sizes: $(I, J, K) \in \{(20, 30, 40), (20, 300, 400), (20, 3000, 4000)\}$. We use $R = 2$ as the number of components.

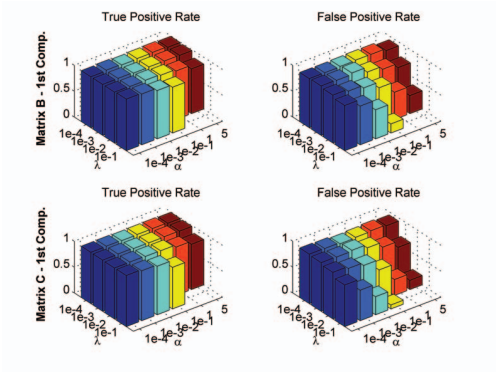
Once coupled matrices are generated, CMF-SPOPT is used to capture $\hat{\mathbf{A}} \in \mathbb{R}^{I \times R_{ext}}, \hat{\mathbf{B}} \in \mathbb{R}^{J \times R_{ext}}$ and $\hat{\mathbf{C}} \in \mathbb{R}^{K \times R_{ext}}$ for different values of penalty parameters: $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ and $\alpha \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 5\}$. R_{ext} indicates the number of extracted components.

We compare the extracted matrices $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$ with the original sparse matrices $\mathbf{B}$ and $\mathbf{C}$ used to generate the coupled data, in terms of sparsity patterns. For instance, the first column of $\mathbf{B}$, $\mathbf{b}_1$, is compared with the matching[1] column of $\hat{\mathbf{B}}$, e.g., $\hat{\mathbf{b}}_1$. If a nonzero in $\mathbf{b}_1$ corresponds to a nonzero in $\hat{\mathbf{b}}_1$, then it is a true-positive; if a zero in $\mathbf{b}_1$ corresponds to a nonzero in $\hat{\mathbf{b}}_1$, it is a false-positive.

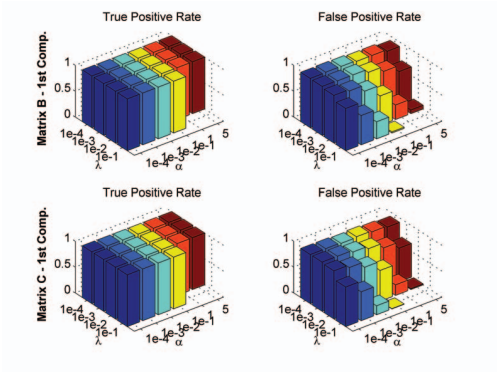
As stopping conditions, CMF-SPOPT uses the relative change in function value (set to 10^{-10}) and the 2-norm of the gradient divided by the number of entries in the gradient (set to 10^{-10}). For initialization, we use multiple random starts and choose the run with the minimum function value.

2) Results: CMF-SPOPT can capture the underlying sparse patterns accurately for varying levels of sparsity; in particular, the recovery is perfect for higher sparsity. We illustrate the performance of CMF-SPOPT in terms of

[1]Due to the permutation ambiguity, we look for the best permutation to match the columns.

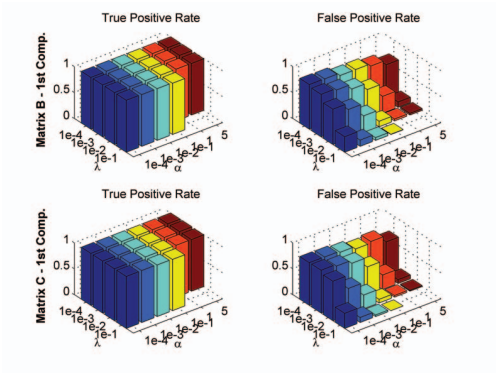


(a) $S = 30$

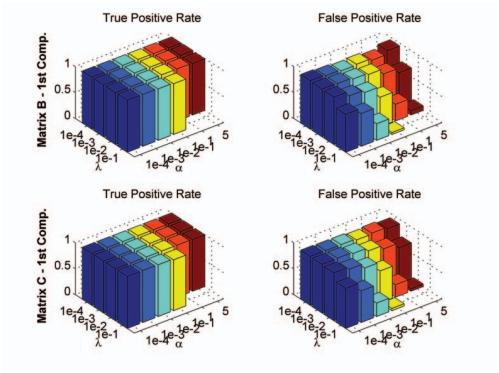


(b) $S = 70$

Figure 1. Performance of CMF-SPOPT for different levels of sparsity and different values of penalty parameters.



(a) $\eta = 0.1$



(b) $\eta = 0.5$

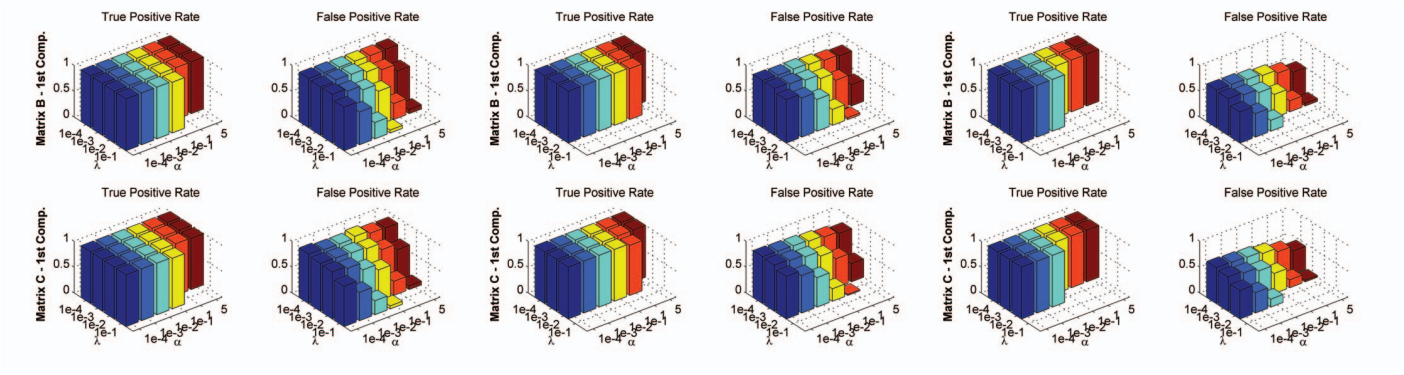
Figure 2. Performance of CMF-SPOPT for different levels of noise and different values of penalty parameters.

true-positive rates (TPR) and false-positive rates (FPR) for different sparsity levels in Figure 1. The best performance, i.e., exact recovery of the underlying sparsity patterns, corresponds to TPR=1 and FPR =0. The top and bottom rows of Figure 1(a) show the performance of CMF-SPOPT in terms of capturing the sparsity pattern of the first column of $\mathbf{B}$ and $\mathbf{C}$, respectively, for sparsity level $S = 30$. We observe that underlying patterns can be captured accurately but not perfectly as the best FPR values are around 0.1 - 0.2 with corresponding TPR values around 0.8-0.9. However, for higher sparsity, underlying sparsity patterns can be perfectly captured (Figure 1(b)). For all sparsity levels, the best performance is achieved for $\alpha = 0.1$ and $\lambda = 0.1$. Here, we set $(I, J, K) = (20, 30, 40)$, $\eta = 0.5$ and $R_{ext} = 2$, and present the average performance on 15 different sets of data.

CMF-SPOPT performs well in terms of capturing the underlying sparse patterns even at high amounts of noise. Figure 2 shows the performance of CMF-SPOPT at different noise levels. While TPR is high and FPR is low for low noise level, i.e., $\eta = 0.1$, with increasing noise we observe the degradation in performance. However, TPR is still high and FPR is low when $\eta = 0.5$. Here, we set $(I, J, K) = (20, 30, 40)$, $S = 50$, and $R_{ext} = 2$, and again report the average performance on 15 sets of data.

As we change data set sizes, best performing penalty parameters change drastically. Figure 3 shows the performance of CMF-SPOPT for varying sizes of coupled data sets for $S = 50, \eta = 0.5$, and $R_{ext} = 2$. We observe that for small number of dimensions in the variables mode, i.e., small values of J and K, $\alpha = 0.1$ and $\lambda = 0.1$ can accurately capture the sparse factors in $\mathbf{B}$ and $\mathbf{C}$. As J and K increase, though, higher α and lower λ values become effective.

We have only reported the results for the first component of $\mathbf{B}$ and $\mathbf{C}$. Results for the second component are similar and omitted here. Also note that matrix factorizations have



(a) $(I, J, K) = (20, 30, 40)$ (b) $(I, J, K) = (20, 300, 400)$ (c) $(I, J, K) = (20, 3000, 4000)$

Figure 3. Performance of CMF-SPOPT for varying sizes of coupled data sets and different values of penalty parameters.

Table I
PERFORMANCE OF CMF-SPOPT FOR $R_{ext} \in \{2, 3, 4\}$ WHEN $R = 2$.

	Component Weight ($\hat{\sigma}_r$)				Matrix B				Matrix C			
					Component 1		Component 2		Component 1		Component 2	
R_{ext}	1	2	3	4	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
2	0.98	0.92			0.99	0.07	0.98	0.04	0.99	0.04	0.98	0.02
3	0.98	0.92	0.00		0.99	0.07	0.97	0.04	0.99	0.04	0.98	0.02
4	0.98	0.92	0.00	0.00	0.99	0.07	0.98	0.04	0.99	0.04	0.98	0.02

rotational ambiguity; in other words, they can capture the factor matrices uniquely only up to a rotation. For certain combination of penalty values, the factor matrices, though, are uniquely captured by CMF-SPOPT, i.e., unique up to scaling and permutation. Reported TPR and FPR values correspond to those cases, where we can uniquely capture the factor matrices up to scaling and permutation.

Finally, we show that CMF-SPOPT is robust to the selection of the component number. Here, we generate data using $R = 2$ but fit the model using $R_{ext} \in \{2, 3, 4\}$. We set $\eta = 0.5$, $S = 50$, $\lambda = \alpha = 0.1$, $(I, J, K) = (20, 30, 40)$. Table I shows the weight of each *coupled* component, calculated as follows: We can rewrite $\mathbf{X} = \mathbf{A}\mathbf{B}^\mathsf{T}$ and $\mathbf{Y} = \mathbf{A}\mathbf{C}^\mathsf{T}$ as $\mathbf{X} = \sum_{r=1}^{R} \beta_r \mathbf{a}_r \mathbf{b}_r^\mathsf{T}$ and $\mathbf{Y} = \sum_{r=1}^{R} \gamma_r \mathbf{a}_r \mathbf{c}_r^\mathsf{T}$, where β_r and γ_r are the weights of component r in $\mathbf{X}$ and $\mathbf{Y}$, respectively, and $\| \mathbf{a}_r \| = \| \mathbf{b}_r \| = \| \mathbf{c}_r \| = 1$, for $r = 1, 2, ...R$. We define the weight of a coupled component r as $\sigma_r = \beta_r + \gamma_r$. Similarly, when $\hat{\mathbf{A}}, \hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$ are extracted using CMF-SPOPT, columns are normalized and $\hat{\sigma}_r$ is computed. Table I shows that when there are two common components, i.e., $R = 2$, and data sets are overfactored using $R_{ext} = 3, 4$, weights of the extra components are 0. Besides, sparsity patterns of common components are still accurately captured as indicated by high TPR and low FPR.

In summary, simulation studies demonstrate that CMF-SPOPT is quite effective in terms of capturing the underlying sparse patterns in coupled data; however, we also observe that the method is sensitive to penalty parameter values.

B. Metabolomics Data Analysis

Next, we use CMF-SPOPT to jointly analyze metabolomics data measured using different analytical techniques and identify potential markers for apple intake.

1) Data: The data consists of blood samples collected from a group of rats, which was part of a study on the effect of apple feeding on colon carcinogenesis [16]. Here, we use the samples from forty-six male Fisher 344 rats (5-8 weeks old) obtained from Charles River (Sulzfeld, Germany). After one week of adaptation on a purified diet, the animals were randomized to two experimental groups: fed either the same purified diet (group 1: Apple 0) or the purified diet added 10 g raw whole apple (group 2: Apple 10) for 13 weeks. At the end of the study, rats were sacrificed after an overnight fasting (16hrs). Animal experiments were carried out under the supervision of the Danish National Agency for Protection of Experimental Animals.

The rat plasma samples were analyzed by untargeted liquid chromatography - time-of-flight (LC-QTOF) mass spectrometry [17] and NMR [18]. In LC-MS analysis, raw data is converted into a feature set, where each feature is denoted by the mass over charge (m/z) ratio and a retention time (see [17] for details). In NMR analysis, the spectra were preprocessed (see [18] for details) and then converted into a set of peaks using an in-house automated peak detection algorithm. We also have a third data set containing Total

cholesterol (chol), low density cholesterol (LDL), very low density cholesterol (VLDL) and high density cholesterol (HDL) lipoproteins (computed based on the NMR data [18]) and triacylglycerol (TG) concentrations (measured using the rat plasma samples). In summary, our data can be represented using the following three matrices:

- $\mathbf{X} \in \mathbb{R}^{I \times J}$ of type *samples* by *features* corresponding to LC-MS data, where $I = 46$, and $J = 1086$.
- $\mathbf{Y} \in \mathbb{R}^{I \times K}$ of type *samples* by *chemical shifts* corresponding to NMR measurements, where $K = 115$.
- $\mathbf{Z} \in \mathbb{R}^{I \times M}$ of type *samples* by *quality variables* corresponding to quality measurements, where $M = 9$. Matrix $\mathbf{Z}$ has missing entries.

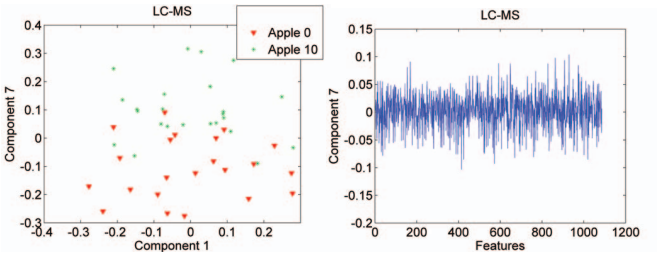
2) Model: Based on the formulation in (4), we jointly analyze $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ by minimizing the following objective:

$$
\begin{aligned}
&f_{\mathbf{W}}(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}) \\
&= \left\| \mathbf{X} - \mathbf{A}\mathbf{B}^{\mathsf{T}} \right\|^2 + \left\| \mathbf{Y} - \mathbf{A}\mathbf{C}^{\mathsf{T}} \right\|^2 + \left\| \mathbf{W} * (\mathbf{Z} - \mathbf{A}\mathbf{D}^{\mathsf{T}}) \right\|^2 \\
&+ \lambda \sum_{r=1}^{R} \sum_{j=1}^{J} \sqrt{b_{jr}^2 + \epsilon} + \lambda \sum_{r=1}^{R} \sum_{k=1}^{K} \sqrt{c_{kr}^2 + \epsilon} \\
&+ \lambda \sum_{r=1}^{R} \sum_{m=1}^{M} \sqrt{d_{mr}^2 + \epsilon} + \alpha \sum_{r=1}^{R} \| \mathbf{a}_r \|^2
\end{aligned}
$$

and extract the factor matrices $\mathbf{A} \in \mathbb{R}^{I \times R}$, $\mathbf{B} \in \mathbb{R}^{J \times R}$, $\mathbf{C} \in \mathbb{R}^{K \times R}$ and $\mathbf{D} \in \mathbb{R}^{M \times R}$ corresponding to the samples, features, chemical shifts and quality variables, respectively. Using simulation data of similar sizes (with sparsity levels of $S = 50$ and $S = 70$), best performing penalty parameter values are determined as $\lambda = 0.01$ and $\alpha = 0.1$.

3) Results: Before discussing the sparse patterns captured using CMF-SPOPT, we first illustrate the factors extracted using the Singular Value Decomposition (SVD) of matrix $\mathbf{X}$. SVD decomposes $\mathbf{X}$ as $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{T}}$, where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices corresponding to the left and right singular vectors, respectively, and $\mathbf{\Sigma}$ is a diagonal matrix with singular values on the diagonal. Figure 4(a) shows the scatter plot of $\mathbf{u}_1$ and $\mathbf{u}_7$ demonstrating that two apple groups can be almost separated using the *seventh* left singular vector. The goal in metabolomics studies is often to understand the reason for the separation; in other words, the metabolites responsible for the separation. Therefore, we plot the *seventh* right singular vector in Figure 4(b) to identify the significant features. However, capturing the significant features is difficult since this vector is dense.

In Figure 5, we illustrate the performance of CMF-SPOPT in terms of apple group separation by coupled analysis of $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$. The scatter plot of $\mathbf{a}_1$ vs. $\mathbf{a}_3$ in Figure 5(a) shows that the first component can almost separate the two groups. In Figure 5, we can see the sparse patterns, i.e., $\mathbf{b}_1$, $\mathbf{c}_1$ and $\mathbf{d}_1$, responsible for this separation. Unlike Figure 4(b), we can clearly identify the significant features



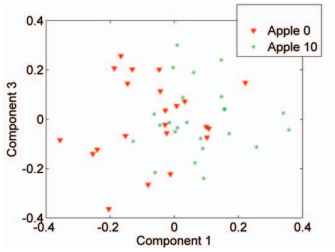
(a) Scatter plot of $\mathbf{u}_1$ vs. $\mathbf{u}_7$. (b) The right singular vector $\mathbf{v}_7$.

Figure 4. Separation of apple groups using SVD of $\mathbf{X}$.

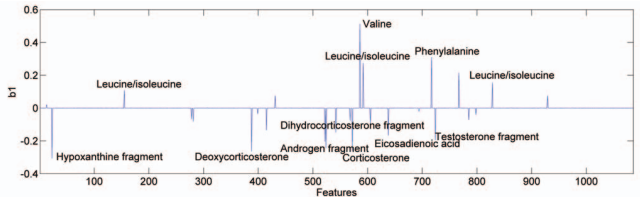
in Figure 5(b). Through coupled analysis, we also get the sparse patterns relevant to apple groups in each data set. Results illustrated in Figure 5 are based on a 5-component CMF-SPOPT model, i.e., $R = 5$. If we decrease R, none of the components can separate the apple groups. For $R = 5$ and $R = 6$, we get almost the exact same component for apple separation. As R increases, we lose the component responsible for the separation. In real data, unlike simulation studies, there are both common and uncommon components in coupled data sets; therefore, robustness of CMF-SPOPT to overfactoring (shown in Table I) is not enough to deal with the problem of determining R. In this study, since we are interested in apple markers, we use a component number that captures the apple group separation.

In order to make sure that the sparse pattern in Figure 5(b) is really meaningful, we form a small matrix, $\bar{\mathbf{X}} \in \mathbb{R}^{I \times L}$, using only the features identified in $\mathbf{b}_1$, where $L = 14$, and check the separation achieved by its SVD. We observe that using only 14 out of 1086 features, we can still separate the apple groups (results not shown but separation is similar to Figure 5(a)); therefore, these features are potential candidates for markers of apple intake.

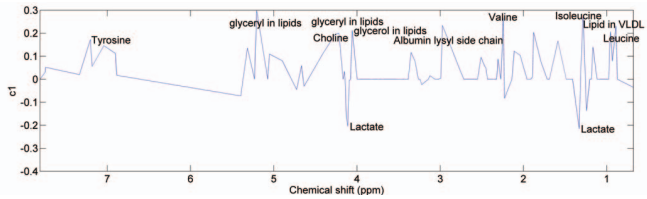
We further study the sparse patterns captured by CMF-SPOPT from a biological perspective. Metabolites identified in the sparse patterns are shown in Figure 5(b) and Figure 5(c). Some of these have been verified by chemical standards while some of them are tentative identifications, further to be explored. Based on the identifications, we find that patterns in Figure 5 are related to apple-induced changes in the endogenous metabolism. These changes include an increase in circulating branched-chain and aromatic amino acids, an increase in circulating glycerol- and choline containing lipids, a decrease in corticosteroids and possibly in androgens, and a decrease in lactate, hypoxanthin and free fatty acids. Several elements of this pattern indicate that the transition from the postprandial to the fasting state was delayed in apple-fed rats, with a slower increase in lactate and free fatty acids and a slower loss of amino acids and lipids from the blood. Moreover, apple feeding seems to



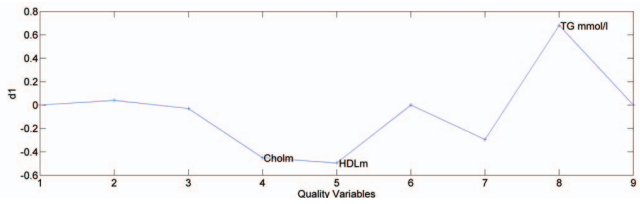
(a) Scatter plot of $\mathbf{a}_1$ vs. $\mathbf{a}_3$.



(b) Sparse pattern $\mathbf{b}_1$ extracted from LC-MS.



(c) Sparse pattern $\mathbf{c}_1$ extracted from NMR.



(d) Sparse pattern $\mathbf{d}_1$ extracted from Quality Measurements.

Figure 5. Separation of apple groups using CMF-SPOPT.

suppress the increase in corticosteroids and possibly also of androgens with fasting in support of an effect on genes involved in steroid metabolism that we have observed in these rats.

Using CMF-SPOPT, we were able to extract meaningful sparse patterns from LC-MS and NMR complementing each other and describing apple-induced changes in the metabolism.

IV. Related Work

Simultaneous analysis of multiple matrices dates back to one of the earliest models aiming to capture the common variation in data sets, i.e., Canonical Correlation Analysis (CCA) [19]. CCA looks for the patterns in each data set that correlate well and it is, in that sense, different from coupled matrix factorization. This difference has been illustrated in a recent metabolomics study [20].

More in line with the formulation in (1), Levin [21] studied simultaneous factorization of Gramian matrices. Similarly, in signal processing, joint diagonalization of symmetric and Hermitian matrices has been a topic of interest [22]. Furthermore, principal component analysis of multiple matrices has been widely studied in chemometrics using various models, some with clear objective functions while some are based on heuristic multi-level approaches [5]. Badea [23] extended the formulation in (1) to simultaneous nonnegative matrix factorizations by extracting nonnegative factor matrices. Another line of work related to simultaneous matrix factorization is Generalized SVD and its extention to multiple matrices [24].

With the increasing interest in the analysis of multi-relational data, Singh and Gordon [8] and Long et al. [7] studied Collective Matrix Factorization for joint factorization of matrices. We can also consider tensor factorizations as simultaneous factorization of multiple matrices (see a recent survey for various tensor models [25]).

While coupled matrix factorization has been widely studied in many disciplines, a recent study by Deun et al. [26] is the only study that enforces sparsity on the factors within the coupled matrix factorization framework, to the best of our knowledge. This work considers various penalty schemes such as the lasso, elastic net, group lasso, etc., and it is the most related to what we propose, or more specificially to (2). The main differences are (i) we do not enforce orthogonality constraints on factor matrix $\mathbf{A}$, as in [26], (ii) while alternating least squares is used in [26], we use an all-at-once approach solving a smooth approximation of the objective in (2), and (iii) we extend our formulation to joint analysis of incomplete data as in (4).

V. Conclusions

While we can collect huge amounts of data using different platforms in metabolomics, we are still lacking the data mining tools for the fusion and analysis of these data sets. In this paper, we have formulated data fusion as a coupled matrix factorization model with penalties to enforce sparsity with a goal of capturing the underlying sparse patterns in coupled data sets. We have also discussed the extension of the proposed model to coupled analysis of incomplete data. In order to fit the model to coupled data sets, we have developed a gradient-based optimization algorithm solving for all factor matrices simultaneously. Using numerical experiments on simulated data, effectiveness of the proposed

approach in terms of capturing the underlying sparse patterns is demonstrated. We have also illustrated the usefulness of the proposed method in a metabolomics application, where potential markers for apple intake are identified through coupled analysis of LC-MS and NMR data. The main limitation of our formulation is to impose the same level of sparsity on different data sets. We plan to extend our model to different levels of sparsity in coupled data sets; in other words, to use different λ values for different matrices. This may require reformulation of the model in order to deal with the scaling ambiguity problem.

VI. Acknowledgments

This work is funded by the Danish Council for Independent Research — Technology and Production Sciences and Sapere Aude Program under the projects 11-116328 and 11-120947.

References

[1] S. E. Richards, M.-E. Dumas, J. M. Fonville, T. M. Ebbels, E. Holmes, and J. K. Nicholson, "Intra- and inter-omic fusion of metabolic profiling data in a systems biology framework," *Chemometr Intell Lab*, vol. 104, pp. 121–131, 2010.

[2] H. Ma, H. Yang, M. R. Lyu, and I. King, "Sorec: Social recommendation using probabilistic matrix factorization," in *Proc. CIKM'08*, 2008, pp. 931–940.

[3] A. Banerjee, S. Basu, and S. Merugu, "Multi-way clustering on relation graphs," in *Proc. SDM'07*, 2007, pp. 145–156.

[4] Y.-R. Lin, J. Sun, P. Castro, R. Konuru, H. Sundaram, and A. Kelliher, "Metafac: community discovery via relational hypergraph factorization," in *Proc. KDD'09*, 2009, pp. 527–536.

[5] A. K. Smilde, J. A. Westerhuis, and S. de Jong, "A framework for sequential multiblock component methods," *J Chemometr*, vol. 17, pp. 323–337, 2003.

[6] T. G. Doeswijk, A. K. Smilde, J. A. Hageman, J. A. Westerhuis, and F. A. van Eeuwijk, "On the increase of predictive performance with high-level data fusion," *Anal Chim Acta*, vol. 705, pp. 41–47, 2011.

[7] B. Long, Z. Zhang, X. Wu, and P. S. Yu, "Spectral clustering for multi-type relational data," in *Proc. ICML'06*, 2006, pp. 585–592.

[8] A. P. Singh and G. J. Gordon, "Relational learning via collective matrix factorization," in *Proc. KDD'08*, 2008, pp. 650–658.

[9] K. van Deun, A. K. Smilde, M. J. van der Werf, H. A. Kiers, and I. van Mechelen, "A structured overview of simultaneous component based data integration," *BMC Bioinformatics*, vol. 10, p. 246, 2009.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J R Stat Soc. Series B (Methodological)*, vol. 58, pp. 267–288, 1996.

[11] S. Lee, H. Lee, P. Abbeel, and A. Y. Ng, "Efficient l1 regularized logistic regression," in *Proc. AAAI'06*, 2006, pp. 401–408.

[12] J. Nocedal and S. J. Wright, *Numerical Optimization*. Springer, 2006.

[13] A. M. Buchanan and A. W. Fitzgibbon, "Damped newton algorithms for matrix factorization with missing data," in *Proc. CVPR'05*, vol. 2, 2005, pp. 316–322.

[14] E. Acar, D. M. Dunlavy, and T. G. Kolda, "A scalable optimization approach for fitting canonical tensor decompositions," *J Chemometr*, vol. 25, pp. 67–86, 2011.

[15] D. M. Dunlavy, T. G. Kolda, and E. Acar, "Poblano v1.0: A Matlab toolbox for gradient-based optimization," Sandia National Laboratories, Tech. Rep. SAND2010-1422, 2010.

[16] M. Poulsen, A. Mortensen, M. L. Binderup, S. Langkilde, J. Markowski, and L. O. Dragsted, "The effect of apple feeding on markers of colon carcinogenesis," *Nutr Cancer*, vol. 63, pp. 402–409, 2011.

[17] G. Gürdeniz, M. Kristensen, T. Skov, and L. O. Dragsted, "The effect of lc-ms data preprocessing methods on the selection of plasma biomarkers in fed vs. fasted rats," *Metabolites*, vol. 2, pp. 77–99, 2012.

[18] M. Kristensen, F. Savorani, G. Ravn-Haren, M. Poulsen, J. Markowski, F. H. Larsen, L. O. Dragsted, and S. B. Engelsen, "Nmr and ipls are reliable methods for determination of cholesterol in rodent lipoprotein fractions," *Metabolomics*, vol. 6, pp. 129–136, 2010.

[19] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1936.

[20] R. A. van den Berg, C. M. Rubingh, J. A. Westerhuis, M. J. van der Werf, and A. K. Smilde, "Metabolomics data exploration guided by prior knowledge," *Anal Chim Acta*, vol. 651, pp. 173–181, 2009.

[21] J. Levin, "Simultaneous factor analysis of several gramian matrices," *Psychometrika*, vol. 31, pp. 413–419, 1966.

[22] A. Yeredor, "Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation," *IEEE T Signal Proces*, vol. 50, pp. 1545–1553, 2002.

[23] L. Badea, "Extracting gene expression profiles common to colon and pancreatic adenocarcinoma using simultaneous nonnegative matrix factorization," in *Pacific Symposium on Biocomputing*, vol. 13, 2008, pp. 279–290.

[24] S. P. Ponnapalli, M. A. Saunders, C. F. V. Loan, and O. Alter, "A higher-order generalized singular value decomposition for comparison of global mrna expression from multiple organisms," *PLoS One*, vol. 6, p. e28072, 2011.

[25] E. Acar and B. Yener, "Unsupervised multiway data analysis: A literature survey," *IEEE T Knowl Data En*, vol. 21, no. 1, pp. 6–20, 2009.

[26] K. van Deun, T. F. Wilderjans, R. A. van den Berg, A. Antoniadis, and I. van Mechelen, "A flexible framework for sparse simultaneous component based data integration," *BMC Bioinformatics*, vol. 12, p. 448, 2011.